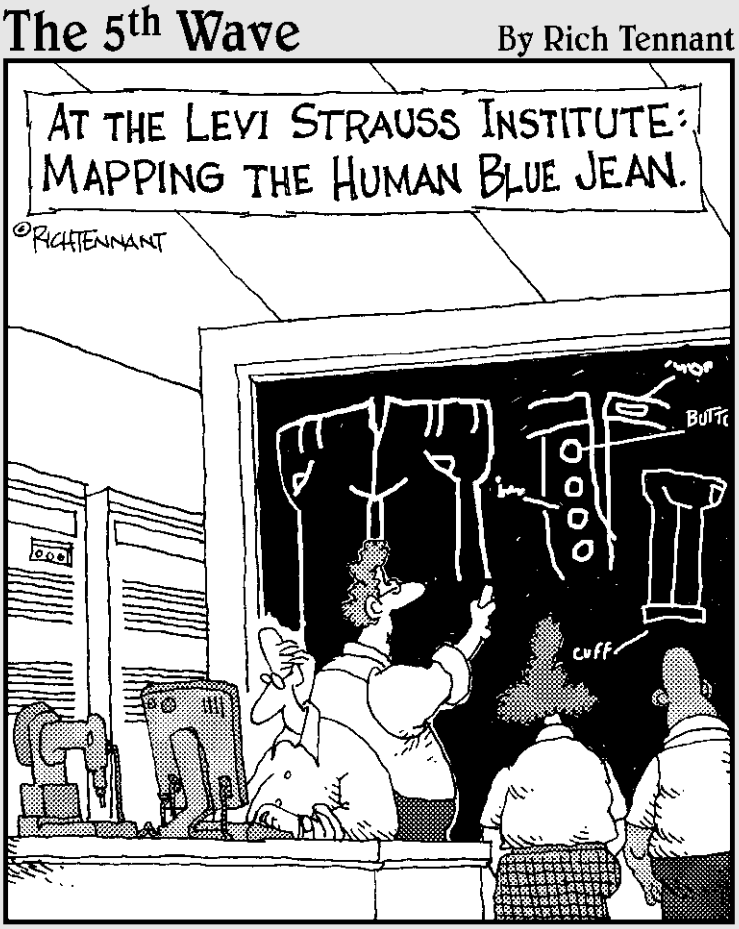
Part V

Genetics: Why We Are What We Are



In this part . . .

We roll up our sleeves and return to the subject of genes and DNA to look at them much more closely. We cover the way DNA replicates itself and look at a number of applications related to DNA sequencing. Then it's off to RNA transcription and protein synthesis and translation. At the very end we mention some of the goals and questions sought by the Human Genome Project.

Chapter 15

Photocopying DNA

In This Chapter

- Learning about replication
- Checking out recombinant DNA
- Examining DNA ssequencing
- Discussing ethical issues

In 1958, Francis Crick postulated what became the "central dogma of molecular biology." In this postulate, he, and later others, reasoned that DNA was the central source of genetic information and that it passed on some of this information to form RNA, which, in turn, passed this information on to form proteins. This central dogma is an extension of the one-gene one-protein hypothesis. To achieve this, the DNA must be able to pass on its information both to later generations (*replication*), and to RNA (*transcription*). RNA must finish the series by forming the appropriate proteins (*translation*).

Some RNA, especially some viral RNA, can undergo replication and even reverse-transcription — thus, RNA can produce both RNA and DNA. Genetic researchers initially thought this was in conflict with the central dogma; but Crick reasoned that RNA creating DNA was an extension of this postulate.

Many of the viruses capable of reverse-transcription are cancer causing.

The primary structure of DNA consists of two polynucleotide strands held together by hydrogen bonds. Adenine forms hydrogen bonds to thymine, and cytosine forms hydrogen bonds to guanine (Figure 15-1). The sequence of nitrogen bases contains the genetic information. The DNA molecules wrap around a protein called a *histone* — the combination of eight histones with the associated DNA is a *nucleosome.* (We talk more about histones in Chapter 16.)

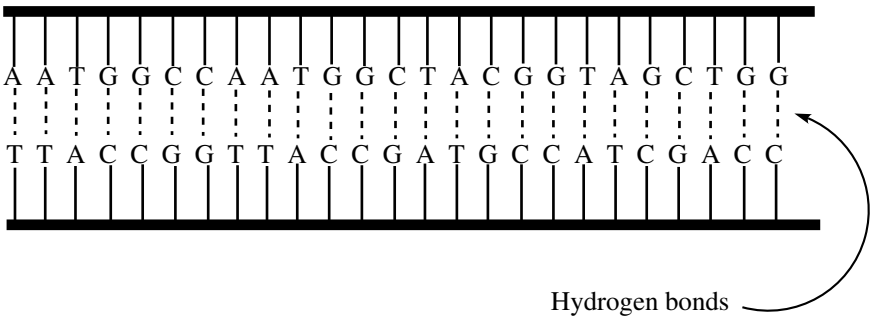


Figure 15-1: A schematic illustration of the base pairs present in a segment of DNA.

A *gene* is a portion of a DNA molecule that carries specific information. The portion of the gene coding for that specific information is called an *exon*. The portion of a gene that does not code for specific information is an *intron*.

Let's Do It Again: Replication

Replication is the process that produces new DNA molecules. One DNA molecule produce two DNA molecules in a process where the DNA must unwind and open — kind of like a zipper. New nucleotides bind to the backbone of each strand of the opened DNA by forming hydrogen bonds to the nucleotides (the zipper's "teeth") that are already present. The process proceeds along the opening DNA strand until each half of the original DNA has a complementary strand hydrogen bonded to it. The result is two DNA double helices each with half old DNA, and half new. It doesn't sound like much fun, but it works for DNA. Replication is illustrated in Figure 15-2.

The specific hydrogen bonding forces the new strands to contain a nucleotide sequence that is complementary to the nucleotide sequence in the old strand. Therefore, it can create an exact duplicate of the original DNA.

This description of replication is a simplification. It barely scratches the surface of this complicated process, but it should give you enough background information in order to understand what comes next.

The first step in understanding replication was the discovery of DNA polymerase from *Escherichia coli*. Subsequent studies showed that this enzyme needed a DNA template and all four deoxyriboside triphosphates (dATP, dCTP, dGTP, and dTTP). In addition, a short section of RNA called a *primer* is also needed. The enzyme prefers a single DNA strand for the template in order to produce a complementary strand.

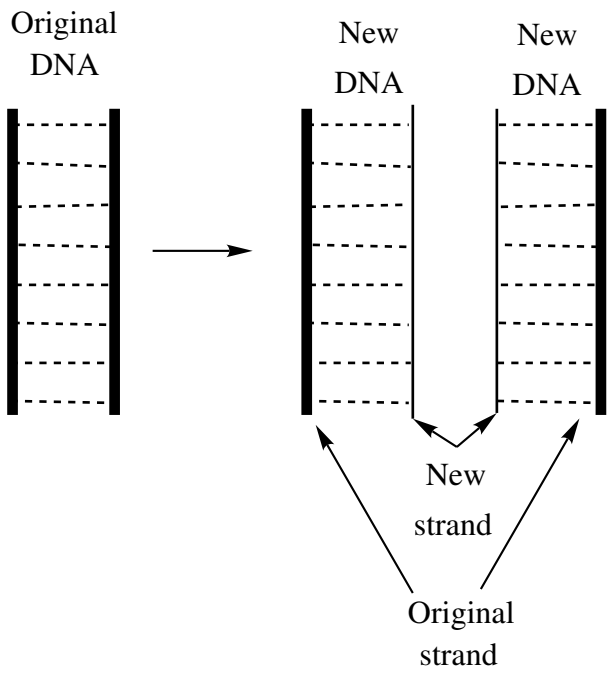


Figure 15-2: A simplified representation of replication.

During replication, simultaneous duplication of the two strands of DNA occurs. Because the two strands of DNA are anti-parallel, the mode of synthesis is different for each strand, but the overall process is the same: moving from one end to the other. For one strand the synthesis is from 5' → 3'. On the other strand it appears to be from 3' → 5', but in actuality it is also 5' → 3'. There is a complication on the 3' → 5' strand (which we discuss later in this section). See Figure 15-3.

The initiation of replication begins at a particular site, and, once initiated, a series of fragments form discontinuously along one strand and continuously along the other strand. These discontinuous fragments, known as *Okazaki fragments,* contain from 1,000 to 2,000 nucleotides. The synthesis of the fragments is always in the 5' → 3' direction. See Figure 15-4. Note that Figures 15-3 and 15-4 appear to be different at first glance. Figure 15-3 represents a simplified view of the overall process, whereas Figure 15-4 illustrates in more detail how this overall process occurs.

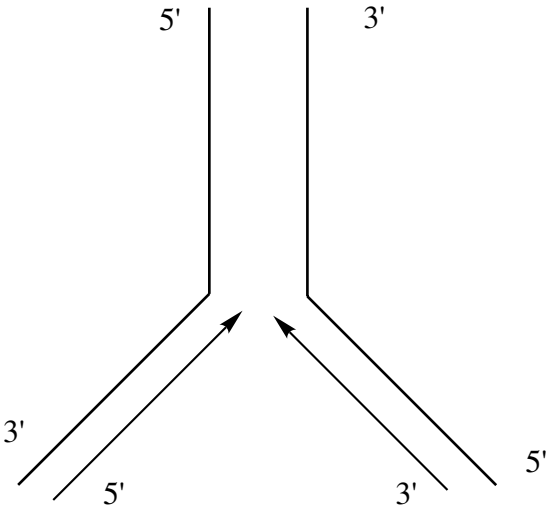


Figure 15-3: A simplified scheme of the replication of DNA.

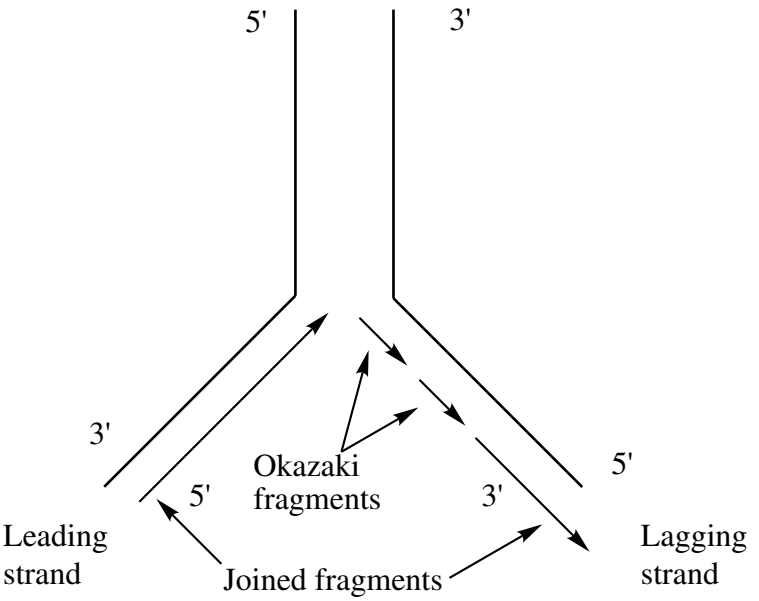


Figure 15-4: Detailed scheme of the replication of DNA.

Researchers unexpected found that RNA synthesis is a prerequisite for the replication of DNA. Initially, an RNA primer, typically 20 to 30 nucleotides in length, forms on a single DNA strand. Once formed, deoxyribonucleotide nucleotides add to the 3' terminus. Later, it is necessary to remove the RNA primer and attach the appropriate DNA fragment to produce the completed DNA.

At least a portion of the double-stranded DNA must be separated before replication can occur, and the separated portions can serve as templates. Enzymes known as *helicases* are responsible for this separation. The energy needed comes from the hydrolysis of ATP. The mechanism of separation is not well understood and is still under investigation. Apparently, the helicase binds more strongly to one strand of the DNA than the other so that the enzyme squeezes in and pushes the other strand away. ATP hydrolysis provides the energy necessary to cause the enzyme to move along the one strand nucleotide by nucleotide. This results in regions of the DNA opening like the afore-mentioned zipper.

DNA polymerases

DNA polymerases are the enzymes responsible for joining the nucleotide triphosphate fragments to produce a strand of DNA, acting as the bricklayers and carpenters in its construction. This process will only occur in the presence of a DNA template (parent DNA). Before the enzyme can connect a nucleotide, the nucleotide must bind to the appropriate site on the template.

There may be more than one DNA polymerase present in a cell. For example, in *E. coli* three different enzymes perform this task. These enzymes may also act as exonucleases. An *exonuclease* has the opposite function as a polymerase; that is, it removes nucleotides from the DNA strand.

The addition of the nucleotides is always to the 3' end of a polynucleotide chain. DNA polymerases cannot start building a nucleotide from scratch — there must be a polynucleotide already present. In contrast, RNA polymerase *can* begin from scratch. RNA polymerase generates the RNA primer, using ribonucleotides, at the beginning of replication. DNA polymerase then takes over the task and adds deoxynucleotides to the RNA primer. The polymerization requires the presence of two metal ions to enable the joining of the nucleotide to the polynucleotide.

Replication of DNA needs to be error free to ensure proper transmission of genetic information, and DNA polymerases are extremely effective in reducing errors. The enzyme binds tightly to the template and to the incoming nucleotide. This nucleotide is initially bound to the template through hydrogen bonding. If the wrong nucleotide is present, the subsequent binding to the polymerase is ineffective, and the nucleotide is "rejected." In addition to this checking, DNA polymerase also proofreads the preceding nucleotide to make sure it is correct. If the wrong nucleotide is present, it does not fit properly, making it necessary to remove the erroneous nucleotide from the polynucleotide so that the correct nucleotide may enter. The exonuclease portion of the polymerases performs this function. The polymerase proofreads the polynucleotide chain as polymerization proceeds. Proofreading is in the reverse direction (3' $\rightarrow$ 5'). There must be a nucleotide already in place before the polymerase can proofread. (We hope that our proofreader is as good as the DNA polymerases.)

The current model of DNA replication

In vitro studies show that in *E. coli,* replication begins when a protein binds a region of the DNA containing four specific binding sites. This is the *origin of replication* site. Once this protein binds, a helicase enzyme attacks the DNA and begins to unwind and separate the two strands. A third protein enters and holds the DNA strands open so that replication can continue. This third protein is the single-strand binding protein. The partially opened DNA and associated proteins are called the *prepriming complex* (Figure 15-5).

It is necessary to expose the DNA templates in this manner. A DNA strand may have more than one origin of replication site — this allows replication to occur in many places at one time. Simultaneous replication allows the cell to replicate the entire strand in less time.

Replication cannot continue until the exposed template is primed. A type of RNA polymerase known as *primase* binds to the prepriming complex in a region known as the *primosome.* Primase synthesizes a short RNA segment of about five nucleotides. Primase is capable of performing this function because its proofreading ability is not as efficient as that of DNA polymerase. For this reason, a nucleotide doesn't need to already be present to be checked. Because the primer consists of ribonucleotides instead of deoxyribonucleotides, it is temporary and will be detected and removed later. Once removed, the appropriate deoxyribonucleotides join to complete the DNA strand (Figure 15-6).

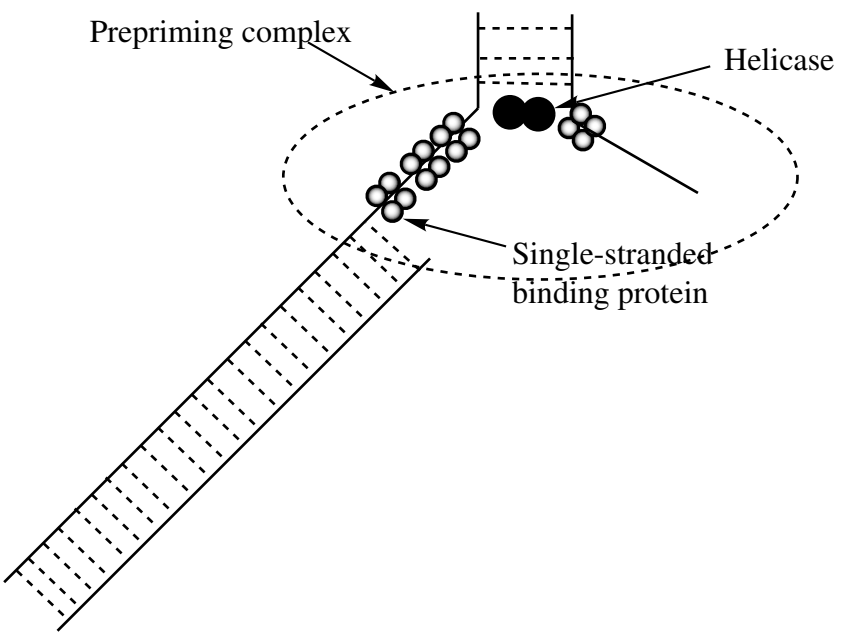


Figure 15-5: A simplified view of the prepriming complex.

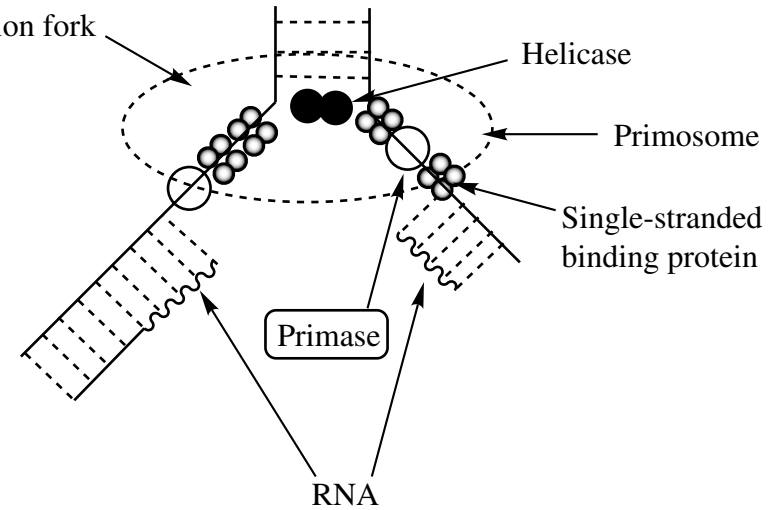


Figure 15-6: Formation of the RNA primer.

Although both strands of DNA serve as templates, the replication process differs on each strand. The point where the strands split and replication occurs is the *replication fork.* Because the two strands are anti-parallel, and DNA polymerase only works in the 5' → 3' direction, direct replication only works on one strand — called the *leading strand.* The other strand is the *lagging strand.*

As the DNA strands separate, eventually there is enough room to begin synthesis in the reverse direction on the lagging strand. (The reverse direction on the anti-parallel lagging strand is still 5' → 3'.) Replication on the lagging strand is discontinuous, and fragments of about 1,000 nucleotides form, called, as we have already noted, Okazaki fragments. DNA ligase then joins the fragments to produce a continuous strand.

DNA polymerase III holoenzyme (complete enzyme) simultaneously produces DNA on both the leading and lagging strands, though the mechanisms on the two strands are different. On the leading strand the process is continuous, whereas on the lagging strand it is discontinuous and more complex. To carry out the polymerization on the lagging strand, this strand loops around so that polymerization in the 5' → 3' direction can take place. After about 1,000 nucleotides — an Okazaki fragment — the polymerase releases the loop and begins a new loop and fragment. Each Okazaki fragment has a RNA primer. DNA polymerase I synthesizes DNA in the gaps between the fragments and removes the primer section. DNA ligase then joins the fragments (Figure 15-7). Wow! John wishes the carpenters who built his new house were that efficient!

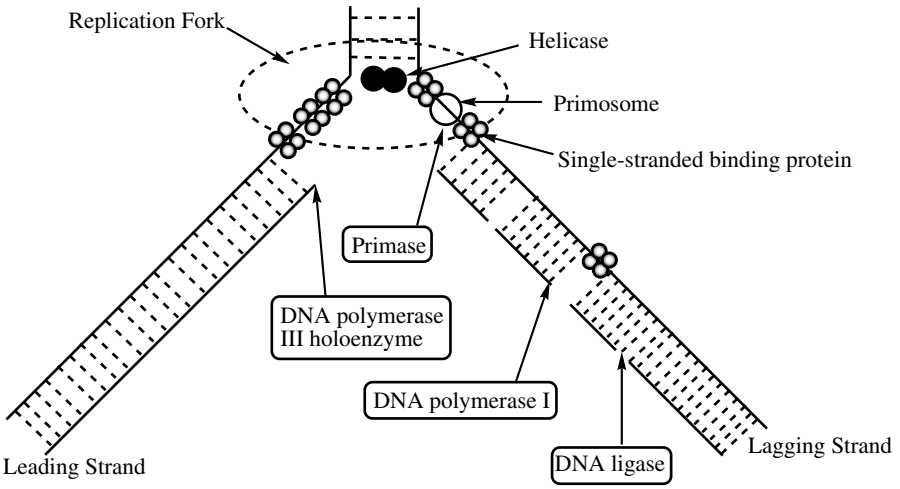


Figure 15-7: An expanded representation of the replication fork.

The ends of the DNA strands require a different procedure than does the majority of the strand, and this procedure is especially important on the lagging strand. If care were not taken, each replication cycle would result in a shorter DNA strand, eventually leading to the loss of important genetic material. To resolve this problem the ends of the DNA strands contain telomeres. *Telomeres* are DNA segments containing hundreds of repeating units. In humans, the repeating units are the hexanucleotide AGGGTT. The enzyme *telomerase,* in humans, detects the primer sequence GGTT and repeatedly attaches the hexanucleotide units, completing the DNA strand.

Mechanisms of DNA repair

All cells have a variety of DNA repair mechanisms, which are necessary to repair defective DNA and ensure retention of genetic information. Damage to DNA may occur during replication or by the action of radiation or chemicals. There is a rare error known as *xeroderma pigmentosum,* which impairs these repair mechanisms. Individuals suffering from it are extremely susceptible to cancers, especially skin cancers. Eventually the skin cancers metastasize, leading to death. The three general types of repair mechanisms are

- Direct repair
- Base-excision repair
- Nucleotide-excision repair

One example of damage needing repair is the formation of a thymine dimer (Figure 15-8) by ultraviolet (UV) light. The *thymine dimer* is an example of a

pyrimidine dimer, and its presence causes distortion of the DNA in the region. Other problems include base mismatches and missing or additional bases.

Direct repair

Here, the correction of the problem occurs in place. The photoreactivating enzyme, DNA photolyase, binds to the cyclobutane ring present in a thymine dimer, using light energy to cleave this dimer into the original bases.

Base-excision repair

In base-excision repair, the correction of the problem involves removal and replacement of the base. This is necessary whenever a modified base is present. There are various causes of modified bases, such as radiation or certain chemicals. The presence of a modified base normally results in a recognizable distortion in the DNA molecule. An enzyme, behaving as a glycosylase, cleaves the glycosidic bond to release the base from the deoxyribose. The result is an AP site (AP meaning *apurinic* or *apyrimidinic*). With apurinic, the purine base is absent, in apyrimidinic the pyrimidine base is absent. An AP endonuclease recognizes this site and cuts the DNA backbone adjacent to the site. Next, a deoxyribose phosphodiesterase completes the removal of the remaining deoxyribose phosphate. DNA polymerase I then inserts a replacement nucleotide to match the nucleotide in the complementary DNA strand. Finally, DNA ligase connects the units to yield the repaired strand — kind of like an electrician cutting out a bad circuit and splicing a good one in its place.

Nucleotide-excision repair

In nucleotide-excision repair, the correction of the problem involves the removal of a segment of DNA around the problem followed by its replacement. When this mechanism occurs, a DNA strand on both sides of the error is cut from the DNA strand. Typically, an exonuclease removes a 12-nucleotide section. DNA polymerase I then synthesizes a replacement segment of the strand. DNA ligase then finishes the repair.

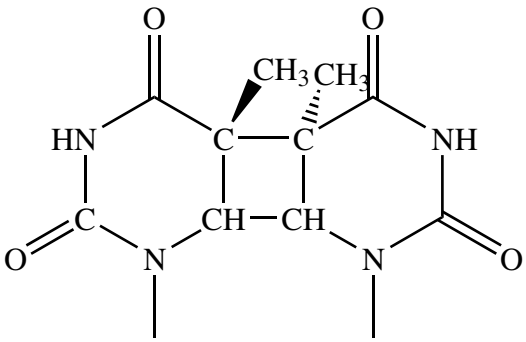


Figure 15-8: Structure of a thymine dimer.

Mutation: The good, the bad, and the ugly

Several types of mutations are known. DNA repair mechanisms try to prevent new mutations — however, such mechanisms are not always effective. Known mutations include the substitution of one base pair for another, the insertion of one or more base pairs, and the deletion of one or more base pairs. Changes, especially subtle ones, may occur during or after replication.

The substitution of one base for another is a common mutation. There are two types of substitutions. The replacement may be of a purine by the other purine (Figure 15-9) or the replacement of a pyrimidine by the other pyrimidine (Figure 15-10). This type of error is a *transition.* The other type of substitution is the replacement of a purine for a pyrimidine or vice versa. This latter type is a *transversion.*

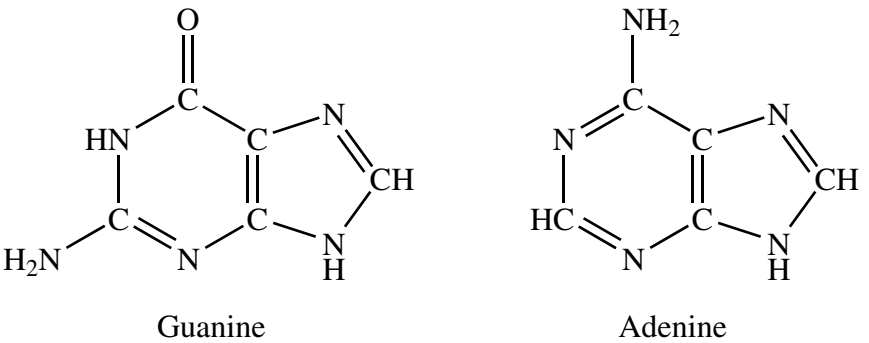


Figure 15-9: The purines.

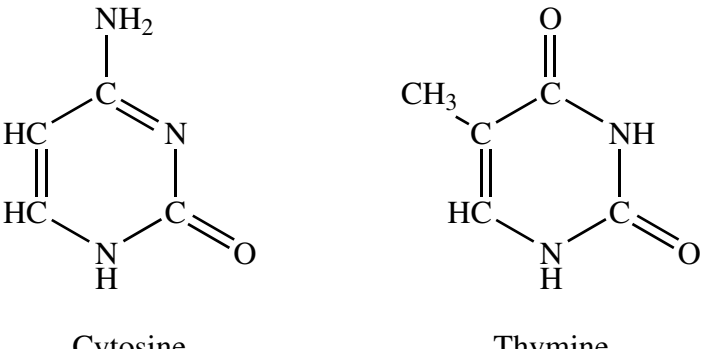


Figure 15-10: The pyrimidines.

Any uncorrected discrepancy in the genetic code will become "normal" in all future generations. The new genetic code is a *mutation.* The change in the base sequence may or may not affect the amino acid for which the codon codes. For example, changing from GTT (coding for leucine) to GTG (also coding for leucine) results in no change. However, if the change results in coding for a different amino acid, the resultant protein will function differently. If the new protein exhibits improved function, the organism benefits from the change. But if the new protein exhibits impaired function — the more likely situation — the organism suffers from the change. Problems from impaired function are genetic diseases. Table 15-1 lists some of these.

Table 15-1 Some Genetic Diseases in Humans

Disease	*Defective Protein*
Acatalasia	Catalase
Albinism	Tyrosinase
Cystic fibrosis	CF transmembrane conductance regulator
Fabray's Disease	α-Galactosidase
Gaucher's Disease	Glucocerebrosidase
Goiter	Iodotyrosine dehalogenase
Hemochromatosis	Hemochromatosis
Hemophilia	Antihemophilic factor (factor VIII)
Hyperammonemia	Ornothine transcarbamylase
McArdle's Syndrome	Muscle phosphorylase
Niemann-Pick Disease	Sphingomyelinase
Phenylketonuria	Phenylalanine hydroxylase
Pulmonary emphysema	α-Globulin of blood
Sickle cell anemia	Hemoglobin
Tay-Sachs Disease	Hexosaminidase A
Wilson's Disease	Ceruloplasmin (blood protein)

Restriction enzymes

Although not directly related to replication, restriction enzymes are important tools in genetic research. *Restriction enzymes,* or *restriction nucleases,* are capable of cutting DNA into fragments. These were first found in prokaryote cells like *E. coli* where these enzymes locate and destroy invading DNA, such as that of a bacteriophage, but leave the cell's own DNA alone. Recent research focuses on the fact that it is possible to manipulate these fragments so that DNA ligases can join the fragments into new DNA. Restriction enzymes are important in vitro biochemical tools that act as very accurate molecular scalpels. Cleavage may leave both DNA strands of equal length or one strand may be longer than the other (a *staggered* cut).

More than 100 restriction enzymes have been identified and are available for research. These enzymes recognize specific regions in the DNA and cleave DNA molecules into specific fragments. Because these fragments are smaller than the parent DNA is, they are easier to manipulate and analyze. Testing a strand of DNA with a series of restriction enzymes can provide a fingerprint of cleaved fragments. In fact, you can map the structure of DNA.

Many times in reading descriptions of genetic determination and modification, you will run across the terms *in vivo* and *in vitro. In vivo* means in the cell, whereas *in vitro* means in a test tube.

Mendel Rolling Over: Recombinant DNA

Recombinant DNA technology allows the synthesis of DNA strands that contain one or more genes not originally present. The addition of new genes enables an organism to produce new biochemicals. For example, *E. coli* has been engineered to produce human insulin. Recombinant DNA technology also allows biochemists to add a gene to compensate for a defective gene.

Restriction enzymes are capable of removing DNA fragments of interest. It is then necessary to join one of these fragments to another DNA strand for replication. The DNA to which the fragment of interest is attached is the *vector.* Common vectors include plasmids. A *plasmid* is a naturally occurring DNA circle. The first step in adding the fragment is to create a staggered cut in the DNA of the vector. The longer end of the staggered cut is a "sticky" or cohesive end. It is possible to attach any DNA fragment to the sticky end if it has the complementary DNA sequence. The complementary sticky end will be present if the same restriction enzyme was used to excise the fragment of interest. DNA ligase completes the joining of the fragment to the vector (Figure 15-11).

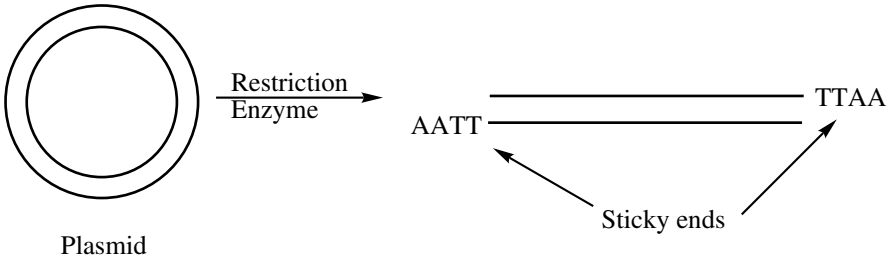


Figure 15-11: Opening of a plasmid by a restriction enzyme such as Eco RI.

It is possible to bond a DNA linker to a DNA molecule to make it susceptible to a particular restriction enzyme. By this method, the cohesive ends characteristic of any restriction enzyme may be added to almost any DNA molecule. The completed DNA can undergo replication.

Plasmids are, to a certain extent, accessory chromosomes. They can replicate independently of the host chromosomes. Thus, there may be multiple copies of a particular plasmid within a cell. This replication, in general, makes plasmids more useful as vectors than host chromosomes. Thus far, these plasmids have only been shown to be relevant in bacterial organisms.

The addition of "new" genes to an organism produces an organism that may be considered a new species. There is somewhat of a risk that these organisms could infect humans and lead to a new disease for which there is no known treatment. To minimize potential risks posed by these organisms, researchers either use enfeebled (weakened) organisms or ones that do not infect humans.

Patterns: Determining DNA Sequences

Restriction enzymes are a major tool in the determination of the base sequence in DNA. The cleaved DNA fragments are significantly smaller than the parent DNA is, making manipulation and analysis significantly easier. To separate the fragments after cleavage, gel electrophoresis is often used.

Gel electrophoresis is a biochemical technique used to separate and purify proteins and nucleic acids that differ in charge, size, or confirmation. The sample is placed into wells within a gel — a polymer that is specifically formulated for the type of analysis. This gel is in the shape of a thin slab. When separating proteins or small nucleic acids (DNA, RNA, and so on) cross-linked polyacrylamide is used. For separating larger nucleic acids, agarose, an extract from seaweed, is used. These gels have the consistency of Jell-O, but probably don't taste nearly as good.

The gel is immersed in a buffer solution, and an electrical current is applied to the ends of the gel. The charged species within the sample migrate toward one or the other of the electrodes. Proteins may have either a positive or a negative charge, but, at the proper pH, nucleic acids have only a negative charge. The positively charged species move toward the negatively charged end of the gel, and the negatively charged species move toward the positively charged end. Normally a buffer adjusts the pH so that all the species of interest have either a positive or negative charge (Figure 15-12).

Different molecules move at different speeds through the gel. When the smaller, faster molecules have about reached the end of it, the process is stopped, and the molecules are stained to make them visible. Sometimes, agents are added to cause the molecules to fluoresce (glow) under UV light. Then a photograph of the gel may be taken as it is exposed to the UV light. When several samples, including a known sample, are run side by side, the molecular weight of a sample component may be determined. This is one step in the identification of unknown components.

The separation of DNA fragments by gel electrophoresis readily distinguishes even minor differences between the fragments. Different gels are useful in separating large fragments than are useful in the separation of small fragments. In some gels, it is possible to distinguish between fragments differing by one base in several hundred. Modification of the electrophoresis method provides further separation. Each type of DNA gives a different pattern, making it possible to distinguish between two different samples. Two samples giving identical patterns must be from the same source or from identical twins.

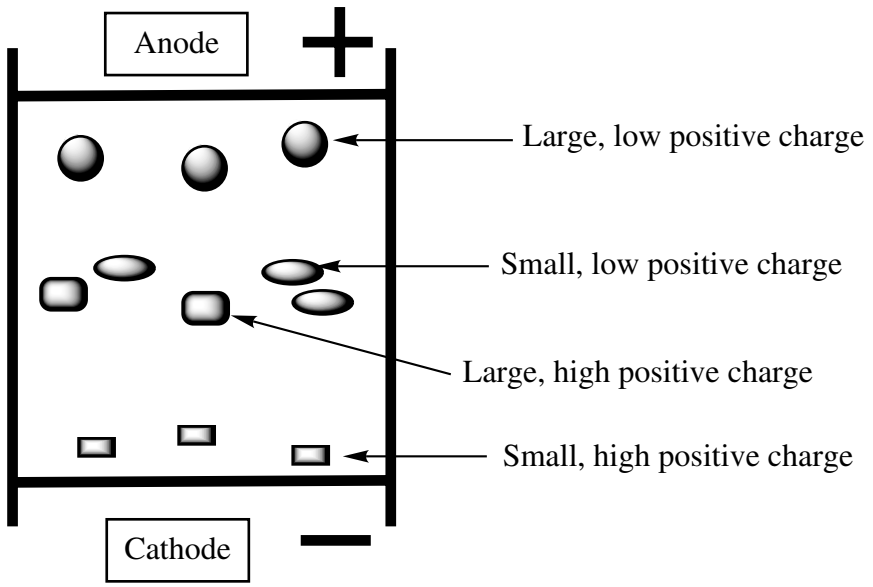


Figure 15-12: Gel electrophoresis.

In the analysis and manipulation of genetic material, it is advantageous to be able to identify whether a certain sequence of nucleotides is present. The general method for finding a particular sequence of nucleotides in DNA was developed by Edwin Southern. It is called *Southern blotting.* This method uses radioactive ^{32}P as a label that is easily detectible. This radioisotope is incorporated into the phosphate in some of the nucleotides. Determination of a particular nucleotide sequence in RNA is achieved through *Northern blotting,* and protein identification through *Western blotting.* (The names Northern and Western do not refer to persons with that name, but are by analogy to Southern. We don't know what happened to Eastern.) Alternatively, Southern, Northern, and Western blotting are DNA, RNA, and protein blots, respectively.

Determining the base sequence

Since the first isolation of DNA, a number of methods have been developed to determine the base sequence. In general, the *Sanger dideoxy method* has replaced all others. It employs the controlled termination of replication with modified nucleotides containing dideoxyribose in place of deoxyribose.

DNA fragments produced by employing restriction enzymes are denatured to give single-stranded DNA. (*Denaturing* typically involves heating a DNA-containing solution to 96°C for a few seconds.) Four samples of this DNA are treated separately to produce double-stranded DNA through replication, with each sample containing a small quantity of a different dideoxy nucleotide. The dideoxy nucleotide contains dideoxyribose (Figure 15-13). The absence of an additional oxygen atom in dideoxyribose means that there is no 3' hydroxyl group available to continue replication. Thus, the incorporation of a dideoxynucleotide terminates the DNA chain.

One of the four samples will contain a small quantity of the dideoxy analog of the nucleotide dGTP. This "defective" unit enters the new DNA strand as the complement to a cytosine base in the original fragment. Separation of the new material from the original strand material gives a set of DNA fragments of varying length. These fragments are then separated by electrophoresis according to length (size). The length of each of these fragments locates the position of each C in the original strand. The other three samples give the positions of all A, T, and G bases in the original strand.

Fluorescence tagging is a useful modification to this method. Each of the dideoxy nucleotides has a different fluorescent tag attached. After attaching the tags, it is possible to conduct all four experiments in one container. Separating the fragments by electrophoresis and examining the tags gives a colored pattern showing all the bases in sequence. This method works for fragments of up to 500 bases.

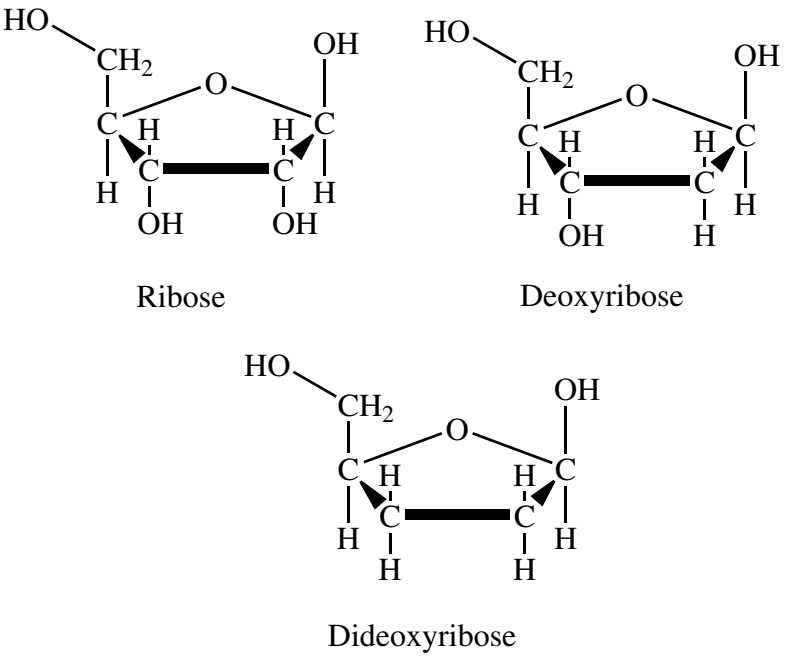


Figure 15-13: Structures of ribose, deoxyribose, and dideoxy-ribose.

To do the studies described, you need a sufficient amount of genetic material. Lack of sufficient quantities of sample has been a problem, especially with forensic evidence. Therefore, ways of quickly duplicating sufficient quantities of identical DNA fragments or producing a number of DNA strands from a very small sample were developed. *Polymerase chain reaction* (PCR), is a useful method to amplify specific DNA sequences. It is an in vitro procedure where it is necessary to know the base sequences, the flanking sequences, adjacent to a particular target sequence. However, it is not necessary to know the base sequence in the target region. Denaturation of a DNA sample provides two separate strands. Two primers are added to the mixture and one primer will attach to the flanking sequence of each strand. DNA polymerase begins replication starting at each of these primers. Repeating these steps quickly generates a large quantity of DNA. After 30 or so cycles, a billion-fold amplification occurs. Thirty cycles take less than one hour.

The butler did it: Forensic applications

Scientists can identify a species by the isolation and examination of the DNA sequences unique to that species. For example, DNA analysis is useful in the identification of organisms, such as bacteria, that may be polluting our water, food, and other samples. It has been used to establish pedigrees for livestock

breeds as well as identify endangered species in the prosecution of poachers. However, the application that has received the most publicity is in the area of *forensics*.

Because an individual's DNA comes from both the mother and father, it is unique to that individual (except in the case of identical twins). Even brothers and sisters, including fraternal twins, with the same parents show some variation in their DNA. This fact makes DNA analysis very valuable in forensics investigations (as anyone who ever watched an episode of *CSI* can attest).

In order to identify an individual, forensic investigators examine 13 regions (markers) of the DNA sample that vary significantly from individual to individual. There is a very small chance that two individuals might have the same DNA pattern at these 13 regions, but it is only about one chance in a billion. The investigation of additional markers can improve the discriminating ability of the procedure. Investigators then combine the results into a DNA profile — also known as a DNA fingerprint — of the individual.

You can isolate DNA samples from blood, hair, bone, fingernails, teeth, and any type of bodily fluid. In a typical crime scene analysis, samples are taken from the evidence and suspects; the DNA is extracted and then analyzed for the specific markers. A match of a single marker does not prove that an individual was at the crime scene, but the matching of four or five markers indicates a very high probability that the individual was present. PCR may be necessary if the sample is very small (see preceding section).

Methods of analysis

Several techniques are used in DNA analysis. The three most common are RFLP, PCR, and STR. In RFLP (*Restriction fragment length polymorphism*), the DNA sample is digested with a specific enzyme, a restriction endonuclease. This enzyme cuts DNA at a specific sequence pattern. The presence or absence of these sites in a DNA sample leads to variable lengths of DNA fragments. Gel electrophoresis then separates these fragments.

RFLP was one of the original forensic DNA analysis techniques. However, it requires relatively large amounts of DNA and samples contaminated with dirt and mold are difficult to analyze with RFLP. It has been somewhat replaced with polymerase chain reaction (PCR) enhancement, followed by STR analysis.

PCR (discussed in the earlier section "Determining the base sequence") is a useful technique that reduces the sample size requirement of RFLP — in essence it is a DNA amplifier. PCR quickly makes millions of exact copies of the DNA sample. Using PCR, DNA analysis can be done on a sample as small as a few cells and on samples that are extensively degraded. After PCR treatment, it is possible to analyze the sample with RFLP or STR.

In STR (*Short tandem repeat*) analysis, the DNA sample is quickly examined for 13 specific regions. The FBI uses this standard STR profile in its CODIS (Combined DNA Index System) program, which links national, state, and local databases of DNA profiles from felons, missing persons, and unsolved crime scenes. CODIS has an index of more than 3 million DNA profiles.

Paternity testing

Along with crime scene analysis, paternity testing is one of the most widely used applications of DNA testing. The procedure begins with the collection of DNA samples from the mother, child, and alleged father(s). The DNA profiles of the child and mother are first determined. The markers not inherited from the mother must have come from the biological father. The alleged father's DNA profile is then compared to the child. If the man's DNA profile contains markers common to the child but not the mother, then the probability that he is the biological father is great. Figure 15-14 indicates that Alleged Father 2 is more likely to be the biological father than Alleged Father 1.

Genetic Diseases and Other DNA Testing Applications

DNA testing always seems to find new ways of being useful. It has been used for a number of years, for example, in determining the gender of athletes. In addition to gender testing, the NFL used a strand of synthetic DNA to mark all the Super Bowl XXXIV footballs as a way to combat fraud associated with sports memorabilia. In a different situation, a section of DNA was added to the ink used to imprint all official goods marketed at the 2000 Summer Olympics Games. This same technology is used to tag original artwork, in addition to sports souvenirs.

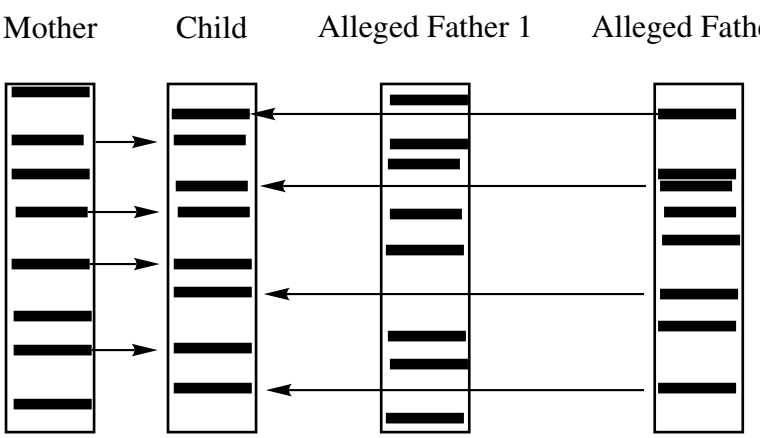


Figure 15-14: Comparison of results for paternity testing.

Genetic diseases are the result of an abnormal pattern in the DNA of an individual. These diseases are inherited, though some individuals are only carriers and not sufferers. Recently quite a bit of research has been done in determining the genetic pattern that is causing the disease, and ways to detect the probability of passing on the disease to offspring. However, methods of treatment for most all of these disease are limited. It is the dream of researchers to find the means of correcting these genetic diseases through genetic modifications. Researchers have investigated several of these genetic diseases in detail. In this section we briefly examine a few of the more well-known genetic diseases.

Sickle cell anemia

Sickle cell anemia is an inherited genetic disease of the blood's *hemoglobin,* a component of red blood cells. Sickle cell anemia is the result of the change of a single amino acid in the protein sequence of hemoglobin. This change involves the substitution of valine (non-polar) for glutamic acid (polar). The condition affects millions of individuals throughout the world, especially those whose ancestors came from Africa, South America, Cuba, Saudi Arabia, and a few other countries. In the United States, it affects about 72,000 people. Sickle cell happens in about 1 in 500 African American births and about 1 in 1,200 Hispanic American births.

Hemoglobin is responsible for carrying oxygen from the lungs to the cells. In an individual with sickle cell anemia, the defective hemoglobin molecules clump together, causing the red blood cells to assume a sickle shape, hence the name. These abnormal cells have trouble squeezing through small blood vessels, causing oxygen depletion in organs and extremities along with episodes of pain. These sickle cells also have a much shorter lifetime in the body, leaving the individual with chronic anemia. Many states now test newborns for sickle cell disease.

Hemochromatosis

Hemochromatosis, one of the most common genetic diseases in the United States, is an inherited disease that causes the body to absorb and store far too much iron. This excess iron is stored in organs, such as the liver, pancreas, and skin (yes, the skin is considered an organ!). It is due to a mutation in the HFE gene, the gene that regulates the absorption of iron from food. If this defective gene is inherited from both parents, then the person will develop hemochromatosis. If the individual inherits the mutated gene from only one parent, the person will be a carrier but will not necessarily develop the disease. About 5 Caucasian people in 1,000 carry both mutated genes, and 1 in 10 is a carrier. Genetic testing can detect it about 90 percent of the time.

Cystic fibrosis

Cystic fibrosis is a chronic and normally fatal genetic disease affecting the body's mucus glands. It targets the digestive and respiratory systems. About 55,000 individuals worldwide have cystic fibrosis. Most of these individuals are Caucasians who have ancestors who came from northern Europe. For the disease to appear it is necessary to inherit the mutated gene responsible for cystic fibrosis from both parents. Estimates are that 1 in 20 Americans carry the abnormal gene. Most of these individuals are not aware that they are carriers. Genetic testing is only about 80 percent accurate.

Hemophilia

Hemophilia is a genetic disorder caused by the lack of the blood-clotting factor stemming from a defective gene on the X chromosome. Females have two X chromosomes, so if there is a defective gene on one, there is little chance that the other one is also defective. However, she will be a carrier. Males, however, only have one X chromosome, so if it is defective, then the individual will develop hemophilia. If a woman is a carrier, she will have a 50 percent chance that her sons will have hemophilia and a 50 percent chance that her daughters will be carriers. Daughters of a hemophilic male will be carriers. Genetic testing can detect the presence of the abnormal gene.

Tay-Sachs

Tay-Sachs is an inherited disease in which a fatty-acid derivative, a lipid called *ganglioside,* accumulates in the brain — the result of a mutation of a specific gene. Although found primarily in the Jewish population, some French Canadians and Louisiana Cajuns also carry the abnormal gene. The symptoms most commonly appear in infants. Death normally occurs before the age of five. Although Tay-Sachs is a very rare disease, it was one of the first genetic diseases for which extensive and inexpensive genetic screening was developed. Screening tests were developed in the 1970s, and Israel offered free genetic screening and counseling. Because of this aggressive testing and counseling, the disease has been almost totally eradicated from Jewish families worldwide.

Ethics of genetic modification and testing

The emerging field of *bioengineering* has raised many ethical questions. One has only to listen to the debates over stem-cell research, gender selection of children, genetic modification to enhance certain traits such as athletic ability, and so on. Public policy decisions related to cost are also being debated, as genetic modification and screening are generally expensive. Is it to be available to only those who can pay, or should there be equal access? Many gray areas concerning genetic modification exist in the field of patent law. There are many questions and concerns but no quick answers.

Although the success in eradicating Tay-Sachs is directly related to genetic testing, such testing is not without its ethical questions. The major concern is one of privacy. DNA samples and profiles can be used to determine parentage and susceptibility to certain genetic diseases. Many people fear that the government, insurance companies, employers, banks, schools, and other organizations could use such information for genetic discrimination. In fact, in the United Kingdom, a man was denied treatment for hemochromatosis because his insurance company claimed it was a preexisting condition. Individuals applying for life insurance have reported other cases of genetic discrimination. Who gets to request the genetic screening and who has access to the results? These are just a few of the questions we will be debating for many years to come.

Chapter 16

Transcribe This! RNA Transcription

In This Chapter

- Finding out what's in your genes
- Breaking the genetic code
- Modeling gene regulation

Cells utilize a number of types of ribonucleic acid — RNA.

Messenger RNA (mRNA), a form that is not very stable, carries information from the cell nucleus (DNA) into the cell and must migrate to the ribosomes. Messenger RNA carries the actual genetic information necessary for the synthesis of a specific protein; however, the other forms of RNA are necessary to complete the process.

Transfer RNA (tRNA) transfers amino acids to the ribosomes for protein synthesis. This is a relatively small form of ribonucleic acid, typically containing from 73 to 93 nucleotides.

The relatively large ribosomal RNA (rRNA) resides in the ribosomes and has a direct influence on the synthesis of proteins. This form of RNA has protein components. There are three types of rRNA (called 23S, 16S, and 5S), and all three must be present in each ribosome.

Finally, small nuclear RNA (snRNA) serves a number of ancillary functions. In this chapter we concentrate upon the synthesis of RNA, which is called transcription.

Protein synthesis begins with *transcription,* the process whereby DNA produces mRNA. First, a portion (a gene) of a DNA double helix opens. Nucleotides can then bind to the exposed DNA nucleotides through a process similar to replication. However, this process differs from replication in that only a portion of the

DNA opens, and the entering nucleotides contain uracil in place of thymine. One gene yields one mRNA, which, in turn, may lead to the synthesis of one or more proteins.

The enzyme RNA polymerase joins the nucleotides to produce RNA in a process that occurs within the cell nucleus. The process begins as an initiation signal toward the 5' end of RNA and goes toward a termination sequence nearer the 3' end.

RNA Polymerase Requirements

Three requirements are needed for RNA polymerase to operate. It requires activated precursors of each of the four ribonucleoside triphosphates (ATP, CTP, GTP, and UTP) from which to produce the new RNA. (See Figure 16-1.) A divalent metal ion, either magnesium or manganese, is necessary. Finally, a template must be present. Single-stranded DNA will work; however, the preferred template is double-stranded DNA. However, the DNA strands must open (separate) in order to allow the RNA polymerase access.

There are many similarities between replication and transcription. In both processes, the direction of synthesis is 5' → 3'. Elongation occurs as the 3'-OH group of the chain attacks the innermost phosphate of the entering nucleoside triphosphate. This is called a *nucleophilic attack.* The hydrolysis of pyrophosphate provides the impetus to drive the process forward. However, there are differences. Unlike its DNA counterpart, RNA polymerase is not capable of "reviewing" its work and then eliminating a mismatched nucleotide. RNA polymerase does not require a primer.

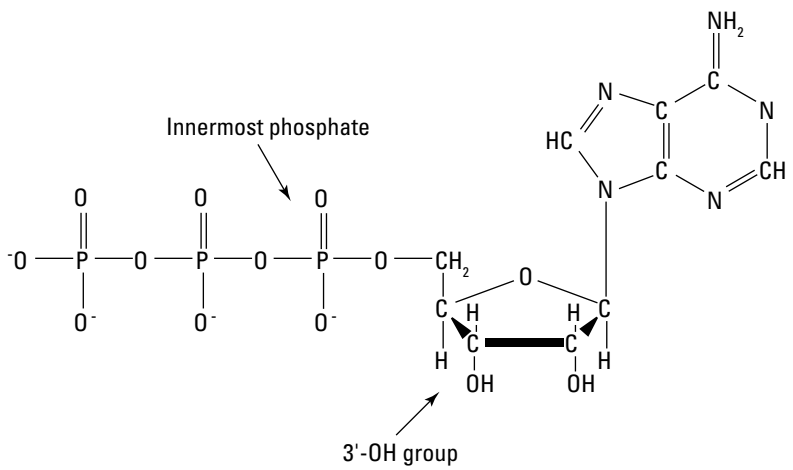


Figure 16-1: Structure of ATP.

In simple organisms, such as *E. coli*, one type of RNA polymerase synthesizes all forms of RNA. In more advanced organisms, like human beings, there are different types of RNA polymerase. Usually at least three different types are present in mammalian cells.

Making RNA: The Basics

The region of a DNA molecule that codes for a protein is a *structural gene.* Other regions are present to regulate the activity of this gene. (We examine these regulatory regions later in this chapter.) To begin transcription, it is necessary for RNA polymerase to detect one particular gene present in a long DNA strand. Detection begins with the enzyme locating a region on the DNA strand known as a *promoter site,* which is "upstream" from the actual gene. (Upstream means on the 5' side.) RNA polymerase tightly binds to the promoter site, and once in place, transcription can begin.

In prokaryotic cells the promoter sites are centered at –10 (the Pribnow box) and in the –35 region. The Pribnow box has the consensus sequence TATAAT centered at –10. The other site has the consensus sequence TTGACA. (Not all organisms have the same consensus sequence.) In eukaryotic cells a promoter is centered at about –25 (the TATA box or Hogness box), and sometimes centered near –75 (the CAAT box). The consensus sequence in the Hogness box is TATAAA. The CAAT box has the sequence GGXCAATCT. In addition, eukaryotic genes may have enhancer sequences up to several thousand bases away from the start site and on either side (Figure 16-2).

The position of sequences along the DNA chain begins at the beginning of a gene. This position is 0. The first nucleotide of the gene is +1. Counting upstream (towards the 5' terminus) is negative. Thus, ten nucleotides before the beginning of the gene would be –10.

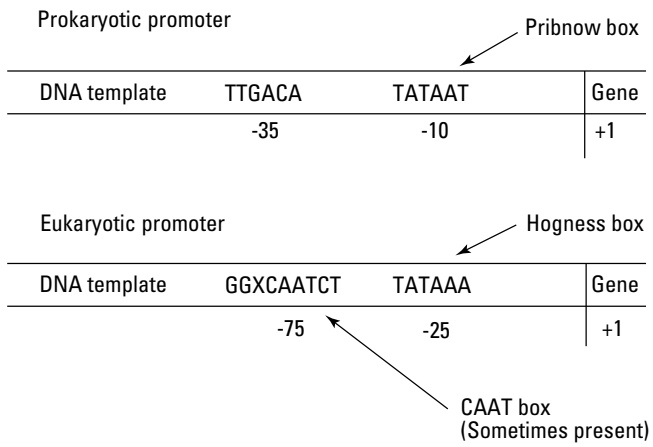


Figure 16-2: Prokaryotic and eukaryotic promoter sites.

Transcription proceeds as an RNA polymerase moves along the DNA strand. Eventually, the enzyme encounters a termination signal. In prokaryotic cells, there are two termination signals. The first is a *base-paired hairpin,* which consists of a self-complementary sequence rich in C and G followed by a sequence of several instances of U. After the sequence forms, the new RNA detaches from the template. The other method uses a *rho protein.*

The termination in eukaryotic cells is not very well understood. In eukaryotic cell, mRNA undergoes further modification after transcription. A "cap" is attached to the 5' end of the RNA, and a poly(A) tail goes onto the other end. These modifications increase the lifetime of mRNA.

The stages in RNA synthesis are *initiation*, *elongation*, and *termination*. To accomplish these tasks, RNA polymerase must perform a series of functions. The enzyme must travel along a DNA strand until it encounters a promoter site. As it "sticks" to the promoter site it unwinds a short segment of the DNA double helix and separates the strands to reach the template. Then the appropriate ribonucleoside triphosphate enters, and hydrolysis of the phosphate occurs in order to supply the needed energy. Each ribonucleoside triphosphate is brought in as the RNA polymerase moves along the DNA strand. (The DNA unwinds as the enzyme passes, and rewinds after the enzymes has passed.) This continues until the RNA polymerase finds a termination signal. The enzyme also must interact with transcription factors or trans-acting factors — proteins that act as activators or repressors — to regulate the rate of transcription initiation.

The best understood operation of RNA polymerase comes from studies of the prokaryotic cells of *E. coli.* Eukaryotic cells behave in a similar, though more complicated, manner. One major difference between the two is that in prokaryotic cells, transcription and translation (protein synthesis) may occur almost simultaneously, whereas in eukaryotic cells there is a gap between the two processes while the mRNA moves from the nucleus to the ribosome. The other major difference is that RNA in eukaryotic cells almost always requires processing after synthesis. Prokaryotic RNA is usually ready immediately after synthesis. Processing includes adding a cap, adding a poly(A) tail, and — in nearly all cases — splicing to remove introns.

Prokaryotic cells

RNA polymerase in *E. coli* contains four subunits that combine to form a holoenzyme designated $\alpha_2\beta\beta'\sigma$. The purpose of the σ subunit is to help find the promoter and to help initiate RNA synthesis. Once synthesis begins, this unit leaves the remainder, the core enzyme. The catalytic site in the core enzyme contains two divalent metal ions, one of which stays with the core and one that enters with the ribonucleoside triphosphate and leaves with

the cleaved pyrophosphate. Three aspartate residues aid in the binding of the metal ions. Although DNA polymerase and RNA polymerase have very different overall structures, their active sites are similar.

In the absence of the σ subunit, RNA polymerase would bind tightly to DNA at any point. When this unit is present, binding at other than a promoter site is significantly lower. Due to its reduced affinity, the holoenzyme can slide along the DNA strand until a σ subunit detects a promoter site. It binds to this site more strongly than to other positions on the DNA strand. The efficiency of this binding is one form of regulation. A number of σ subunits are present, each designed to recognize a different promoter site.

It's almost like tying knots in an anchor rope. A diver could swim upward holding on to the rope, but a knot signals a spot to stop and decompress.

Once the RNA polymerase arrives at a promoter site, it becomes necessary to unwind a 17 base-pair segment of the double helix and to unpair the bases. This unwinding converts a closed promoter complex to an open promoter complex. RNA polymerase is now ready to begin the RNA chain by incorporating the first nucleotide triphosphate. (Unlike DNA replication, no primer is necessary.) This first nucleotide triphosphate is usually a pppG or a pppA, which remains throughout transcription. This tap is at the 5' end of the new RNA molecule, and growth begins when a new nucleotide links to the 3' position (Figure 16-3).

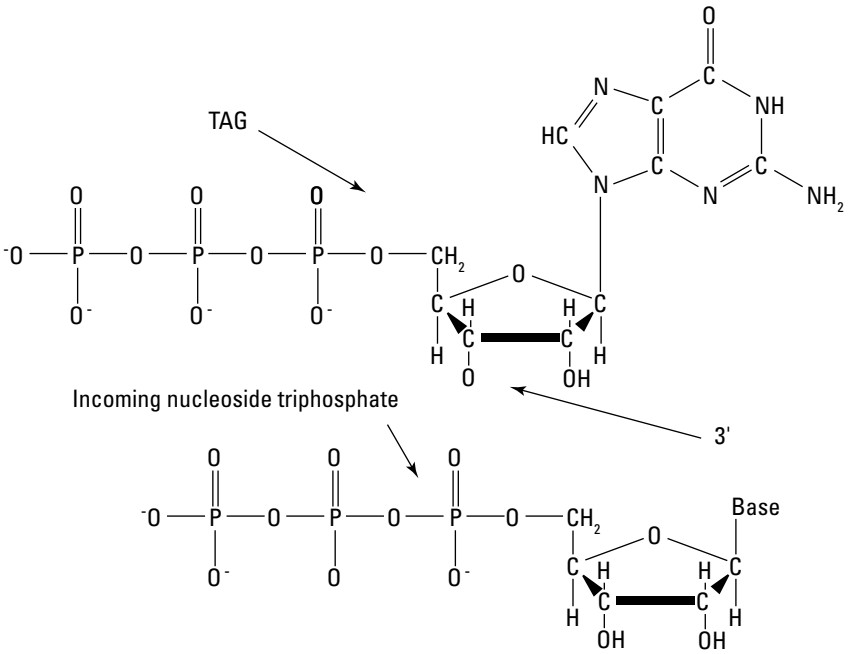


Figure 16-3: Linking of the second nucleotide to the tag, using pppG as an example (top), and linked nucleotides at the beginning of the chain (bottom).

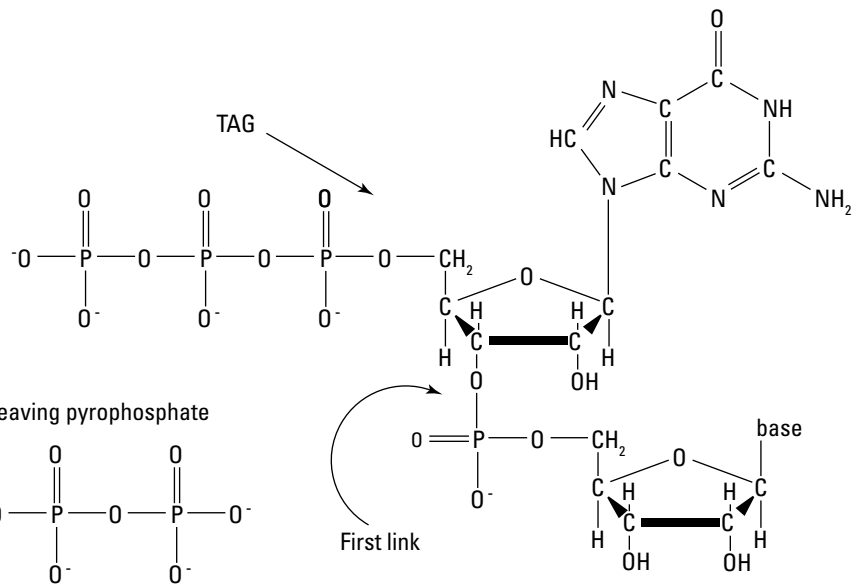


Figure 16-3: continued

Once the first two nucleotides link (through the formation of the linking phosphate diester) the σ subunit leaves. This allows the core enzyme to bind more tightly to the substrate. A transcription bubble now forms that contains the RNA polymerase, the unwound portion of the DNA, and the rapidly forming nascent RNA. Initially a short segment of the new RNA forms a hybrid helix with the DNA. This segment normally consists of about eight base pairs or one turn of the double helix. The growth rate is on the order of 50 nucleotides per second. (Compare this to DNA replication, which proceeds at about 800 nucleotides per second.) It is important to note that RNA polymerase does not "proofread" the new RNA. Thus, errors creep in at a higher rate than in replication. However, because the products do not pass to the next generation, there is no mutation or lasting effect. In any case, the next RNA stand to form lacks this defect and behaves correctly. One bad RNA in several hundred or more copies of the same gene is likely to have a minimal influence on the cell.

Elongation proceeds until the RNA polymerase encounters a termination signal, initiating a series of actions. At this point, formation of new phosphate diesters ceases, the RNA-DNA hybrid separates, the portion of the DNA chain that is still open rewinds, and the RNA polymerase separates from the DNA. There are different termination signals. One simple one is a palindromic (reading the same forward or backwards) GC-rich region followed by an AT rich region. The palindromic region is self-complementry, and these bases hydrogen bond to form a hairpin loop. The AT-rich region results in a number of U_{RNA}-A_{DNA} pairs, which have the weakest hydrogen bond interactions of all types of pairs. The formation of this hairpin and the AT region destabilizes, and the RNA-DNA hybrid and the nascent RNA begin to leave. See Figure 16-4.

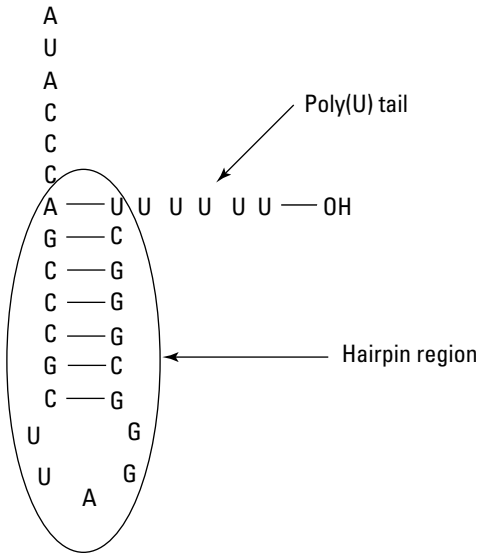


Figure 16-4: The hairpin and subsequent portion of the RNA.

Not all termination signals contain a hairpin and a U-rich segment. In at least some cases, RNA polymerase needs help. Evidence for this came from the observation that in vitro RNA chains were often longer than in vivo chains for the same RNA. Clearly, the in vitro RNA polymerase was unable to terminate elongation. The missing aid was a protein known as the *rho factor* (ρ). This protein wraps about the nascent RNA soon after the RNA exits the transcription bubble. In the presence of RNA, the ρ protein hydrolyzes ATP, which supplies energy. The protein first attaches to an RNA segment that is poor in guanine and rich in cytosine. Rho moves along the nascent RNA until it encounters the transcription bubble. At this point, it breaks the RNA-DNA hybrid and separates the nascent RNA. Other proteins serve a similar function as the rho factor.

In prokaryotic cells, mRNA is either ready or nearly ready to function immediately after release from the transcription (translation may begin before transcription terminates). However, both tRNA and rRNA require cleavage and other modifications of the nascent RNA chain. Various nucleases cleave the RNA in a very precise manner. It is possible to get more than one gene from a long nascent RNA strand. Processing may require the connection of a number of nucleotides — for example, all tRNA molecules need a CCA tail to function correctly. In some cases, there may be modification of the bases or ribose units.

Eukaryotic cells

Unlike prokaryotic cells, transcription and translation occur in different regions of the cell, leading to greater control of gene expression. Another difference is that eukaryotic cells extensively process mRNA in addition to rRNA

and tRNA. After RNA polymerase action, mRNA acquires a cap and a poly(A) tail. Nearly all mRNA molecules are spliced. Splicing involves removal of introns with the remaining exons being connected. Ninety percent of the nascent RNA may be introns.

Eukaryotic cells typically contain three types of RNA polymerase. Type I RNA polymerase (in the nucleolus) produces most forms of rRNA. Type II (in the nucleoplasm) produces mRNA and snRNA. And type III (in the nucleoplasm) produces tRNA and small rRNA molecules. (Actually, these polymerases only produce the pre-RNA forms of these molecules.)

Each of the three polymerases has a distinct type of promoter. These promoters may be in the same upstream sites as in prokaryotic cells, in downstream sites, or within the genes themselves. In addition to promoters, there may be enhancers. Enhancers, though not promoters, increase the effectiveness of a promoter. Enhancers for a single promoter may occur in different positions on the DNA chain and are important for gene regulation. Both promoters and most enhancers are on the same side of the DNA chain as the gene they regulate; for this reason, they are cis-acting elements. The promoters, as discussed earlier, are typically a TATA box (usually between –30 and –100), the CAAT box, and the GC box (both are usually between –40 and –150). Enhancers may appear upstream, downstream, or within the gene about to undergo transcription. Enhancers that are present on the opposite DNA chain are trans-acting factors, known as transcription factors, on the other DNA chain.

The typical series of events is that the transcription factor TFIID binds to the TATA box (TF stands for *transcription factor,* and the II means *RNA polymerase II*). Binding is the result of a small component of TFIID known as TBP (TATA-box-binding protein), which has an extremely high affinity for the TATA-box. When TBP binds to the DNA, substantial changes occur in DNA, including some degree of unwinding.

Other components utilized in transcription later attach to the TBP. These are, in order: TFIIA, TFIIB, TFIIF, RNA polymerase II, and finally TFIIE. This final group is the basal transcription complex. This example illustrates only one of numerous transcription factor initiations.

In eukaryotic cells, nearly all, if not all, products of transcription (precursors) undergo further processing before they reach their final active form. In general, tRNA precursors need to have the 5' leader removed, splicing to remove any and all introns, replacement of the poly(U) tail with a CCA sequence, and possible modification of some of the bases. Each of these processes requires one or more enzymes.

The precursors to the various forms of mRNA normally require the most modification. These precursors need, amongst other things, a 5' cap and a 3' poly(A) tail. The caps are cap 0, cap 1, and cap 2 — the numbers refer to the number of methylated ribose sugars (Figure 16-5). Caps are not present on tRNA, snRNA, or rRNA.

Most mRNA has a poly(A) tail not encoded by DNA. Usually, addition of this tail is preceded by cleavage of an intron portion of the mRNA precursor. The series AAUAAA signals where the cleavage will occur. This series is only part of the signal — the other part is uncertain. After cleavage, a poly(A) polymerase adds about 250 adenylate residues to the 3' end. The exact purpose of the tail is uncertain. It appears to enhance translation and increase the lifetime of the mRNA molecule.

In some cases, it is necessary to edit some mRNA precursors. *Editing* refers to an alteration of the base sequence other than that caused by splicing. An example is to chemically change one base into another. An example of editing occurs in the mRNA that encodes for apolipoprotein B (apo B). The entire protein contains 4,536 residues. However, a related 2,152-residue form is also important. The longer form, synthesized in the liver, is useful in the transport of lipids within the liver. The smaller form, synthesized in the small intestine, interacts with dietary fats. The same mRNA is responsible for both forms of the protein. In the small intestine, a deaminase acts on a specific cytosine and converts it to a uracil, which changes a CAA codon (Gln) to a UAA codon (stop) — which truncates the protein chain to yield the smaller form.

Splicing is a very common form of modification of all forms of RNA. *Splicing* involves the removal of introns and the joining of the exons to yield the final RNA molecule. Splicing must be very precise, as a miss by one base alters the entire sequence of codons present.

A number of different introns need to be removed. In eukaryotic cells, the intron begins with a GU and ends with an AG. Further refinement is present in vertebrates, where GU is the end of the sequence AGGUAAGU. A variety of AG sequences are found in higher eukaryotic cells. In general, one end of the intron loops about and connects to a point (the branch point) on the intron chain. Joining of the exons present then proceeds.

Spliceosomes are important in the splicing of mRNA precursors. These assemblages contain the mRNA precursors, several snRNAs, and proteins known as splicing factors. A group of snRNAs labeled U1, U2, U4, U5, and U6 are important. U1 binds to the 5' end of the splice site and then to the 3' end. U2 binds to the branch point, U4 blocks U6 until the appropriate moment, U5 binds to the 5' splice site, and U6 catalyzes the splicing. There are alternate splicing procedures.

It should be noted that alternate splicing can lead to production of different proteins from the same RNA.

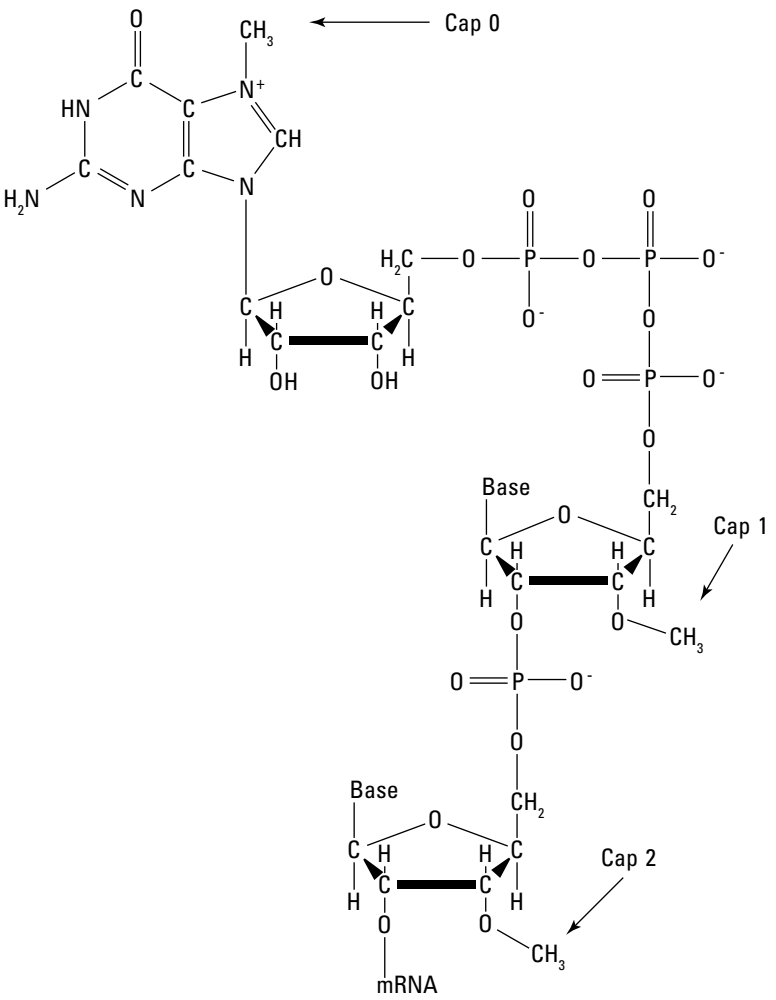


Figure 16-5: The general structure of a mRNA cap.

To Heck with Da Vinci: The Genetic Code

Just as DNA serves as the template for the generation of RNA, mRNA serves as the template for the generation of protein. In order to synthesize the appropriate protein, there needs to be a species that interacts with this template to assure the incorporation of the correct amino acid. The interaction species is tRNA. This relatively small form of RNA has two important regions: a template recognition site and the appropriate amino acid. The template recognition site is an anticodon, which corresponds to a codon on the mRNA. Attachment of the amino acid to the tRNA is by the action of an aminoacyl-tRNA synthetase. Each of the 20 amino acids has at least one

specific synthetase. This enzyme attaches the specific amino acid to the 3' terminal adenosine of the tRNA (Figure 16-6).

Codons

The genetic code contains the information necessary for the synthesis of proteins and consists of a set of three-letter words made from an alphabet containing four letters. Each three-letter word is a *codon.* This vocabulary is universal as it applies to all known living organisms.

The four letters are as follows:

- A, for adenine
- C, for cytosine

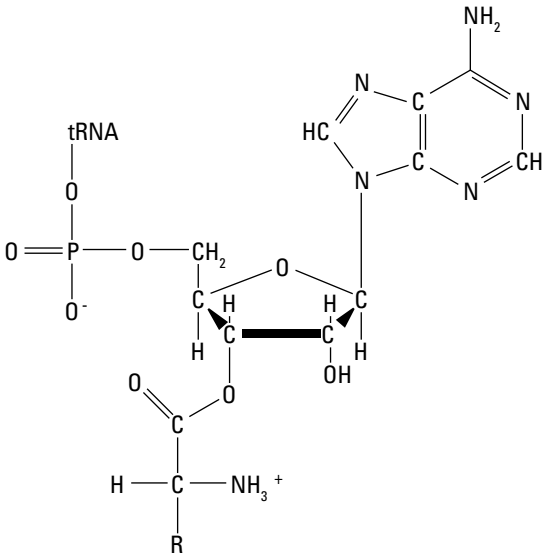


Figure 16-6: The attachment of an amino acid to the terminal adenosine.

- G, for guanine
- U, for uracil

The four letters give a total dictionary containing 64 words. Sixty-one of these words code for specific amino acids, and the remaining three words code for no amino acid. The codons coding for no amino acid are the "stop" signals.

Because there are only 20 amino acids to code for, the presence of 61 codons means that some amino acids can come from more than one codon.

Table 16-1 lists the genetic code.

Table 16-1 The Standard Genetic Code

Codon	*Amino Acid*	*Codon*	*Amino Acid*	*Codon*	*Amino Acid*	*Codon*	*Amino Acid*
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
UUA	Leu	UCA	Ser	UAA	***Stop***	UGA	***Stop***
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUG	Leu	UCG	Ser	UAG	***Stop***	UGG	Trp
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys

Analysis of the genetic code shows that two amino acids — methionine (see Figure 16-7) and tryptophan — only have one codon each. At the other extreme, three amino acids — arginine, leucine, and serine — each have six codons. The remaining 15 amino acids have at least two codons each. Amino acids with more codons are more abundant in proteins. Examining Table 16-1 shows that most *synonyms* (codons coding for the same amino acid) are grouped together and differ by a single base, usually the last base in the codon. Other correlations are present in the table. See what others you can find. The similarity of synonyms limits potential damage due to mutations.

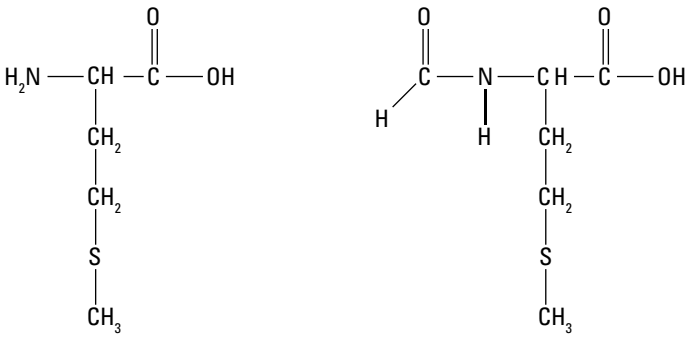


Figure 16-7: Structures of methionine and formylmethionine.

Alpha and omega

Although tRNA does not read the termination sequences, UAA, UAG, and UGA, specific proteins known as *release factors* read them. When a release factor binds to the ribosome, it triggers the release of the new protein, and release of the protein signals new synthesis to begin.

The stop signals are rather obvious on the Table 16-1, but what about the start? What signals the initiation of protein synthesis? The initiation sequence is usually AUG, the codon for methionine (Figure 16-8). In eukaryotic cells, additional factors come into play. In many bacteria, fMet (formylmethionine) is the initial amino acid (refer to Figure 16-7), which AUG usually codes for; however, GUG works sometimes.

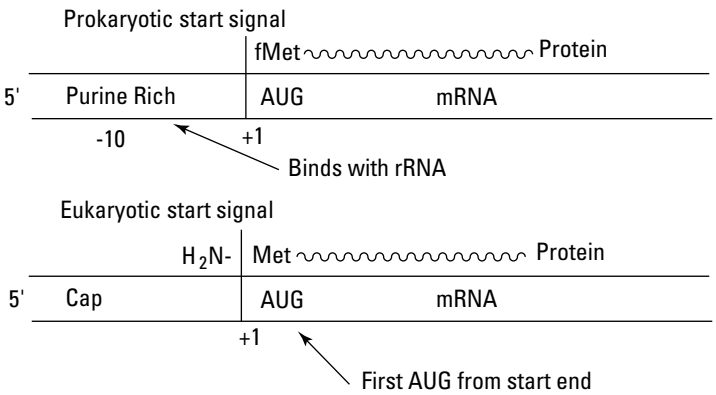


Figure 16-8: The start signals.

The genetic code is nearly universal — the codons correspond to the same amino acid in most cases. A few exceptions are known. For example, the code in mitochondrial DNA has several differences from normal DNA. In mitochondrial DNA, UGA is not a stop signal, but a codon for tryptophan.

In prokaryotic cells, coding for proteins is continuous, but this is not always true in the case of eukaryotic cells. In some mammals and birds, most genes are discontinuous. For example in the gene encoding for β-globin, there are regions that do not encode for a portion of protein. The gene contains about 1,660 base pairs — about 250 pairs on each end, plus an additional 500 pair segment code for the protein. These coding segments are *exons.* Two segments, one of about 120 base pairs and one of about 550 base pairs, do not code for protein. These non-coding regions are *introns.* The entire gene has, in sequence, a 240-pair exon, a 120-pair intron, a 500-pair exon, a 550-pair intron, and a 250-pair exon.

If an mRNA forms from a gene containing introns, it needs to undergo modification before it is of use. It is necessary to cut the intron regions from the mRNA and to splice the exon ends together to form the final mRNA molecule. In most cases, the intron portion begins with a GU and ends with a pyrimidine-rich segment ending with an AG. This combination signals the intron domain.

Models of Gene Regulation

The organism does not need to produce all the different proteins all the time. To control which proteins form at which time requires some form of *gene regulation.* When the organism requires a specific protein, it is necessary to "switch on" a certain gene — and once there is a sufficient quantity of that protein the gene must be "switched off." Control may occur either at the transcription level (gene regulation) or at the translation level.

In this section we examine processes in prokaryotic cells and then move on to the more complicated processes that take place in eukaryotic cells. The examination of the simpler mechanisms in prokaryotic cells gives insight into the processes in eukaryotic cells — the basic processes are similar.

As usual, our prokaryotic example is *E. coli.* Insight on gene regulation came when the diet of the *E. coli* was changed from glucose-rich to lactose-rich. For the cells to utilize this alternate energy source, they must generate the enzyme β-galactosidase. This enzyme is normally available at very low levels — a situation that quickly changes after replacing the glucose with lactose. One clue to the mechanism was that as the levels of β-galactosidase increased, so did the levels of galactoside permease (which transports lactose into the cell) and thiogalactoside transacetylase (which detoxifies other materials transported by galactoside permease). Thus, one change in the environment triggered multiple enzymes. This coordinated triggering of gene expression is

called an *operon*. Francois Jacob and Jacques Monad proposed the operon model to explain gene regulation.

The Jacob-Monod (operon) model

The simultaneous change in the levels of three different enzymes by one change in the environment suggested a link between the control mechanisms, and the operon model was created to account for this link. This model requires a regulator gene that affects a number of structural genes and an operator site. The operator and associated structural genes constitute the operon. The regulator gene is responsible for producing a *repressor protein*. The repressor protein binds to the operator site and prevents expression of the structural genes, as shown in Figure 16-9. The lac operon is one of the better understood operons.

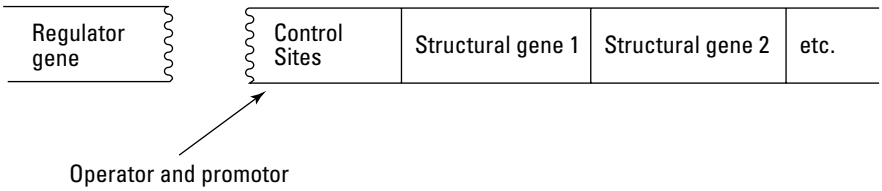


Figure 16-9: Diagram of a generic operon.

The multiple structural genes produce one large mRNA, and this single RNA strand is capable of generating a set of proteins. An mRNA that is capable of encoding for multiple proteins is *polygenic* or *polycistronic*.

The lac operon

The lac operon is the model regulatory system that, since its discovery in 1961, has provided extensive insight into how a cell regulates its genome. Figure 16-10 illustrates the lac operon.

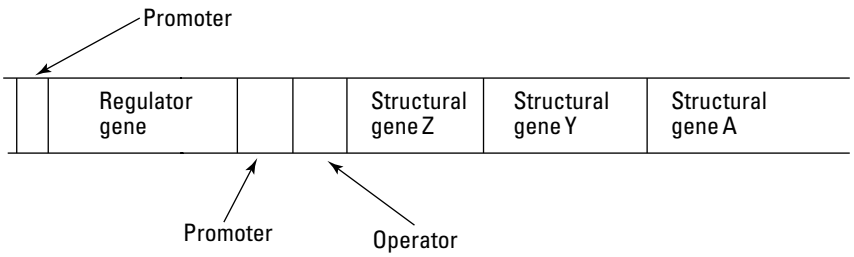


Figure 16-10: Diagram of the lac operon.

The *lac operator* is a palindromic DNA sequence with a twofold symmetry axis. The repeat is not always a perfect palindrome. (Many protein-DNA interactions involve a matching of symmetry.) The lac operator is as follows, with the center axis in bold:

TGTGTGGAATTGTGAGC**G**GATAACAATTTCACACA

ACACACCTTAACACTCG**C**CTAATGTTAAAGTGTGT

The *lac repressor* is a dimeric protein that can join to form a tetramer. In the absence of lactose, the repressor tightly binds to the operator. The presence of the repressor prevents RNA polymerase from unwinding the DNA strand to initiate transcription.

The presence of lactose is not the direct trigger of the lac operon; the trigger is allolactose. Both lactose and allolactose are disaccharides composed of galactose and glucose (Figure 16-11). In lactose there is an α-1,4 linkage, whereas in allolactose the linkage is an α-1,6. Allolactose results when the few molecules of β-galactosidase that are normally present in the cell first encounter lactose. This disaccharide along with a few similar molecules is an inducer of the lac operon. The inducer binds to the repressor and reduces the affinity of the latter to operator on the DNA. With its affinity reduced, the repressor detaches from the operator, and the DNA segment is now open for business.

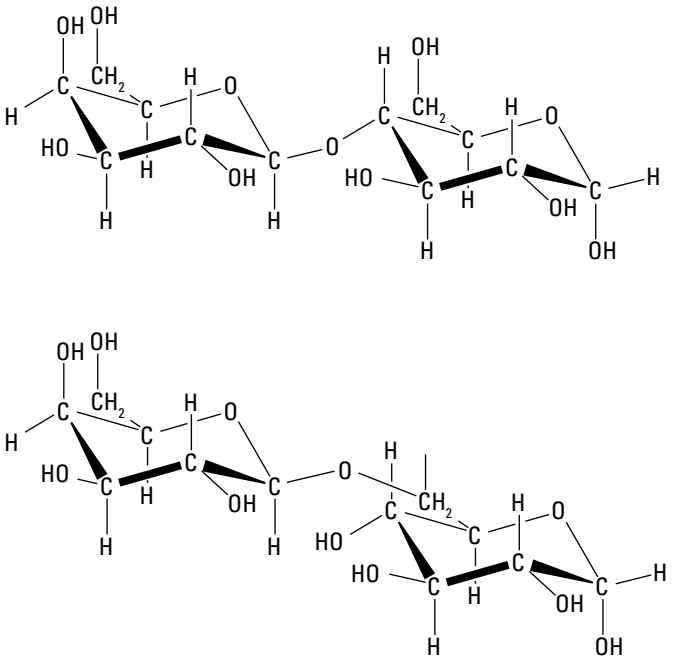


Figure 16-11: Structures of lactose and allolactose.

When transcription begins, all three structural genes become active, and the cell begins producing β-galactosidase, galactoside permease, and thiogalactoside transacetylase. This continues until the lactose and hence the allolactose concentration falls so that the repressor proteins are available to reattach to the DNA.

Other prokaryotic regulators

The *pur repressor* affects the genes responsible for the biosynthesis of purines and, to a lesser extent, pyrimidines. This protein is similar in structure to the lac repressor; however, the pur repressor only binds to the operator after another molecule binds to the repressor. Therefore, while the binding of another molecule releases the lac repressor, the binding of another molecule causes the pur repressor to bind. The other molecule has an opposite affect. In the case of the pur repressor, the other molecule is a *corepressor.*

There are also regulators that stimulate transcription instead of repressing it. The catabolite activator protein (CAP) is one example. This protein interacts with the promoter and, along with two cAMP molecules, interacts with RNA polymerase. This interaction leads to stimulating the initiation of transcription of certain genes.

Regulation of eucaryotic genes

Although there are similarities, the regulation of genes in eukaryotic cells is more complex than in prokaryotic cells. One reason for this is that the typical eukaryotic genome is much larger than the typical prokaryotic genome. Another source of complexity is that many eukaryotic cells are part of a larger organism and do not serve the same purpose as other cells do within the same organism. For example, although some of the proteins are the same, a liver cell must produce a different overall set of proteins than a heart cell does.

Histones

Eukaryotic DNA has a group of proteins associated with it. These small, basic proteins are called *histones.* They are basic because approximately 25 percent of the amino acid residues present are either arginine or lysine. These are tightly bound to the DNA and total approximately half of the mass of a chromosome. A complex of the cell's DNA and associated protein is a *chromatin,* and there are five important histones present in chromatin: H1 — and four that associate with each other: H2A, H2B, H3, and H4.

A chromatin apparently consists of repeat units consisting of two copies each of H2A, H2B, H3, and H4, with a strand of DNA consisting of about 200 base pairs tightly wrapped around this histone octomer. Each of these repeating units is a nucleosome. The wrapping of the DNA to form a nucleosome yields a significant compaction of the DNA. Research indicates that about 145 of

the 200 base pairs are actually associated with the histone octomer, and the remaining base pairs are linker DNA that link one histone octomer to the next. Histone H1 usually binds to linker DNA.

The eight histones in a histone octomer are arranged into a tetramer with the composition $(H3)_2(H4)_2$ and two dimers each with the composition (H2A)(H2B). All the histone proteins have long tails rich in arginine and lysine residues that extend out of the core. Modification of these tails is important for gene regulation.

The structure of chromatin is a factor in eukaryotic gene regulation. For a gene to be available for transcription, the tightly packed chromatin structure must open. In addition, the structure regulates access to regulatory sites on DNA. Enhancers disturb this structure, explaining why enhancers can have an effect on the expression of a gene even though the enhancer site may be thousands of base pairs away from the gene. Certain enhancers only occur in specific types of cells. Thus, the genes they enhance are only expressed in these cells. For example, the gene to produce insulin is expressed only in pancreatic cells.

A modification of DNA can also inhibit gene expression. Approximately 70 percent of the 5'-CpG-3' sequences in mammals have the cytosine methylated. The distribution of the methylated cytosines (Figure 16-12) varies with cell type. Regions in chromatins necessary for gene expression in that cell are hypomethylated (have fewer methylated cytosines), relative to similar regions in cells where no expression of the gene occurs. The presence of the methyl group interferes with the binding of enhancers and promoters.

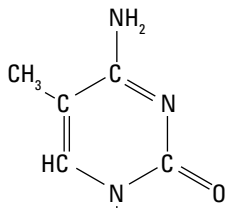


Figure 16-12: Structure of methylated cytosine.

Mediating transcription

Eukaryotic cells require a variety of transcription factors to initiate transcription — no factor can carry out the entire process on its own. This is called *combinatorial control,* and it is necessary in organisms with multiple cell types and helpful in other eukaryotic cells.

A number of nonpolar molecules, such as the steroid hormones, can easily pass through the hydrophobic cell membrane and bind to receptor proteins. They are very specific. Estrogen (Figure 16-13) is one example of a steroid hormone. Such molecules are known as *ligands.*

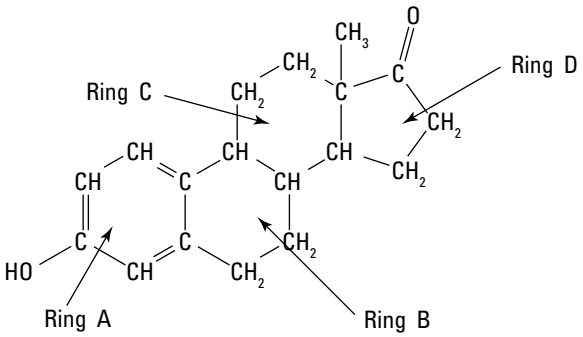


Figure 16-13: Structure of estrogen.

The ligand binds to a specific site — called, helpfully, the *ligand-binding site* — which is present near the end of a receptor protein. This portion of the protein contains many nonpolar residues, which have an affinity for hydrophobic molecules. Receptor proteins that bind hormones are called *nuclear hormone receptors.* There is a DNA binding site near the center of the protein that contains eight cysteine residues, which are necessary to bind zinc ions, four residues for each. The presence of the zinc ions stabilize structure and led to the name *zinc finger domains.* (There are other cysteine residues and zinc ions nearby.) The binding of a molecule to the ligand-binding site causes a significant structural rearrangement of the protein. This situation would seem to be similar to the lac repressor in prokaryotic cells; however, experiments indicate that there is no significant alteration in binding affinity

The next part of the puzzle involves a number of small proteins known as *coactivators.* Near the center of each of these are three regions with the pattern Leu-X-X-Leu-Leu. Each of these regions generates a short hydrophobic α-helix. These three helices bind to a hydrophobic region on the ligand-binding region. The presence of the ligand appears to enhance the binding of a coactivator. (A receptor protein may act as a repressor, especially in the presence of a corepressor.)

Just what are the roles of coactivators and corepressors? Their effectiveness appears to be linked to their ability to covalently bond to the tails of the histones. Histone acetyltransferases (HATs) catalyze this modification of the histone tails (a process that is reversed by histone deacetylase enzymes — see Figure 16-14). This process changes a very polar (positively charged lysine) to a much less polar (neutral) amide, resulting in a significant reduction in the affinity of the tail to the associated DNA. To a lesser degree, it reduces the affinity of the entire histone to the associated DNA. The reduction in the affinity allows access of a portion of the DNA to transcription.

The acetyated lysine residues also affect the acetyllysine-binding domain (the *bromodomain*) present in many of the eukaryotic transcription regulatory proteins.

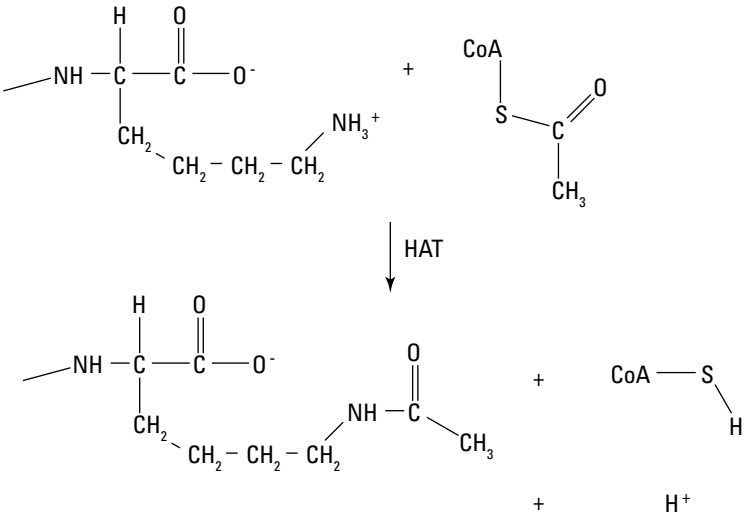


Figure 16-14: Reaction catalyzed by histone acetyltransferases (HATs).

There are two important bromodomain-containing proteins: One of these is a large complex of ten proteins that binds to the TATA-box-binding protein that is responsible for the transcription of many genes. The other proteins containing bromodomains are part of large complexes known as *chromatin-remodeling engines.* As the name implies, these proteins alter the structure of the chromatins, which changes the behavior of the chromatin.

All these factors alter the availability of portions of the DNA structure to transcription. Once the DNA becomes open, the procedures discussed earlier in the chapter come into play.

Chapter 17

Translation: Protein Synthesis

In This Chapter

- Talking about translation
- Looking at protein synthesis
- Examining eukaryotic cells
- Discussing the Human Genome Project

You are no doubt familiar with the process of translation — converting text from one language into another. The process of translation in biochemistry does exactly the same thing.

Hopefully Not Lost in Translation

Translation is the process where the four-letter alphabet of the nucleic acids becomes the twenty-letter alphabet of proteins. In doing so, genetic information is passed on. Translation occurs in the cell's ribosomes, which contain ribosomal RNA (rRNA). The information necessary for translation travels from the cell nucleus to the ribosomes via messenger RNA (mRNA). The messenger RNA binds to the smaller ribosomal body, and transfer RNA (tRNA) brings amino acids to it.

Why translation is necessary

The purpose of translation is to put together specific amino acids in a specific order to produce a specific protein. Messenger RNA provides the template or blueprint for this process. To utilize this template, something must bring the amino acids to the mRNA, and that thing is transfer RNA (tRNA). Transfer RNA has two important sites. One site is for the attachment of a specific amino acid. For example, only one specific type of tRNA will transfer the amino acid methionine. The other site is the recognition site, which contains an

anticodon. An *anticodon* is a sequence of three bases that match a codon on the mRNA. A codon sequence of AUG on the mRNA matches the UAC anticodon on a tRNA. All of this takes place in the ribosome, home of rRNA.

Home, home in the ribosome

The *ribosome* is the factory that produces proteins. Thousands of ribosomes are present in even the simplest of cells. They are complex units composed of RNA and protein. It is possible to dissociate a prokaryotic ribosome into two units. One unit is the 50S, or large unit, and the other is the 30S, or small unit. The large unit contains 34 different proteins, labeled L1 through L34, and two RNA molecules, labeled 23S and 5S. The small unit contains 21 different proteins, labeled S1 through S21, and an RNA molecule labeled 16S. The RNA molecules in the ribosomes act as transfer RNA in translation.

A prokaryotic ribosome contains three rRNA molecules (23S, 16S, and 5S), one copy of proteins S1 through S21, two copies of L7 and L12, and one each of the other L1-L34 proteins. L7 and L12 are identical except that L7 has an acetylated amino terminus. S20 and L26 are identical. Mixing the constituents in vitro leads to the two subunits reconstituting themselves. A version of the structure of the 16S form of ribosomal RNA appears in Figure 17-1.

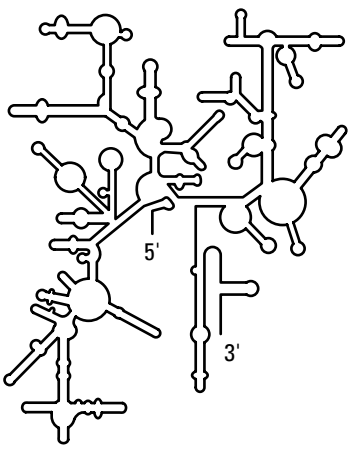


Figure 17-1: Simplified schematic of the structure of the 16S form of ribosomal RNA.

The Translation Team

A number of players, along with the rRNA, must interact in order to form a protein molecule. In addition, the structure of the ribosome is important to controlling protein synthesis. Both the rRNA and protein molecules control this structure. One possibly helpful analogy is the game of football.

The team captain: rRNA

RNA makes up approximately two-thirds of the mass of a ribosome. The three mRNA units play a key role in the shape and function of the ribosome (the proteins apparently fine-tune the shape and structure of the ribosome). The three mRNA form from the cleaving and processing of transcribed 30S RNA. A significant portion of each of the rRNA molecules have numerous duplex regions (short stretches of base-paired RNA).

The 30S and 50S rRNA subunits combine to form a 70S ribosome, which holds an mRNA in place during translation. There are also three sites for various tRNA molecules: the E, P, and A sites. The E site is the exit site. A tRNA occupies this position after delivery of its amino acid and just before exiting the ribosome. The P site is the peptidyl site, which holds the tRNA containing either the initial amino acid or the C-terminal amino acid of a protein chain. Finally, the A site is the aminoacyl site, which holds the tRNA attached to the next amino acid in sequence. When the 30S and 50S subunits join, they create A and E sites at the interface of the subunits. The P site of the 50S unit is the opening of a tunnel through which the growing protein chain passes out of the ribosome.

Here's the snap: mRNA

The base sequence of the mRNA is read in the 5' $\rightarrow$ 3' direction, and transcription occurs in this same direction. (Prokaryotic cells sometimes take advantage of this by beginning translation before transcription is over. This situation cannot occur in eukaryotic cells because the E sites of transcription and translation are physically separated.) The mRNA resulting from transcription gains a cap and a poly(A) tail before it ventures out of the nucleus on its trip to the ribosome.

Translation does not begin at the 5' terminus of the mRNA molecule. Just as there is a "stop" signal to terminate translation, there is a "start" signal. The 5' terminus base-pairs with the 3' terminus of the 16S rRNA. This region is normally about 30 nucleotides in length (a portion of this region, called the Shine-Dalgarno sequence, is purine-rich).

Shortly after this sequence is the start signal. In most cases, the start signal is AUG (methionine), though there are instances where the signal is GUG (valine). In *E. coli,* the first amino acid is formylmethionine instead of methionine. The formylmethionine is usually removed soon after translation begins. In prokaryotic cells, there may be more than one start and stop signal because many of the mRNA molecules are polygenic (polycistronic) — that is, they produce more than one protein. The structures of methionine and formylmethionine attached to tRNA are illustrated in Figure 17-2.

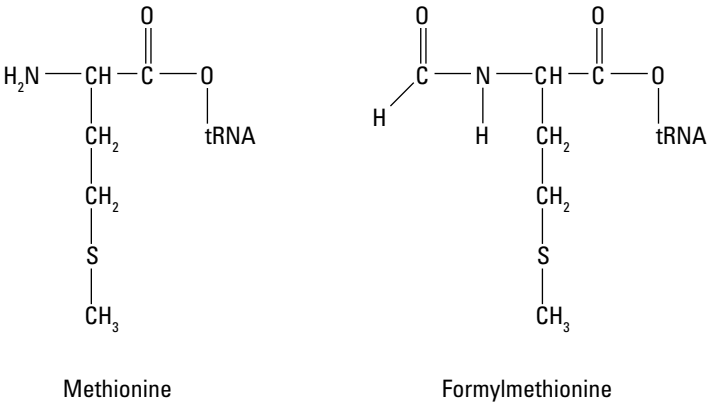


Figure 17-2: The structures of methionine and formylmethionine attached to tRNA.

Carrying the ball: tRNA

Several features are common to all forms of tRNA. Each form of tRNA is a single strand containing between 73 and 93 nucleotides. There are between seven and fifteen unusual bases (not one of the usual four, A, C, G, or U) in each molecule. Approximately half of the nucleotides present are base-paired. The activated amino acid is attached to the hydroxyl group at the 3'-end of the chain. The hydroxyl group is on the adenosine residue of a CCA segment. The other end, the 5'-end, is phosphorylated. The phosphorylation usually is a pG. The anticodon is contained in a loop near the center of the molecule.

Many of the unusual bases are methylated or dimethylated forms of A, C, G, or U, which are usually the result of post-transcription modification of the molecule. The presence of the methyl groups interferes with the formation of some base pairs, which prevents certain additional interactions. Methyl groups are nonpolar, so their presence makes regions of the tRNA hydrophobic, which affects their interaction with ribosomal proteins and syntheses. The unusual bases include dihydrouridine, dimethylguanosine, inosine, methylguanosine, methylinosine, pseudouridine, and ribothymidine. Inosine, shown in Figure 17-3, is part of the anticodon. Many of these are in or near the bends in structure of tRNA.

There are five regions, shown in Figure 17-4, that are not base-paired. (Note that the structure of tRNA shown in Figure 17-4 is not the actual three-dimensional structure of tRNA.) Starting at the 5'-end, the unpaired regions are, in order, the DHU loop, the anticodon loop, the extra arm, the TψC loop, and the 3-CCA terminus. (The name of the DHU loop derives from the presence of several dihydrouracil residues. The anticodon loop contains the segment that recognizes the codon on the mRNA, and the extra arm contains a variable number of

residues. The TψC loop derives its name from the presence of the sequence thymine-pseudouracil-cytosine.) These loops make each tRNA different, even though the overall structure is the same.

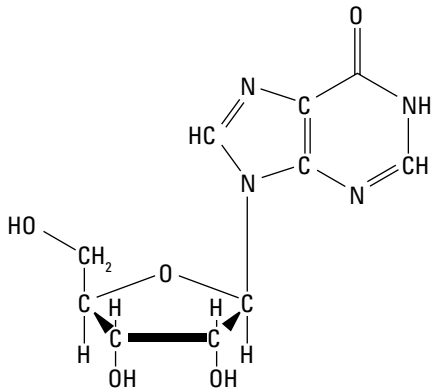


Figure 17-3: The structure of inosine.

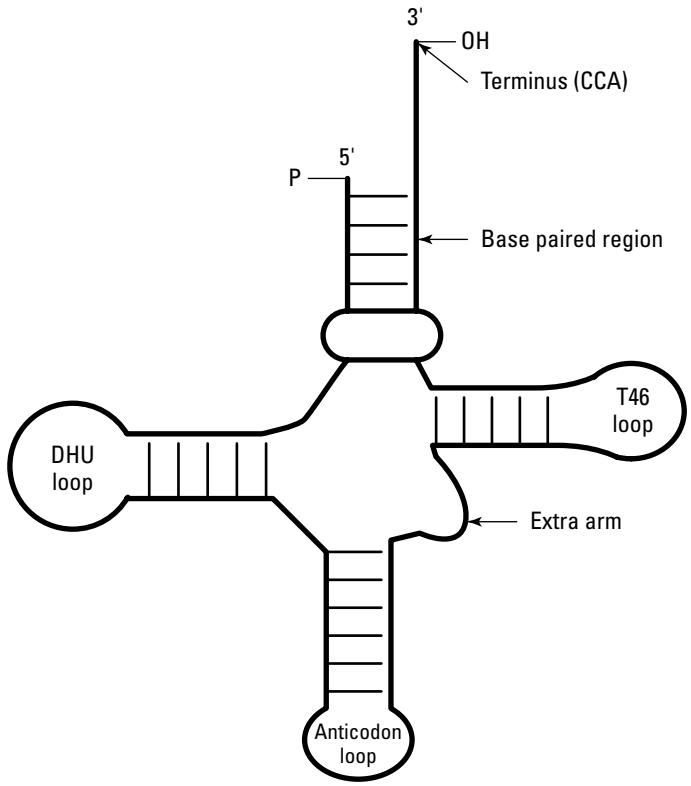


Figure 17-4: Some aspects of the structure of tRNA.

The anticodon is present in the 5' → 3' direction, and it base-pairs to a codon in the 3' → 5' direction. This matches the first base of the anticodon with the third base of the codon. (Don't forget the convention of writing base sequences in the 5' → 3' direction.)

Charging up the middle: Amino acid activation

It is imperative that the correct amino acid attaches to the tRNA because the presence of an incorrect amino acid or the absence of any amino acid would be devastating to translation. Connection of the amino acid to the tRNA activates the amino acid. Joining free amino acids is a nonspontaneous process, however, connecting the amino acid to the tRNA changes the free amino acid to a more reactive amino acid ester. The amino acid-tRNA combination is an aminoacyl-tRNA or a charged tRNA (Figure 17-5).

Specific aminoacyl-tRNA synthetases, called *activating enzymes,* catalyze the activation reaction. The process begins with an amino acid and an ATP forming an aminoacyl adenylate, which leads to the release of a pyrophosphate. Figure 17-6 shows an aminoacyl adenylate.

There is a separate aminoacyl-tRNA synthetase for each amino acid.

The two classes of aminoacyl-tRNA synthetases are denoted Class I (monomeric) and Class II (usually dimeric). Each class is responsible for ten amino acids. The CCA arm adopts different structures when interacting with members of the different classes, and ATP adopts a different conformation when interacting with members of different classes. Most Class II examples attach the amino acid as illustrated back in Figure 17-5, whereas Class I examples attach the amino acid to the alternate linking site. Some aspects of the structure of tRNA appear in the schematic structure shown in Figure 17-4.

The conversion of an aminoacyl adenylate, once formed, remains tightly bound to the synthetase until it can form an aminoacyl-tRNA.

In order to make sure that the aminoacyl-tRNA synthetase incorporates the correct amino acid, the enzyme must take advantage of specific properties of the amino acids. Examining the amino acids serine, valine, and threonine can give some insight into the selection process. These three amino acids appear in Figure 17-7, where they are drawn to emphasize similarities in the side-chain. (Recall that the threonine side-chain is chiral, but the others are not.) It is important to realize that there are size differences (–H for $–CH_3$) and

hydrogen bonding differences (–OH can, but $–CH_3$ cannot). The recognition site has the proper size and composition to take advantage of these specific properties. A significant species in this site is a zinc ion, which coordinates to the enzyme and the amino acid.

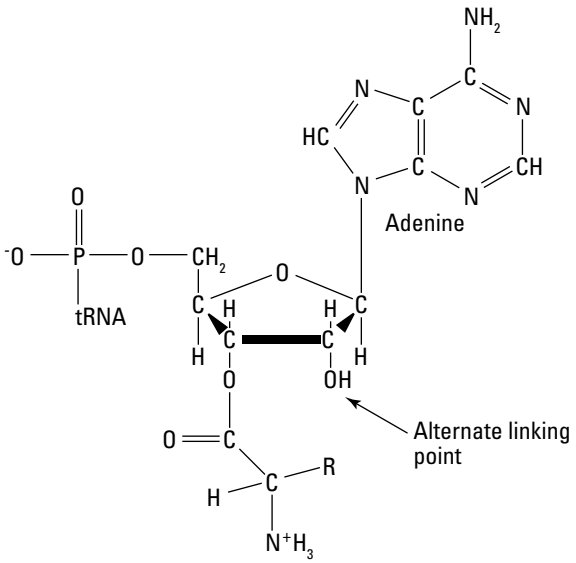


Figure 17-5: An example of an aminoacyl-tRNA.

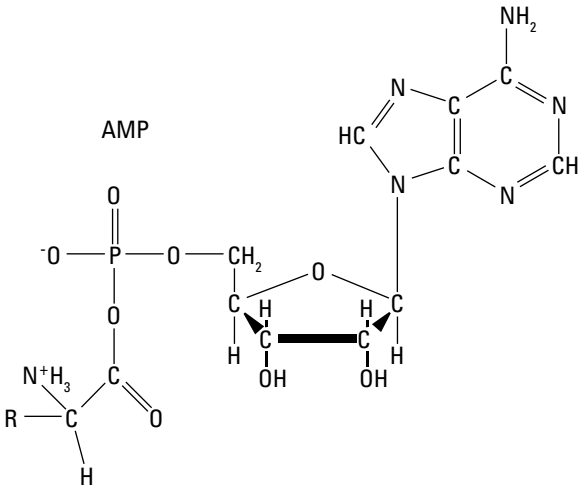


Figure 17-6: Structure of an aminoacyl adenylate.

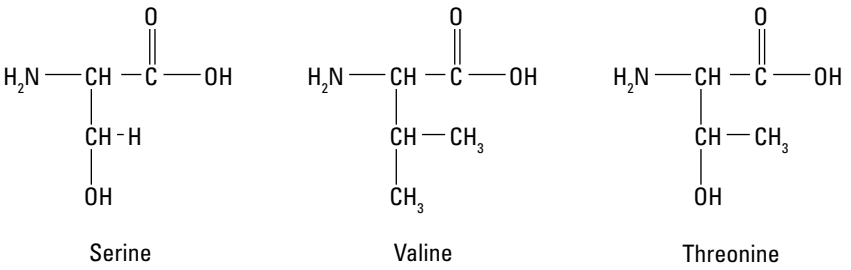


Figure 17-7: Structures of serine, valine, and threonine.

Even with these differences, serine sometimes replaces threonine. Fortunately, the enzyme includes an editing feature — the editing site is near the reaction site, but it is not the same. Similar editing occurs in other aminoacyl-tRNA synthetases. Amino acids, such as tryptophan, do not have closely similar analogues; thus, editing is far less important in these cases.

The aminoacyl-tRNA synthetases need to be able to recognize the anticodon present to make sure they interact with the appropriate tRNA, matching it to the correct amino acid. The enzymes may recognize other features of the tRNA structure. These features include the size of the extra arm and the hydrophobic character imparted by methylating some of the ribonucleotides.

Hooking Up: Protein Synthesis

The major steps in protein (polypeptide) synthesis are as follows:

- Activation
- Initiation
- Elongation
- Termination

These basics apply to all living organisms — there are no differences between human translation, fungi translation, or tulip translation. Synthesis proceeds from the amino to the carboxyl direction of the protein.

In this section we discuss these in greater detail. These steps involve tRNA, mRNA, and rRNA — along with a number of protein factors.

Activation

As mentioned earlier in this chapter, during *activation* an amino acid reacts with ATP to give aminoacyl adenylate. The aminoacyl adenylate then reacts

with a specific tRNA to give aminoacyl-tRNA plus AMP. This constitutes one of the players necessary for the translation game.

Initiation

During *initiation,* an mRNA attaches to a ribosome by interacting, through the Shine-Delgarno sequence, to the 30S rRNA subunit. Then the anticodon of the first tRNA attaches to the AUG (or GUG) codon on the mRNA. This occupies the P site of the 30S subunit. The amino acid extends into the P site of the 50S subunit. The 30S and 50S portions of the rRNA combine to produce the 70S ribosome. The combination of the two subunits allows the tRNA to interact with both parts.

In order to initiate translation it is necessary to bring the mRNA and the first tRNA to the ribosome. Three proteins, known as *initiation factors,* accomplish this task: IF1, IF2, and IF3. First, the 30S ribosome subunit, IF1 and IF3, form a complex. The two initiation factors bound to the 30S subunit interfere with a premature joining of the 30S and 50S subunit without the necessary mRNA. The remaining initiation factor, IF2, binds to GTP. The IF2-GTP combination binds to the initiator-tRNA, and the IF2-GTP-initiator-tRNA unit binds to the mRNA. Interaction of the Shine-Dalgarno sequence and the 16S rRNA manipulates the incoming group into the correct position.

Combining all these units with the 30S subunit gives the 30S initiation complex. Hydrolysis of the GTP as the 50S subunit approaches leads to expulsion of the initiation factors. With the initiation factors out of the way, the remaining moieties join to give the 70S initiation complex. (Wow, trying say that three times fast!) Once this complex forms, elongation can begin.

Elongation

During *elongation,* a second activated tRNA comes into the A site (which is adjacent to the P site) on the 30S subunit, where it binds to the appropriate codon. The activated tRNA is brought to the A site by a protein known as *elongation factor Tu* or *EF-Tu.* EF-Tu forms a complex with the activated tRNA (in the GTP form), and this complex protects the ester linkage holding the amino acid to the tRNA. In addition, the complex does not allow the activated tRNA to enter the A site if there is not a codon-anticodon match. EF-Tu interacts with all tRNAs except the initiator-tRNA. The energy needed for the EF-Tu to leave the tRNA in the ribosome comes from the hydrolysis of the GTP unit induced by the protein known as *elongation factor Ts.*

The two amino acids extend into the peptidyl transferase center of the ribosome. The amino group of the aminoacyl-tRNA from the A site is held in position to attack the ester linkage of the aminoacyl-tRNA in the P site. The catalyzed formation of the peptide bond occurs, accompanied by separation from the tRNA in the P site. The protein is now attached to the A site (30S).

With the loss of its amino acid, the tRNA no longer interacts in the same way with the ribosome. The tRNA moves to the E site of the 50S subunit as the next RNA, with its attached polypeptide, moves to the P (tunnel) site of the same subunit. The ribosome must now move over (the fancy way to say it is *translocate*) by one codon. For translocation to occur, it is necessary to employ the elongation factor G enzyme (EF-G or translocase is the protein that aids translocation). The hydrolysis of GTP to GDP supplies the energy for the move. This move places the polypeptide-tRNA into the P site of the 30S subunit. At the same time, the amino acid–stripped tRNA disengages from the mRNA and moves into the E site of the same subunit. Throughout this process, the polypeptide chain remains in the P site of the 50S subunit.

The first tRNA leaves the E site. Now the elongation cycle can begin again with the entry of another tRNA carrying the next amino acid. The process continuously cycles until a "stop" signal codon.

Termination

A "stop" signals *termination,* which results in the release of the protein, the last tRNA, and the mRNA.

Recall that the stop signals are UAA, UGA, and UAG.

Normal cells do not contain tRNAs with anticodons complementary to these codons. However, proteins known as *release factors* (RF) recognize these three codons. Release factor 1, RF1, recognizes UAA and UAG. Release factor 2, RF2, recognizes UAA and UGA. Release factor 3, RF3, is an intermediary between RF1, RF2, and the ribosome. The release factors carry a water molecule into the ribosome in place of an amino acid. The final reaction, the one that releases the newly formed protein, is the hydrolysis of the last ester linkage to a tRNA. The water brought in by the release factors is necessary for this hydrolysis.

The 70S ribosome remains together for a short time. Dissociation of the complex is mediated by a ribosome release factor and EF-G. GTP supplies the energy for this process.

The wobble hypothesis

Experimental studies have found that even pure tRNA molecules are capable of recognizing more than one codon. Biochemists developed the *wobble hypothesis* to explain this behavior, and subsequent work has firmly established this hypothesis.

The presence of the unusual base, inosine (shown back in Figure 17-3), in the anticodon loop is the key to understanding the wobble hypothesis. This base is capable of base pairing with adenine, cytosine, or uracil, allowing for some variation, or wobble, in the matching of codon to anticodon. The presence of inosine increases the number of different codons a particular tRNA can read. The first two bases in the codon pair to the corresponding bases in the anticodon. The third base is the wobble position.

Review the table of codons (Table 16-1 in Chapter 16) and see which amino acids depend only on the first two bases. Hint: Look at valine.

The base-pairing rules for the wobble hypothesis are shown in Table 17-1. The presence of an A or C as the first base allows the reading of only one codon. The presence of a G or U allows the reading of two codons, whereas an I allows the reading of three codons. Inosine is a useful base for allowing wobble; however, as Table 17-1 shows, it is only when the first anticodon base is an A or a C that there is no wobble. In general, the base in the wobble position forms weaker hydrogen bonds than normal because of the strain in the environment. The weaker hydrogen bonding aids in the loss of the tRNA after it delivers its amino acid.

Table 17-1 Base-pairing Rules for the Wobble Hypothesis

Base on Anticodon (1st Base)	*Bases Recognized on Codon (3rd Base)*
A	U
C	G
G	U, C
U	A, G
I	U, C, A

Four codons code for valine, comprising a four-codon family. If you examine three of the codons for valine: GUU, GUC, and GUA, they would all pair to the anticodon CAI instead of the anticodons CAA, CAG, and CAU. For this reason,

one CAI anticodon replaces three other anticodons. The remaining valine codon is GUG, which requires the synthesis of only two types of tRNA instead of four. Other four-codon families also work this way.

The only cases where the codons for a particular amino acid differ in the first two bases are the six-codon families, which are those of arginine, leucine, and serine. These families require three different tRNAs.

The presence of wobble reduces the number of necessary tRNAs in a cell from 61 to 31. However, cells usually have some number of tRNAs between these extremes. All the tRNAs coding for a specific amino acid require only one aminoacyl-tRNA synthetase.

Variation in Eukaryotic Cells

All cells follow the same basic pattern for translation. However, eukaryotic cells show some variations. More proteins are necessary to mediate translation, and the steps are, in general, more complicated.

Ribosomes

In eukaryotic cells, the ribosomes contain a 60S subunit and a 40S subunit, which combine to produce an 80S ribosome. The 40S subunit contains an 18S rRNA analogous to the 16S in the 30S subunit. There are three rRNA components in the 60S subunit: a 5S and a 23S, analogous to the 5S and the 23S of the prokaryotic 50S subunit, and a unique 5.8S rRNA.

Initiator tRNA

In eukaryotic cells, the initiator amino acid is methionine instead of formylmethionine. As in prokaryotic cells, a special tRNA is necessary for the first tRNA — a modification of the normal methionine-carrying tRNA.

Initiation

AUG is the only initiator codon in eukaryotic cells, and this is always the AUG nearest the 5' end of the mRNA. There is no purine-rich sequence immediately before this as in prokaryotic cells. The 40S ribosome subunit attaches to the mRNA cap and moves base by base in the 3' direction until it reaches

The Human Genome Project

The U.S. Human Genome Project was begun in 1990. It was originally scheduled to last for 15 years but because of rapid advances in the field of biotechnology it finished two years ahead of schedule in 2003. The U.S. Department of Energy and the National Institutes of Health coordinated the projects.

Goals

The Project had the following goals:

- Identify all the 20,000–25,000 genes in human DNA.
- Determine the sequences of the approximately 3 billion base pairs in human DNA.
- Store the information in databases.
- Improve data analysis tools.
- Transfer the developed technology to the private sector.
- Address the ethical, legal, and social issues associated with the project.

In addition to human DNA, researchers also studied the genetic blueprints of *E. coli*, a common bacterium found in humans as well as mice and fruit flies. The goal of transferring the technology to the private sector was included to develop the infant biotechnology industry and encourage the development of new medical applications.

Potential Benefits

Some potential benefits of the Human Genome Project include the following:

- Improved disease diagnosis
- Earlier detection of genetic predispositions to disease
- Drug design and gene therapy
- Creation of new biofuels
- More effective ways of detecting environmental pollutants
- Studying evolution through mutations in lineages
- Forensic identification of subjects through DNA analysis
- Establishing paternity
- Matching organ donors and patients
- Creation of insect- and disease-resistant crops
- Creation of biopesticides
- Increased productivity of crops and farm animals

Many of these potential benefits are showing up in our everyday life already.

Ethical, Legal, and Social Issues

One of the unique aspects of the Human Genome Project was that it was the first large scientific project that studied and addressed potential ethical, legal, and social implications that arose from the data generated from the study. Questions such as the following were addressed:

- Who should have access to personal genetic information?
- Who controls and owns genetic information?
- How reliable and useful is fetal genetic testing?
- How will genetic tests be checked for reliability and accuracy?
- Do parents have the right to test their children for adult-onset diseases?
- Do people's genes influence their behavior?
- Where is the line between medical treatment and enhancement?
- Are genetically modified foods safe for humans?

Many questions have been raised — but, as yet, few answers have resulted.

an AUG codon. The hydrolysis of ATP by helicases powers this process. Many more initiation factors are present in eukaryotic cells. A eukaryotic initiation factor has the symbol eIF instead of IF.

Elongation and termination

The EF-Tu and EF-Ts prokaryotic elongation factors have the eukaryotic counterparts EF1α and EF1$\beta\gamma$. Translocation is driven by eukaryotic EF2 with the aid of GTP. Only one release factor, eRF1, is present in eukaryotic cells instead of the two factors in prokaryotic cells. To prevent the reassembly of the two ribosome subunits, eIF3 functions like the IF3 protein in prokaryotic cells.