

SECTION IV

Structure, Function, & Replication of Informational Macromolecules

Nucleotides

33

Victor W. Rodwell, PhD

BIOMEDICAL IMPORTANCE

Nucleotides—the monomer units or building blocks of nucleic acids—serve multiple additional functions. They form a part of many coenzymes and serve as donors of phosphoryl groups (eg, ATP or GTP), of sugars (eg, UDP- or GDP-sugars), or of lipid (eg, CDP-acylglycerol). Regulatory nucleotides include the second messengers cAMP and cGMP, the control by ADP of oxidative phosphorylation, and allosteric regulation of enzyme activity by ATP, AMP, and CTP. Synthetic purine and pyrimidine analogs that contain halogens, thiols, or additional nitrogen are employed for chemotherapy of cancer and AIDS and as suppressors of the immune response during organ transplantation.

PURINES, PYRIMIDINES, NUCLEOSIDES, & NUCLEOTIDES

Purines and pyrimidines are nitrogen-containing heterocycles, cyclic compounds whose rings contain both carbon and other elements (hetero atoms). Note that the smaller pyrimidine has the *longer* name and the larger purine the *shorter* name and that their six-atom rings are numbered in opposite directions (Figure 33–1). The planar character of purines and pyrimidines facilitates their close association, or “stacking,” which stabilizes double-stranded DNA (Chapter 36). The oxo and amino groups of purines and pyrimidines exhibit keto-enol and amine-imine tautomerism (Figure 33–2), but physiologic conditions strongly favor the amino and oxo forms.

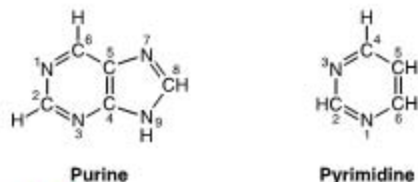


Figure 33–1. Purine and pyrimidine. The atoms are numbered according to the international system.

Nucleosides & Nucleotides

Nucleosides are derivatives of purines and pyrimidines that have a sugar linked to a ring nitrogen. Numerals with a prime (eg, 2' or 3') distinguish atoms of the sugar from those of the heterocyclic base. The sugar in **ribonucleosides** is D-ribose, and in **deoxyribonucleosides** it is 2-deoxy-D-ribose. The sugar is linked to the heterocyclic base via a **β -N-glycosidic bond**, almost always to N-1 of a pyrimidine or to N-9 of a purine (Figure 33–3).



Figure 33–2. Tautomerism of the oxo and amino functional groups of purines and pyrimidines.

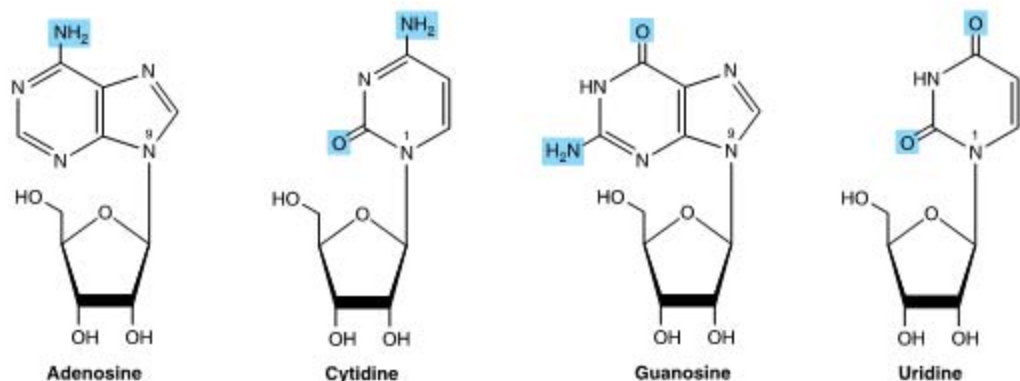


Figure 33-3. Ribonucleosides, drawn as the syn conformers.

Mononucleotides are nucleosides with a phosphoryl group esterified to a hydroxyl group of the sugar. The 3'- and 5'-nucleotides are nucleosides with a phosphoryl group on the 3'- or 5'-hydroxyl group of the sugar, respectively. Since most nucleotides are 5', the prefix "5'-" is usually omitted when naming them. UMP and dAMP thus represent nucleotides with a phosphoryl group on C-5 of the pentose. Additional phosphoryl groups linked by **acid anhydride bonds** to the phosphoryl group of a mononucleotide form nucleoside **diphosphates** and **triphosphates** (Figure 33-4).

Steric hindrance by the base restricts rotation about the β -N-glycosidic bond of nucleosides and nu-

cleotides. Both therefore exist as syn or anti conformers (Figure 33-5). While both conformers occur in nature, anti conformers predominate. Table 33-1 lists the major purines and pyrimidines and their nucleoside and nucleotide derivatives. Single-letter abbreviations are used to identify adenine (A), guanine (G), cytosine (C), thymine (T), and uracil (U), whether free or present in nucleosides or nucleotides. The prefix "d" (deoxy) indicates that the sugar is 2'-deoxy-D-ribose (eg, dGTP) (Figure 33-6).

Nucleic Acids Also Contain Additional Bases

Small quantities of additional purines and pyrimidines occur in DNA and RNAs. Examples include 5-methylcytosine of bacterial and human DNA, 5-hydroxymethylcytosine of bacterial and viral nucleic acids, and mono- and di-N-methylated adenine and guanine of

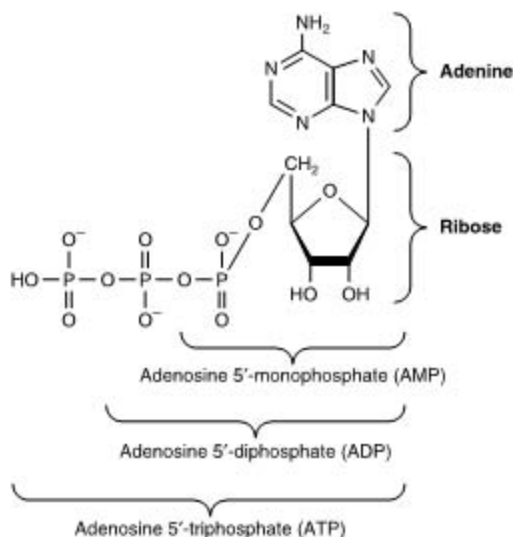


Figure 33-4. ATP, its diphosphate, and its monophosphate.

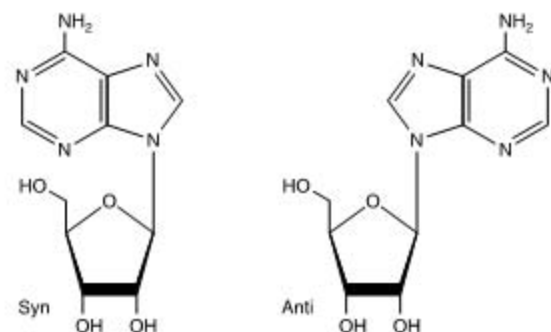
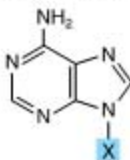
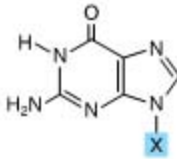
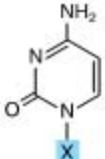
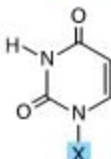
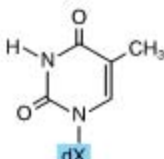
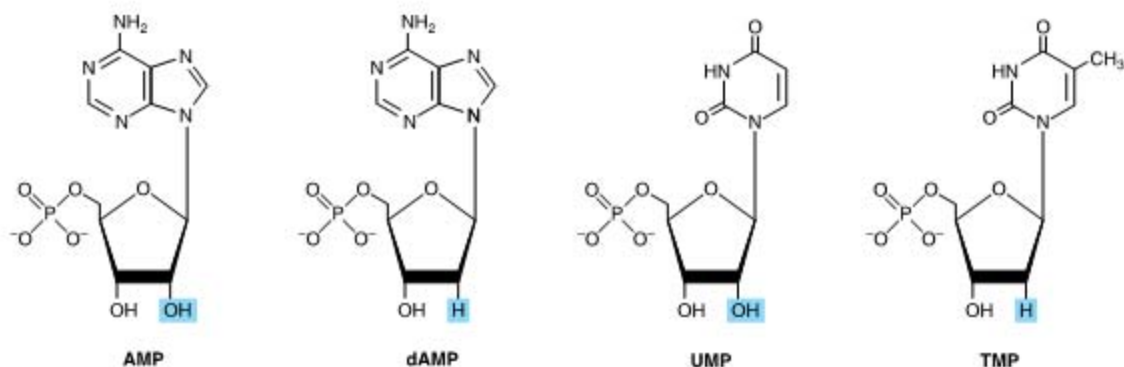


Figure 33-5. The syn and anti conformers of adenosine differ with respect to orientation about the N-glycosidic bond.

Table 33–1. Bases, nucleosides, and nucleotides.

Base Formula	Base X = H	Nucleoside X = Ribose or Deoxyribose	Nucleotide, Where X = Ribose Phosphate
	Adenine A	Adenosine A	Adenosine monophosphate AMP
	Guanine G	Guanosine G	Guanosine monophosphate GMP
	Cytosine C	Cytidine C	Cytidine monophosphate CMP
	Uracil U	Uridine U	Uridine monophosphate UMP
	Thymine T	Thymidine T	Thymidine monophosphate TMP

**Figure 33–6.** AMP, dAMP, UMP, and TMP.

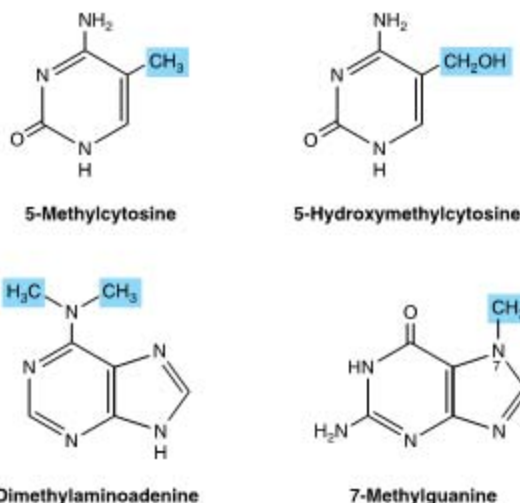


Figure 33-7. Four uncommon naturally occurring pyrimidines and purines.

mammalian messenger RNAs (Figure 33-7). These atypical bases function in oligonucleotide recognition and in regulating the half-lives of RNAs. Free nucleotides include hypoxanthine, xanthine, and uric acid (see Figure 34-8), intermediates in the catabolism of adenine and guanine. Methylated heterocyclic bases of plants include the xanthine derivatives caffeine of coffee, theophylline of tea, and theobromine of cocoa (Figure 33-8).

Posttranslational modification of preformed polynucleotides can generate additional bases such as pseudouridine, in which D-ribose is linked to C-5 of uracil by a carbon-to-carbon bond rather than by a β -N-glycosidic bond. The nucleotide pseudouridylic acid Ψ arises by rearrangement of UMP of a preformed tRNA. Similarly, methylation by S-adenosylmethionine of a UMP of preformed tRNA forms TMP (thymidine monophosphate), which contains ribose rather than deoxyribose.

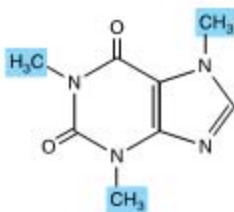


Figure 33-8. Caffeine, a trimethylxanthine. The dimethylxanthines theobromine and theophylline are similar but lack the methyl group at N-1 and at N-7, respectively.

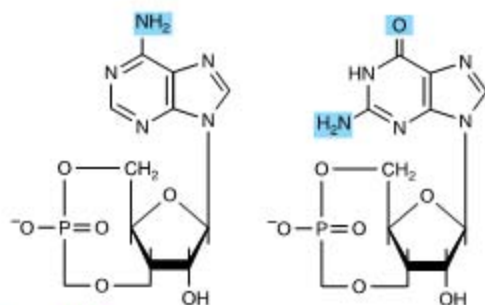


Figure 33-9. cAMP, 3',5'-cyclic AMP, and cGMP.

Nucleotides Serve Diverse Physiologic Functions

Nucleotides participate in reactions that fulfill physiologic functions as diverse as protein synthesis, nucleic acid synthesis, regulatory cascades, and signal transduction pathways.

Nucleoside Triphosphates Have High Group Transfer Potential

Acid anhydrides, unlike phosphate esters, have high group transfer potential. $\Delta G'$ for the hydrolysis of each of the terminal phosphates of nucleoside triphosphates is about -7 kcal/mol (-30 kJ/mol). The high group transfer potential of purine and pyrimidine nucleoside triphosphates permits them to function as group transfer reagents. Cleavage of an acid anhydride bond typically is coupled with a highly endergonic process such as covalent bond synthesis—eg, polymerization of nucleoside triphosphates to form a nucleic acid.

In addition to their roles as precursors of nucleic acids, ATP, GTP, UTP, CTP, and their derivatives each serve unique physiologic functions discussed in other chapters. Selected examples include the role of ATP as the principal biologic transducer of free energy; the second messenger cAMP (Figure 33-9); adenosine 3'-phosphate-5'-phosphosulfate (Figure 33-10), the sulfate donor for sulfated proteoglycans (Chapter 48) and for sulfate conjugates of drugs; and the methyl group donor S-adenosylmethionine (Figure 33-11).

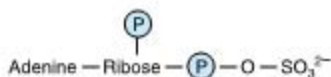


Figure 33-10. Adenosine 3'-phosphate-5'-phosphosulfate.

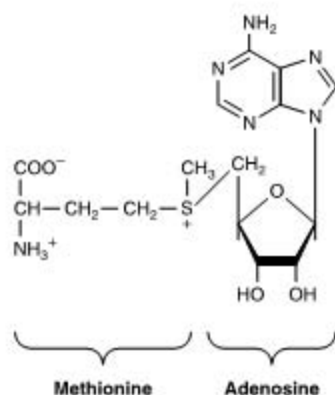


Figure 33–11. S-Adenosylmethionine.

GTP serves as an allosteric regulator and as an energy source for protein synthesis, and cGMP (Figure 33–9) serves as a second messenger in response to nitric oxide (NO) during relaxation of smooth muscle (Chapter 48). UDP-sugar derivatives participate in sugar epimerizations and in biosynthesis of glycogen, glucosyl disaccharides, and the oligosaccharides of glycoproteins and proteoglycans (Chapters 47 and 48). UDP-glucuronic acid forms the urinary glucuronide conjugates of bilirubin (Chapter 32) and of drugs such as aspirin. CTP participates in biosynthesis of phosphoglycerides, sphingomyelin, and other substituted sphingosines (Chapter 24). Finally, many coenzymes incorporate nucleotides as well as structures similar to purine and pyrimidine nucleotides (see Table 33–2).

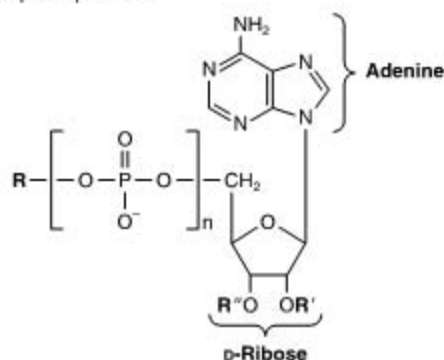
Nucleotides Are Polyfunctional Acids

Nucleosides or free purine or pyrimidine bases are uncharged at physiologic pH. By contrast, the primary phosphoryl groups (pK about 1.0) and secondary phosphoryl groups (pK about 6.2) of nucleotides ensure that they bear a negative charge at physiologic pH. Nucleotides can, however, act as proton donors or acceptors at pH values two or more units above or below neutrality.

Nucleotides Absorb Ultraviolet Light

The conjugated double bonds of purine and pyrimidine derivatives absorb ultraviolet light. The mutagenic effect of ultraviolet light results from its absorption by nucleotides in DNA with accompanying chemical changes. While spectra are pH-dependent, at pH 7.0 all the common nucleotides absorb light at a wavelength close to 260 nm. The concentration of nucleotides and

Table 33–2. Many coenzymes and related compounds are derivatives of adenosine monophosphate.



Coenzyme	R	R'	R''	n
Active methionine	Methionine*	H	H	0
Amino acid adenylates	Amino acid	H	H	1
Active sulfate	SO ₃ ²⁻	H	PO ₃ ²⁻	1
3',5'-Cyclic AMP		H	PO ₃ ²⁻	1
NAD*	†	H	H	2
NADP*	†	PO ₃ ²⁻	H	2
FAD	†	H	H	2
CoASH	†	H	PO ₃ ²⁻	2

*Replaces phosphoryl group.

†R is a B vitamin derivative.

nucleic acids thus often is expressed in terms of “absorbance at 260 nm.”

SYNTHETIC NUCLEOTIDE ANALOGS ARE USED IN CHEMOTHERAPY

Synthetic analogs of purines, pyrimidines, nucleosides, and nucleotides altered in either the heterocyclic ring or the sugar moiety have numerous applications in clinical medicine. Their toxic effects reflect either inhibition of enzymes essential for nucleic acid synthesis or their incorporation into nucleic acids with resulting disruption of base-pairing. Oncologists employ 5-fluoro- or 5-iodouracil, 3-deoxyuridine, 6-thioguanine and 6-mercaptopurine, 5- or 6-azauridine, 5- or 6-azacytidine, and 8-azaguanine (Figure 33–12), which are incorporated into DNA prior to cell division. The purine analog allopurinol, used in treatment of hyperuricemia and gout, inhibits purine biosynthesis and xanthine oxidase activity. Cytarabine is used in chemotherapy of cancer. Finally, azathioprine, which is catabolized to 6-mercaptopurine, is employed during organ transplantation to suppress immunologic rejection.

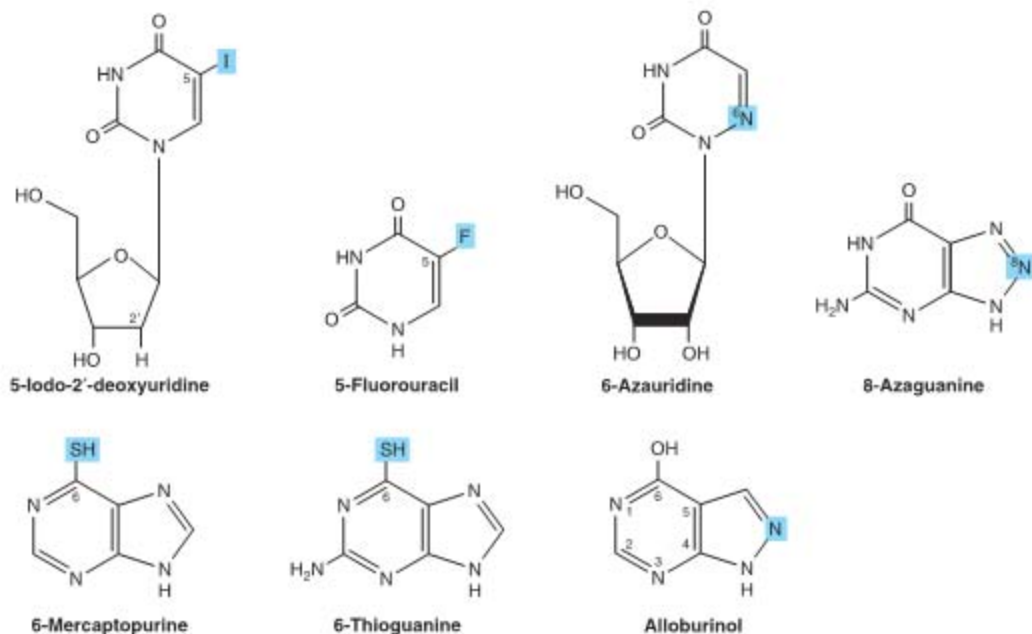


Figure 33-12. Selected synthetic pyrimidine and purine analogs.

Nonhydrolyzable Nucleoside Triphosphate Analogs Serve as Research Tools

Synthetic nonhydrolyzable analogs of nucleoside triphosphates (Figure 33-13) allow investigators to distinguish the effects of nucleotides due to phosphoryl transfer from effects mediated by occupancy of allosteric nucleotide-binding sites on regulated enzymes.

POLYNUCLEOTIDES

The 5'-phosphoryl group of a mononucleotide can esterify a second —OH group, forming a **phosphodiester**. Most commonly, this second —OH group is the 3'-OH of the pentose of a second nucleotide. This forms a **dinucleotide** in which the pentose moieties are linked by a 3' → 5' phosphodiester bond to form the “backbone” of RNA and DNA.

While formation of a dinucleotide may be represented as the elimination of water between two monomers, the reaction in fact strongly favors phosphodiester hydrolysis. **Phosphodiesterases** rapidly catalyze the hydrolysis of phosphodiester bonds whose spontaneous hydrolysis is an extremely slow process. Consequently, DNA persists for considerable periods and has been detected even in fossils. RNAs are far less stable than DNA since the 2'-hydroxyl group of RNA

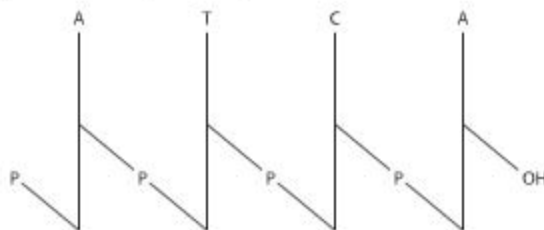
(absent from DNA) functions as a nucleophile during hydrolysis of the 3',5'-phosphodiester bond.

Polynucleotides Are Directional Macromolecules

Phosphodiester bonds link the 3'- and 5'-carbons of adjacent monomers. Each end of a nucleotide polymer thus is distinct. We therefore refer to the “5'-end” or the “3'-end” of polynucleotides, the 5'-end being the one with a free or phosphorylated 5'-hydroxyl.

Polynucleotides Have Primary Structure

The base sequence or **primary structure** of a polynucleotide can be represented as shown below. The phosphodiester bond is represented by P or p, bases by a single letter, and pentoses by a vertical line.



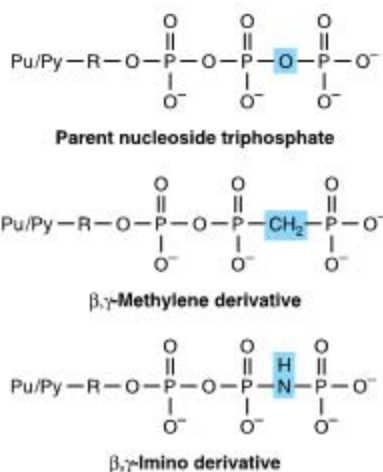


Figure 33-13. Synthetic derivatives of nucleoside triphosphates incapable of undergoing hydrolytic release of the terminal phosphoryl group. (Pu/Py, a purine or pyrimidine base; R, ribose or deoxyribose.) Shown are the parent (hydrolyzable) nucleoside triphosphate (**top**) and the unhydrolyzable β -methylene (**center**) and γ -imino derivatives (**bottom**).

Where all the phosphodiester bonds are $5' \rightarrow 3'$, a more compact notation is possible:



This representation indicates that the $5'$ -hydroxyl—but not the $3'$ -hydroxyl—is phosphorylated.

The most compact representation shows only the base sequence with the $5'$ -end on the left and the $3'$ -end on the right. The phosphoryl groups are assumed but not shown:



SUMMARY

- Under physiologic conditions, the amino and oxo tautomers of purines, pyrimidines, and their derivatives predominate.
- Nucleic acids contain, in addition to A, G, C, T, and U, traces of 5-methylcytosine, 5-hydroxymethylcytosine, pseudouridine (Ψ), or N-methylated bases.
- Most nucleosides contain D-ribose or 2-deoxy-D-ribose linked to N-1 of a pyrimidine or to N-9 of a purine by a β -glycosidic bond whose syn conformers predominate.
- A primed numeral locates the position of the phosphate on the sugars of mononucleotides (eg, $3'$ -GMP, $5'$ -dCMP). Additional phosphoryl groups linked to the first by acid anhydride bonds form nucleoside diphosphates and triphosphates.
- Nucleoside triphosphates have high group transfer potential and participate in covalent bond syntheses. The cyclic phosphodiester cAMP and cGMP function as intracellular second messengers.
- Mononucleotides linked by $3' \rightarrow 5'$ -phosphodiester bonds form polynucleotides, directional macromolecules with distinct $3'$ - and $5'$ -ends. For pTpGpTp or TGCATCA, the $5'$ -end is at the left, and all phosphodiester bonds are $3' \rightarrow 5'$.
- Synthetic analogs of purine and pyrimidine bases and their derivatives serve as anticancer drugs either by inhibiting an enzyme of nucleotide biosynthesis or by being incorporated into DNA or RNA.

REFERENCES

- Adams RLP, Knowler JT, Leader DP: *The Biochemistry of the Nucleic Acids*, 11th ed. Chapman & Hall, 1992.
- Blackburn GM, Gait MJ: *Nucleic Acids in Chemistry & Biology*. IRL Press, 1990.
- Bugg CE, Carson WM, Montgomery JA: Drugs by design. *Sci Am* 1992;269(6):92.

Metabolism of Purine & Pyrimidine Nucleotides

34

Victor W. Rodwell, PhD

BIOMEDICAL IMPORTANCE

The biosynthesis of purines and pyrimidines is stringently regulated and coordinated by feedback mechanisms that ensure their production in quantities and at times appropriate to varying physiologic demand. Genetic diseases of purine metabolism include gout, Lesch-Nyhan syndrome, adenosine deaminase deficiency, and purine nucleoside phosphorylase deficiency. By contrast, apart from the orotic acidurias, there are few clinically significant disorders of pyrimidine catabolism.

PURINES & PYRIMIDINES ARE DIETARILY NONESSENTIAL

Human tissues can synthesize purines and pyrimidines from amphibolic intermediates. Ingested nucleic acids and nucleotides, which therefore are dietarily nonessential, are degraded in the intestinal tract to mononucleotides, which may be absorbed or converted to purine and pyrimidine bases. The purine bases are then oxidized to uric acid, which may be absorbed and excreted in the urine. While little or no dietary purine or pyrimidine is incorporated into tissue nucleic acids, injected compounds are incorporated. The incorporation of injected [^3H]thymidine into newly synthesized DNA thus is used to measure the rate of DNA synthesis.

BIOSYNTHESIS OF PURINE NUCLEOTIDES

Purine and pyrimidine nucleotides are synthesized *in vivo* at rates consistent with physiologic need. Intracellular mechanisms sense and regulate the pool sizes of nucleotide triphosphates (NTPs), which rise during growth or tissue regeneration when cells are rapidly dividing. Early investigations of nucleotide biosynthesis employed birds, and later ones used *Escherichia coli*. Isotopic precursors fed to pigeons established the source of each atom of a purine base (Figure 34-1) and initiated study of the intermediates of purine biosynthesis.

Three processes contribute to purine nucleotide biosynthesis. These are, in order of decreasing importance: (1) synthesis from amphibolic intermediates

(synthesis *de novo*), (2) phosphoribosylation of purines, and (3) phosphorylation of purine nucleosides.

INOSINE MONOPHOSPHATE (IMP) IS SYNTHESIZED FROM AMPHIBOLIC INTERMEDIATES

Figure 34-2 illustrates the intermediates and reactions for conversion of α -D-ribose 5-phosphate to inosine monophosphate (IMP). Separate branches then lead to AMP and GMP (Figure 34-3). Subsequent phosphoryl transfer from ATP converts AMP and GMP to ADP and GDP. Conversion of GDP to GTP involves a second phosphoryl transfer from ATP, whereas conversion of ADP to ATP is achieved primarily by oxidative phosphorylation (see Chapter 12).

Multifunctional Catalysts Participate in Purine Nucleotide Biosynthesis

In prokaryotes, each reaction of Figure 34-2 is catalyzed by a different polypeptide. By contrast, in eukaryotes, the enzymes are polypeptides with multiple catalytic activities whose adjacent catalytic sites facilitate channeling of intermediates between sites. Three distinct multifunctional enzymes catalyze reactions 3, 4, and 6, reactions 7 and 8, and reactions 10 and 11 of Figure 34-2.

Antifolate Drugs or Glutamine Analogs Block Purine Nucleotide Biosynthesis

The carbons added in reactions 4 and 5 of Figure 34-2 are contributed by derivatives of tetrahydrofolate. Purine deficiency states, which are rare in humans, generally reflect a deficiency of folic acid. Compounds that inhibit formation of tetrahydrofolates and therefore block purine synthesis have been used in cancer chemotherapy. Inhibitory compounds and the reactions they inhibit include azaserine (reaction 5, Figure 34-2), diazanorleucine (reaction 2), 6-mercaptopurine (reactions 13 and 14), and mycophenolic acid (reaction 14).

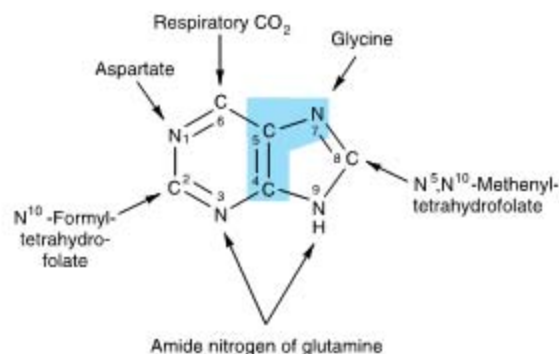


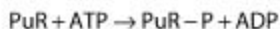
Figure 34-1. Sources of the nitrogen and carbon atoms of the purine ring. Atoms 4, 5, and 7 (shaded) derive from glycine.

“SALVAGE REACTIONS” CONVERT PURINES & THEIR NUCLEOSIDES TO MONONUCLEOTIDES

Conversion of purines, their ribonucleosides, and their deoxyribonucleosides to mononucleotides involves so-called “salvage reactions” that require far less energy than *de novo* synthesis. The more important mechanism involves phosphoribosylation by PRPP (structure II, Figure 34-2) of a free purine (Pu) to form a purine 5′-mononucleotide (Pu-RP).



Two phosphoribosyl transferases then convert adenine to AMP and hypoxanthine and guanine to IMP or GMP (Figure 34-4). A second salvage mechanism involves phosphoryl transfer from ATP to a purine ribonucleoside (PuR):



Adenosine kinase catalyzes phosphorylation of adenosine and deoxyadenosine to AMP and dAMP, and deoxycytidine kinase phosphorylates deoxycytidine and 2′-deoxyguanosine to dCMP and dGMP.

Liver, the major site of purine nucleotide biosynthesis, provides purines and purine nucleosides for salvage and utilization by tissues incapable of their biosynthesis. For example, human brain has a low level of PRPP amidotransferase (reaction 2, Figure 34-2) and hence depends in part on exogenous purines. Erythrocytes and polymorphonuclear leukocytes cannot synthesize 5-phosphoribosylamine (structure III, Figure 34-2)

and therefore utilize exogenous purines to form nucleotides.

AMP & GMP Feedback-Regulate PRPP Glutamyl Amidotransferase

Since biosynthesis of IMP consumes glycine, glutamine, tetrahydrofolate derivatives, aspartate, and ATP, it is advantageous to regulate purine biosynthesis. The major determinant of the rate of *de novo* purine nucleotide biosynthesis is the concentration of PRPP, whose pool size depends on its rates of synthesis, utilization, and degradation. The rate of PRPP synthesis depends on the availability of ribose 5-phosphate and on the activity of PRPP synthase, an enzyme sensitive to feedback inhibition by AMP, ADP, GMP, and GDP.

AMP & GMP Feedback-Regulate Their Formation From IMP

Two mechanisms regulate conversion of IMP to GMP and AMP. AMP and GMP feedback-inhibit adenylosuccinate synthase and IMP dehydrogenase (reactions 12 and 14, Figure 34-3), respectively. Furthermore, conversion of IMP to adenylosuccinate en route to AMP requires GTP, and conversion of xanthinylate (XMP) to GMP requires ATP. This cross-regulation between the pathways of IMP metabolism thus serves to decrease synthesis of one purine nucleotide when there is a deficiency of the other nucleotide. AMP and GMP also inhibit hypoxanthine-guanine phosphoribosyltransferase, which converts hypoxanthine and guanine to IMP and GMP (Figure 34-4), and GMP feedback-inhibits PRPP glutamyl amidotransferase (reaction 2, Figure 34-2).

REDUCTION OF RIBONUCLEOSIDE DIPHOSPHATES FORMS DEOXYRIBONUCLEOSIDE DIPHOSPHATES

Reduction of the 2′-hydroxyl of purine and pyrimidine ribonucleotides, catalyzed by the **ribonucleotide reductase complex** (Figure 34-5), forms deoxyribonucleoside diphosphates (dNDPs). The enzyme complex is active only when cells are actively synthesizing DNA. Reduction requires thioredoxin, thioredoxin reductase, and NADPH. The immediate reductant, reduced thioredoxin, is produced by NADPH:thioredoxin reductase (Figure 34-5). Reduction of ribonucleoside diphosphates (NDPs) to deoxyribonucleoside diphosphates (dNDPs) is subject to complex regulatory controls that achieve balanced production of deoxyribonucleotides for synthesis of DNA (Figure 34-6).

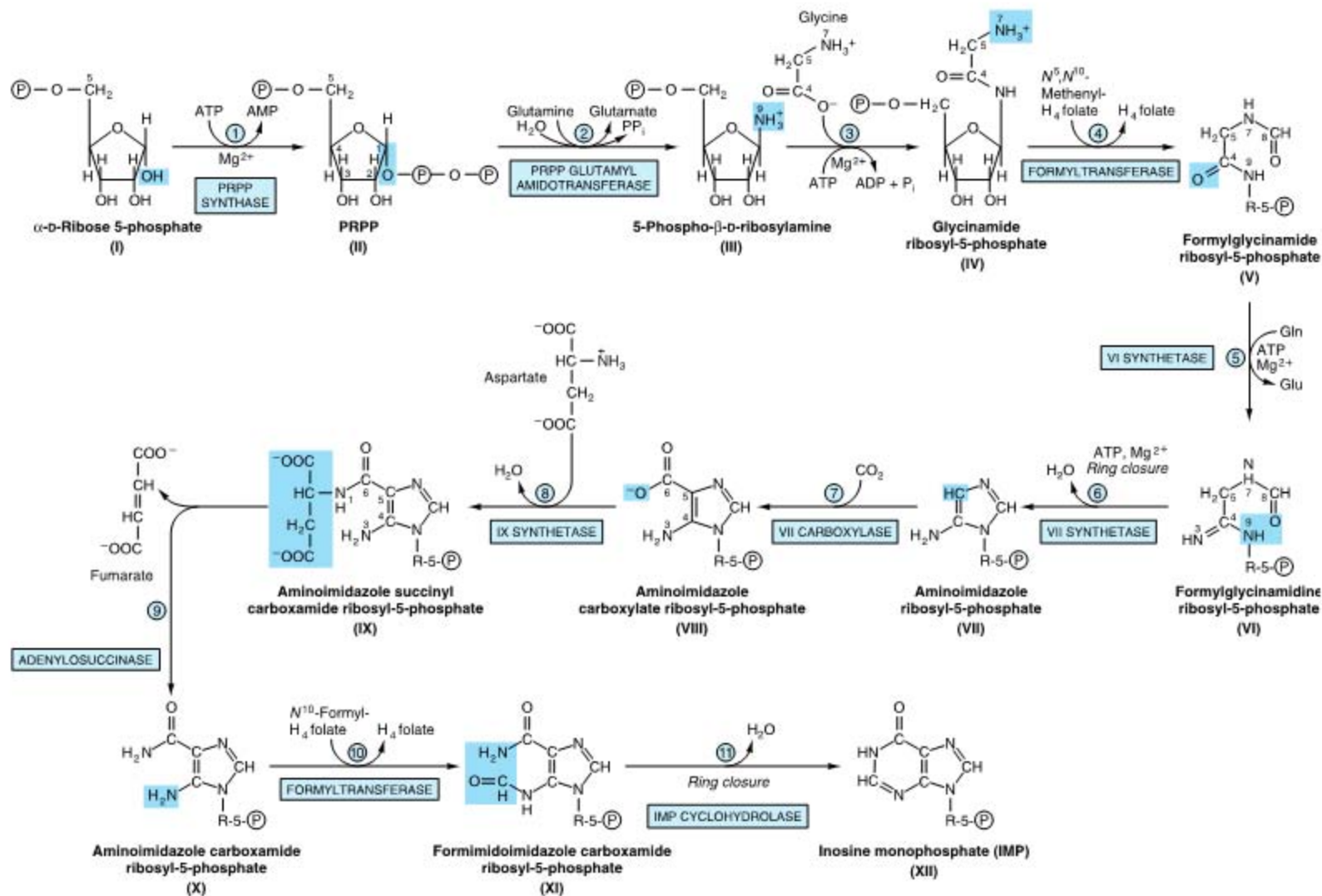


Figure 34-2. Purine biosynthesis from ribose 5-phosphate and ATP. See text for explanations. (P, PO_3^{2-} or PO_2^- .)

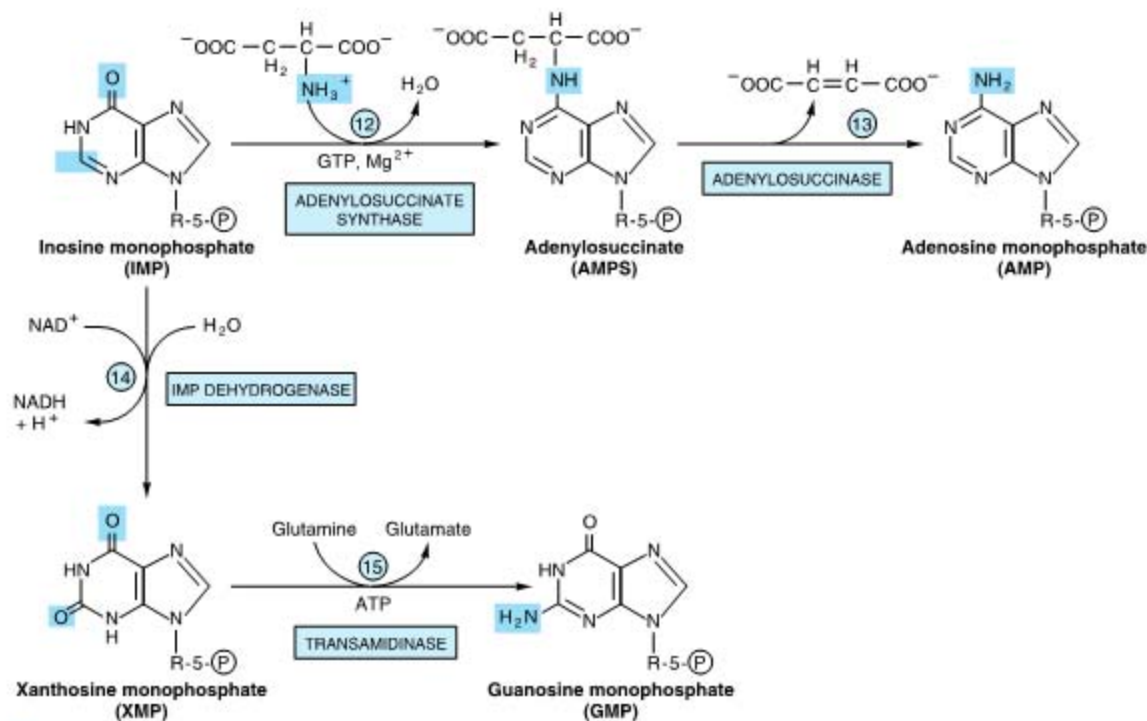


Figure 34-3. Conversion of IMP to AMP and GMP.

BIOSYNTHESIS OF PYRIMIDINE NUCLEOTIDES

Figure 34-7 summarizes the roles of the intermediates and enzymes of pyrimidine nucleotide biosynthesis. The catalyst for the initial reaction is *cytosolic* carbamoyl phosphate synthase II, a different enzyme from the *mitochondrial* carbamoyl phosphate synthase I of urea synthesis (Figure 29-9). Compartmentation thus provides two independent pools of carbamoyl phosphate. PRPP, an early participant in purine nucleotide synthesis (Figure 34-2), is a much later participant in pyrimidine biosynthesis.

Multifunctional Proteins Catalyze the Early Reactions of Pyrimidine Biosynthesis

Five of the first six enzyme activities of pyrimidine biosynthesis reside on multifunctional polypeptides. One such polypeptide catalyzes the first three reactions of Figure 34-2 and ensures efficient channeling of carbamoyl phosphate to pyrimidine biosynthesis. A second bifunctional enzyme catalyzes reactions 5 and 6.

THE DEOXYRIBONUCLEOSIDES OF URACIL & CYTOSINE ARE SALVAGED

While mammalian cells reutilize few free pyrimidines, “salvage reactions” convert the ribonucleosides uridine and cytidine and the deoxyribonucleosides thymidine and deoxycytidine to their respective nucleotides. ATP-dependent phosphoryltransferases (kinases) catalyze the phosphorylation of the nucleoside diphosphates 2'-deoxycytidine, 2'-deoxyguanosine, and 2'-deoxyadenosine to their corresponding nucleoside triphosphates. In addition, orotate phosphoryltransferase (reaction 5, Figure 34-7), an enzyme of pyrimidine nucleotide synthesis, salvages orotic acid by converting it to orotidine monophosphate (OMP).

Methotrexate Blocks Reduction of Dihydrofolate

Reaction 12 of Figure 34-7 is the only reaction of pyrimidine nucleotide biosynthesis that requires a tetrahydrofolate derivative. The methylene group of N^5,N^{10} -methylene-tetrahydrofolate is reduced to the methyl group that is transferred, and tetrahydrofolate is oxidized to dihydro-

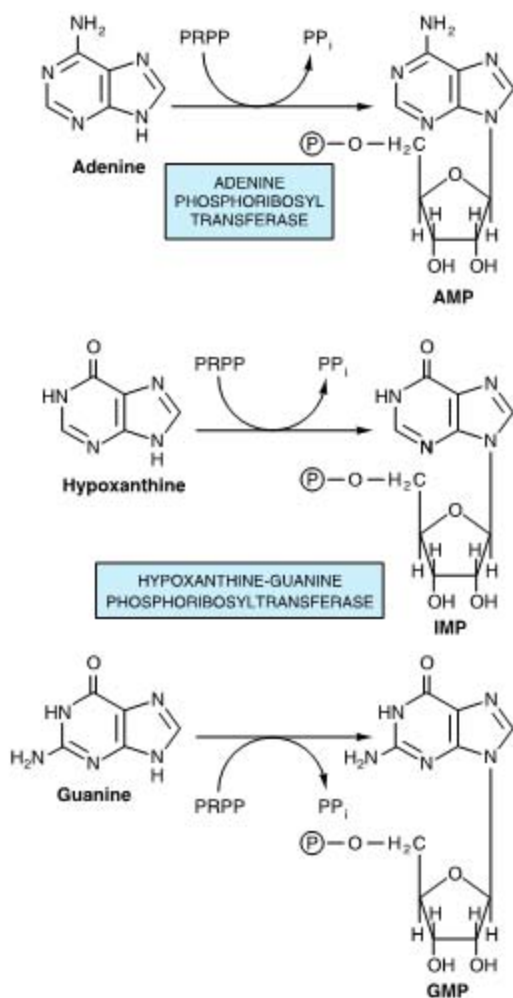


Figure 34-4. Phosphoribosylation of adenine, hypoxanthine, and guanine to form AMP, IMP, and GMP, respectively.

folate. For further pyrimidine synthesis to occur, dihydrofolate must be reduced back to tetrahydrofolate, a reaction catalyzed by dihydrofolate reductase. Dividing cells, which must generate TMP and dihydrofolate, thus are especially sensitive to inhibitors of dihydrofolate reductase such as the anticancer drug **methotrexate**.

Certain Pyrimidine Analogs Are Substrates for Enzymes of Pyrimidine Nucleotide Biosynthesis

Orotate phosphoribosyltransferase (reaction 5, Figure 34-7) converts the drug **allopurinol** (Figure 33-12) to

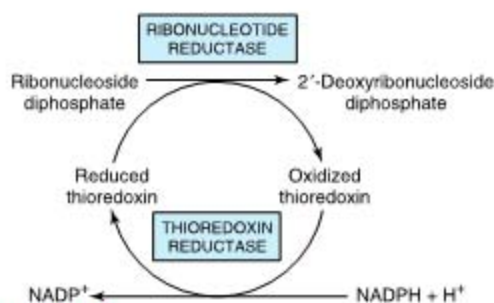


Figure 34-5. Reduction of ribonucleoside diphosphates to 2'-deoxyribonucleoside diphosphates.

a nucleotide in which the ribosyl phosphate is attached to N-1 of the pyrimidine ring. The anticancer drug **5-fluorouracil** (Figure 33-12) is also phosphoribosylated by orotate phosphoribosyl transferase.

REGULATION OF PYRIMIDINE NUCLEOTIDE BIOSYNTHESIS

Gene Expression & Enzyme Activity Both Are Regulated

The activities of the first and second enzymes of pyrimidine nucleotide biosynthesis are controlled by allosteric

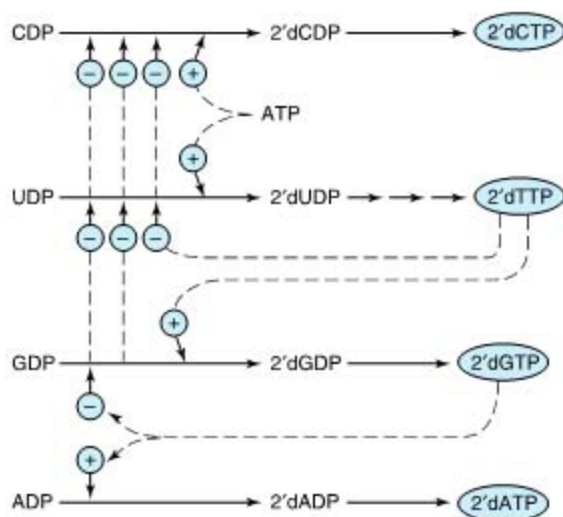


Figure 34-6. Regulation of the reduction of purine and pyrimidine ribonucleotides to their respective 2'-deoxyribonucleotides. Solid lines represent chemical flow. Broken lines show negative (⊖) or positive (⊕) feedback regulation.

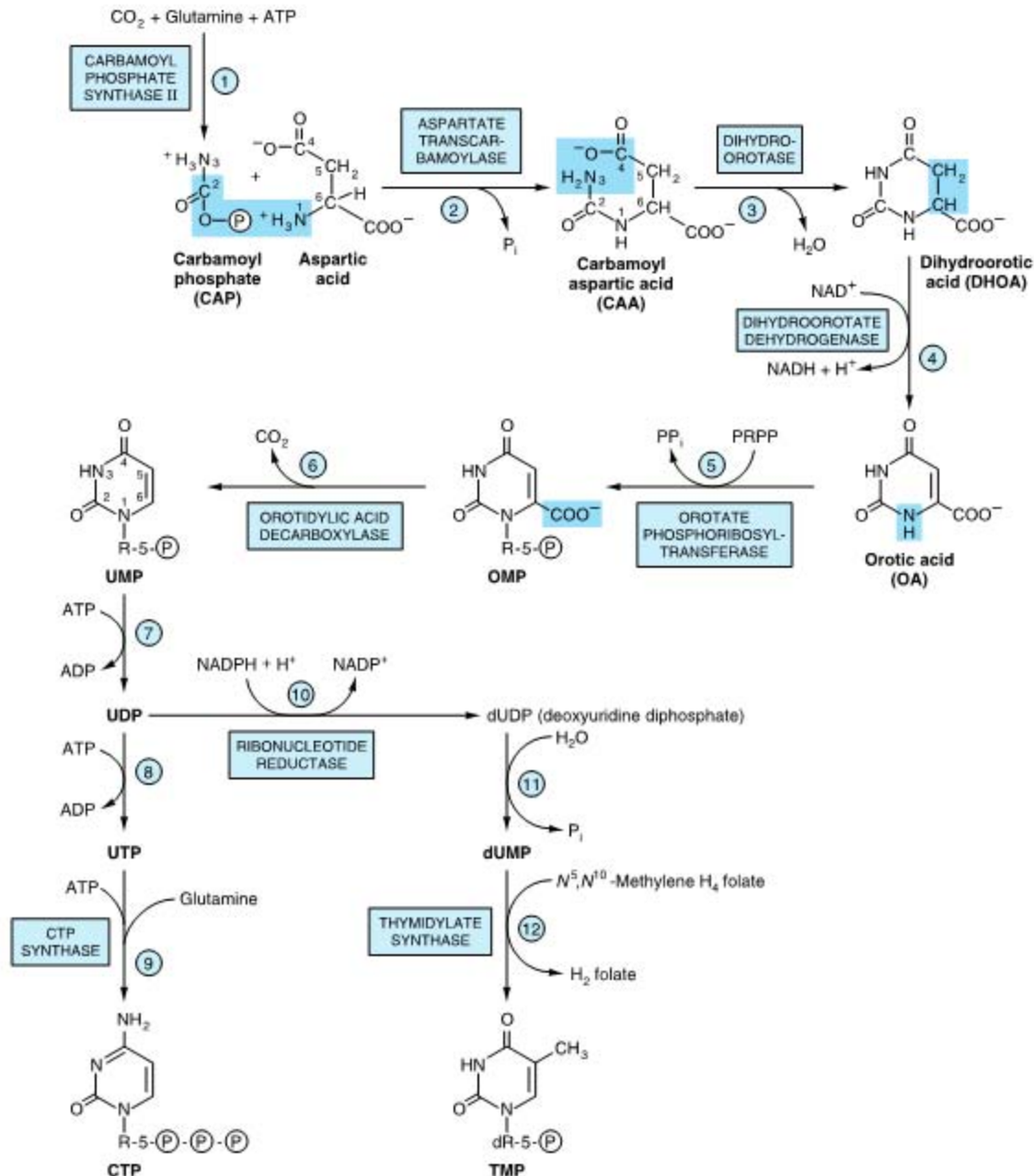


Figure 34-7. The biosynthetic pathway for pyrimidine nucleotides.

regulation. Carbamoyl phosphate synthase II (reaction 1, Figure 34-7) is inhibited by UTP and purine nucleotides but activated by PRPP. Aspartate transcarbamoylase (reaction 2, Figure 34-7) is inhibited by CTP but activated by ATP. In addition, the first three and the last two enzymes of the pathway are regulated by coordinate repression and derepression.

Purine & Pyrimidine Nucleotide Biosynthesis Are Coordinately Regulated

Purine and pyrimidine biosynthesis parallel one another mole for mole, suggesting coordinated control of their biosynthesis. Several sites of cross-regulation characterize purine and pyrimidine nucleotide biosynthesis. The PRPP synthase reaction (reaction 1, Figure 34-2), which forms a precursor essential for both processes, is feedback-inhibited by both purine and pyrimidine nucleotides.

HUMANS CATABOLIZE PURINES TO URIC ACID

Humans convert adenosine and guanosine to uric acid (Figure 34-8). Adenosine is first converted to inosine by adenosine deaminase. In mammals other than higher primates, uricase converts uric acid to the water-soluble product allantoin. However, since humans lack uricase, the end product of purine catabolism in humans is uric acid.

GOUT IS A METABOLIC DISORDER OF PURINE CATABOLISM

Various genetic defects in PRPP synthetase (reaction 1, Figure 34-2) present clinically as gout. Each defect—eg, an elevated V_{max} , increased affinity for ribose 5-phosphate, or resistance to feedback inhibition—results in overproduction and overexcretion of purine catabolites. When serum urate levels exceed the solubility limit, sodium urate crystallizes in soft tissues and joints and causes an inflammatory reaction, **gouty arthritis**. However, most cases of gout reflect abnormalities in renal handling of uric acid.

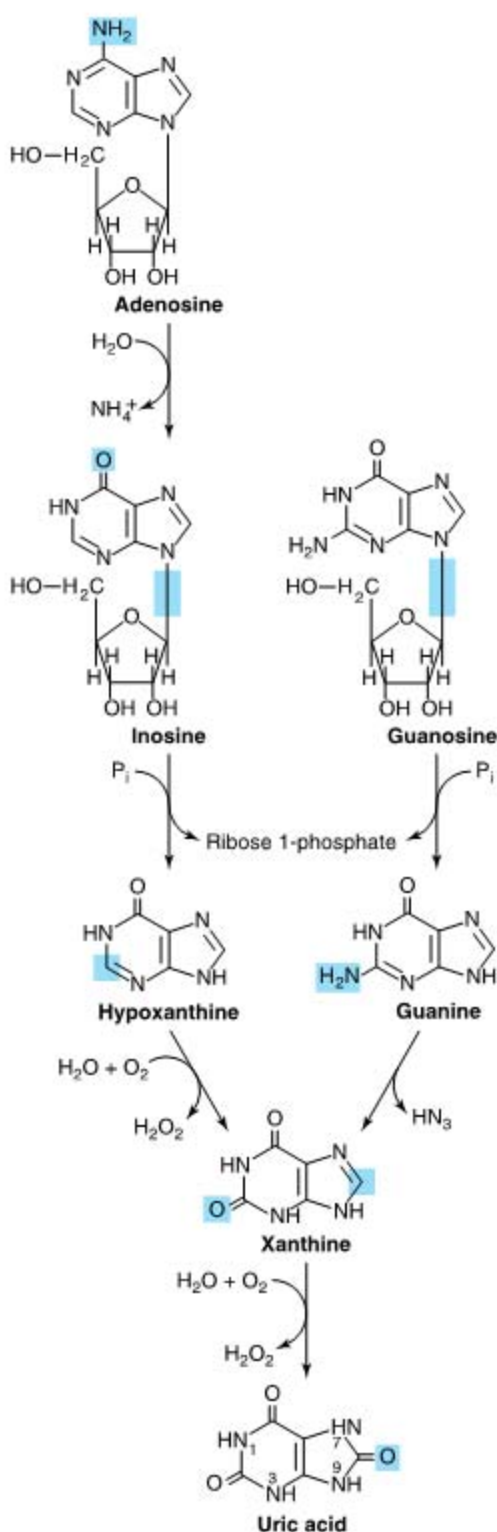


Figure 34-8. Formation of uric acid from purine nucleosides by way of the purine bases hypoxanthine, xanthine, and guanine. Purine deoxyribonucleosides are degraded by the same catabolic pathway and enzymes, all of which exist in the mucosa of the mammalian gastrointestinal tract.

OTHER DISORDERS OF PURINE CATABOLISM

While purine deficiency states are rare in human subjects, there are numerous genetic disorders of purine catabolism. **Hyperuricemias** may be differentiated based on whether patients excrete normal or excessive quantities of total urates. Some hyperuricemias reflect specific enzyme defects. Others are secondary to diseases such as cancer or psoriasis that enhance tissue turnover.

Lesch-Nyhan Syndrome

Lesch-Nyhan syndrome, an overproduction hyperuricemia characterized by frequent episodes of uric acid lithiasis and a bizarre syndrome of self-mutilation, reflects a defect in **hypoxanthine-guanine phosphoribosyl transferase**, an enzyme of purine salvage (Figure 34-4). The accompanying rise in intracellular PRPP results in purine overproduction. Mutations that decrease or abolish hypoxanthine-guanine phosphoribosyltransferase activity include deletions, frameshift mutations, base substitutions, and aberrant mRNA splicing.

Von Gierke's Disease

Purine overproduction and hyperuricemia in von Gierke's disease (**glucose-6-phosphatase deficiency**) occurs secondary to enhanced generation of the PRPP precursor ribose 5-phosphate. An associated lactic acidosis elevates the renal threshold for urate, elevating total body urates.

Hypouricemia

Hypouricemia and increased excretion of hypoxanthine and xanthine are associated with **xanthine oxidase deficiency** due to a genetic defect or to severe liver damage. Patients with a severe enzyme deficiency may exhibit xanthinuria and xanthine lithiasis.

Adenosine Deaminase & Purine Nucleoside Phosphorylase Deficiency

Adenosine deaminase deficiency is associated with an immunodeficiency disease in which both thymus-derived lymphocytes (T cells) and bone marrow-derived lymphocytes (B cells) are sparse and dysfunctional. **Purine nucleoside phosphorylase deficiency** is associated with a severe deficiency of T cells but apparently normal B cell function. Immune dysfunctions appear to result from accumulation of dGTP and dATP, which inhibit ribonucleotide reductase and thereby deplete cells of DNA precursors.

CATABOLISM OF PYRIMIDINES PRODUCES WATER-SOLUBLE METABOLITES

Unlike the end products of purine catabolism, those of pyrimidine catabolism are highly water-soluble: CO_2 , NH_3 , β -alanine, and β -aminoisobutyrate (Figure 34-9). Excretion of β -aminoisobutyrate increases in leukemia and severe x-ray radiation exposure due to increased destruction of DNA. However, many persons of Chinese or Japanese ancestry routinely excrete β -aminoisobutyrate. Humans probably transaminate β -aminoisobutyrate to methylmalonate semialdehyde, which then forms succinyl-CoA (Figure 19-2).

Pseudouridine Is Excreted Unchanged

Since no human enzyme catalyzes hydrolysis or phosphorolysis of pseudouridine, this unusual nucleoside is excreted unchanged in the urine of normal subjects.

OVERPRODUCTION OF PYRIMIDINE CATABOLITES IS ONLY RARELY ASSOCIATED WITH CLINICALLY SIGNIFICANT ABNORMALITIES

Since the end products of pyrimidine catabolism are highly water-soluble, pyrimidine overproduction results in few clinical signs or symptoms. In hyperuricemia associated with severe overproduction of PRPP, there is overproduction of pyrimidine nucleotides and increased excretion of β -alanine. Since N^5,N^{10} -methylene-tetrahydrofolate is required for thymidylate synthesis, disorders of folate and vitamin B_{12} metabolism result in deficiencies of TMP.

Orotic Acidurias

The orotic aciduria that accompanies **Reye's syndrome** probably is a consequence of the inability of severely damaged mitochondria to utilize carbamoyl phosphate, which then becomes available for cytosolic overproduction of orotic acid. **Type I orotic aciduria** reflects a deficiency of both orotate phosphoribosyltransferase and orotidylate decarboxylase (reactions 5 and 6, Figure 34-7); the rarer **type II orotic aciduria** is due to a deficiency only of orotidylate decarboxylase (reaction 6, Figure 34-7).

Deficiency of a Urea Cycle Enzyme Results in Excretion of Pyrimidine Precursors

Increased excretion of orotic acid, uracil, and uridine accompanies a deficiency in liver mitochondrial ornithine transcarbamoylase (reaction 2, Figure 29-9).

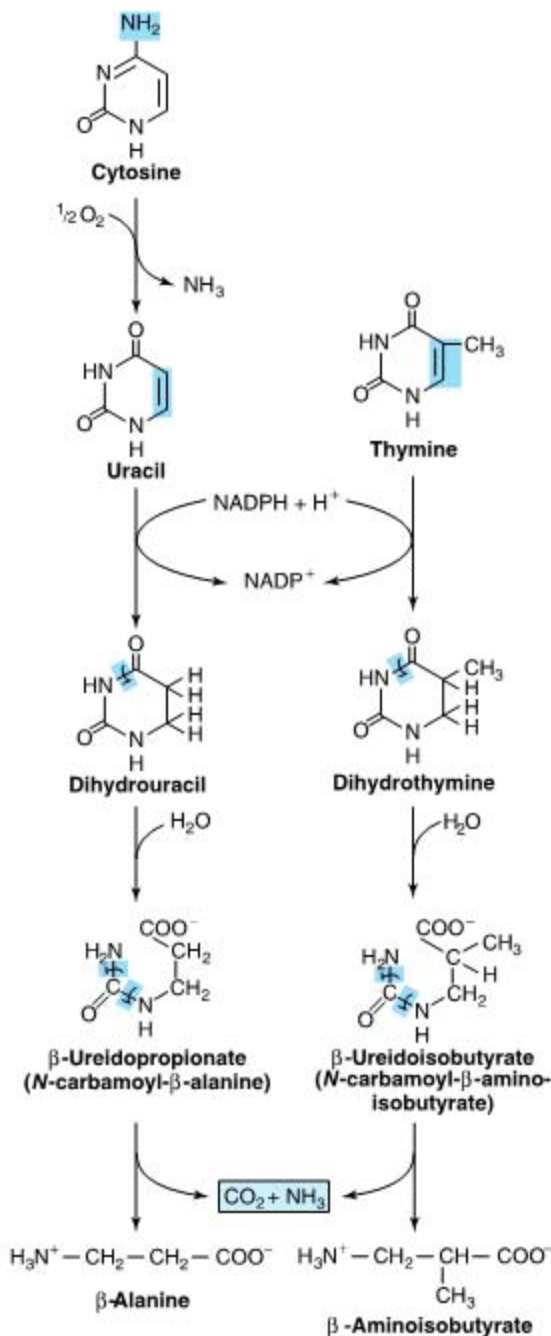


Figure 34-9. Catabolism of pyrimidines.

Excess carbamoyl phosphate exits to the cytosol, where it stimulates pyrimidine nucleotide biosynthesis. The resulting mild **orotic aciduria** is increased by high-nitrogen foods.

Drugs May Precipitate Orotic Aciduria

Allopurinol (Figure 33-12), an alternative substrate for orotate phosphoribosyltransferase (reaction 5, Figure 34-7), competes with orotic acid. The resulting nucleotide product also inhibits orotidylate decarboxylase (reaction 6, Figure 34-7), resulting in **orotic aciduria** and **orotidinuria**. 6-Azauridine, following conversion to 6-azauridylate, also competitively inhibits orotidylate decarboxylase (reaction 6, Figure 34-7), enhancing excretion of orotic acid and orotidine.

SUMMARY

- Ingested nucleic acids are degraded to purines and pyrimidines. New purines and pyrimidines are formed from amphibolic intermediates and thus are dietarily nonessential.
- Several reactions of IMP biosynthesis require folate derivatives and glutamine. Consequently, antifolate drugs and glutamine analogs inhibit purine biosynthesis.
- Oxidation and amination of IMP forms AMP and GMP, and subsequent phosphoryl transfer from ATP forms ADP and GDP. Further phosphoryl transfer from ATP to GDP forms GTP. ADP is converted to ATP by oxidative phosphorylation. Reduction of NDPs forms dNDPs.
- Hepatic purine nucleotide biosynthesis is stringently regulated by the pool size of PRPP and by feedback inhibition of PRPP-glutamyl amidotransferase by AMP and GMP.
- Coordinated regulation of purine and pyrimidine nucleotide biosynthesis ensures their presence in proportions appropriate for nucleic acid biosynthesis and other metabolic needs.
- Humans catabolize purines to uric acid (pK_a 5.8), present as the relatively insoluble acid at acidic pH or as its more soluble sodium urate salt at a pH near neutrality. Urate crystals are diagnostic of gout. Other disorders of purine catabolism include Lesch-Nyhan syndrome, von Gierke's disease, and hypouricemias.
- Since pyrimidine catabolites are water-soluble, their overproduction does not result in clinical abnormalities. Excretion of pyrimidine precursors can, however, result from a deficiency of ornithine transcarbamoylase because excess carbamoyl phosphate is available for pyrimidine biosynthesis.

REFERENCES

- Benkovic SJ: The transformylase enzymes in de novo purine biosynthesis. *Trends Biochem Sci* 1994;9:320.
- Brooks EM et al: Molecular description of three macro-deletions and an Alu-Alu recombination-mediated duplication in the HPRT gene in four patients with Lesch-Nyhan disease. *Mutat Res* 2001;476:43.
- Curto R, Voit EO, Cascante M: Analysis of abnormalities in purine metabolism leading to gout and to neurological dysfunctions in man. *Biochem J* 1998;329:477.
- Harris MD, Siegel LB, Alloway JA: Gout and hyperuricemia. *Am Family Physician* 1999;59:925.
- Lipkowitz MS et al: Functional reconstitution, membrane targeting, genomic structure, and chromosomal localization of a human urate transporter. *J Clin Invest* 2001;107:1103.
- Martinez J et al: Human genetic disorders, a phylogenetic perspective. *J Mol Biol* 2001;308:587.
- Puig JG et al: Gout: new questions for an ancient disease. *Adv Exp Med Biol* 1998;431:1.
- Scriver CR et al (editors): *The Metabolic and Molecular Bases of Inherited Disease*, 8th ed. McGraw-Hill, 2001.
- Tvrđik T et al: Molecular characterization of two deletion events involving Alu-sequences, one novel base substitution and two tentative hotspot mutations in the hypoxanthine phosphoribosyltransferase gene in five patients with Lesch-Nyhan syndrome. *Hum Genet* 1998;103:311.
- Zalkin H, Dixon JE: De novo purine nucleotide synthesis. *Prog Nucleic Acid Res Mol Biol* 1992;42:259.

Daryl K. Granner, MD

BIOMEDICAL IMPORTANCE

The discovery that genetic information is coded along the length of a polymeric molecule composed of only four types of monomeric units was one of the major scientific achievements of the twentieth century. This polymeric molecule, **DNA**, is the chemical basis of heredity and is organized into genes, the fundamental units of genetic information. The basic information pathway—ie, DNA directs the synthesis of RNA, which in turn directs protein synthesis—has been elucidated. Genes do not function autonomously; their replication and function are controlled by various gene products, often in collaboration with components of various signal transduction pathways. Knowledge of the structure and function of nucleic acids is essential in understanding genetics and many aspects of pathophysiology as well as the genetic basis of disease.

DNA CONTAINS THE GENETIC INFORMATION

The demonstration that DNA contained the genetic information was first made in 1944 in a series of experiments by Avery, MacLeod, and McCarty. They showed that the genetic determination of the character (type) of the capsule of a specific pneumococcus could be transmitted to another of a different capsular type by introducing purified DNA from the former coccus into the latter. These authors referred to the agent (later shown to be DNA) accomplishing the change as “transforming factor.” Subsequently, this type of genetic manipulation has become commonplace. Similar experiments have recently been performed utilizing yeast, cultured mammalian cells, and insect and mammalian embryos as recipients and cloned DNA as the donor of genetic information.

DNA Contains Four Deoxynucleotides

The chemical nature of the monomeric deoxynucleotide units of DNA—**deoxyadenylate**, **deoxyguanylate**, **deoxycytidylate**, and **thymidylate**—is described in Chapter 33. These monomeric units of DNA are held in polymeric form by 3',5'-phosphodiester bridges constituting a single strand, as depicted in Figure 35-1.

The informational content of DNA (the genetic code) resides in the sequence in which these monomers—purine and pyrimidine deoxyribonucleotides—are ordered. The polymer as depicted possesses a polarity; one end has a 5'-hydroxyl or phosphate terminal while the other has a 3'-phosphate or hydroxyl terminal. The importance of this polarity will become evident. Since the genetic information resides in the order of the monomeric units within the polymers, there must exist a mechanism of reproducing or replicating this specific information with a high degree of fidelity. That requirement, together with x-ray diffraction data from the DNA molecule and the observation of Chargaff that in DNA molecules the concentration of deoxyadenosine (A) nucleotides equals that of thymidine (T) nucleotides (A = T), while the concentration of deoxyguanosine (G) nucleotides equals that of deoxycytidine (C) nucleotides (G = C), led Watson, Crick, and Wilkins to propose in the early 1950s a model of a double-stranded DNA molecule. The model they proposed is depicted in Figure 35-2. The two strands of this double-stranded helix are held in register by **hydrogen bonds** between the purine and pyrimidine bases of the respective linear molecules. The pairings between the purine and pyrimidine nucleotides on the opposite strands are very specific and are dependent upon hydrogen bonding of **A with T** and **G with C** (Figure 35-3).

This common form of DNA is said to be right-handed because as one looks down the double helix the base residues form a spiral in a clockwise direction. In the double-stranded molecule, restrictions imposed by the rotation about the phosphodiester bond, the favored anti configuration of the glycosidic bond (Figure 33-8), and the predominant tautomers (see Figure 33-3) of the four bases (A, G, T, and C) allow A to pair only with T and G only with C, as depicted in Figure 35-3. This base-pairing restriction explains the earlier observation that in a double-stranded DNA molecule the content of A equals that of T and the content of G equals that of C. The two strands of the double-helical molecule, each of which possesses a polarity, are **antiparallel**; ie, one strand runs in the 5' to 3' direction and the other in the 3' to 5' direction. This is analogous to two parallel streets, each running one way but carrying traffic in opposite directions. In the double-stranded DNA molecules, the genetic information re-

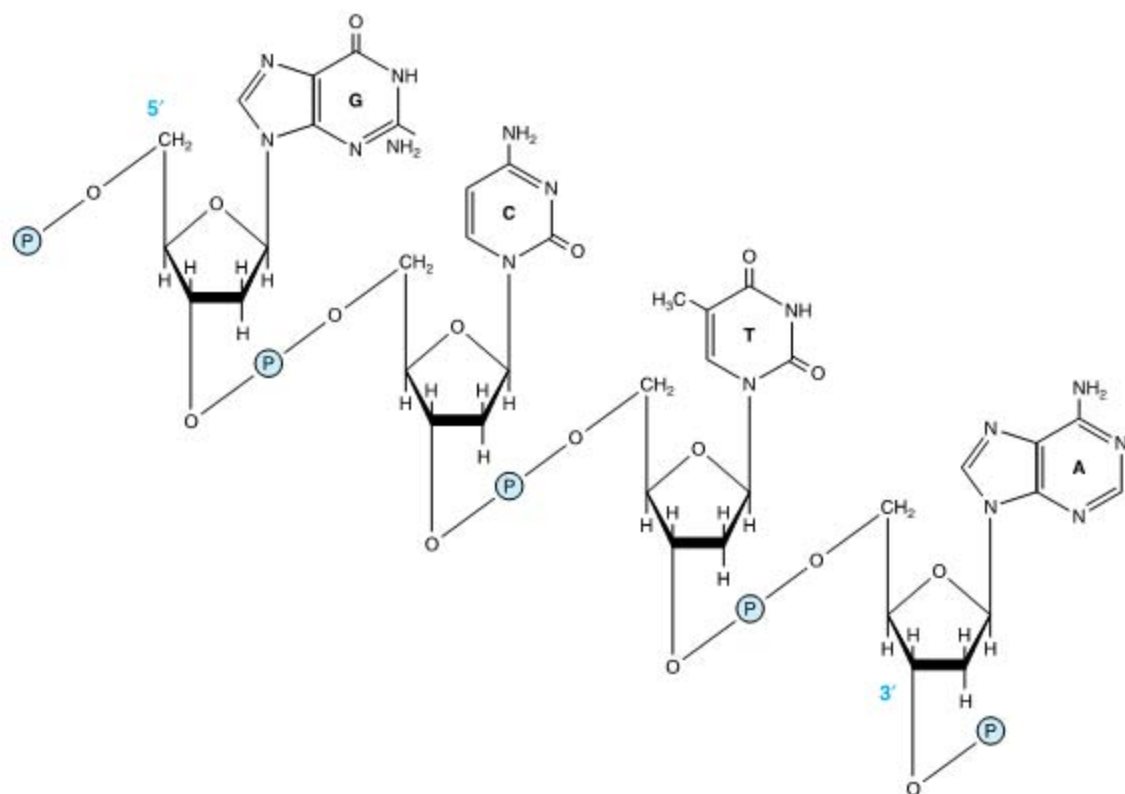


Figure 35-1. A segment of one strand of a DNA molecule in which the purine and pyrimidine bases guanine (G), cytosine (C), thymine (T), and adenine (A) are held together by a phosphodiester backbone between 2'-deoxyribosyl moieties attached to the nucleobases by an *N*-glycosidic bond. Note that the backbone has a polarity (ie, a direction). Convention dictates that a single-stranded DNA sequence is written in the 5' to 3' direction (ie, pGpCpTpA, where G, C, T, and A represent the four bases and p represents the interconnecting phosphates).

sides in the sequence of nucleotides on one strand, the **template strand**. This is the strand of DNA that is copied during nucleic acid synthesis. It is sometimes referred to as the **noncoding strand**. The opposite strand is considered the **coding strand** because it matches the RNA transcript that encodes the protein.

The two strands, in which opposing bases are held together by hydrogen bonds, wind around a central axis in the form of a **double helix**. Double-stranded DNA exists in at least six forms (A–E and Z). The B form is usually found under physiologic conditions (low salt, high degree of hydration). A single turn of B-DNA about the axis of the molecule contains ten base pairs. The distance spanned by one turn of B-DNA is 3.4 nm. The width (helical diameter) of the double helix in B-DNA is 2 nm.

As depicted in Figure 35-3, three hydrogen bonds hold the deoxyguanosine nucleotide to the deoxycyti-

dine nucleotide, whereas the other pair, the A–T pair, is held together by two hydrogen bonds. Thus, the G–C bonds are much more resistant to denaturation, or “melting,” than A–T-rich regions.

The Denaturation (Melting) of DNA Is Used to Analyze Its Structure

The double-stranded structure of DNA can be separated into two component strands (melted) in solution by increasing the temperature or decreasing the salt concentration. Not only do the two stacks of bases pull apart but the bases themselves unstack while still connected in the polymer by the phosphodiester backbone. Concomitant with this denaturation of the DNA molecule is an increase in the optical absorbance of the purine and pyrimidine bases—a phenomenon referred to as **hyperchromicity** of denaturation. Because of the

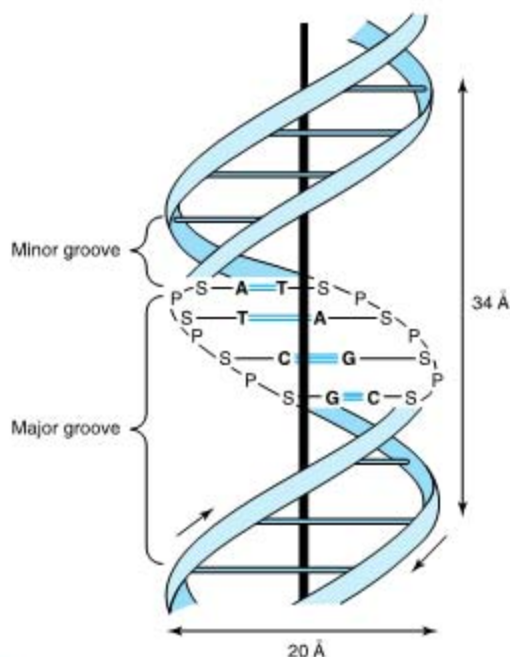


Figure 35-2. A diagrammatic representation of the Watson and Crick model of the double-helical structure of the B form of DNA. The horizontal arrow indicates the width of the double helix (20 Å), and the vertical arrow indicates the distance spanned by one complete turn of the double helix (34 Å). One turn of B-DNA includes ten base pairs (bp), so the rise is 3.4 Å per bp. The central axis of the double helix is indicated by the vertical rod. The short arrows designate the polarity of the antiparallel strands. The major and minor grooves are depicted. (A, adenine; C, cytosine; G, guanine; T, thymine; P, phosphate; S, sugar [deoxyribose].)

stacking of the bases and the hydrogen bonding between the stacks, the double-stranded DNA molecule exhibits properties of a rigid rod and in solution is a viscous material that loses its viscosity upon denaturation.

The strands of a given molecule of DNA separate over a temperature range. The midpoint is called the **melting temperature, or T_m** . The T_m is influenced by the base composition of the DNA and by the salt concentration of the solution. DNA rich in G–C pairs, which have three hydrogen bonds, melts at a higher temperature than that rich in A–T pairs, which have two hydrogen bonds. A tenfold increase of monovalent cation concentration increases the T_m by 16.6 °C. Formamide, which is commonly used in recombinant DNA experiments, destabilizes hydrogen bonding between bases, thereby lowering the T_m . This allows the strands of DNA

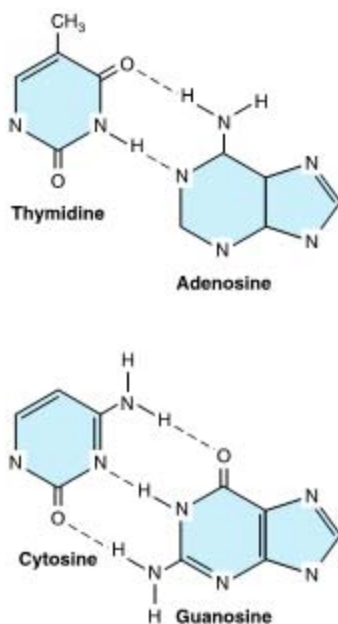


Figure 35-3. Base pairing between deoxyadenosine and thymine involves the formation of two hydrogen bonds. Three such bonds form between deoxycytidine and deoxyguanosine. The broken lines represent hydrogen bonds.

or DNA–RNA hybrids to be separated at much lower temperatures and minimizes the phosphodiester bond breakage that occurs at high temperatures.

Renaturation of DNA Requires Base Pair Matching

Separated strands of DNA will renature or reassociate when appropriate physiologic temperature and salt conditions are achieved. The rate of reassociation depends upon the concentration of the complementary strands. Reassociation of the two complementary DNA strands of a chromosome after DNA replication is a physiologic example of renaturation (see below). At a given temperature and salt concentration, a particular nucleic acid strand will associate tightly only with a complementary strand. Hybrid molecules will also form under appropriate conditions. For example, DNA will form a hybrid with a complementary DNA (cDNA) or with a cognate messenger RNA (mRNA; see below). When combined with gel electrophoresis techniques that separate hybrid molecules by size and radioactive labeling to provide a detectable signal, the resulting analytic techniques are called **Southern (DNA/cDNA)** and **Northern blotting (DNA/RNA)**, respectively. These proce-

dures allow for very specific identification of hybrids from mixtures of DNA or RNA (see Chapter 40).

There Are Grooves in the DNA Molecule

Careful examination of the model depicted in Figure 35-2 reveals a **major groove** and a **minor groove** winding along the molecule parallel to the phosphodiester backbones. In these grooves, proteins can interact specifically with exposed atoms of the nucleotides (usually by H bonding) and thereby recognize and bind to specific nucleotide sequences without disrupting the base pairing of the double-helical DNA molecule. As discussed in Chapters 37 and 39, regulatory proteins control the expression of specific genes via such interactions.

DNA Exists in Relaxed & Supercoiled Forms

In some organisms such as bacteria, bacteriophages, and many DNA-containing animal viruses, the ends of the DNA molecules are joined to create a closed circle with no covalently free ends. This of course does not destroy the polarity of the molecules, but it eliminates all free 3' and 5' hydroxyl and phosphoryl groups. Closed circles exist in relaxed or supercoiled forms. Supercoils are introduced when a closed circle is twisted around its own axis or when a linear piece of duplex DNA, whose ends are fixed, is twisted. This energy-requiring process puts the molecule under stress, and the greater the number of supercoils, the greater the stress or torsion (test this by twisting a rubber band). **Negative supercoils** are formed when the molecule is twisted in the direction opposite from the clockwise turns of the right-handed double helix found in B-DNA. Such DNA is said to be underwound. The energy required to achieve this state is, in a sense, stored in the supercoils. The transition to another form that requires energy is thereby facilitated by the underwinding. One such transition is strand separation, which is a prerequisite for DNA replication and transcription. Supercoiled DNA is therefore a preferred form in biologic systems. Enzymes that catalyze topologic changes of DNA are called **topoisomerases**. Topoisomerases can relax or insert supercoils. The best-characterized example is **bacterial gyrase**, which induces negative supercoiling in DNA using ATP as energy source. Homologs of this enzyme exist in all organisms and are important targets for cancer chemotherapy.

DNA PROVIDES A TEMPLATE FOR REPLICATION & TRANSCRIPTION

The genetic information stored in the nucleotide sequence of DNA serves two purposes. It is the source of information for the synthesis of all protein molecules of

the cell and organism, and it provides the information inherited by daughter cells or offspring. Both of these functions require that the DNA molecule serve as a template—in the first case for the transcription of the information into RNA and in the second case for the replication of the information into daughter DNA molecules.

The complementarity of the Watson and Crick double-stranded model of DNA strongly suggests that replication of the DNA molecule occurs in a semiconservative manner. Thus, when each strand of the double-stranded parental DNA molecule separates from its complement during replication, each serves as a template on which a new complementary strand is synthesized (Figure 35-4). The two newly formed double-stranded daughter DNA molecules, each containing one strand (but complementary rather than identical) from the parent double-stranded DNA molecule, are then sorted between the two daughter cells (Figure 35-5). Each daughter cell contains DNA molecules with information identical to that which the parent possessed; yet in each daughter cell the DNA molecule of the parent cell has been only semiconserved.

THE CHEMICAL NATURE OF RNA DIFFERS FROM THAT OF DNA

Ribonucleic acid (RNA) is a polymer of purine and pyrimidine ribonucleotides linked together by 3',5'-phosphodiester bridges analogous to those in DNA (Figure 35-6). Although sharing many features with DNA, RNA possesses several specific differences:

- (1) In RNA, the sugar moiety to which the phosphates and purine and pyrimidine bases are attached is ribose rather than the 2'-deoxyribose of DNA.
- (2) The pyrimidine components of RNA differ from those of DNA. Although RNA contains the ribonucleotides of adenine, guanine, and cytosine, it does not possess thymine except in the rare case mentioned below. Instead of thymine, RNA contains the ribonucleotide of uracil.
- (3) RNA exists as a single strand, whereas DNA exists as a double-stranded helical molecule. However, given the proper complementary base sequence with opposite polarity, the single strand of RNA—as demonstrated in Figure 35-7—is capable of folding back on itself like a hairpin and thus acquiring double-stranded characteristics.
- (4) Since the RNA molecule is a single strand complementary to only one of the two strands of a gene, its guanine content does not necessarily equal its cytosine content, nor does its adenine content necessarily equal its uracil content.

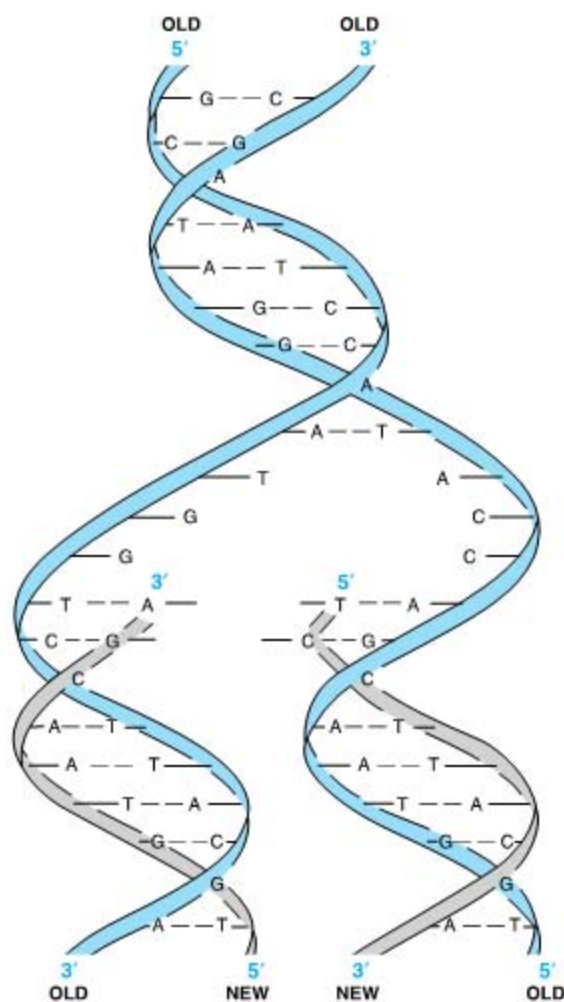


Figure 35-4. The double-stranded structure of DNA and the template function of each old strand (dark shading) on which a new (light shading) complementary strand is synthesized.

(5) RNA can be hydrolyzed by alkali to 2',3' cyclic diesters of the mononucleotides, compounds that cannot be formed from alkali-treated DNA because of the absence of a 2'-hydroxyl group. The alkali lability of RNA is useful both diagnostically and analytically.

Information within the single strand of RNA is contained in its sequence ("primary structure") of purine and pyrimidine nucleotides within the polymer. The sequence is complementary to the template strand of the gene from which it was transcribed. Because of this

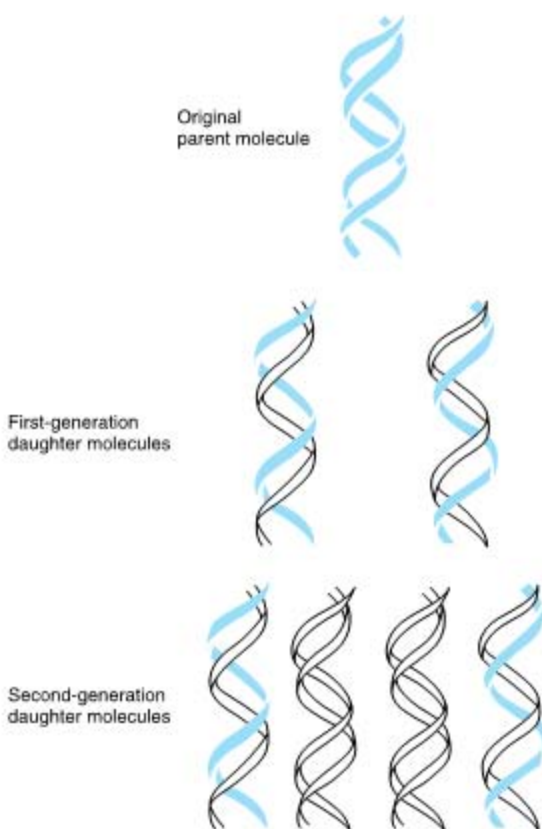


Figure 35-5. DNA replication is semiconservative. During a round of replication, each of the two strands of DNA is used as a template for synthesis of a new, complementary strand.

complementarity, an RNA molecule can bind specifically via the base-pairing rules to its template DNA strand; it will not bind ("hybridize") with the other (coding) strand of its gene. The sequence of the RNA molecule (except for U replacing T) is the same as that of the coding strand of the gene (Figure 35-8).

Nearly All of the Several Species of RNA Are Involved in Some Aspect of Protein Synthesis

Those cytoplasmic RNA molecules that serve as templates for protein synthesis (i.e., that transfer genetic information from DNA to the protein-synthesizing machinery) are designated **messenger RNAs**, or **mRNAs**. Many other cytoplasmic RNA molecules (**ribosomal RNAs**; **rRNAs**) have structural roles wherein they con-

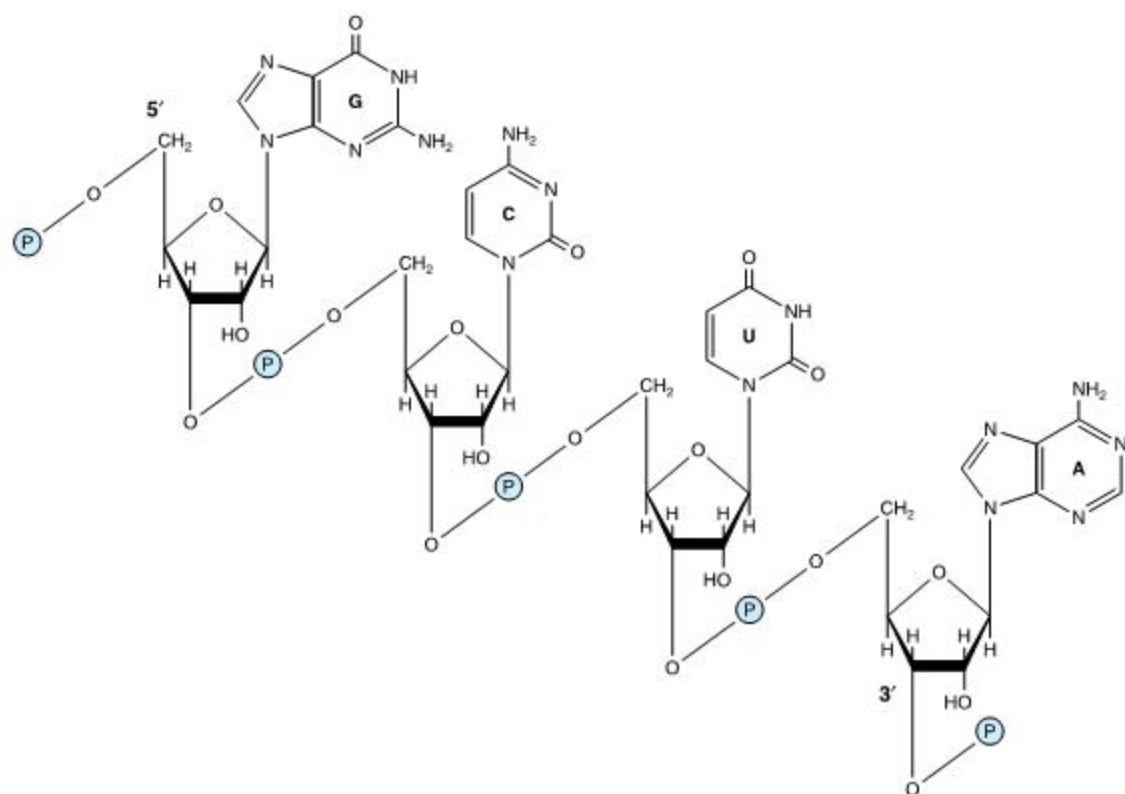


Figure 35–6. A segment of a ribonucleic acid (RNA) molecule in which the purine and pyrimidine bases—guanine (G), cytosine (C), uracil (U), and adenine (A)—are held together by phosphodiester bonds between ribosyl moieties attached to the nucleobases by *N*-glycosidic bonds. Note that the polymer has a polarity as indicated by the labeled 3′- and 5′-attached phosphates.

tribute to the formation and function of ribosomes (the organellar machinery for protein synthesis) or serve as adapter molecules (**transfer RNAs; tRNAs**) for the translation of RNA information into specific sequences of polymerized amino acids.

Some RNA molecules have intrinsic catalytic activity. The activity of these **ribozymes** often involves the cleavage of a nucleic acid. An example is the role of RNA in catalyzing the processing of the primary transcript of a gene into mature messenger RNA.

Much of the RNA synthesized from DNA templates in eukaryotic cells, including mammalian cells, is degraded within the nucleus, and it never serves as either a structural or an informational entity within the cellular cytoplasm.

In all eukaryotic cells there are **small nuclear RNA (snRNA)** species that are not directly involved in protein synthesis but play pivotal roles in RNA processing. These relatively small molecules vary in size from 90 to about 300 nucleotides (Table 35–1).

The genetic material for some animal and plant viruses is RNA rather than DNA. Although some RNA viruses never have their information transcribed into a DNA molecule, many animal RNA viruses—specifically, the retroviruses (the HIV virus, for example)—are transcribed by an RNA-dependent DNA polymerase, the so-called **reverse transcriptase**, to produce a double-stranded DNA copy of their RNA genome. In many cases, the resulting double-stranded DNA transcript is integrated into the host genome and subsequently serves as a template for gene expression and from which new viral RNA genomes can be transcribed.

RNA Is Organized in Several Unique Structures

In all prokaryotic and eukaryotic organisms, three main classes of RNA molecules exist: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA

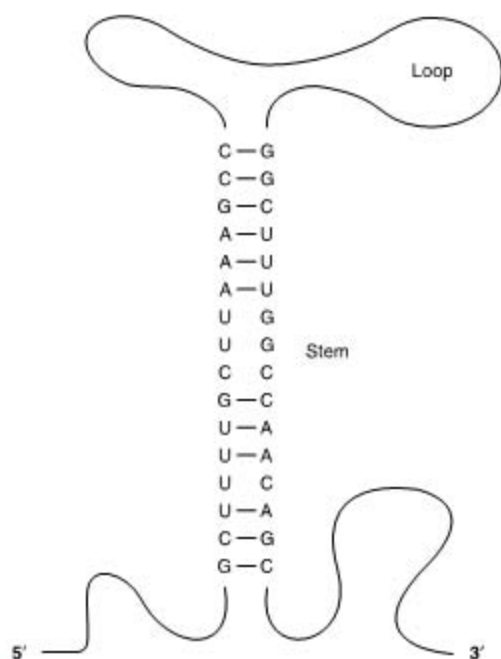


Figure 35-7. Diagrammatic representation of the secondary structure of a single-stranded RNA molecule in which a stem loop, or "hairpin," has been formed and is dependent upon the intramolecular base pairing. Note that A forms hydrogen bonds with U in RNA.

(rRNA). Each differs from the others by size, function, and general stability.

A. MESSENGER RNA (mRNA)

This class is the most heterogeneous in size and stability. All members of the class function as messengers conveying the information in a gene to the protein-synthesizing machinery, where each serves as a template on which a specific sequence of amino acids is polymerized to form a specific protein molecule, the ultimate gene product (Figure 35-9).

Table 35-1. Some of the species of small stable RNAs found in mammalian cells.

Name	Length (nucleotides)	Molecules per Cell	Localization
U1	165	1×10^6	Nucleoplasm/hnRNA
U2	188	5×10^5	Nucleoplasm
U3	216	3×10^5	Nucleolus
U4	139	1×10^5	Nucleoplasm
U5	118	2×10^5	Nucleoplasm
U6	106	3×10^5	Perichromatin granules
4.5S	91-95	3×10^5	Nucleus and cytoplasm
7S	280	5×10^5	Nucleus and cytoplasm
7-2	290	1×10^5	Nucleus and cytoplasm
7-3	300	2×10^5	Nucleus

Messenger RNAs, particularly in eukaryotes, have some unique chemical characteristics. The 5' terminal of mRNA is "capped" by a 7-methylguanosine triphosphate that is linked to an adjacent 2'-O-methyl ribonucleoside at its 5'-hydroxyl through the three phosphates (Figure 35-10). The mRNA molecules frequently contain internal 6-methyladenylates and other 2'-O-ribose methylated nucleotides. The cap is involved in the recognition of mRNA by the translating machinery, and it probably helps stabilize the mRNA by preventing the attack of 5'-exonucleases. The protein-synthesizing machinery begins translating the mRNA into proteins beginning downstream of the 5' or capped terminal. The other end of most mRNA molecules, the 3'-hydroxyl terminal, has an attached polymer of adenylate residues 20-250 nucleotides in length. The specific function of the **poly(A)** "tail" at the 3'-hydroxyl terminal of mRNAs is not fully understood, but it seems that it maintains the intracellular stability of the specific mRNA by preventing the attack of 3'-exonucleases. Some mRNAs, including those for some histones, do not contain poly(A). The poly(A) tail, because it will form a base pair with oligodeoxythymidine polymers attached to a solid substrate like cellulose, can be used to separate mRNA from other species of RNA, including mRNA molecules that lack this tail.

DNA strands:



Figure 35-8. The relationship between the sequences of an RNA transcript and its gene, in which the coding and template strands are shown with their polarities. The RNA transcript with a 5' to 3' polarity is complementary to the template strand with its 3' to 5' polarity. Note that the sequence in the RNA transcript and its polarity is the same as that in the coding strand, except that the U of the transcript replaces the T of the gene.

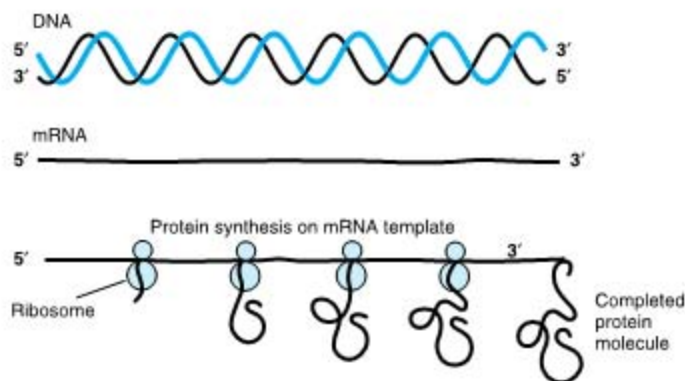


Figure 35–9. The expression of genetic information in DNA into the form of an mRNA transcript. This is subsequently translated by ribosomes into a specific protein molecule.

In mammalian cells, including cells of humans, the mRNA molecules present in the cytoplasm are not the RNA products immediately synthesized from the DNA template but must be formed by processing from a precursor molecule before entering the cytoplasm. Thus, in mammalian nuclei, the immediate products of gene transcription constitute a fourth class of RNA molecules. These nuclear RNA molecules are very heterogeneous in size and are quite large. The **heterogeneous nuclear RNA (hnRNA)** molecules may have a molecular weight in excess of 10^7 , whereas the molecular weight of mRNA molecules is generally less than 2×10^6 . As discussed in Chapter 37, hnRNA molecules are processed to generate the mRNA molecules which then enter the cytoplasm to serve as templates for protein synthesis.

B. TRANSFER RNA (tRNA)

tRNA molecules vary in length from 74 to 95 nucleotides. They also are generated by nuclear processing of a precursor molecule (Chapter 37). The tRNA molecules serve as adapters for the translation of the information in the sequence of nucleotides of the mRNA into specific amino acids. There are at least 20 species of tRNA molecules in every cell, at least one (and often several) corresponding to each of the 20 amino acids required for protein synthesis. Although each specific tRNA differs from the others in its sequence of nucleotides, the tRNA molecules as a class have many features in common. The primary structure—i.e., the nucleotide sequence—of all tRNA molecules allows extensive folding and intrastrand complementarity to generate a secondary structure that appears like a cloverleaf (Figure 35–11).

All tRNA molecules contain four main arms. The **acceptor arm** terminates in the nucleotides CpCpAOH. These three nucleotides are added posttranscriptionally. The tRNA-appropriate amino acid is attached to the 3'-OH group of the A moiety of the acceptor arm.

The **D**, **TΨC**, and **extra arms** help define a specific tRNA.

Although tRNAs are quite stable in prokaryotes, they are somewhat less stable in eukaryotes. The opposite is true for mRNAs, which are quite unstable in prokaryotes but generally stable in eukaryotic organisms.

C. RIBOSOMAL RNA (rRNA)

A ribosome is a cytoplasmic nucleoprotein structure that acts as the machinery for the synthesis of proteins from the mRNA templates. On the ribosomes, the mRNA and tRNA molecules interact to translate into a specific protein molecule information transcribed from the gene. In active protein synthesis, many ribosomes are associated with an mRNA molecule in an assembly called the **polysome**.

The components of the mammalian ribosome, which has a molecular weight of about 4.2×10^6 and a sedimentation velocity of 80S (Svedberg units), are shown in Table 35–2. The mammalian ribosome contains two major nucleoprotein subunits—a larger one with a molecular weight of 2.8×10^6 (60S) and a smaller subunit with a molecular weight of 1.4×10^6 (40S). The 60S subunit contains a 5S ribosomal RNA (rRNA), a 5.8S rRNA, and a 28S rRNA; there are also probably more than 50 specific polypeptides. The 40S subunit is smaller and contains a single 18S rRNA and approximately 30 distinct polypeptide chains. All of the ribosomal RNA molecules except the 5S rRNA are processed from a single 45S precursor RNA molecule in the nucleolus (Chapter 37). 5S rRNA is independently transcribed. The highly methylated ribosomal RNA molecules are packaged in the nucleolus with the specific ribosomal proteins. In the cytoplasm, the ribosomes remain quite stable and capable of many translation cycles. The functions of the ribosomal RNA molecules in the ribosomal particle are not fully understood, but they are necessary for ribosomal assembly and seem to play key roles in the binding of mRNA to

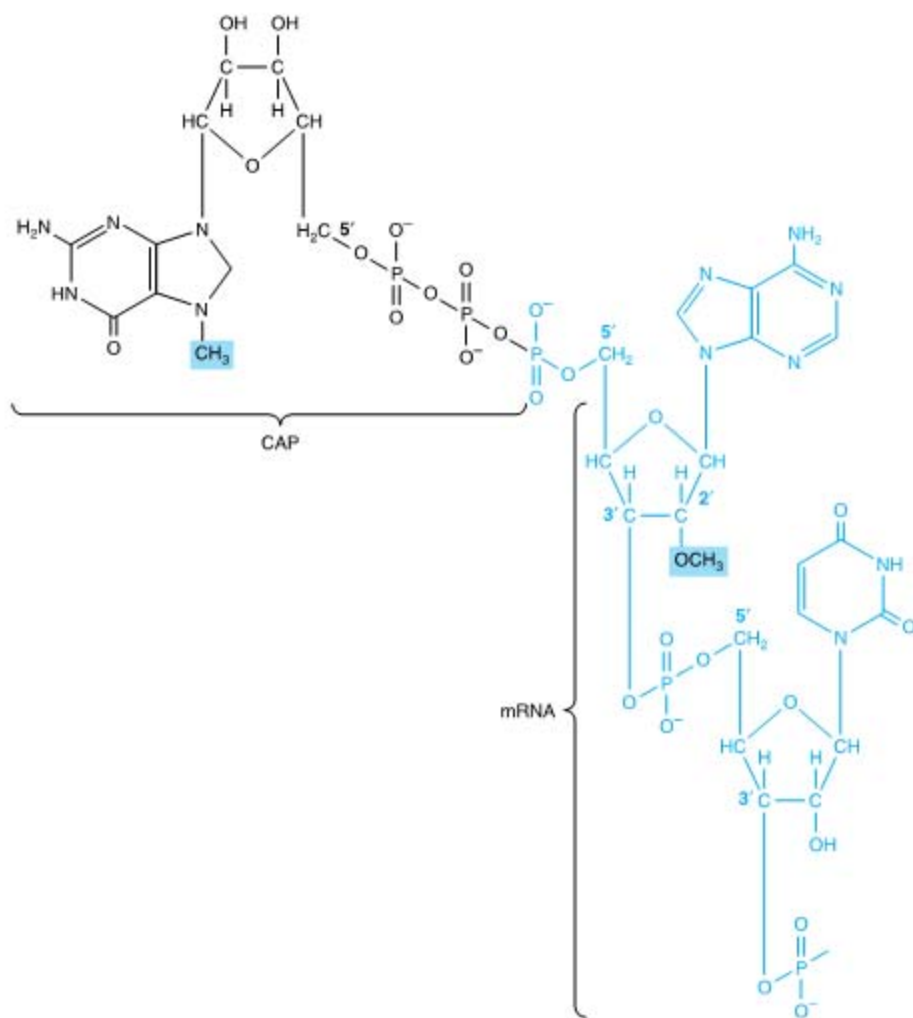


Figure 35–10. The cap structure attached to the 5' terminal of most eukaryotic messenger RNA molecules. A 7-methylguanosine triphosphate (black) is attached at the 5' terminal of the mRNA (shown in blue), which usually contains a 2'-O-methylpurine nucleotide. These modifications (the cap and methyl group) are added after the mRNA is transcribed from DNA.

ribosomes and its translation. Recent studies suggest that an rRNA component performs the peptidyl transferase activity and thus is an enzyme (a ribozyme).

D. SMALL STABLE RNA

A large number of discrete, highly conserved, and small stable RNA species are found in eukaryotic cells. The majority of these molecules are complexed with proteins to form ribonucleoproteins and are distributed in the nucleus, in the cytoplasm, or in both. They range in

size from 90 to 300 nucleotides and are present in 100,000–1,000,000 copies per cell.

Small nuclear RNAs (snRNAs), a subset of these RNAs, are significantly involved in mRNA processing and gene regulation. Of the several snRNAs, U1, U2, U4, U5, and U6 are involved in intron removal and the processing of hnRNA into mRNA (Chapter 37). The U7 snRNA may be involved in production of the correct 3' ends of histone mRNA—which lacks a poly(A) tail. The U4 and U6 snRNAs may also be required for poly(A) processing.

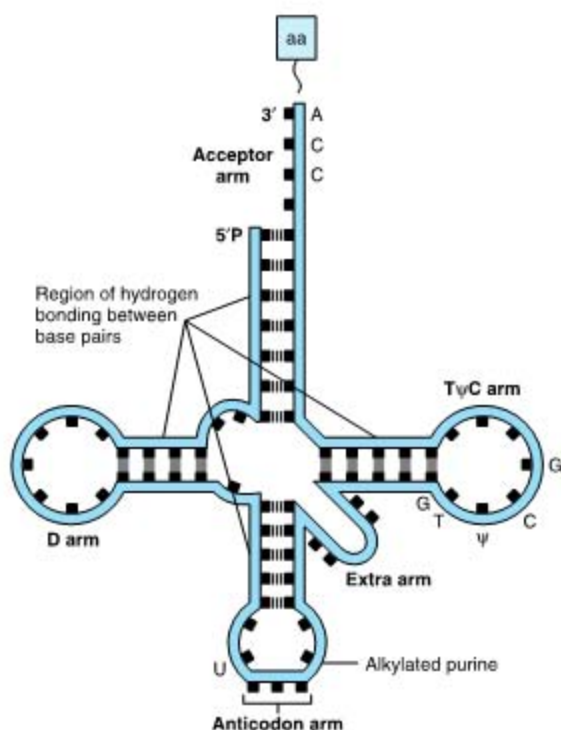


Figure 35–11. Typical aminoacyl tRNA in which the amino acid (aa) is attached to the 3' CCA terminal. The anticodon, TΨC, and dihydrouracil (D) arms are indicated, as are the positions of the intramolecular hydrogen bonding between these base pairs. (From Watson JD: *Molecular Biology of the Gene*, 3rd ed. Copyright © 1976, 1970, 1965, by W.A. Benjamin, Inc., Menlo Park, California.)

SPECIFIC NUCLEASES DIGEST NUCLEIC ACIDS

Enzymes capable of degrading nucleic acids have been recognized for many years. These nucleases can be classified in several ways. Those which exhibit specificity for deoxyribonucleic acid are referred to as **deoxyribonucleases**. Those which specifically hydrolyze ribonucleic acids are **ribonucleases**. Within both of these classes are enzymes capable of cleaving internal phosphodiester bonds to produce either 3'-hydroxyl and 5'-phosphoryl terminals or 5'-hydroxyl and 3'-phosphoryl terminals. These are referred to as **endonucleases**. Some are capable of hydrolyzing both strands of a **double-stranded** molecule, whereas others can only cleave **single strands** of nucleic acids. Some nucleases can hydrolyze only unpaired single strands, while others are capable of hydrolyzing single strands participating in the formation of a double-stranded molecule. There exist classes of endonucleases that recognize specific sequences in DNA; the majority of these are the **restriction endonucleases**, which have in recent years become important tools in molecular genetics and medical sciences. A list of some currently recognized restriction endonucleases is presented in Table 40–2.

Some nucleases are capable of hydrolyzing a nucleotide only when it is present at a terminal of a molecule; these are referred to as **exonucleases**. Exonucleases act in one direction (3' → 5' or 5' → 3') only. In bacteria, a 3' → 5' exonuclease is an integral part of the DNA replication machinery and there serves to edit—or proofread—the most recently added deoxynucleotide for base-pairing errors.

Table 35–2. Components of mammalian ribosomes.¹

Component	Mass (mw)	Protein Number	Protein Mass	RNA Size	RNA Mass	Bases
40S subunit	1.4×10^6	~35	7×10^5	18S	7×10^5	1900
60S subunit	2.8×10^6	~50	1×10^6	5S	35,000	120
				5.8S	45,000	160
				28S	1.6×10^6	4700

¹The ribosomal subunits are defined according to their sedimentation velocity in Svedberg units (40S or 60S). This table illustrates the total mass (MW) of each. The number of unique proteins and their total mass (MW) and the RNA components of each subunit in size (Svedberg units), mass, and number of bases are listed.

SUMMARY

- DNA consists of four bases—A, G, C, and T—which are held in linear array by phosphodiester bonds through the 3' and 5' positions of adjacent deoxyribose moieties.
- DNA is organized into two strands by the pairing of bases A to T and G to C on complementary strands. These strands form a double helix around a central axis.
- The 3×10^9 base pairs of DNA in humans are organized into the haploid complement of 23 chromosomes. The exact sequence of these 3 billion nucleotides defines the uniqueness of each individual.
- DNA provides a template for its own replication and thus maintenance of the genotype and for the transcription of the 30,000–50,000 genes into a variety of RNA molecules.
- RNA exists in several different single-stranded structures, most of which are involved in protein synthe-

sis. The linear array of nucleotides in RNA consists of A, G, C, and U, and the sugar moiety is ribose.

- The major forms of RNA include messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA). Certain RNA molecules act as catalysts (ribozymes).

REFERENCES

- Green R, Noller HF: Ribosomes and translation. *Annu Rev Biochem* 1997;66:689.
- Guthrie C, Patterson B: Spliceosomal snRNAs. *Ann Rev Genet* 1988;22:387.
- Hunt T: *DNA Makes RNA Makes Protein*. Elsevier, 1983.
- Watson JD, Crick FHC: Molecular structure of nucleic acids. *Nature* 1953;171:737.
- Watson JD: *The Double Helix*. Atheneum, 1968.
- Watson JD et al: *Molecular Biology of the Gene*, 5th ed. Benjamin-Cummings, 2000.

DNA Organization, Replication, & Repair

36

Daryl K. Granner, MD, & P. Anthony Weil, PhD

BIOMEDICAL IMPORTANCE*

The genetic information in the DNA of a chromosome can be transmitted by exact replication or it can be exchanged by a number of processes, including crossing over, recombination, transposition, and conversion. These provide a means of ensuring adaptability and diversity for the organism but, when these processes go awry, can also result in disease. A number of enzyme systems are involved in DNA replication, alteration, and repair. Mutations are due to a change in the base sequence of DNA and may result from the faulty replication, movement, or repair of DNA and occur with a frequency of about one in every 10^6 cell divisions. Abnormalities in gene products (either in protein function or amount) can be the result of mutations that occur in coding or regulatory-region DNA. A mutation in a germ cell will be transmitted to offspring (so-called vertical transmission of hereditary disease). A number of factors, including viruses, chemicals, ultraviolet light, and ionizing radiation, increase the rate of mutation. Mutations often affect somatic cells and so are passed on to successive generations of cells, but only within an organism. It is becoming apparent that a number of diseases—and perhaps most cancers—are due to the combined effects of vertical transmission of mutations as well as horizontal transmission of induced mutations.

CHROMATIN IS THE CHROMOSOMAL MATERIAL EXTRACTED FROM NUCLEI OF CELLS OF EUKARYOTIC ORGANISMS

Chromatin consists of very long double-stranded DNA molecules and a nearly equal mass of rather small basic proteins termed **histones** as well as a smaller amount of **nonhistone proteins** (most of which are acidic and

larger than histones) and a small quantity of **RNA**. The nonhistone proteins include enzymes involved in DNA replication, such as DNA topoisomerases. Also included are proteins involved in transcription, such as the RNA polymerase complex. The double-stranded DNA helix in each chromosome has a length that is thousands of times the diameter of the cell nucleus. One purpose of the molecules that comprise chromatin, particularly the histones, is to condense the DNA. Electron microscopic studies of chromatin have demonstrated dense spherical particles called **nucleosomes**, which are approximately 10 nm in diameter and connected by DNA filaments (Figure 36-1). Nucleosomes are composed of DNA wound around a collection of histone molecules.

Histones Are the Most Abundant Chromatin Proteins

The histones are a small family of closely related basic proteins. **H1 histones** are the ones least tightly bound to chromatin (Figure 36-1) and are, therefore, easily removed with a salt solution, after which chromatin becomes soluble. The organizational unit of this soluble chromatin is the nucleosome. Nucleosomes contain four classes of histones: **H2A, H2B, H3, and H4**. The structures of all four histones—H2A, H2B, H3, and H4, the so-called core histones forming the nucleosome—have been highly conserved between species. This extreme conservation implies that the function of histones is identical in all eukaryotes and that the entire molecule is involved quite specifically in carrying out this function. The carboxyl terminal two-thirds of the molecules have a typical random amino acid composition, while their amino terminal thirds are particularly rich in basic amino acids. **These four core histones are subject to at least five types of covalent modification:** acetylation, methylation, phosphorylation, ADP-ribosylation, and covalent linkage (H2A only) to ubiquitin. These histone modifications probably play an important role in chromatin structure and function as illustrated in Table 36-1.

The histones interact with each other in very specific ways. **H3 and H4 form a tetramer** containing two mol-

*So far as is possible, the discussion in this chapter and in Chapters 37, 38, and 39 will pertain to mammalian organisms, which are, of course, among the higher eukaryotes. At times it will be necessary to refer to observations in prokaryotic organisms such as bacteria and viruses, but in such cases the information will be of a kind that can be extrapolated to mammalian organisms.

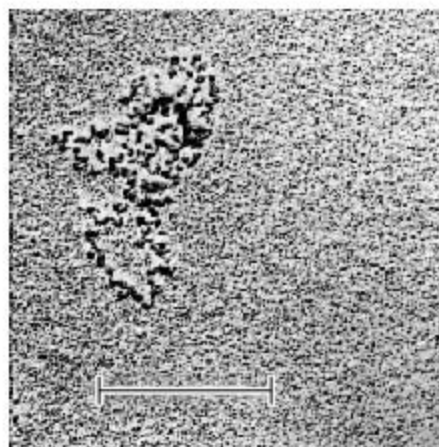


Figure 36–1. Electron micrograph of nucleosomes attached by strands of nucleic acid. (The bar represents 2.5 μm .) (Reproduced, with permission, from Oudet P, Gross-Bellard M, Chambon P: Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell* 1975;4:281.)

ecules of each $(\text{H3/H4})_2$, while **H2A and H2B form dimers** (H2A-H2B). Under physiologic conditions, these histone oligomers associate to form the **histone octamer** of the composition $(\text{H3/H4})_2-(\text{H2A-H2B})_2$.

The Nucleosome Contains Histone & DNA

When the histone octamer is mixed with purified, double-stranded DNA, the same x-ray diffraction pattern is formed as that observed in freshly isolated chromatin. Electron microscopic studies confirm the existence of reconstituted nucleosomes. Furthermore, the reconstitution of nucleosomes from DNA and histones H2A, H2B, H3, and H4 is independent of the organismal or cellular origin of the various components. The histone H1 and the nonhistone proteins are not necessary for the reconstitution of the nucleosome core.

Table 36–1. Possible roles of modified histones.

1. Acetylation of histones H3 and H4 is associated with the activation or inactivation of gene transcription (Chapter 37).
2. Acetylation of core histones is associated with chromosomal assembly during DNA replication.
3. Phosphorylation of histone H1 is associated with the condensation of chromosomes during the replication cycle.
4. ADP-ribosylation of histones is associated with DNA repair.
5. Methylation of histones is correlated with activation and repression of gene transcription.

In the nucleosome, the DNA is supercoiled in a left-handed helix over the surface of the disk-shaped histone octamer (Figure 36–2). The majority of core histone proteins interact with the DNA on the inside of the supercoil without protruding, though the amino terminal tails of all the histones probably protrude outside of this structure and are available for regulatory covalent modifications (see Table 36–1).

The $(\text{H3/H4})_2$ tetramer itself can confer nucleosome-like properties on DNA and thus has a central role in the formation of the nucleosome. The addition of two H2A-H2B dimers stabilizes the primary particle and firmly binds two additional half-turns of DNA previously bound only loosely to the $(\text{H3/H4})_2$. Thus, 1.75 superhelical turns of DNA are wrapped around the surface of the histone octamer, protecting 146 base pairs of DNA and forming the nucleosome core particle (Figure 36–2). The core particles are separated by an about 30-bp linker region of DNA. Most of the DNA is in a repeating series of these structures, giving the so-called “beads-on-a-string” appearance when examined by electron microscopy (see Figure 36–1).

The assembly of nucleosomes is mediated by one of several chromatin assembly factors facilitated by histone chaperones, proteins such as the anionic nuclear protein **nucleoplasmin**. As the nucleosome is assembled, histones are released from the histone chaperones. Nucleosomes appear to exhibit preference for certain regions on specific DNA molecules, but the basis for this nonrandom distribution, termed **phasing**, is not completely

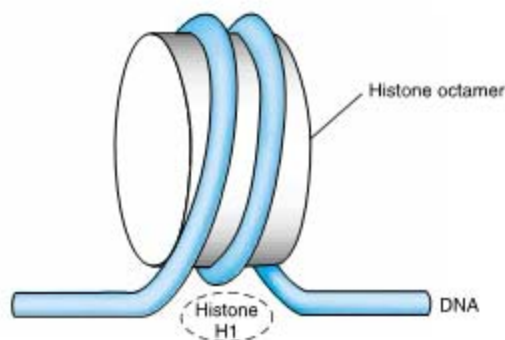


Figure 36–2. Model for the structure of the nucleosome, in which DNA is wrapped around the surface of a flat protein cylinder consisting of two each of histones H2A, H2B, H3, and H4 that form the histone octamer. The 146 base pairs of DNA, consisting of 1.75 superhelical turns, are in contact with the histone octamer. This protects the DNA from digestion by a nuclease. The position of histone H1, when it is present, is indicated by the dashed outline at the bottom of the figure.

understood. It is probably related to the relative physical flexibility of certain nucleotide sequences that are able to accommodate the regions of kinking within the supercoil as well as the presence of other DNA-bound factors that limit the sites of nucleosome deposition.

The super-packing of nucleosomes in nuclei is seemingly dependent upon the interaction of the H1 histones with adjacent nucleosomes.

HIGHER-ORDER STRUCTURES PROVIDE FOR THE COMPACTION OF CHROMATIN

Electron microscopy of chromatin reveals two higher orders of structure—the 10-nm fibril and the 30-nm chromatin fiber—beyond that of the nucleosome itself. The disk-like nucleosome structure has a 10-nm diameter and a height of 5 nm. The **10-nm fibril** consists of nucleosomes arranged with their edges separated by a small distance (30 bp of DNA) with their flat faces parallel with the fibril axis (Figure 36-3). The 10-nm fibril is probably further supercoiled with six or seven nucleosomes per turn to form the **30-nm chromatin fiber** (Figure 36-3). Each turn of the supercoil is relatively flat, and the faces of the nucleosomes of successive turns would be nearly parallel to each other. H1 histones appear to stabilize the 30-nm fiber, but their position and that of the variable length spacer DNA are not clear. It is probable that nucleosomes can form a variety of packed structures. In order to form a mitotic chromosome, the 30-nm fiber must be compacted in length another 100-fold (see below).

In **interphase chromosomes**, chromatin fibers appear to be organized into 30,000–100,000 bp **loops or domains** anchored in a scaffolding (or supporting matrix) within the nucleus. Within these domains, some DNA sequences may be located nonrandomly. It has been suggested that each looped domain of chromatin corresponds to one or more separate genetic functions, containing both coding and noncoding regions of the cognate gene or genes.

SOME REGIONS OF CHROMATIN ARE “ACTIVE” & OTHERS ARE “INACTIVE”

Generally, every cell of an individual metazoan organism contains the same genetic information. Thus, the differences between cell types within an organism must be explained by differential expression of the common genetic information. Chromatin containing active genes (i.e., transcriptionally active chromatin) has been shown to differ in several ways from that of inactive regions. The nucleosome structure of active chromatin appears to be altered, sometimes quite extensively, in highly active regions. DNA in active chromatin contains large regions (about 100,000 bases long) that are **sensitive to diges-**

tion by a nuclease such as DNase I. DNase I makes single-strand cuts in any segment of DNA (no sequence specificity). It will digest DNA not protected by protein into its component deoxynucleotides. The sensitivity to DNase I of chromatin regions being actively transcribed reflects only a potential for transcription rather than transcription itself and in several systems can be correlated with a relative lack of 5-methyldeoxycytidine in the DNA and particular histone covalent modifications (phosphorylation, acetylation, etc; see Table 36-1).

Within the large regions of active chromatin there exist shorter stretches of 100–300 nucleotides that exhibit an even greater (another tenfold) sensitivity to DNase I. These **hypersensitive sites** probably result from a structural conformation that favors access of the nuclease to the DNA. These regions are often located immediately upstream from the active gene and are the location of interrupted nucleosomal structure caused by the binding of nonhistone regulatory transcription factor proteins. (See Chapters 37 and 39.) In many cases, it seems that if a gene is capable of being transcribed, it very often has a DNase-hypersensitive site(s) in the chromatin immediately upstream. As noted above, nonhistone regulatory proteins involved in transcription control and those involved in maintaining access to the template strand lead to the formation of hypersensitive sites. Hypersensitive sites often provide the first clue about the presence and location of a transcription control element.

Transcriptionally inactive chromatin is densely packed during interphase as observed by electron microscopic studies and is referred to as **heterochromatin**; transcriptionally active chromatin stains less densely and is referred to as **euchromatin**. Generally, euchromatin is replicated earlier than heterochromatin in the mammalian cell cycle (see below).

There are two types of heterochromatin: constitutive and facultative. **Constitutive heterochromatin** is always condensed and thus inactive. It is found in the regions near the chromosomal centromere and at chromosomal ends (telomeres). **Facultative heterochromatin** is at times condensed, but at other times it is actively transcribed and, thus, uncondensed and appears as euchromatin. Of the two members of the X chromosome pair in mammalian females, one X chromosome is almost completely inactive transcriptionally and is heterochromatic. However, the heterochromatic X chromosome decondenses during gametogenesis and becomes transcriptionally active during early embryogenesis—thus, it is facultative heterochromatin.

Certain cells of insects, eg, *Chironomus*, contain giant chromosomes that have been replicated for ten cycles without separation of daughter chromatids. These copies of DNA line up side by side in precise register and produce a banded chromosome containing regions of condensed chromatin and lighter bands of

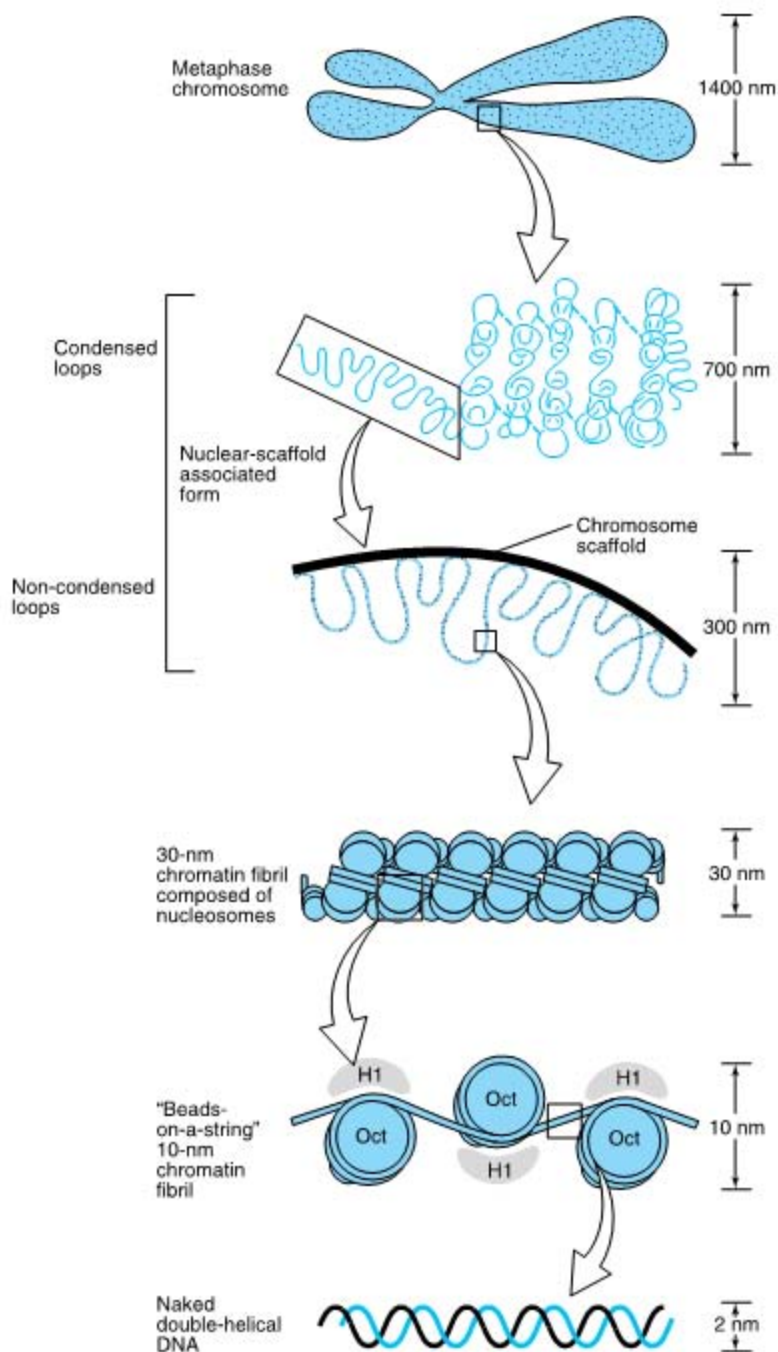


Figure 36-3. Shown is the extent of DNA packaging in metaphase chromosomes (*top*) to naked duplex DNA (*bottom*). Chromosomal DNA is packaged and organized at several levels as shown (see Table 36-2). Each phase of condensation or compaction and organization (*bottom to top*) decreases overall DNA accessibility to an extent that the DNA sequences in metaphase chromosomes are almost totally transcriptionally inert. In toto, these five levels of DNA compaction result in nearly a 10^4 -fold linear decrease in end-to-end DNA length. Complete condensation and decondensation of the linear DNA in chromosomes occur in the space of hours during the normal replicative cell cycle (see Figure 36-20).

more extended chromatin. Transcriptionally active regions of these **polytene chromosomes** are especially decondensed into “**puffs**” that can be shown to contain the enzymes responsible for transcription and to be the sites of RNA synthesis (Figure 36-4).

DNA IS ORGANIZED INTO CHROMOSOMES

At metaphase, mammalian **chromosomes** possess a twofold symmetry, with the identical duplicated **sister chromatids** connected at a **centromere**, the relative po-

sition of which is characteristic for a given chromosome (Figure 36-5). The centromere is an adenine-thymine (A-T) rich region ranging in size from 10^2 (brewer's yeast) to 10^6 (mammals) base pairs. It binds several proteins with high affinity. This complex, called the **kinetochore**, provides the anchor for the mitotic spindle. It thus is an essential structure for chromosomal segregation during mitosis.

The ends of each chromosome contain structures called **telomeres**. Telomeres consist of short, repeat TG-rich sequences. Human telomeres have a variable number of repeats of the sequence 5'-TTAGGG-3', which can extend for several kilobases. **Telomerase**, a multisubunit RNA-containing complex related to viral RNA-dependent DNA polymerases (reverse transcriptases), is the enzyme responsible for telomere synthesis and thus for maintaining the length of the telomere. Since telomere shortening has been associated with both malignant transformation and aging, telomerase has become an attractive target for cancer chemotherapy and drug development. Each sister chromatid contains one double-stranded DNA molecule. During interphase, the packing of the DNA molecule is less dense than it is in the condensed chromosome during metaphase. Metaphase chromosomes are nearly completely transcriptionally inactive.

The human haploid genome consists of about 3×10^9 bp and about 1.7×10^7 nucleosomes. Thus, each of the 23 chromatids in the human haploid genome would contain on the average 1.3×10^8 nucleotides in one double-stranded DNA molecule. The length of each DNA molecule must be compressed about 8000-fold to generate the structure of a condensed metaphase chromosome! In metaphase chromosomes, the 30-nm chromatin fibers are also folded into a series of **looped domains**, the proximal portions of which are anchored to a nonhistone proteinaceous scaffolding within the nucleus (Figure 36-3). The packing ratios of each of the orders of DNA structure are summarized in Table 36-2.

The packaging of nucleoproteins within chromatids is not random, as evidenced by the characteristic patterns observed when chromosomes are stained with specific dyes such as quinacrine or Giemsa's stain (Figure 36-6).

From individual to individual within a single species, the pattern of staining (banding) of the entire chromosome complement is highly reproducible; nonetheless, it differs significantly from other species, even those closely related. Thus, the packaging of the nucleoproteins in chromosomes of higher eukaryotes must in some way be dependent upon species-specific characteristics of the DNA molecules.

A combination of specialized staining techniques and high-resolution microscopy has allowed geneticists

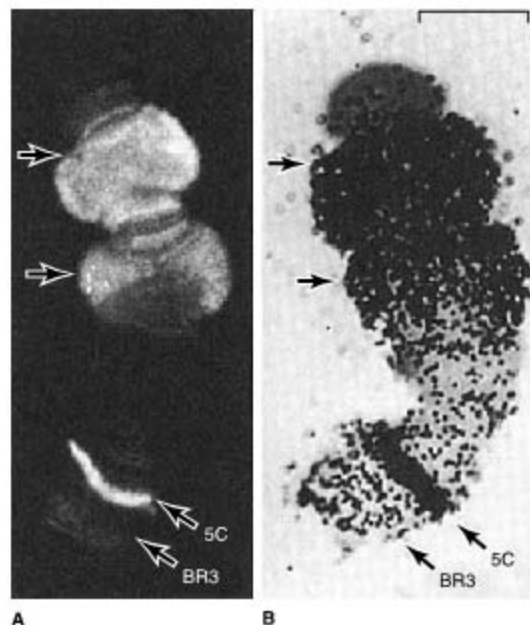
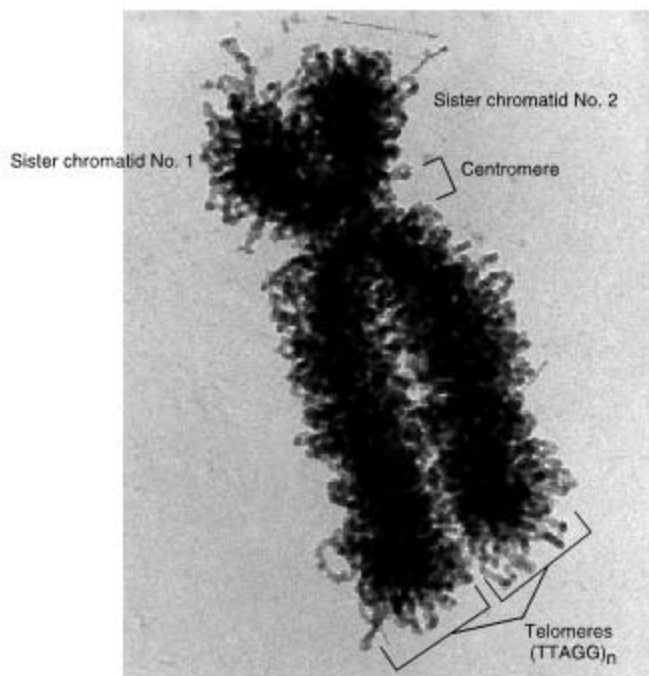


Figure 36-4. Illustration of the tight correlation between the presence of RNA polymerase II and RNA synthesis. A number of genes are activated when *Chironomus tentans* larvae are subjected to heat shock (39 °C for 30 minutes). **A:** Distribution of RNA polymerase II (also called type B) in isolated chromosome IV from the salivary gland (at arrows). The enzyme was detected by immunofluorescence using an antibody directed against the polymerase. The 5C and BR3 are specific bands of chromosome IV, and the arrows indicate puffs. **B:** Autoradiogram of a chromosome IV that was incubated in ^3H -uridine to label the RNA. Note the correspondence of the immunofluorescence and presence of the radioactive RNA (black dots). Bar = 7 μm . (Reproduced, with permission, from Sass H: RNA polymerase B in polytene chromosomes. Cell 1982;28:274. Copyright © 1982 by the Massachusetts Institute of Technology.)

Figure 36–5. The two sister chromatids of human chromosome 12 ($\times 27,850$). The location of the A+T-rich centromeric region connecting sister chromatids is indicated, as are two of the four telomeres residing at the very ends of the chromatids that are attached one to the other at the centromere. (Modified and reproduced, with permission, from DuPrav EJ: *DNA and Chromosomes*. Holt, Rinehart, and Winston, 1970.)



to quite precisely map thousands of genes to specific regions of mouse and human chromosomes. With the recent elucidation of the human and mouse genome sequences, it has become clear that many of these visual mapping methods were remarkably accurate.

Coding Regions Are Often Interrupted by Intervening Sequences

The **protein coding regions of DNA**, the transcripts of which ultimately appear in the cytoplasm as single mRNA molecules, are usually **interrupted in the eukaryotic genome by large intervening sequences of**

nonprotein coding DNA. Accordingly, the primary transcripts of DNA (mRNA precursors, originally termed hnRNA because this species of RNA was quite heterogeneous in size [length] and mostly restricted to the nucleus), contain noncoding intervening sequences of RNA that must be removed in a process which also joins together the appropriate coding segments to form the mature mRNA. Most coding sequences for a single mRNA are interrupted in the genome (and thus in the primary transcript) by at least one—and in some cases as many as 50—noncoding intervening sequences (**introns**). In most cases, the introns are much longer than the continuous coding regions (**exons**). The processing of the primary transcript, which involves removal of introns and splicing of adjacent exons, is described in detail in Chapter 37.

The function of the intervening sequences, or introns, is not clear. They may serve to separate functional domains (exons) of coding information in a form that permits genetic rearrangement by recombination to occur more rapidly than if all coding regions for a given genetic function were contiguous. Such an enhanced rate of genetic rearrangement of functional domains might allow more rapid evolution of biologic function. The relationships among chromosomal DNA, gene clusters on the chromosome, the exon-intron structure of genes, and the final mRNA product are illustrated in Figure 36–7.

Table 36–2. The packing ratios of each of the orders of DNA structure.

Chromatin Form	Packing Ratio
Naked double-helical DNA	~1.0
10-nm fibril of nucleosomes	7–10
25- to 30-nm chromatin fiber of superhelical nucleosomes	40–60
Condensed metaphase chromosome of loops	8000

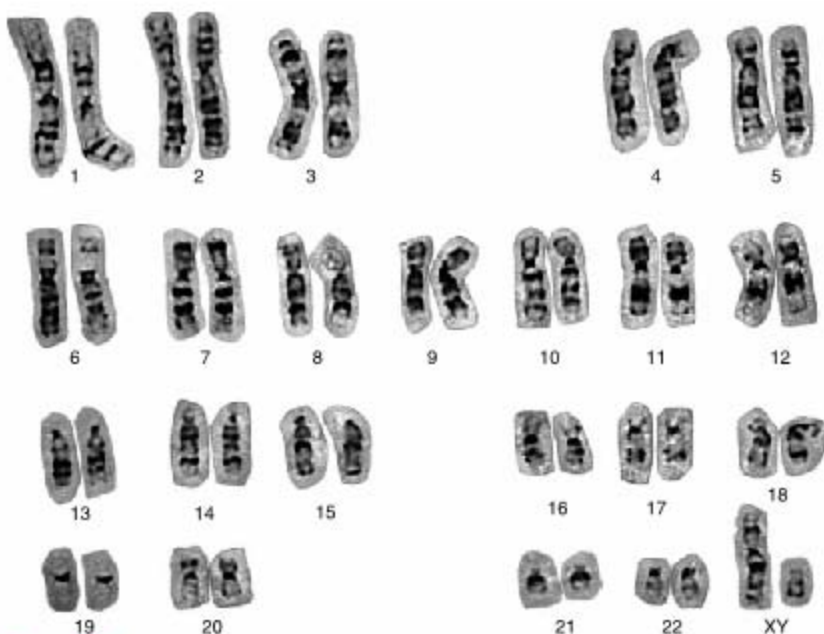


Figure 36–6. A human karyotype (of a man with a normal 46,XY constitution), in which the metaphase chromosomes have been stained by the Giemsa method and aligned according to the Paris Convention. (Courtesy of H Lawce and F Conte.)

MUCH OF THE MAMMALIAN GENOME IS REDUNDANT & MUCH IS NOT TRANSCRIBED

The haploid genome of each human cell consists of 3×10^9 base pairs of DNA subdivided into 23 chromosomes. The entire haploid genome contains sufficient DNA to code for nearly 1.5 million average-sized genes. However, studies of mutation rates and of the complexities of the genomes of higher organisms strongly suggest that humans have $< 100,000$ proteins encoded by the $\sim 1.1\%$ of the human genome that is composed of exonic DNA. This implies that most of the DNA is noncoding—ie, its information is never translated into an amino acid sequence of a protein molecule. Certainly, some of the excess DNA sequences serve to regulate the expression of genes during development, differentiation, and adaptation to the environment. Some excess clearly makes up the intervening sequences or introns (24% of the total human genome) that split the coding regions of genes, but much of the excess appears to be composed of many families of repeated sequences for which no functions have been clearly defined. A summary of the salient features of the human genome is presented in Chapter 40.

The DNA in a eukaryotic genome can be divided into different “sequence classes.” These are **unique-sequence**, or **nonrepetitive**, DNA and **repetitive-sequence DNA**. In the haploid genome, unique-sequence DNA generally includes the single copy genes that code for proteins. The repetitive DNA in the haploid genome includes sequences that vary in copy number from two to as many as 10^7 copies per cell.

More Than Half the DNA in Eukaryotic Organisms Is in Unique or Nonrepetitive Sequences

This estimation (and the distribution of repetitive-sequence DNA) is based on a variety of DNA-RNA hybridization techniques and, more recently, on direct DNA sequencing. Similar techniques are used to estimate the number of active genes in a population of unique-sequence DNA. In brewers’ yeast (*Saccharomyces cerevisiae*, a lower eukaryote), about two thirds of its 6200 genes are expressed. In typical tissues in a higher eukaryote (eg, mammalian liver and kidney), between 10,000 and 15,000 genes are expressed. Different combinations of genes are expressed in each tissue,

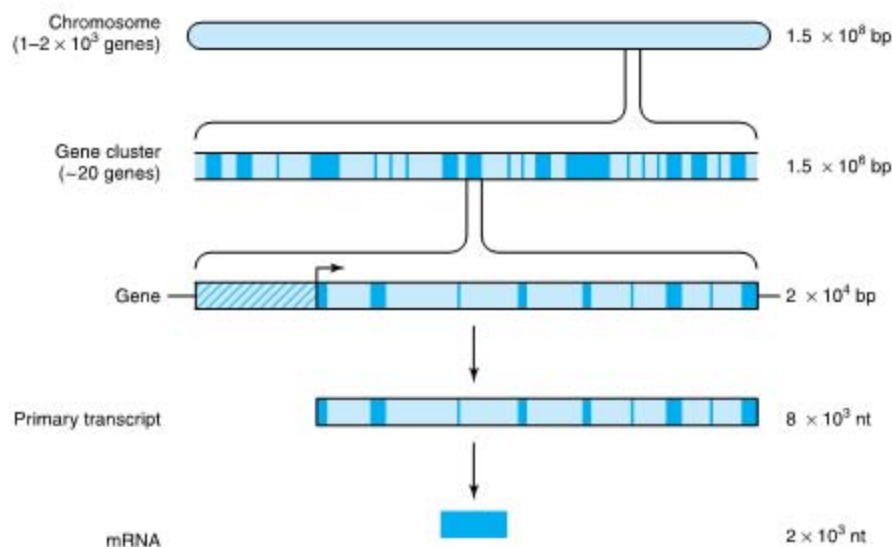


Figure 36-7. The relationship between chromosomal DNA and mRNA. The human haploid DNA complement of 3×10^9 base pairs (bp) is distributed between 23 chromosomes. Genes are clustered on these chromosomes. An average gene is 2×10^4 bp in length, including the regulatory region (hatched area), which is usually located at the 5' end of the gene. The regulatory region is shown here as being adjacent to the transcription initiation site (arrow). Most eukaryotic genes have alternating exons and introns. In this example, there are nine exons (dark blue areas) and eight introns (light blue areas). The introns are removed from the primary transcript by the processing reaction, and the exons are ligated together in sequence to form the mature mRNA. (nt, nucleotides.)

of course, and how this is accomplished is one of the major unanswered questions in biology.

In Human DNA, at Least 30% of the Genome Consists of Repetitive Sequences

Repetitive-sequence DNA can be broadly classified as moderately repetitive or as highly repetitive. The highly repetitive sequences consist of 5–500 base pair lengths repeated many times in tandem. These sequences are usually clustered in centromeres and telomeres of the chromosome and are present in about 1–10 million copies per haploid genome. These sequences are transcriptionally inactive and may play a structural role in the chromosome (see Chapter 40).

The moderately repetitive sequences, which are defined as being present in numbers of less than 10^6 copies per haploid genome, are not clustered but are interspersed with unique sequences. In many cases, these long interspersed repeats are transcribed by RNA polymerase II and contain caps indistinguishable from those on mRNA.

Depending on their length, moderately repetitive sequences are classified as **long interspersed repeat sequences (LINEs)** or **short interspersed repeat sequences (SINEs)**. Both types appear to be **retrotransposons**, i.e., they arose from movement from one location to another (**transposition**) through an RNA intermediate by the action of reverse transcriptase that transcribes an RNA template into DNA. Mammalian genomes contain 20–50 thousand copies of the 6–7 kb LINEs. These represent species-specific families of repeat elements. SINEs are shorter (70–300 bp), and there may be more than 100,000 copies per genome. Of the SINEs in the human genome, one family, the **Alu family**, is present in about 500,000 copies per haploid genome and accounts for at least 5–6% of the human genome. Members of the human Alu family and their closely related analogs in other animals are transcribed as integral components of hnRNA or as discrete RNA molecules, including the well-studied 4.5S RNA and 7S RNA. These particular family members are highly conserved within a species as well as between mammalian species. Components of the short inter-

scattered repeats, including the members of the Alu family, may be mobile elements, capable of jumping into and out of various sites within the genome (see below). This can have disastrous results, as exemplified by the insertion of Alu sequences into a gene, which, when so mutated, causes neurofibromatosis.

Microsatellite Repeat Sequences

One category of repeat sequences exists as both dispersed and grouped tandem arrays. The sequences consist of 2–6 bp repeated up to 50 times. These **microsatellite sequences** most commonly are found as dinucleotide repeats of AC on one strand and TG on the opposite strand, but several other forms occur, including CG, AT, and CA. The AC repeat sequences are estimated to occur at 50,000–100,000 locations in the genome. At any locus, the number of these repeats may vary on the two chromosomes, thus providing heterozygosity of the number of copies of a particular microsatellite number in an individual. This is a heritable trait, and, because of their number and the ease of detecting them using the polymerase chain reaction (PCR) (Chapter 40), AC repeats are very useful in constructing genetic linkage maps. Most genes are associated with one or more microsatellite markers, so the relative position of genes on chromosomes can be assessed, as can the association of a gene with a disease. Using PCR, a large number of family members can be rapidly screened for a certain **microsatellite polymorphism**. The association of a specific polymorphism

with a gene in affected family members—and the lack of this association in unaffected members—may be the first clue about the genetic basis of a disease.

Trinucleotide sequences that increase in number (microsatellite instability) can cause disease. The unstable $p(CGG)_n$ repeat sequence is associated with the fragile X syndrome. Other trinucleotide repeats that undergo dynamic mutation (usually an increase) are associated with Huntington's chorea (CAG), myotonic dystrophy (CTG), spinobulbar muscular atrophy (CAG), and Kennedy's disease (CAG).

ONE PERCENT OF CELLULAR DNA IS IN MITOCHONDRIA

The majority of the peptides in mitochondria (about 54 out of 67) are coded by nuclear genes. The rest are coded by genes found in mitochondrial (mt) DNA. Human mitochondria contain two to ten copies of a small circular double-stranded DNA molecule that makes up approximately 1% of total cellular DNA. This mtDNA codes for mt ribosomal and transfer RNAs and for 13 proteins that play key roles in the respiratory chain. The linearized structural map of the human mitochondrial genes is shown in Figure 36–8. Some of the features of mtDNA are shown in Table 36–3.

An important feature of human mitochondrial mtDNA is that—because all mitochondria are contributed by the ovum during zygote formation—it is transmitted by maternal nonmendelian inheritance.

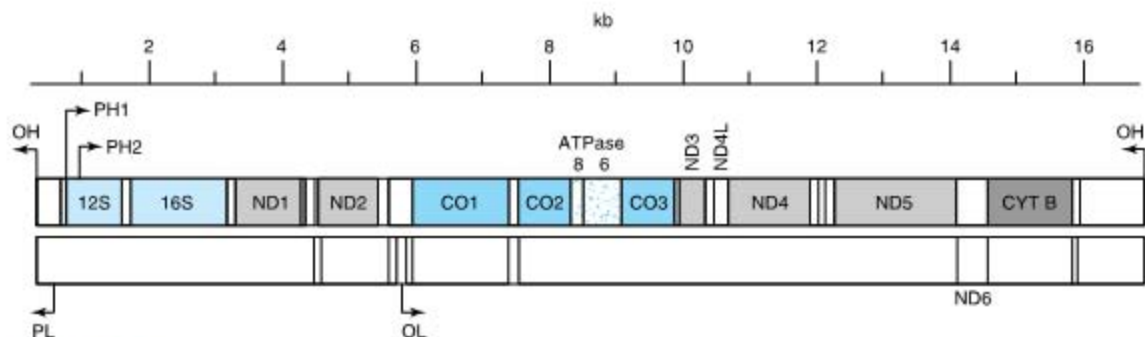


Figure 36–8. Maps of human mitochondrial genes. The maps represent the heavy (upper strand) and light (lower map) strands of linearized mitochondrial (mt) DNA, showing the genes for the subunits of NADH-coenzyme Q oxidoreductase (ND1 through ND6), cytochrome *c* oxidase (CO1 through CO3), cytochrome *b* (CYT B), and ATP synthase (ATPase 8 and 6) and for the 12S and 16S ribosomal mt rRNAs. The transfer RNAs are denoted by small open boxes. The origin of heavy-strand (OH) and light-strand (OL) replication and the promoters for the initiation of heavy-strand (PH1 and PH2) and light-strand (PL) transcription are indicated by arrows. (Reproduced, with permission, from Moraes CT et al: Mitochondrial DNA deletions in progressive external ophthalmoplegia and Kearns-Sayre syndrome. *N Engl J Med* 1989;320:1293.)

Table 36–3. Some major features of the structure and function of human mitochondrial DNA.¹

- Is circular, double-stranded, and composed of heavy (H) and a light (L) chains or strands.
- Contains 16,569 bp.
- Encodes 13 protein subunits of the respiratory chain (of a total of about 67):
 - Seven subunits of NADH dehydrogenase (complex I)
 - Cytochrome *b* of complex III
 - Three subunits of cytochrome oxidase (complex IV)
 - Two subunits of ATP synthase
- Encodes large (16S) and small (12S) mt ribosomal RNAs.
- Encodes 22 mt tRNA molecules.
- Genetic code differs slightly from the standard code:
 - UGA (standard stop codon) is read as Trp.
 - AGA and AGG (standard codons for Arg) are read as stop codons.
- Contains very few untranslated sequences.
- High mutation rate (five to ten times that of nuclear DNA).
- Comparisons of mtDNA sequences provide evidence about evolutionary origins of primates and other species.

¹Adapted from Harding AE: Neurological disease and mitochondrial genes. *Trends Neurol Sci* 1991;14:132.

Thus, in diseases resulting from mutations of mtDNA, an affected mother would in theory pass the disease to all of her children but only her daughters would transmit the trait. However, in some cases, deletions in mtDNA occur during oogenesis and thus are not inherited from the mother. A number of diseases have now been shown to be due to mutations of mtDNA. These include a variety of myopathies, neurologic disorders, and some cases of diabetes mellitus.

GENETIC MATERIAL CAN BE ALTERED & REARRANGED

An alteration in the sequence of purine and pyrimidine bases in a gene due to a change—a removal or an insertion—of one or more bases may result in an altered gene product. Such alteration in the genetic material results in a **mutation** whose consequences are discussed in detail in Chapter 38.

Chromosomal Recombination Is One Way of Rearranging Genetic Material

Genetic information can be exchanged between similar or homologous chromosomes. The exchange or **recombination** event occurs primarily during meiosis in mammalian cells and requires alignment of homologous metaphase chromosomes, an alignment that almost always occurs with great exactness. A process of

crossing over occurs as shown in Figure 36–9. This usually results in an equal and reciprocal exchange of genetic information between homologous chromosomes. If the homologous chromosomes possess different alleles of the same genes, the crossover may produce noticeable and heritable genetic linkage differences. In the rare case where the alignment of homologous chromosomes is not exact, the crossing over or recombination event may result in an unequal exchange of information. One chromosome may receive less genetic material and thus a deletion, while the other partner of the chromosome pair receives more genetic material and thus an insertion or duplication (Figure 36–9). Unequal crossing over does occur in humans, as evidenced by the existence of hemoglobins designated Lepore and anti-Lepore (Figure 36–10). The farther apart two sequences are on an individual chromosome, the greater the likelihood of a crossover recombination

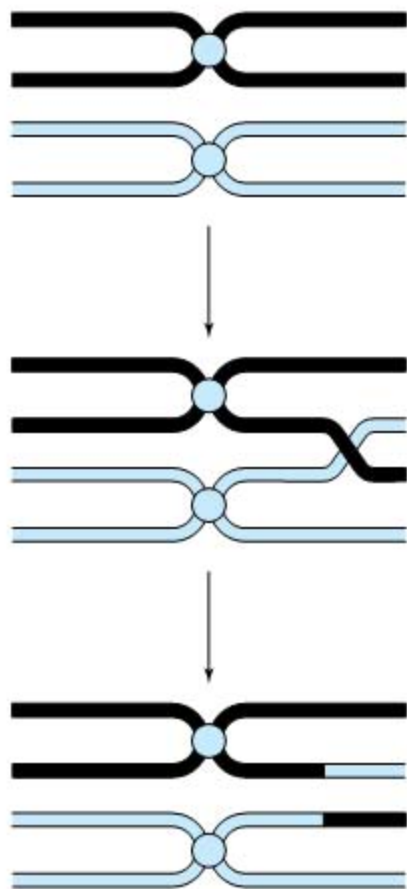


Figure 36–9. The process of crossing-over between homologous metaphase chromosomes to generate recombinant chromosomes. See also Figure 36–12.

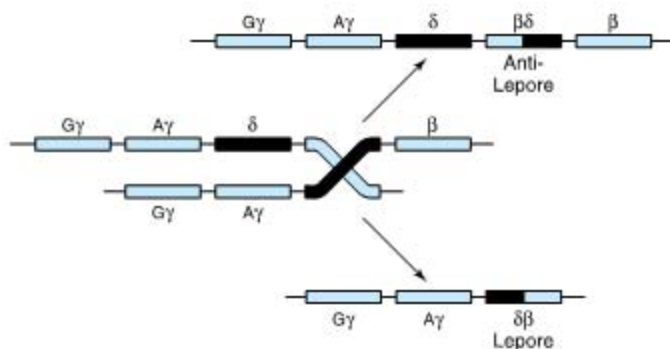


Figure 36–10. The process of unequal crossover in the region of the mammalian genome that harbors the structural genes encoding hemoglobins and the generation of the unequal recombinant products hemoglobin delta-beta Lepore and beta-delta anti-Lepore. The examples given show the locations of the crossover regions between amino acid residues. (Redrawn and reproduced, with permission, from Clegg JB, Weatherall DJ: β^0 Thalassemia: Time for a reappraisal? *Lancet* 1974;2:133.)

event. This is the basis for genetic mapping methods. **Unequal crossover** affects tandem arrays of repeated DNAs whether they are related globin genes, as in Figure 36–10, or more abundant repetitive DNA. Unequal crossover through slippage in the pairing can result in expansion or contraction in the copy number of the repeat family and may contribute to the expansion and fixation of variant members throughout the array.

Chromosomal Integration Occurs With Some Viruses

Some bacterial viruses (bacteriophages) are capable of recombining with the DNA of a bacterial host in such a way that the genetic information of the bacteriophage is incorporated in a linear fashion into the genetic information of the host. This integration, which is a form of recombination, occurs by the mechanism illustrated in Figure 36–11. The backbone of the circular bacteriophage genome is broken, as is that of the DNA molecule of the host; the appropriate ends are resealed with the proper polarity. The bacteriophage DNA is figuratively straightened out (“linearized”) as it is integrated into the bacterial DNA molecule—frequently a closed circle as well. The site at which the bacteriophage genome integrates or recombines with the bacterial genome is chosen by one of two mechanisms. If the bacteriophage contains a DNA sequence **homologous** to a sequence in the host DNA molecule, then a recombination event analogous to that occurring between homologous chromosomes can occur. However, some bacteriophages synthesize proteins that bind specific sites on bacterial chromosomes to a **nonhomologous** site characteristic of the bacteriophage DNA molecule. Integration occurs at the site and is said to be “**site-specific**.”

Many animal viruses, particularly the oncogenic viruses—either directly or, in the case of RNA viruses such as HIV that causes AIDS, their DNA transcripts generated by the action of the viral RNA-dependent

DNA polymerase, or reverse transcriptase—can be integrated into chromosomes of the mammalian cell. The integration of the animal virus DNA into the animal genome generally is not “site-specific” but does display site preferences.

Transposition Can Produce Processed Genes

In eukaryotic cells, small DNA elements that clearly are not viruses are capable of transposing themselves in and

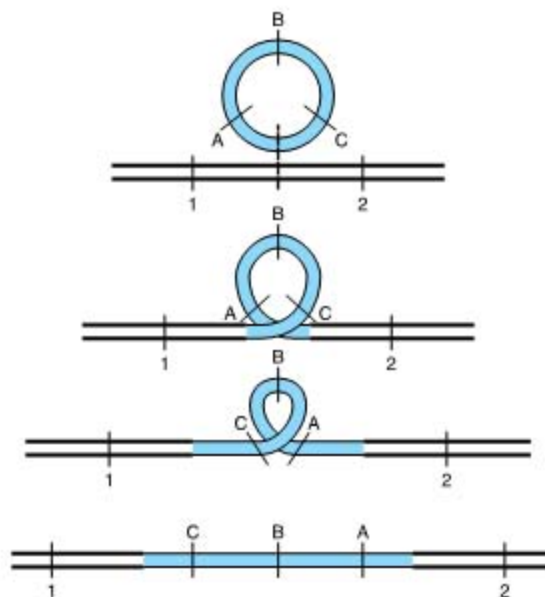


Figure 36–11. The integration of a circular genome from a virus (with genes A, B, and C) into the DNA molecule of a host (with genes 1 and 2) and the consequent ordering of the genes.

out of the host genome in ways that affect the function of neighboring DNA sequences. These mobile elements, sometimes called “jumping DNA,” can carry flanking regions of DNA and, therefore, profoundly affect evolution. As mentioned above, the Alu family of moderately repeated DNA sequences has structural characteristics similar to the termini of retroviruses, which would account for the ability of the latter to move into and out of the mammalian genome.

Direct evidence for the transposition of other small DNA elements into the human genome has been provided by the discovery of “processed genes” for immunoglobulin molecules, α -globin molecules, and several others. These **processed genes** consist of DNA sequences identical or nearly identical to those of the messenger RNA for the appropriate gene product. That is, the 5′ nontranscribed region, the coding region without intron representation, and the 3′ poly(A) tail are all present contiguously. This particular DNA sequence arrangement must have resulted from the reverse transcription of an appropriately processed messenger RNA molecule from which the intron regions had been removed and the poly(A) tail added. The only recognized mechanism this reverse transcript could have used to integrate into the genome would have been a transposition event. In fact, these “processed genes” have short terminal repeats at each end, as do known transposed sequences in lower organisms. In the absence of their transcription and thus genetic selection for function, many of the processed genes have been randomly altered through evolution so that they now contain nonsense codons which preclude their ability to encode a functional, intact protein (see Chapter 38). Thus, they are referred to as “**pseudogenes**.”

Gene Conversion Produces Rearrangements

Besides unequal crossover and transposition, a third mechanism can effect rapid changes in the genetic material. Similar sequences on homologous or nonhomologous chromosomes may occasionally pair up and eliminate any mismatched sequences between them. This may lead to the accidental fixation of one variant or another throughout a family of repeated sequences and thereby homogenize the sequences of the members of repetitive DNA families. This latter process is referred to as **gene conversion**.

Sister Chromatids Exchange

In diploid eukaryotic organisms such as humans, after cells progress through the S phase they contain a tetraploid content of DNA. This is in the form of sister chromatids of chromosome pairs. Each of these sister

chromatids contains identical genetic information since each is a product of the semiconservative replication of the original parent DNA molecule of that chromosome. Crossing over occurs between these genetically identical sister chromatids. Of course, these **sister chromatid exchanges** (Figure 36–12) have no genetic consequence as long as the exchange is the result of an equal crossover.

Immunoglobulin Genes Rearrange

In mammalian cells, some interesting gene rearrangements occur normally during development and differentiation. For example, in mice the V_L and C_L genes for a single immunoglobulin molecule (see Chapter 39) are widely separated in the germ line DNA. In the DNA of a differentiated immunoglobulin-producing (plasma) cell, the same V_L and C_L genes have been moved physically closer together in the genome and into the same transcription unit. However, even then, this rearrangement of DNA during differentiation does not bring the V_L and C_L genes into contiguity in the DNA. Instead, the DNA



Figure 36–12. Sister chromatid exchanges between human chromosomes. These are detectable by Giemsa staining of the chromosomes of cells replicated for two cycles in the presence of bromodeoxyuridine. The arrows indicate some regions of exchange. (Courtesy of S Wolff and J Bodycote.)

contains an interspersed or interruption sequence of about 1200 base pairs at or near the junction of the V and C regions. The interspersed sequence is transcribed into RNA along with the V_L and C_L genes, and the interspersed information is removed from the RNA during its nuclear processing (Chapters 37 and 39).

DNA SYNTHESIS & REPLICATION ARE RIGIDLY CONTROLLED

The primary function of DNA replication is understood to be the provision of progeny with the genetic information possessed by the parent. Thus, the replication of DNA must be complete and carried out in such a way as to maintain genetic stability within the organism and the species. The process of DNA replication is complex and involves many cellular functions and several verification procedures to ensure fidelity in replication. About 30 proteins are involved in the replication of the *E. coli* chromosome, and this process is almost certainly more complex in eukaryotic organisms. The first enzymologic observations on DNA replication were made in *E. coli* by Kornberg, who described in that organism the existence of an enzyme now called DNA polymerase I. This enzyme has multiple catalytic activities, a complex structure, and a requirement for the triphosphates of the four deoxyribonucleosides of adenine, guanine, cytosine, and thymine. The polymerization reaction catalyzed by DNA polymerase I of *E. coli* has served as a prototype for all DNA polymerases of both prokaryotes and eukaryotes, even though it is now recognized that the major role of this polymerase is to complete replication on the lagging strand.

In all cells, replication can occur only from a single-stranded DNA (ssDNA) template. Mechanisms must exist to target the site of initiation of replication and to unwind the double-stranded DNA (dsDNA) in that region. The replication complex must then form. After replication is complete in an area, the parent and daughter strands must re-form dsDNA. In eukaryotic cells, an additional step must occur. The dsDNA must precisely re-form the chromatin structure, including nucleosomes, that existed prior to the onset of replication. Although this entire process is not well understood in eukaryotic cells, replication has been quite precisely described in prokaryotic cells, and the general principles are thought to be the same in both. The major steps are listed in Table 36-4, illustrated in Figure 36-13, and discussed, in sequence, below. A number of proteins, most with specific enzymatic action, are involved in this process (Table 36-5).

The Origin of Replication

At the **origin of replication (ori)**, there is an association of sequence-specific dsDNA-binding proteins with

Table 36-4. Steps involved in DNA replication in eukaryotes.

1. Identification of the origins of replication.
2. Unwinding (denaturation) of dsDNA to provide an ssDNA template.
3. Formation of the replication fork.
4. Initiation of DNA synthesis and elongation.
5. Formation of replication bubbles with ligation of the newly synthesized DNA segments.
6. Reconstitution of chromatin structure.

a series of direct repeat DNA sequences. In bacteriophage λ , the $ori\lambda$ is bound by the λ -encoded O protein to four adjacent sites. In *E. coli*, the $oriC$ is bound by the protein *dnaA*. In both cases, a complex is formed consisting of 150–250 bp of DNA and multimers of the DNA-binding protein. This leads to the local denaturation and unwinding of an adjacent A+T-rich region of DNA. Functionally similar **autonomously replicating sequences (ARS)** have been identified in yeast cells. The ARS contains a somewhat degenerate 11-bp sequence called the **origin replication element (ORE)**. The ORE binds a set of proteins, analogous to the *dnaA* protein of *E. coli*, which is collectively called the **origin recognition complex (ORC)**. The ORE is located adjacent to an approximately 80-bp A+T-rich sequence that is easy to unwind. This is called the **DNA unwinding element (DUE)**. The DUE is the origin of replication in yeast.

Consensus sequences similar to *ori* or ARS in structure or function have not been precisely defined in mammalian cells, though several of the proteins that participate in *ori* recognition and function have been identified and appear quite similar to their yeast counterparts in both amino acid sequence and function.

Unwinding of DNA

The interaction of proteins with *ori* defines the start site of replication and provides a short region of ssDNA essential for initiation of synthesis of the nascent DNA strand. This process requires the formation of a number of protein-protein and protein-DNA interactions. A critical step is provided by a DNA helicase that allows for processive unwinding of DNA. In uninfected *E. coli*, this function is provided by a complex of *dnaB* helicase and the *dnaC* protein. Single-stranded DNA-binding proteins (SSBs) stabilize this complex. In λ phage-infected *E. coli*, the phage protein P binds to *dnaB* and the P/*dnaB* complex binds to $ori\lambda$ by interacting with the O protein. *dnaB* is not an active helicase when in the P/*dnaB*/O complex. Three *E. coli* heat shock proteins (*dnaK*, *dnaJ*, and *GrpE*) cooperate to remove the P

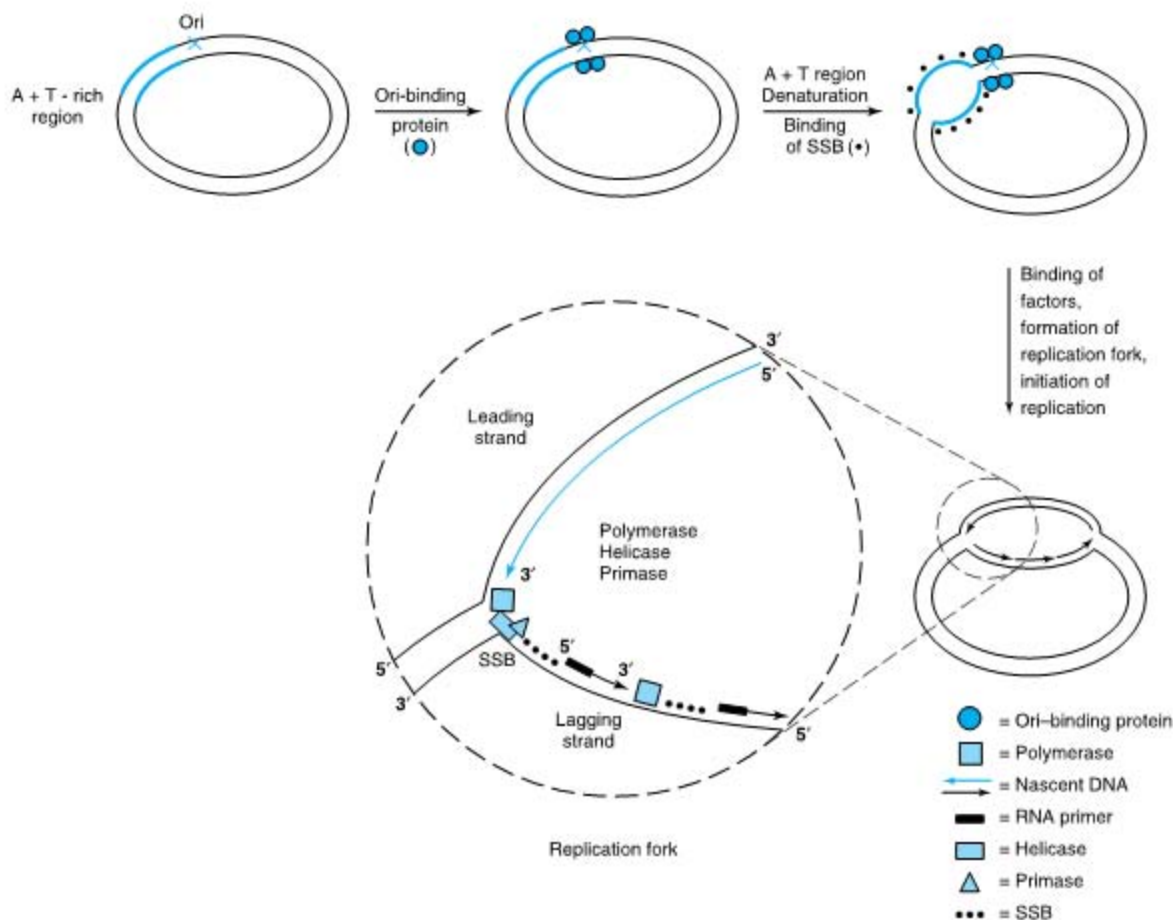


Figure 36-13. Steps involved in DNA replication. This figure describes DNA replication in an *E. coli* cell, but the general steps are similar in eukaryotes. A specific interaction of a protein (the O protein) to the origin of replication (ori) results in local unwinding of DNA at an adjacent A+T-rich region. The DNA in this area is maintained in the single-strand conformation (ssDNA) by single-strand-binding proteins (SSBs). This allows a variety of proteins, including helicase, primase, and DNA polymerase, to bind and to initiate DNA synthesis. The replication fork proceeds as DNA synthesis occurs continuously (long arrow) on the leading strand and discontinuously (short arrows) on the lagging strand. The nascent DNA is always synthesized in the 5' to 3' direction, as DNA polymerases can add a nucleotide only to the 3' end of a DNA strand.

protein and activate the *dnaB* helicase. In cooperation with SSB, this leads to DNA unwinding and active replication. In this way, the replication of the λ phage is accomplished at the expense of replication of the host *E. coli* cell.

Formation of the Replication Fork

A replication fork consists of four components that form in the following sequence: (1) the DNA helicase unwinds a short segment of the parental duplex DNA;

(2) a primase initiates synthesis of an RNA molecule that is essential for priming DNA synthesis; (3) the DNA polymerase initiates nascent, daughter strand synthesis; and (4) SSBs bind to ssDNA and prevent premature reannealing of ssDNA to dsDNA. These reactions are illustrated in Figure 36-13.

The polymerase III holoenzyme (the *dnaE* gene product in *E. coli*) binds to template DNA as part of a multiprotein complex that consists of several polymerase accessory factors (β , γ , δ , δ' , and ϵ). DNA polymerases only synthesize DNA in the 5' to 3' direction,

Table 36–5. Classes of proteins involved in replication.

Protein	Function
DNA polymerases	Deoxynucleotide polymerization
Helicases	Processive unwinding of DNA
Topoisomerases	Relieve torsional strain that results from helicase-induced unwinding
DNA primase	Initiates synthesis of RNA primers
Single-strand binding proteins	Prevent premature reannealing of dsDNA
DNA ligase	Seals the single strand nick between the nascent chain and Okazaki fragments on lagging strand

and only one of the several different types of polymerases is involved at the replication fork. Because the DNA strands are antiparallel (Chapter 35), the polymerase functions asymmetrically. On the **leading (forward) strand**, the DNA is synthesized continuously. On the **lagging (retrograde) strand**, the DNA is synthesized in short (1–5 kb; see Figure 36–16) fragments, the so-called **Okazaki fragments**. Several Okazaki fragments (up to 250) must be synthesized, in sequence, for each replication fork. To ensure that this happens, the helicase acts on the lagging strand to unwind dsDNA in a 5' to 3' direction. The helicase associates with the primase to afford the latter proper access to the template. This allows the RNA primer to be made and, in turn, the polymerase to begin replicating the DNA. This is an important reaction sequence since DNA polymerases cannot initiate DNA synthesis *de novo*. The mobile complex between helicase and primase has been called a **primosome**. As the synthesis of an Okazaki fragment is completed and the polymerase is released, a new primer has been synthesized. The same polymerase molecule remains associated with the replication fork and proceeds to synthesize the next Okazaki fragment.

The DNA Polymerase Complex

A number of different DNA polymerase molecules engage in DNA replication. These share three important properties: (1) **chain elongation**, (2) **processivity**, and (3) **proofreading**. Chain elongation accounts for the rate (in nucleotides per second) at which polymerization occurs. Processivity is an expression of the number of nucleotides added to the nascent chain before the polymerase disengages from the template. The proofreading function identifies copying errors and corrects them. In *E. coli*, polymerase III (pol III) functions at the

replication fork. Of all polymerases, it catalyzes the highest rate of chain elongation and is the most processive. It is capable of polymerizing 0.5 Mb of DNA during one cycle on the leading strand. Pol III is a large (> 1 MDa), ten-subunit protein complex in *E. coli*. The two identical β subunits of pol III encircle the DNA template in a sliding “clamp,” which accounts for the stability of the complex and for the high degree of processivity the enzyme exhibits.

Polymerase II (pol II) is mostly involved in proofreading and DNA repair. Polymerase I (pol I) completes chain synthesis between Okazaki fragments on the lagging strand. Eukaryotic cells have counterparts for each of these enzymes plus some additional ones. A comparison is shown in Table 36–6.

In mammalian cells, the polymerase is capable of polymerizing about 100 nucleotides per second, a rate at least tenfold slower than the rate of polymerization of deoxynucleotides by the bacterial DNA polymerase complex. This reduced rate may result from interference by nucleosomes. It is not known how the replication complex negotiates nucleosomes.

Initiation & Elongation of DNA Synthesis

The initiation of DNA synthesis (Figure 36–14) requires **priming by a short length of RNA**, about 10–200 nucleotides long. This priming process involves the nucleophilic attack by the 3'-hydroxyl group of the RNA primer on the α phosphate of the first entering deoxynucleoside triphosphate (N in Figure 36–14) with the splitting off of pyrophosphate. The 3'-hydroxyl group of the recently attached deoxyribonucleoside monophosphate is then free to carry out a **nucleophilic attack** on the next entering deoxyribonucleoside triphosphate (N + 1 in Figure 36–14), again at its α phosphate moiety, with the splitting off of pyrophosphate. Of course, selection of the proper deoxyribonucleotide whose terminal 3'-hydroxyl group is to be attacked is dependent upon **proper base pairing**

Table 36–6. A comparison of prokaryotic and eukaryotic DNA polymerases.

<i>E. coli</i>	Mammalian	Function
I	α	Gap filling and synthesis of lagging strand
II	ϵ	DNA proofreading and repair
	β	DNA repair
	γ	Mitochondrial DNA synthesis
III	δ	Processive, leading strand synthesis

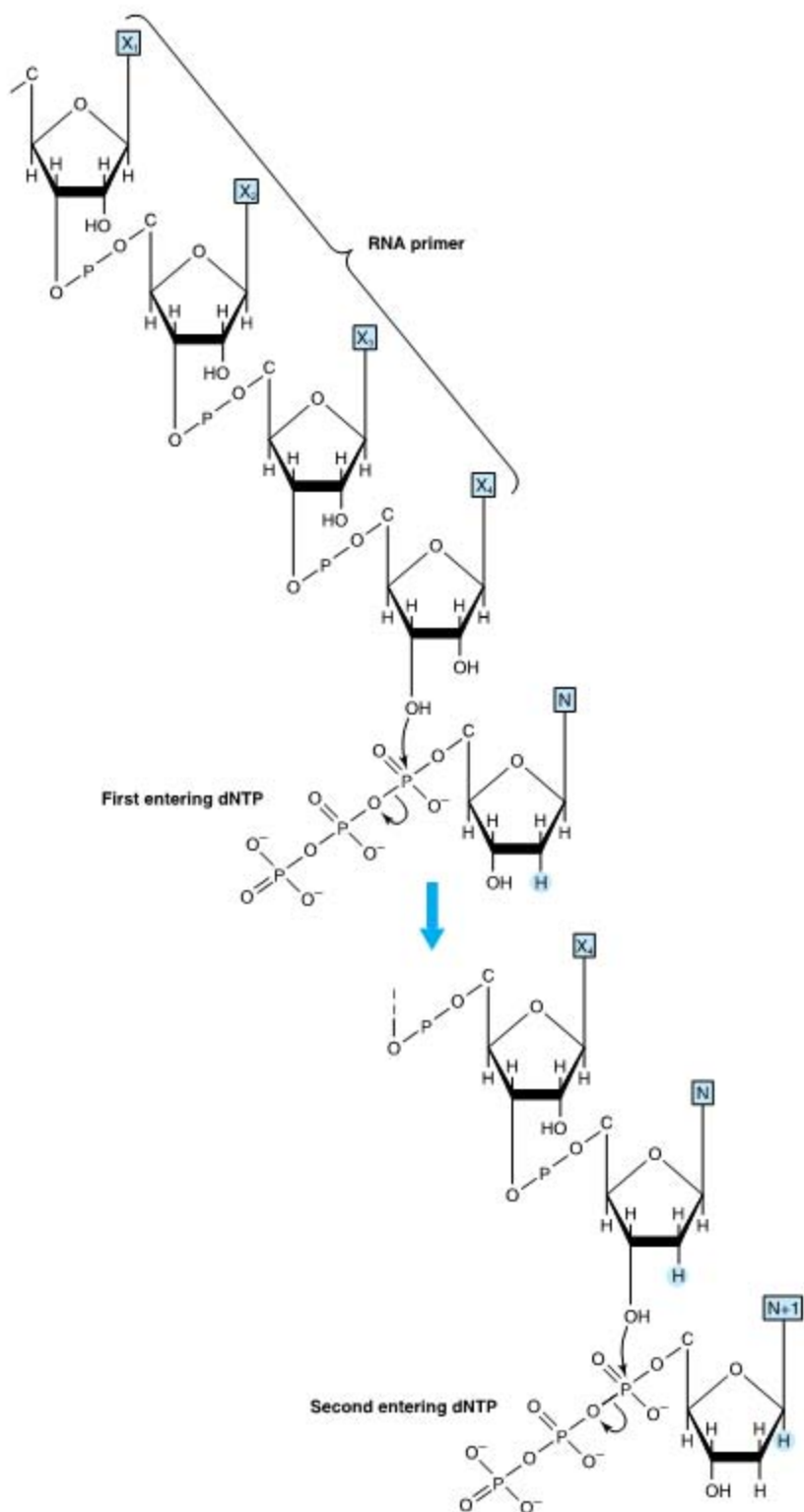


Figure 36–14. The initiation of DNA synthesis upon a primer of RNA and the subsequent attachment of the second deoxyribonucleoside triphosphate.

with the other strand of the DNA molecule according to the rules proposed originally by Watson and Crick (Figure 36-15). When an adenine deoxyribonucleoside monophosphoryl moiety is in the template position, a thymidine triphosphate will enter and its α phosphate will be attacked by the 3'-hydroxyl group of the deoxyribonucleoside monophosphoryl most recently added to the polymer. By this stepwise process, the template dictates which deoxyribonucleoside triphosphate is complementary and by hydrogen bonding holds it in place while the 3'-hydroxyl group of the growing strand attacks and incorporates the new nucleotide into the polymer. These segments of DNA attached to an RNA initiator component are the **Okazaki fragments** (Figure 36-16). In mammals, after many Okazaki fragments are generated, the replication complex begins to remove the RNA primers, to fill in the gaps left by their removal with the proper base-paired deoxynucleotide, and then to seal the fragments

of newly synthesized DNA by enzymes referred to as **DNA ligases**.

Replication Exhibits Polarity

As has already been noted, DNA molecules are double-stranded and the two strands are antiparallel, ie, running in opposite directions. The replication of DNA in prokaryotes and eukaryotes occurs on both strands simultaneously. However, an enzyme capable of polymerizing DNA in the 3' to 5' direction does not exist in any organism, so that both of the newly replicated DNA strands cannot grow in the same direction simultaneously. Nevertheless, the same enzyme does replicate both strands at the same time. The single enzyme replicates one strand ("leading strand") in a continuous manner in the 5' to 3' direction, with the same overall forward direction. It replicates the other strand ("lagging strand") discontinuously while polymerizing the

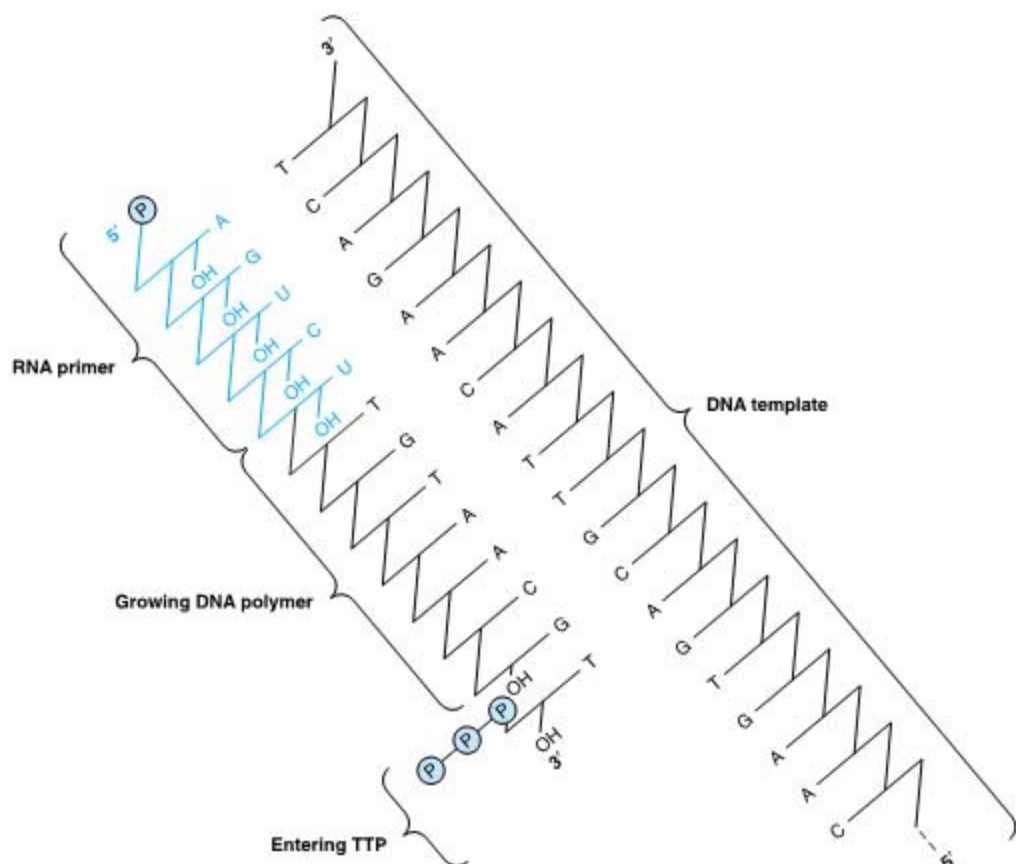


Figure 36-15. The RNA-primed synthesis of DNA demonstrating the template function of the complementary strand of parental DNA.

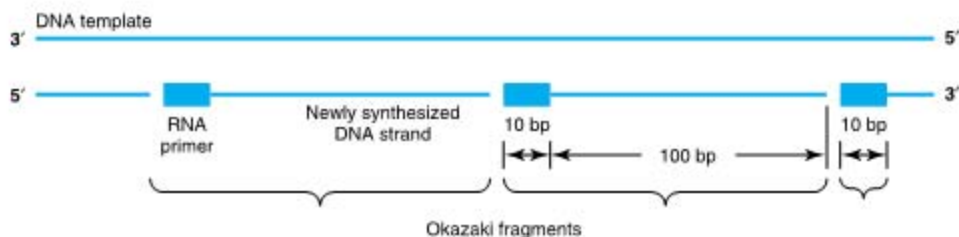


Figure 36–16. The discontinuous polymerization of deoxyribonucleotides on the lagging strand; formation of Okazaki fragments during lagging strand DNA synthesis is illustrated. Okazaki fragments are 100–250 nt long in eukaryotes, 1000–2000 bp in prokaryotes.

nucleotides in short spurts of 150–250 nucleotides, again in the 5′ to 3′ direction, but at the same time it faces toward the back end of the preceding RNA primer rather than toward the unreplicated portion. This process of **semidiscontinuous DNA synthesis** is shown diagrammatically in Figures 36–13 and 36–16.

In the mammalian nuclear genome, most of the RNA primers are eventually removed as part of the replication process, whereas after replication of the mitochondrial genome the small piece of RNA remains as an integral part of the closed circular DNA structure.

Formation of Replication Bubbles

Replication proceeds from a single *ori* in the circular bacterial chromosome, composed of roughly 6×10^6 bp of DNA. This process is completed in about 30 minutes, a replication rate of 3×10^5 bp/min. The entire mammalian genome replicates in approximately 9 hours, the average period required for formation of a tetraploid genome from a diploid genome in a replicating cell. If a mammalian genome (3×10^9 bp) replicated at the same rate as bacteria (ie, 3×10^5 bp/min) from but a single *ori*, replication would take over 150

hours! Metazoan organisms get around this problem using two strategies. First, replication is bidirectional. Second, replication proceeds from multiple origins in each chromosome (a total of as many as 100 in humans). Thus, replication occurs in both directions along all of the chromosomes, and both strands are replicated simultaneously. This replication process generates “**replication bubbles**” (Figure 36–17).

The multiple sites that serve as origins for DNA replication in eukaryotes are poorly defined except in a few animal viruses and in yeast. However, it is clear that initiation is regulated both spatially and temporally, since clusters of adjacent sites initiate replication synchronously. There are suggestions that functional domains of chromatin replicate as intact units, implying that the origins of replication are specifically located with respect to transcription units.

During the replication of DNA, there must be a separation of the two strands to allow each to serve as a template by hydrogen bonding its nucleotide bases to the incoming deoxynucleoside triphosphate. The separation of the DNA double helix is promoted by SSBs, specific protein molecules that stabilize the single-stranded structure as the replication fork progresses. These stabi-

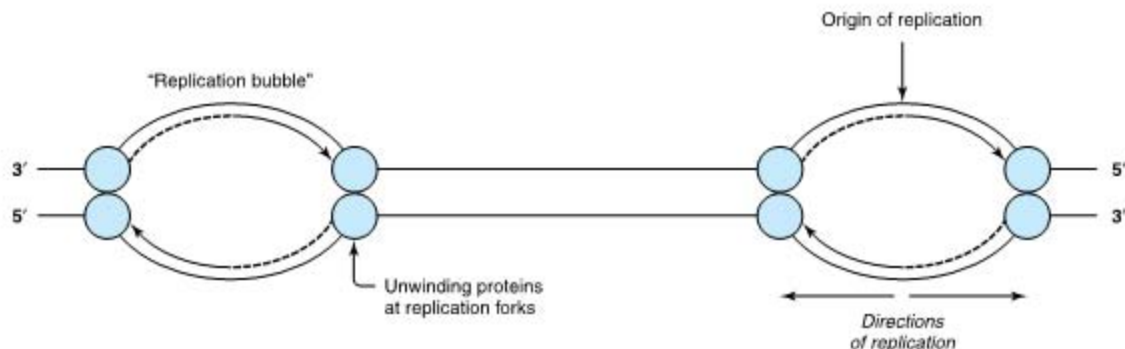


Figure 36–17. The generation of “replication bubbles” during the process of DNA synthesis. The bidirectional replication and the proposed positions of unwinding proteins at the replication forks are depicted.

lizing proteins bind cooperatively and stoichiometrically to the single strands without interfering with the abilities of the nucleotides to serve as templates (Figure 36–13). In addition to separating the two strands of the double helix, there must be an unwinding of the molecule (once every 10 nucleotide pairs) to allow strand separation. This must happen in segments, given the time during which DNA replication occurs. There are multiple “swivels” interspersed in the DNA molecules of all organisms. The swivel function is provided by specific enzymes that introduce “**nicks**” in one strand of the **unwinding double helix**, thereby allowing the unwinding process to proceed. The nicks are quickly resealed

without requiring energy input, because of the formation of a high-energy covalent bond between the nicked phosphodiester backbone and the nicking-sealing enzyme. The nicking-resealing enzymes are called **DNA topoisomerases**. This process is depicted diagrammatically in Figure 36–18 and there compared with the ATP-dependent resealing carried out by the DNA ligases. Topoisomerases are also capable of unwinding supercoiled DNA. Supercoiled DNA is a higher-ordered structure occurring in circular DNA molecules wrapped around a core, as depicted in Figure 36–19.

There exists in one species of animal viruses (retroviruses) a class of enzymes capable of synthesizing a sin-

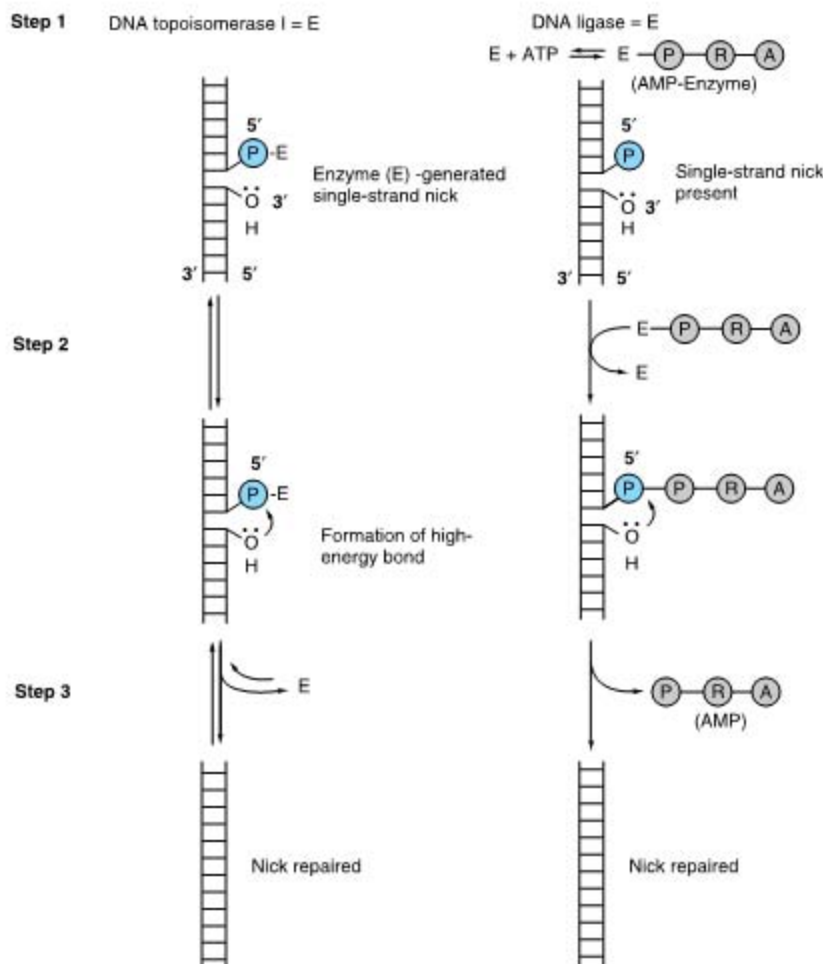


Figure 36–18. Comparison of two types of nick-sealing reactions on DNA. The series of reactions at left is catalyzed by DNA topoisomerase I, that at right by DNA ligase; P = phosphate, R = ribose, A = adenine. (Slightly modified and reproduced, with permission, from Lehninger AL: *Biochemistry*, 2nd ed. Worth, 1975.)

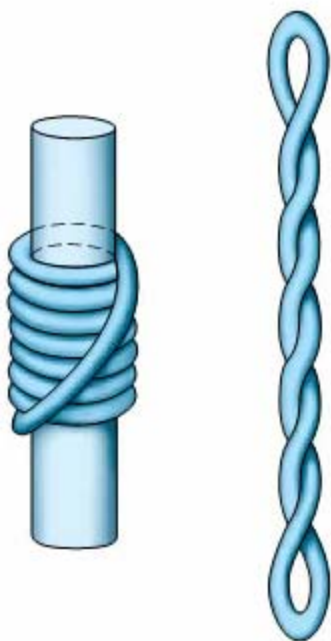


Figure 36-19. Supercoiling of DNA. A left-handed toroidal (solenoidal) supercoil, at left, will convert to a right-handed interwound supercoil, at right, when the cylindric core is removed. Such a transition is analogous to that which occurs when nucleosomes are disrupted by the high salt extraction of histones from chromatin.

gle-stranded and then a double-stranded DNA molecule from a single-stranded RNA template. This polymerase, RNA-dependent DNA polymerase, or “**reverse transcriptase**,” first synthesizes a DNA-RNA hybrid molecule utilizing the RNA genome as a template. A specific nuclease, RNase H, degrades the RNA strand, and the remaining DNA strand in turn serves as a template to form a double-stranded DNA molecule containing the information originally present in the RNA genome of the animal virus.

Reconstitution of Chromatin Structure

There is evidence that nuclear organization and chromatin structure are involved in determining the regulation and initiation of DNA synthesis. As noted above, the rate of polymerization in eukaryotic cells, which have chromatin and nucleosomes, is tenfold slower than that in prokaryotic cells, which have naked DNA. It is also clear that chromatin structure must be re-formed after replication. Newly replicated DNA is rapidly assembled into nucleosomes, and the

preexisting and newly assembled histone octamers are randomly distributed to each arm of the replication fork.

DNA Synthesis Occurs During the S Phase of the Cell Cycle

In animal cells, including human cells, the replication of the DNA genome occurs only at a specified time during the life span of the cell. This period is referred to as the synthetic or S phase. This is usually temporally separated from the mitotic phase by nonsynthetic periods referred to as gap 1 (G1) and gap 2 (G2), occurring before and after the S phase, respectively (Figure 36-20). Among other things, the cell prepares for DNA synthesis in G1 and for mitosis in G2. The cell regulates its DNA synthesis grossly by allowing it to occur only at specific times and mostly in cells preparing to divide by a mitotic process.

It appears that all eukaryotic cells have gene products that govern the transition from one phase of the cell cycle to another. The **cyclins** are a family of proteins whose concentration increases and decreases throughout the cell cycle—thus their name. The cyclins turn on, at the appropriate time, different **cyclin-dependent protein kinases (CDKs)** that phosphorylate substrates essential for progression through the cell cycle (Figure 36-21). For example, cyclin D levels rise in late G1 phase and allow progression beyond the **start (yeast)** or **restriction point (mammals)**, the point beyond which cells irrevocably proceed into the S or DNA synthesis phase.

The D cyclins activate CDK4 and CDK6. These two kinases are also synthesized during G1 in cells undergoing active division. The D cyclins and CDK4 and CDK6 are nuclear proteins that assemble as a complex in late G1 phase. The complex is an active serine-threonine protein kinase. One substrate for this kinase is the retinoblastoma (Rb) protein. Rb is a cell cycle regulator because it binds to and inactivates a transcription factor (E2F) necessary for the transcription of certain genes (histone genes, DNA replication proteins, etc) needed for progression from G1 to S phase. The phosphorylation of Rb by CDK4 or CDK6 results in the release of E2F from Rb-mediated transcription repression—thus, gene activation ensues and cell cycle progression takes place.

Other cyclins and CDKs are involved in different aspects of cell cycle progression (Table 36-7). Cyclin E and CDK2 form a complex in late G1. Cyclin E is rapidly degraded, and the released CDK2 then forms a complex with cyclin A. This sequence is necessary for the initiation of DNA synthesis in S phase. A complex between cyclin B and CDK1 is rate-limiting for the G2/M transition in eukaryotic cells.

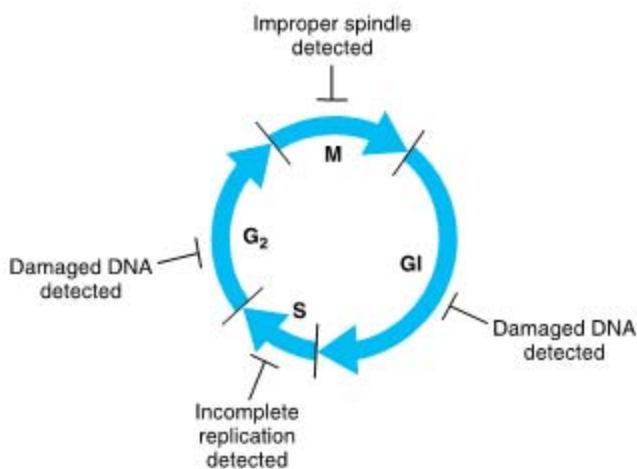


Figure 36–20. Mammalian cell cycle and cell cycle checkpoints. DNA, chromosome, and chromosome segregation integrity is continuously monitored throughout the cell cycle. If DNA damage is detected in either the G1 or the G2 phase of the cell cycle, if the genome is incompletely replicated, or if normal chromosome segregation machinery is incomplete (ie, a defective spindle), cells will not progress through the phase of the cycle in which defects are detected. In some cases, if the damage cannot be repaired, such cells undergo programmed cell death (apoptosis).

Many of the cancer-causing viruses (oncoviruses) and cancer-inducing genes (oncogenes) are capable of alleviating or disrupting the apparent restriction that normally controls the entry of mammalian cells from G1 into the S phase. From the foregoing, one might have surmised that excessive production of a cyclin—or production at an inappropriate time—might result in abnormal or unrestrained cell division. In this context it is noteworthy that the *bcl* oncogene associated with B cell lymphoma appears to be the cyclin D1 gene. Similarly, the oncoproteins (or transforming proteins) pro-

duced by several DNA viruses target the Rb transcription repressor for inactivation, inducing cell division inappropriately.

During the S phase, mammalian cells contain greater quantities of DNA polymerase than during the nonsynthetic phases of the cell cycle. Furthermore, those enzymes responsible for formation of the substrates for DNA synthesis—ie, deoxyribonucleoside triphosphates—are also increased in activity, and their activity will diminish following the synthetic phase until the reappearance of the signal for renewed DNA

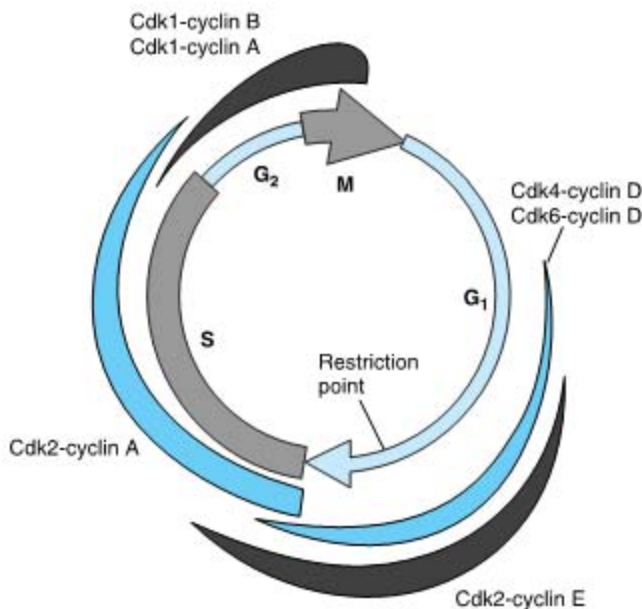


Figure 36–21. Schematic illustration of the points during the mammalian cell cycle during which the indicated cyclins and cyclin-dependent kinases are activated. The thickness of the various colored lines is indicative of the extent of activity.

Table 36–7. Cyclins and cyclin-dependent kinases involved in cell cycle progression.

Cyclin	Kinase	Function
D	CDK4, CDK6	Progression past restriction point at G1/S boundary
E, A	CDK2	Initiation of DNA synthesis in early S phase
B	CDK1	Transition from G2 to M

synthesis. During the S phase, the nuclear DNA is **completely replicated once and only once**. It seems that once chromatin has been replicated, it is marked so as to prevent its further replication until it again passes through mitosis. The molecular mechanisms for this phenomenon have yet to be elucidated.

In general, a given pair of chromosomes will replicate simultaneously and within a fixed portion of the S phase upon every replication. On a chromosome, clusters of replication units replicate coordinately. The nature of the signals that regulate DNA synthesis at these levels is unknown, but the regulation does appear to be an intrinsic property of each individual chromosome.

Enzymes Repair Damaged DNA

The maintenance of the integrity of the information in DNA molecules is of utmost importance to the survival of a particular organism as well as to survival of the species. Thus, it can be concluded that surviving species have evolved mechanisms for repairing DNA damage occurring as a result of either replication errors or environmental insults.

As described in Chapter 35, the major responsibility for the fidelity of replication resides in the specific pairing of nucleotide bases. Proper pairing is dependent upon the presence of the favored tautomers of the purine and pyrimidine nucleotides, but the equilibrium whereby one tautomer is more stable than another is only about 10^4 or 10^5 in favor of that with the greater stability. Although this is not favorable enough to ensure the high fidelity that is necessary, favoring of the preferred tautomers—and thus of the proper base pairing—could be ensured by monitoring the base pairing twice. Such double monitoring does appear to occur in both bacterial and mammalian systems: once at the time of insertion of the deoxyribonucleoside triphosphates, and later by a follow-up energy-requiring mechanism that removes all improper bases which may occur in the newly formed strand. This “proofreading” prevents tautomer-induced misincorporation from occur-

ring more frequently than once every 10^8 – 10^{10} base pairs of DNA synthesized. The mechanisms responsible for this monitoring mechanism in *E. coli* include the 3' to 5' exonuclease activities of one of the subunits of the pol III complex and of the pol I molecule. The analogous mammalian enzymes (δ and α) do not seem to possess such a nuclease proofreading function. Other enzymes provide this repair function.

Replication errors, even with a very efficient repair system, lead to the accumulation of mutations. A human has 10^{14} nucleated cells each with 3×10^9 base pairs of DNA. If about 10^{16} cell divisions occur in a lifetime and 10^{-10} mutations per base pair per cell generation escape repair, there may eventually be as many as one mutation per 10^6 bp in the genome. Fortunately, most of these will probably occur in DNA that does not encode proteins or will not affect the function of encoded proteins and so are of no consequence. In addition, spontaneous and chemically induced damage to DNA must be repaired.

Damage to DNA by environmental, physical, and chemical agents may be classified into four types (Table 36–8). Abnormal regions of DNA, either from copying errors or DNA damage, are repaired by four mechanisms: (1) mismatch repair, (2) base excision-repair, (3) nucleotide excision-repair, and (4) double-strand break repair (Table 36–9). These mechanisms exploit the redundancy of information inherent in the double helical DNA structure. The defective region in one strand can be returned to its original form by relying on the complementary information stored in the unaffected strand.

Table 36–8. Types of damage to DNA.

- | |
|--|
| I. Single-base alteration |
| A. Depurination |
| B. Deamination of cytosine to uracil |
| C. Deamination of adenine to hypoxanthine |
| D. Alkylation of base |
| E. Insertion or deletion of nucleotide |
| F. Base-analog incorporation |
| II. Two-base alteration |
| A. UV light-induced thymine-thymine (pyrimidine) dimer |
| B. Bifunctional alkylating agent cross-linkage |
| III. Chain breaks |
| A. Ionizing radiation |
| B. Radioactive disintegration of backbone element |
| C. Oxidative free radical formation |
| IV. Cross-linkage |
| A. Between bases in same or opposite strands |
| B. Between DNA and protein molecules (eg, histones) |

Table 36–9. Mechanism of DNA repair

Mechanism	Problem	Solution
Mismatch repair	Copying errors (single base or two- to five-base unpaired loops)	Methyl-directed strand cutting, exonuclease digestion, and replacement
Base excision-repair	Spontaneous, chemical, or radiation damage to a single base	Base removal by <i>N</i> -glycosylase, abasic sugar removal, replacement
Nucleotide excision-repair	Spontaneous, chemical, or radiation damage to a DNA segment	Removal of an approximately 30-nucleotide oligomer and replacement
Double-strand break repair	Ionizing radiation, chemotherapy, oxidative free radicals	Synapsis, unwinding, alignment, ligation

Mismatch Repair

Mismatch repair corrects errors made when DNA is copied. For example, a C could be inserted opposite an A, or the polymerase could slip or stutter and insert two to five extra unpaired bases. Specific proteins scan the newly synthesized DNA, using adenine methylation within a GATC sequence as the point of reference (Figure 36–22). The template strand is methylated, and the newly synthesized strand is not. This difference allows the repair enzymes to identify the strand that contains the errant nucleotide which requires replacement. If a mismatch or small loop is found, a GATC endonuclease cuts the strand bearing the mutation at a site corresponding to the GATC. An exonuclease then digests this strand from the GATC through the mutation, thus removing the faulty DNA. This can occur from either end if the defect is bracketed by two GATC sites. This defect is then filled in by normal cellular enzymes according to base pairing rules. In *E. coli*, three proteins (Mut S, Mut C, and Mut H) are required for recognition of the mutation and nicking of the strand. Other cellular enzymes, including ligase, polymerase, and SSBs, remove and replace the strand. The process is somewhat more complicated in mammalian cells, as about six proteins are involved in the first steps.

Faulty mismatch repair has been linked to hereditary nonpolyposis colon cancer (HNPCC), one of the most common inherited cancers. Genetic studies linked HNPCC in some families to a region of chromosome 2. The gene located, designated *hMSH2*, was subsequently shown to encode the human analog of the

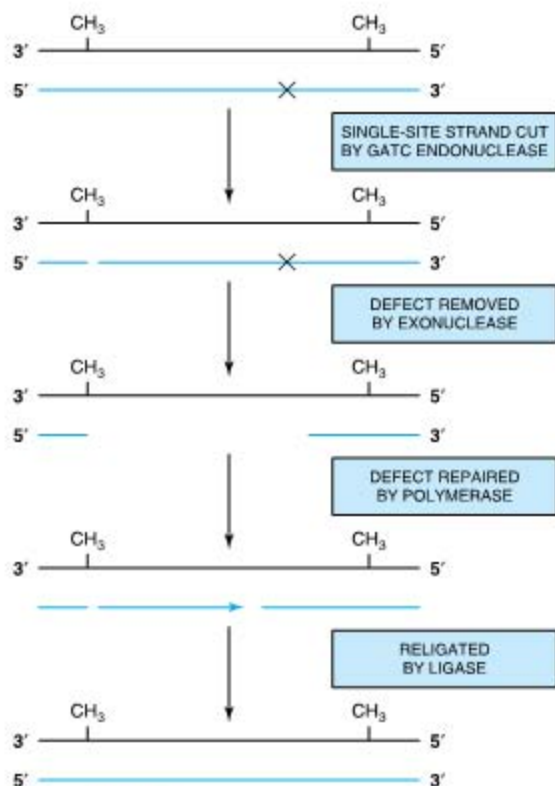


Figure 36–22. Mismatch repair of DNA. This mechanism corrects a single mismatch base pair (eg, C to A rather than T to A) or a short region of unpaired DNA. The defective region is recognized by an endonuclease that makes a single-strand cut at an adjacent methylated GATC sequence. The DNA strand is removed through the mutation, replaced, and religated.

E. coli MutS protein that is involved in mismatch repair (see above). Mutations of *hMSH2* account for 50–60% of HNPCC cases. Another gene, *hMLH1*, is associated with most of the other cases. *hMLH1* is the human analog of the bacterial mismatch repair gene *MutL*. How does faulty mismatch repair result in colon cancer? The human genes were localized because microsatellite instability was detected. That is, the cancer cells had a microsatellite of a length different from that found in the normal cells of the individual. It appears that the affected cells, which harbor a mutated *hMSH2* or *hMLH1* mismatch repair enzyme, are unable to remove small loops of unpaired DNA, and the microsatellite thus increases in size. Ultimately, microsatellite DNA expansion must affect either the expression or the function of a protein critical in surveillance of the cell cycle in these colon cells.

Base Excision-Repair

The **depurination of DNA**, which happens spontaneously owing to the thermal lability of the purine N-glycosidic bond, occurs at a rate of 5000–10,000/cell/d at 37 °C. Specific enzymes recognize a depurinated site and replace the appropriate purine directly, without interruption of the phosphodiester backbone.

Cytosine, adenine, and guanine bases in DNA spontaneously form uracil, hypoxanthine, or xanthine, respectively. Since none of these normally exist in DNA, it is not surprising that specific **N-glycosylases** can recognize these abnormal bases and remove the base itself from the DNA. This removal marks the site of the defect and allows an **apurinic or apyrimidinic endonuclease** to excise the abasic sugar. The proper base is then replaced by a repair DNA polymerase, and a **ligase** returns the DNA to its original state (Figure 36–23). This series of events is called **base excision-repair**. By a similar series of steps involving initially the recognition of the defect, alkylated bases and base analogs can be removed from DNA and the DNA returned to its original informational content. This mechanism is suitable for replacement of a single base but is not effective at replacing regions of damaged DNA.

Nucleotide Excision-Repair

This mechanism is used to replace regions of damaged DNA up to 30 bases in length. Common examples of DNA damage include ultraviolet (UV) light, which induces the formation of cyclobutane pyrimidine-pyrimidine dimers, and smoking, which causes formation of benzo[a]pyrene-guanine adducts. Ionizing radiation, cancer chemotherapeutic agents, and a variety of chemicals found in the environment cause base modification, strand breaks, cross-linkage between bases on opposite strands or between DNA and protein, and numerous other defects. These are repaired by a process called nucleotide excision-repair (Figure 36–24). This complex process, which involves more gene products than the two other types of repair, essentially involves the hydrolysis of two phosphodiester bonds on the strand containing the defect. A special excision nuclease (exonuclease), consisting of at least three subunits in *E. coli* and 16 polypeptides in humans, accomplishes this task. In eukaryotic cells the enzymes cut between the third to fifth phosphodiester bond 3' from the lesion, and on the 5' side the cut is somewhere between the twenty-first and twenty-fifth bonds. Thus, a fragment of DNA 27–29 nucleotides long is excised. After the strand is removed it is replaced, again by exact base pairing, through the action of yet another polymerase (δ/ϵ in humans), and the ends are joined to the existing strands by DNA ligase.

Xeroderma pigmentosum (XP) is an autosomal recessive genetic disease. The clinical syndrome includes

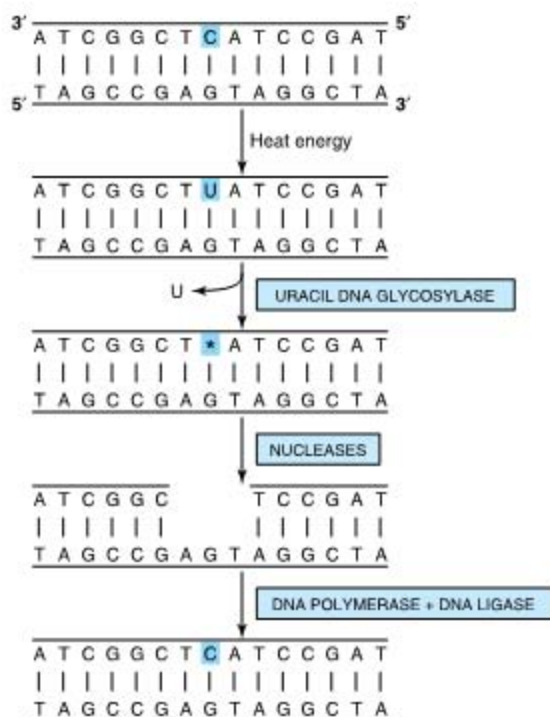


Figure 36–23. Base excision-repair of DNA. The enzyme uracil DNA glycosylase removes the uracil created by spontaneous deamination of cytosine in the DNA. An endonuclease cuts the backbone near the defect; then, after an endonuclease removes a few bases, the defect is filled in by the action of a repair polymerase and the strand is rejoined by a ligase. (Courtesy of B. Alberts.)

marked sensitivity to sunlight (ultraviolet) with subsequent formation of multiple skin cancers and premature death. The risk of developing skin cancer is increased 1000- to 2000-fold. The inherited defect seems to involve the repair of damaged DNA, particularly thymine dimers. Cells cultured from patients with xeroderma pigmentosum exhibit low activity for the nucleotide excision-repair process. Seven complementation groups have been identified using hybrid cell analyses, so at least seven gene products (XPA–XPG) are involved. Two of these (XPA and XPC) are involved in recognition and excision. XPB and XPD are helicases and, interestingly, are subunits of the transcription factor TFIIH (see Chapter 37).

Double-Strand Break Repair

The repair of double-strand breaks is part of the physiologic process of immunoglobulin gene rearrangement. It

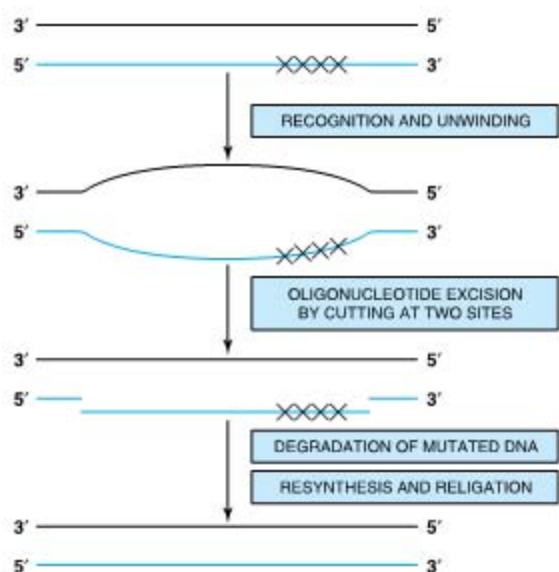


Figure 36–24. Nucleotide excision-repair. This mechanism is employed to correct larger defects in DNA and generally involves more proteins than either mismatch or base excision-repair. After defect recognition (indicated by XXXX) and unwinding of the DNA encompassing the defect, an excision nuclease (exinuclease) cuts the DNA upstream and downstream of the defective region. This gap is then filled in by a polymerase (δ/ϵ in humans) and religated.

is also an important mechanism for repairing damaged DNA, such as occurs as a result of ionizing radiation or oxidative free radical generation. Some chemotherapeutic agents destroy cells by causing ds breaks or preventing their repair.

Two proteins are initially involved in the nonhomologous rejoining of a ds break. **Ku**, a heterodimer of 70 kDa and 86 kDa subunits, binds to free DNA ends and has latent ATP-dependent helicase activity. The DNA-bound Ku heterodimer recruits a unique protein kinase, **DNA-dependent protein kinase (DNA-PK)**. DNA-PK has a binding site for DNA free ends and another for dsDNA just inside these ends. It therefore allows for the approximation of the two separated ends. The free end DNA-Ku-DNA-PK complex activates the kinase activity in the latter. DNA-PK reciprocally phosphorylates Ku and the other DNA-PK molecule, on the opposing strand, in trans. DNA-PK then dissociates from the DNA and Ku, resulting in activation of the Ku helicase. This results in unwinding of the two ends. The unwound, approximated DNA forms base pairs; the extra nucleotide tails are removed by an exonuclease;

and the gaps are filled and closed by DNA ligase. This repair mechanism is illustrated in Figure 36–25.

Some Repair Enzymes Are Multifunctional

Somewhat surprising is the recent observation that DNA repair proteins can serve other purposes. For example, some repair enzymes are also found as components of the large TFIIF complex that plays a central role in gene transcription (Chapter 37). Another component of TFIIF is involved in cell cycle regulation. Thus, three critical cellular processes may be linked through use of common proteins. There is also good evidence that some repair enzymes are involved in gene rearrangements that occur normally.

In patients with **ataxia-telangiectasia**, an autosomal recessive disease in humans resulting in the development of cerebellar ataxia and lymphoreticular neoplasms, there appears to exist an increased sensitivity to damage by x-ray. Patients with **Fanconi's anemia**, an autosomal recessive anemia characterized also by an increased frequency of cancer and by chromosomal instability, probably have defective repair of cross-linking damage.

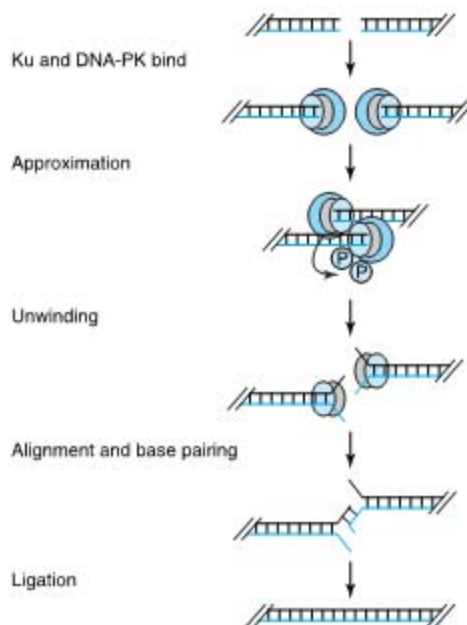


Figure 36–25. Double-strand break repair of DNA. The proteins Ku and DNA-dependent protein kinase combine to approximate the two strands and unwind them. The aligned fragments form base pairs; the extra ends are removed, probably by a DNA-PK-associated endo- or exonuclease, and the gaps are filled in; and continuity is restored by ligation.

All three of these clinical syndromes are associated with an increased frequency of cancer. It is likely that other human diseases resulting from disordered DNA repair capabilities will be found in the future.

DNA & Chromosome Integrity Is Monitored Throughout the Cell Cycle

Given the importance of normal DNA and chromosome function to survival, it is not surprising that eukaryotic cells have developed elaborate mechanisms to monitor the integrity of the genetic material. As detailed above, a number of complex multi-subunit enzyme systems have evolved to repair damaged DNA at the nucleotide sequence level. Similarly, DNA mishaps at the chromosome level are also monitored and repaired. As shown in Figure 36–20, DNA integrity and chromosomal integrity are continuously monitored throughout the cell cycle. The four specific steps at which this monitoring occurs have been termed **checkpoint controls**. If problems are detected at any of these checkpoints, progression through the cycle is interrupted and transit through the cell cycle is halted until the damage is repaired. The molecular mechanisms underlying detection of DNA damage during the G1 and G2 phases of the cycle are understood better than those operative during S and M phases.

The **tumor suppressor p53**, a protein of MW 53 kDa, plays a key role in both G1 and G2 checkpoint control. Normally a very unstable protein, p53 is a DNA binding transcription factor, one of a family of related proteins, that is somehow stabilized in response to DNA damage, perhaps by direct p53-DNA interactions. Increased levels of p53 activate transcription of an ensemble of genes that collectively serve to delay transit through the cycle. One of these induced proteins, p21^{CIP}, is a potent CDK-cyclin inhibitor (CKI) that is capable of efficiently inhibiting the action of all CDKs. Clearly, inhibition of CDKs will halt progression through the cell cycle (see Figures 36–19 and 36–20). If DNA damage is too extensive to repair, the affected cells undergo **apoptosis** (programmed cell death) in a p53-dependent fashion. In this case, p53 induces the activation of a collection of genes that induce apoptosis. Cells lacking functional p53 fail to undergo apoptosis in response to high levels of radiation or DNA-active chemotherapeutic agents. It may come as no surprise, then, that p53 is one of the most frequently mutated genes in human cancers. Additional research into the mechanisms of checkpoint control will prove invaluable for the development of effective anticancer therapeutic options.

SUMMARY

- DNA in eukaryotic cells is associated with a variety of proteins, resulting in a structure called chromatin.

- Much of the DNA is associated with histone proteins to form a structure called the nucleosome. Nucleosomes are composed of an octamer of histones and 150 bp of DNA.
- Nucleosomes and higher-order structures formed from them serve to compact the DNA.
- As much as 90% of DNA may be transcriptionally inactive as a result of being nuclease-resistant, highly compacted, and nucleosome-associated.
- DNA in transcriptionally active regions is sensitive to nuclease attack; some regions are exceptionally sensitive and are often found to contain transcription control sites.
- Transcriptionally active DNA (the genes) is often clustered in regions of each chromosome. Within these regions, genes may be separated by inactive DNA in nucleosomal structures. The transcription unit—that portion of a gene that is copied by RNA polymerase—consists of coding regions of DNA (exons) interrupted by intervening sequences of non-coding DNA (introns).
- After transcription, during RNA processing, introns are removed and the exons are ligated together to form the mature mRNA that appears in the cytoplasm.
- DNA in each chromosome is exactly replicated according to the rules of base pairing during the S phase of the cell cycle.
- Each strand of the double helix is replicated simultaneously but by somewhat different mechanisms. A complex of proteins, including DNA polymerase, replicates the leading strand continuously in the 5' to 3' direction. The lagging strand is replicated discontinuously, in short pieces of 150–250 nucleotides, in the 3' to 5' direction.
- DNA replication occurs at several sites—called replication bubbles—in each chromosome. The entire process takes about 9 hours in a typical cell.
- A variety of mechanisms employing different enzymes repair damaged DNA, as after exposure to chemical mutagens or ultraviolet radiation.

REFERENCES

- DePamphilis ML: Origins of DNA replication in metazoan chromosomes. *J Biol Chem* 1993;268:1.
- Hartwell LH, Kastan MB: Cell cycle control and cancer. *Science* 1994;266:1821.
- Jenuwein T, Allis CD: Translating the histone code. *Science* 2001; 293:1074.
- Lander ES et al: Initial sequencing and analysis of the human genome. *Nature* 2001;409:860.
- Luger L et al: Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 1997;398:251.

- Marians KJ: Prokaryotic DNA replication. *Annu Rev Biochem* 1992;61:673.
- Michelson RJ, Weinart T: Closing the gaps among a web of DNA repair disorders. *Bioessays J* 2002;22:966.
- Moll UM, Erster S, Zaika A: p53, p63 and p73—solos, alliances and feuds among family members. *Biochim Biophys Acta* 2001;1552:47.
- Mouse Genome Sequencing Consortium: Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;420:520.
- Narlikar GJ et al: Cooperation between complexes that regulate chromatin structure and transcription. *Cell* 2002;108:475.
- Sullivan et al: Determining centromere identity: cyclical stories and forking paths. *Nat Rev Genet* 2001;2:584.
- van Holde K, Zlatanova J: Chromatin higher order: chasing a mirage? *J Biol Chem* 1995;270:8373.
- Venter JC et al: The sequence of the human genome. *Science* 2002;291:1304.
- Wallace DC: Mitochondrial DNA in aging and disease. *Sci Am* 1997 Aug;277:40.
- Wood RD: Nucleotide excision repair in mammalian cells. *J Biol Chem* 1997;272:23465.

RNA Synthesis, Processing, & Modification

37

Daryl K. Granner, MD, & P. Anthony Weil, PhD

BIOMEDICAL IMPORTANCE

The synthesis of an RNA molecule from DNA is a complex process involving one of the group of RNA polymerase enzymes and a number of associated proteins. The general steps required to synthesize the primary transcript are initiation, elongation, and termination. Most is known about initiation. A number of DNA regions (generally located upstream from the initiation site) and protein factors that bind to these sequences to regulate the initiation of transcription have been identified. Certain RNAs—mRNAs in particular—have very different life spans in a cell. It is important to understand the basic principles of messenger RNA synthesis and metabolism, for modulation of this process results in altered rates of protein synthesis and thus a variety of metabolic changes. This is how all organisms adapt to changes of environment. It is also how differentiated cell structures and functions are established and maintained. The RNA molecules synthesized in mammalian cells are made as precursor molecules that have to be processed into mature, active RNA. Errors or changes in synthesis, processing, and splicing of mRNA transcripts are a cause of disease.

RNA EXISTS IN FOUR MAJOR CLASSES

All eukaryotic cells have four major classes of RNA: ribosomal RNA (rRNA), messenger RNA (mRNA), transfer RNA (tRNA), and small nuclear RNA (snRNA). The first three are involved in protein synthesis, and snRNA is involved in mRNA splicing. As shown in Table 37-1, these various classes of RNA are different in their diversity, stability, and abundance in cells.

RNA IS SYNTHESIZED FROM A DNA TEMPLATE BY AN RNA POLYMERASE

The processes of DNA and RNA synthesis are similar in that they involve (1) the general steps of initiation, elongation, and termination with 5' to 3' polarity; (2) large, multicomponent initiation complexes; and (3) adherence to Watson-Crick base-pairing rules. These processes differ in several important ways, including the

following: (1) ribonucleotides are used in RNA synthesis rather than deoxyribonucleotides; (2) U replaces T as the complementary base pair for A in RNA; (3) a primer is not involved in RNA synthesis; (4) only a very small portion of the genome is transcribed or copied into RNA, whereas the entire genome must be copied during DNA replication; and (5) there is no proofreading function during RNA transcription.

The process of synthesizing RNA from a DNA template has been characterized best in prokaryotes. Although in mammalian cells the regulation of RNA synthesis and the processing of the RNA transcripts are different from those in prokaryotes, the process of RNA synthesis per se is quite similar in these two classes of organisms. Therefore, the description of RNA synthesis in prokaryotes, where it is better understood, is applicable to eukaryotes even though the enzymes involved and the regulatory signals are different.

The Template Strand of DNA Is Transcribed

The sequence of ribonucleotides in an RNA molecule is complementary to the sequence of deoxyribonucleotides in one strand of the double-stranded DNA molecule (Figure 35-8). The strand that is transcribed or copied into an RNA molecule is referred to as the **template strand** of the DNA. The other DNA strand is frequently referred to as the **coding strand** of that gene. It is called this because, with the exception of T for U changes, it corresponds exactly to the sequence of the primary transcript, which encodes the protein product of the gene. In the case of a double-stranded DNA molecule containing many genes, the template strand for each gene will not necessarily be the same strand of the DNA double helix (Figure 37-1). Thus, a given strand of a double-stranded DNA molecule will serve as the template strand for some genes and the coding strand of other genes. Note that the nucleotide sequence of an RNA transcript will be the same (except for U replacing T) as that of the coding strand. The information in the template strand is read out in the 3' to 5' direction.

Table 37-1. Classes of eukaryotic RNA.

RNA	Types	Abundance	Stability
Ribosomal (rRNA)	28S, 18S, 5.8S, 5S	80% of total	Very stable
Messenger (mRNA)	~10 ⁵ different species	2–5% of total	Unstable to very stable
Transfer (tRNA)	~60 different species	~15% of total	Very stable
Small nuclear (snRNA)	~30 different species	≤ 1% of total	Very stable

DNA-Dependent RNA Polymerase Initiates Transcription at a Distinct Site, the Promoter

DNA-dependent RNA polymerase is the enzyme responsible for the polymerization of ribonucleotides into a sequence complementary to the template strand of the gene (see Figures 37-2 and 37-3). The enzyme attaches at a specific site—the promoter—on the template strand. This is followed by initiation of RNA synthesis at the starting point, and the process continues until a termination sequence is reached (Figure 37-3). A **transcription unit** is defined as that region of DNA that includes the signals for transcription initiation, elongation, and termination. The RNA product, which is synthesized in the 5' to 3' direction, is the **primary transcript**. In prokaryotes, this can represent the product of several contiguous genes; in mammalian cells, it usually represents the product of a single gene. The 5' terminals of the primary RNA transcript and the mature cytoplasmic RNA are identical. Thus, the starting point of transcription corresponds to the 5' nucleotide of the mRNA. This is designated position +1, as is the corresponding nucleotide in the DNA. The

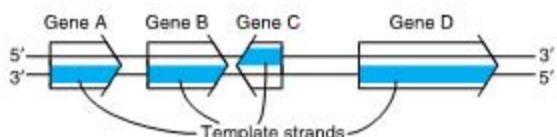


Figure 37-1. This figure illustrates that genes can be transcribed off both strands of DNA. The arrowheads indicate the direction of transcription (polarity). Note that the template strand is always read in the 3' to 5' direction. The opposite strand is called the coding strand because it is identical (except for T for U changes) to the mRNA transcript (the primary transcript in eukaryotic cells) that encodes the protein product of the gene.

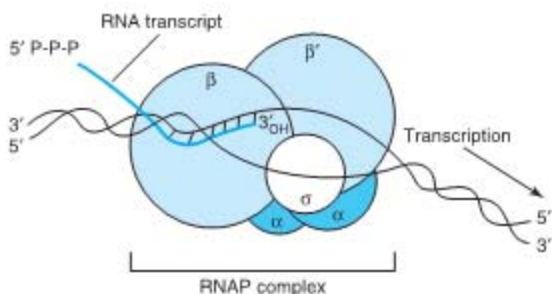


Figure 37-2. RNA polymerase (RNAP) catalyzes the polymerization of ribonucleotides into an RNA sequence that is complementary to the template strand of the gene. The RNA transcript has the same polarity (5' to 3') as the coding strand but contains U rather than T. *E. coli* RNAP consists of a core complex of two α subunits and two β subunits (β and β'). The holoenzyme contains the σ subunit bound to the $\alpha_2\beta\beta'$ core assembly. The ω subunit is not shown. The transcription “bubble” is an approximately 20-bp area of melted DNA, and the entire complex covers 30–75 bp, depending on the conformation of RNAP.

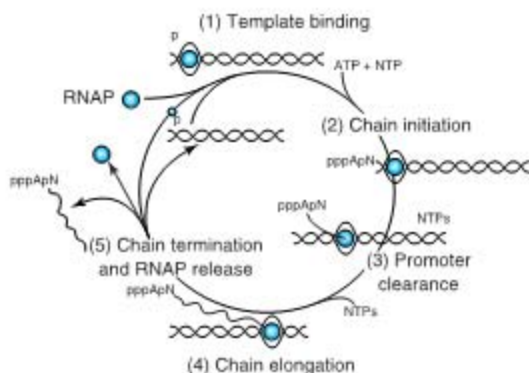


Figure 37-3. The transcription cycle in bacteria. Bacterial RNA transcription is described in four steps: **(1) Template binding:** RNA polymerase (RNAP) binds to DNA and locates a promoter (P) melts the two DNA strands to form a preinitiation complex (PIC). **(2) Chain initiation:** RNAP holoenzyme (core + one of multiple sigma factors) catalyzes the coupling of the first base (usually ATP or GTP) to a second ribonucleoside triphosphate to form a dinucleotide. **(3) Chain elongation:** Successive residues are added to the 3'-OH terminus of the nascent RNA molecule. **(4) Chain termination and release:** The completed RNA chain and RNAP are released from the template. The RNAP holoenzyme re-forms, finds a promoter, and the cycle is repeated.

numbers increase as the sequence proceeds *downstream*. This convention makes it easy to locate particular regions, such as intron and exon boundaries. The nucleotide in the promoter adjacent to the transcription initiation site is designated -1, and these negative numbers increase as the sequence proceeds *upstream*, away from the initiation site. This provides a conventional way of defining the location of regulatory elements in the promoter.

The primary transcripts generated by RNA polymerase II—one of three distinct nuclear DNA-dependent RNA polymerases in eukaryotes—are promptly capped by 7-methylguanosine triphosphate caps (Figure 35-10) that persist and eventually appear on the 5' end of mature cytoplasmic mRNA. These caps are necessary for the subsequent processing of the primary transcript to mRNA, for the translation of the mRNA, and for protection of the mRNA against exonucleolytic attack.

Bacterial DNA-Dependent RNA Polymerase Is a Multisubunit Enzyme

The DNA-dependent RNA polymerase (RNAP) of the bacterium *Escherichia coli* exists as an approximately 400 kDa core complex consisting of two identical α subunits, similar but not identical β and β' subunits, and an ω subunit. Beta is thought to be the catalytic subunit (Figure 37-2). RNAP, a metalloenzyme, also contains two zinc molecules. The core RNA polymerase associates with a specific protein factor (the sigma [σ] factor) that helps the core enzyme recognize and bind to the specific deoxynucleotide sequence of the promoter region (Figure 37-5) to form the preinitiation complex (PIC). Sigma factors have a dual role in the process of promoter recognition; σ association with core RNA polymerase decreases its affinity for nonpromoter DNA while simultaneously increasing holoenzyme affinity for promoter DNA. Bacteria contain multiple σ factors, each of which acts as a regulatory protein that modifies the **promoter recognition specificity** of the RNA polymerase. The appearance of different σ factors can be correlated temporally with various programs of gene expression in prokaryotic systems such as bacteriophage development, sporulation, and the response to heat shock.

Mammalian Cells Possess Three Distinct Nuclear DNA-Dependent RNA Polymerases

The properties of mammalian polymerases are described in Table 37-2. Each of these DNA-dependent RNA polymerases is responsible for transcription of dif-

Table 37-2. Nomenclature and properties of mammalian nuclear DNA-dependent RNA polymerases.

Form of RNA Polymerase	Sensitivity to α -Amanitin	Major Products
I (A)	Insensitive	rRNA
II (B)	High sensitivity	mRNA
III (C)	Intermediate sensitivity	tRNA/5S rRNA

ferent sets of genes. The sizes of the RNA polymerases range from MW 500,000 to MW 600,000. These enzymes are much more complex than prokaryotic RNA polymerases. They all have two large subunits and a number of smaller subunits—as many as 14 in the case of RNA pol III. The eukaryotic RNA polymerases have extensive amino acid homologies with prokaryotic RNA polymerases. This homology has been shown recently to extend to the level of three-dimensional structures. The functions of each of the subunits are not yet fully understood. Many could have regulatory functions, such as serving to assist the polymerase in the recognition of specific sequences like promoters and termination signals.

One peptide toxin from the mushroom *Amanita phalloides*, α -amanitin, is a specific differential inhibitor of the eukaryotic nuclear DNA-dependent RNA polymerases and as such has proved to be a powerful research tool (Table 37-2). α -Amanitin blocks the translocation of RNA polymerase during transcription.

RNA SYNTHESIS IS A CYCLICAL PROCESS & INVOLVES INITIATION, ELONGATION, & TERMINATION

The process of RNA synthesis in bacteria—depicted in Figure 37-3—involves first the binding of the RNA holopolymerase molecule to the template at the promoter site to form a PIC. Binding is followed by a conformational change of the RNAP, and the first nucleotide (almost always a purine) then associates with the initiation site on the β subunit of the enzyme. In the presence of the appropriate nucleotide, the RNAP catalyzes the formation of a phosphodiester bond, and the nascent chain is now attached to the polymerization site on the β subunit of RNAP. (The analogy to the A and P sites on the ribosome should be noted; see Figure 38-9.)

Initiation of formation of the RNA molecule at its 5' end then follows, while elongation of the RNA mole-

cule from the 5' to its 3' end continues cyclically, antiparallel to its template. The enzyme polymerizes the ribonucleotides in a specific sequence dictated by the template strand and interpreted by Watson-Crick base-pairing rules. Pyrophosphate is released in the polymerization reaction. This pyrophosphate (PP_i) is rapidly degraded to 2 mol of inorganic phosphate (P_i) by ubiquitous pyrophosphatases, thereby providing irreversibility on the overall synthetic reaction. In both prokaryotes and eukaryotes, a purine ribonucleotide is usually the first to be polymerized into the RNA molecule. As with eukaryotes, 5' triphosphate of this first nucleotide is maintained in prokaryotic mRNA.

As the **elongation** complex containing the core RNA polymerase progresses along the DNA molecule, **DNA unwinding** must occur in order to provide access for the appropriate base pairing to the nucleotides of the coding strand. The extent of this transcription bubble (ie, DNA unwinding) is constant throughout transcription and has been estimated to be about 20 base pairs per polymerase molecule. Thus, it appears that the size of the unwound DNA region is dictated by the polymerase and is independent of the DNA sequence in the complex. This suggests that RNA polymerase has associated with it an "unwindase" activity that opens the DNA helix. The fact that the DNA double helix must unwind and the strands part at least transiently for transcription implies some disruption of the nucleosome structure of eukaryotic cells. Topoisomerase both precedes and follows the progressing RNAP to prevent the formation of superhelical complexes.

Termination of the synthesis of the RNA molecule in bacteria is signaled by a sequence in the template strand of the DNA molecule—a signal that is recognized by a termination protein, the rho (ρ) factor. Rho is an ATP-dependent RNA-stimulated helicase that disrupts the nascent RNA-DNA complex. After termination of synthesis of the RNA molecule, the enzyme separates from the DNA template and probably dissociates to free core enzyme and free σ factor. With the assistance of another σ factor, the core enzyme then recognizes a promoter at which the synthesis of a new RNA molecule commences. In eukaryotic cells, termination is less well defined. It appears to be somehow linked both to initiation and to addition of the 3' polyA tail of mRNA and could involve destabilization of the RNA-DNA complex at a region of A-U base pairs. More than one RNA polymerase molecule may transcribe the same template strand of a gene simultaneously, but the process is phased and spaced in such a way that at any one moment each is transcribing a different portion of the DNA sequence. An electron micrograph of extremely active RNA synthesis is shown in Figure 37-4.



Figure 37-4. Electron photomicrograph of multiple copies of amphibian ribosomal RNA genes in the process of being transcribed. The magnification is about 6000 \times . Note that the length of the transcripts increases as the RNA polymerase molecules progress along the individual ribosomal RNA genes; transcription start sites (filled circles) to transcription termination sites (open circles). RNA polymerase I (not visualized here) is at the base of the nascent rRNA transcripts. Thus, the proximal end of the transcribed gene has short transcripts attached to it, while much longer transcripts are attached to the distal end of the gene. The arrows indicate the direction (5' to 3') of transcription. (Reproduced with permission, from Miller OL Jr, Beatty BR: Portrait of a gene. *J Cell Physiol* 1969;74[Suppl 1]:225.)

THE FIDELITY & FREQUENCY OF TRANSCRIPTION IS CONTROLLED BY PROTEINS BOUND TO CERTAIN DNA SEQUENCES

The DNA sequence analysis of specific genes has allowed the recognition of a number of sequences important in gene transcription. From the large number of bacterial genes studied it is possible to construct consensus models of transcription initiation and termination signals.

The question, “How does RNAP find the correct site to initiate transcription?” is not trivial when the complexity of the genome is considered. *E. coli* has 4×10^3 transcription initiation sites in 4×10^6 base pairs (bp) of DNA. The situation is even more complex in humans, where perhaps 10^5 transcription initiation sites are distributed throughout in 3×10^9 bp of DNA. RNAP can bind to many regions of DNA, but it scans the DNA sequence—at a rate of $\geq 10^3$ bp/s—until it recognizes certain specific regions of DNA to which it binds with higher affinity. This region is called the promoter, and it is the association of RNAP with the promoter that ensures accurate initiation of transcription. The promoter recognition-utilization process is the target for regulation in both bacteria and humans.

Bacterial Promoters Are Relatively Simple

Bacterial promoters are approximately 40 nucleotides (40 bp or four turns of the DNA double helix) in length, a region small enough to be covered by an *E. coli* RNA holopolymerase molecule. In this consensus promoter region are two short, conserved sequence elements. Approximately 35 bp upstream of the transcrip-

tion start site there is a consensus sequence of eight nucleotide pairs (5'-TGTTGACA-3') to which the RNAP binds to form the so-called **closed complex**. More proximal to the transcription start site—about ten nucleotides upstream—is a six-nucleotide-pair A+T-rich sequence (5'-TATAAT-3'). These conserved sequence elements comprising the promoter are shown schematically in Figure 37-5. The latter sequence has a low melting temperature because of its deficiency of GC nucleotide pairs. Thus, the **TATA box** is thought to ease the dissociation between the two DNA strands so that RNA polymerase bound to the promoter region can have access to the nucleotide sequence of its immediately downstream template strand. Once this process occurs, the combination of RNA polymerase plus promoter is called the **open complex**. Other bacteria have slightly different consensus sequences in their promoters, but all generally have two components to the promoter; these tend to be in the same position relative to the transcription start site, and in all cases the sequences between the boxes have no similarity but still provide critical spacing functions facilitating recognition of -35 and -10 sequence by RNA polymerase holoenzyme. Within a bacterial cell, different sets of genes are often

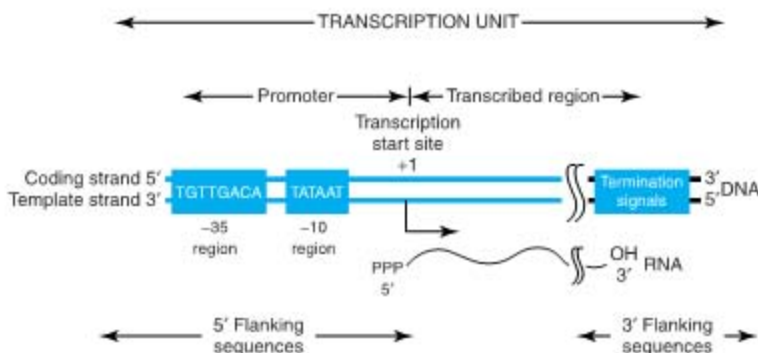


Figure 37-5. Bacterial promoters, such as that from *E. coli* shown here, share two regions of highly conserved nucleotide sequence. These regions are located 35 and 10 bp upstream (in the 5' direction of the coding strand) from the start site of transcription, which is indicated as +1. By convention, all nucleotides upstream of the transcription initiation site (at +1) are numbered in a negative sense and are referred to as 5'-flanking sequences. Also by convention, the DNA regulatory sequence elements (TATA box, etc) are described in the 5' to 3' direction and as being on the coding strand. These elements function only in double-stranded DNA, however. Note that the transcript produced from this transcription unit has the same polarity or “sense” (ie, 5' to 3' orientation) as the coding strand. Termination *cis*-elements reside at the end of the transcription unit (see Figure 37-6 for more detail). By convention the sequences downstream of the site at which transcription termination occurs are termed 3'-flanking sequences.

coordinately regulated. One important way that this is accomplished is through the fact that these co-regulated genes share unique -35 and -10 promoter sequences. These unique promoters are recognized by different σ factors bound to core RNA polymerase.

Rho-dependent transcription **termination signals** in *E. coli* also appear to have a distinct consensus sequence, as shown in Figure 37-6. The conserved consensus sequence, which is about 40 nucleotide pairs in length, can be seen to contain a hyphenated or interrupted inverted repeat followed by a series of AT base pairs. As transcription proceeds through the hyphenated, inverted repeat, the generated transcript can form the intramolecular hairpin structure, also depicted in Figure 37-6.

Transcription continues into the AT region, and with the aid of the ρ termination protein the RNA polymerase stops, dissociates from the DNA template, and releases the nascent transcript.

Eukaryotic Promoters Are More Complex

It is clear that the signals in DNA which control transcription in eukaryotic cells are of several types. Two types of sequence elements are promoter-proximal. One of these defines **where transcription is to commence** along the DNA, and the other contributes to the mechanisms that control **how frequently** this event is to occur. For example, in the thymidine kinase gene of the herpes

simplex virus, which utilizes transcription factors of its mammalian host for gene expression, there is a single unique transcription start site, and accurate transcription from this start site depends upon a nucleotide sequence located 32 nucleotides upstream from the start site (ie, at -32) (Figure 37-7). This region has the sequence of **TATAAAAG** and bears remarkable similarity to the functionally related TATA box that is located about 10 bp upstream from the prokaryotic mRNA start site (Figure 37-5). Mutation or inactivation of the TATA box markedly reduces transcription of this and many other genes that contain this consensus *cis* element (see Figures 37-7, 37-8). Most mammalian genes have a TATA box that is usually located 25–30 bp upstream from the transcription start site. The consensus sequence for a TATA box is TATAAA, though numerous variations have been characterized. The TATA box is bound by 34 kDa **TATA binding protein (TBP)**, which in turn binds several other proteins called **TBP-associated factors (TAFs)**. This complex of TBP and TAFs is referred to as **TFIID**. Binding of TFIID to the TATA box sequence is thought to represent the first step in the formation of the transcription complex on the promoter.

A small number of genes lack a TATA box. In such instances, two additional *cis* elements, an **initiator sequence (Inr)** and the so-called **downstream promoter element (DPE)**, direct RNA polymerase II to the promoter and in so doing provide basal transcription starting from the correct site. The Inr element spans the start

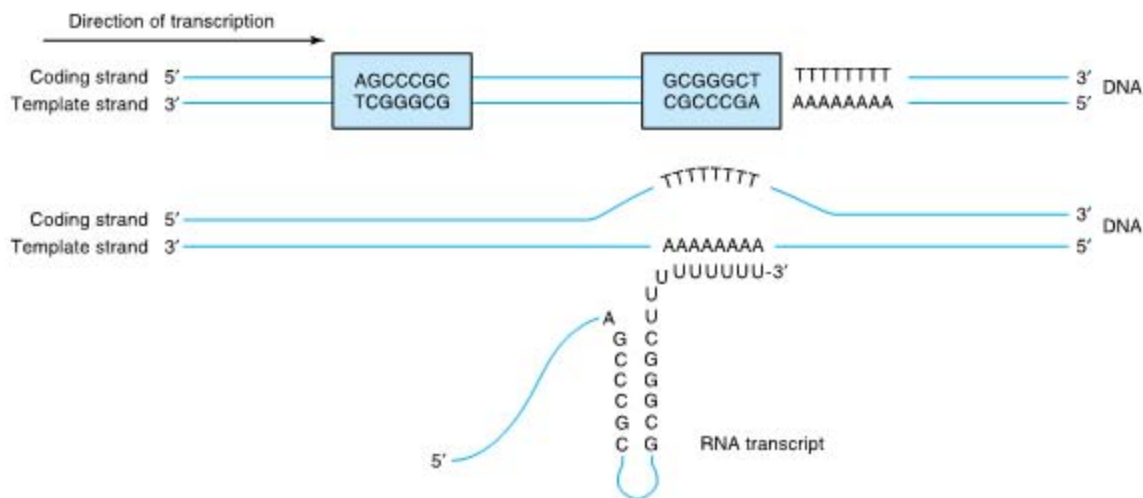


Figure 37-6. The predominant bacterial transcription termination signal contains an inverted, hyphenated repeat (the two boxed areas) followed by a stretch of AT base pairs (top figure). The inverted repeat, when transcribed into RNA, can generate the secondary structure in the RNA transcript shown at the bottom of the figure. Formation of this RNA hairpin causes RNA polymerase to pause and subsequently the ρ termination factor interacts with the paused polymerase and somehow induces chain termination.

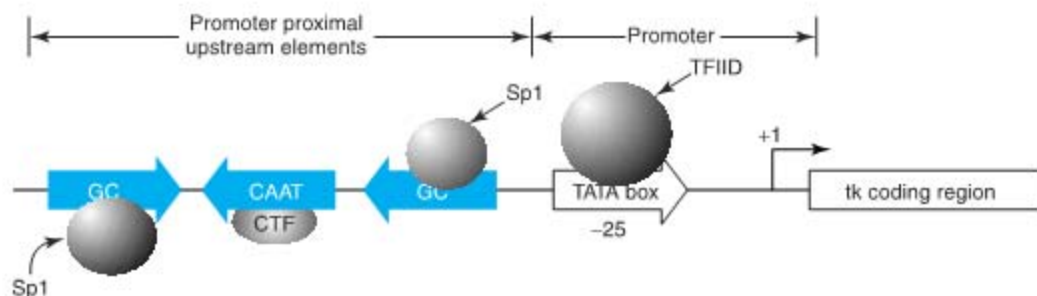


Figure 37-7. Transcription elements and binding factors in the herpes simplex virus thymidine kinase (*tk*) gene. DNA-dependent RNA polymerase II binds to the region of the TATA box (which is bound by transcription factor TFIID) to form a multicomponent preinitiation complex capable of initiating transcription at a single nucleotide (+1). The frequency of this event is increased by the presence of upstream *cis*-acting elements (the GC and CAAT boxes). These elements bind *trans*-acting transcription factors, in this example Sp1 and CTF (also called C/EBP, NF1, NFY). These *cis* elements can function independently of orientation (arrows).

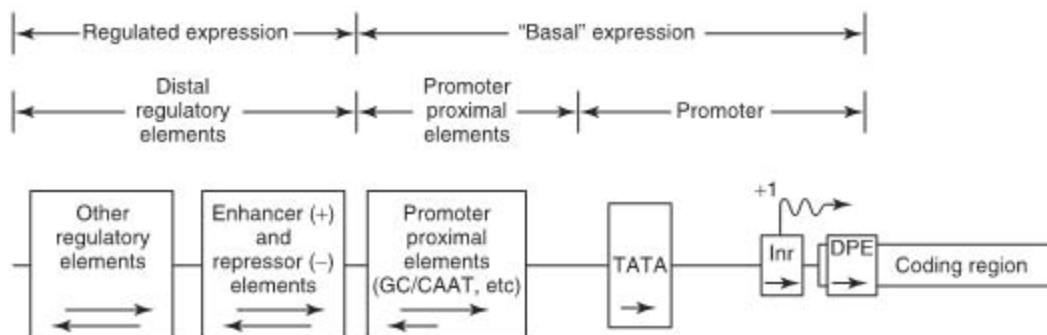


Figure 37-8. Schematic diagram showing the transcription control regions in a hypothetical class II (mRNA-producing) eukaryotic gene. Such a gene can be divided into its coding and regulatory regions, as defined by the transcription start site (arrow; +1). The coding region contains the DNA sequence that is transcribed into mRNA, which is ultimately translated into protein. The regulatory region consists of two classes of elements. One class is responsible for ensuring basal expression. These elements generally have two components. The proximal component, generally the TATA box, or Inr or DPE elements direct RNA polymerase II to the correct site (fidelity). In TATA-less promoters, an initiator (Inr) element that spans the initiation site (+1) may direct the polymerase to this site. Another component, the upstream elements, specifies the frequency of initiation. Among the best studied of these is the CAAT box, but several other elements (Sp1, NF1, AP1, etc) may be used in various genes. A second class of regulatory *cis*-acting elements is responsible for regulated expression. This class consists of elements that enhance or repress expression and of others that mediate the response to various signals, including hormones, heat shock, heavy metals, and chemicals. Tissue-specific expression also involves specific sequences of this sort. The orientation dependence of all the elements is indicated by the arrows within the boxes. For example, the proximal element (the TATA box) must be in the 5' to 3' orientation. The upstream elements work best in the 5' to 3' orientation, but some of them can be reversed. The locations of some elements are not fixed with respect to the transcription start site. Indeed, some elements responsible for regulated expression can be located either interspersed with the upstream elements, or they can be located downstream from the start site.

site (from -3 to +5) and consists of the general consensus sequence TCA₋₁ G/T T T/C which is similar to the initiation site sequence per se. (A+1 indicates the first nucleotide transcribed.) The proteins that bind to Inr in order to direct pol II binding include TFIID. Promoters that have both a TATA box and an Inr may be stronger than those that have just one of these elements. The DPE has the consensus sequence A/GGA/T CGTG and is localized about 25 bp downstream of the +1 start site. Like the Inr, DPE sequences are also bound by the TAF subunits of TFIID. In a survey of over 200 eukaryotic genes, roughly 30% contained a TATA box and Inr, 25% contained Inr and DPE, 15% contained all three elements, while ~30% contained just the Inr.

Sequences farther upstream from the start site determine how frequently the transcription event occurs. Mutations in these regions reduce the frequency of transcriptional starts tenfold to twentyfold. Typical of these DNA elements are the GC and CAAT boxes, so named because of the DNA sequences involved. As illustrated in Figure 37-7, each of these boxes binds a protein, Sp1 in the case of the GC box and CTF (or C/EPB, NF1, NFY) by the CAAT box; both bind through their distinct DNA binding domains (DBDs). The frequency of transcription initiation is a consequence of these protein-DNA interactions and complex interactions between particular domains of the transcription factors (distinct from the DBD domains—so-called activation domains; ADs) of these proteins and the rest of the transcription machinery (RNA polymerase II and the basal factors TFIIA, B, D, E, F). (See

below and Figures 37-9 and 37-10). The protein-DNA interaction at the TATA box involving RNA polymerase II and other components of the basal transcription machinery ensures the fidelity of initiation.

Together, then, the promoter and promoter-proximal *cis*-active upstream elements confer fidelity and frequency of initiation upon a gene. The TATA box has a particularly rigid requirement for both position and orientation. Single-base changes in any of these *cis* elements have dramatic effects on function by reducing the binding affinity of the cognate *trans* factors (either TFIID/TBP or Sp1, CTF, and similar factors). The spacing of these elements with respect to the transcription start site can also be critical. This is particularly true for the TATA box Inr and DPE.

A third class of sequence elements can either increase or decrease the rate of transcription initiation of eukaryotic genes. These elements are called either **enhancers** or **repressors (or silencers)**, depending on which effect they have. They have been found in a variety of locations both upstream and downstream of the transcription start site and even within the transcribed portions of some genes. In contrast to proximal and upstream promoter elements, enhancers and silencers can exert their effects when located hundreds or even thousands of bases away from transcription units located on the same chromosome. Surprisingly, enhancers and silencers can function in an orientation-independent fashion. Literally hundreds of these elements have been described. In some cases, the sequence requirements for binding are rigidly constrained; in others, considerable sequence variation is

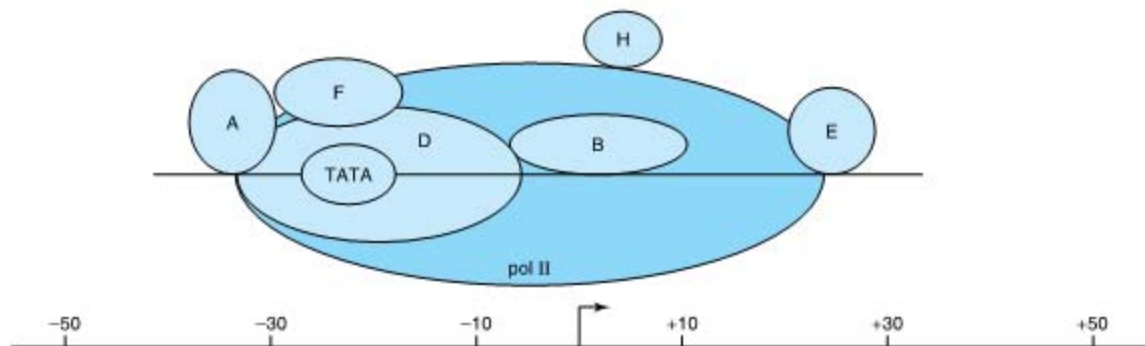


Figure 37-9. The eukaryotic basal transcription complex. Formation of the basal transcription complex begins when TFIID binds to the TATA box. It directs the assembly of several other components by protein-DNA and protein-protein interactions. The entire complex spans DNA from position -30 to +30 relative to the initiation site (+1, marked by bent arrow). The atomic level, x-ray-derived structures of RNA polymerase II alone and of TBP bound to TATA promoter DNA in the presence of either TFIIB or TFIIA have all been solved at 3 Å resolution. The structure of TFIID complexes have been determined by electron microscopy at 30 Å resolution. Thus, the molecular structures of the transcription machinery are beginning to be elucidated. Much of this structural information is consistent with the models presented here.

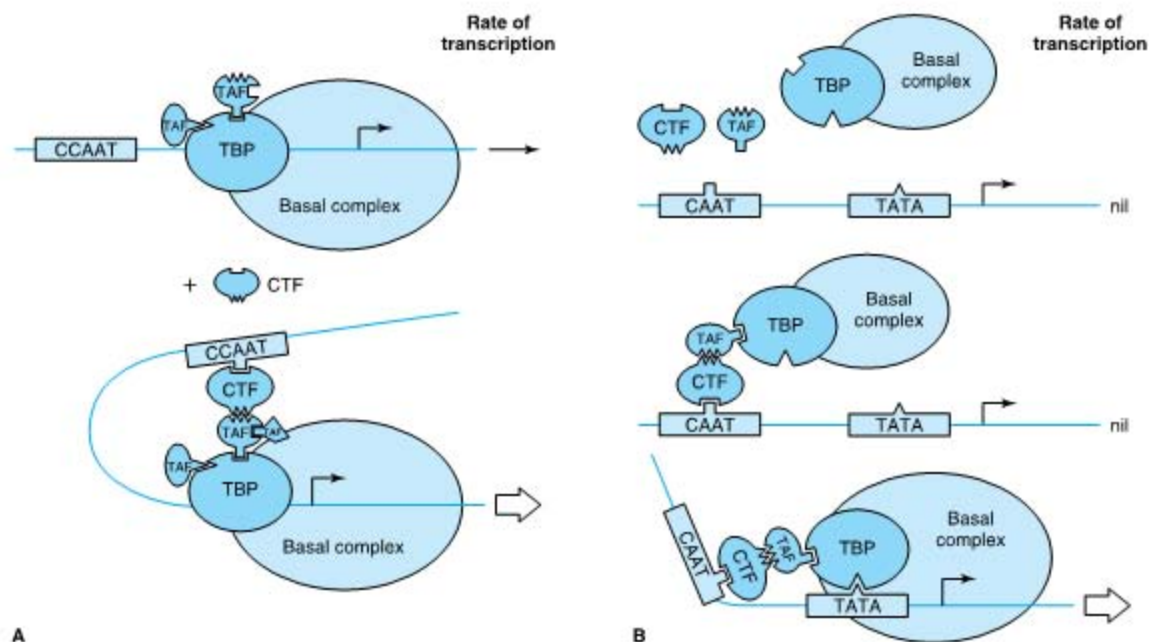


Figure 37-10. Two models for assembly of the active transcription complex and for how activators and coactivators might enhance transcription. Shown here as a small oval is TBP, which contains TFIID, a large oval that contains all the components of the basal transcription complex illustrated in Figure 37-9 (ie, RNAP II and TFIIA, TFIIB, TFIIE, TFIIIF, and TFIIF). **Panel A:** The basal transcription complex is assembled on the promoter after the TBP subunit of TFIID is bound to the TATA box. Several TAFs (coactivators) are associated with TBP. In this example, a transcription activator, CTF, is shown bound to the CAAT box, forming a loop complex by interacting with a TAF bound to TBP. **Panel B:** The recruitment model. The transcription activator CTF binds to the CAAT box and interacts with a coactivator (TAF in this case). This allows for an interaction with the preformed TBP-basal transcription complex. TBP can now bind to the TATA box, and the assembled complex is fully active.

allowed. Some sequences bind only a single protein, but the majority bind several different proteins. Similarly, a single protein can bind to more than one element.

Hormone response elements (for steroids, T_3 , retinoic acid, peptides, etc) act as—or in conjunction with—enhancers or silencers (Chapter 43). Other processes that enhance or silence gene expression—such as the response to heat shock, heavy metals (Cd^{2+} and Zn^{2+}), and some toxic chemicals (eg, dioxin)—are mediated through specific regulatory elements. Tissue-specific expression of genes (eg, the albumin gene in liver, the hemoglobin gene in reticulocytes) is also mediated by specific DNA sequences.

Specific Signals Regulate Transcription Termination

The signals for the termination of transcription by eukaryotic RNA polymerase II are very poorly under-

stood. However, it appears that the termination signals exist far downstream of the coding sequence of eukaryotic genes. For example, the transcription termination signal for mouse β -globin occurs at several positions 1000–2000 bases beyond the site at which the poly(A) tail will eventually be added. Little is known about the termination process or whether specific termination factors similar to the bacterial ρ factor are involved. However, it is known that the mRNA 3' terminal is generated posttranscriptionally, is somehow coupled to events or structures formed at the time and site of initiation, depends on a special structure in one of the subunits of RNA polymerase II (the CTD; see below), and appears to involve at least two steps. After RNA polymerase II has traversed the region of the transcription unit encoding the 3' end of the transcript, an RNA endonuclease cleaves the primary transcript at a position about 15 bases 3' of the consensus sequence AAUAAA that serves in eukaryotic transcripts as a cleavage signal.

Finally, this newly formed 3' terminal is polyadenylated in the nucleoplasm, as described below.

THE EUKARYOTIC TRANSCRIPTION COMPLEX

A complex apparatus consisting of as many as 50 unique proteins provides accurate and regulatable transcription of eukaryotic genes. The RNA polymerase enzymes (pol I, pol II, and pol III for class I, II, and III genes, respectively) transcribe information contained in the template strand of DNA into RNA. These polymerases must recognize a specific site in the promoter in order to initiate transcription at the proper nucleotide. In contrast to the situation in prokaryotes, eukaryotic RNA polymerases alone are not able to discriminate between promoter sequences and other regions of DNA; thus, other proteins known as general transcription factors or GTFs facilitate promoter-specific binding of these enzymes and formation of the preinitiation complex (PIC). This combination of components can catalyze basal or (non)-unregulated transcription *in vitro*. Another set of proteins—coactivators—help regulate the rate of transcription initiation by interacting with transcription activators that bind to upstream DNA elements (see below).

Formation of the Basal Transcription Complex

In bacteria, a σ factor–polymerase complex selectively binds to DNA in the promoter forming the PIC. The situation is more complex in eukaryotic genes. Class II genes—those transcribed by pol II to make mRNA—are described as an example. In class II genes, the function of σ factors is assumed by a number of proteins. **Basal transcription requires, in addition to pol II, a number of GTFs called TFIIA, TFIIB, TFIID, TFIIE, TFIIIF, and TFIIH.** These GTFs serve to promote RNA polymerase II transcription on essentially all genes. Some of these GTFs are composed of multiple subunits. **TFIID, which binds to the TATA box promoter element, is the only one of these factors capable of binding to specific sequences of DNA.** As described above, TFIID consists of TATA binding protein (TBP) and 14 TBP-associated factors (TAFs).

TBP binds to the TATA box in the minor groove of DNA (most transcription factors bind in the major groove) and causes an approximately 100-degree bend or kink of the DNA helix. This bending is thought to facilitate the interaction of TBP-associated factors with other components of the transcription initiation complex and possibly with factors bound to upstream elements. Although defined as a component of class II gene promoters, TBP, by virtue of its association with

distinct, polymerase-specific sets of TAFs, is also an important component of class I and class III initiation complexes even if they do not contain TATA boxes.

The binding of TBP marks a specific promoter for transcription and is the only step in the assembly process that is entirely dependent on specific, high-affinity protein–DNA interaction. Of several subsequent *in vitro* steps, the first is the binding of TFIIB to the TFIID–promoter complex. This results in a stable ternary complex which is then more precisely located and more tightly bound at the transcription initiation site. This complex then attracts and tethers the pol II–TFIIF complex to the promoter. TFIIF is structurally and functionally similar to the bacterial σ factor and is required for the delivery of pol II to the promoter. TFIIA binds to this assembly and may allow the complex to respond to activators, perhaps by the displacement of repressors. Addition of TFIIE and TFIIH is the final step in the assembly of the PIC. TFIIE appears to join the complex with pol II–TFIIF, and TFIIH is then recruited. Each of these binding events extends the size of the complex so that finally about 60 bp (from –30 to +30 relative to +1, the nucleotide from which transcription commences) are covered (Figure 37–9). The PIC is now complete and capable of basal transcription initiated from the correct nucleotide. In genes that lack a TATA box, the same factors, including TBP, are required. In such cases, an Inr or the DPEs (see Figure 37–8) position the complex for accurate initiation of transcription.

Phosphorylation Activates Pol II

Eukaryotic pol II consists of 12 subunits. The two largest subunits, both about 200 kDa, are homologous to the bacterial β and β' subunits. In addition to the increased number of subunits, eukaryotic pol II differs from its prokaryotic counterpart in that it has a series of heptad repeats with consensus sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser at the carboxyl terminal of the largest pol II subunit. This **carboxyl terminal repeat domain (CTD)** has 26 repeated units in brewers' yeast and 52 units in mammalian cells. The CTD is both a substrate for several kinases, including the kinase component of TFIIH, and a binding site for a wide array of proteins. The CTD has been shown to interact with RNA processing enzymes; such binding may be involved with RNA polyadenylation. The association of the factors with the CTD of RNA polymerase II (and other components of the basal machinery) somehow serves to couple initiation with mRNA 3' end formation. Pol II is activated when phosphorylated on the Ser and Thr residues and displays reduced activity when the CTD is dephosphorylated. Pol II lacking the CTD tail is incapable of activating transcription, which underscores the importance of this domain.

Pol II associates with other proteins to form a holoenzyme complex. In yeast, at least nine gene products—called Srb (for suppressor of RNA polymerase B)—bind to the CTD. The Srb proteins—or mediators, as they are also called—are essential for pol II transcription, though their exact role in this process has not been defined. Related proteins comprising even more complex forms of RNA polymerase II have been described in human cells.

The Role of Transcription Activators & Coactivators

TFIID was originally considered to be a single protein. However, several pieces of evidence led to the important discovery that TFIID is actually a complex consisting of TBP and the 14 TAFs. The first evidence that TFIID was more complex than just the TBP molecules came from the observation that TBP binds to a 10-bp segment of DNA, immediately over the TATA box of the gene, whereas native holo-TFIID covers a 35 bp or larger region (Figure 37–9). Second, TBP has a molecular mass of 20–40 kDa (depending on the species), whereas the TFIID complex has a mass of about 1000 kDa. Finally, and perhaps most importantly, TBP supports basal transcription but not the augmented transcription provided by certain activators, eg, Sp1 bound to the GC box. TFIID, on the other hand, supports both basal and enhanced transcription by Sp1, Oct1, AP1, CTF, ATF, etc. (Table 37–3). The TAFs are essential for this activator-enhanced transcription. It is not yet clear whether there are one or several forms of TFIID that might differ slightly in their complement of

Table 37–3. Some of the transcription control elements, their consensus sequences, and the factors that bind to them which are found in mammalian genes transcribed by RNA polymerase II. A complete list would include dozens of examples. The asterisks mean that there are several members of this family.

Element	Consensus Sequence	Factor
TATA box	TATAAA	TBP
CAAT box	CCAATC	C/EBP*, NF-Y*
GC box	GGGCGG	Sp1*
	CAACTGAC	Myo D
	T/CGGA/CN ₅ GCCAA	NF1*
Ig octamer	ATGCAAT	Oct1, 2, 4, 6*
AP1	TGAG/CTC/AA	Jun, Fos, ATF*
Serum response	GATGCCATA	SRF
Heat shock	(NGAAN) ₃	HSF

TAFs. It is conceivable that different combinations of TAFs with TBP—or one of several recently discovered TBP-like factors (TLFs)—may bind to different promoters, and recent reports suggest that this may account for selective activation noted in various promoters and for the different strengths of certain promoters. **TAFs, since they are required for the action of activators, are often called coactivators.** There are thus three classes of transcription factors involved in the regulation of class II genes: basal factors, coactivators, and activator-repressors (Table 37–4). How these classes of proteins interact to govern both the site and frequency of transcription is a question of central importance.

Two Models Explain the Assembly of the Preinitiation Complex

The formation of the PIC described above is based on the sequential addition of purified components in *in vitro* experiments. An essential feature of this model is that the assembly takes place on the DNA template. Accordingly, transcription activators, which have autonomous DNA binding and activation domains (see Chapter 39), are thought to function by stimulating either PIC formation or PIC function. The TAF coactivators are viewed as bridging factors that communicate between the upstream activators, the proteins associated with pol II, or the many other components of TFIID. This view, which assumes that there is **stepwise assembly** of the PIC—promoted by various interactions between activators, coactivators, and PIC components—is illustrated in panel A of Figure 37–10. This model was supported by observations that many of these proteins could indeed bind to one another *in vitro*.

Recent evidence suggests that there is another possible mechanism of PIC formation and transcription regulation. First, large preassembled complexes of GTFs and pol II are found in cell extracts, and this complex can associate with a promoter in a single step. Second, the rate of transcription achieved when activators are added to limiting concentrations of pol II holoenzyme can be matched by increasing the concentration of the pol II holoenzyme in the absence of activators. Thus,

Table 37–4. Three classes of transcription factors in class II genes.

General Mechanisms	Specific Components
Basal components	TBP, TFIIB, E, F, and H
Coactivators	TAFs (TBP + TAFs) = TFIID; Srfbs
Activators	SP1, ATF, CTF, AP1, etc

activators are not in themselves absolutely essential for PIC formation. These observations led to the “**recruitment hypothesis**,” which has now been tested experimentally. Simply stated, the role of activators and coactivators may be solely to recruit a preformed holoenzyme-GTF complex to the promoter. The requirement for an activation domain is circumvented when either a component of TFIID or the pol II holoenzyme is artificially tethered, using recombinant DNA techniques, to the DNA binding domain (DBD) of an activator. This anchoring, through the DBD component of the activator molecule, leads to a transcriptionally competent structure, and there is no further requirement for the activation domain of the activator. In this view, the role of activation domains and TAFs is to form an assembly that directs the preformed holoenzyme-GTF complex to the promoter; they do not assist in PIC assembly (see panel B, Figure 37–10). The efficiency of this recruitment process determines the rate of transcription at a given promoter.

Hormones—and other effectors that serve to transmit information related to the extracellular environment—modulate gene expression by influencing the assembly and activity of the activator and coactivator complexes and the subsequent formation of the PIC at the promoter of target genes (see Chapter 43). The numerous components involved provide for an abundance of possible combinations and therefore a range of transcriptional activity of a given gene. It is important to note that the two models are not mutually exclusive—stepwise versus holoenzyme-mediated PIC formation. Indeed, one can envision various more complex models invoking elements of both models operating on a gene.

RNA MOLECULES ARE USUALLY PROCESSED BEFORE THEY BECOME FUNCTIONAL

In prokaryotic organisms, the primary transcripts of mRNA-encoding genes begin to serve as translation templates even before their transcription has been completed. This is because the site of transcription is not compartmentalized into a nucleus as it is in eukaryotic organisms. Thus, transcription and translation are coupled in prokaryotic cells. Consequently, prokaryotic mRNAs are subjected to little processing prior to carrying out their intended function in protein synthesis. Indeed, appropriate regulation of some genes (eg, the *Trp* operon) relies upon this coupling of transcription and translation. Prokaryotic rRNA and tRNA molecules are transcribed in units considerably longer than the ultimate molecule. In fact, many of the tRNA transcription units contain more than one molecule. Thus, in prokaryotes the processing of these rRNA and tRNA

precursor molecules is required for the generation of the mature functional molecules.

Nearly all eukaryotic RNA primary transcripts undergo extensive processing between the time they are synthesized and the time at which they serve their ultimate function, whether it be as mRNA or as a component of the translation machinery such as rRNA, 5S RNA, or tRNA or RNA processing machinery, snRNAs. Processing occurs primarily within the nucleus and includes nucleolytic cleavage to smaller molecules and coupled **nucleolytic and ligation reactions (splicing of exons)**. In mammalian cells, 50–75% of the nuclear RNA does not contribute to the cytoplasmic mRNA. This nuclear RNA loss is significantly greater than can be reasonably accounted for by the loss of intervening sequences alone (see below). Thus, the exact function of the seemingly excessive transcripts in the nucleus of a mammalian cell is not known.

The Coding Portions (Exons) of Most Eukaryotic Genes Are Interrupted by Introns

Interspersed within the amino acid-coding portions (**exons**) of many genes are long sequences of DNA that do not contribute to the genetic information ultimately translated into the amino acid sequence of a protein molecule (see Chapter 36). In fact, these sequences actually interrupt the coding region of structural genes. These **intervening sequences (introns)** exist within most but not all mRNA encoding genes of higher eukaryotes. The primary transcripts of the structural genes contain RNA complementary to the interspersed sequences. However, the intron RNA sequences are cleaved out of the transcript, and the exons of the transcript are appropriately spliced together in the nucleus before the resulting mRNA molecule appears in the cytoplasm for translation (Figures 37–11 and 37–12). One speculation is that exons, which often encode an activity domain of a protein, represent a convenient means of shuffling genetic information, permitting organisms to quickly test the results of combining novel protein functional domains.

Introns Are Removed & Exons Are Spliced Together

The mechanisms whereby introns are removed from the primary transcript in the nucleus, exons are ligated to form the mRNA molecule, and the mRNA molecule is transported to the cytoplasm are being elucidated. Four different splicing reaction mechanisms have been described. The one most frequently used in eukaryotic cells is described below. Although the sequences of nu-

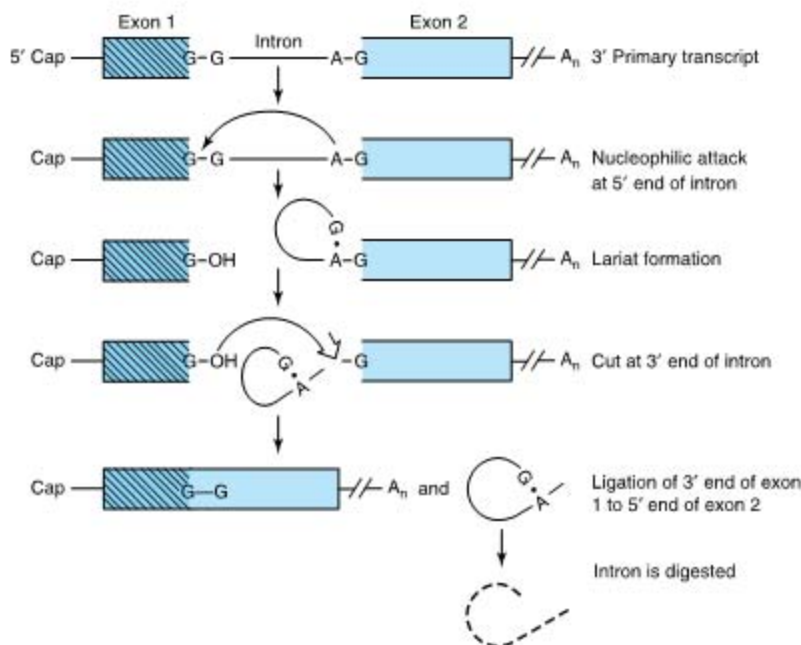


Figure 37-11. The processing of the primary transcript to mRNA. In this hypothetical transcript, the 5' (left) end of the intron is cut (\downarrow) and a lariat forms between the G at the 5' end of the intron and an A near the 3' end, in the consensus sequence UACUAAC. This sequence is called the branch site, and it is the 3' most A that forms the 5'-2' bond with the G. The 3' (right) end of the intron is then cut (\downarrow). This releases the lariat, which is digested, and exon 1 is joined to exon 2 at G residues.

cleotides in the introns of the various eukaryotic transcripts—and even those within a single transcript—are quite heterogeneous, there are reasonably conserved sequences at each of the two exon-intron (splice) junctions and at the branch site, which is located 20–40 nucleotides upstream from the 3' splice site (see consensus sequences in Figure 37-12). A special structure, the **spliceosome**, is involved in converting the primary transcript into mRNA. Spliceosomes consist of the pri-

mary transcript, five small nuclear RNAs (U1, U2, U5, U4, and U6) and more than 60 proteins. Collectively, these form a **small nucleoprotein (snRNP) complex**, sometimes called a “**snurp**.” It is likely that this pentasnrNP spliceosome forms prior to interaction with mRNA precursors. Snurps are thought to position the RNA segments for the necessary splicing reactions. The splicing reaction starts with a cut at the junction of the 5' exon (donor or left) and intron (Figure 37-11). This

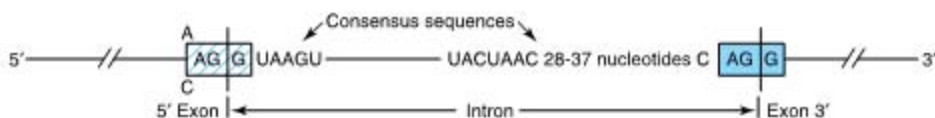


Figure 37-12. Consensus sequences at splice junctions. The 5' (donor or left) and 3' (acceptor or right) sequences are shown. Also shown is the yeast consensus sequence (UACUAAC) for the branch site. In mammalian cells, this consensus sequence is PyNPpyPuAPy, where Py is a pyrimidine, Pu is a purine, and N is any nucleotide. The branch site is located 20–40 nucleotides upstream from the 3' site.

is accomplished by a nucleophilic attack by an adenyl residue in the branch point sequence located just upstream from the 3' end of this intron. The free 5' terminal then forms a loop or lariat structure that is linked by an unusual 5'-2' phosphodiester bond to the reactive A in the PyNPYPuAPy branch site sequence (Figure 37-12). This adenyl residue is typically located 28-37 nucleotides upstream from the 3' end of the intron being removed. The branch site identifies the 3' splice site. A second cut is made at the junction of the intron with the 3' exon (donor on right). In this second transesterification reaction, the 3' hydroxyl of the upstream exon attacks the 5' phosphate at the downstream exon-intron boundary, and the lariat structure containing the intron is released and hydrolyzed. The 5' and 3' exons are ligated to form a continuous sequence.

The snRNAs and associated proteins are required for formation of the various structures and intermediates. U1 within the snRNP complex binds first by base pairing to the 5' exon-intron boundary. U2 within the snRNP complex then binds by base pairing to the branch site, and this exposes the nucleophilic A residue. U5/U4/U6 within the snRNP complex mediates an ATP-dependent protein-mediated unwinding that results in disruption of the base-paired U4-U6 complex with the release of U4. U6 is then able to interact first with U2, then with U1. These interactions serve to approximate the 5' splice site, the branch point with its reactive A, and the 3' splice site. This alignment is enhanced by U5. This process also results in the formation of the loop or lariat structure. **The two ends are cleaved, probably by the U2-U6 within the snRNP complex.** U6 is certainly essential, since yeasts deficient in this snRNA are not viable. It is important to note that RNA serves as the catalytic agent. This sequence is then repeated in genes containing multiple introns. In such cases, a definite pattern is followed for each gene, and the introns are not necessarily removed in sequence—1, then 2, then 3, etc.

The relationship between hnRNA and the corresponding mature mRNA in eukaryotic cells is now apparent. The hnRNA molecules are the primary transcripts plus their early processed products, which, after the addition of caps and poly(A) tails and removal of the portion corresponding to the introns, are transported to the cytoplasm as mature mRNA molecules.

Alternative Splicing Provides for Different mRNAs

The processing of hnRNA molecules is a site for regulation of gene expression. Alternative patterns of RNA splicing result from tissue-specific adaptive and developmental control mechanisms. As mentioned

above, the sequence of exon-intron splicing events generally follows a hierarchical order for a given gene. The fact that very complex RNA structures are formed during splicing—and that a number of snRNAs and proteins are involved—affords numerous possibilities for a change of this order and for the generation of different mRNAs. Similarly, the use of alternative termination-cleavage-polyadenylation sites also results in mRNA heterogeneity. Some schematic examples of these processes, all of which occur in nature, are shown in Figure 37-13.

Faulty splicing can cause disease. At least one form of β -thalassemia, a disease in which the β -globin gene of hemoglobin is severely underexpressed, appears to result from a nucleotide change at an exon-intron junction, precluding removal of the intron and therefore leading to diminished or absent synthesis of the β -chain protein. This is a consequence of the fact that the normal translation reading frame of the mRNA is disrupted—a defect in this fundamental process (splicing) that underscores the accuracy which the process of RNA-RNA splicing must achieve.

Alternative Promoter Utilization Provides a Form of Regulation

Tissue-specific regulation of gene expression can be provided by control elements in the promoter or by the

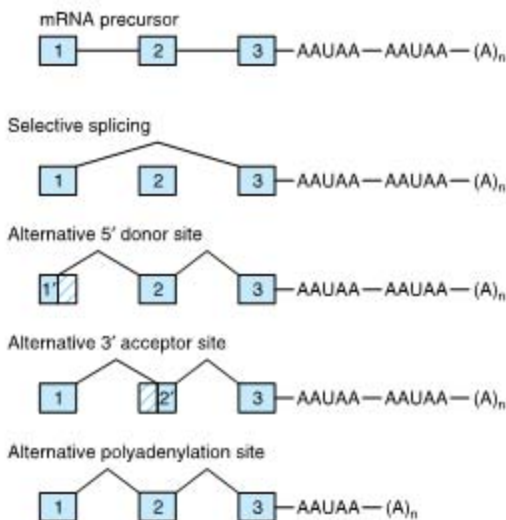


Figure 37-13. Mechanisms of alternative processing of mRNA precursors. This form of RNA processing involves the selective inclusion or exclusion of exons, the use of alternative 5' donor or 3' acceptor sites, and the use of different polyadenylation sites.

use of alternative promoters. The glucokinase (*GK*) gene consists of ten exons interrupted by nine introns. The sequence of exons 2–10 is identical in liver and pancreatic B cells, the primary tissues in which GK protein is expressed. Expression of the *GK* gene is regulated very differently—by two different promoters—in these two tissues. The liver promoter and exon 1L are located near exons 2–10; exon 1L is ligated directly to exon 2. In contrast, the pancreatic B cell promoter is located about 30 kbp upstream. In this case, the 3' boundary of exon 1B is ligated to the 5' boundary of exon 2. The liver promoter and exon 1L are excluded and removed during the splicing reaction (see Figure 37–14). The existence of multiple distinct promoters allows for cell- and tissue-specific expression patterns of a particular gene (mRNA).

Both Ribosomal RNAs & Most Transfer RNAs Are Processed From Larger Precursors

In mammalian cells, the three rRNA molecules are transcribed as part of a single large precursor molecule. **The precursor is subsequently processed in the nucleolus** to provide the RNA component for the ribosome subunits found in the cytoplasm. The rRNA genes are located in the nucleoli of mammalian cells. Hundreds of copies of these genes are present in every cell. This large number of genes is required to synthesize sufficient copies of each type of rRNA to form the 10^7 ribosomes required for each cell replication. Whereas a single mRNA molecule may be copied into 10^5 protein molecules, providing a large amplification, the rRNAs are end products. This lack of amplification requires a large number of genes. Similarly, transfer RNAs are often synthesized as precursors, with extra sequences both 5' and 3' of the sequences comprising the

mature tRNA. A small fraction of tRNAs even contain introns.

RNAs CAN BE EXTENSIVELY MODIFIED

Essentially all RNAs are covalently modified after transcription. It is clear that at least some of these modifications are regulatory.

Messenger RNA (mRNA) Is Modified at the 5' & 3' Ends

As mentioned above, mammalian mRNA molecules contain a 7-methylguanosine cap structure at their 5' terminal, and most have a poly(A) tail at the 3' terminal. The cap structure is added to the 5' end of the newly transcribed mRNA precursor in the nucleus prior to transport of the mRNA molecule to the cytoplasm. The 5' cap of the RNA transcript is required both for efficient translation initiation and protection of the 5' end of mRNA from attack by 5' → 3' exonucleases. The secondary methylations of mRNA molecules, those on the 2'-hydroxy and the N⁶ of adenylyl residues, occur after the mRNA molecule has appeared in the cytoplasm.

Poly(A) tails are added to the 3' end of mRNA molecules in a posttranscriptional processing step. The mRNA is first cleaved about 20 nucleotides downstream from an AAUAA recognition sequence. Another enzyme, poly(A) polymerase, adds a poly(A) tail which is subsequently extended to as many as 200 A residues. The **poly(A) tail** appears to protect the 3' end of mRNA from 3' → 5' exonuclease attack. The presence or absence of the poly(A) tail does not determine whether a precursor molecule in the nucleus appears in the cytoplasm, because all poly(A)-tailed hnRNA molecules do not contribute to cytoplasmic mRNA, nor do all cytoplasmic mRNA molecules contain poly(A) tails

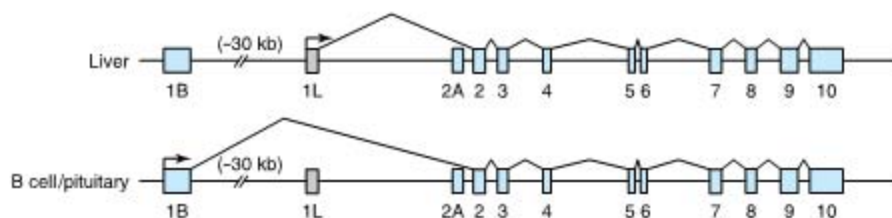


Figure 37–14. Alternative promoter use in the liver and pancreatic B cell glucokinase genes. Differential regulation of the glucokinase (*GK*) gene is accomplished by the use of tissue-specific promoters. The B cell *GK* gene promoter and exon 1B are located about 30 kbp upstream from the liver promoter and exon 1L. Each promoter has a unique structure and is regulated differently. Exons 2–10 are identical in the two genes, and the *GK* proteins encoded by the liver and B cell mRNAs have identical kinetic properties.

(the histones are most notable in this regard). Cytoplasmic enzymes in mammalian cells can both add and remove adenyl residues from the poly(A) tails; this process has been associated with an alteration of mRNA stability and translatability.

The size of some cytoplasmic mRNA molecules, even after the poly(A) tail is removed, is still considerably greater than the size required to code for the specific protein for which it is a template, often by a factor of 2 or 3. **The extra nucleotides occur in untranslated (non-protein coding) regions** both 5' and 3' of the coding region; the longest untranslated sequences are usually at the 3' end. The exact function of these sequences is unknown, but they have been implicated in RNA processing, transport, degradation, and translation; each of these reactions potentially contributes additional levels of control of gene expression.

RNA Editing Changes mRNA After Transcription

The central dogma states that for a given gene and gene product there is a linear relationship between the coding sequence in DNA, the mRNA sequence, and the protein sequence (Figure 36–7). Changes in the DNA sequence should be reflected in a change in the mRNA sequence and, depending on codon usage, in protein sequence. However, exceptions to this dogma have been recently documented. Coding information can be changed at the mRNA level by **RNA editing**. In such cases, the coding sequence of the mRNA differs from that in the cognate DNA. An example is the apolipoprotein B (*apoB*) gene and mRNA. In liver, the single *apoB* gene is transcribed into an mRNA that directs the synthesis of a 100-kDa protein, apoB100. In the intestine, the same gene directs the synthesis of the primary transcript; however, a cytidine deaminase converts a CAA codon in the mRNA to UAA at a single specific site. Rather than encoding glutamine, this codon becomes a termination signal, and a 48-kDa protein (apoB48) is the result. ApoB100 and apoB48 have different functions in the two organs. A growing number of other examples include a glutamine to arginine change in the glutamate receptor and several changes in trypanosome mitochondrial mRNAs, generally involving the addition or deletion of uridine. The exact extent of RNA editing is unknown, but current estimates suggest that < 0.01% of mRNAs are edited in this fashion.

Transfer RNA (tRNA) Is Extensively Processed & Modified

As described in Chapters 35 and 38, the tRNA molecules serve as adapter molecules for the translation of

mRNA into protein sequences. The tRNAs contain many modifications of the standard bases A, U, G, and C, including methylation, reduction, deamination, and rearranged glycosidic bonds. Further modification of the tRNA molecules includes nucleotide alkylations and the attachment of the characteristic CpCpAOH terminal at the 3' end of the molecule by the enzyme nucleotidyl transferase. The 3' OH of the A ribose is the point of attachment for the specific amino acid that is to enter into the polymerization reaction of protein synthesis. The methylation of mammalian tRNA precursors probably occurs in the nucleus, whereas the cleavage and attachment of CpCpAOH are cytoplasmic functions, since the terminals turn over more rapidly than do the tRNA molecules themselves. Enzymes within the cytoplasm of mammalian cells are required for the attachment of amino acids to the CpCpAOH residues. (See Chapter 38.)

RNA CAN ACT AS A CATALYST

In addition to the catalytic action served by the snRNAs in the formation of mRNA, several other enzymatic functions have been attributed to RNA. **Ribozymes** are RNA molecules with catalytic activity. These generally involve transesterification reactions, and most are concerned with RNA metabolism (splicing and endoribonuclease). Recently, a ribosomal RNA component was noted to hydrolyze an aminoacyl ester and thus to play a central role in peptide bond function (peptidyl transferases; see Chapter 38). These observations, made in organelles from plants, yeast, viruses, and higher eukaryotic cells, show that RNA can act as an enzyme. This has revolutionized thinking about enzyme action and the origin of life itself.

SUMMARY

- RNA is synthesized from a DNA template by the enzyme RNA polymerase.
- There are three distinct nuclear DNA-dependent RNA polymerases in mammals: RNA polymerases I, II, and III. These enzymes control the transcriptional function—the transcription of rRNA, mRNA, and small RNA (tRNA/5S rRNA, snRNA) genes, respectively.
- RNA polymerases interact with unique *cis*-active regions of genes, termed promoters, in order to form preinitiation complexes (PICs) capable of initiation. In eukaryotes the process of PIC formation is facilitated by multiple general transcription factors (GTFs), TFIIA, B, D, E, F, and H.
- Eukaryotic PIC formation can occur either stepwise—by the sequential, ordered interactions of

GTFs and RNA polymerase with promoters—or in one step by the recognition of the promoter by a preformed GTF-RNA polymerase holoenzyme complex.

- Transcription exhibits three phases: initiation, elongation, and termination. All are dependent upon distinct DNA *cis*-elements and can be modulated by distinct *trans*-acting protein factors.
- Most eukaryotic RNAs are synthesized as precursors that contain excess sequences which are removed prior to the generation of mature, functional RNA.
- Eukaryotic mRNA synthesis results in a pre-mRNA precursor that contains extensive amounts of excess RNA (introns) that must be precisely removed by RNA splicing to generate functional, translatable mRNA composed of exonic coding and noncoding sequences.
- All steps—from changes in DNA template, sequence, and accessibility in chromatin to RNA stability—are subject to modulation and hence are potential control sites for eukaryotic gene regulation.

REFERENCES

- Busby S, Ebright RH: Promoter structure, promoter recognition, and transcription activation in prokaryotes. *Cell* 1994;79:743.
- Cramer P, Bushnell DA, Kornberg R: Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* 2001;292:1863.
- Fedor MJ: Ribozymes. *Curr Biol* 1998;8:R441.
- Gott JM, Emeson RB: Functions and mechanisms of RNA editing. *Ann Rev Genet* 2000;34:499.
- Hirose Y, Manley JL: RNA polymerase II and the integration of nuclear events. *Genes Dev* 2000;14:1415.
- Keaveney M, Struhl K: Activator-mediated recruitment of the RNA polymerase machinery is the predominant mechanism for transcriptional activation in yeast. *Mol Cell* 1998;1:917.
- Lemon B, Tjian R: Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* 2000;14:2551.
- Maniatis T, Reed R: An extensive network of coupling among gene expression machines. *Nature* 2002;416:499.
- Orphanides G, Reinberg D: A unified theory of gene expression. *Cell* 2002;108:439.
- Shatkin AJ, Manley JL: The ends of the affair: capping and polyadenylation. *Nat Struct Biol* 2000;7:838.
- Stevens SW et al: Composition and functional characterization of the yeast spliceosomal penta-snRNP. *Mol Cell* 2002;9:31.
- Tucker M, Parker R: Mechanisms and control of mRNA decapping in *Saccharomyces cerevisiae*. *Ann Rev Biochem* 2000;69:571.
- Woychik NA, Hampsey M: The RNA polymerase II machinery: structure illuminates function. *Cell* 2002;108:453.

Protein Synthesis & the Genetic Code

38

Daryl K. Granner, MD

BIOMEDICAL IMPORTANCE

The letters A, G, T, and C correspond to the nucleotides found in DNA. They are organized into three-letter code words called **codons**, and the collection of these codons makes up the **genetic code**. It was impossible to understand protein synthesis—or to explain mutations—before the genetic code was elucidated. The code provides a foundation for explaining the way in which protein defects may cause genetic disease and for the diagnosis and perhaps the treatment of these disorders. In addition, the pathophysiology of many viral infections is related to the ability of these agents to disrupt host cell protein synthesis. Many antibacterial agents are effective because they selectively disrupt protein synthesis in the invading bacterial cell but do not affect protein synthesis in eukaryotic cells.

GENETIC INFORMATION FLOWS FROM DNA TO RNA TO PROTEIN

The genetic information within the nucleotide sequence of DNA is transcribed in the nucleus into the specific nucleotide sequence of an RNA molecule. The sequence of nucleotides in the RNA transcript is complementary to the nucleotide sequence of the template strand of its gene in accordance with the base-pairing rules. Several different classes of RNA combine to direct the synthesis of proteins.

In prokaryotes there is a linear correspondence between the gene, the **messenger RNA (mRNA)** transcribed from the gene, and the polypeptide product. The situation is more complicated in higher eukaryotic cells, in which the primary transcript is much larger than the mature mRNA. The large mRNA precursors contain coding regions (**exons**) that will form the mature mRNA and long intervening sequences (**introns**) that separate the exons. The hnRNA is processed within the nucleus, and the introns, which often make up much more of this RNA than the exons, are removed. Exons are spliced together to form mature mRNA, which is transported to the cytoplasm, where it is translated into protein.

The cell must possess the machinery necessary to translate information accurately and efficiently from the nucleotide sequence of an mRNA into the sequence of amino acids of the corresponding specific protein. Clarification of our understanding of this process, which is termed **translation**, awaited deciphering of the genetic code. It was realized early that mRNA molecules themselves have no affinity for amino acids and, therefore, that the translation of the information in the mRNA nucleotide sequence into the amino acid sequence of a protein requires an intermediate adapter molecule. This adapter molecule must recognize a specific nucleotide sequence on the one hand as well as a specific amino acid on the other. With such an adapter molecule, the cell can direct a specific amino acid into the proper sequential position of a protein during its synthesis as dictated by the nucleotide sequence of the specific mRNA. In fact, the functional groups of the amino acids do not themselves actually come into contact with the mRNA template.

THE NUCLEOTIDE SEQUENCE OF AN mRNA MOLECULE CONSISTS OF A SERIES OF CODONS THAT SPECIFY THE AMINO ACID SEQUENCE OF THE ENCODED PROTEIN

Twenty different amino acids are required for the synthesis of the cellular complement of proteins; thus, there must be at least 20 distinct codons that make up the genetic code. Since there are only four different nucleotides in mRNA, each codon must consist of more than a single purine or pyrimidine nucleotide. Codons consisting of two nucleotides each could provide for only 16 (4^2) specific codons, whereas codons of three nucleotides could provide 64 (4^3) specific codons.

It is now known that each codon consists of a sequence of three nucleotides; ie, it is a **triplet code** (see Table 38-1). The deciphering of the genetic code depended heavily on the chemical synthesis of nucleotide polymers, particularly triplets in repeated sequence.

Table 38–1. The genetic code (codon assignments in mammalian messenger RNA).¹

First Nucleotide	Second Nucleotide				Third Nucleotide
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Term	Term ²	A
	Leu	Ser	Term	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile ²	Thr	Lys	Arg ²	A
	Met	Thr	Lys	Arg ²	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

¹The terms first, second, and third nucleotide refer to the individual nucleotides of a triplet codon. U, uridine nucleotide; C, cytosine nucleotide; A, adenine nucleotide; G, guanine nucleotide; Term, chain terminator codon. AUG, which codes for Met, serves as the initiator codon in mammalian cells and encodes for internal methionines in a protein. (Abbreviations of amino acids are explained in Chapter 3.)

²In mammalian mitochondria, AUA codes for Met and UGA for Trp, and AGA and AGG serve as chain terminators.

THE GENETIC CODE IS DEGENERATE, UNAMBIGUOUS, NONOVERLAPPING, WITHOUT PUNCTUATION, & UNIVERSAL

Three of the 64 possible codons do not code for specific amino acids; these have been termed **nonsense codons**. These nonsense codons are utilized in the cell as **termination signals**; they specify where the polymerization of amino acids into a protein molecule is to stop. The remaining 61 codons code for 20 amino acids (Table 38–1). Thus, there must be “**degeneracy**” in the genetic code—ie, multiple codons must decode the same amino acid. Some amino acids are encoded by several codons; for example, six different codons specify serine. Other amino acids, such as methionine and tryptophan, have a single codon. In general, the third nucleotide in a codon is less important than the first two in determining the specific amino acid to be incorporated, and this accounts for most of the degeneracy of

the code. However, for any specific codon, only a single amino acid is indicated; with rare exceptions, the genetic code is **unambiguous**—ie, given a specific codon, only a single amino acid is indicated. **The distinction between ambiguity and degeneracy is an important concept.**

The unambiguous but degenerate code can be explained in molecular terms. The recognition of specific codons in the mRNA by the tRNA adapter molecules is dependent upon their **anticodon region** and specific base-pairing rules. Each tRNA molecule contains a specific sequence, complementary to a codon, which is termed its anticodon. For a given codon in the mRNA, only a single species of tRNA molecule possesses the proper anticodon. Since each tRNA molecule can be charged with only one specific amino acid, each codon therefore specifies only one amino acid. However, some tRNA molecules can utilize the anticodon to recognize more than one codon. **With few exceptions, given a specific codon, only a specific amino acid will be incorporated—although, given a specific amino acid, more than one codon may be used.**

As discussed below, the reading of the genetic code during the process of protein synthesis does not involve any overlap of codons. **Thus, the genetic code is nonoverlapping.** Furthermore, once the reading is commenced at a specific codon, there is **no punctuation** between codons, and the message is read in a continuing sequence of nucleotide triplets until a translation stop codon is reached.

Until recently, the genetic code was thought to be universal. It has now been shown that the set of tRNA molecules in mitochondria (which contain their own separate and distinct set of translation machinery) from lower and higher eukaryotes, including humans, reads four codons differently from the tRNA molecules in the cytoplasm of even the same cells. As noted in Table 38–1, the codon AUA is read as Met, and UGA codes for Trp in mammalian mitochondria. In addition, in mitochondria, the codons AGA and AGG are read as stop or chain terminator codons rather than as Arg. As a result, mitochondria require only 22 tRNA molecules to read their genetic code, whereas the cytoplasmic translation system possesses a full complement of 31 tRNA species. These exceptions noted, **the genetic code is universal.** The frequency of use of each amino acid codon varies considerably between species and among different tissues within a species. The specific tRNA levels generally mirror these codon usage biases. Thus, a particular abundantly used codon is decoded by a similarly abundant specific tRNA which recognizes that particular codon. Tables of **codon usage** are becoming more accurate as more genes are sequenced. This is of considerable importance because investigators

Table 38–2. Features of the genetic code.

-
- Degenerate
 - Unambiguous
 - Nonoverlapping
 - Not punctuated
 - Universal
-

often need to deduce mRNA structure from the primary sequence of a portion of protein in order to synthesize an oligonucleotide probe and initiate a recombinant DNA cloning project. The main features of the genetic code are listed in Table 38–2.

AT LEAST ONE SPECIES OF TRANSFER RNA (tRNA) EXISTS FOR EACH OF THE 20 AMINO ACIDS

tRNA molecules have extraordinarily similar functions and three-dimensional structures. The adapter function of the tRNA molecules requires the charging of each specific tRNA with its specific amino acid. Since there is no affinity of nucleic acids for specific functional groups of amino acids, this recognition must be carried out by a protein molecule capable of recognizing both a specific tRNA molecule and a specific amino acid. At least 20 specific enzymes are required for these specific recognition functions and for the proper attachment of the 20 amino acids to specific tRNA molecules. The process of **recognition and attachment (charging)** proceeds in two steps by one enzyme for each of the 20 amino acids. These enzymes are termed **aminoacyl-**

tRNA synthetases. They form an activated intermediate of aminoacyl-AMP-enzyme complex (Figure 38–1). The specific aminoacyl-AMP-enzyme complex then recognizes a specific tRNA to which it attaches the aminoacyl moiety at the 3'-hydroxyl adenosyl terminal. The charging reactions have an error rate of less than 10^{-4} and so are extremely accurate. The amino acid remains attached to its specific tRNA in an ester linkage until it is polymerized at a specific position in the fabrication of a polypeptide precursor of a protein molecule.

The regions of the tRNA molecule referred to in Chapter 35 (and illustrated in Figure 35–11) now become important. The thymidine-pseudouridine-cytidine (T Ψ C) arm is involved in binding of the aminoacyl-tRNA to the ribosomal surface at the site of protein synthesis. The D arm is one of the sites important for the proper recognition of a given tRNA species by its proper aminoacyl-tRNA synthetase. The acceptor arm, located at the 3'-hydroxyl adenosyl terminal, is the site of attachment of the specific amino acid.

The anticodon region consists of seven nucleotides, and it recognizes the three-letter codon in mRNA (Figure 38–2). The sequence read from the 3' to 5' direction in that anticodon loop consists of a variable base-modified purine-XYZ-pyrimidine-pyrimidine-5'. Note that this direction of reading the anticodon is 3' to 5', whereas the genetic code in Table 38–1 is read 5' to 3', since the codon and the anticodon loop of the mRNA and tRNA molecules, respectively, are **antiparallel** in their complementarity just like all other intermolecular interactions between nucleic acid strands.

The degeneracy of the genetic code resides mostly in the last nucleotide of the codon triplet, suggesting that the base pairing between this last nucleotide and the corresponding nucleotide of the anticodon is not strictly

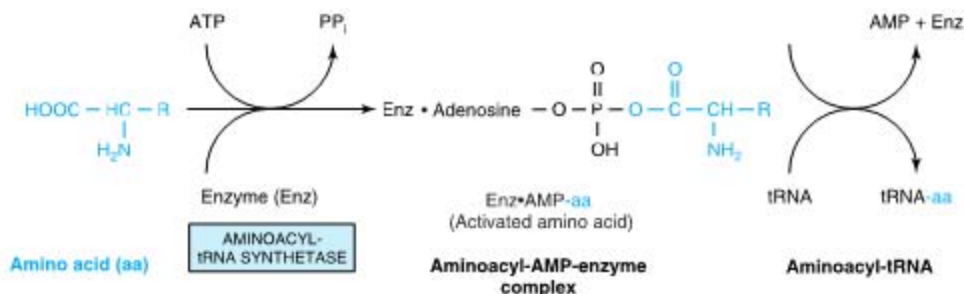


Figure 38–1. Formation of aminoacyl-tRNA. A two-step reaction, involving the enzyme aminoacyl-tRNA synthetase, results in the formation of aminoacyl-tRNA. The first reaction involves the formation of an AMP-amino acid-enzyme complex. This activated amino acid is next transferred to the corresponding tRNA molecule. The AMP and enzyme are released, and the latter can be reutilized. The charging reactions have an error rate of less than 10^{-4} and so are extremely accurate.

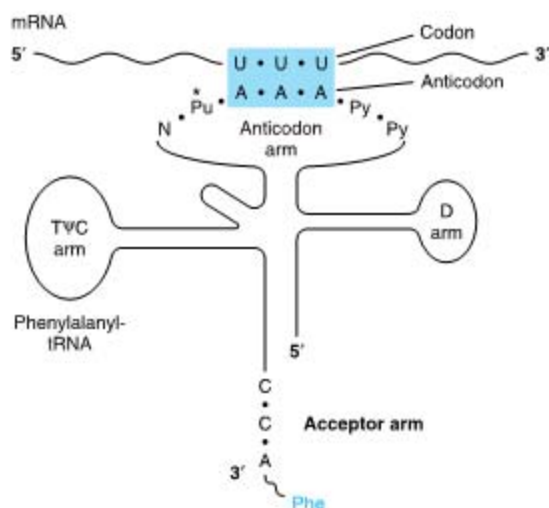


Figure 38–2. Recognition of the codon by the anticodon. One of the codons for phenylalanine is UUU. tRNA charged with phenylalanine (Phe) has the complementary sequence AAA; hence, it forms a base-pair complex with the codon. The anticodon region typically consists of a sequence of seven nucleotides: variable (N), modified purine ((Pu*), X, Y, Z, and two pyrimidines (Py) in the 3' to 5' direction.

by the Watson-Crick rule. This is called **wobble**; the pairing of the codon and anticodon can “wobble” at this specific nucleotide-to-nucleotide pairing site. For example, the two codons for arginine, AGA and AGG, can bind to the same anticodon having a uracil at its 5' end (UCU). Similarly, three codons for glycine—GGU, GGC, and GGA—can form a base pair from one anticodon, CCI. I is an inosine nucleotide, another of the peculiar bases appearing in tRNA molecules.

MUTATIONS RESULT WHEN CHANGES OCCUR IN THE NUCLEOTIDE SEQUENCE

Although the initial change may not occur in the template strand of the double-stranded DNA molecule for that gene, after replication, daughter DNA molecules with mutations in the template strand will segregate and appear in the population of organisms.

Some Mutations Occur by Base Substitution

Single-base changes (**point mutations**) may be **transitions** or **transversions**. In the former, a given pyrimidine is changed to the other pyrimidine or a given

purine is changed to the other purine. Transversions are changes from a purine to either of the two pyrimidines or the change of a pyrimidine into either of the two purines, as shown in Figure 38–3.

If the nucleotide sequence of the gene containing the mutation is transcribed into an RNA molecule, then the RNA molecule will possess a complementary base change at this corresponding locus.

Single-base changes in the mRNA molecules may have one of several effects when translated into protein:

(1) There may be no detectable effect because of the degeneracy of the code. This would be more likely if the changed base in the mRNA molecule were to be at the third nucleotide of a codon; such mutations are often referred to as **silent mutations**. Because of wobble, the translation of a codon is least sensitive to a change at the third position.

(2) A **missense effect** will occur when a different amino acid is incorporated at the corresponding site in the protein molecule. This mistaken amino acid—or missense, depending upon its location in the specific protein—might be acceptable, partially acceptable, or unacceptable to the function of that protein molecule. From a careful examination of the genetic code, one can conclude that most single-base changes would result in the replacement of one amino acid by another with rather similar functional groups. This is an effective mechanism to avoid drastic change in the physical properties of a protein molecule. If an acceptable missense effect occurs, the resulting protein molecule may not be distinguishable from the normal one. A partially acceptable missense will result in a protein molecule with partial but abnormal function. If an unacceptable missense effect occurs, then the protein molecule will not be capable of functioning in its assigned role.

(3) A **nonsense** codon may appear that would then result in the **premature termination** of amino acid incorporation into a peptide chain and the production of only a fragment of the intended protein molecule. The probability is high that a prematurely terminated protein molecule or peptide fragment will not function in its assigned role.

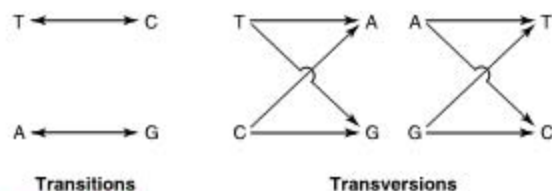


Figure 38–3. Diagrammatic representation of transition mutations and transversion mutations.

Hemoglobin Illustrates the Effects of Single-Base Changes in Structural Genes

Some mutations have no apparent effect. The gene system that encodes hemoglobin is one of the best-studied in humans. The lack of effect of a single-base change is demonstrable only by sequencing the nucleotides in the mRNA molecules or structural genes. The sequencing of a large number of hemoglobin mRNAs and genes from many individuals has shown that the codon for valine at position 67 of the β chain of hemoglobin is not identical in all persons who possess a normally functional β chain of hemoglobin. Hemoglobin Milwaukee has at position 67 a glutamic acid; hemoglobin Bristol contains aspartic acid at position 67. In order to account for the amino acid change by the change of a single nucleotide residue in the codon for amino acid 67, one must infer that the mRNA encoding hemoglobin Bristol possessed a GUU or GUC codon prior to a later change to GAU or GAC, both codons for aspartic acid. However, the mRNA encoding hemoglobin Milwaukee would have

to possess at position 67 a codon GUA or GUG in order that a single nucleotide change could provide for the appearance of the glutamic acid codons GAA or GAG. Hemoglobin Sydney, which contains an alanine at position 67, could have arisen by the change of a single nucleotide in any of the four codons for valine (GUU, GUC, GUA, or GUG) to the alanine codons (GCU, GCC, GCA, or GCG, respectively).

Substitution of Amino Acids Causes Missense Mutations

A. ACCEPTABLE MISSENSE MUTATIONS

An example of an acceptable missense mutation (Figure 38–4, top) in the structural gene for the β chain of hemoglobin could be detected by the presence of an electrophoretically altered hemoglobin in the red cells of an apparently healthy individual. Hemoglobin Hikari has been found in at least two families of Japanese people. This hemoglobin has asparagine substituted for lysine at the 61 position in the β chain. The corresponding

	Protein molecule	Amino acid	Codons
Acceptable missense	Hb A, β chain ↓ Hb Hikari, β chain	61 Lysine ↓ Asparagine	<div style="display: flex; align-items: center;"> <div style="text-align: center;">AAA ↓ AAU</div> <div style="margin: 0 10px;">or</div> <div style="text-align: center;">AAG ↓ AAC</div> </div>
Partially acceptable missense	Hb A, β chain ↓ Hb S, β chain	6 Glutamate ↓ Valine	<div style="display: flex; align-items: center;"> <div style="text-align: center;">GAA ↓ GUA</div> <div style="margin: 0 10px;">or</div> <div style="text-align: center;">GAG ↓ GUG</div> </div>
Unacceptable missense	Hb A, α chain ↓ Hb M (Boston), α chain	58 Histidine ↓ Tyrosine	<div style="display: flex; align-items: center;"> <div style="text-align: center;">CAU ↓ UAU</div> <div style="margin: 0 10px;">or</div> <div style="text-align: center;">CAC ↓ UAC</div> </div>

Figure 38–4. Examples of three types of missense mutations resulting in abnormal hemoglobin chains. The amino acid alterations and possible alterations in the respective codons are indicated. The hemoglobin Hikari β -chain mutation has apparently normal physiologic properties but is electrophoretically altered. Hemoglobin S has a β -chain mutation and partial function; hemoglobin S binds oxygen but precipitates when deoxygenated. Hemoglobin M Boston, an α -chain mutation, permits the oxidation of the heme ferrous iron to the ferric state and so will not bind oxygen at all.

transversion might be either AAA or AAG changed to either AAU or AAC. The replacement of the specific lysine with asparagine apparently does not alter the normal function of the β chain in these individuals.

B. PARTIALLY ACCEPTABLE MISSENSE MUTATIONS

A partially acceptable missense mutation (Figure 38–4, center) is best exemplified by **hemoglobin S**, which is found in sickle cell anemia. Here glutamic acid, the normal amino acid in position 6 of the β chain, has been replaced by valine. The corresponding single nucleotide change within the codon would be GAA or GAG of glutamic acid to GUA or GUG of valine. Clearly, this missense mutation hinders normal function and results in sickle cell anemia when the mutant gene is present in the homozygous state. The glutamate-to-valine change may be considered to be partially acceptable because hemoglobin S does bind and release oxygen, although abnormally.

C. UNACCEPTABLE MISSENSE MUTATIONS

An unacceptable missense mutation (Figure 38–4, bottom) in a hemoglobin gene generates a nonfunctioning hemoglobin molecule. For example, the hemoglobin M mutations generate molecules that allow the Fe^{2+} of the heme moiety to be oxidized to Fe^{3+} , producing methemoglobin. Methemoglobin cannot transport oxygen (see Chapter 6).

Frameshift Mutations Result From Deletion or Insertion of Nucleotides in DNA That Generates Altered mRNAs

The deletion of a single nucleotide from the coding strand of a gene results in an altered reading frame in the mRNA. The machinery translating the mRNA does not recognize that a base was missing, since there is no punctuation in the reading of codons. Thus, a major alteration in the sequence of polymerized amino acids, as depicted in example 1, Figure 38–5, results. Altering the reading frame results in a garbled translation of the mRNA distal to the single nucleotide deletion. Not only is the sequence of amino acids distal to this deletion garbled, but reading of the message can also result in the appearance of a nonsense codon and thus the production of a polypeptide both garbled and prematurely terminated (example 3, Figure 38–5).

If three nucleotides or a multiple of three are deleted from a coding region, the corresponding mRNA when translated will provide a protein from which is missing the corresponding number of amino acids (example 2, Figure 38–5). Because the reading frame is a triplet, the reading phase will not be disturbed for those codons distal to the deletion. If, however, deletion of one or two nucleotides occurs just prior to or within the nor-

mal termination codon (nonsense codon), the reading of the normal termination signal is disturbed. Such a deletion might result in reading through a termination signal until another nonsense codon is encountered (example 1, Figure 38–5). Examples of this phenomenon are described in discussions of hemoglobinopathies.

Insertions of one or two or nonmultiples of three nucleotides into a gene result in an mRNA in which the reading frame is distorted upon translation, and the same effects that occur with deletions are reflected in the mRNA translation. This may result in garbled amino acid sequences distal to the insertion and the generation of a **nonsense codon** at or distal to the insertion, or perhaps reading through the normal termination codon. Following a deletion in a gene, an insertion (or vice versa) can reestablish the proper reading frame (example 4, Figure 38–5). The corresponding mRNA, when translated, would contain a garbled amino acid sequence between the insertion and deletion. Beyond the reestablishment of the reading frame, the amino acid sequence would be correct. One can imagine that different combinations of deletions, of insertions, or of both would result in formation of a protein wherein a portion is abnormal, but this portion is surrounded by the normal amino acid sequences. Such phenomena have been demonstrated convincingly in a number of diseases.

Suppressor Mutations Can Counteract Some of the Effects of Missense, Nonsense, & Frameshift Mutations

The above discussion of the altered protein products of gene mutations is based on the presence of normally functioning tRNA molecules. However, in prokaryotic and lower eukaryotic organisms, abnormally functioning tRNA molecules have been discovered that are themselves the results of mutations. Some of these abnormal tRNA molecules are capable of binding to and decoding altered codons, thereby suppressing the effects of mutations in distant structural genes. These **suppressor tRNA molecules**, usually formed as the result of alterations in their anticodon regions, are capable of suppressing missense mutations, nonsense mutations, and frameshift mutations. However, since the suppressor tRNA molecules are not capable of distinguishing between a normal codon and one resulting from a gene mutation, their presence in a cell usually results in decreased viability. For instance, the nonsense suppressor tRNA molecules can suppress the normal termination signals to allow a read-through when it is not desirable. Frameshift suppressor tRNA molecules may read a normal codon plus a component of a juxtaposed codon to provide a frameshift, also when it is not desirable. Suppressor tRNA molecules may exist in mammalian cells, since read-through transcription occurs.

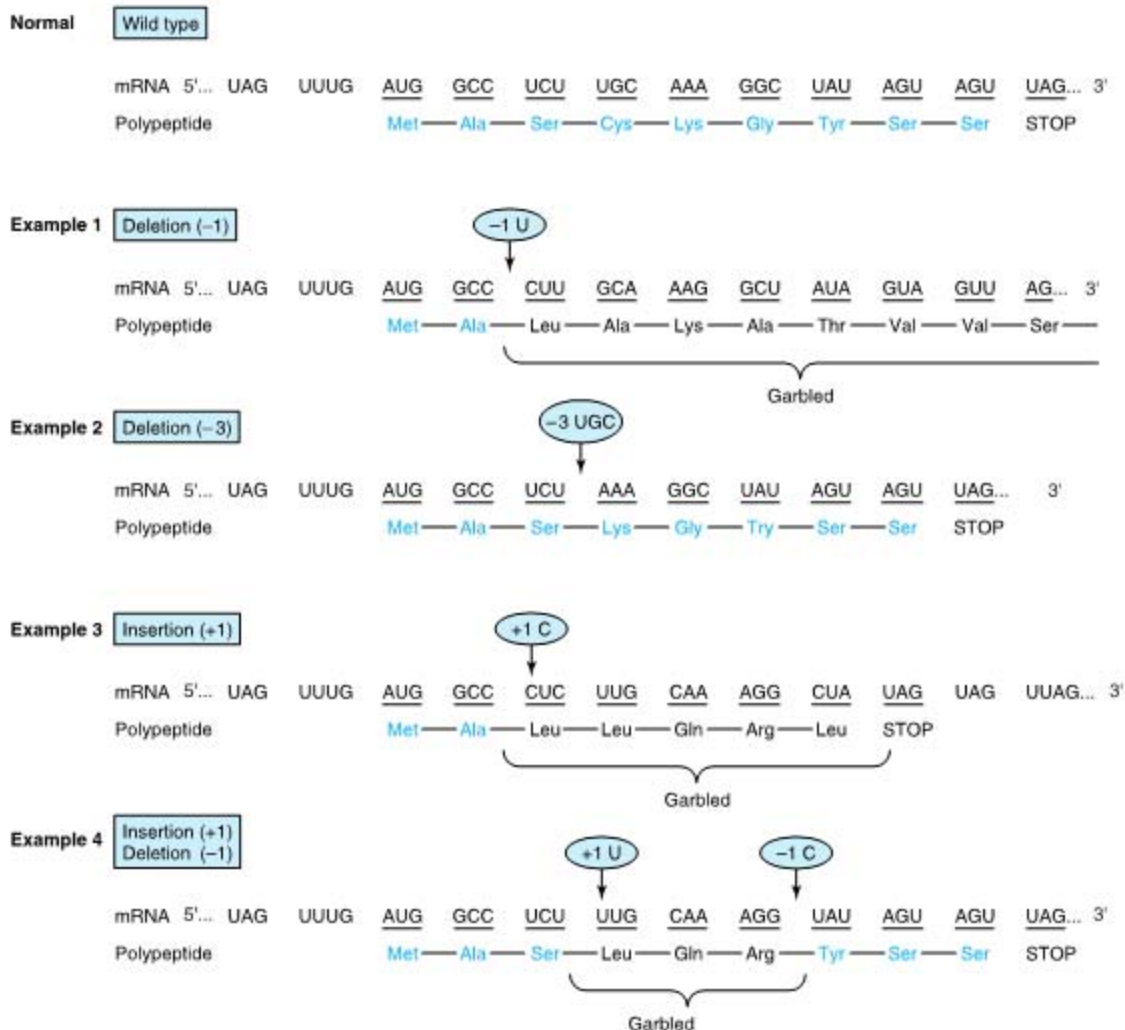


Figure 38–5. Examples of the effects of deletions and insertions in a gene on the sequence of the mRNA transcript and of the polypeptide chain translated therefrom. The arrows indicate the sites of deletions or insertions, and the numbers in the ovals indicate the number of nucleotide residues deleted or inserted. Blue type indicates amino acids in correct order.

LIKE TRANSCRIPTION, PROTEIN SYNTHESIS CAN BE DESCRIBED IN THREE PHASES: INITIATION, ELONGATION, & TERMINATION

The general structural characteristics of ribosomes and their self-assembly process are discussed in Chapter 37. These particulate entities serve as the machinery on which the mRNA nucleotide sequence is translated into the sequence of amino acids of the specified protein.

The translation of the mRNA commences near its 5' terminal with the formation of the corresponding amino terminal of the protein molecule. The message is read from 5' to 3', concluding with the formation of the carboxyl terminal of the protein. Again, the concept of **polarity** is apparent. As described in Chapter 37, the transcription of a gene into the corresponding mRNA or its precursor first forms the 5' terminal of the RNA molecule. In prokaryotes, this allows for the beginning of mRNA translation before the transcription of the gene is completed. In eukaryotic organisms, the process

of transcription is a nuclear one; mRNA translation occurs in the cytoplasm. This precludes simultaneous transcription and translation in eukaryotic organisms and makes possible the processing necessary to generate mature mRNA from the primary transcript—hnRNA.

Initiation Involves Several Protein-RNA Complexes (Figure 38–6)

Initiation of protein synthesis requires that an mRNA molecule be selected for translation by a ribosome. Once the mRNA binds to the ribosome, the latter finds the correct reading frame on the mRNA, and translation begins. This process involves tRNA, rRNA, mRNA, and at least ten eukaryotic initiation factors (eIFs), some of which have multiple (three to eight) subunits. Also involved are GTP, ATP, and amino acids. Initiation can be divided into four steps: (1) dissociation of the ribosome into its 40S and 60S subunits; (2) binding of a ternary complex consisting of met-tRNAⁱ, GTP, and eIF-2 to the 40S ribosome to form a preinitiation complex; (3) binding of mRNA to the 40S preinitiation complex to form a 43S initiation complex; and (4) combination of the 43S initiation complex with the 60S ribosomal subunit to form the 80S initiation complex.

A. RIBOSOMAL DISSOCIATION

Two initiation factors, eIF-3 and eIF-1A, bind to the newly dissociated 40S ribosomal subunit. This delays its reassociation with the 60S subunit and allows other translation initiation factors to associate with the 40S subunit.

B. FORMATION OF THE 43S PREINITIATION COMPLEX

The first step in this process involves the binding of GTP by eIF-2. This binary complex then binds to met-tRNAⁱ, a tRNA specifically involved in binding to the initiation codon AUG. (There are two tRNAs for methionine. One specifies methionine for the initiator codon, the other for internal methionines. Each has a unique nucleotide sequence.) This ternary complex binds to the 40S ribosomal subunit to form the 43S preinitiation complex, which is stabilized by association with eIF-3 and eIF-1A.

eIF-2 is one of two control points for protein synthesis initiation in eukaryotic cells. eIF-2 consists of α , β , and γ subunits. eIF-2 α is phosphorylated (on serine 51) by at least four different protein kinases (HCR, PKR, PERK, and GCN2) that are activated when a cell is under stress and when the energy expenditure required for protein synthesis would be deleterious. Such conditions include amino acid and glucose starvation, virus infection, misfolded proteins, serum deprivation, hyperosmolality, and heat shock. PKR is

particularly interesting in this regard. This kinase is activated by viruses and provides a host defense mechanism that decreases protein synthesis, thereby inhibiting viral replication. Phosphorylated eIF-2 α binds tightly to and inactivates the GTP-GDP recycling protein eIF-2B. This prevents formation of the 43S preinitiation complex and blocks protein synthesis.

C. FORMATION OF THE 43S INITIATION COMPLEX

The 5' terminals of most mRNA molecules in eukaryotic cells are "capped," as described in Chapter 37. This methyl-guanosyl triphosphate cap facilitates the binding of mRNA to the 43S preinitiation complex. A cap binding protein complex, eIF-4F (4F), which consists of eIF-4E and the eIF-4G (4G)-eIF4A (4A) complex, binds to the cap through the 4E protein. Then eIF-4A (4A) and eIF-4B (4B) bind and reduce the complex secondary structure of the 5' end of the mRNA through ATPase and ATP-dependent helicase activities. The association of mRNA with the 43S preinitiation complex to form the 48S initiation complex requires ATP hydrolysis. eIF-3 is a key protein because it binds with high affinity to the 4G component of 4F, and it links this complex to the 40S ribosomal subunit. Following association of the 43S preinitiation complex with the mRNA cap and reduction ("melting") of the secondary structure near the 5' end of the mRNA, the complex scans the mRNA for a suitable initiation codon. Generally this is the 5'-most AUG, but the precise initiation codon is determined by so-called **Kozak consensus sequences** that surround the AUG:



Most preferred is the presence of a purine at positions -3 and +4 relative to the AUG.

D. ROLE OF THE POLY(A) TAIL IN INITIATION

Biochemical and genetic experiments in yeast have revealed that the 3' poly(A) tail and its binding protein, Pab1p, are required for efficient initiation of protein synthesis. Further studies showed that the poly(A) tail stimulates recruitment of the 40S ribosomal subunit to the mRNA through a complex set of interactions. Pab1p, bound to the poly(A) tail, interacts with eIF-4G, which in turn binds to eIF-4E that is bound to the cap structure. It is possible that a circular structure is formed and that this helps direct the 40S ribosomal subunit to the 5' end of the mRNA. This helps explain how the cap and poly(A) tail structures have a synergistic effect on protein synthesis. It appears that a similar mechanism is at work in mammalian cells.

The diagram illustrates the three-step process of translation initiation:

- Ternary complex formation:** Met-tRNA^{Met} (carrying a methionine amino acid) binds to eIF-2C (a subunit of eIF-2) and GTP. This step involves the release of GDP and the formation of a ternary complex consisting of Met-tRNA^{Met}, eIF-2C, and GTP.
- Formation of the 80S initiation complex:** The 80S ribosome dissociates into a 60S subunit and a 40S subunit. The 40S subunit binds to the ternary complex, forming a 43S preinitiation complex. This complex then binds to the 5' cap (4F) of the mRNA, forming a 48S initiation complex. The 48S complex then binds to the 60S subunit, forming the 80S initiation complex. The 80S initiation complex is then activated by eIF-5, which hydrolyzes GTP to GDP and P_i, leading to the release of eIF-2C and the formation of the 80S initiation complex. The 80S initiation complex is then ready for elongation.
- Activation of mRNA:** The 48S initiation complex binds to the 5' cap (4F) of the mRNA, forming a 48S initiation complex. This complex then binds to the 60S subunit, forming the 80S initiation complex. The 80S initiation complex is then activated by eIF-5, which hydrolyzes GTP to GDP and P_i, leading to the release of eIF-2C and the formation of the 80S initiation complex. The 80S initiation complex is then ready for elongation.

E. FORMATION OF THE 80S INITIATION COMPLEX

The binding of the 60S ribosomal subunit to the 48S initiation complex involves hydrolysis of the GTP bound to eIF-2 by eIF-5. This reaction results in release of the initiation factors bound to the 48S initiation complex (these factors then are recycled) and the rapid association of the 40S and 60S subunits to form the 80S ribosome. At this point, the met-tRNAⁱ is on the P site of the ribosome, ready for the elongation cycle to commence.

The Regulation of eIF-4E Controls the Rate of Initiation

The 4F complex is particularly important in controlling the rate of protein translation. As described above, 4F is a complex consisting of 4E, which binds to the m⁷G cap structure at the 5' end of the mRNA, and 4G, which serves as a scaffolding protein. In addition to binding 4E, 4G binds to eIF-3, which links the complex to the 40S ribosomal subunit. It also binds 4A and 4B, the ATPase-helicase complex that helps unwind the RNA (Figure 38-7).

4E is responsible for recognition of the mRNA cap structure, which is a rate-limiting step in translation. This process is regulated at two levels. Insulin and mitogenic growth factors result in the phosphorylation of 4E on ser 209 (or thr 210). Phosphorylated 4E binds to the cap much more avidly than does the nonphosphorylated form, thus enhancing the rate of initiation. A component of the MAP kinase pathway (see Figure 43-8) appears to be involved in this phosphorylation reaction.

The activity of 4E is regulated in a second way, and this also involves phosphorylation. A recently discovered set of proteins bind to and inactivate 4E. These proteins include 4E-BP1 (BP1, also known as PHAS-1) and the closely related proteins 4E-BP2 and 4E-BP3. BP1 binds with high affinity to 4E. The [4E]•[BP1] association prevents 4E from binding to 4G (to form 4F). Since this interaction is essential for the binding of 4F to the ribosomal 40S subunit and for correctly positioning this on the capped mRNA, BP-1 effectively inhibits translation initiation.

Insulin and other growth factors result in the phosphorylation of BP-1 at five unique sites. Phosphorylation of BP-1 results in its dissociation from 4E, and it cannot rebind until critical sites are dephosphorylated. The protein kinase responsible has not been identified, but it appears to be different from the one that phosphorylates 4E. A kinase in the mammalian target of rapamycin (mTOR) pathway, perhaps mTOR itself, is involved. These effects on the activation of 4E explain in part how insulin causes a marked posttranscriptional

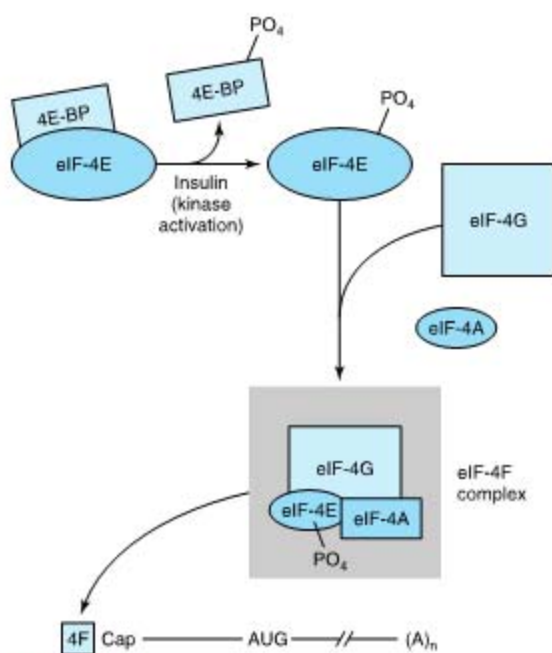
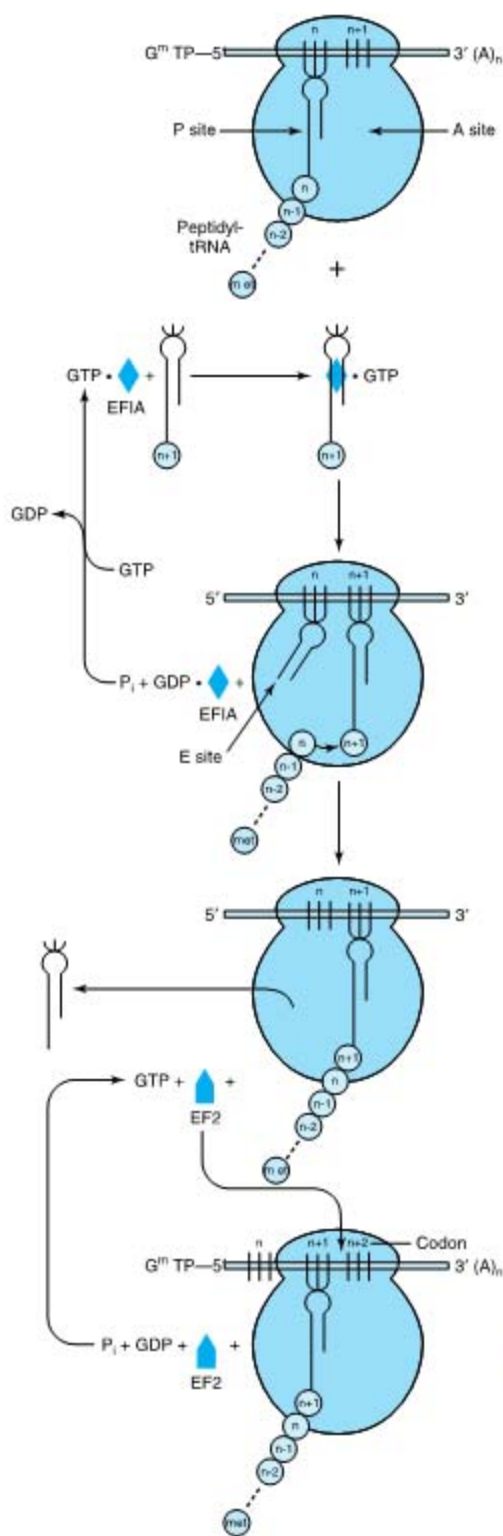


Figure 38-7. Activation of eIF-4E by insulin and formation of the cap binding eIF-4F complex. The 4F-cap mRNA complex is depicted as in Figure 38-6. The 4F complex consists of eIF-4E (4E), eIF-4A, and eIF-4G. 4E is inactive when bound by one of a family of binding proteins (4E-BPs). Insulin and mitogenic factors (eg, IGF-1, PDGF, interleukin-2, and angiotensin II) activate a serine protein kinase in the mTOR pathway, and this results in the phosphorylation of 4E-BP. Phosphorylated 4E-BP dissociates from 4E, and the latter is then able to form the 4F complex and bind to the mRNA cap. These growth peptides also phosphorylate 4E itself by activating a component of the MAP kinase pathway. Phosphorylated 4E binds much more avidly to the cap than does nonphosphorylated 4E.

increase of protein synthesis in liver, adipose tissue, and muscle.

Elongation Also Is a Multistep Process (Figure 38-8)

Elongation is a cyclic process on the ribosome in which one amino acid at a time is added to the nascent peptide chain. The peptide sequence is determined by the order of the codons in the mRNA. Elongation involves several steps catalyzed by proteins called elongation factors (EFs). These steps are (1) binding of aminoacyl-tRNA to the A site, (2) peptide bond formation, and (3) translocation.



A. BINDING OF AMINOACYL-tRNA TO THE A SITE

In the complete 80S ribosome formed during the process of initiation, the A site (aminoacyl or acceptor site) is free. The binding of the proper aminoacyl-tRNA in the A site requires proper codon recognition. **Elongation factor EF1A** forms a ternary complex with GTP and the entering aminoacyl-tRNA (Figure 38-8). This complex then allows the aminoacyl-tRNA to enter the A site with the release of EF1A•GDP and phosphate. GTP hydrolysis is catalyzed by an active site on the ribosome. As shown in Figure 38-8, EF1A-GDP then recycles to EF1A-GTP with the aid of other soluble protein factors and GTP.

B. PEPTIDE BOND FORMATION

The α -amino group of the new aminoacyl-tRNA in the A site carries out a nucleophilic attack on the esterified carboxyl group of the peptidyl-tRNA occupying the P site (peptidyl or polypeptide site). At initiation, this site is occupied by aminoacyl-tRNA met¹. This reaction is catalyzed by a **peptidyltransferase**, a component of the 23S RNA of the 60S ribosomal subunit. This is another example of ribozyme activity and indicates an important—and previously unsuspected—direct role for RNA in protein synthesis (Table 38-3). Because the amino acid on the aminoacyl-tRNA is already “activated,” no further energy source is required for this reaction. The reaction results in attachment of the growing peptide chain to the tRNA in the A site.

C. TRANSLOCATION

The now deacylated tRNA is attached by its anticodon to the P site at one end and by the open CCA tail to an **exit (E) site** on the large ribosomal subunit (Figure 38-8). At this point, **elongation factor 2 (EF2)** binds to and displaces the peptidyl tRNA from the A site to the P site. In turn, the deacylated tRNA is on the E site, from which it leaves the ribosome. The EF2-GTP complex is hydrolyzed to EF2-GDP, effectively moving the mRNA forward by one codon and leaving the A site open for occupancy by another ternary complex of amino acid tRNA-EF1A-GTP and another cycle of elongation.

Figure 38-8. Diagrammatic representation of the peptide elongation process of protein synthesis. The small circles labeled $n-1$, n , $n+1$, etc., represent the amino acid residues of the newly formed protein molecule. EF1A and EF2 represent elongation factors 1 and 2, respectively. The peptidyl-tRNA and aminoacyl-tRNA sites on the ribosome are represented by P site and A site, respectively.

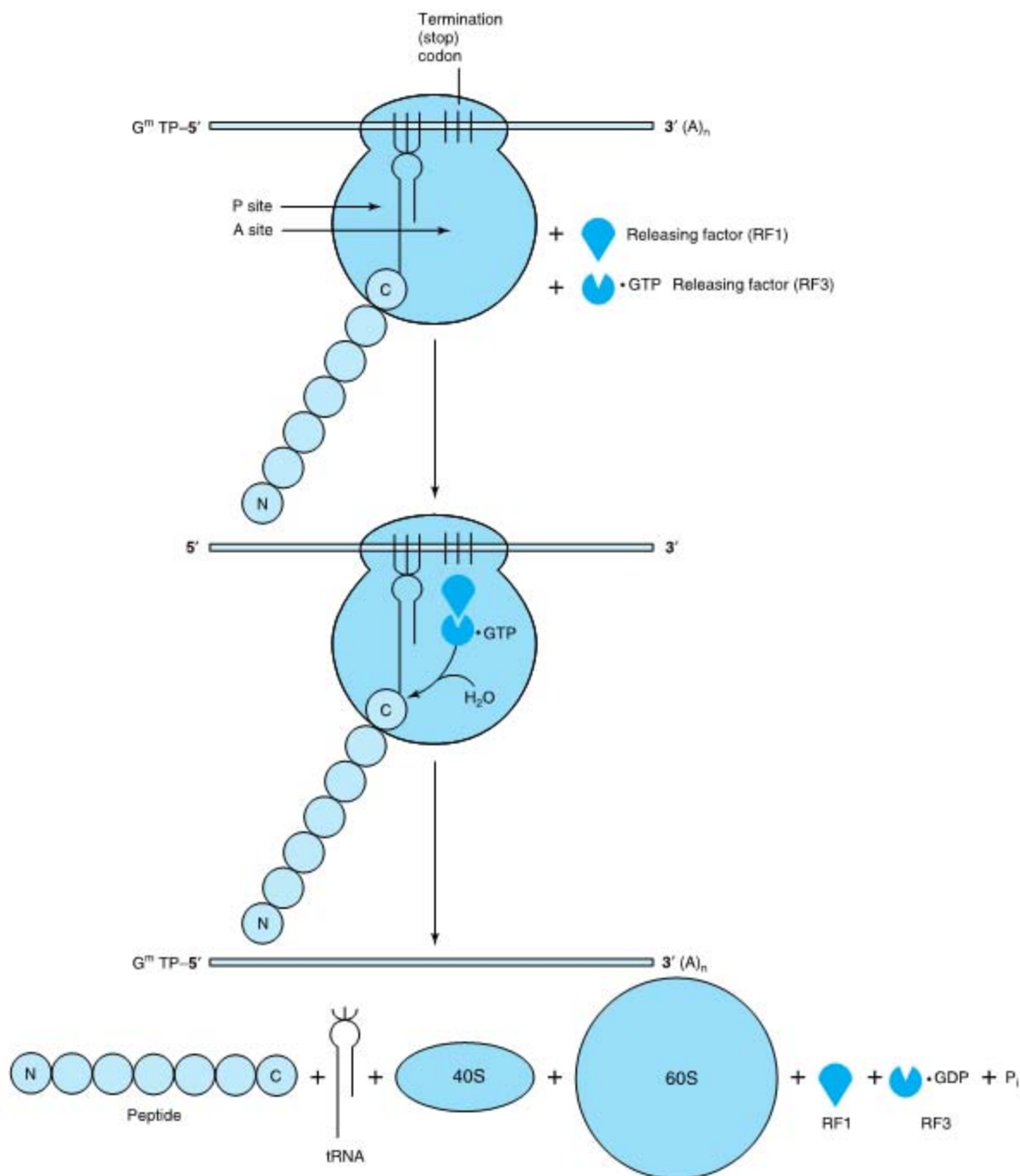


Figure 38–9. Diagrammatic representation of the termination process of protein synthesis. The peptidyl-tRNA and aminoacyl-tRNA sites are indicated as P site and A site, respectively. The termination (stop) codon is indicated by the three vertical bars. Releasing factor RF1 binds to the stop codon. Releasing factor RF3, with bound GTP, binds to RF1. Hydrolysis of the peptidyl-tRNA complex is shown by the entry of H₂O. N and C indicate the amino and carboxyl terminal amino acids, respectively, and illustrate the polarity of protein synthesis.

Table 38–3. Evidence that rRNA is peptidyltransferase.

- Ribosomes can make peptide bonds even when proteins are removed or inactivated.
- Certain parts of the rRNA sequence are highly conserved in all species.
- These conserved regions are on the surface of the RNA molecule.
- RNA can be catalytic.
- Mutations that result in antibiotic resistance at the level of protein synthesis are more often found in rRNA than in the protein components of the ribosome.

The charging of the tRNA molecule with the aminoacyl moiety requires the hydrolysis of an ATP to an AMP, equivalent to the hydrolysis of two ATPs to two ADPs and phosphates. The entry of the aminoacyl-tRNA into the A site results in the hydrolysis of one GTP to GDP. Translocation of the newly formed peptidyl-tRNA in the A site into the P site by EF2 similarly results in hydrolysis of GTP to GDP and phosphate. Thus, the energy requirements for the formation of one peptide bond include the equivalent of the hydrolysis of two ATP molecules to ADP and of two GTP molecules to GDP, or the hydrolysis of four high-energy phosphate bonds. A eukaryotic ribosome can incorporate as many as six amino acids per second; prokaryotic ribosomes incorporate as many as 18 per second. Thus, the process of peptide synthesis occurs with great speed and accuracy until a termination codon is reached.

Termination Occurs When a Stop Codon Is Recognized (Figure 38–9)

In comparison to initiation and elongation, termination is a relatively simple process. After multiple cycles of elongation culminating in polymerization of the specific amino acids into a protein molecule, the stop or terminating codon of mRNA (UAA, UAG, UGA) appears in the A site. Normally, there is no tRNA with an anticodon capable of recognizing such a termination signal. **Releasing factor RF1** recognizes that a stop codon resides in the A site (Figure 38–9). RF1 is bound by a complex consisting of **releasing factor RF3** with bound GTP. This complex, with the peptidyl transferase, promotes hydrolysis of the bond between the peptide and the tRNA occupying the P site. Thus, a water molecule rather than an amino acid is added. This hydrolysis releases the protein and the tRNA from the P site. Upon hydrolysis and release, the **80S ribosome dissociates** into its 40S and 60S subunits, which are then recycled. Therefore, the releasing factors are

proteins that hydrolyze the peptidyl-tRNA bond when a stop codon occupies the A site. The mRNA is then released from the ribosome, which dissociates into its component 40S and 60S subunits, and another cycle can be repeated.

Polysomes Are Assemblies of Ribosomes

Many ribosomes can translate the same mRNA molecule simultaneously. Because of their relatively large size, the ribosome particles cannot attach to an mRNA any closer than 35 nucleotides apart. Multiple ribosomes on the same mRNA molecule form a **polysome**, or “polysome.” In an unrestricted system, the number of ribosomes attached to an mRNA (and thus the size of polyribosomes) correlates positively with the length of the mRNA molecule. The mass of the mRNA molecule is, of course, quite small compared with the mass of even a single ribosome.

A single mammalian ribosome is capable of synthesizing about 400 peptide bonds each minute. Polyribosomes actively synthesizing proteins can exist as free particles in the cellular cytoplasm or may be attached to sheets of membranous cytoplasmic material referred to as **endoplasmic reticulum**. Attachment of the particulate polyribosomes to the endoplasmic reticulum is responsible for its “rough” appearance as seen by electron microscopy. The proteins synthesized by the attached polyribosomes are extruded into the cisternal space between the sheets of rough endoplasmic reticulum and are exported from there. Some of the protein products of the rough endoplasmic reticulum are packaged by the Golgi apparatus into zymogen particles for eventual export (see Chapter 46). The polyribosomal particles free in the cytosol are responsible for the synthesis of proteins required for intracellular functions.

The Machinery of Protein Synthesis Can Respond to Environmental Threats

Ferritin, an iron-binding protein, prevents ionized iron (Fe^{2+}) from reaching toxic levels within cells. Elemental iron stimulates ferritin synthesis by causing the release of a cytoplasmic protein that binds to a specific region in the 5' nontranslated region of ferritin mRNA. Disruption of this protein-mRNA interaction activates ferritin mRNA and results in its translation. This mechanism provides for rapid control of the synthesis of a protein that sequesters Fe^{2+} , a potentially toxic molecule.

Many Viruses Co-opt the Host Cell Protein Synthesis Machinery

The protein synthesis machinery can also be modified in deleterious ways. **Viruses replicate by using host**

cell processes, including those involved in protein synthesis. Some viral mRNAs are translated much more efficiently than those of the host cell (eg, encephalomyocarditis virus). Others, such as reovirus and vesicular stomatitis virus, replicate abundantly, and their mRNAs have a competitive advantage over host cell mRNAs for limited translation factors. Other viruses inhibit host cell protein synthesis by preventing the association of mRNA with the 40S ribosome.

Poliovirus and other picornaviruses gain a selective advantage by disrupting the function of the 4F complex to their advantage. The mRNAs of these viruses do not have a cap structure to direct the binding of the 40S ribosomal subunit (see above). Instead, the 40S ribosomal subunit contacts an **internal ribosomal entry site (IRES)** in a reaction that requires 4G but not 4E. The virus gains a selective advantage by having a protease that attacks 4G and removes the amino terminal 4E binding site. Now the 4E-4G complex (4F) cannot form, so the 40S ribosomal subunit cannot be directed to capped mRNAs. Host cell translation is thus abolished. The 4G fragment can direct binding of the 40S ribosomal subunit to IRES-containing mRNAs, so viral mRNA translation is very efficient (Figure 38–10). These viruses also promote the dephosphorylation of BP1 (PHAS-1), thereby decreasing cap (4E)-dependent translation.

POSTTRANSLATIONAL PROCESSING AFFECTS THE ACTIVITY OF MANY PROTEINS

Some animal viruses, notably poliovirus and hepatitis A virus, synthesize long polycistronic proteins from one long mRNA molecule. These protein molecules are subsequently cleaved at specific sites to provide the several specific proteins required for viral function. In animal cells, many proteins are synthesized from the mRNA template as a precursor molecule, which then must be modified to achieve the active protein. The prototype is insulin, which is a low-molecular-weight protein having two polypeptide chains with interchain and intrachain disulfide bridges. The molecule is synthesized as a single chain precursor, or **prohormone**, which folds to allow the disulfide bridges to form. A specific protease then clips out the segment that connects the two chains which form the functional insulin molecule (see Figure 42–12).

Many other peptides are synthesized as preproteins that require modifications before attaining biologic activity. Many of the posttranslational modifications involve the removal of amino terminal amino acid residues by specific aminopeptidases. Collagen, an abundant protein in the extracellular spaces of higher eukaryotes, is synthesized as procollagen. Three procoll-

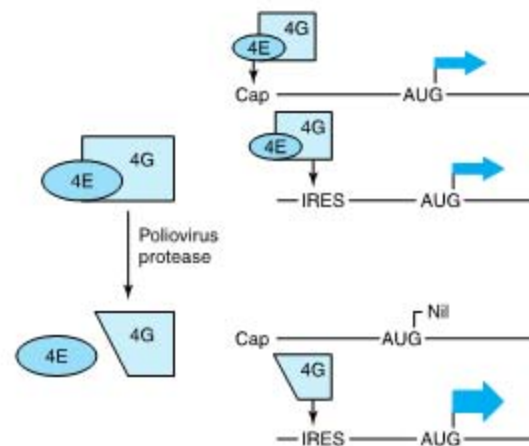


Figure 38–10. Picornaviruses disrupt the 4F complex. The 4E-4G complex (4F) directs the 40S ribosomal subunit to the typical capped mRNA (see text). 4G alone is sufficient for targeting the 40S subunit to the internal ribosomal entry site (IRES) of viral mRNAs. To gain selective advantage, certain viruses (eg, poliovirus) have a protease that cleaves the 4E binding site from the amino terminal end of 4G. This truncated 4G can direct the 40S ribosomal subunit to mRNAs that have an IRES but not to those that have a cap. The widths of the arrows indicate the rate of translation initiation from the AUG codon in each example.

lagen polypeptide molecules, frequently not identical in sequence, align themselves in a particular way that is dependent upon the existence of specific amino terminal peptides. Specific enzymes then carry out hydroxylations and oxidations of specific amino acid residues within the procollagen molecules to provide cross-links for greater stability. Amino terminal peptides are cleaved off the molecule to form the final product—a strong, insoluble collagen molecule. Many other posttranslational modifications of proteins occur. Covalent modification by acetylation, phosphorylation, methylation, ubiquitinylation, and glycosylation is common, for example.

MANY ANTIBIOTICS WORK BECAUSE THEY SELECTIVELY INHIBIT PROTEIN SYNTHESIS IN BACTERIA

Ribosomes in bacteria and in the mitochondria of higher eukaryotic cells differ from the mammalian ribosome described in Chapter 35. The bacterial ribosome is smaller (70S rather than 80S) and has a different, somewhat simpler complement of RNA and protein

molecules. This difference is exploited for clinical purposes because many effective antibiotics interact specifically with the proteins and RNAs of prokaryotic ribosomes and thus inhibit protein synthesis. This results in growth arrest or death of the bacterium. The most useful members of this class of antibiotics (eg, tetracyclines, lincomycin, erythromycin, and chloramphenicol) do not interact with components of eukaryotic ribosomal particles and thus are not toxic to eukaryotes. Tetracycline prevents the binding of aminoacyl-tRNAs to the A site. Chloramphenicol and the macrolide class of antibiotics work by binding to 23S rRNA, which is interesting in view of the newly appreciated role of rRNA in peptide bond formation through its peptidyltransferase activity. It should be mentioned that the close similarity between prokaryotic and mitochondrial ribosomes can lead to complications in the use of some antibiotics.

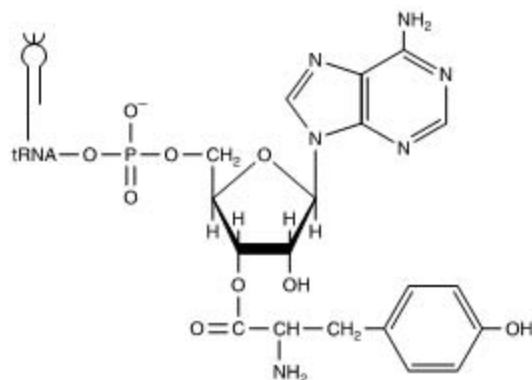
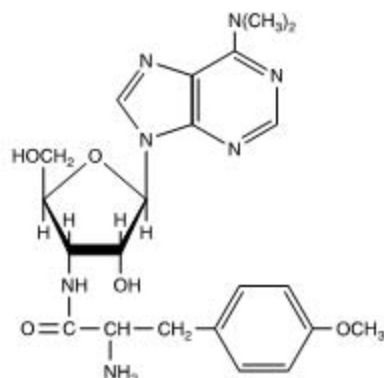


Figure 38–11. The comparative structures of the antibiotic puromycin (top) and the 3' terminal portion of tyrosyl-tRNA (bottom).

Other antibiotics inhibit protein synthesis on all ribosomes (**puromycin**) or only on those of eukaryotic cells (**cycloheximide**). Puromycin (Figure 38–11) is a structural analog of tyrosyl-tRNA. Puromycin is incorporated via the A site on the ribosome into the carboxyl terminal position of a peptide but causes the premature release of the polypeptide. Puromycin, as a tyrosyl-tRNA analog, effectively inhibits protein synthesis in both prokaryotes and eukaryotes. Cycloheximide inhibits peptidyltransferase in the 60S ribosomal subunit in eukaryotes, presumably by binding to an rRNA component.

Diphtheria toxin, an exotoxin of *Corynebacterium diphtheriae* infected with a specific lysogenic phage, catalyzes the ADP-ribosylation of EF-2 on the unique amino acid diphthamide in mammalian cells. This modification inactivates EF-2 and thereby specifically inhibits mammalian protein synthesis. Many animals (eg, mice) are resistant to diphtheria toxin. This resistance is due to inability of diphtheria toxin to cross the cell membrane rather than to insensitivity of mouse EF-2 to diphtheria toxin-catalyzed ADP-ribosylation by NAD.

Ricin, an extremely toxic molecule isolated from the castor bean, inactivates eukaryotic 28S ribosomal RNA by providing the N-glycolytic cleavage or removal of a single adenine.

Many of these compounds—puromycin and cycloheximide in particular—are not clinically useful but have been important in elucidating the role of protein synthesis in the regulation of metabolic processes, particularly enzyme induction by hormones.

SUMMARY

- The flow of genetic information follows the sequence DNA → RNA → protein.
- The genetic information in the structural region of a gene is transcribed into an RNA molecule such that the sequence of the latter is complementary to that in the DNA.
- Several different types of RNA, including ribosomal RNA (rRNA), transfer RNA (tRNA), and messenger RNA (mRNA), are involved in protein synthesis.
- The information in mRNA is in a tandem array of codons, each of which is three nucleotides long.
- The mRNA is read continuously from a start codon (AUG) to a termination codon (UAA, UAG, UGA).
- The open reading frame of the mRNA is the series of codons, each specifying a certain amino acid, that determines the precise amino acid sequence of the protein.
- Protein synthesis, like DNA and RNA synthesis, follows a 5' to 3' polarity and can be divided into three

processes: initiation, elongation, and termination. Mutant proteins arise when single-base substitutions result in codons that specify a different amino acid at a given position, when a stop codon results in a truncated protein, or when base additions or deletions alter the reading frame, so different codons are read.

- A variety of compounds, including several antibiotics, inhibit protein synthesis by affecting one or more of the steps involved in protein synthesis.

REFERENCES

- Crick F et al: The genetic code. *Nature* 1961;192:1227.
- Green R, Noller HF: Ribosomes and translation. *Annu Rev Biochem* 1997;66:679.
- Kozak M: Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J Biol Chem* 1991;266:1986.
- Lawrence JC, Abraham RT: PHAS/4E-BPs as regulators of mRNA translation and cell proliferation. *Trends Biochem Sci* 1997;22:345.
- Sachs AB, Buratowski S: Common themes in translational and transcriptional regulation. *Trends Biochem Sci* 1997;22:189.
- Sachs AB, Sarnow P, Hentze MW: Starting at the beginning, middle and end: translation initiation in eukaryotes. *Cell* 1997; 98:831.
- Weatherall DJ et al: The hemoglobinopathies. In: *The Metabolic and Molecular Bases of Inherited Disease*, 8th ed. Scriver CR et al (editors). McGraw-Hill, 2001.

Regulation of Gene Expression

39

Daryl K. Granner, MD, & P. Anthony Weil, PhD

BIOMEDICAL IMPORTANCE

Organisms adapt to environmental changes by altering gene expression. The process of alteration of gene expression has been studied in detail and often involves modulation of gene transcription. Control of transcription ultimately results from changes in the interaction of specific binding regulatory proteins with various regions of DNA in the controlled gene. This can have a positive or negative effect on transcription. Transcription control can result in tissue-specific gene expression, and gene regulation is influenced by hormones, heavy metals, and chemicals. In addition to transcription level controls, gene expression can also be modulated by gene amplification, gene rearrangement, post-transcriptional modifications, and RNA stabilization. Many of the mechanisms that control gene expression are used to respond to hormones and therapeutic agents. Thus, a molecular understanding of these processes will lead to development of agents that alter pathophysiologic mechanisms or inhibit the function or arrest the growth of pathogenic organisms.

REGULATED EXPRESSION OF GENES IS REQUIRED FOR DEVELOPMENT, DIFFERENTIATION, & ADAPTATION

The genetic information present in each somatic cell of a metazoan organism is practically identical. The exceptions are found in those few cells that have amplified or rearranged genes in order to perform specialized cellular functions. Expression of the genetic information must be regulated during ontogeny and differentiation of the organism and its cellular components. Furthermore, in order for the organism to adapt to its environment and to conserve energy and nutrients, the expression of genetic information must be cued to extrinsic signals and respond only when necessary. As organisms have evolved, more sophisticated regulatory mechanisms have appeared which provide the organism and its cells with the responsiveness necessary for survival in a complex environment. Mammalian cells possess about 1000 times more genetic information than does the bacterium *Escherichia coli*. Much of this additional genetic information is probably involved in regulation of gene expression during the differentiation of tissues and biologic processes in the multicellular organism and in en-

suring that the organism can respond to complex environmental challenges.

In simple terms, there are only two types of gene regulation: **positive regulation** and **negative regulation** (Table 39-1). When the expression of genetic information is quantitatively increased by the presence of a specific regulatory element, regulation is said to be positive; when the expression of genetic information is diminished by the presence of a specific regulatory element, regulation is said to be negative. The element or molecule mediating negative regulation is said to be a negative regulator or **repressor**; that mediating positive regulation is a positive regulator or **activator**. However, a **double negative** has the effect of acting as a positive. Thus, an effector that inhibits the function of a negative regulator will appear to bring about a positive regulation. Many regulated systems that appear to be induced are in fact **derepressed** at the molecular level. (See Chapter 9 for explanation of these terms.)

BIOLOGIC SYSTEMS EXHIBIT THREE TYPES OF TEMPORAL RESPONSES TO A REGULATORY SIGNAL

Figure 39-1 depicts the extent or amount of gene expression in three types of temporal response to an inducing signal. A **type A response** is characterized by an increased extent of gene expression that is dependent upon the continued presence of the inducing signal. When the inducing signal is removed, the amount of gene expression diminishes to its basal level, but the amount repeatedly increases in response to the reappearance of the specific signal. This type of response is commonly observed in prokaryotes in response to sudden changes of the intracellular concentration of a nutrient. It is also observed in many higher organisms after exposure to inducers such as hormones, nutrients, or growth factors (Chapter 43).

A **type B response** exhibits an increased amount of gene expression that is transient even in the continued presence of the regulatory signal. After the regulatory signal has terminated and the cell has been allowed to recover, a second transient response to a subsequent regulatory signal may be observed. This phenomenon of response-desensitization-recovery characterizes the action of many pharmacologic agents, but it is also a

Table 39–1. Effects of positive and negative regulation on gene expression.

	Rate of Gene Expression	
	Negative Regulation	Positive Regulation
Regulator present	Decreased	Increased
Regulator absent	Increased	Decreased

feature of many naturally occurring processes. This type of response commonly occurs during development of an organism, when only the transient appearance of a specific gene product is required although the signal persists.

The **type C response** pattern exhibits, in response to the regulatory signal, an increased extent of gene expression that persists indefinitely even after termination of the signal. The signal acts as a trigger in this pattern. Once expression of the gene is initiated in the cell, it cannot be terminated even in the daughter cells; it is therefore an irreversible and inherited alteration. This type of response typically occurs during the development of differentiated function in a tissue or organ.

Prokaryotes Provide Models for the Study of Gene Expression in Mammalian Cells

Analysis of the regulation of gene expression in prokaryotic cells helped establish the principle that information flows from the gene to a messenger RNA to a specific protein molecule. These studies were aided by the advanced genetic analyses that could be performed in prokaryotic and lower eukaryotic organisms. In recent years, the principles established in these early studies, coupled with a variety of molecular biology techniques, have led to remarkable progress in the analysis of gene regulation in higher eukaryotic organisms, including mammals. In this chapter, the initial discussion will center on prokaryotic systems. The impressive genetic studies will not be described, but the physiology of gene expression will be discussed. However, nearly all of the conclusions about this physiology have been derived from genetic studies and confirmed by molecular genetic and biochemical studies.

Some Features of Prokaryotic Gene Expression Are Unique

Before the physiology of gene expression can be explained, a few specialized genetic and regulatory terms must be defined for prokaryotic systems. In prokaryotes, the genes involved in a metabolic pathway are often present in a linear array called an **operon**, eg, the *lac* operon. An operon can be regulated by a single promoter or regulatory region. The **cistron** is the smallest unit of genetic expression. As described in Chapter 9, some enzymes and other protein molecules are composed of two or more nonidentical subunits. Thus, the “one gene, one enzyme” concept is not necessarily valid. The cistron is the genetic unit coding for the structure of the subunit of a protein molecule, acting as it does as the smallest unit of genetic expression. Thus, the one gene, one enzyme idea might more accurately

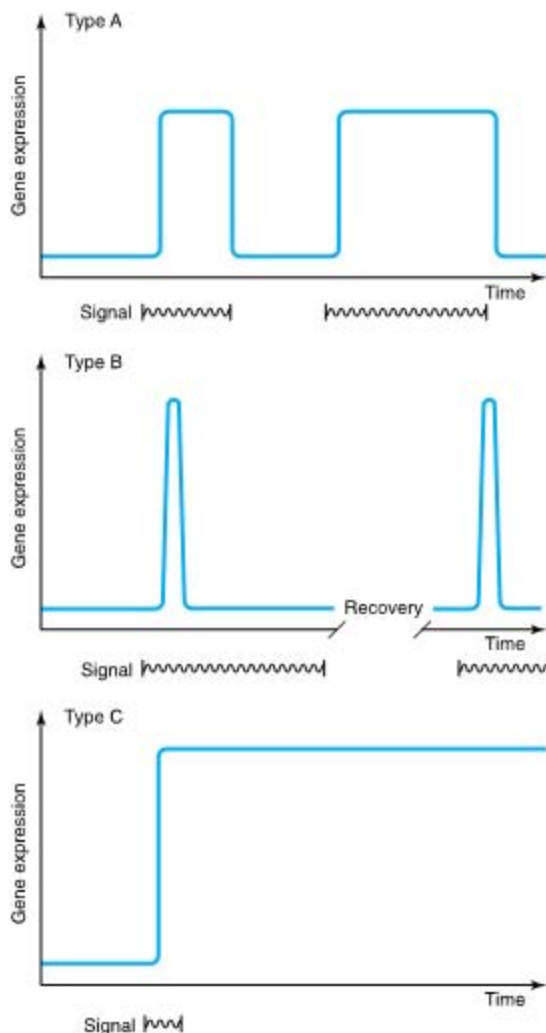


Figure 39–1. Diagrammatic representations of the responses of the extent of expression of a gene to specific regulatory signals such as a hormone.

be regarded as a **one cistron, one subunit** concept. A single mRNA that encodes more than one separately translated protein is referred to as a **polycistronic mRNA**. For example, the polycistronic *lac* operon mRNA is translated into three separate proteins (see below). Operons and polycistronic mRNAs are common in bacteria but not in eukaryotes.

An **inducible gene** is one whose expression increases in response to an **inducer** or **activator**, a specific positive regulatory signal. In general, inducible genes have relatively low basal rates of transcription. By contrast, genes with high basal rates of transcription are often subject to down-regulation by repressors.

The expression of some genes is **constitutive**, meaning that they are expressed at a reasonably constant rate and not known to be subject to regulation. These are often referred to as **housekeeping genes**. As a result of mutation, some inducible gene products become constitutively expressed. A mutation resulting in constitutive expression of what was formerly a regulated gene is called a **constitutive mutation**.

Analysis of Lactose Metabolism in *E. coli* Led to the Operon Hypothesis

Jacob and Monod in 1961 described their **operon model** in a classic paper. Their hypothesis was to a large extent based on observations on the regulation of lactose metabolism by the intestinal bacterium *E. coli*. The molecular mechanisms responsible for the regulation of the genes involved in the metabolism of lactose are now among the best-understood in any organism. β -Galactosidase hydrolyzes the β -galactoside lactose to galactose and glucose. The structural gene for β -galactosidase (*lacZ*) is clustered with the genes responsible for the permeation of galactose into the cell (*lacY*) and for thiogalactoside transacetylase (*lacA*). The structural genes for these three enzymes, along with the *lac* promoter and *lac* operator (a regulatory region), are physically associated to constitute the ***lac* operon** as depicted in Figure 39-2. This genetic arrangement of the structural genes and their regulatory genes allows for **coordinate expression** of the three enzymes concerned with lactose metabolism. Each of these linked genes is transcribed into one large mRNA molecule that contains multiple independent translation start (AUG) and stop (UAA) codons for each cistron. Thus, each protein is translated separately, and they are not processed from a single large precursor protein. This type of mRNA molecule is called a **polycistronic mRNA**. Polycistronic mRNAs are predominantly found in prokaryotic organisms.

It is now conventional to consider that a gene includes regulatory sequences as well as the region that

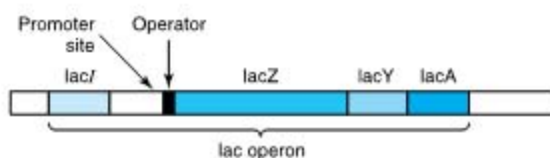


Figure 39-2. The positional relationships of the structural and regulatory genes of the *lac* operon. *lacZ* encodes β -galactosidase, *lacY* encodes a permease, and *lacA* encodes a thiogalactoside transacetylase. *lacI* encodes the *lac* operon repressor protein.

encodes the primary transcript. Although there are many historical exceptions, a gene is generally italicized in lower case and the encoded protein, when abbreviated, is expressed in roman type with the first letter capitalized. For example, the gene *lacI* encodes the repressor protein LacI. When *E. coli* is presented with lactose or some specific lactose analogs under appropriate non-repressing conditions (eg, high concentrations of lactose, no or very low glucose in media; see below), the expression of the activities of β -galactosidase, galactoside permease, and thiogalactoside transacetylase is increased 100-fold to 1000-fold. This is a type A response, as depicted in Figure 39-1. The kinetics of induction can be quite rapid; *lac*-specific mRNAs are fully induced within 5–6 minutes after addition of lactose to a culture; β -galactosidase protein is maximal within 10 minutes. Under fully induced conditions, there can be up to 5000 β -galactosidase molecules per cell, an amount about 1000 times greater than the basal, uninduced level. Upon removal of the signal, ie, the inducer, the synthesis of these three enzymes declines.

When *E. coli* is exposed to both lactose and glucose as sources of carbon, the organisms first metabolize the glucose and then temporarily stop growing until the genes of the *lac* operon become induced to provide the ability to metabolize lactose as a usable energy source. Although lactose is present from the beginning of the bacterial growth phase, the cell does not induce those enzymes necessary for catabolism of lactose until the glucose has been exhausted. This phenomenon was first thought to be attributable to repression of the *lac* operon by some catabolite of glucose; hence, it was termed catabolite repression. It is now known that catabolite repression is in fact mediated by a **catabolite gene activator protein (CAP)** in conjunction with cAMP (Figure 18-5). This protein is also referred to as the cAMP regulatory protein (CRP). The expression of many inducible enzyme systems or operons in *E. coli* and other prokaryotes is sensitive to catabolite repression, as discussed below.

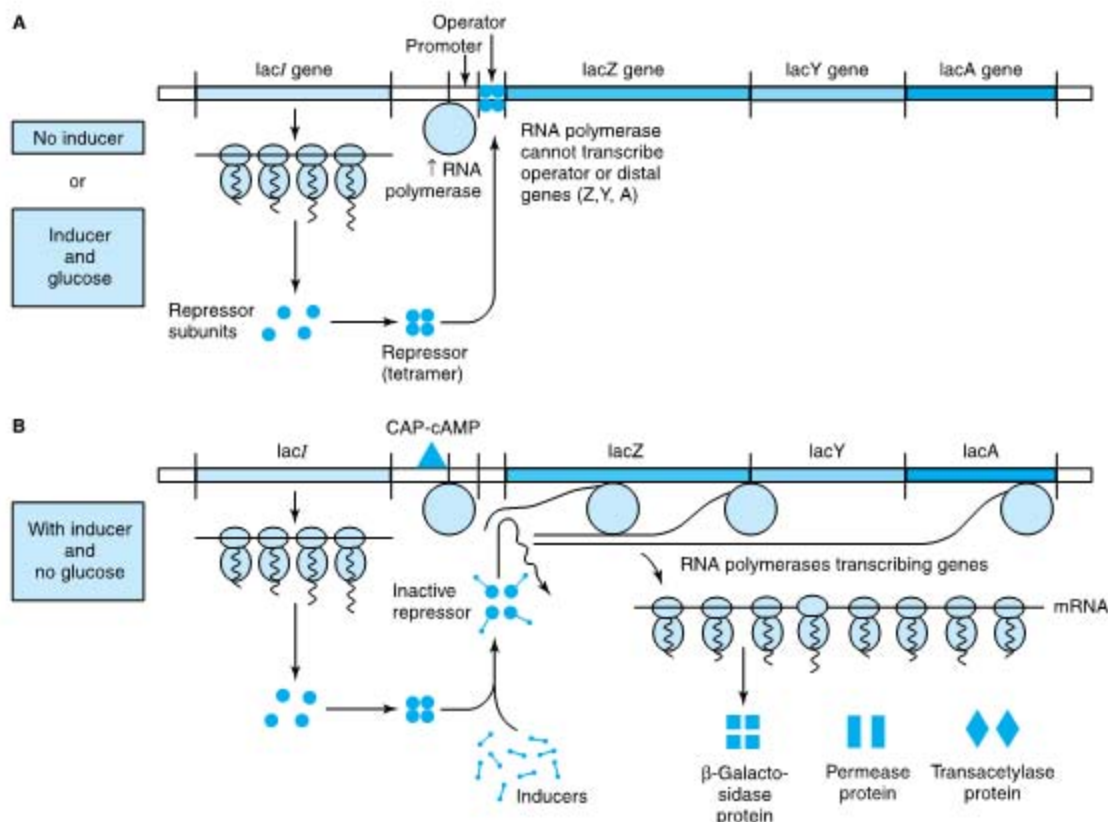
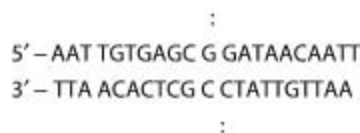


Figure 39-3. The mechanism of repression and derepression of the *lac* operon. When either no inducer is present or inducer is present with glucose (**A**), the *lacI* gene products that are synthesized constitutively form a repressor tetramer molecule which binds at the operator locus to prevent the efficient initiation of transcription by RNA polymerase at the promoter locus and thus to prevent the subsequent transcription of the *lacZ*, *lacY*, and *lacA* structural genes. When inducer is present (**B**), the constitutively expressed *lacI* gene forms repressor molecules that are conformationally altered by the inducer and cannot efficiently bind to the operator locus (affinity of binding reduced > 1000-fold). In the presence of cAMP and its binding protein (CAP), the RNA polymerase can transcribe the structural genes *lacZ*, *lacY*, and *lacA*, and the polycistronic mRNA molecule formed can be translated into the corresponding protein molecules β -galactosidase, permease, and transacetylase, allowing for the catabolism of lactose.

The physiology of induction of the *lac* operon is well understood at the molecular level (Figure 39-3). Expression of the normal *lacI* gene of the *lac* operon is constitutive; it is expressed at a constant rate, resulting in formation of the subunits of the *lac* repressor. Four identical subunits with molecular weights of 38,000 assemble into a *lac* repressor molecule. The LacI repressor protein molecule, the product of *lacI*, has a high affinity (K_d about 10^{-13} mol/L) for the operator locus. The **operator locus** is a region of double-stranded DNA 27 base pairs long with a twofold rotational symmetry and

an inverted palindrome (indicated by solid lines about the dotted axis) in a region that is 21 base pairs long, as shown below:



The minimum effective size of an operator for LacI repressor binding is 17 base pairs (boldface letters in the

above sequence). At any one time, only two subunits of the repressors appear to bind to the operator, and within the 17-base-pair region at least one base of each base pair is involved in *LacI* recognition and binding. The binding occurs mostly in the **major groove** without interrupting the base-paired, double-helical nature of the operator DNA. The **operator locus** is between the **promoter site**, at which the DNA-dependent RNA polymerase attaches to commence transcription, and the transcription initiation site of the ***lacZ* gene**, the structural gene for β -galactosidase (Figure 39-2). When attached to the operator locus, the *LacI* repressor molecule prevents transcription of the operator locus as well as of the distal structural genes, *lacZ*, *lacY*, and *lacA*. Thus, the *LacI* repressor molecule is a **negative regulator**; in its presence (and in the absence of inducer; see below), expression from the *lacZ*, *lacY*, and *lacA* genes is prevented. There are normally 20–40 repressor tetramer molecules in the cell, a concentration of tetramer sufficient to effect, at any given time, > 95% occupancy of the one *lac* operator element in a bacterium, thus ensuring low (but not zero) basal *lac* operon gene transcription in the absence of inducing signals.

A lactose analog that is capable of inducing the *lac* operon while not itself serving as a substrate for β -galactosidase is an example of a **gratuitous inducer**. An example is isopropylthiogalactoside (IPTG). The addition of lactose or of a gratuitous inducer such as IPTG to bacteria growing on a poorly utilized carbon source (such as succinate) results in prompt induction of the *lac* operon enzymes. Small amounts of the gratuitous inducer or of lactose are able to enter the cell even in the absence of permease. The *LacI* repressor molecules—both those attached to the operator loci and those free in the cytosol—have a high affinity for the inducer. Binding of the inducer to a repressor molecule attached to the operator locus induces a conformational change in the structure of the repressor and causes it to dissociate from the DNA because its affinity for the operator is now 10^3 times lower (K_d about 10^{-9} mol/L) than that of *LacI* in the absence of IPTG. If DNA-dependent RNA polymerase has already attached to the coding strand at the promoter site, transcription will begin. The polymerase generates a polycistronic mRNA whose 5' terminal is complementary to the template strand of the operator. In such a manner, **an inducer derepresses the *lac* operon** and allows transcription of the structural genes for β -galactosidase, galactoside permease, and thiogalactoside transacetylase. Translation of the polycistronic mRNA can occur even before transcription is completed. Derepression of the *lac* operon allows the cell to synthesize the enzymes necessary to catabolize lactose as an energy source. Based on the physiology just described, IPTG-induced expression of transfected plasmids bearing the *lac* operator-promoter ligated to appro-

priate bioengineered constructs is commonly used to express mammalian recombinant proteins in *E. coli*.

In order for the RNA polymerase to efficiently form a PIC at the promoter site, there must also be present the **catabolite gene activator protein (CAP)** to which cAMP is bound. By an independent mechanism, the bacterium accumulates cAMP only when it is starved for a source of carbon. In the presence of glucose—or of glycerol in concentrations sufficient for growth—the bacteria will lack sufficient cAMP to bind to CAP because the glucose inhibits adenylyl cyclase, the enzyme that converts ATP to cAMP (see Chapter 42). Thus, in the presence of glucose or glycerol, cAMP-saturated CAP is lacking, so that the DNA-dependent RNA polymerase cannot initiate transcription of the *lac* operon. In the presence of the CAP-cAMP complex, which binds to DNA just upstream of the promoter site, transcription then occurs (Figure 39-3). Studies indicate that a region of CAP contacts the RNA polymerase α subunit and facilitates binding of this enzyme to the promoter. Thus, the CAP-cAMP regulator is acting as a **positive regulator** because its presence is required for gene expression. The *lac* operon is therefore controlled by two distinct, ligand-modulated DNA binding trans factors; one that acts positively (cAMP-CRP complex) and one that acts negatively (*LacI* repressor). Maximal activity of the *lac* operon occurs when glucose levels are low (high cAMP with CAP activation) and lactose is present (*LacI* is prevented from binding to the operator).

When the *lacI* gene has been mutated so that its product, *LacI*, is not capable of binding to operator DNA, the organism will exhibit **constitutive expression** of the *lac* operon. In a contrary manner, an organism with a *lacI* gene mutation that produces a *LacI* protein which prevents the binding of an inducer to the repressor will remain repressed even in the presence of the inducer molecule, because the inducer cannot bind to the repressor on the operator locus in order to derepress the operon. Similarly, bacteria harboring mutations in their *lac* operator locus such that the operator sequence will not bind a normal repressor molecule constitutively express the *lac* operon genes. Mechanisms of positive and negative regulation comparable to those described here for the *lac* system have been observed in eukaryotic cells (see below).

The Genetic Switch of Bacteriophage Lambda (λ) Provides a Paradigm for Protein-DNA Interactions in Eukaryotic Cells

Like some eukaryotic viruses (eg, herpes simplex, HIV), some bacterial viruses can either reside in a dormant state within the host chromosomes or can replicate

within the bacterium and eventually lead to lysis and killing of the bacterial host. Some *E. coli* harbor such a “temperate” virus, bacteriophage lambda (λ). When lambda infects an organism of that species it injects its 45,000-bp, double-stranded, linear DNA genome into the cell (Figure 39–4). Depending upon the nutritional state of the cell, the lambda DNA will either **integrate** into the host genome (**lysogenic pathway**) and remain dormant until activated (see below), or it will commence **replicating** until it has made about 100 copies of complete, protein-packaged virus, at which point it causes lysis of its host (**lytic pathway**). The newly generated virus particles can then infect other susceptible hosts.

When integrated into the host genome in its dormant state, lambda will remain in that state until activated by exposure of its lysogenic bacterial host to DNA-damaging agents. In response to such a noxious stimulus, the dormant bacteriophage becomes “induced” and begins to transcribe and subsequently translate those genes of its own genome which are necessary for its excision from the host chromosome, its DNA replication, and the synthesis of its protein coat and lysis enzymes. This event acts like a trigger or type C (Figure 39–1) response; i.e., once lambda has committed itself to induction, there is no turning back until the cell is lysed and the replicated bacteriophage released. This switch from a dormant or **prophage state** to a **lytic infection** is well understood at the genetic and molecular levels and will be described in detail here.

The switching event in lambda is centered around an 80-bp region in its double-stranded DNA genome referred to as the “right operator” (O_R) (Figure 39–5A). The **right operator** is flanked on its left side by the structural gene for the lambda repressor protein, the **cI protein**, and on its right side by the structural gene encoding another regulatory protein called **Cro**. When lambda is in its prophage state—i.e., integrated into the host genome—the **cI** repressor gene is the *only* lambda gene **cI** protein that is expressed. When the bacteriophage is undergoing lytic growth, the **cI** repressor gene is not expressed, but the **cro** gene—as well as many other genes in lambda—is expressed. That is, **when the repressor gene is on, the cro gene is off, and when the cro gene is on, the repressor gene is off**. As we shall see, these two genes regulate each other’s expression and thus, ultimately, the decision between lytic and lysogenic growth of lambda. **This decision between repressor gene transcription and cro gene transcription is a paradigmatic example of a molecular switch.**

The 80-bp λ right operator, O_R , can be subdivided into three discrete, evenly spaced, 17-bp cis-active DNA elements that represent the binding sites for either of two bacteriophage λ regulatory proteins. Import-

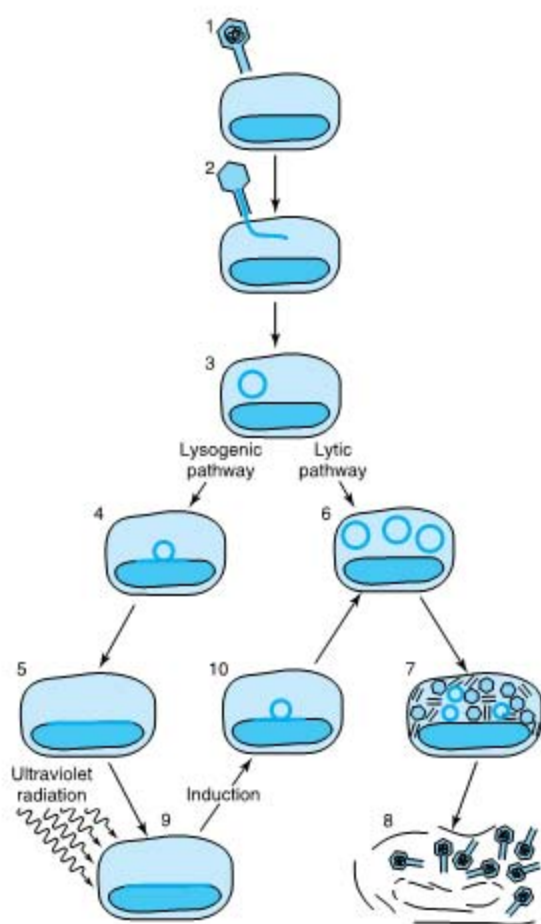


Figure 39–4. Infection of the bacterium *E. coli* by phage lambda begins when a virus particle attaches itself to the bacterial cell (1) and injects its DNA (shaded line) into the cell (2, 3). Infection can take either of two courses depending on which of two sets of viral genes is turned on. In the lysogenic pathway, the viral DNA becomes integrated into the bacterial chromosome (4, 5), where it replicates passively as the bacterial cell divides. The dormant virus is called a prophage, and the cell that harbors it is called a lysogen. In the alternative lytic mode of infection, the viral DNA replicates itself (6) and directs the synthesis of viral proteins (7). About 100 new virus particles are formed. The proliferating viruses lyse, or burst, the cell (8). A prophage can be “induced” by a DNA damaging agent such as ultraviolet radiation (9). The inducing agent throws a switch, so that a different set of genes is turned on. Viral DNA loops out of the chromosome (10) and replicates; the virus proceeds along the lytic pathway. (Reproduced, with permission, from Ptashne M, Johnson AD, Pabo CO: A genetic switch in a bacterial virus. *Sci Am* [Nov] 1982;247:128.)

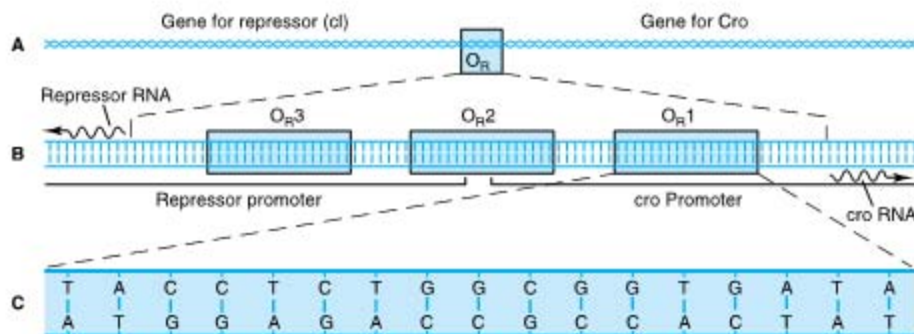


Figure 39-5. Right operator (O_R) is shown in increasing detail in this series of drawings. The operator is a region of the viral DNA some 80 base pairs long (A). To its left lies the gene encoding lambda repressor (*cl*), to its right the gene (*cro*) encoding the regulator protein Cro. When the operator region is enlarged (B), it is seen to include three subregions, O_{R1} , O_{R2} , and O_{R3} , each 17 base pairs long. They are recognition sites to which both repressor and Cro can bind. The recognition sites overlap two promoters—sequences of bases to which RNA polymerase binds in order to transcribe these genes into mRNA (wavy lines), that are translated into protein. Site O_{R1} is enlarged (C) to show its base sequence. Note that in this region of the λ chromosome, both strands of DNA act as a template for transcription (Chapter 39). (Reproduced, with permission, from Ptashne M, Johnson AD, Pabo CO: A genetic switch in a bacterial virus. *Sci Am* [Nov] 1982;247:128.)

tantly, the nucleotide sequences of these three tandemly arranged sites are similar but not identical (Figure 39-5B). The three related *cis* elements, termed operators O_{R1} , O_{R2} , and O_{R3} , can be bound by either *cl* or Cro proteins. However, the relative affinities of *cl* and Cro for each of the sites varies, and this differential binding affinity is central to the appropriate operation of the λ phage lytic or lysogenic “molecular switch.” The DNA region between the *cro* and repressor genes also contains two promoter sequences that direct the binding of RNA polymerase in a specified orientation, where it commences transcribing adjacent genes. One promoter directs RNA polymerase to transcribe in the **rightward direction** and, thus, to transcribe *cro* and other distal genes, while the other promoter directs the transcription of the **repressor gene** in the **leftward direction** (Figure 39-5B).

The product of the repressor gene, the 236-amino-acid, 27 kDa **repressor protein**, exists as a **two-domain** molecule in which the **amino terminal domain binds to operator DNA** and the **carboxyl terminal domain promotes the association** of one repressor protein with another to form a dimer. A **dimer** of repressor molecules binds to **operator DNA** much more tightly than does the monomeric form (Figure 39-6A to 39-6C).

The product of the *cro* gene, the 66-amino-acid, 9 kDa **Cro protein**, has a single domain but also binds the operator DNA more tightly as a **dimer** (Figure

39-6D). The Cro protein’s single domain mediates both operator binding and dimerization.

In a lysogenic bacterium—ie, a bacterium containing a lambda prophage—the lambda repressor dimer binds **preferentially to O_{R1}** but in so doing, by a cooperative interaction, enhances the binding (by a factor of 10) of another repressor dimer to O_{R2} (Figure 39-7). The affinity of repressor for O_{R3} is the least of the three operator subregions. The binding of repressor to O_{R1} has two major effects. The occupation of O_{R1} by repressor **blocks the binding of RNA polymerase to the rightward promoter** and in that way prevents expression of *cro*. Second, as mentioned above, repressor dimer bound to O_{R1} enhances the binding of repressor dimer to O_{R2} . The binding of repressor to O_{R2} has the important added effect of **enhancing the binding of RNA polymerase to the leftward promoter** that overlaps O_{R2} and thereby enhances transcription and subsequent expression of the repressor gene. This enhancement of transcription is apparently mediated through direct protein-protein interactions between promoter-bound RNA polymerase and O_{R2} -bound repressor. Thus, the lambda repressor is both a **negative regulator**, by preventing transcription of *cro*, and a **positive regulator**, by enhancing transcription of its own gene, the repressor gene. This dual effect of repressor is responsible for the stable state of the dormant lambda bacteriophage; not only does the repressor prevent expression of the genes necessary for lysis, but it also promotes expression of itself to

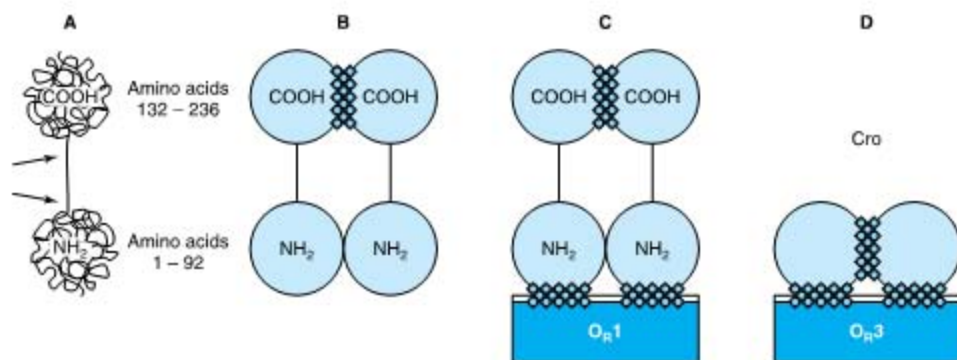


Figure 39-6. Schematic molecular structures of λ (lambda repressor, shown in **A**, **B**, and **C**) and Cro (**D**). Lambda repressor protein is a polypeptide chain 236 amino acids long. The chain folds itself into a dumbbell shape with two substructures: an amino terminal (NH₂) domain and a carboxyl terminal (COOH) domain. The two domains are linked by a region of the chain that is susceptible to cleavage by proteases (indicated by the two arrows in **A**). Single repressor molecules (monomers) tend to associate to form dimers (**B**); a dimer can dissociate to form monomers again. A dimer is held together mainly by contact between the carboxyl terminal domains (hatching). Repressor dimers bind to (and can dissociate from) the recognition sites in the operator region; they display the greatest affinity for site O_R1 (**C**). It is the amino terminal domain of the repressor molecule that makes contact with the DNA (hatching). Cro (**D**) has a single domain with sites that promote dimerization and other sites that promote binding of dimers to operator, preferentially to O_R3. (Reproduced, with permission, from Ptashne M, Johnson AD, Pabo CO. A genetic switch in a bacterial virus. *Sci Am* [Nov] 1982;247:128.)

stabilize this state of differentiation. In the event that intracellular repressor protein concentration becomes very high, this excess repressor will bind to O_R3 and by so doing diminish transcription of the repressor gene from the leftward promoter until the repressor concentration drops and repressor dissociates itself from O_R3.

With such a stable, repressive, λ -mediated, lysogenic state, one might wonder how the lytic cycle could ever be entered. However, this process does occur quite efficiently. When a DNA-damaging signal, such as ultraviolet light, strikes the lysogenic host bacterium, fragments of single-stranded DNA are generated that activate a specific **protease** coded by a bacterial gene and referred to as **recA** (Figure 39-7). The activated recA protease hydrolyzes the portion of the repressor protein that connects the amino terminal and carboxyl terminal domains of that molecule (see Figure 39-6A). Such cleavage of the repressor domains causes the **repressor dimers to dissociate**, which in turn causes **dissociation of the repressor molecules from O_R2** and eventually from O_R1. The effects of removal of repressor from O_R1 and O_R2 are predictable. RNA polymerase immediately has access to the rightward promoter and commences transcribing the **cro gene**, and the enhancement effect of the repressor at O_R2 on leftward transcription is lost (Figure 39-7).

The resulting newly synthesized Cro protein also binds to the operator region as a dimer, but its order of preference is opposite to that of repressor (Figure 39-7). That is, **Cro binds most tightly to O_R3**, but there is no cooperative effect of Cro at O_R3 on the binding of Cro to O_R2. At increasingly higher concentrations of Cro, the protein will bind to O_R2 and eventually to O_R1.

Occupancy of O_R3 by Cro immediately turns off transcription from the leftward promoter and in that way **prevents any further expression of the repressor gene**. The molecular switch is thus completely “thrown” in the lytic direction. The *cro* gene is now expressed, and the repressor gene is fully turned off. This event is irreversible, and the expression of other lambda genes begins as part of the lytic cycle. When Cro repressor concentration becomes quite high, it will eventually occupy O_R1 and in so doing reduce the expression of its own gene, a process that is necessary in order to effect the final stages of the lytic cycle.

The three-dimensional structures of Cro and of the lambda repressor protein have been determined by x-ray crystallography, and models for their binding and effecting the above-described molecular and genetic events have been proposed and tested. Both bind to DNA using helix-turn-helix DNA binding domain motifs (see below).

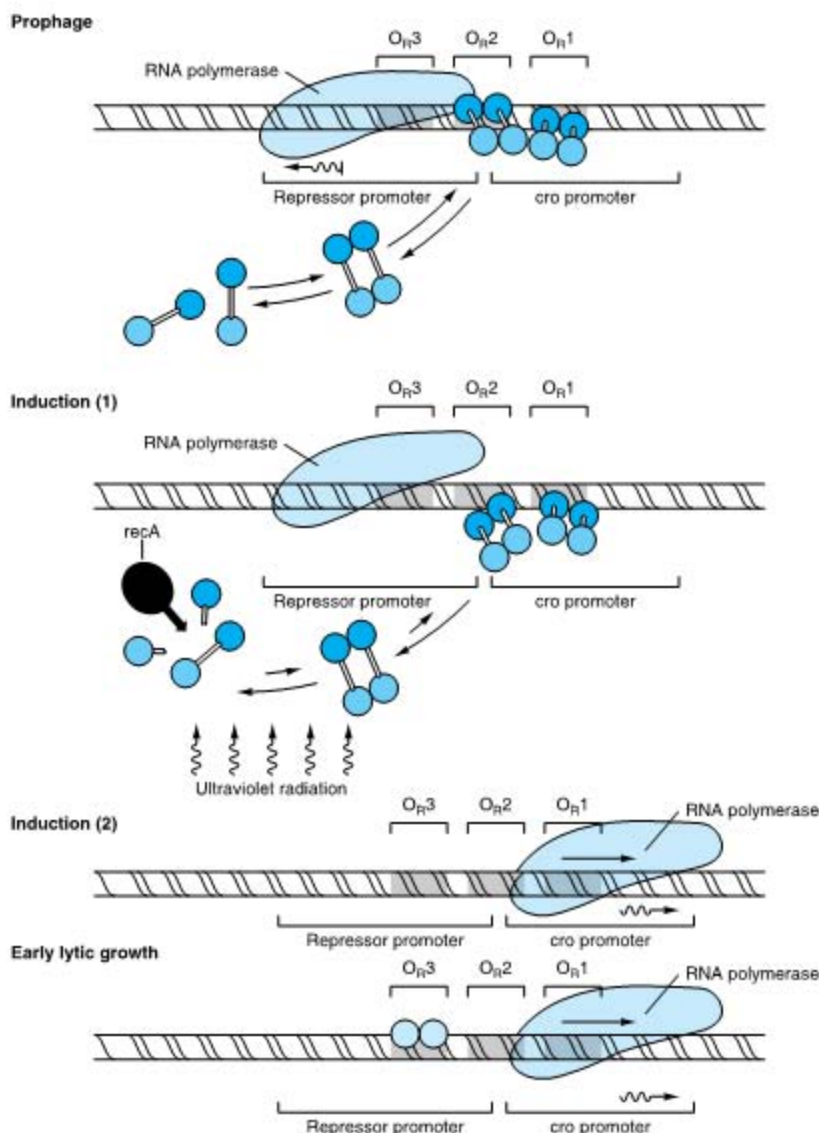


Figure 39-7. Configuration of the switch is shown at four stages of lambda's life cycle. The lysogenic pathway (in which the virus remains dormant as a prophage) is selected when a repressor dimer binds to O_R1 , thereby making it likely that O_R2 will be filled immediately by another dimer. In the prophage (top), the repressor dimers bound at O_R1 and O_R2 prevent RNA polymerase from binding to the rightward promoter and so block the synthesis of Cro (negative control). The repressors also enhance the binding of polymerase to the leftward promoter (positive control), with the result that the repressor gene is transcribed into RNA (wavy line) and more repressor is synthesized, maintaining the lysogenic state. The prophage is induced when ultraviolet radiation activates the protease *recA*, which cleaves repressor monomers. The equilibrium of free monomers, free dimers, and bound dimers is thereby shifted, and dimers leave the operator sites. RNA polymerase is no longer encouraged to bind to the leftward promoter, so that repressor is no longer synthesized. As induction proceeds, all the operator sites become vacant, and so polymerase can bind to the rightward promoter and Cro is synthesized. During early lytic growth, a single Cro dimer binds to O_R3 (shaded circles), the site for which it has the highest affinity. Consequently, RNA polymerase cannot bind to the leftward promoter, but the rightward promoter remains accessible. Polymerase continues to bind there, transcribing *cro* and other early lytic genes. Lytic growth ensues. (Reproduced, with permission, from Ptashne M, Johnson AD, Pabo CO: A genetic switch in a bacterial virus. *Sci Am* [Nov] 1982;247:128.)

To date, this system provides the best understanding of the molecular events involved in gene regulation.

Detailed analysis of the lambda repressor led to the important concept that transcription regulatory proteins have several functional domains. For example, lambda repressor binds to DNA with high affinity. Repressor monomers form dimers, dimers interact with each other, and repressor interacts with RNA polymerase. The protein-DNA interface and the three protein-protein interfaces all involve separate and distinct domains of the repressor molecule. As will be noted below (see Figure 39-17), this is a characteristic shared by most (perhaps all) molecules that regulate transcription.

SPECIAL FEATURES ARE INVOLVED IN REGULATION OF EUKARYOTIC GENE TRANSCRIPTION

Most of the DNA in prokaryotic cells is organized into genes, and the templates can always be transcribed. A very different situation exists in mammalian cells, in which relatively little of the total DNA is organized into genes and their associated regulatory regions. The function of the extra DNA is unknown. In addition, as described in Chapter 36, the DNA in eukaryotic cells is extensively folded and packed into the protein-DNA complex called chromatin. Histones are an important part of this complex since they both form the structures known as nucleosomes (see Chapter 36) and also factor significantly into gene regulatory mechanisms as outlined below.

Chromatin Remodeling Is an Important Aspect of Eukaryotic Gene Expression

Chromatin structure provides an additional level of control of gene transcription. As discussed in Chapter 36, large regions of chromatin are transcriptionally inactive while others are either active or potentially active. With few exceptions, each cell contains the same complement of genes (antibody-producing cells are a notable exception). The development of specialized organs, tissues, and cells and their function in the intact organism depend upon the differential expression of genes.

Some of this differential expression is achieved by having different regions of chromatin available for transcription in cells from various tissues. For example, the DNA containing the β -globin gene cluster is in “**active**” chromatin in the reticulocyte but in “**inactive**” chromatin in muscle cells. All the factors involved in the determination of active chromatin have not been elucidated. The presence of nucleosomes and of complexes of histones and DNA (see Chapter 36) certainly provides a barrier against the ready association of transcription fac-

tors with specific DNA regions. The dynamics of the formation and disruption of nucleosome structure are therefore an important part of eukaryotic gene regulation.

Histone acetylation and deacetylation is an important determinant of gene activity. The surprising discovery that histone acetylase activity is associated with TAFs and the coactivators involved in hormonal regulation of gene transcription (see Chapter 43) has provided a new concept of gene regulation. Acetylation is known to occur on lysine residues in the amino terminal tails of histone molecules. This modification reduces the positive charge of these tails and decreases the binding affinity of histone for the negatively charged DNA. Accordingly, the acetylation of histone could result in disruption of nucleosomal structure and allow readier access of transcription factors to cognate regulatory DNA elements. As discussed previously, this would enhance binding of the basal transcription machinery to the promoter. Histone deacetylation would have the opposite effect. Different proteins with specific acetylase and deacetylase activities are associated with various components of the transcription apparatus. The specificity of these processes is under investigation, as are a variety of mechanisms of action. Some specific examples are illustrated in Chapter 43.

There is evidence that the **methylation of deoxycytidine residues** (in the sequence 5'-mCpG-3') in DNA may effect gross changes in chromatin so as to preclude its active transcription, as described in Chapter 36. For example, in mouse liver, only the unmethylated ribosomal genes can be expressed, and there is evidence that many animal viruses are not transcribed when their DNA is methylated. Acute demethylation of deoxycytidine residues in a specific region of the tyrosine aminotransferase gene—in response to glucocorticoid hormones—has been associated with an increased rate of transcription of the gene. However, it is not possible to generalize that methylated DNA is transcriptionally inactive, that all inactive chromatin is methylated, or that active DNA is not methylated.

Finally, the binding of specific transcription factors to cognate DNA elements may result in disruption of nucleosomal structure. Many eukaryotic genes have multiple protein-binding DNA elements. The serial binding of transcription factors to these elements—in a combinatorial fashion—may either directly disrupt the structure of the nucleosome or prevent its re-formation or recruit, via protein-protein interactions, multiprotein coactivator complexes that have the ability to covalently modify or remodel nucleosomes. These reactions result in chromatin-level structural changes that in the end increase DNA accessibility to other factors and the transcription machinery.

Eukaryotic DNA that is in an “active” region of chromatin can be transcribed. As in prokaryotic cells, a

promoter dictates where the RNA polymerase will initiate transcription, but this promoter cannot be neatly defined as containing a -35 and -10 box, particularly in mammalian cells (Chapter 37). In addition, the trans-acting factors generally come from other chromosomes (and so act in trans), whereas this consideration is moot in the case of the single chromosome-containing prokaryotic cells. Additional complexity is added by elements or factors that enhance or repress transcription, define tissue-specific expression, and modulate the actions of many effector molecules.

Certain DNA Elements Enhance or Repress Transcription of Eukaryotic Genes

In addition to gross changes in chromatin affecting transcriptional activity, certain DNA elements facilitate or enhance initiation at the promoter. For example, in simian virus 40 (SV40) there exists about 200 bp upstream from the promoter of the early genes a region of two identical, tandem 72-bp lengths that can greatly increase the expression of genes in vivo. Each of these 72-bp elements can be subdivided into a series of smaller elements; therefore, some enhancers have a very complex structure. **Enhancer elements** differ from the promoter in two remarkable ways. They can exert their positive influence on transcription even when separated by thousands of base pairs from a promoter; they work when oriented in either direction; and they can work upstream (5') or downstream (3') from the promoter. Enhancers are promiscuous; they can stimulate any promoter in the vicinity and may act on more than one promoter. The SV40 enhancer element can exert an influence on, for example, the transcription of β -globin by increasing its transcription 200-fold in cells containing both the enhancer and the β -globin gene on the same plasmid (see below and Figure 39-8). The enhancer element does not produce a product that in turn acts on the promoter, since it is active only when it exists within the same DNA molecule as (ie, *cis* to) the promoter. Enhancer binding proteins are responsible for this effect. The exact mechanisms by which these transcription activators work are subject to much debate. Certainly, enhancer binding trans factors have been shown to interact with a plethora of other transcription proteins. These interactions include chromatin-modifying coactivators as well as the individual components of the basal RNA polymerase II transcription machinery. Ultimately, trans-factor-enhancer DNA binding events result in an increase in the binding of the basal transcription machinery to the promoter. Enhancer elements and associated binding proteins often convey nuclease hypersensitivity to those regions where they reside (Chapter 36). A summary of the properties of enhancers is presented in Table 39-2. One of the

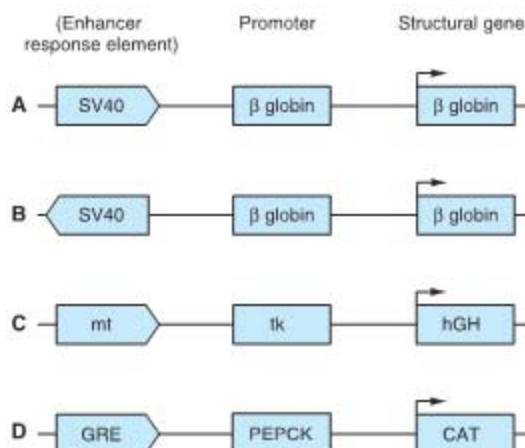


Figure 39-8. A schematic explanation of the action of enhancers and other *cis*-acting regulatory elements. These model chimeric genes consist of a reporter (structural) gene that encodes a protein which can be readily assayed, a promoter that ensures accurate initiation of transcription, and the putative regulatory elements. In all cases, high-level transcription from the indicated chimeras depends upon the presence of enhancers, which stimulate transcription ≥ 100 -fold over basal transcriptional levels (ie, transcription of the same chimeric genes containing just promoters fused to the structural genes). Examples **A** and **B** illustrate the fact that enhancers (eg, SV40) work in either orientation and upon a heterologous promoter. Example **C** illustrates that the metallothionein (mt) regulatory element (which under the influence of cadmium or zinc induces transcription of the endogenous mt gene and hence the metal-binding mt protein) will work through the thymidine kinase (tk) promoter to enhance transcription of the human growth hormone (hGH) gene. The engineered genetic constructions were introduced into the male pronuclei of single-cell mouse embryos and the embryos placed into the uterus of a surrogate mother to develop as transgenic animals. Offspring have been generated under these conditions, and in some the addition of zinc ions to their drinking water effects an increase in liver growth hormone. In this case, these transgenic animals have responded to the high levels of growth hormone by becoming twice as large as their normal litter mates. Example **D** illustrates that a glucocorticoid response element (GRE) will work through homologous (PEPCK gene) or heterologous promoters (not shown; ie, tk promoter, SV40 promoter, β -globin promoter, etc).

Table 39–2. Summary of the properties of enhancers.

-
- Work when located long distances from the promoter
 - Work when upstream or downstream from the promoter
 - Work when oriented in either direction
 - Work through heterologous promoters
 - Work by binding one or more proteins
 - Work by facilitating binding of the basal transcription complex to the promoter
-

best-understood mammalian enhancer systems is that of the β -interferon gene. This gene is induced upon viral infection of mammalian cells. One goal of the cell, once virally infected, is to attempt to mount an antiviral response—if not to save the infected cell, then to help to save the entire organism from viral infection. Interferon production is one mechanism by which this is accomplished. This family of proteins is secreted by virally infected cells. They interact with neighboring cells to cause an inhibition of viral replication by a variety of mechanisms, thereby limiting the extent of viral infection. The enhancer element controlling induction of this gene, located between nucleotides –110 and –45 of the β -interferon gene, is well characterized. This enhancer is composed of four distinct clustered cis elements, each of which is bound by distinct trans factors. One cis element is bound by the trans-acting factor NF- κ B, one by a member of the IRF (interferon regulatory factor) family of trans factors, and a third by the heterodimeric leucine zipper factor ATF-2/c-Jun. The fourth factor is the ubiquitous, architectural transcription factor known as HMG I(Y). Upon binding to its degenerate, A+T-rich binding sites, HMG I(Y) induces a significant bend in the DNA. There are four such HMG I(Y) binding sites interspersed throughout the enhancer. These sites play a critical role in forming the enhanceosome, along with the aforementioned three trans factors, by inducing a series of critically spaced DNA bends. Consequently, HMG I(Y) induces the cooperative formation of a unique, stereospecific, three dimensional structure within which all four factors are active when viral infection signals are sensed by the cell. The structure formed by the cooperative assembly of these four factors is termed the β -interferon enhanceosome (see Figure 39–9), so named because of its obvious structural similarity to the nucleosome, also a unique three-dimensional protein DNA structure that wraps DNA about an assembly of proteins (see Figures 36–1 and 36–2). The enhanceosome, once formed, induces a large increase in β -interferon gene transcription upon virus infection. It is not simply the protein occupancy of the linearly apposed cis element sites that in-

duces β -interferon gene transcription—rather, it is the formation of the enhanceosome proper that provides appropriate surfaces for the recruitment of coactivators that results in the enhanced formation of the PIC on the cis-linked promoter and thus transcription activation.

The cis-acting elements that decrease or **repress** the expression of specific genes have also been identified. Because fewer of these elements have been studied, it is not possible to formulate generalizations about their mechanism of action—though again, as for gene activation, chromatin level covalent modifications of histones and other proteins by (repressor)-recruited multisubunit corepressors have been implicated.

Tissue-Specific Expression May Result From the Action of Enhancers or Repressors

Many genes are now recognized to harbor enhancer or activator elements in various locations relative to their coding regions. In addition to being able to enhance gene transcription, some of these enhancer elements clearly possess the ability to do so in a tissue-specific manner. Thus, the enhancer element associated with the immunoglobulin genes between the J and C regions enhances the expression of those genes preferentially in lymphoid cells. Similarly to the SV40 enhancer, which is capable of promiscuously activating a variety of cis-linked genes, enhancer elements associated with the genes for pancreatic enzymes are capable of enhancing even unrelated but physically linked genes preferentially in the pancreatic cells of mice into which the specifically engineered gene constructions were introduced microscopically at the single-cell embryo stage. This **transgenic animal** approach has proved useful in studying tissue-specific gene expression. For example, DNA containing a pancreatic B cell tissue-specific enhancer (from the insulin gene), when ligated in a vector to polyoma large-T antigen, an oncogene, produced B cell tumors in transgenic mice. Tumors did not develop in any other tissue. Tissue-specific gene expression may therefore be mediated by enhancers or enhancer-like elements.

Reporter Genes Are Used to Define Enhancers & Other Regulatory Elements

By ligating regions of DNA suspected of harboring regulatory sequences to various reporter genes (the **reporter** or **chimeric gene approach**) (Figures 39–10 and 39–11), one can determine which regions in the vicinity of structural genes have an influence on their expression. Pieces of DNA thought to harbor regulatory elements are ligated to a suitable reporter gene and

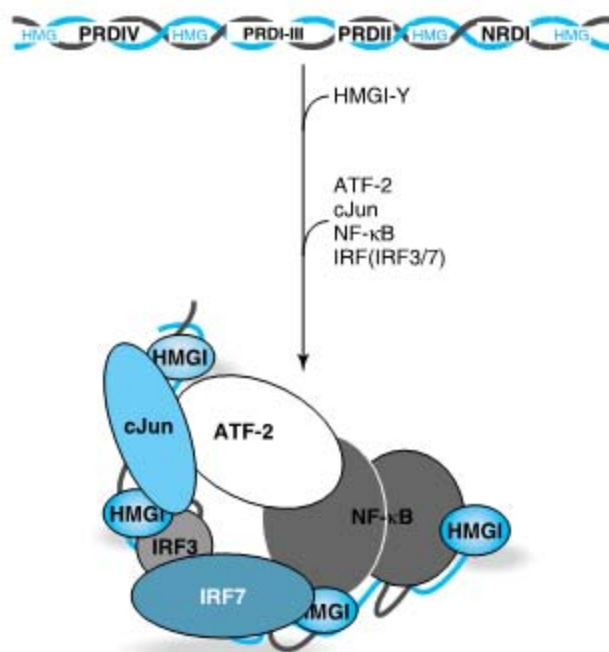


Figure 39–9. Formation and putative structure of the enhanceosome formed on the human β -interferon gene enhancer. Diagrammatically represented at the top is the distribution of the multiple cis-elements (HMG, PRDIV, PRDI-III, PRDII, NRDI) composing the β -interferon gene enhancer. The intact enhancer mediates transcriptional induction of the β -interferon gene (over 100-fold) upon virus infection of human cells. The cis-elements of this modular enhancer represent the binding sites for the trans-factors HMG I(Y), cJun-ATF-2, IRF3, IRF7, and NF- κ B, respectively. The factors interact with these DNA elements in an obligatory, ordered, and highly cooperative fashion as indicated by the arrow. Initial binding of four HMG I(Y) proteins induces sharp DNA bends in the enhancer, causing the entire 70–80 bp region to assume a high level of curvature. This curvature is integral to the subsequent highly cooperative binding of the other trans-factors since this enables the DNA-bound factors to make important, direct protein-protein interactions that both contribute to the formation and stability of the enhanceosome and generate a unique three-dimensional surface that serves to recruit chromatin-modifying activities (eg, Swi/Snf and P/CAF) as well as the general transcription machinery (RNA polymerase II and GTFs). Although four of the five cis-elements (PRDIV, PRDI-III, PRDII, NRDI) independently can modestly stimulate (\sim tenfold) transcription of a reporter gene in transfected cells (see Figures 39–10 and 39–12), all five cis-elements, in appropriate order, are required to form an enhancer that can appropriately stimulate mRNA gene transcription (ie, \geq 100-fold) in response to viral infection of a human cell. This distinction indicates the strict requirement for appropriate enhanceosome architecture for efficient trans-activation. Similar enhanceosomes, involving distinct cis- and trans-factors, are proposed to form on many other mammalian genes.

introduced into a host cell (Figure 39–10). Basal expression of the reporter gene will be increased if the DNA contains an enhancer. Addition of a hormone or heavy metal to the culture medium will increase expression of the reporter gene if the DNA contains a hormone or metal response element (Figure 39–11). The location of the element can be pinpointed by using progressively shorter pieces of DNA, deletions, or point mutations (Figure 39–11).

This strategy, **using transfected cells in culture and transgenic animals**, has led to the identification of dozens of enhancers, repressors, tissue-specific elements, and hormone, heavy metal, and drug-response elements. The activity of a gene at any moment reflects the interaction of these numerous cis-acting DNA elements with their respective trans-acting factors. The challenge now is to figure out how this occurs.

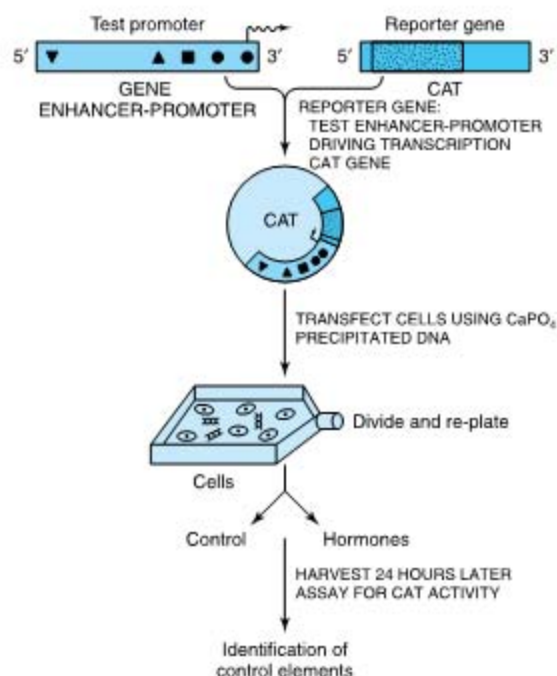


Figure 39–10. The use of reporter genes to define DNA regulatory elements. A DNA fragment from the gene in question—in this example, approximately 2 kb of 5′-flanking DNA and cognate promoter—is ligated into a plasmid vector that contains a suitable reporter gene—in this case, the bacterial enzyme chloramphenicol transferase (CAT). The enzyme luciferase (abbreviated LUC) is another popular reporter gene. Neither LUC nor CAT is present in mammalian cells; hence, detection of these activities in a cell extract means that the cell was successfully transfected by the plasmid. An increase of CAT activity over the basal level, eg, after addition of one or more hormones, means that the region of DNA inserted into the reporter gene plasmid contains functional hormone response elements (HRE). Progressively shorter pieces of DNA, regions with internal deletions, or regions with point mutations can be constructed and inserted to pinpoint the response element (see Figure 39–11 for deletion mapping of the relevant HREs).

Combinations of DNA Elements & Associated Proteins Provide Diversity in Responses

Prokaryotic genes are often regulated in an on-off manner in response to simple environmental cues. Some eukaryotic genes are regulated in the simple on-off man-

ner, but the process in most genes, especially in mammals, is much more complicated. Signals representing a number of complex environmental stimuli may converge on a single gene. The response of the gene to these signals can have several physiologic characteristics. First, the response may extend over a considerable range. This is accomplished by having additive and synergistic positive responses counterbalanced by negative or repressing effects. In some cases, either the positive or the negative response can be dominant. Also required is a mechanism whereby an effector such as a hormone can activate some genes in a cell while repressing others and leaving still others unaffected. When all of these processes are coupled with tissue-specific element factors, considerable flexibility is afforded. These physiologic variables obviously require an arrangement much more complicated than an on-off switch. The array of DNA elements in a promoter specifies—with associated factors—how a given gene will respond. Some simple examples are illustrated in Figure 39–12.

Transcription Domains Can Be Defined by Locus Control Regions & Insulators

The large number of genes in eukaryotic cells and the complex arrays of transcription regulatory factors presents an organizational problem. Why are some genes available for transcription in a given cell whereas others are not? If enhancers can regulate several genes and are not position- and orientation-dependent, how are they prevented from triggering transcription randomly? Part of the solution to these problems is arrived at by having the chromatin arranged in functional units that restrict patterns of gene expression. This may be achieved by having the chromatin form a structure with the nuclear matrix or other physical entity, or compartments within the nucleus. Alternatively, some regions are controlled by complex DNA elements called **locus control regions (LCRs)**. An LCR—with associated bound proteins—controls the expression of a cluster of genes. The best-defined LCR regulates expression of the globin gene family over a large region of DNA. Another mechanism is provided by **insulators**. These DNA elements, also in association with one or more proteins, prevent an enhancer from acting on a promoter on the other side of an insulator in another transcription domain.

SEVERAL MOTIFS MEDIATE THE BINDING OF REGULATORY PROTEINS TO DNA

The specificity involved in the control of transcription requires that regulatory proteins bind with high affinity to the correct region of DNA. Three unique motifs—the **helix-turn-helix**, the **zinc finger**, and the **leucine**

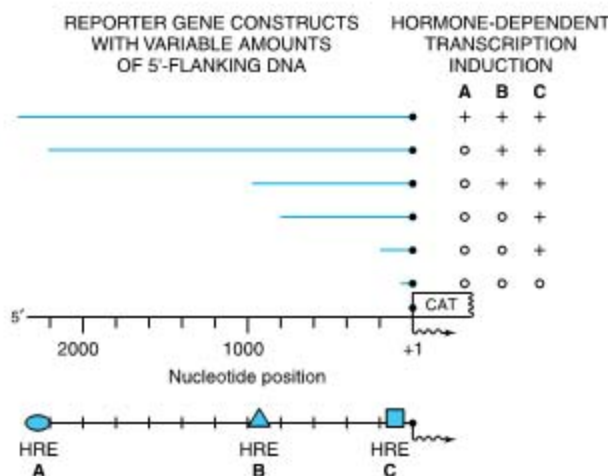


Figure 39-11. Location of hormone response elements (HREs) A, B, and C using the reporter gene-transfection approach. A family of reporter genes, constructed as described in Figure 39-10, can be transfected individually into a recipient cell. By analyzing when certain hormone responses are lost in comparison to the 5' deletion, specific hormone-responsive elements can be located.

zipper—account for many of these specific protein-DNA interactions. Examples of proteins containing these motifs are given in Table 39-3.

Comparison of the binding activities of the proteins that contain these motifs leads to several important generalizations.

- (1) Binding must be of high affinity to the specific site and of low affinity to other DNA.
- (2) Small regions of the protein make direct contact with DNA; the rest of the protein, in addition to pro-

Table 39-3. Examples of transcription regulatory proteins that contain the various binding motifs.

Binding Motif	Organism	Regulatory Protein
Helix-turn-helix	<i>E. coli</i>	lac repressor
	Phage	CAP
	Mammals	λ ci, cro, and tryptophan and 434 repressors homeo box proteins Pit-1, Oct1, Oct2
Zinc finger	<i>E. coli</i>	Gene 32 protein
	Yeast	Gal4
	<i>Drosophila</i>	Serendipity, Hunchback
	Xenopus Mammals	TFIIIA steroid receptor family, Sp1
Leucine zipper	Yeast	GCN4
	Mammals	C/EBP, fos, Jun, Fra-1, CRE binding protein, c-myc, n-myc, l-myc

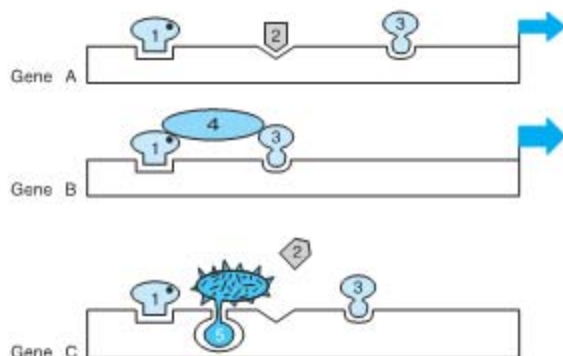


Figure 39-12. Combinations of DNA elements and proteins provide diversity in the response of a gene. Gene A is activated (the width of the arrow indicates the extent) by the combination of activators 1, 2, and 3 (probably with coactivators, as shown in Figure 37-10). Gene B is activated, in this case more effectively, by the combination of 1, 3, and 4; note that 4 does not contact DNA directly in this example. The activators could form a linear bridge that links the basal machinery to the promoter, or this could be accomplished by looping out of the DNA. In either case, the purpose is to direct the basal transcription machinery to the promoter. Gene C is inactivated by the combination of 1, 5, and 3; in this case, factor 5 is shown to preclude the essential binding of factor 2 to DNA, as occurs in example A. If activator 1 helps repressor 5 bind and if activator 1 binding requires a ligand (solid dot), it can be seen how the ligand could activate one gene in a cell (gene A) and repress another (gene C).

viding the trans-activation domains, may be involved in the dimerization of monomers of the binding protein, may provide a contact surface for the formation of heterodimers, may provide one or more ligand-binding sites, or may provide surfaces for interaction with coactivators or corepressors.

(3) The protein-DNA interactions are maintained by hydrogen bonds and van der Waals forces.

(4) The motifs found in these proteins are unique; their presence in a protein of unknown function suggests that the protein may bind to DNA.

(5) Proteins with the helix-turn-helix or leucine zipper motifs form symmetric dimers, and their respective DNA binding sites are symmetric palindromes. In proteins with the zinc finger motif, the binding site is repeated two to nine times. These features allow for cooperative interactions between binding sites and enhance the degree and affinity of binding.

The Helix-Turn-Helix Motif

The first motif described—and the one studied most extensively—is the helix-turn-helix. Analysis of the three-dimensional structure of the λ Cro transcription regulator has revealed that each monomer consists of three antiparallel β sheets and three α helices (Figure 39–13). The dimer forms by association of the antiparallel β_3 sheets. The α_3 helices form the DNA recognition surface, and the rest of the molecule appears to be involved in stabilizing these structures. The average diameter of an α helix is 1.2 nm, which is the approximate width of the major groove in the B form of DNA. The DNA recognition domain of each Cro monomer interacts with 5 bp and the dimer binding sites span 3.4 nm, allowing fit into successive half turns of the major groove on the same surface (Figure 39–13). X-ray analyses of the λ cI repressor, CAP (the cAMP receptor

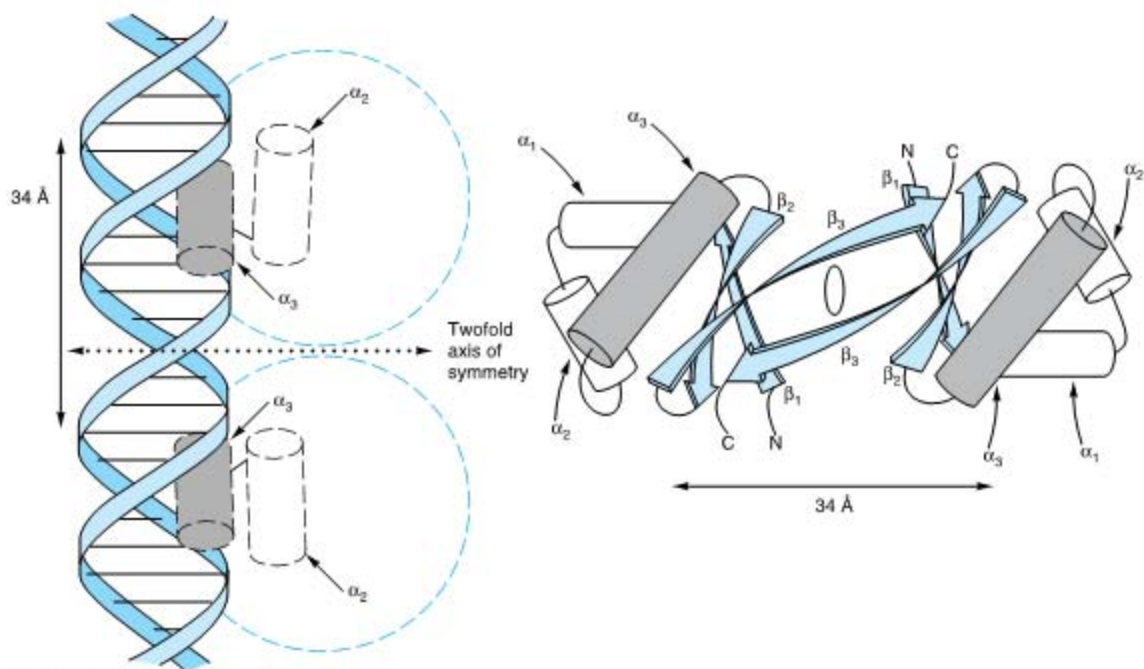


Figure 39–13. A schematic representation of the three-dimensional structure of Cro protein and its binding to DNA by its helix-turn-helix motif. The Cro monomer consists of three antiparallel β sheets (β_1 – β_3) and three α -helices (α_1 – α_3). The helix-turn-helix motif is formed because the α_3 and α_2 helices are held at about 90 degrees to each other by a turn of four amino acids. The α_3 helix of Cro is the DNA recognition surface (shaded). Two monomers associate through the antiparallel β_3 sheets to form a dimer that has a twofold axis of symmetry (right). A Cro dimer binds to DNA through its α_3 helices, each of which contacts about 5 bp on the same surface of the major groove. The distance between comparable points on the two DNA α -helices is 34 Å, which is the distance required for one complete turn of the double helix. (Courtesy of B Mathews.)

protein of *E. coli*), tryptophan repressor, and phage 434 repressor all also display this dimeric helix-turn-helix structure that is present in eukaryotic DNA proteins as well (see Table 39–3).

The Zinc Finger Motif

The zinc finger was the second DNA binding motif whose atomic structure was elucidated. It was known that the protein TFIIIA, a positive regulator of 5S RNA transcription, required zinc for activity. Structural and biophysical analyses revealed that each TFIIIA molecule contains nine zinc ions in a repeating coordination complex formed by closely spaced cysteine-cysteine residues followed 12–13 amino acids later by a histidine-histidine pair (Figure 39–14). In some instances—notably the steroid-thyroid receptor family—the His-His doublet is replaced by a second Cys-Cys pair. The protein containing zinc fingers appears to lie on one face of the DNA helix, with successive fingers alternatively positioned in one turn in the major groove. As is the case with the recognition domain in the helix-turn-helix protein, each TFIIIA zinc finger contacts about 5 bp of DNA. The importance of this motif in the action of steroid hormones is underscored by an “experiment of nature.” A single amino acid mutation in either of the two zinc fingers of the $1,25(\text{OH})_2\text{-D}_3$ receptor protein results in resistance to the action of this hormone and the clinical syndrome of rickets.

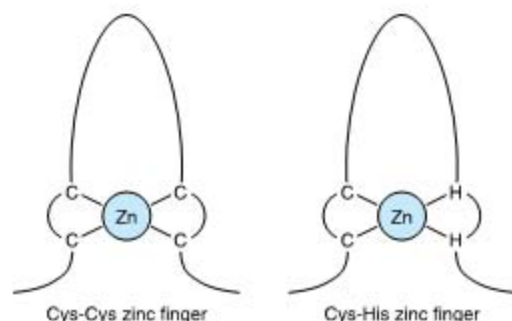


Figure 39–14. Zinc fingers are a series of repeated domains (two to nine) in which each is centered on a tetrahedral coordination with zinc. In the case of TFIIIA, the coordination is provided by a pair of cysteine residues (C) separated by 12–13 amino acids from a pair of histidine (H) residues. In other zinc finger proteins, the second pair also consists of C residues. Zinc fingers bind in the major groove, with adjacent fingers making contact with 5 bp along the same face of the helix.

The Leucine Zipper Motif

Careful analysis of a 30-amino-acid sequence in the carboxyl terminal region of the enhancer binding protein C/EBP revealed a novel structure. As illustrated in Figure 39–15, this region of the protein forms an α helix in which there is a periodic repeat of leucine residues at every seventh position. This occurs for eight helical turns and four leucine repeats. Similar structures have been found in a number of other proteins associated with the regulation of transcription in mammalian and yeast cells. It is thought that this structure allows two identical monomers or heterodimers (eg, Fos-Jun or Jun-Jun) to “zip together” in a coiled coil and form a tight dimeric complex (Figure 39–15). This protein-protein interaction may serve to enhance the association of the separate DNA binding domains with their target (Figure 39–15).

THE DNA BINDING & TRANS-ACTIVATION DOMAINS OF MOST REGULATORY PROTEINS ARE SEPARATE & NONINTERACTIVE

DNA binding could result in a general conformational change that allows the bound protein to activate transcription, or these two functions could be served by separate and independent domains. Domain swap experiments suggest that the latter is the case.

The *GAL1* gene product is involved in galactose metabolism in yeast. Transcription of this gene is positively regulated by the GAL4 protein, which binds to an upstream activator sequence (UAS), or enhancer, through an amino terminal domain. The amino terminal 73-amino-acid DNA-binding domain (DBD) of GAL4 was removed and replaced with the DBD of LexA, an *E. coli* DNA-binding protein. This domain swap resulted in a molecule that did not bind to the *GAL1* UAS and, of course, did not activate the *GAL1* gene (Figure 39–16). If, however, the *lexA* operator—the DNA sequence normally bound by the *lexA* DBD—was inserted into the promoter region of the *GAL* gene, the hybrid protein bound to this promoter (at the *lexA* operator) and it activated transcription of *GAL1*. This experiment, which has been repeated a number of times, affords solid evidence that the carboxyl terminal region of GAL4 causes transcriptional activation. These data also demonstrate that the DNA-binding DBD and trans-activation domains (ADs) are independent and noninteractive. The hierarchy involved in assembling gene transcription activating complexes includes proteins that bind DNA and trans-activate; others that form protein-protein complexes which bridge DNA-binding proteins to trans-activating proteins; and others that form protein-protein complexes with components of the basal transcription

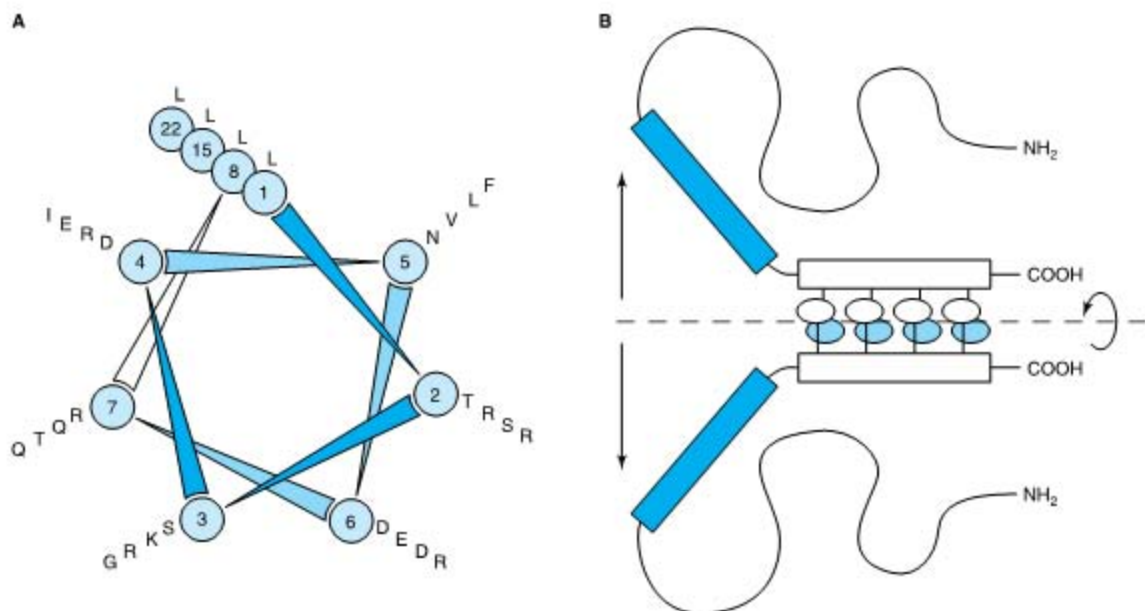


Figure 39-15. The leucine zipper motif. **A** shows a helical wheel analysis of a carboxyl terminal portion of the DNA binding protein C/EBP. The amino acid sequence is displayed end-to-end down the axis of a schematic α -helix. The helical wheel consists of seven spokes that correspond to the seven amino acids that comprise every two turns of the α -helix. Note that leucine residues (L) occur at every seventh position. Other proteins with "leucine zippers" have a similar helical wheel pattern. **B** is a schematic model of the DNA binding domain of C/EBP. Two identical C/EBP polypeptide chains are held in dimer formation by the leucine zipper domain of each polypeptide (denoted by the rectangles and attached ovals). This association is apparently required to hold the DNA binding domains of each polypeptide (the shaded rectangles) in the proper conformation for DNA binding. (Courtesy of S McKnight.)

apparatus. A given protein may thus have several surfaces or domains that serve different functions (see Figure 39-17). As described in Chapter 37, the primary purpose of these complex assemblies is to facilitate the assembly of the basal transcription apparatus on the cis-linked promoter.

GENE REGULATION IN PROKARYOTES & EUKARYOTES DIFFERS IN IMPORTANT RESPECTS

In addition to transcription, eukaryotic cells employ a variety of mechanisms to regulate gene expression (Table 39-4). The nuclear membrane of eukaryotic cells physically segregates gene transcription from translation, since ribosomes exist only in the cytoplasm. Many more steps, especially in RNA processing, are involved in the expression of eukaryotic genes than of prokaryotic genes, and these steps provide additional sites for regulatory influences that cannot exist in

prokaryotes. These RNA processing steps in eukaryotes, described in detail in Chapter 37, include capping of the 5' ends of the primary transcripts, addition of a polyadenylate tail to the 3' ends of transcripts, and excision of intron regions to generate spliced exons in the mature mRNA molecule. To date, analyses of eukaryotic gene expression provide evidence that regulation occurs at the level of **transcription**, **nuclear RNA processing**, and **mRNA stability**. In addition, gene amplification and rearrangement influence gene expression.

Owing to the advent of recombinant DNA technology, much progress has been made in recent years in the understanding of eukaryotic gene expression. However, because most eukaryotic organisms contain so much more genetic information than do prokaryotes and because manipulation of their genes is so much more limited, molecular aspects of eukaryotic gene regulation are less well understood than the examples discussed earlier in this chapter. This section briefly describes a few different types of eukaryotic gene regulation.

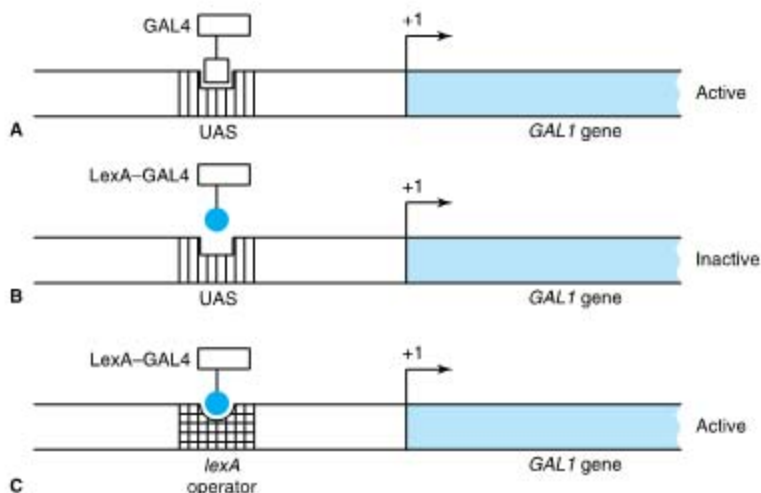


Figure 39-16. Domain-swap experiments demonstrate the independent nature of DNA binding and transcription activation domains. The *GAL1* gene promoter contains an upstream activating sequence (UAS) or enhancer that binds the regulatory protein GAL4 (A). This interaction results in a stimulation of *GAL1* gene transcription. A chimeric protein, in which the amino terminal DNA binding domain of GAL4 is removed and replaced with the DNA binding region of the *E. coli* protein LexA, fails to stimulate *GAL1* transcription because the LexA domain cannot bind to the UAS (B). The LexA-GAL4 fusion protein does increase *GAL1* transcription when the *lexA* operator (its natural target) is inserted into the *GAL1* promoter region (C).

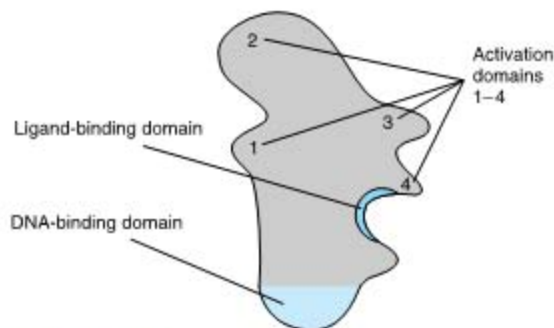


Figure 39-17. Proteins that regulate transcription have several domains. This hypothetical transcription factor has a DNA-binding domain (DBD) that is distinct from a ligand-binding domain (LBD) and several activation domains (ADs) (1-4). Other proteins may lack the DBD or LBD and all may have variable numbers of domains that contact other proteins, including co-regulators and those of the basal transcription complex (see also Chapters 42 and 43).

Eukaryotic Genes Can Be Amplified or Rearranged During Development or in Response to Drugs

During early development of metazoans, there is an abrupt increase in the need for specific molecules such as ribosomal RNA and messenger RNA molecules for proteins that make up such organs as the eggshell. One way to increase the rate at which such molecules can be formed is to increase the number of genes available for transcription of these specific molecules. Among the repetitive DNA sequences are hundreds of copies of ribosomal RNA genes and tRNA genes. These genes pre-exist repetitively in the genomic material of the gametes

Table 39-4. Gene expression is regulated by transcription and in numerous other ways in eukaryotic cells.

- | |
|---|
| • Gene amplification |
| • Gene rearrangement |
| • RNA processing |
| • Alternate mRNA splicing |
| • Transport of mRNA from nucleus to cytoplasm |
| • Regulation of mRNA stability |

and thus are transmitted in high copy numbers from generation to generation. In some specific organisms such as the fruit fly (*Drosophila*), there occurs during oogenesis an amplification of a few preexisting genes such as those for the chorion (eggshell) proteins. Subsequently, these amplified genes, presumably generated by a process of repeated initiations during DNA synthesis, provide multiple sites for gene transcription (Figures 36–16 and 39–18).

As noted in Chapter 37, the coding sequences responsible for the generation of specific protein molecules are frequently not contiguous in the mammalian genome. In the case of antibody encoding genes, this is particularly true. As described in detail in Chapter 50, immunoglobulins are composed of two polypeptides, the so-called heavy (about 50 kDa) and light (about 25 kDa) chains. The mRNAs encoding these two protein subunits are encoded by gene sequences that are subjected to extensive DNA sequence-coding changes. These DNA coding changes are integral to generating the requisite recognition diversity central to appropriate immune function.

IgG heavy and light chain mRNAs are encoded by several different segments that are tandemly repeated in the germline. Thus, for example, the IgG light chain is composed of variable (V_L), joining (J_L), and constant (C_L) domains or segments. For particular subsets of IgG light chains, there are roughly 300 tandemly repeated V_L gene coding segments, five tandemly arranged J_L coding sequences, and roughly ten C_L gene coding segments. All of these multiple, distinct coding regions are located in the same region of the same chromosome, and each type of coding segment (V_L , J_L , and C_L) is tandemly repeated in head-to-tail fashion within the segment repeat region. By having multiple V_L , J_L , and C_L segments to choose from, an immune cell has a greater repertoire of sequences to work with to develop

both immunologic flexibility and specificity. However, a given functional IgG light chain transcription unit—like all other “normal” mammalian transcription units—contains only the coding sequences for a single protein. Thus, before a particular IgG light chain can be expressed, *single* V_L , J_L , and C_L coding sequences must be recombined to generate a *single*, contiguous transcription unit excluding the multiple nonutilized segments (i.e., the other approximately 300 unused V_L segments, the other four unused J_L segments, and the other nine unused C_L segments). This deletion of unused genetic information is accomplished by selective DNA recombination that removes the unwanted coding DNA while retaining the required coding sequences: one V_L , one J_L , and one C_L sequence. (V_L sequences are subjected to additional point mutagenesis to generate even more variability—hence the name.) The newly recombined sequences thus form a single transcription unit that is competent for RNA polymerase II-mediated transcription. Although the IgG genes represent one of the best-studied instances of directed DNA rearrangement modulating gene expression, other cases of gene regulatory DNA rearrangement have been described in the literature. Indeed, as detailed below, drug-induced gene amplification is an important complication of cancer chemotherapy.

In recent years, it has been possible to promote the amplification of specific genetic regions in cultured mammalian cells. In some cases, a several thousand-fold increase in the copy number of specific genes can be achieved over a period of time involving increasing doses of selective drugs. In fact, it has been demonstrated in patients receiving methotrexate for cancer that malignant cells can develop **drug resistance** by increasing the number of genes for dihydrofolate reductase, the target of methotrexate. Gene amplification events such as these occur spontaneously *in vivo*—i.e., in the absence of exogenously supplied selective agents—and these unscheduled extra rounds of replication can become “frozen” in the genome under appropriate selective pressures.

Alternative RNA Processing Is Another Control Mechanism

In addition to affecting the efficiency of promoter utilization, eukaryotic cells employ alternative RNA processing to control gene expression. This can result when alternative promoters, intron-exon splice sites, or polyadenylation sites are used. Occasionally, heterogeneity within a cell results, but more commonly the same primary transcript is processed differently in different tissues. A few examples of each of these types of regulation are presented below.

The use of alternative **transcription start sites** results in a different 5' exon on mRNAs corresponding to

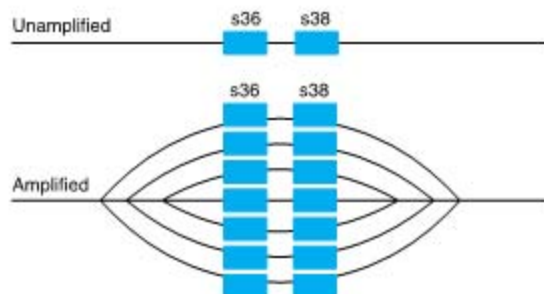


Figure 39–18. Schematic representation of the amplification of chorion protein genes s36 and s38. (Reproduced, with permission, from Chisholm R: Gene amplification during development. *Trends Biochem Sci* 1982;7:161.)

mouse amylase and myosin light chain, rat glucokinase, and drosophila alcohol dehydrogenase and actin. **Alternative polyadenylation sites** in the μ immunoglobulin heavy chain primary transcript result in mRNAs that are either 2700 bases long (μ_m) or 2400 bases long (μ_s). This results in a different carboxyl terminal region of the encoded proteins such that the μ_m protein remains attached to the membrane of the B lymphocyte and the μ_s immunoglobulin is secreted. **Alternative splicing and processing** results in the formation of seven unique α -tropomyosin mRNAs in seven different tissues. It is not clear how these processing-splicing decisions are made or whether these steps can be regulated.

Regulation of Messenger RNA Stability Provides Another Control Mechanism

Although most mRNAs in mammalian cells are very stable (half-lives measured in hours), some turn over very rapidly (half-lives of 10–30 minutes). In certain instances, mRNA stability is subject to regulation. This has important implications since there is usually a direct relationship between mRNA amount and the translation of that mRNA into its cognate protein. Changes in the stability of a specific mRNA can therefore have major effects on biologic processes.

Messenger RNAs exist in the cytoplasm as ribonucleoprotein particles (RNPs). Some of these proteins protect the mRNA from digestion by nucleases, while others may under certain conditions promote nuclease attack. It is thought that mRNAs are stabilized or destabilized by the interaction of proteins with these various structures or sequences. Certain effectors, such as hormones, may regulate mRNA stability by increasing or decreasing the amount of these proteins.

It appears that **the ends of mRNA molecules are involved in mRNA stability** (Figure 39–19). The 5'

cap structure in eukaryotic mRNA prevents attack by 5' exonucleases, and the poly(A) tail prohibits the action of 3' exonucleases. In mRNA molecules with those structures, it is presumed that a single endonucleolytic cut allows exonucleases to attack and digest the entire molecule. Other structures (sequences) in the 5' non-coding sequence (5' NCS), the coding region, and the 3' NCS are thought to promote or prevent this initial endonucleolytic action (Figure 39–19). A few illustrative examples will be cited.

Deletion of the 5' NCS results in a threefold to fivefold prolongation of the half-life of *c-myc* mRNA. Shortening the coding region of histone mRNA results in a prolonged half-life. A form of autoregulation of mRNA stability indirectly involves the coding region. Free tubulin binds to the first four amino acids of a nascent chain of tubulin as it emerges from the ribosome. This appears to activate an RNase associated with the ribosome (RNP) which then digests the tubulin mRNA.

Structures at the 3' end, including the poly(A) tail, enhance or diminish the stability of specific mRNAs. The absence of a poly(A) tail is associated with rapid degradation of mRNA, and the removal of poly(A) from some RNAs results in their destabilization. Histone mRNAs lack a poly(A) tail but have a sequence near the 3' terminal that can form a stem-loop structure, and this appears to provide resistance to exonucleolytic attack. Histone H4 mRNA, for example, is degraded in the 3' to 5' direction but only after a single endonucleolytic cut occurs about nine nucleotides from the 3' end in the region of the putative stem-loop structure. Stem-loop structures in the 3' noncoding sequence are also critical for the regulation, by iron, of the mRNA encoding the transferrin receptor. Stem-loop structures are also associated with mRNA stability in bacteria, suggesting that this mechanism may be commonly employed.

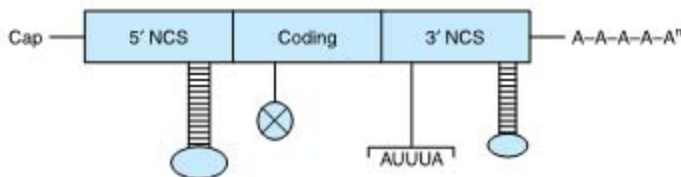


Figure 39–19. Structure of a typical eukaryotic mRNA showing elements that are involved in regulating mRNA stability. The typical eukaryotic mRNA has a 5' noncoding sequence (5' NCS), a coding region, and a 3' NCS. All are capped at the 5' end, and most have a polyadenylate sequence at the 3' end. The 5' cap and 3' poly(A) tail protect the mRNA against exonuclease attack. Stem-loop structures in the 5' and 3' NCS, features in the coding sequence, and the AU-rich region in the 3' NCS are thought to play roles in mRNA stability.

Other sequences in the 3' ends of certain eukaryotic mRNAs appear to be involved in the destabilization of these molecules. Of particular interest are AU-rich regions, many of which contain the sequence AUUUA. This sequence appears in mRNAs that have a very short half-life, including some encoding oncogene proteins and cytokines. The importance of this region is underscored by an experiment in which a sequence corresponding to the 3' noncoding region of the short-half-life colony-stimulating factor (CSF) mRNA, which contains the AUUUA motif, was added to the 3' end of the β -globin mRNA. Instead of becoming very stable, this hybrid β -globin mRNA now had the short-half-life characteristic of CSF mRNA.

From the few examples cited, it is clear that a number of mechanisms are used to regulate mRNA stability—just as several mechanisms are used to regulate the synthesis of mRNA. Coordinate regulation of these two processes confers on the cell remarkable adaptability.

SUMMARY

- The genetic constitutions of nearly all metazoan somatic cells are identical.
- Phenotype (tissue or cell specificity) is dictated by differences in gene expression of this complement of genes.
- Alterations in gene expression allow a cell to adapt to environmental changes.
- Gene expression can be controlled at multiple levels by changes in transcription, RNA processing, localization, and stability or utilization. Gene amplification and rearrangements also influence gene expression.
- Transcription controls operate at the level of protein-DNA and protein-protein interactions. These interactions display protein domain modularity and high specificity.
- Several different classes of DNA-binding domains have been identified in transcription factors.
- Chromatin modifications are important in eukaryotic transcription control.

REFERENCES

- Albright SR, Tjian R: TAFs revisited: more data reveal new twists and confirm old ideas. *Gene* 2000;242:1.
- Bird AP, Wolffe AP: Methylation-induced repression—belts, braces and chromatin. *Cell* 1999;99:451.
- Berger SL, Felsenfeld G: Chromatin goes global. *Mol Cell* 2001; 8:263.
- Busby S, Ebright RH: Promoter structure, promoter recognition, and transcription activation in prokaryotes. *Cell* 1994;79: 743.
- Busby S, Ebright RH: Transcription activation by catabolite activator protein (CAP). *J Mol Biol* 1999;293:199.
- Cowell IG: Repression versus activation in the control of gene transcription. *Trends Biochem Sci* 1994;19:38.
- Ebright RH: RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J Mol Biol* 2000;304:687.
- Fugman SD: RAG1 and RAG2 in V(D)J recombination and transposition. *Immunol Res* 2001;23:23.
- Jacob F, Monod J: Generic regulatory mechanisms in protein synthesis. *J Mol Biol* 1961;3:318.
- Lemon B, Tjian R: Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* 2000;14:2551.
- Letchman DS: Transcription factor mutations and disease. *N Engl J Med* 1996;334:28.
- Merika M, Thanos D: Enhanceosomes. *Curr Opin Genet Dev* 2001;11:205.
- Naar AM, Lemon BD, Tjian R: Transcriptional coactivator complexes. *Annu Rev Biochem* 2001;70:475.
- Narlikar GJ, Fan HY, Kingston RE: Cooperation between complexes that regulate chromatin structure and transcription. *Cell* 2002;108:475.
- Oltz EM: Regulation of antigen receptor gene assembly in lymphocytes. *Immunol Res* 2001;23:121.
- Ptashne M: Control of gene transcription: an outline. *Nat Med* 1997;3:1069.
- Ptashne M: *A Genetic Switch*, 2nd ed. Cell Press and Blackwell Scientific Publications, 1992.
- Sternier DE, Berger SL: Acetylation of histones and transcription-related factors. *Microbiol Mol Biol Rev* 2000;64:435.
- Wu R, Bahl CP, Narang SA: Lactose operator-repressor interaction. *Curr Top Cell Regul* 1978;13:137.

Molecular Genetics, Recombinant DNA, & Genomic Technology

40

Daryl K. Granner, MD, & P. Anthony Weil, PhD

BIOMEDICAL IMPORTANCE*

The development of recombinant DNA, high-density, high-throughput screening, and other molecular genetic methodologies has revolutionized biology and is having an increasing impact on clinical medicine. Much has been learned about human genetic disease from pedigree analysis and study of affected proteins, but in many cases where the specific genetic defect is unknown, these approaches cannot be used. The new technologies circumvent these limitations by going directly to the DNA molecule for information. Manipulation of a DNA sequence and the construction of chimeric molecules—so-called genetic engineering—provides a means of studying how a specific segment of DNA works. Novel molecular genetic tools allow investigators to query and manipulate genomic sequences as well as to examine both cellular mRNA and protein profiles at the molecular level.

Understanding this technology is important for several reasons: (1) It offers a rational approach to understanding the molecular basis of a number of diseases (eg, familial hypercholesterolemia, sickle cell disease, the thalassemias, cystic fibrosis, muscular dystrophy). (2) Human proteins can be produced in abundance for therapy (eg, insulin, growth hormone, tissue plasminogen activator). (3) Proteins for vaccines (eg, hepatitis B) and for diagnostic testing (eg, AIDS tests) can be obtained. (4) This technology is used to diagnose existing diseases and predict the risk of developing a given disease. (5) Special techniques have led to remarkable advances in forensic medicine. (6) Gene therapy for sickle cell disease, the thalassemias, adenosine deaminase deficiency, and other diseases may be devised.

* See glossary of terms at the end of this chapter.

ELUCIDATION OF THE BASIC FEATURES OF DNA LED TO RECOMBINANT DNA TECHNOLOGY

DNA Is a Complex Biopolymer Organized as a Double Helix

The fundamental organizational element is the sequence of purine (adenine [A] or guanine [G]) and pyrimidine (cytosine [C] or thymine [T]) bases. These bases are attached to the C-1' position of the sugar deoxyribose, and the bases are linked together through joining of the sugar moieties at their 3' and 5' positions via a phosphodiester bond (Figure 35-1). The alternating deoxyribose and phosphate groups form the backbone of the double helix (Figure 35-2). These 3'-5' linkages also define the orientation of a given strand of the DNA molecule, and, since the two strands run in opposite directions, they are said to be antiparallel.

Base Pairing Is a Fundamental Concept of DNA Structure & Function

Adenine and thymine always pair, by hydrogen bonding, as do guanine and cytosine (Figure 35-3). These base pairs are said to be **complementary**, and the guanine content of a fragment of double-stranded DNA will always equal its cytosine content; likewise, the thymine and adenine contents are equal. Base pairing and hydrophobic base-stacking interactions hold the two DNA strands together. These interactions can be reduced by heating the DNA to denature it. The laws of base pairing predict that two complementary DNA strands will reanneal exactly in register upon renaturation, as happens when the temperature of the solution is slowly reduced to normal. Indeed, the degree of base-pair matching (or mismatching) can be estimated from the temperature re-

quired for denaturation-renaturation. Segments of DNA with high degrees of base-pair matching require more energy input (heat) to accomplish denaturation—or, to put it another way, a closely matched segment will withstand more heat before the strands separate. This reaction is used to determine whether there are significant differences between two DNA sequences, and it underlies the concept of **hybridization**, which is fundamental to the processes described below.

There are about 3×10^9 base pairs (bp) in each human haploid genome. If an average gene length is 3×10^3 bp (3 kilobases [kb]), the genome could consist of 10^6 genes, assuming that there is no overlap and that transcription proceeds in only one direction. It is thought that there are $< 10^5$ genes in the human and that only 1–2% of the DNA codes for proteins. The exact function of the remaining ~98% of the human genome has not yet been defined.

The double-helical DNA is packaged into a more compact structure by a number of proteins, most notably the basic proteins called histones. This condensation may serve a regulatory role and certainly has a practical purpose. The DNA present within the nucleus of a cell, if simply extended, would be about 1 meter long. The chromosomal proteins compact this long strand of DNA so that it can be packaged into a nucleus with a volume of a few cubic micrometers.

DNA Is Organized Into Genes

In general, prokaryotic genes consist of a small regulatory region (100–500 bp) and a large protein-coding segment (500–10,000 bp). Several genes are often controlled by a single regulatory unit. Most mammalian genes are more complicated in that the coding regions are interrupted by noncoding regions that are eliminated when the primary RNA transcript is processed into mature **messenger RNA (mRNA)**. The **coding regions** (those regions that appear in the mature RNA species) are called **exons**, and the **noncoding regions**, which interpose or intervene between the exons, are called **introns** (Figure 40–1). Introns are always removed from precursor RNA before transport into the cytoplasm occurs. The process by which introns are removed from precursor RNA and by which exons are ligated together is called **RNA splicing**. Incorrect processing of the primary transcript into the mature mRNA can result in disease in humans (see below); this underscores the importance of these posttranscriptional processing steps. The variation in size and complexity of some human genes is illustrated in Table 40–1. Although there is a 300-fold difference in the sizes of the genes illustrated, the mRNA sizes vary only about 20-fold. This is because most of the DNA in genes is present as introns, and introns tend to be much larger than

exons. Regulatory regions for specific eukaryotic genes are usually located in the DNA that flanks the transcription initiation site at its 5' end (**5' flanking-sequence DNA**). Occasionally, such sequences are found within the gene itself or in the region that flanks the 3' end of the gene. In mammalian cells, each gene has its own regulatory region. Many eukaryotic genes (and some viruses that replicate in mammalian cells) have special regions, called **enhancers**, that increase the rate of transcription. Some genes also have DNA sequences, known as **silencers**, that repress transcription. Mammalian genes are obviously complicated, multi-component structures.

Genes Are Transcribed Into RNA

Information generally flows from DNA to mRNA to protein, as illustrated in Figure 40–1 and discussed in more detail in Chapter 39. This is a rigidly controlled process involving a number of complex steps, each of which no doubt is regulated by one or more enzymes or factors; faulty function at any of these steps can cause disease.

RECOMBINANT DNA TECHNOLOGY INVOLVES ISOLATION & MANIPULATION OF DNA TO MAKE CHIMERIC MOLECULES

Isolation and manipulation of DNA, including end-to-end joining of sequences from very different sources to make chimeric molecules (eg, molecules containing both human and bacterial DNA sequences in a sequence-independent fashion), is the essence of recombinant DNA research. This involves several unique techniques and reagents.

Restriction Enzymes Cut DNA Chains at Specific Locations

Certain endonucleases—enzymes that cut DNA at specific DNA sequences within the molecule (as opposed to exonucleases, which digest from the ends of DNA molecules)—are a key tool in recombinant DNA research. These enzymes were called **restriction enzymes** because their presence in a given bacterium restricted the growth of certain bacterial viruses called bacteriophages. Restriction enzymes cut DNA of any source into short pieces in a sequence-specific manner—in contrast to most other enzymatic, chemical, or physical methods, which break DNA randomly. These defensive enzymes (hundreds have been discovered) protect the host bacterial DNA from DNA from foreign organisms (primarily infective phages). However, they are present only in cells that also have a companion enzyme which methylates the host DNA, rendering it an unsuitable substrate for digestion by the restriction enzyme. Thus,

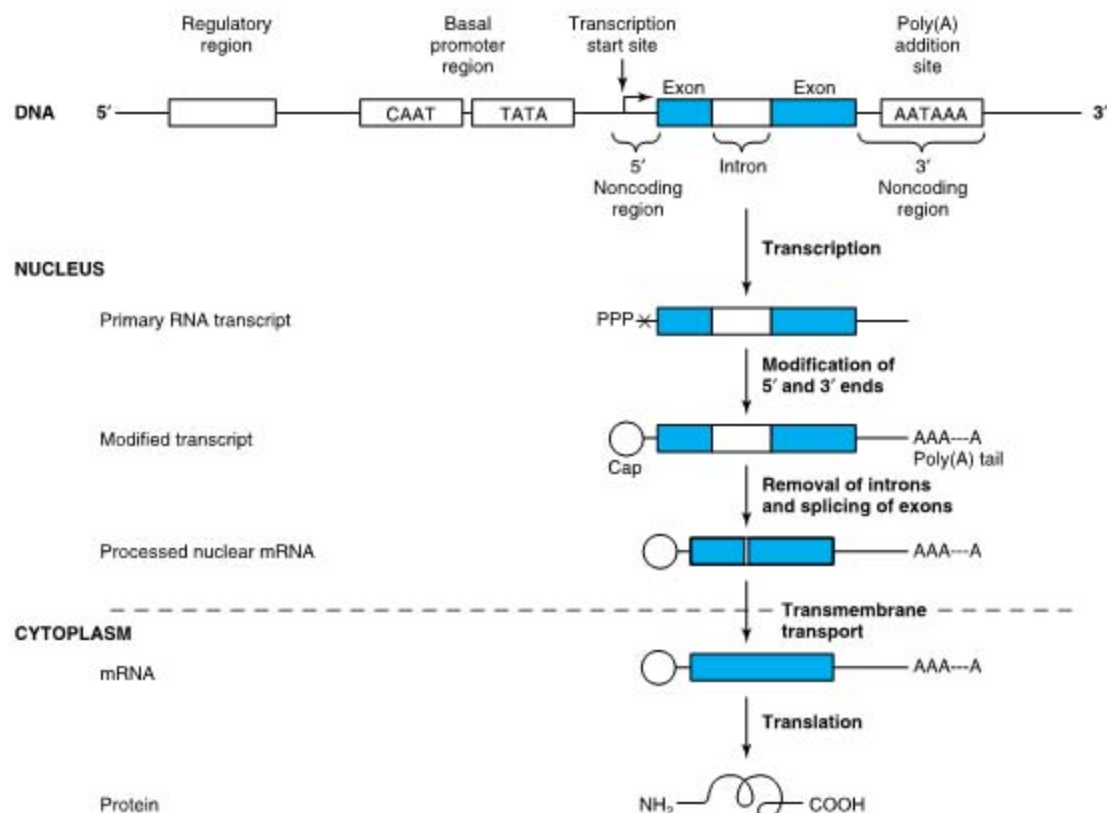


Figure 40-1. Organization of a eukaryotic transcription unit and the pathway of eukaryotic gene expression. Eukaryotic genes have structural and regulatory regions. The structural region consists of the coding DNA and 5' and 3' noncoding DNA sequences. The coding regions are divided into two parts: (1) exons, which eventually are ligated together to become mature RNA, and (2) introns, which are processed out of the primary transcript. The structural region is bounded at its 5' end by the transcription initiation site and at its 3' end by the polyadenylate addition or termination site. The promoter region, which contains specific DNA sequences that interact with various protein factors to regulate transcription, is discussed in detail in Chapters 37 and 39. The primary transcript has a special structure, a cap, at the 5' end and a stretch of As at the 3' end. This transcript is processed to remove the introns; and the mature mRNA is then transported to the cytoplasm, where it is translated into protein.

site-specific DNA methylases and restriction enzymes always exist in pairs in a bacterium.

Restriction enzymes are named after the bacterium from which they are isolated. For example, *EcoRI* is from *Escherichia coli*, and *BamHI* is from *Bacillus amyloliquefaciens* (Table 40-2). The first three letters in the restriction enzyme name consist of the first letter of the genus (E) and the first two letters of the species (co). These may be followed by a strain designation (R) and a roman numeral (I) to indicate the order of discovery (eg, *EcoRI*, *EcoRII*). Each enzyme recognizes and cleaves a specific double-stranded DNA sequence that is 4–7 bp long. These DNA cuts result in **blunt ends** (eg,

HpaI) or overlapping (**sticky ends**) (eg, *BamHI*) (Figure 40-2), depending on the mechanism used by the enzyme. Sticky ends are particularly useful in constructing hybrid or chimeric DNA molecules (see below). If the four nucleotides are distributed randomly in a given DNA molecule, one can calculate how frequently a given enzyme will cut a length of DNA. For each position in the DNA molecule, there are four possibilities (A, C, G, and T); therefore, a restriction enzyme that recognizes a 4-bp sequence cuts, on average, once every 256 bp (4^4), whereas another enzyme that recognizes a 6-bp sequence cuts once every 4096 bp (4^6). A given piece of DNA has a characteristic linear array of sites for

Table 40-1. Variations in the size and complexity of some human genes and mRNAs.¹

Gene	Gene Size (kb)	Number of Introns	mRNA Size (kb)
β -Globin	1.5	2	0.6
Insulin	1.7	2	0.4
β -Adrenergic receptor	3	0	2.2
Albumin	25	14	2.1
LDL receptor	45	17	5.5
Factor VIII	186	25	9.0
Thyroglobulin	300	36	8.7

¹The sizes are given in kilobases (kb). The sizes of the genes include some proximal promoter and regulatory region sequences; these are generally about the same size for all genes. Genes vary in size from about 1500 base pairs (bp) to over 2×10^6 bp. There is also great variation in the number of introns and exons. The β -adrenergic receptor gene is intronless, and the thyroglobulin gene has 36 introns. As noted by the smaller difference in mRNA sizes, introns comprise most of the gene sequence.

the various enzymes dictated by the linear sequence of its bases; hence, a **restriction map** can be constructed. When DNA is digested with a given enzyme, the ends of all the fragments have the same DNA sequence. The fragments produced can be isolated by electrophoresis on agarose or polyacrylamide gels (see the discussion of blot transfer, below); this is an essential step in cloning and a major use of these enzymes.

A number of other enzymes that act on DNA and RNA are an important part of recombinant DNA technology. Many of these are referred to in this and subsequent chapters (Table 40-3).

Restriction Enzymes & DNA Ligase Are Used to Prepare Chimeric DNA Molecules

Sticky-end ligation is technically easy, but some special techniques are often required to overcome problems inherent in this approach. Sticky ends of a vector may reconnect with themselves, with no net gain of DNA. Sticky ends of fragments can also anneal, so that tandem heterogeneous inserts form. Also, sticky-end sites may not be available or in a convenient position. To circumvent these problems, an enzyme that generates blunt ends is used, and new ends are added using the enzyme terminal transferase. If poly d(G) is added to the 3' ends of the vector and poly d(C) is added to the 3' ends of the foreign DNA, the two molecules can only anneal to each other, thus circumventing the problems listed above. This procedure is called homopolymer tailing. Sometimes, synthetic blunt-ended duplex oligonucleotide linkers with a convenient restriction enzyme se-

Table 40-2. Selected restriction endonucleases and their sequence specificities.¹

Endonuclease	Sequence Recognized Cleavage Sites Shown	Bacterial Source
<i>Bam</i> HI	↓ GGATCC CCTAGG ↑	<i>Bacillus amyloliquefaciens</i> H
<i>Bgl</i> II	↓ AGATCT TCTAGA ↑	<i>Bacillus globigii</i>
<i>Eco</i> RI	↓ GAATTC CTTAAG ↑	<i>Escherichia coli</i> RY13
<i>Eco</i> RII	↓ CCTGG GGACC ↑	<i>Escherichia coli</i> R245
<i>Hind</i> III	↓ AAGCTT TTCGAA ↑	<i>Haemophilus influenzae</i> R ₆
<i>Hha</i> I	↓ GCGC CGCG ↑	<i>Haemophilus haemolyticus</i>
<i>Hpa</i> I	↓ GTTAAC CAATTG ↑	<i>Haemophilus parainfluenzae</i>
<i>Mst</i> II	↓ CCTNAGG GGAN ¹ TCC ↑	Microcoleus strain
<i>Pst</i> I	↓ CTGCAG GACGTC ↑	<i>Providencia stuartii</i> 164
<i>Taq</i> I	↓ TCGA AGCT ↑	<i>Thermus aquaticus</i> YTI

¹A, adenine; C, cytosine; G, guanine, T, thymine. Arrows show the site of cleavage; depending on the site, sticky ends (*Bam*HI) or blunt ends (*Hpa*I) may result. The length of the recognition sequence can be 4 bp (*Taq*I), 5 bp (*Eco*RII), 6 bp (*Eco*RI), or 7 bp (*Mst*II) or longer. By convention, these are written in the 5' or 3' direction for the upper strand of each recognition sequence, and the lower strand is shown with the opposite (ie, 3' or 5') polarity. Note that most recognition sequences are palindromes (ie, the sequence reads the same in opposite directions on the two strands). A residue designated N means that any nucleotide is permitted.

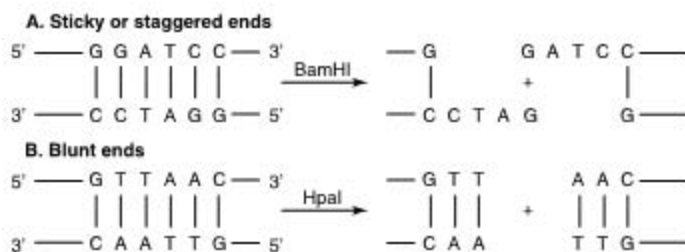


Figure 40-2. Results of restriction endonuclease digestion. Digestion with a restriction endonuclease can result in the formation of DNA fragments with sticky, or cohesive, ends (**A**) or blunt ends (**B**). This is an important consideration in devising cloning strategies.

quence are ligated to the blunt-ended DNA. Direct blunt-end ligation is accomplished using the enzyme bacteriophage T4 DNA ligase. This technique, though less efficient than sticky-end ligation, has the advantage of joining together any pairs of ends. The disadvantages are that there is no control over the orientation of insertion or the number of molecules annealed together, and there is no easy way to retrieve the insert.

Cloning Amplifies DNA

A **clone** is a large population of identical molecules, bacteria, or cells that arise from a common ancestor. Molecular cloning allows for the production of a large number of identical DNA molecules, which can then be charac-

terized or used for other purposes. This technique is based on the fact that chimeric or hybrid DNA molecules can be constructed in **cloning vectors**—typically bacterial plasmids, phages, or cosmids—which then continue to replicate in a host cell under their own control systems. In this way, the chimeric DNA is amplified. The general procedure is illustrated in Figure 40-3.

Bacterial **plasmids** are small, circular, duplex DNA molecules whose natural function is to confer antibiotic resistance to the host cell. Plasmids have several properties that make them extremely useful as cloning vectors. They exist as single or multiple copies within the bacterium and replicate independently from the bacterial DNA. The complete DNA sequence of many plasmids is known; hence, the precise location of restriction enzyme

Table 40-3. Some of the enzymes used in recombinant DNA research.¹

Enzyme	Reaction	Primary Use
Alkaline phosphatase	Dephosphorylates 5' ends of RNA and DNA.	Removal of 5'-PO ₄ groups prior to kinase labeling to prevent self-ligation.
BAL 31 nuclease	Degrades both the 3' and 5' ends of DNA.	Progressive shortening of DNA molecules.
DNA ligase	Catalyzes bonds between DNA molecules.	Joining of DNA molecules.
DNA polymerase I	Synthesizes double-stranded DNA from single-stranded DNA.	Synthesis of double-stranded cDNA; nick translation; generation of blunt ends from sticky ends.
DNase I	Under appropriate conditions, produces single-stranded nicks in DNA.	Nick translation; mapping of hypersensitive sites; mapping protein-DNA interactions.
Exonuclease III	Removes nucleotides from 3' ends of DNA.	DNA sequencing; mapping of DNA-protein interactions.
λ exonuclease	Removes nucleotides from 5' ends of DNA.	DNA sequencing.
Polynucleotide kinase	Transfers terminal phosphate (γ position) from ATP to 5'-OH groups of DNA or RNA.	³² P labeling of DNA or RNA.
Reverse transcriptase	Synthesizes DNA from RNA template.	Synthesis of cDNA from mRNA; RNA (5' end) mapping studies.
S1 nuclease	Degrades single-stranded DNA.	Removal of "hairpin" in synthesis of cDNA; RNA mapping studies (both 5' and 3' ends).
Terminal transferase	Adds nucleotides to the 3' ends of DNA.	Homopolymer tailing.

¹Adapted and reproduced, with permission, from Emery AEH: Page 41 in: *An Introduction to Recombinant DNA*. Wiley, 1984.

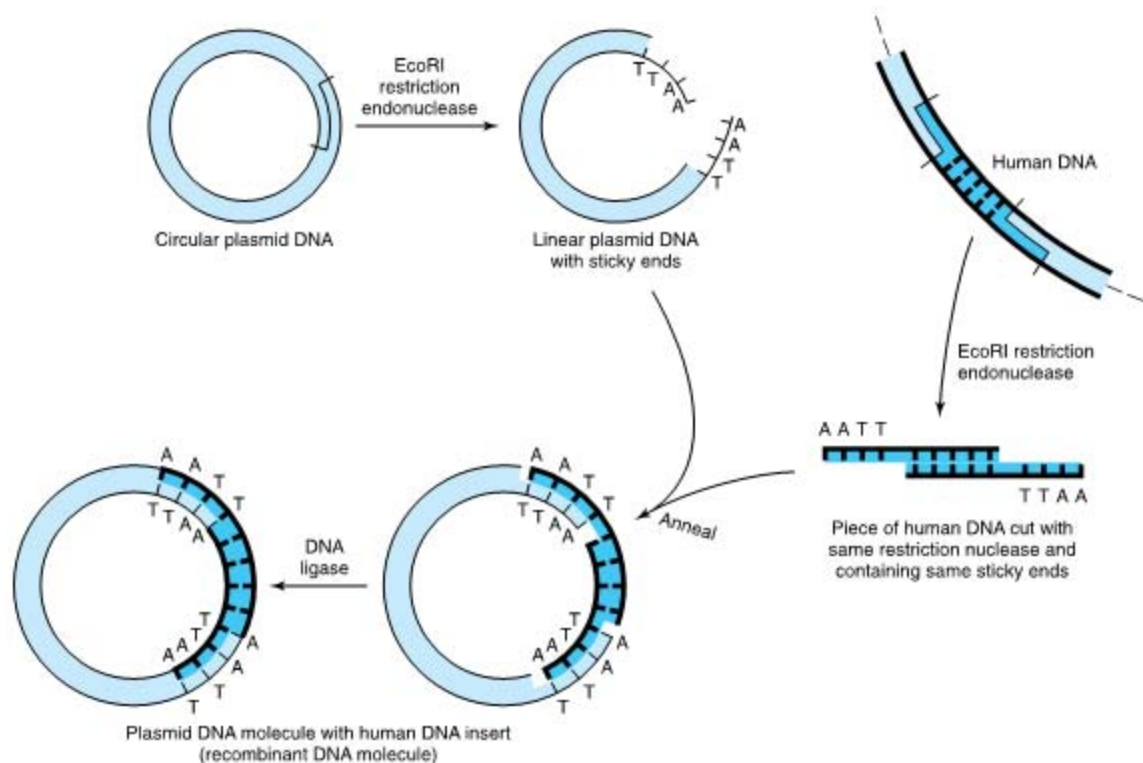


Figure 40-3. Use of restriction nucleases to make new recombinant or chimeric DNA molecules. When inserted back into a bacterial cell (by the process called transformation), typically only a single plasmid is taken up by a single cell, and the plasmid DNA replicates not only itself but also the physically linked new DNA insert. Since recombining the sticky ends, as indicated, regenerates the same DNA sequence recognized by the original restriction enzyme, the cloned DNA insert can be cleanly cut back out of the recombinant plasmid circle with this endonuclease. If a mixture of all of the DNA pieces created by treatment of total human DNA with a single restriction nuclease is used as the source of human DNA, a million or so different types of recombinant DNA molecules can be obtained, each pure in its own bacterial clone. (Modified and reproduced, with permission, from Cohen SN: The manipulation of genes. *Sci Am* [July] 1975;233:34.)

cleavage sites for inserting the foreign DNA is available. Plasmids are smaller than the host chromosome and are therefore easily separated from the latter, and the desired plasmid-inserted DNA is readily removed by cutting the plasmid with the enzyme specific for the restriction site into which the original piece of DNA was inserted.

Phages usually have linear DNA molecules into which foreign DNA can be inserted at several restriction enzyme sites. The chimeric DNA is collected after the phage proceeds through its lytic cycle and produces mature, infective phage particles. A major advantage of phage vectors is that while plasmids accept DNA pieces about 6–10 kb long, phages can accept DNA fragments 10–20 kb long, a limitation imposed by the amount of DNA that can be packed into the phage head.

Larger fragments of DNA can be cloned in **cosmids**, which combine the best features of plasmids and phages. Cosmids are plasmids that contain the DNA sequences, so-called **cos sites**, required for packaging lambda DNA into the phage particle. These vectors grow in the plasmid form in bacteria, but since much of the unnecessary lambda DNA has been removed, more chimeric DNA can be packaged into the particle head. It is not unusual for cosmids to carry inserts of chimeric DNA that are 35–50 kb long. Even larger pieces of DNA can be incorporated into bacterial artificial chromosome (BAC), yeast artificial chromosome (YAC), or *E. coli* bacteriophage P1-based (PAC) vectors. These vectors will accept and propagate DNA inserts of several hundred kilobases or more and have largely re-

Table 40–4. Cloning capacities of common cloning vectors.

Vector	DNA Insert Size
Plasmid pBR322	0.01–10 kb
Lambda charon 4A	10–20 kb
Cosmids	35–50 kb
BAC, P1	50–250 kb
YAC	500–3000 kb

placed the plasmid, phage, and cosmid vectors for some cloning and gene mapping applications. A comparison of these vectors is shown in Table 40–4.

Because insertion of DNA into a functional region of the vector will interfere with the action of this region, care must be taken not to interrupt an essential function of the vector. This concept can be exploited, however, to provide a selection technique. For example, the common plasmid vector **pBR322** has both **tetracycline (tet)** and **ampicillin (amp)** resistance genes. A single *Pst*I restriction enzyme site within the amp resistance gene is commonly used as the insertion site for a piece of foreign DNA. In addition to having sticky ends (Table 40–2 and Figure 40–2), the DNA inserted at this site disrupts the amp resistance gene and makes the bacterium carrying this plasmid amp-sensitive (Figure 40–4). Thus, the parental plasmid, which provides resistance to both antibiotics, can be readily separated from the chimeric plasmid, which is resistant only to tetracycline. YACs contain replication and segregation functions that work in both bacteria and yeast cells and therefore can be propagated in either organism.

In addition to the vectors described in Table 40–4 that are designed primarily for propagation in bacterial cells, vectors for mammalian cell propagation and insert gene (cDNA)/protein expression have also been developed. These vectors are all based upon various eukaryotic viruses that are composed of RNA or DNA genomes. Notable examples of such viral vectors are those utilizing adenoviral (DNA-based) and retroviral (RNA-based) genomes. Though somewhat limited in the size of DNA sequences that can be inserted, such mammalian viral cloning vectors make up for this shortcoming because they will efficiently infect a wide range of different cell types. For this reason, various mammalian viral vectors are being investigated for use in gene therapy experiments.

A Library Is a Collection of Recombinant Clones

The combination of restriction enzymes and various cloning vectors allows the entire genome of an organism to be packed into a vector. A collection of these dif-

ferent recombinant clones is called a library. A **genomic library** is prepared from the total DNA of a cell line or tissue. A **cDNA library** comprises complementary DNA copies of the population of mRNAs in a tissue. Genomic DNA libraries are often prepared by performing partial digestion of total DNA with a restriction enzyme that cuts DNA frequently (eg, a four base cutter such as *Taq*I). The idea is to generate rather large fragments so that most genes will be left intact. The BAC, YAC, and P1 vectors are preferred since they can accept very large fragments of DNA and thus offer a better chance of isolating an intact gene on a single DNA fragment.

A vector in which the protein coded by the gene introduced by recombinant DNA technology is actually synthesized is known as an **expression vector**. Such vectors are now commonly used to detect specific cDNA molecules in libraries and to produce proteins by genetic engineering techniques. These vectors are specially constructed to contain very active inducible promoters, proper in-phase translation initiation codons, both transcription and translation termination signals, and appropriate protein processing signals, if needed. Some expression vectors even contain genes that code for protease inhibitors, so that the final yield of product is enhanced.

Probes Search Libraries for Specific Genes or cDNA Molecules

A variety of molecules can be used to “probe” libraries in search of a specific gene or cDNA molecule or to define and quantitate DNA or RNA separated by electrophoresis through various gels. Probes are generally pieces of DNA or RNA labeled with a ^{32}P -containing nucleotide—or fluorescently labeled nucleotides (more commonly now). Importantly, neither modification (^{32}P or fluorescent-label) affects the hybridization properties of the resulting labeled nucleic acid probes. The probe must recognize a complementary sequence to be effective. A cDNA synthesized from a specific mRNA can be used to screen either a cDNA library for a longer cDNA or a genomic library for a complementary sequence in the coding region of a gene. A popular technique for finding specific genes entails taking a short amino acid sequence and, employing the codon usage for that species (see Chapter 38), making an oligonucleotide probe that will detect the corresponding DNA fragment in a genomic library. If the sequences match exactly, probes 15–20 nucleotides long will hybridize. cDNA probes are used to detect DNA fragments on Southern blot transfers and to detect and quantitate RNA on Northern blot transfers. Specific antibodies can also be used as probes provided that the vector used synthesizes protein molecules that are recognized by them.

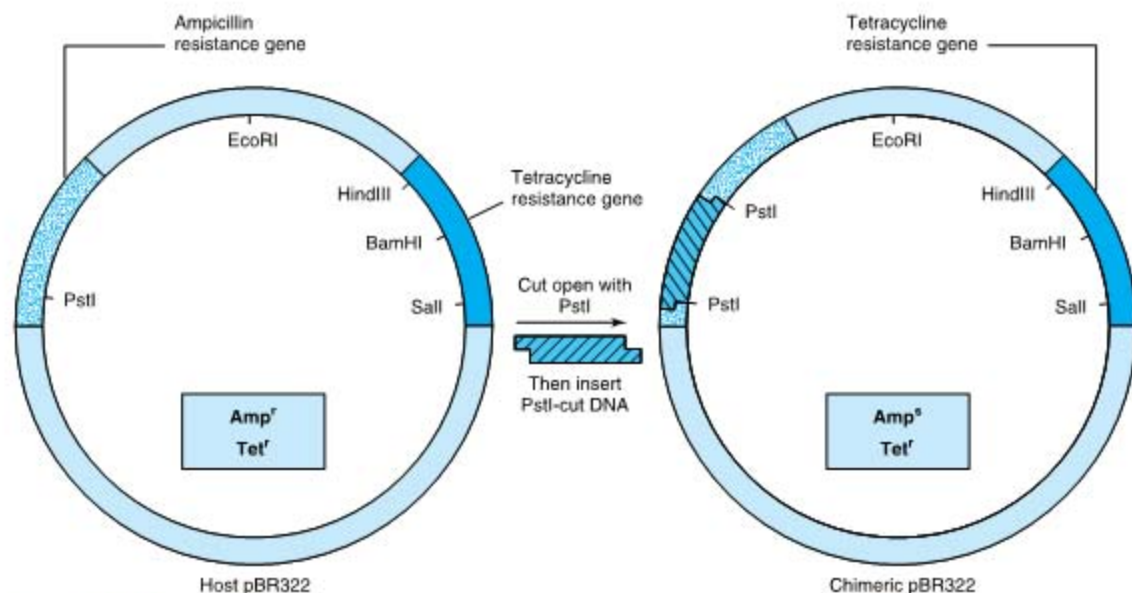


Figure 40-4. A method of screening recombinants for inserted DNA fragments. Using the plasmid pBR322, a piece of DNA is inserted into the unique *PstI* site. This insertion disrupts the gene coding for a protein that provides ampicillin resistance to the host bacterium. Hence, the chimeric plasmid will no longer survive when plated on a substrate medium that contains this antibiotic. The differential sensitivity to tetracycline and ampicillin can therefore be used to distinguish clones of plasmid that contain an insert. A similar scheme relying upon production of an in-frame fusion of a newly inserted DNA producing a peptide fragment capable of complementing an inactive, deleted form of the enzyme β -galactosidase allows for blue-white colony formation on agar plates containing a dye hydrolyzable by β -galactosidase. β -Galactosidase-positive colonies are blue.

Blotting & Hybridization Techniques Allow Visualization of Specific Fragments

Visualization of a specific DNA or RNA fragment among the many thousands of “contaminating” molecules requires the convergence of a number of techniques, collectively termed **blot transfer**. Figure 40-5 illustrates the **Southern** (DNA), **Northern** (RNA), and **Western** (protein) blot transfer procedures. (The first is named for the person who devised the technique, and the other names began as laboratory jargon but are now accepted terms.) These procedures are useful in determining how many copies of a gene are in a given tissue or whether there are any gross alterations in a gene (deletions, insertions, or rearrangements). Occasionally, if a specific base is changed and a restriction site is altered, these procedures can detect a point mutation. The Northern and Western blot transfer techniques are used to size and quantitate specific RNA and protein molecules, respectively. A fourth hybridization technique, the Southwestern blot, examines protein•DNA interactions. Proteins are separated by electrophoresis,

renatured, and analyzed for an interaction by hybridization with a specific labeled DNA probe.

Colony or plaque hybridization is the method by which specific clones are identified and purified. Bacteria are grown on colonies on an agar plate and overlaid with nitrocellulose filter paper. Cells from each colony stick to the filter and are permanently fixed thereto by heat, which with NaOH treatment also lyses the cells and denatures the DNA so that it will hybridize with the probe. A radioactive probe is added to the filter, and (after washing) the hybrid complex is localized by exposing the filter to x-ray film. By matching the spot on the autoradiograph to a colony, the latter can be picked from the plate. A similar strategy is used to identify fragments in phage libraries. Successive rounds of this procedure result in a clonal isolate (bacterial colony) or individual phage plaque.

All of the hybridization procedures discussed in this section depend on the specific base-pairing properties of complementary nucleic acid strands described above. Perfect matches hybridize readily and withstand high temperatures in the hybridization and washing reac-

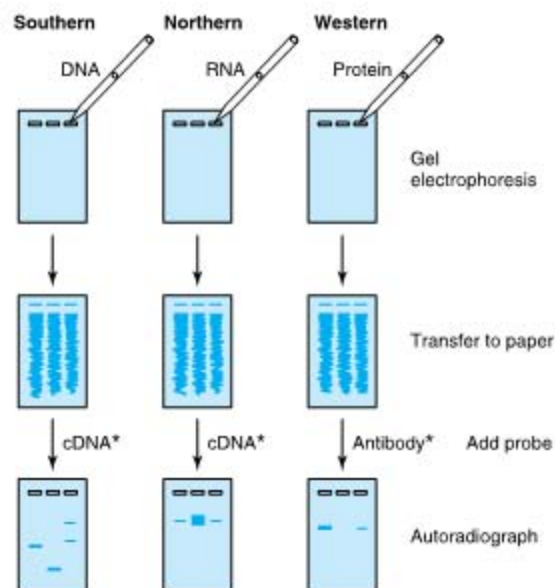


Figure 40-5. The blot transfer procedure. In a Southern, or DNA, blot transfer, DNA isolated from a cell line or tissue is digested with one or more restriction enzymes. This mixture is pipetted into a well in an agarose or polyacrylamide gel and exposed to a direct electrical current. DNA, being negatively charged, migrates toward the anode; the smaller fragments move the most rapidly. After a suitable time, the DNA is denatured by exposure to mild alkali and transferred to nitrocellulose or nylon paper, in an exact replica of the pattern on the gel, by the blotting technique devised by Southern. The DNA is bound to the paper by exposure to heat, and the paper is then exposed to the labeled cDNA probe, which hybridizes to complementary fragments on the filter. After thorough washing, the paper is exposed to x-ray film, which is developed to reveal several specific bands corresponding to the DNA fragment that recognized the sequences in the cDNA probe. The RNA, or Northern, blot is conceptually similar. RNA is subjected to electrophoresis before blot transfer. This requires some different steps from those of DNA transfer, primarily to ensure that the RNA remains intact, and is generally somewhat more difficult. In the protein, or Western, blot, proteins are electrophoresed and transferred to nitrocellulose and then probed with a specific antibody or other probe molecule. (Asterisks signify labeling, either radioactive or fluorescent.)

tions. Specific complexes also form in the presence of low salt concentrations. Less than perfect matches do not tolerate these **stringent conditions** (ie, elevated temperatures and low salt concentrations); thus, hybridization either never occurs or is disrupted during the washing step. Gene families, in which there is some degree of homology, can be detected by varying the stringency of the hybridization and washing steps. Cross-species comparisons of a given gene can also be made using this approach. Hybridization conditions capable of detecting just a single base pair mismatch between probe and target have been devised.

Manual & Automatic Techniques Are Available to Determine the Sequence of DNA

The segments of specific DNA molecules obtained by recombinant DNA technology can be analyzed to determine their nucleotide sequence. This method depends upon having a large number of identical DNA molecules. This requirement can be satisfied by cloning the fragment of interest, using the techniques described above. The **manual enzymatic method (Sanger)** employs specific dideoxynucleotides that terminate DNA strand synthesis at specific nucleotides as the strand is synthesized on purified template nucleic acid. The reactions are adjusted so that a population of DNA fragments representing termination at every nucleotide is obtained. By having a radioactive label incorporated at the end opposite the termination site, one can separate the fragments according to size using polyacrylamide gel electrophoresis. An autoradiograph is made, and each of the fragments produces an image (band) on an x-ray film. These are read in order to give the DNA sequence (Figure 40-6). Another manual method, that of **Maxam and Gilbert**, employs **chemical methods** to cleave the DNA molecules where they contain the specific nucleotides. Techniques that do not require the use of radioisotopes are commonly employed in automated DNA sequencing. Most commonly employed is an automated procedure in which four different fluorescent labels—one representing each nucleotide—are used. Each emits a specific signal upon excitation by a laser beam, and this can be recorded by a computer.

Oligonucleotide Synthesis Is Now Routine

The automated chemical synthesis of moderately long oligonucleotides (about 100 nucleotides) of precise sequence is now a routine laboratory procedure. Each synthetic cycle takes but a few minutes, so an entire molecule can be made by synthesizing relatively short segments that can then be ligated to one another. Oligonucleotides are now indispensable for DNA se-

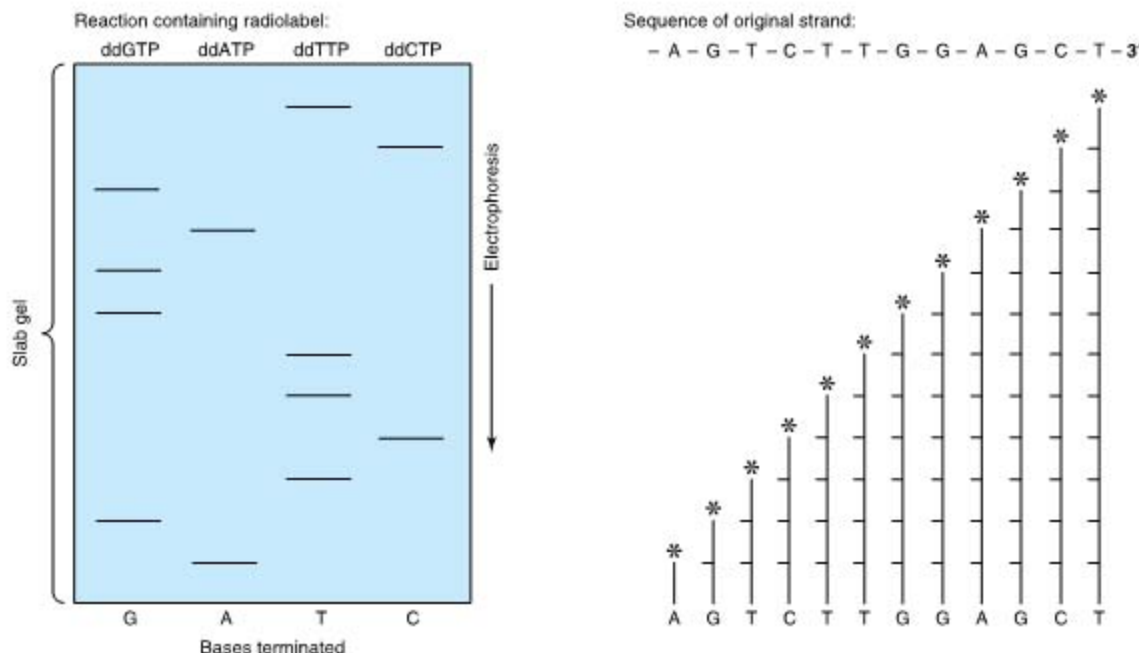


Figure 40-6. Sequencing of DNA by the method devised by Sanger. The ladder-like arrays represent from bottom to top all of the successively longer fragments of the original DNA strand. Knowing which specific dideoxynucleotide reaction was conducted to produce each mixture of fragments, one can determine the sequence of nucleotides from the labeled end (asterisk) toward the unlabeled end by reading up the gel. Automated sequencing involves the reading of chemically modified deoxynucleotides. The base-pairing rules of Watson and Crick (A-T, G-C) dictate the sequence of the other (complementary) strand. (Asterisks signify radiolabeling.)

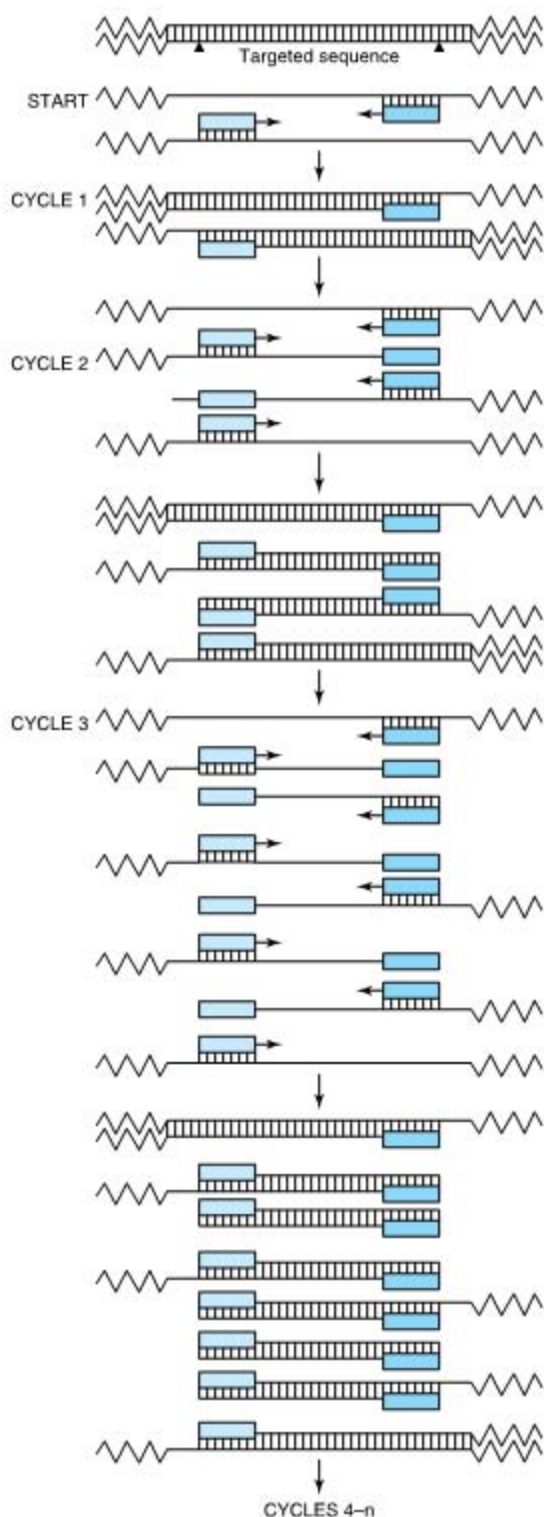
quencing, library screening, protein-DNA binding, DNA mobility shift assays, the polymerase chain reaction (see below), site-directed mutagenesis, and numerous other applications.

The Polymerase Chain Reaction (PCR) Amplifies DNA Sequences

The polymerase chain reaction (PCR) is a method of amplifying a target sequence of DNA. PCR provides a sensitive, selective, and extremely rapid means of amplifying a desired sequence of DNA. Specificity is based on the use of two oligonucleotide primers that hybridize to complementary sequences on opposite strands of DNA and flank the target sequence (Figure 40-7). The DNA sample is first heated to separate the two strands; the primers are allowed to bind to the DNA; and each strand is copied by a DNA polymerase, starting at the primer site. The two DNA strands each serve as a template for the synthesis of new DNA from the two primers. Repeated cycles of heat denaturation, annealing of the primers to their complementary se-

quences, and extension of the annealed primers with DNA polymerase result in the exponential amplification of DNA segments of defined length. Early PCR reactions used an *E. coli* DNA polymerase that was destroyed by each heat denaturation cycle. Substitution of a heat-stable DNA polymerase from *Thermus aquaticus* (or the corresponding DNA polymerase from other thermophilic bacteria), an organism that lives and replicates at 70–80 °C, obviates this problem and has made possible automation of the reaction, since the polymerase reactions can be run at 70 °C. This has also improved the specificity and the yield of DNA.

DNA sequences as short as 50–100 bp and as long as 10 kb can be amplified. Twenty cycles provide an amplification of 10^6 and 30 cycles of 10^9 . The PCR allows the DNA in a single cell, hair follicle, or spermatozoon to be amplified and analyzed. Thus, the applications of PCR to forensic medicine are obvious. The PCR is also used (1) to detect infectious agents, especially latent viruses; (2) to make prenatal genetic diagnoses; (3) to detect allelic polymorphisms; (4) to establish precise tissue types for transplants; and (5) to study



evolution, using DNA from archeological samples after RNA copying and mRNA quantitation by the so-called RT-PCR method (cDNA copies of mRNA generated by a retroviral reverse transcriptase). There are an equal number of applications of PCR to problems in basic science, and new uses are developed every year.

PRACTICAL APPLICATIONS OF RECOMBINANT DNA TECHNOLOGY ARE NUMEROUS

The isolation of a specific gene from an entire genome requires a technique that will discriminate one part in a million. The identification of a regulatory region that may be only 10 bp in length requires a sensitivity of one part in 3×10^8 ; a disease such as sickle cell anemia is caused by a single base change, or one part in 3×10^9 . Recombinant DNA technology is powerful enough to accomplish all these things.

Gene Mapping Localizes Specific Genes to Distinct Chromosomes

Gene localizing thus can define a map of the human genome. This is already yielding useful information in the definition of human disease. Somatic cell hybridization and *in situ* hybridization are two techniques used to accomplish this. In ***in situ* hybridization**, the simpler and more direct procedure, a radioactive probe is added to a metaphase spread of chromosomes on a glass slide. The exact area of hybridization is localized by layering photographic emulsion over the slide and, after exposure, lining up the grains with some histologic identification of the chromosome. **Fluorescence *in situ* hybridization (FISH)** is a very sensitive technique that is also used for this purpose. This often places the gene at a location on a given band or region on the chromosome. Some of the human genes localized using these techniques are listed in Table 40-5. This table represents only a sampling, since thousands of genes have been mapped as a result of the recent sequencing of the

Figure 40-7. The polymerase chain reaction is used to amplify specific gene sequences. Double-stranded DNA is heated to separate it into individual strands. These bind two distinct primers that are directed at specific sequences on opposite strands and that define the segment to be amplified. DNA polymerase extends the primers in each direction and synthesizes two strands complementary to the original two. This cycle is repeated several times, giving an amplified product of defined length and sequence. Note that the two primers are present in excess.

Table 40-5. Localization of human genes.¹

Gene	Chromosome	Disease
Insulin	11p15	
Prolactin	6p23-q12	
Growth hormone	17q21-qter	Growth hormone deficiency
α -Globin	16p12-pter	α -Thalassemia
β -Globin	11p12	β -Thalassemia, sickle cell
Adenosine deaminase	20q13-qter	Adenosine deaminase deficiency
Phenylalanine hydroxylase	12q24	Phenylketonuria
Hypoxanthine-guanine phosphoribosyltransferase	Xq26-q27	Lesch-Nyhan syndrome
DNA segment G8	4p	Huntington's chorea

¹This table indicates the chromosomal location of several genes and the diseases associated with deficient or abnormal production of the gene products. The chromosome involved is indicated by the first number or letter. The other numbers and letters refer to precise localizations, as defined in McKusick VA: *Mendelian Inheritance in Man*, 6th ed. John Hopkins Univ Press, 1983.

human genome. Once the defect is localized to a region of DNA that has the characteristic structure of a gene (Figure 40-1), a synthetic gene can be constructed and expressed in an appropriate vector and its function can be assessed—or the putative peptide, deduced from the open reading frame in the coding region, can be synthesized. Antibodies directed against this peptide can be used to assess whether this peptide is expressed in normal persons and whether it is absent in those with the genetic syndrome.

Proteins Can Be Produced for Research & Diagnosis

A practical goal of recombinant DNA research is the production of materials for biomedical applications. This technology has two distinct merits: (1) It can supply large amounts of material that could not be obtained by conventional purification methods (eg, interferon, tissue plasminogen activating factor). (2) It can provide human material (eg, insulin, growth hormone). The advantages in both cases are obvious. Although the primary aim is to supply products—generally proteins—for treatment (insulin) and diagnosis (AIDS testing) of human and other animal diseases and for disease prevention (hepatitis B vaccine), there are other potential commercial applications, especially in agriculture. An example of the latter is the attempt to engineer plants that are more resistant to drought or temperature extremes, more efficient at fixing nitrogen, or that produce seeds containing the complete complement of essential amino acids (rice, wheat, corn, etc).

Recombinant DNA Technology Is Used in the Molecular Analysis of Disease

A. NORMAL GENE VARIATIONS

There is a normal variation of DNA sequence just as is true of more obvious aspects of human structure. Variations of DNA sequence, **polymorphisms**, occur approximately once in every 500 nucleotides, or about 10^7 times per genome. There are without doubt deletions and insertions of DNA as well as single-base substitutions. In healthy people, these alterations obviously occur in noncoding regions of DNA or at sites that cause no change in function of the encoded protein. This heritable polymorphism of DNA structure can be associated with certain diseases within a large kindred and can be used to search for the specific gene involved, as is illustrated below. It can also be used in a variety of applications in forensic medicine.

B. GENE VARIATIONS CAUSING DISEASE

Classic genetics taught that most genetic diseases were due to point mutations which resulted in an impaired protein. This may still be true, but if on reading the initial sections of this chapter one predicted that genetic disease could result from derangement of any of the steps illustrated in Figure 40-1, one would have made a proper assessment. This point is nicely illustrated by examination of the β -globin gene. This gene is located in a cluster on chromosome 11 (Figure 40-8), and an expanded version of the gene is illustrated in Figure 40-9. Defective production of β -globin results in a variety of diseases and is due to many

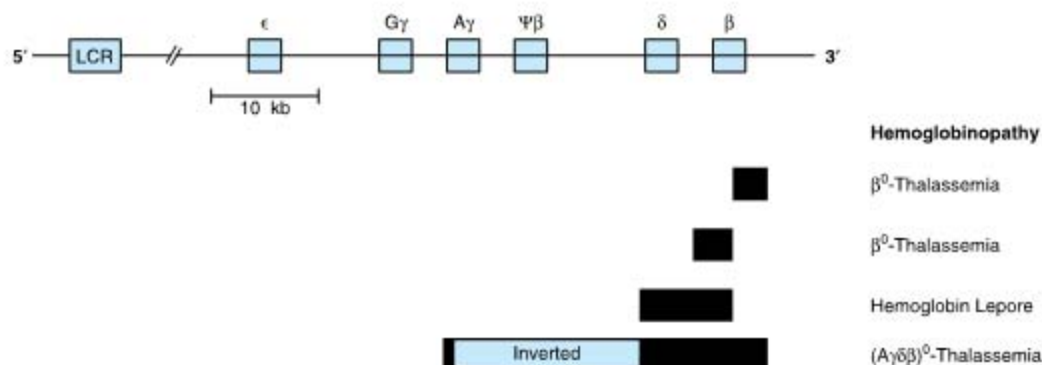


Figure 40-8. Schematic representation of the β -globin gene cluster and of the lesions in some genetic disorders. The β -globin gene is located on chromosome 11 in close association with the two γ -globin genes and the δ -globin gene. The β -gene family is arranged in the order 5'- ϵ - $G\gamma$ - $A\gamma$ - $\Psi\beta$ - δ - β -3'. The ϵ locus is expressed in early embryonic life (as $\alpha_2\epsilon_2$). The γ genes are expressed in fetal life, making fetal hemoglobin (HbF, $\alpha_2\gamma_2$). Adult hemoglobin consists of HbA ($\alpha_2\beta_2$) or HbA₂ ($\alpha_2\delta_2$). The $\Psi\beta$ is a pseudogene that has sequence homology with β but contains mutations that prevent its expression. A locus control region (LCR) located upstream (5') from the ϵ gene controls the rate of transcription of the entire β -globin gene cluster. Deletions (solid bar) of the β locus cause β -thalassemia (deficiency or absence [β^0] of β -globin). A deletion of δ and β causes hemoglobin Lepore (only hemoglobin α is present). An inversion ($A\gamma\delta\beta$)⁰ in this region (colored bar) disrupts gene function and also results in thalassemia (type III). Each type of thalassemia tends to be found in a certain group of people, eg, the ($A\gamma\delta\beta$)⁰ deletion inversion occurs in persons from India. Many more deletions in this region have been mapped, and each causes some type of thalassemia.

different lesions in and around the β -globin gene (Table 40-6).

C. POINT MUTATIONS

The classic example is **sickle cell disease**, which is caused by mutation of a single base out of the 3×10^9 in the genome, a T-to-A DNA substitution, which in

turn results in an A-to-U change in the mRNA corresponding to the sixth codon of the β -globin gene. The altered codon specifies a different amino acid (valine rather than glutamic acid), and this causes a structural abnormality of the β -globin molecule. Other point mutations in and around the β -globin gene result in decreased production or, in some instances, no produc-

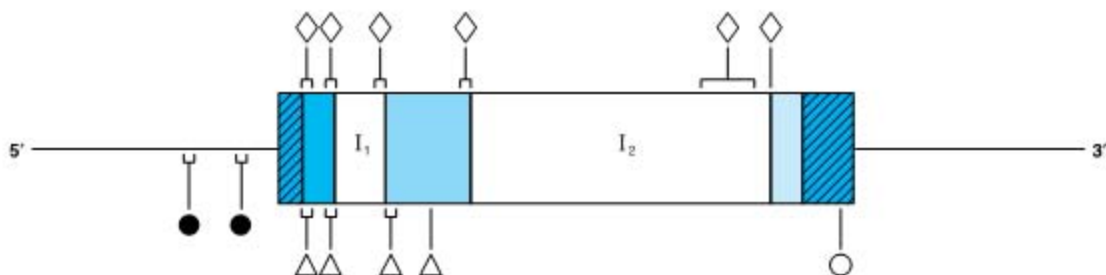


Figure 40-9. Mutations in the β -globin gene causing β -thalassemia. The β -globin gene is shown in the 5' to 3' orientation. The cross-hatched areas indicate the 5' and 3' nontranslated regions. Reading from the 5' to 3' direction, the shaded areas are exons 1-3 and the clear spaces are introns 1 (I_1) and 2 (I_2). Mutations that affect transcription control (●) are located in the 5' flanking-region DNA. Examples of nonsense mutations (Δ), mutations in RNA processing (◇), and RNA cleavage mutations (○) have been identified and are indicated. In some regions, many mutations have been found. These are indicated by the brackets.

Table 40–6. Structural alterations of the β -globin gene.

Alteration	Function Affected	Disease
Point mutations	Protein folding	Sickle cell disease
	Transcriptional control	β -Thalassemia
	Frameshift and non-sense mutations	β -Thalassemia
	RNA processing	β -Thalassemia
Deletion	mRNA production	β^0 -Thalassemia Hemoglobin Lepore
Rearrangement	mRNA production	β -Thalassemia type III

tion of β -globin; β -thalassemia is the result of these mutations. (The thalassemias are characterized by defects in the synthesis of hemoglobin subunits, and so β -thalassemia results when there is insufficient production of β -globin.) Figure 40–9 illustrates that point mutations affecting each of the many processes involved in generating a normal mRNA (and therefore a normal protein) have been implicated as a cause of β -thalassemia.

D. DELETIONS, INSERTIONS, & REARRANGEMENTS OF DNA

Studies of bacteria, viruses, yeasts, and fruit flies show that pieces of DNA can move from one place to another within a genome. The deletion of a critical piece of DNA, the rearrangement of DNA within a gene, or the insertion of a piece of DNA within a coding or regulatory region can all cause changes in gene expression resulting in disease. Again, a molecular analysis of β -thalassemia produces numerous examples of these processes—particularly deletions—as causes of disease (Figure 40–8). The globin gene clusters seem particularly prone to this lesion. Deletions in the α -globin cluster, located on chromosome 16, cause α -thalassemia. There is a strong ethnic association for many of these deletions, so that northern Europeans, Filipinos, blacks, and Mediterranean peoples have different lesions all resulting in the absence of hemoglobin A and α -thalassemia.

A similar analysis could be made for a number of other diseases. Point mutations are usually defined by sequencing the gene in question, though occasionally, if the mutation destroys or creates a restriction enzyme site, the technique of restriction fragment analysis can be used to pinpoint the lesion. Deletions or insertions of DNA larger than 50 bp can often be detected by the Southern blotting procedure.

E. PEDIGREE ANALYSIS

Sickle cell disease again provides an excellent example of how recombinant DNA technology can be applied to the study of human disease. The substitution of T for A in the template strand of DNA in the β -globin gene changes the sequence in the region that corresponds to the sixth codon from

\downarrow
 CCTGAGG Coding strand
 GGAC \textcircled{T} CC Template strand
 \uparrow

to

CCTGTGG Coding strand
 GGAC \textcircled{A} CC Template strand

and destroys a recognition site for the restriction enzyme *MstII* (CCTNAGG; denoted by the small vertical arrows; Table 40–2). Other *MstII* sites 5' and 3' from this site (Figure 40–10) are not affected and so will be cut. Therefore, incubation of DNA from normal (AA), heterozygous (AS), and homozygous (SS) individuals results in three different patterns on Southern blot transfer (Figure 40–10). This illustrates how a DNA pedigree can be established using the principles discussed in this chapter. Pedigree analysis has been applied to a number of genetic diseases and is most useful in those caused by deletions and insertions or the rarer instances in which a restriction endonuclease cleavage site is affected, as in the example cited in this paragraph. The analysis is facilitated by the PCR reaction, which can provide sufficient DNA for analysis from just a few nucleated red blood cells.

F. PRENATAL DIAGNOSIS

If the genetic lesion is understood and a specific probe is available, prenatal diagnosis is possible. DNA from cells collected from as little as 10 mL of amniotic fluid (or by chorionic villus biopsy) can be analyzed by Southern blot transfer. A fetus with the restriction pattern AA in Figure 40–10 does not have sickle cell disease, nor is it a carrier. A fetus with the SS pattern will develop the disease. Probes are now available for this type of analysis of many genetic diseases.

G. RESTRICTION FRAGMENT LENGTH POLYMORPHISM (RFLP)

The differences in DNA sequence cited above can result in variations of restriction sites and thus in the length of restriction fragments. An inherited difference in the pattern of restriction (eg, a DNA variation occurring in more than 1% of the general population) is known as a restriction fragment length polymorphism,

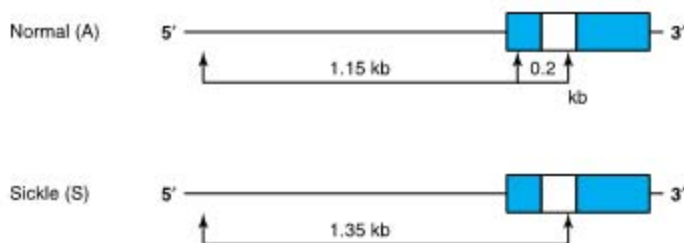
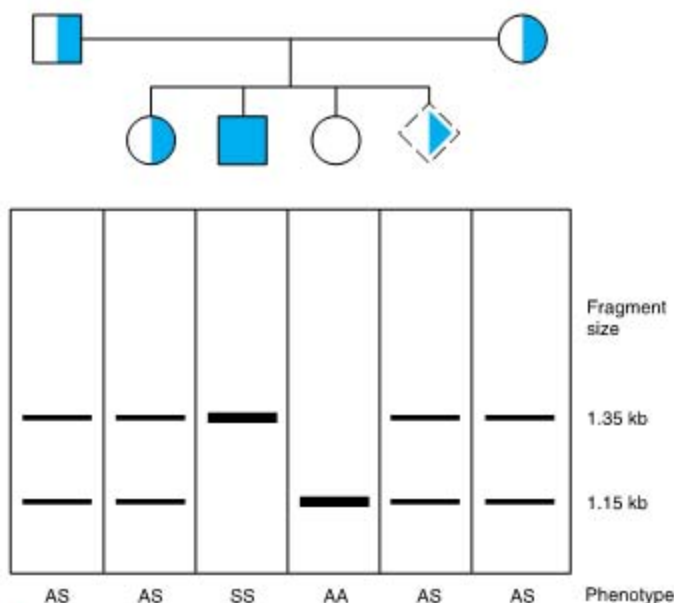
A. MstII restriction sites around and in the β -globin gene**B. Pedigree analysis**

Figure 40–10. Pedigree analysis of sickle cell disease. The top part of the figure (A) shows the first part of the β -globin gene and the *MstII* restriction enzyme sites in the normal (A) and sickle cell (S) β -globin genes. Digestion with the restriction enzyme *MstII* results in DNA fragments 1.15 kb and 0.2 kb long in normal individuals. The T-to-A change in individuals with sickle cell disease abolishes one of the three *MstII* sites around the β -globin gene; hence, a single restriction fragment 1.35 kb in length is generated in response to *MstII*. This size difference is easily detected on a Southern blot. (The 0.2-kb fragment would run off the gel in this illustration.) (B) Pedigree analysis shows three possibilities: AA = normal (open circle); AS = heterozygous (half-solid circles, half-solid square); SS = homozygous (solid square). This approach allows for prenatal diagnosis of sickle cell disease (dash-sided square).

or RFLP. An extensive RFLP map of the human genome has been constructed. This is proving useful in the human genome sequencing project and is an important component of the effort to understand various single-gene and multigenic diseases. RFLPs result from single-base changes (eg, sickle cell disease) or from deletions or insertions of DNA into a restriction fragment (eg, the thalassemias) and have proved to be useful diagnostic tools. They have been found at known gene loci and in sequences that have no known function; thus, RFLPs may disrupt the function of the gene or may have no biologic consequences.

RFLPs are inherited, and they segregate in a mendelian fashion. A major use of RFLPs (thousands are now known) is in the definition of inherited diseases in which the functional deficit is unknown. RFLPs can be used to establish linkage groups, which in turn, by the process of **chromosome walking**, will eventually define the disease locus. In chromosome walking (Figure 40-11), a fragment representing one end of a long piece of DNA is used to isolate another that overlaps but extends the first. The direction of extension is determined by restriction mapping, and the procedure is repeated sequentially until the desired sequence is obtained. The X chromosome-linked disorders are particularly amenable to this approach, since only a single allele is expressed. Hence, 20% of the defined RFLPs are on the X chromosome, and a reasonably complete linkage map of this chromosome exists. The gene for the X-linked disorder, Duchenne-type muscular dystrophy, was found using RFLPs. Likewise, the defect in Huntington's disease was localized to the terminal region of the short arm of chromosome 4, and the defect that causes polycystic kidney disease is linked to the α -globin locus on chromosome 16.

H. MICROSATELLITE DNA POLYMORPHISMS

Short (2–6 bp), inherited, tandem repeat units of DNA occur about 50,000–100,000 times in the human genome (Chapter 36). Because they occur more frequently—and in view of the routine application of sensitive PCR methods—they are replacing RFLPs as the marker loci for various genome searches.

I. RFLPs & VNTRs in FORENSIC MEDICINE

Variable numbers of tandemly repeated (VNTR) units are one common type of "insertion" that results in an RFLP. The VNTRs can be inherited, in which case they are useful in establishing genetic association with a disease in a family or kindred; or they can be unique to an individual and thus serve as a molecular fingerprint of that person.

J. GENE THERAPY

Diseases caused by deficiency of a gene product (Table 40-5) are amenable to replacement therapy. The strategy is to clone a gene (eg, the gene that codes for adenosine deaminase) into a vector that will readily be taken up and incorporated into the genome of a host cell. Bone marrow precursor cells are being investigated for this purpose because they presumably will resettle in the marrow and replicate there. The introduced gene would begin to direct the expression of its protein product, and this would correct the deficiency in the host cell.

K. TRANSGENIC ANIMALS

The somatic cell gene replacement described above would obviously not be passed on to offspring. Other strategies to alter germ cell lines have been devised but have been tested only in experimental animals. A certain

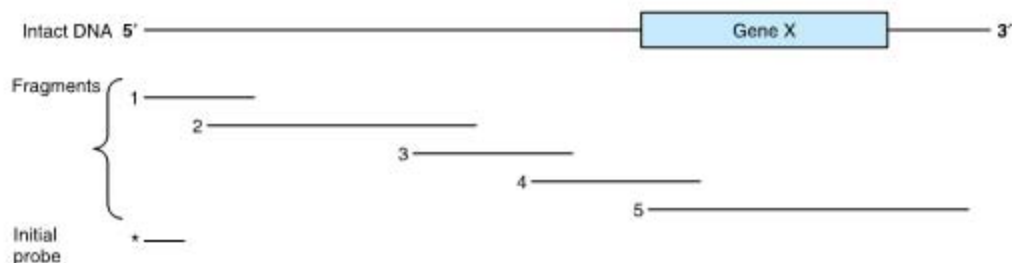


Figure 40-11. The technique of chromosome walking. Gene X is to be isolated from a large piece of DNA. The exact location of this gene is not known, but a probe (*) directed against a fragment of DNA (shown at the 5' end in this representation) is available, as is a library containing a series of overlapping DNA fragments. For the sake of simplicity, only five of these are shown. The initial probe will hybridize only with clones containing fragment 1, which can then be isolated and used as a probe to detect fragment 2. This procedure is repeated until fragment 4 hybridizes with fragment 5, which contains the entire sequence of gene X.

percentage of genes injected into a fertilized mouse ovum will be incorporated into the genome and found in both somatic and germ cells. Hundreds of transgenic animals have been established, and these are useful for analysis of tissue-specific effects on gene expression and effects of overproduction of gene products (eg, those from the growth hormone gene or oncogenes) and in discovering genes involved in development—a process that heretofore has been difficult to study. The transgenic approach has recently been used to correct a genetic deficiency in mice. Fertilized ova obtained from mice with genetic hypogonadism were injected with DNA containing the coding sequence for the gonadotropin-releasing hormone (GnRH) precursor protein. This gene was expressed and regulated normally in the hypothalamus of a certain number of the resultant mice, and these animals were in all respects normal. Their offspring also showed no evidence of GnRH deficiency. This is, therefore, evidence of somatic cell expression of the transgene and of its maintenance in germ cells.

Targeted Gene Disruption or Knockout

In transgenic animals, one is adding one or more copies of a gene to the genome, and there is no way to control where that gene eventually resides. A complementary—and much more difficult—approach involves the selective removal of a gene from the genome. Gene knockout animals (usually mice) are made by creating a mutation that totally disrupts the function of a gene. This is then used to replace one of the two genes in an embryonic stem cell that can be used to create a heterozygous transgenic animal. The mating of two such animals will, by mendelian genetics, result in a homozygous mutation in 25% of offspring. Several hundred strains of mice with knockouts of specific genes have been developed.

RNA Transcript & Protein Profiling

The “-omic” revolution of the last several years has culminated in the determination of the nucleotide sequences of entire genomes, including those of budding and fission yeasts, various bacteria, the fruit fly, the worm *Caenorhabditis elegans*, the mouse and, most notably, humans. Additional genomes are being sequenced at an accelerating pace. The availability of all of this DNA sequence information, coupled with engineering advances, has led to the development of several revolutionary methodologies, most of which are based upon **high-density microarray technology**. We now have the ability to deposit thousands of specific, known, definable DNA sequences (more typically now synthetic oligonucleotides) on a glass microscope-style slide in the space of a few

square centimeters. By coupling such DNA microarrays with highly sensitive detection of hybridized fluorescently labeled nucleic acid probes derived from mRNA, investigators can rapidly and accurately generate profiles of gene expression (eg, specific cellular mRNA content) from cell and tissue samples as small as 1 gram or less. Thus entire **transcriptome information** (the entire collection of cellular mRNAs) for such cell or tissue sources can readily be obtained in only a few days. Transcriptome information allows one to predict the collection of proteins that might be expressed in a particular cell, tissue, or organ in normal and disease states based upon the mRNAs present in those cells. Complementing this high-throughput, transcript-profiling method is the recent development of high-sensitivity, high-throughput **mass spectrometry of complex protein samples**. Newer mass spectrometry methods allow one to identify hundreds to thousands of proteins in proteins extracted from very small numbers of cells (< 1 g). This critical information tells investigators which of the many mRNAs detected in transcript microarray mapping studies are actually translated into protein, generally the ultimate dictator of phenotype. Microarray techniques and mass spectrometric protein identification experiments both lead to the generation of huge amounts of data. Appropriate data management and interpretation of the deluge of information forthcoming from such studies has relied upon statistical methods; and this new technology, coupled with the flood of DNA sequence information, has led to the development of the field of **bioinformatics**, a new discipline whose goal is to help manage, analyze, and integrate this flood of biologically important information. Future work at the intersection of bioinformatics and transcript-protein profiling will revolutionize our understanding of biology and medicine.

SUMMARY

- A variety of very sensitive techniques can now be applied to the isolation and characterization of genes and to the quantitation of gene products.
- In DNA cloning, a particular segment of DNA is removed from its normal environment using one of many restriction endonucleases. This is then ligated into one of several vectors in which the DNA segment can be amplified and produced in abundance.
- The cloned DNA can be used as a probe in one of several types of hybridization reactions to detect other related or adjacent pieces of DNA, or it can be used to quantitate gene products such as mRNA.
- Manipulation of the DNA to change its structure, so-called genetic engineering, is a key element in cloning (eg, the construction of chimeric molecules) and can

also be used to study the function of a certain fragment of DNA and to analyze how genes are regulated.

- Chimeric DNA molecules are introduced into cells to make transfected cells or into the fertilized oocyte to make transgenic animals.
- Techniques involving cloned DNA are used to locate genes to specific regions of chromosomes, to identify the genes responsible for diseases, to study how faulty gene regulation causes disease, to diagnose genetic diseases, and increasingly to treat genetic diseases.

GLOSSARY

ARS: Autonomously replicating sequence; the origin of replication in yeast.

Autoradiography: The detection of radioactive molecules (eg, DNA, RNA, protein) by visualization of their effects on photographic film.

Bacteriophage: A virus that infects a bacterium.

Blunt-ended DNA: Two strands of a DNA duplex having ends that are flush with each other.

cDNA: A single-stranded DNA molecule that is complementary to an mRNA molecule and is synthesized from it by the action of reverse transcriptase.

Chimeric molecule: A molecule (eg, DNA, RNA, protein) containing sequences derived from two different species.

Clone: A large number of organisms, cells or molecules that are identical with a single parental organism cell or molecule.

Cosmid: A plasmid into which the DNA sequences from bacteriophage lambda that are necessary for the packaging of DNA (cos sites) have been inserted; this permits the plasmid DNA to be packaged in vitro.

Endonuclease: An enzyme that cleaves internal bonds in DNA or RNA.

Excinuclease: The excision nuclease involved in nucleotide exchange repair of DNA.

Exon: The sequence of a gene that is represented (expressed) as mRNA.

Exonuclease: An enzyme that cleaves nucleotides from either the 3' or 5' ends of DNA or RNA.

Fingerprinting: The use of RFLPs or repeat sequence DNA to establish a unique pattern of DNA fragments for an individual.

Footprinting: DNA with protein bound is resistant to digestion by DNase enzymes. When a sequencing reaction is performed using such DNA, a protected area, representing the "footprint" of the bound protein, will be detected.

Hairpin: A double-helical stretch formed by base pairing between neighboring complementary se-

quences of a single strand of DNA or RNA.

Hybridization: The specific reassociation of complementary strands of nucleic acids (DNA with DNA, DNA with RNA, or RNA with RNA).

Insert: An additional length of base pairs in DNA, generally introduced by the techniques of recombinant DNA technology.

Intron: The sequence of a gene that is transcribed but excised before translation.

Library: A collection of cloned fragments that represents the entire genome. Libraries may be either genomic DNA (in which both introns and exons are represented) or cDNA (in which only exons are represented).

Ligation: The enzyme-catalyzed joining in phosphodiester linkage of two stretches of DNA or RNA into one; the respective enzymes are DNA and RNA ligases.

Lines: Long interspersed repeat sequences.

Microsatellite polymorphism: Heterozygosity of a certain microsatellite repeat in an individual.

Microsatellite repeat sequences: Dispersed or group repeat sequences of 2–5 bp repeated up to 50 times. May occur at 50–100 thousand locations in the genome.

Nick translation: A technique for labeling DNA based on the ability of the DNA polymerase from *E. coli* to degrade a strand of DNA that has been nicked and then to resynthesize the strand; if a radioactive nucleoside triphosphate is employed, the rebuilt strand becomes labeled and can be used as a radioactive probe.

Northern blot: A method for transferring RNA from an agarose gel to a nitrocellulose filter, on which the RNA can be detected by a suitable probe.

Oligonucleotide: A short, defined sequence of nucleotides joined together in the typical phosphodiester linkage.

Ori: The origin of DNA replication.

PAC: A high capacity (70–95 kb) cloning vector based upon the lytic *E. coli* bacteriophage P1 that replicates in bacteria as an extrachromosomal element.

Palindrome: A sequence of duplex DNA that is the same when the two strands are read in opposite directions.

Plasmid: A small, extrachromosomal, circular molecule of DNA that replicates independently of the host DNA.

Polymerase chain reaction (PCR): An enzymatic method for the repeated copying (and thus amplification) of the two strands of DNA that make up a particular gene sequence.

Primosome: The mobile complex of helicase and primase that is involved in DNA replication.

Probe: A molecule used to detect the presence of a specific fragment of DNA or RNA in, for instance, a bacterial colony that is formed from a genetic library or during analysis by blot transfer techniques; common probes are cDNA molecules, synthetic oligodeoxynucleotides of defined sequence, or antibodies to specific proteins.

Proteome: The entire collection of expressed proteins in an organism.

Pseudogene: An inactive segment of DNA arising by mutation of a parental active gene.

Recombinant DNA: The altered DNA that results from the insertion of a sequence of deoxynucleotides not previously present into an existing molecule of DNA by enzymatic or chemical means.

Restriction enzyme: An endodeoxynuclease that causes cleavage of both strands of DNA at highly specific sites dictated by the base sequence.

Reverse transcription: RNA-directed synthesis of DNA, catalyzed by reverse transcriptase.

RT-PCR: A method used to quantitate mRNA levels that relies upon a first step of cDNA copying of mRNAs prior to PCR amplification and quantitation.

Signal: The end product observed when a specific sequence of DNA or RNA is detected by autoradiography or some other method. Hybridization with a complementary radioactive polynucleotide (eg, by Southern or Northern blotting) is commonly used to generate the signal.

Sines: Short interspersed repeat sequences.

SNP: Single nucleotide polymorphism. Refers to the fact that single nucleotide genetic variation in genome sequence exists at discrete loci throughout the chromosomes. Measurement of allelic SNP differences is useful for gene mapping studies.

snRNA: Small nuclear RNA. This family of RNAs is best known for its role in mRNA processing.

Southern blot: A method for transferring DNA from an agarose gel to nitrocellulose filter, on which the DNA can be detected by a suitable probe (eg, complementary DNA or RNA).

Southwestern blot: A method for detecting protein-DNA interactions by applying a labeled DNA probe to a transfer membrane that contains a renatured protein.

Spliceosome: The macromolecular complex responsible for precursor mRNA splicing. The spliceosome consists of at least five small nuclear RNAs (snRNA; U1, U2, U4, U5, and U6) and many proteins.

Splicing: The removal of introns from RNA accompanied by the joining of its exons.

Sticky-ended DNA: Complementary single strands of DNA that protrude from opposite ends of a DNA duplex or from the ends of different duplex molecules (see also Blunt-ended DNA, above).

Tandem: Used to describe multiple copies of the same sequence (eg, DNA) that lie adjacent to one another.

Terminal transferase: An enzyme that adds nucleotides of one type (eg, deoxyadenonucleotidyl residues) to the 3' end of DNA strands.

Transcription: Template DNA-directed synthesis of nucleic acids; typically DNA-directed synthesis of RNA.

Transcriptome: The entire collection of expressed mRNAs in an organism.

Transgenic: Describing the introduction of new DNA into germ cells by its injection into the nucleus of the ovum.

Translation: Synthesis of protein using mRNA as template.

Vector: A plasmid or bacteriophage into which foreign DNA can be introduced for the purposes of cloning.

Western blot: A method for transferring protein to a nitrocellulose filter, on which the protein can be detected by a suitable probe (eg, an antibody).

REFERENCES

- Lewin B: *Genes VII*. Oxford Univ Press, 1999.
- Martin JB, Gusella JF: Huntington's disease: pathogenesis and management. *N Engl J Med* 1986;315:1267.
- Sambrook J, Fritsch EF, Maniatis T: *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, 1989.
- Spector DL, Goldman RD, Leinwand LA: *Cells: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, 1998.
- Watson JD et al: *Recombinant DNA*, 2nd ed. Scientific American Books, Freeman, 1992.
- Weatherall DJ: *The New Genetics and Clinical Practice*, 3rd ed. Oxford Univ Press, 1991.