CHAPTER 4

DNA Synthesis *In Vivo* and *In Vitro*

Biotechnology

http://dx.doi.org/10.1016/B978-0-12-385015-7.00004-1

INTRODUCTION

Replication copies the entire set of genomic DNA so that the cell can divide in two. During replication, the entire genome must be uncoiled and copied exactly. This elegant process occurs extremely fast in *E. coli*, where DNA polymerase copies about 1000 nucleotides per second. Although the process is slower in eukaryotes, DNA polymerase still copies 50 nucleotides per second. Many biotechnology applications use the principles and ideas behind replication; therefore, this chapter first introduces the basics of DNA replication as it occurs in the cell. We then review some of the most widely used techniques in genetic engineering and biotechnology, including chemical synthesis of DNA, polymerase chain reaction, and DNA sequencing.

REPLICATION OF DNA

To maintain the integrity of an organism, the entire genome must be replicated identically. Even for plasmids, viruses, or transposons, replication is critical for their survival. The complementary two-stranded structure of DNA is the key to understanding its duplication during cell division. The double-stranded helix unwinds, and the hydrogen bonds holding the bases together melt apart to form two single strands. This Y-shaped region of DNA is the **replication fork** (Fig. 4.1). Replication starts at a specific site called an **origin of replication (ori)** on the chromosome. The origin is called *oriC* on the *E. coli* chromosome and covers about 245 base pairs of DNA. The origin has mostly AT base pairs, which require less energy to break than GC base pairs.

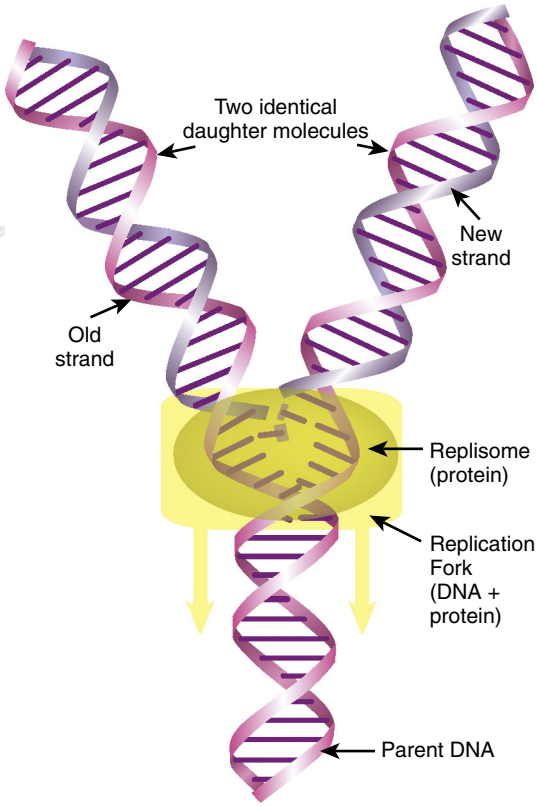

FIGURE 4.1
Replication
Replication enzymes open the double-helix around the origin to make it single-stranded. DNA polymerase adds complementary nucleotides.

In replication, DNA polymerase synthesizes the leading strand as one continuous piece and the lagging strand as Okazaki fragments. Each copy has one strand from the original helix and one new strand.

Once the replication fork is established, a large assembly of enzymes and factors called a **replisome** assembles to synthesize the complementary strands of DNA (Fig. 4.2). The replisome starts synthesizing the complementary strand on one side of the fork by adding complementary bases in a 5′ to 3′ direction. The **leading strand** is synthesized continuously because there is always a free 3′-OH group. Because DNA polymerase synthesizes only in a 5′ to 3′ direction, the other strand, called the **lagging strand**, is synthesized as small fragments called **Okazaki fragments**. As DNA polymerase makes this strand, the clamp loader must continually release and reattach at a new location. This results in the single-stranded region bubbling out from the replisome. The lagging strand fragments are ligated together by an enzyme called **DNA ligase**. Ligase links the 3′-OH and the 5′-PO_4 of neighboring nucleotides, forming a phosphodiester bond. The final step is to add methyl (-CH_3) groups along the new strand (Fig. 4.3). The original double-stranded helix is now two identical double-stranded helices, each containing one strand from the original molecule and one new strand. This is why the process is called **semiconservative replication**.

Uncoiling the DNA

Because DNA is condensed into supercoils in order to fit inside the cell, several different enzymes are needed to open and relax the DNA before replication can start (Fig. 4.2). **DNA helicase** and **DNA gyrase** attach near the replication fork and untwist the strands

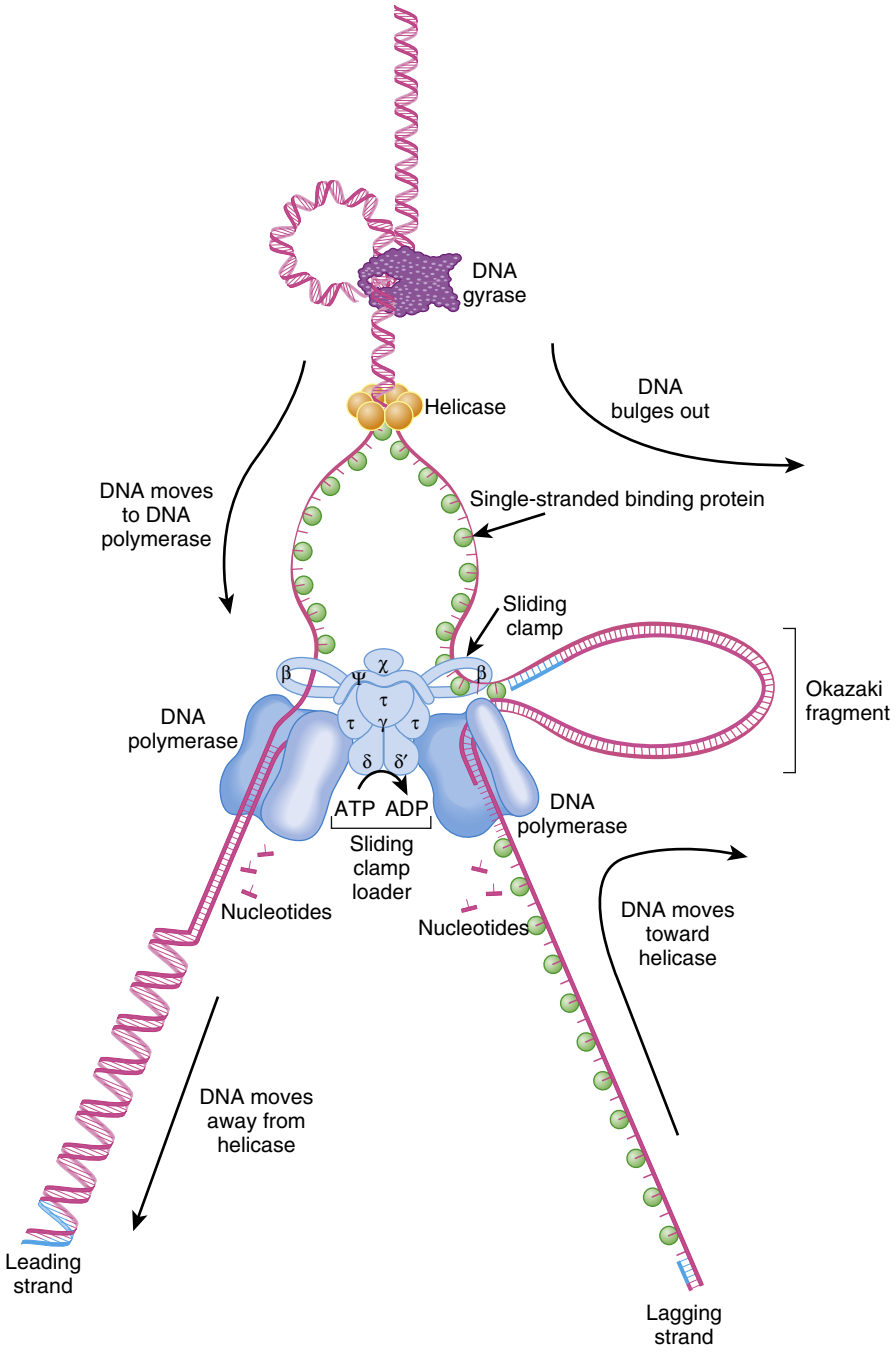

FIGURE 4.2 DNA Polymerase III Replication Assembly During replication, the sliding clamp loader complex makes contacts with single-stranded binding protein and the sliding clamps. This complex stabilizes the two single-stranded DNA strands and provides a stable binding site for two DNA polymerase III molecules. The unwound single-stranded DNA templates move toward the clamp loader complex. On the leading strand (left), the strand is unwinding in a 3′ to 5′ direction, so DNA polymerase can add complementary nucleotides in the 5′ to 3′ direction. On the lagging strand (right), the template strand is antiparallel, and therefore, the strand is unwinding in a 5′ to 3′ direction. Since DNA polymerase III must synthesize the new strand in a 5′ to 3′ direction also, the template strand must move toward the helicase. This causes the lagging strand to bubble out from the complex. Once DNA polymerase III reaches the end of the previous Okazaki fragment, the replicated DNA is released by the clamp loader and a new section of single-stranded DNA is reloaded.

of DNA. DNA gyrase removes the supercoiling, and helicase unwinds the double helix by dissolving the hydrogen bonds between the paired bases. The two strands are kept apart by **single-stranded binding protein**, which coats the single-stranded regions. This prevents the two strands from reannealing so that other enzymes can gain access to the origin and begin replication.

As DNA polymerase travels along the DNA, more positive supercoils are added ahead of the replication fork. Because the bacterial chromosome is negatively supercoiled, initially the new positive supercoils relax the DNA. After about 5% of the genome has been replicated, though, the positive supercoils begin to accumulate and need to be removed. DNA gyrase cancels the positive supercoils by adding negative supercoils. When circular chromosomes are replicated, the two daughter copies may become **catenated**, or connected like two links

of a chain (Fig. 4.4). Topoisomerase IV releases catenated daughter strands by introducing double-stranded nicks into one chromosome. The second copy can then pass through the first, giving two separated molecules.

DNA helicase, DNA gyrase, and topoisomerase IV untwist and untangle the supercoiled DNA during replication.

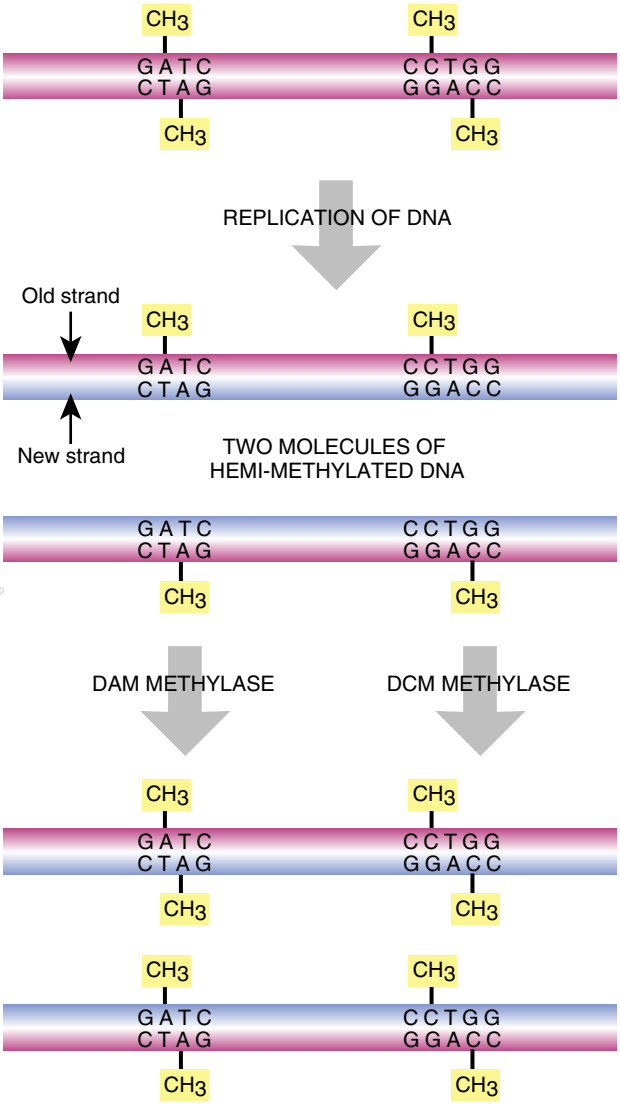

FIGURE 4.3 Hemimethylated DNA: Old Strands versus New
When DNA is replicated, the old strand is methylated, but there is a delay in methylating the new strand, and thus, the DNA double helix is hemimethylated. Dam methylase and dcm methylase add the methyl groups onto the newly synthesized DNA.

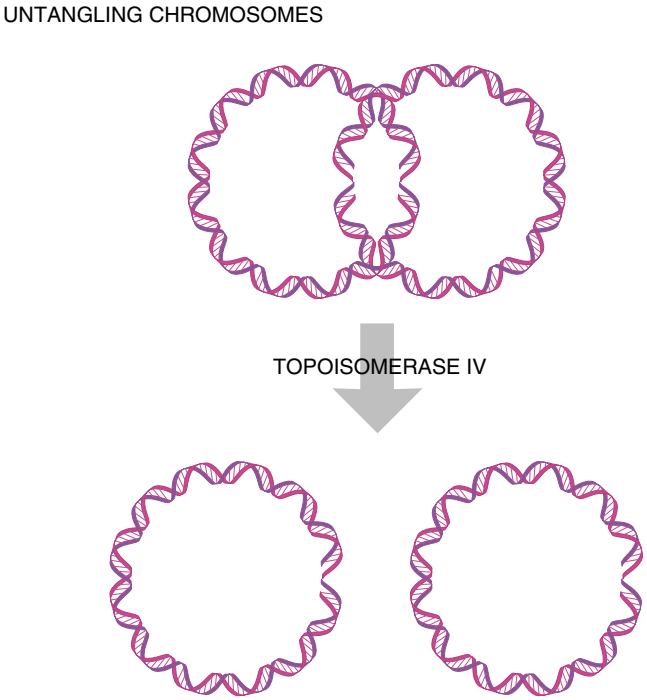

FIGURE 4.4 Untangling Circular Chromosomes
Sometimes after the replication of circular genomes is complete, the two rings are catenated, or linked together like links in a chain. Topoisomerase IV untangles the two chromosomes so they can partition into the daughter cells.

Priming DNA Synthesis

DNA polymerase cannot initiate new strands of nucleic acid synthesis because it can only add a nucleotide onto a pre-existing 3′-OH. Therefore, an 11 to 12 base-pair length of RNA (an RNA primer) is made at the beginning of each new strand of DNA. Since the leading strand is synthesized as a single piece, there is only one RNA primer at the origin. On the lagging strand, each Okazaki fragment begins with a single RNA primer. DNA polymerase then makes DNA starting from each RNA primer. At the origin, a protein called **PriA** displaces the SSB proteins so a special RNA polymerase, called **primase** (DnaG), can enter and synthesize short RNA primers using ribonucleotides. Two molecules of DNA polymerase III bind to the primers on the leading and lagging strands and synthesize new DNA from the 3′ hydroxyls (Fig. 4.5).

Primase, a special RNA polymerase, works with PriA to displace the SSB proteins and synthesize a short RNA primer at the origin. DNA polymerase then starts synthesis of the new DNA strand using the 3′-OH of the RNA primer. This synthesis occurs at multiple locations on the lagging strand.

Structure and Function of DNA Polymerase

DNA polymerase III (PolIII) is the major form of DNA polymerase used to replicate bacterial chromosomes and consists of multiple protein subunits (see Fig. 4.2). The **sliding clamp** is a donut-shaped protein consisting of a dimer of DnaN proteins, also called the β-subunits. Two clamps encircle the two single strands of DNA at the replication fork. A cluster of accessory proteins, called the **clamp loader complex**, loads the clamps onto DNA strands. The two sliding clamps bind two **core enzymes**, one for each strand of DNA. The core enzyme consists of three subunits: DnaE (α subunit), which links the nucleotides together; DnaQ (ε subunit), which proofreads the new strand; and HolE (θ subunit), which stabilizes the two other subunits (not shown in Fig. 4.2). As the α subunit adds new nucleotides, the ε subunit recognizes any distortions and removes any mismatched bases. A correct nucleotide is then added. Bacterial DNA polymerase III can add up to 1000 bases per second, which is an extraordinarily fast rate of enzyme activity.

The multiple subunits of DNA polymerase III work together to synthesize a new strand of DNA. The core has two essential subunits: the α subunit links the nucleotides, and the ε subunit ensures that they are accurate.

Synthesizing the Lagging Strand

After the new lagging strand of DNA has been made, it has many segments of RNA derived from multiple RNA primers, as well as multiple breaks, or **nicks**, along the backbone that need to be sealed (Fig. 4.6). DNA polymerase I removes the RNA primers from the lagging strand. DNA polymerase I has exonuclease activity that removes the RNA bases, and then its polymerase activity fills in the regions with DNA bases. The RNA bases may also be removed by RNaseH, an enzyme that specifically identifies RNA:DNA heteroduplexes and removes the RNA bases. Finally, the DNA fragments of the lagging strand are linked together with a ligation reaction by DNA ligase. DNA polymerase I and DNA ligase are both very important enzymes in molecular biology and are used extensively in biotechnology.

Because the lagging strand is synthesized in small pieces, either DNA polymerase I or RNaseH excise the RNA bases and replace them with DNA. DNA ligase closes the nicks in the sugar/phosphate backbone of the new DNA strand.

Repairing Mistakes after Replication

After replication is complete, the **mismatch repair system** corrects mistakes made by DNA polymerase. If the wrong base is inserted and DNA polymerase does not correct the error itself, there will be a small bulge in the helix at that location. Identifying which of the two bases is correct is critical. The cell assumes that the base on the new strand is wrong and the original parental base is correct. The mismatch repair system of *E. coli* (MutSHL) deciphers which strand is the original by monitoring methylation. Immediately after replication, the DNA is **hemimethylated**; that is, the old strand still has methyl groups attached to various bases, but the new strand has not been methylated yet (see Fig. 4.3). Two different *E. coli* enzymes add methyl groups: **DNA adenine methylase (Dam)** adds a methyl group to the adenine in GATC, and **DNA cytosine methylase (Dcm)** adds a methyl group to the cytosine in CCAGG or CCTGG. These enzymes methylate the new strand after replication, but they are slow. This allows mismatch repair to find and fix any mistakes first.

Three genes of *E. coli* are responsible for mismatch repair: *mutS*, *mutL*, and *mutH* (Fig. 4.7). MutS protein recognizes the bulge or distortion in the sequence. MutH finds the nearest GATC site and nicks the nonmethylated strand—that is, the newly made strand. MutL holds the MutS plus mismatch and the MutH plus GATC site together (these may be far apart on the DNA helix). Finally, the DNA on the new strand is degraded and replaced with the correct sequence by DNA polymerase III.

In *E. coli,* mismatch repair proteins (MutSHL) identify a mistake in replication, excise the new nucleotides around the mistake, and recruit DNA polymerase III to the single-stranded region to make the new strand without a mistake.

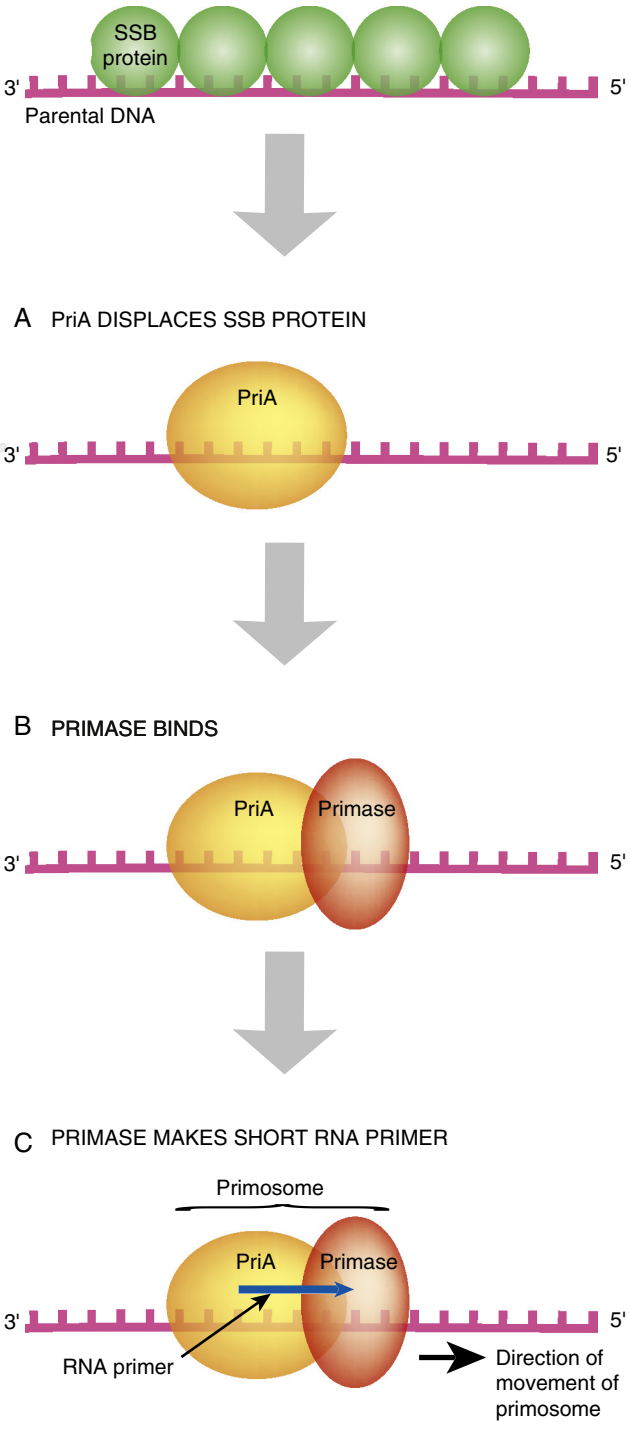

FIGURE 4.5 Strand Initiation Requires an RNA Primer
DNA polymerase cannot synthesize new DNA without a pre-existing 3′-OH. Thus, DNA replication requires an RNA primer to initiate strand formation. (A) First, the PriA protein displaces the SSB proteins. (B) Second, primase associates with the PriA protein. (C) Last, the primase makes the short RNA primer.

COMPARING REPLICATION IN GENE CREATIONS, PROKARYOTES, AND EUKARYOTES

Although the basic mechanism for replication is the same for most organisms, the timing, direction, and sites for initiation and termination are variable. The major differences in replication occur mainly because of the special challenges posed by circular and linear genomes. Normal DNA replication occurs bidirectionally in prokaryotes and eukaryotes, whether the genome is linear or circular. Two replication forks travel in opposite directions, unwinding the DNA helix as they go. In bacteria such as *E. coli,* there is only one origin, *oriC,* and replication occurs in both directions around the circular chromosome until it meets at the other side, the terminus, *terC.* Halfway through this process, the chromosome looks like the Greek letter θ; therefore, this process is often called **theta replication** (Fig. 4.8). The single circular chromosome then becomes two. Theta replication is also used by many plasmids, such as the F plasmid of *E. coli,* when growing and dividing asexually (as opposed to transferring its genome to another cell via conjugation).

Some plasmids and many viruses replicate their genomes by a process called **rolling circle replication** (Fig. 4.9). At the origin of replication, one strand of the DNA is nicked and unrolled. The intact strand thus rolls relative to its partner (hence "rolling circle"). DNA is synthesized from the origin using the circular strand as a template. As DNA polymerase circles the template strand, the new strand of DNA is base-paired to the circular template. Meanwhile the other parental strand is dangling free. This dangling strand is removed, ligated to form another circle, and finally a second strand is synthesized. This process results in two rings of plasmid or viral DNA, each with one strand from the original molecule and one newly synthesized strand.

Some viral genomes use rolling circle replication but continue to make more and more copies of the original circular template. They continue rolling around the circle, synthesizing more and more copies that are all dangling as a long single strand. The long strand of new DNA may be made double-stranded or left single-stranded (depending on the

type of virus). Finally, the dangling strand is chopped into genome-sized units and packaged into viral particles. Some viruses circularize these copies before packaging; others simply leave the genomes linear.

Long linear DNA molecules such as human chromosomes pose several problems for replication. The ends pose a particularly difficult problem because the RNA primer is synthesized at the very end of the lagging strand. When the RNA primer is removed by an exonuclease, there is no upstream 3′-OH for addition of new nucleotides to fill the gap. (In eukaryotes, there is no equivalent to the dual-function DNA polymerase I. A separate exonuclease, MF1, removes the RNA primers, and DNA polymerase δ fills in the gaps for the lagging strand.) Over successive rounds of replication, the ends of linear chromosomes get shorter and shorter. Special structures called **telomeres** are found at the tips of each linear chromosome and prevent chromosome shortening from affecting important genes. Telomeres have multiple tandem repeats of a short sequence (TTAGGG in humans). The enzyme **telomerase** can regenerate the telomere by using an RNA template to synthesize new repeats. This happens only in some cells; in others, the telomeres shorten every time the cell replicates its DNA. One theory regards telomere shortening as a molecular clock, aging the cell, and eventually triggering suicide (see Chapter 20).

The length of linear chromosomes also poses a problem. The time it takes to synthesize an entire human chromosome would be too long if replication began at only one origin. To solve this issue, multiple origins exist, each initiating new strands in both directions. These are elongated until they meet the new strands from the other direction.

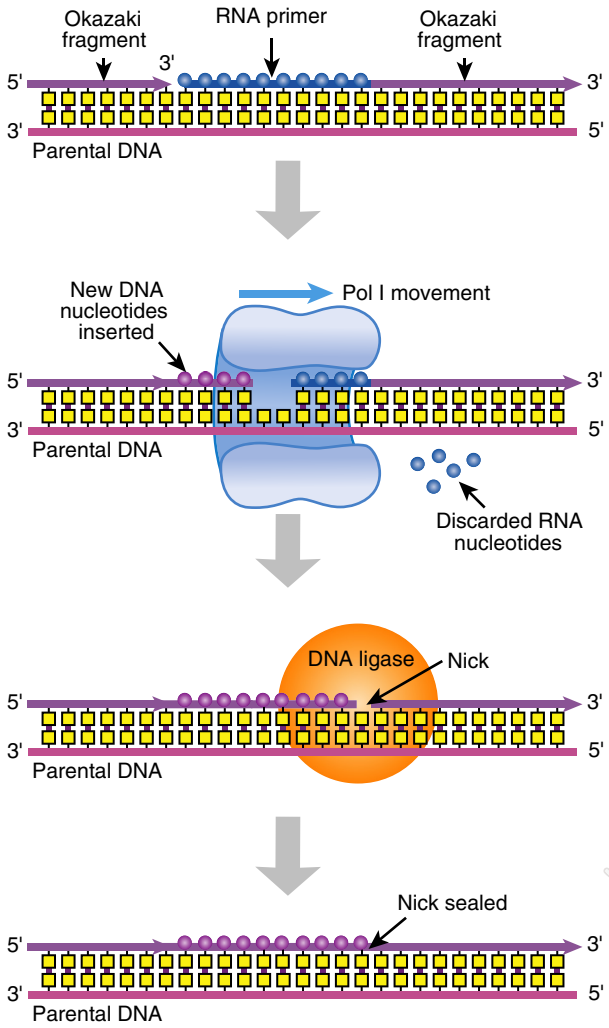

FIGURE 4.6
Joining the Okazaki Fragments
When first made, the lagging strand is composed of alternating Okazaki fragments and RNA primers. Next, DNA polymerase I binds to the primer region, and as it moves forward, it degrades the RNA and replaces it with DNA. Finally, DNA ligase seals the nick in the phosphate backbone.

The cellular structure of eukaryotes also poses some problems for replication. (In bacteria, the chromosome simply replicates, the two copies move to each end of the cell, and a new wall forms in the middle. There are no nuclear membranes or organelles to divide; there is just one chromosome plus, perhaps, some plasmids.) In eukaryotes, the cell has a specific **cell cycle**, with four different phases, and replication occurs at specific points (Fig. 4.10). During G_1, the cell rests for a period before DNA synthesis begins. This period varies, lasting about 25 minutes for yeast. The next phase is S, or synthesis, during which the entire genome is replicated. This is usually the longest phase, lasting about 40 minutes in yeast. The third phase, G_2, is another resting phase before the cell undergoes mitosis, in the M phase. During mitosis, cells divide their walls and membranes into two separate cells, partitioning the new chromosomes and other cellular components into each half. The signal that triggers cell division depends on many factors, including environment, size, and age.

Eukaryotic mitosis is a dynamic process with much movement and repositioning of cellular components. First, the nuclear membrane must be dissolved before the chromosomes can separate. After replication, the two sets of chromosomes are partitioned to separate sides of the cell. The chromosomes attach to long fibers making up the **spindle** via special sequences called **centromeres**. They slide along the spindle fibers until they reach separate ends of the cell. A new cell membrane separating the two halves is then synthesized. Other cellular components including mitochondria, endoplasmic reticulum, lysosomes, and so forth are split between the two daughter cells. Finally, a new nuclear membrane must be assembled around the chromosomes of each new daughter cell. The dynamics of this process are still being investigated, and new proteins and molecules are still being discovered that mediate different parts of mitosis in eukaryotic cells.

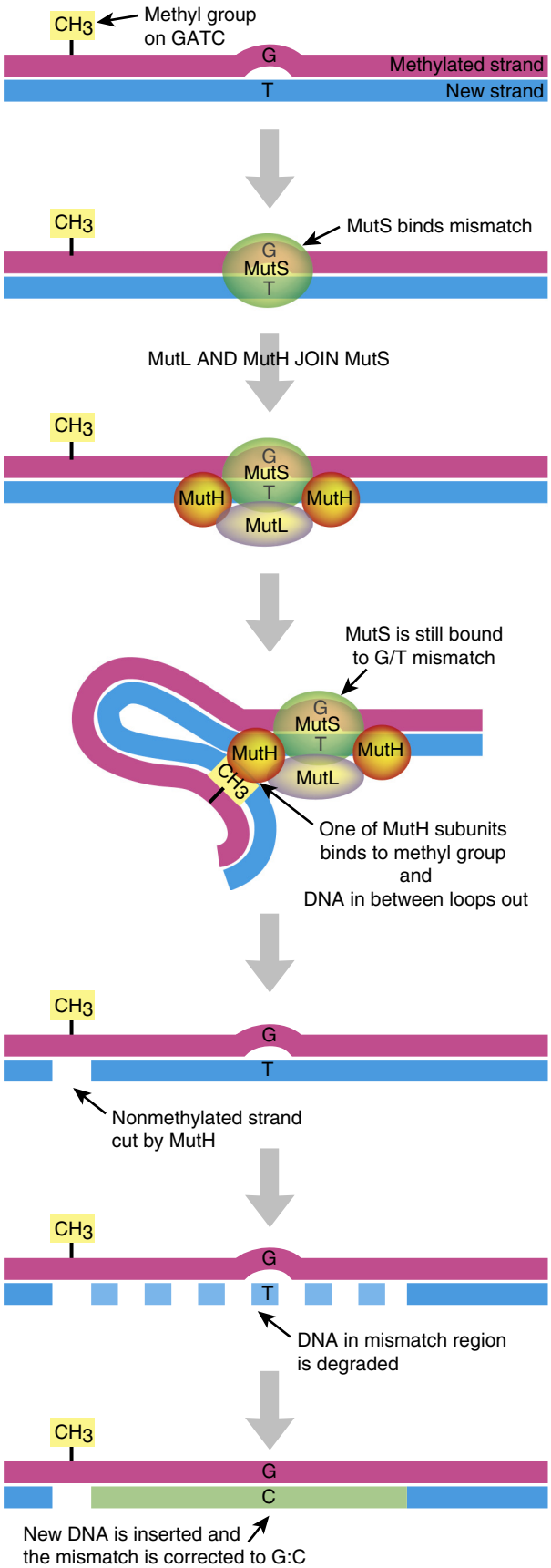

FIGURE 4.7 Mismatch Repair Occurs after Replication

MutS recognizes a mismatch shortly after DNA replication. MutS recruits MutL and two MutH proteins to the mismatch. MutH locates the nearest GATC of the new strand by identifying the methyl group attached to the "parent" strand. MutH cleaves the nonmethylated strand and the DNA between the cut and the mismatch is degraded. The region is replaced, and the mismatch is corrected.

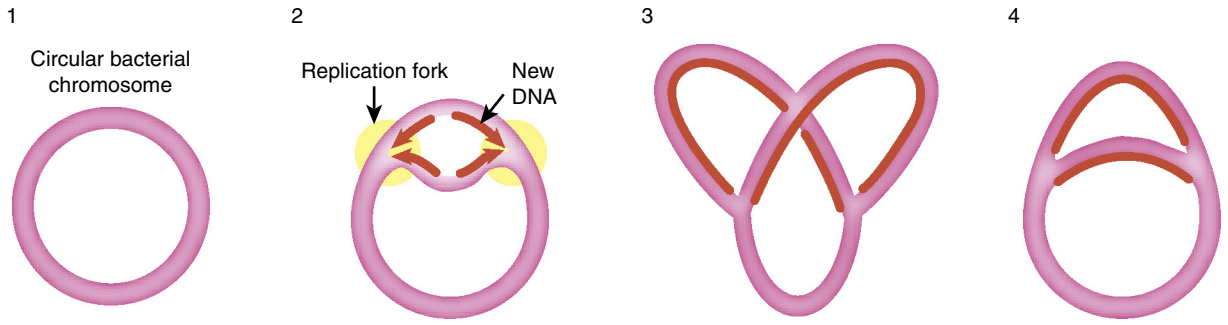

FIGURE 4.8 Theta Replication
In circular genomes or plasmids, replication enzymes recognize the origin of replication, unwind the DNA, and start synthesis of two new strands of DNA, one in each direction. The net result is a replication bubble that makes the chromosome or plasmid look similar to the Greek letter theta (θ). The two replication forks keep moving around the circle until they meet on the opposite side.

Bacteria and viruses use either theta replication or rolling circle replication to create new genomes.

Eukaryotic cells have chromosomes with multiple origins of replication. Telomeres protect the ends of the chromosomes because each round of replication shortens the DNA. Replication in eukaryotes occurs only at a specific point during the cell cycle.

IN VITRO DNA SYNTHESIS

Making DNA in the laboratory relies on the same basic principles outlined for replication (Fig. 4.11). DNA replication needs the following "reagents": enzymes to melt the two template DNA strands apart, an RNA primer with a 3′-hydroxyl for DNA polymerase to synthesize a new DNA strand, a pool of nucleotide precursors, plus DNA polymerase to catalyze the addition of new nucleotides.

To perform DNA replication in the laboratory, the researcher makes a few modifications. First, the enzymes that open and unwind the template DNA are not used. Instead, double-stranded DNA is converted to single-stranded DNA using heat or a strong base to disrupt the hydrogen bonds that hold the two strands together. Alternatively, template DNA can be made by using a virus that packages its DNA in single-stranded form. For example, M13 is a bacteriophage that infects *E. coli*, amplifies its genome using rolling circle replication, and packages the single-stranded DNA in viral particles that are released without lysing open the *E. coli* cell. If template DNA is cloned into the M13 genome, then the template will also be manufactured as in a single-stranded form. This DNA can be isolated directly from the viral particles.

During *in vitro* synthesis of DNA, an RNA primer is not used because RNA is very unstable and degrades easily. Instead, a short single-stranded oligonucleotide of DNA is used as a

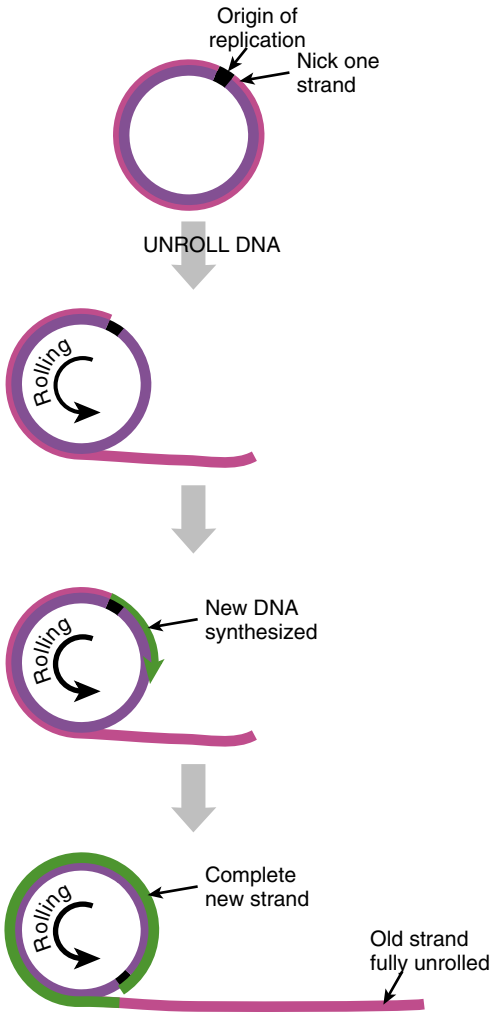

FIGURE 4.9 Rolling Circle Replication
During rolling circle replication, one strand of the plasmid or viral DNA is nicked, and the broken strand (pink) separates from the circular strand (purple). The gap left by the separation is filled in with new DNA starting at the origin of replication (green strand). The newly synthesized DNA keeps displacing the linear strand until the circular strand is completely replicated. The linear single-stranded piece is fully "unrolled" in the process.

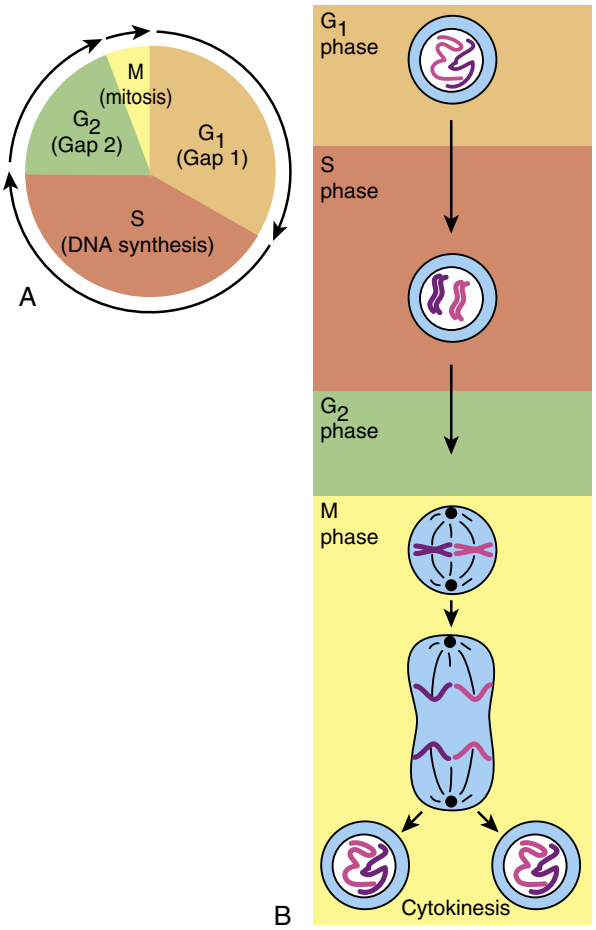

FIGURE 4.10
Eukaryotic Cell Cycle
DNA replication occurs during the S phase of the cell cycle but the chromosomes are actually separated later, during mitosis, or the M phase. The S and M phases are separated by G_1 and G_2.

primer. (As long as the primer has a free 3′-hydroxyl, DNA polymerase will add nucleotides onto the end.) The primers are synthesized chemically (see later discussion) and mixed with the single-stranded template DNA. The oligonucleotide primer has a sequence complementary to a short region on the DNA template. Therefore, at least some sequence information must be known about the template. If the sequence of the template DNA is unknown, it may be cloned into a vector, and the primer is then designed to match sequences of the vector (such as the polylinker region) that are close to the inserted DNA.

Finally, purified DNA polymerase plus a pool of nucleotides (dATP, dCTP, dGTP, and dTTP) is added to the primer and template. The primer anneals to its complementary sequence, and DNA polymerase elongates the primer, creating a new strand of DNA complementary to the template DNA.

In vitro replication requires a single-stranded piece of template DNA, a primer, nucleotide precursors, and DNA polymerase.

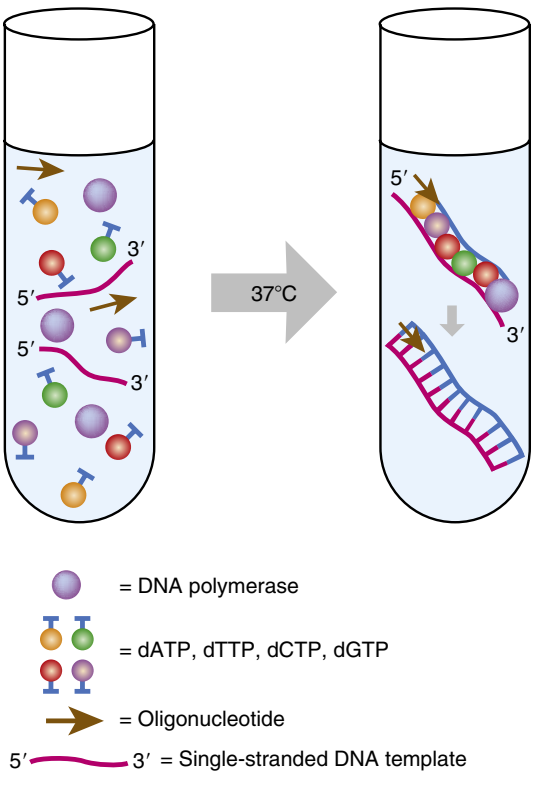

FIGURE 4.11 ***In Vitro*** **DNA Synthesis**
DNA synthesis in the laboratory uses single-stranded template DNA, plus DNA polymerase, an oligonucleotide primer, and nucleotide precursors. After all the components are incubated at the appropriate temperature, double-stranded DNA is made.

CHEMICAL SYNTHESIS OF DNA

Making DNA chemically rather than biologically was one of the first new technologies to be applied by the biotechnology industry. The ability to make short synthetic stretches of DNA is crucial to using DNA replication in laboratory techniques. DNA polymerase cannot synthesize DNA without a free 3′-OH end to elongate. Therefore, to use DNA polymerase *in vitro*, the researcher must supply a short primer. Such primers are used to sequence DNA (see later discussion), to amplify DNA with PCR (see later discussion), and even to find genes in library screening (see Chapter 3). So a short review of how primers are synthesized is included here.

Research into chemical synthesis of DNA began shortly after Watson and Crick published their research on the crystal structure of DNA. H. Gobind Khorana at the University of Chicago was an early pioneer in the study of **oligonucleotide** synthesis (see Box 4.1). Technically, oligonucleotides are

Box 4.1 Khorana, Nirenberg, and Holley

Har Gobind Khorana, Marshall W. Nirenberg, and Robert W. Holley are pioneers in the field of molecular biology. The three scientists received the Nobel Prize in Physiology or Medicine in 1968 for their combined efforts in identifying which triplet codons coded for which amino acid. Khorana originally began chemical synthesis of DNA in order to help elucidate the role of different enzymes. He wanted to understand the mode of action for nucleases and phosphodiesterases, but without being able to chemically synthesize a defined nucleic acid, the work on enzymes would be very difficult.

Khorana's lab determined ways to synthesize dinucleotide, trinucleotide, and tetranucleotide sequences using chemical synthesis. Rather than using single nucleotide additions, his lab focused on synthesizing nucleotides in blocks. His ability to chemically synthesize blocks of DNA was the backbone experiment, but many other discoveries were instrumental in determining the amino acid codes.

Matthaei and Nirenberg (1961) experimentally determined that polyuridylate [poly(U)] mixed with a bacterial cell-free amino acid incorporating system created polyphenylalanine. This experiment determined that the codon UUU encoded for the amino acid phenylalanine. During this time, Robert Holley was working on tRNA. He specifically identified the structure of the tRNA for alanine by purifying tRNA-alanine from yeast, fragmenting the tRNA into pieces with nucleases, and logically piecing together the size of the fragments and the sites at which the enzymes were recognized. Other important discoveries included the purification of DNA polymerase and RNA polymerase.

These experiments were woven into an elegant method of determining which triplet nucleotide sequence encoded which amino acid. First, Khorana's groups began synthesizing dinucleotide, trinucleotide, and tetranucleotide double-stranded DNA fragments. For example, one of these fragments had the following structure:

5′ TCTCTC 3′
3′ AGAGAG 5′

Arthur Kornberg had previously won the Nobel Prize for his discovery and purification of DNA polymerase I. Khorana's group mixed their short synthesized DNA with pure DNA polymerase to create long polydeoxynucleotides with a known sequence. Next, the DNA pieces were mixed with RNA polymerase to create long polyribonucleotides of known sequence. These were mixed with the cell-free system devised by Matthaei and Nirenberg, which made polypeptides. The preceding dinucleotide example resulted in a polypeptide of repeating serine and leucine. The experiment demonstrated that TCT or CTC encoded serine or leucine, respectively. There was no way to determine definitively which codon matched which amino acid, so more experiments were needed.

The final important contributions to make the final assignments were using purified tRNAs labeled with ^{14}C. Nirenberg and Leder (1964) mixed Khorana's synthetic trinucleotides and mixed them with the labeled tRNAs and ribosomes. (Note: The isolation of pure tRNA was not possible without Robert W. Holley's work.) They looked for binding of the labeled tRNA to the trinucleotide sequence. These experiments provided clear answers to many of the trinucleotide sequences, but many times the results were not very clear. It was the combination of these experiments with Khorana's work that determined the direct genetic code.

any piece of DNA less than 20 nucleotides in length, but today, *oligonucleotide* denotes a short piece of DNA that is chemically synthesized. In 1970, Khorana's lab synthesized an active tRNA molecule of 72 nucleotides (Agarwal et al., 1970). The chemistry he used was inefficient and cumbersome, but some of his ideas are still used in current oligonucleotide synthesis. Today chemical synthesis is done with an automated **DNA synthesizer** that creates DNA by sequentially adding one nucleotide after another in the correct sequence order.

Unlike *in vivo* DNA synthesis, artificial synthesis is done in the 3′ to 5′ direction. The first step is attaching the first nucleotide to a porous material made of **controlled pore glass (CPG)**. The first nucleotide is not attached directly but is linked to the surface via a spacer molecule that binds to the 3′-OH of the nucleotide (Fig. 4.12). The column pores allow reagents to be washed through and removed easily. (Using CPG is one improvement over Khorana's technology. He used polymer beads to couple the reaction but found that the polymer swelled as the reagents passed through the column, which inhibited synthesis. CPG is superior because it does not swell.)

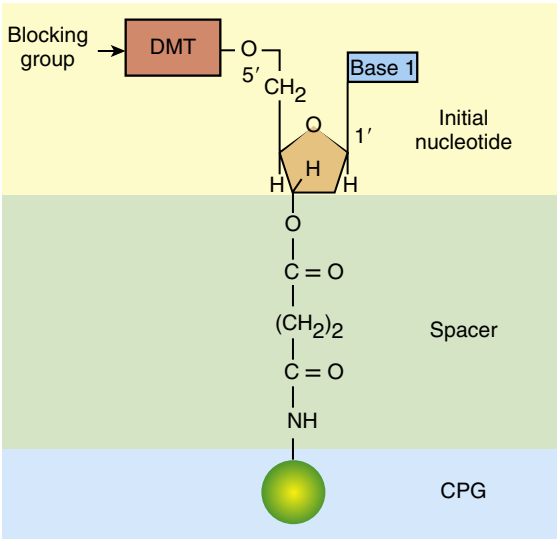

FIGURE 4.12
Addition of a Spacer Molecule and First Base to the CPG
The first nucleotide is linked to a glass bead via a spacer molecule attached to its 3′-OH group. The structure of the spacer varies, but it is important to keep the synthesis away from the glass surface and to allow efficient removal of the completed oligonucleotide.

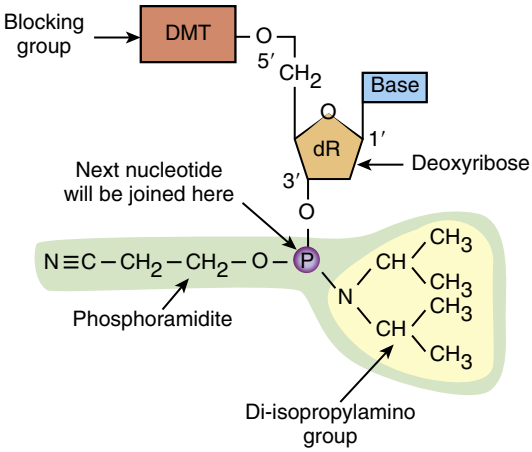

FIGURE 4.13 Nucleoside Phosphoramidites Are Used for Chemical Synthesis of DNA
Nucleotides are modified to ensure that the correct group reacts with the growing oligonucleotide. Each nucleotide has a DMT group blocking its 5′-OH. The 3′-OH is activated by a phosphoramidite group, which is originally also protected by di-isopropylamine.

When the spacer is linked to the nucleotide 3′-OH, a chemical blocking group is attached to the 5′-OH. Thus, the 3′-OH is the only available reactive group. Khorana's early synthesis was revolutionary in this respect because he chose the **dimethyloxytrityl (DMT) group**, which is still used as a blocking group in today's synthesizers. DMT has a strong orange color and is easily removed from the 5′-OH so that another nucleotide can be linked to the first. In practice, the CPG–spacer–first nucleotide is washed, and then the DMT group is removed by mild acid such as trichloroacetic acid (TCA). The 5′-OH is now ready to accept the next nucleotide. The efficiency of removing DMT is critical. If DMT is not removed completely, many of the potential oligonucleotides will fail to elongate. The orange color reveals the efficiency of removal and is easily measured optically.

Each nucleotide is added as a **phosphoramidite**, which is a nucleotide that has a blocking group protecting a 3′-phosphite group (Fig. 4.13). (One problem with early oligonucleotide synthesis technology was branching. Rather than the incoming nucleotide adding to the 5′ end, it sometimes attached to the phosphate linking two nucleotides.) To prevent branching, every added nucleotide has a di-isopropylamine group attached to the 3′ phosphite group, which also stabilizes the nucleotides allowing long-term storage. Before another nucleoside is added, the 3′ phosphite group is activated by tetrazole. The next nucleotide is then added, and it reacts with the phosphite to form a dinucleotide (Fig. 4.14).

If the terminal nucleotide of a growing chain fails to react with an incoming nucleotide, the chain must be capped off to prevent generation of an incorrect sequence by later reactions. The 5′-OH of all unreacted nucleotides is acetylated with acetic anhydride. This terminates the chain so that no other nucleoside phosphoramidites can be added. (Fig. 4.15)

At this stage of the synthesis, the column has CPG–spacer–first nucleoside–phosphite–second nucleoside–DMT. Phosphites are used because they react much faster, but they are unstable. Adding iodine oxidizes the phosphite triester into the normal phosphodiester, which is more stable under acidic conditions (Fig. 4.16).

The column can now be prepared to add the third nucleotide. The DMT is removed with TCA, and the third phosphoramidite nucleotide is added. The chains are capped so that any dinucleotides that failed to react with the third phosphoramidite are prevented from adding any more nucleosides. Finally, the phosphite triester is oxidized to phosphodiester. This process continually repeats until all the desired nucleotides are added and the final oligonucleotide has the correct sequence (Fig. 4.17).

After the final phosphoramidite nucleoside is added, the oligonucleotide still has DMT protecting the 5′-OH, cyanoethyl groups attached to the phosphates, and amino-protecting groups on the bases. (Amino groups would react with the reagents during synthesis; therefore, chemical groups are added to protect the bases before they are added to the column.) All three types of protective groups must be removed. The organic salts of the protecting groups are then removed by desalting, and the final oligonucleotide is cleaved from the CPG surface. Finally, the 5′-OH must be phosphorylated to make the oligonucleotides

biologically active. A kinase from bacteriophage T4 is used to transfer a phosphate group from ATP to the 5′ end of the oligonucleotides. The newly synthesized oligonucleotide is now ready for use.

Chemical synthesis of DNA occurs by successively adding phosphoramidite nucleotides to the previous base attached to controlled pore glass (CPG) columns. Synthesis occurs in a 3′ to 5′ direction by removing the 5′-blocking group from the existing nucleotide and adding the new activated phosphoramidite nucleotide. After coupling, the unreacted nucleotides are capped, and the phosphate triester is oxidized to a phosphodiester group. Synthesis ends by removal of all blocking groups from the bases, removing the cyanoethyl groups, and cleavage from the CPG.

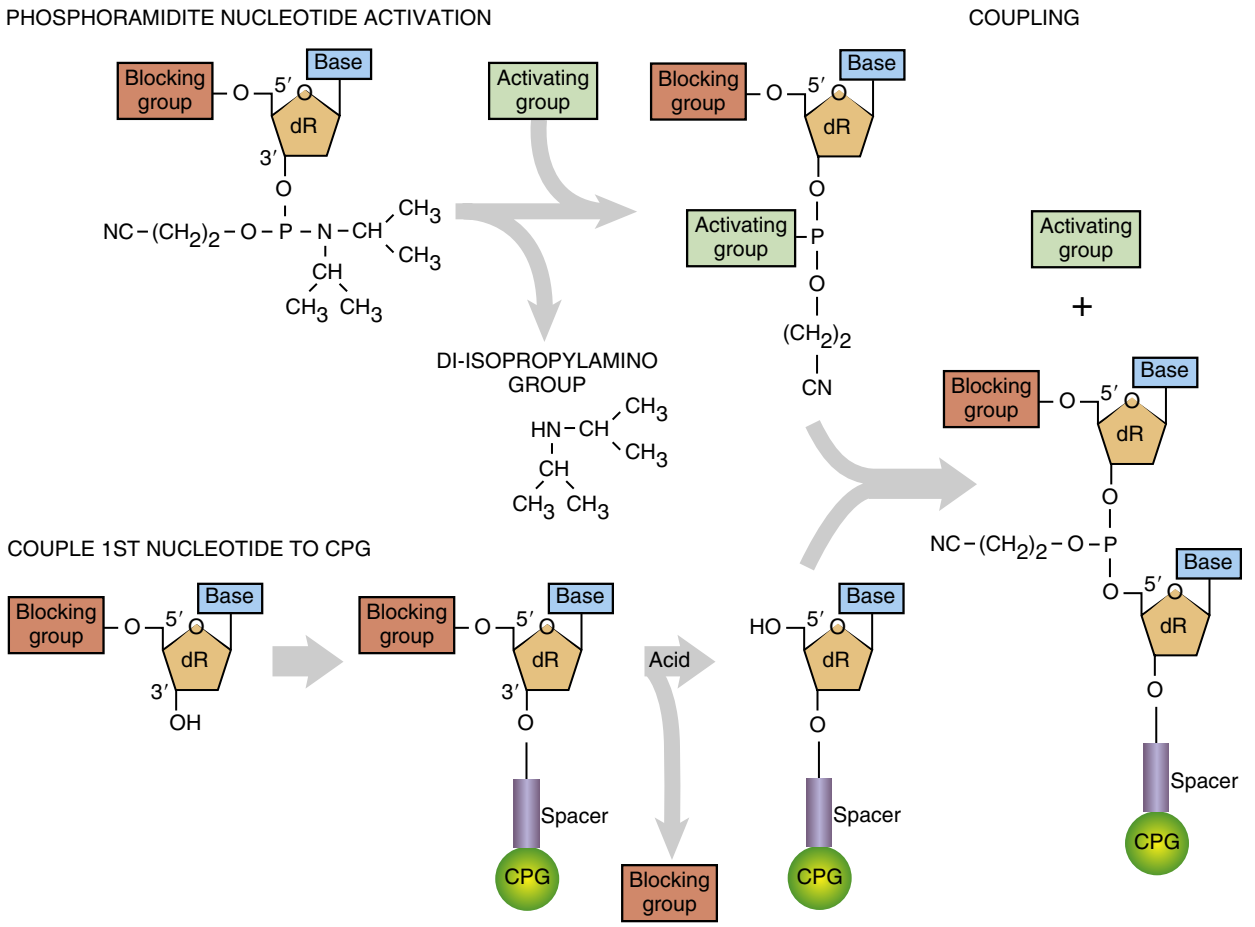

FIGURE 4.14 Adding the Second Nucleotide
During chemical synthesis of DNA, nucleotides are added in a 3′ to 5′ direction (the opposite of *in vivo* DNA synthesis). Therefore, the 3′-OH of an incoming nucleotide must be activated, but the 5′-OH must be blocked (see top nucleotide). For nucleotides already attached to the bead, the opposite must be done. Here, the blocking group on the 5′-OH of nucleotide 1 is removed by treatment with a mild acid. When the second nucleotide is added, it reacts to form a dinucleotide.

CHEMICAL SYNTHESIS OF COMPLETE GENES

As mentioned earlier, at each nucleoside addition in chemical synthesis, a proportion of oligonucleotides do not react with the next base, and these are capped with an acetyl group. The efficiency for nucleoside addition is critical, because if each step has low efficiency, the number of full-length oligonucleotides will decrease exponentially. For example, if the efficiency is 50% at each round, only half of the oligonucleotides add the second base, one-fourth would add the

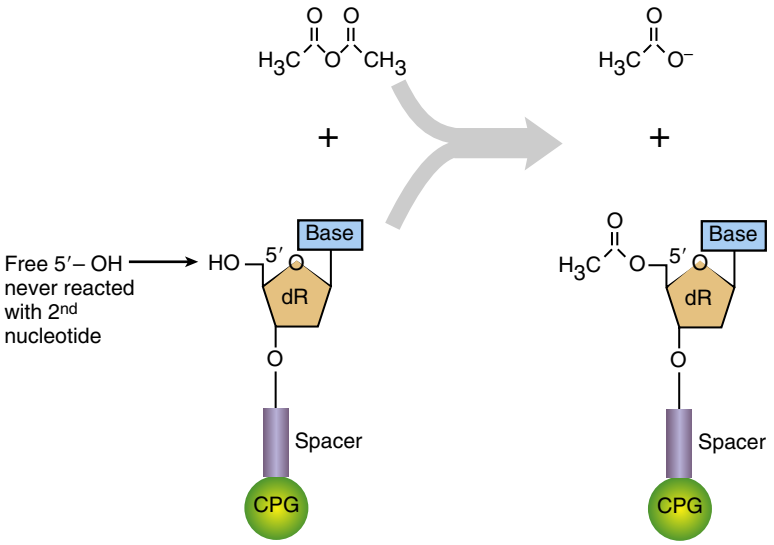

FIGURE 4.15
Capping of Unreacted Nucleotides
If any of the first nucleotides are not coupled to a second nucleotide, these could react with a subsequent nucleotide creating an internal deletion of the oligonucleotide. To prevent this error, the unreacted 5′-OH is capped with an acetyl group from acetic anhydride.

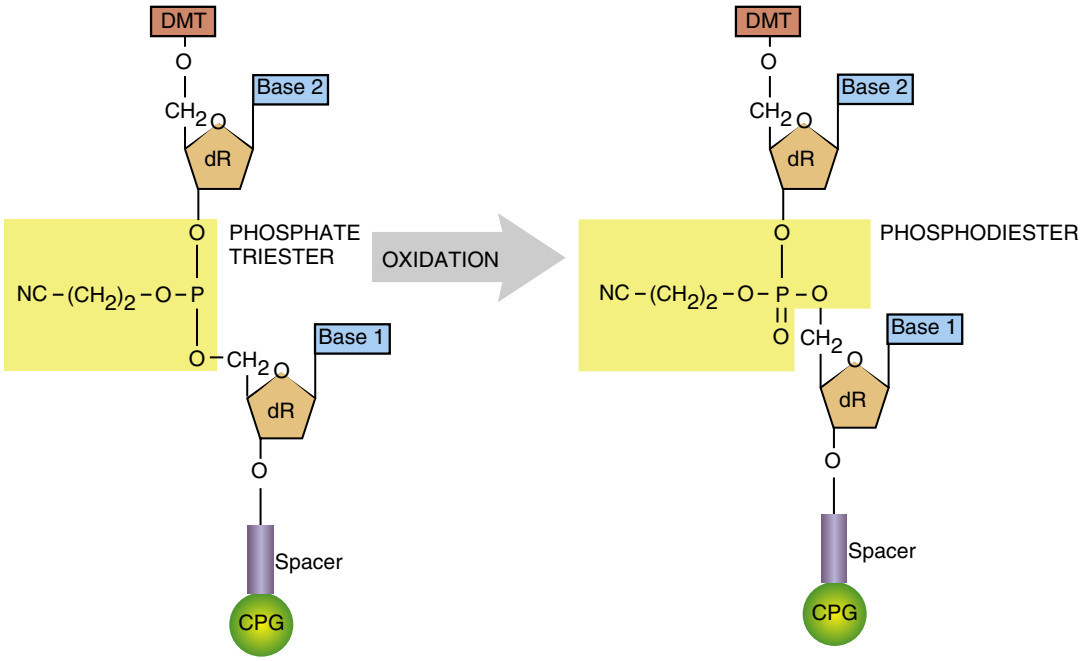

FIGURE 4.16
Oxidation Converts Phosphite Triester into a Phosphodiester
The phosphite triester is oxidized to a phosphodiester by adding iodine. This stabilizes the dinucleotide for further additions.

third base, one-eighth would get four bases; one-sixteenth would get the fifth base, and so on. Even if the final product were merely 10 bases in length, poor coupling would yield minuscule amounts of full-length product. It is critical for DNA synthesizers to have about 99% efficiency in each round, and then truncated products are the minority of the final sample. With high efficiencies, it is possible to synthesize longer segments of DNA. At 99% efficiency, an oligonucleotide that is 100 nucleotides long would give about 30%–40% final yield. If the desired oligonucleotide is separated from the truncated products by electrophoresis (see Chapter 3), it is possible to get plenty of full-length products.

Complete genes can be synthesized by linking smaller oligonucleotides together (Fig. 4.18). If the complete sequence of a gene is known, then long oligonucleotides can be synthesized identical to that sequence. The efficiency of the DNA synthesizer usually limits the length of each segment to about 100 bases; therefore, the gene segments are made with overlapping ends. Because oligonucleotides are single-stranded, both

strands of the gene must be synthesized and annealed to each other, and then the segments are linked using ligase. Another strategy for assembly is to create strands that overlap only partially and then use DNA polymerase I to fill in the large single-stranded gaps.

DNA can be synthesized in long segments provided each base is added efficiently. These long segments can be linked into one complete gene.

POLYMERASE CHAIN REACTION USES *IN VITRO* SYNTHESIS TO AMPLIFY SMALL AMOUNTS OF DNA

The **polymerase chain reaction (PCR)** amplifies small samples of DNA into large amounts, much as a photocopier makes many copies of a sheet of paper. The DNA is amplified using the principles of replication; that is, the DNA is replicated over and over by DNA polymerase until a large amount is manufactured. Kary Mullis invented this technique while working at Cetus in 1983. He later won the Nobel Prize in Chemistry for PCR because of its huge impact on biology and science. PCR is used in forensic medicine to identify victims or criminals by amplifying the minuscule amounts of DNA left at a crime scene (see Chapter 23); PCR can identify infectious diseases such as HIV before symptoms emerge (see Chapter 21); PCR can amplify specific segments of genes without the need for cloning the segment first; in fact, PCR is now used in all aspects of the biological sciences.

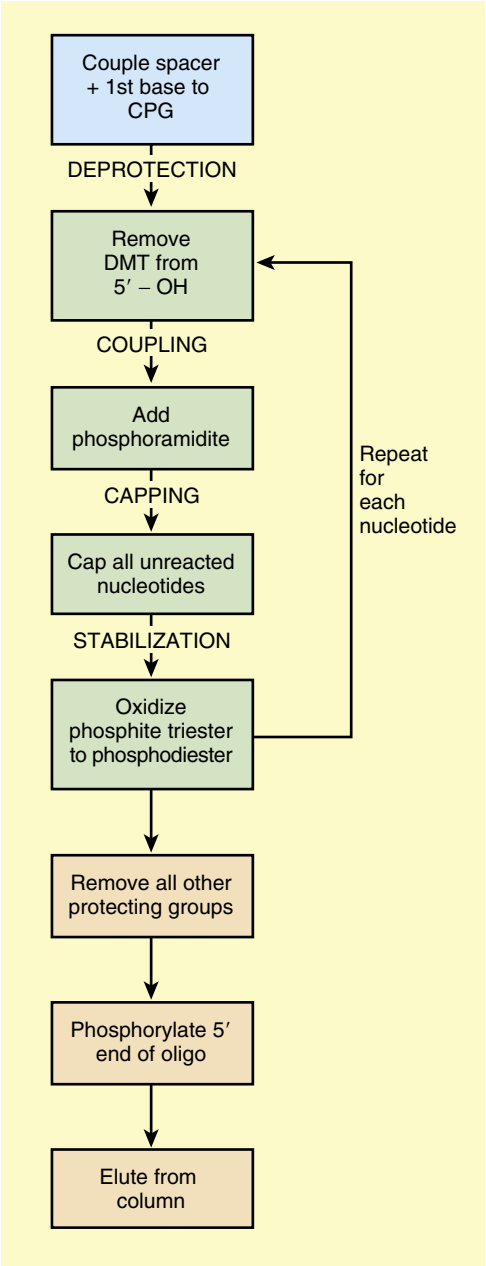

FIGURE 4.17 Flow Chart of Oligonucleotide Synthesis Oligonucleotide synthesis has many steps that are repeated. The first nucleotide is coupled to a bead with a spacer molecule. Next, the 5′-DMT is removed, and activated phosphoramidite nucleotide is added to the 5′ end of the first nucleotide. All the first nucleotides that were not linked to a second nucleotide are capped to prevent any further extension. Next, the phosphite triester is converted to a phosphodiester. These steps (in green) are repeated for the entire length of the oligonucleotide. Once the oligonucleotide has the appropriate length, the steps in tan are performed on the entire molecule.

Just as the photocopier needs more paper, ink, and a machine to make the copies, PCR requires specific reagents. The sample to be copied is called the template DNA, and this is often a known sequence or gene. The template DNA is typically double-stranded, and extremely small quantities are sufficient. The template DNA can be found within a complex mixture such as whole genomic DNA samples or within a fairly simple sample of bacterial plasmid DNA. The second reagent needed for PCR is a pair of oligonucleotide primers, which have sequences complementary to the ends of the template DNA. The DNA primers are oligonucleotides about 8 to 20 nucleotides long. One primer anneals to the 5′ end of the sense strand, and the other anneals to the 3′ end of the antisense strand of the target sequence. The primer sequences specify the exact target region of the DNA sample, thus focusing the reaction on the template DNA even if it is found within a complex mixture of genomic DNA. The third reagent is a supply of

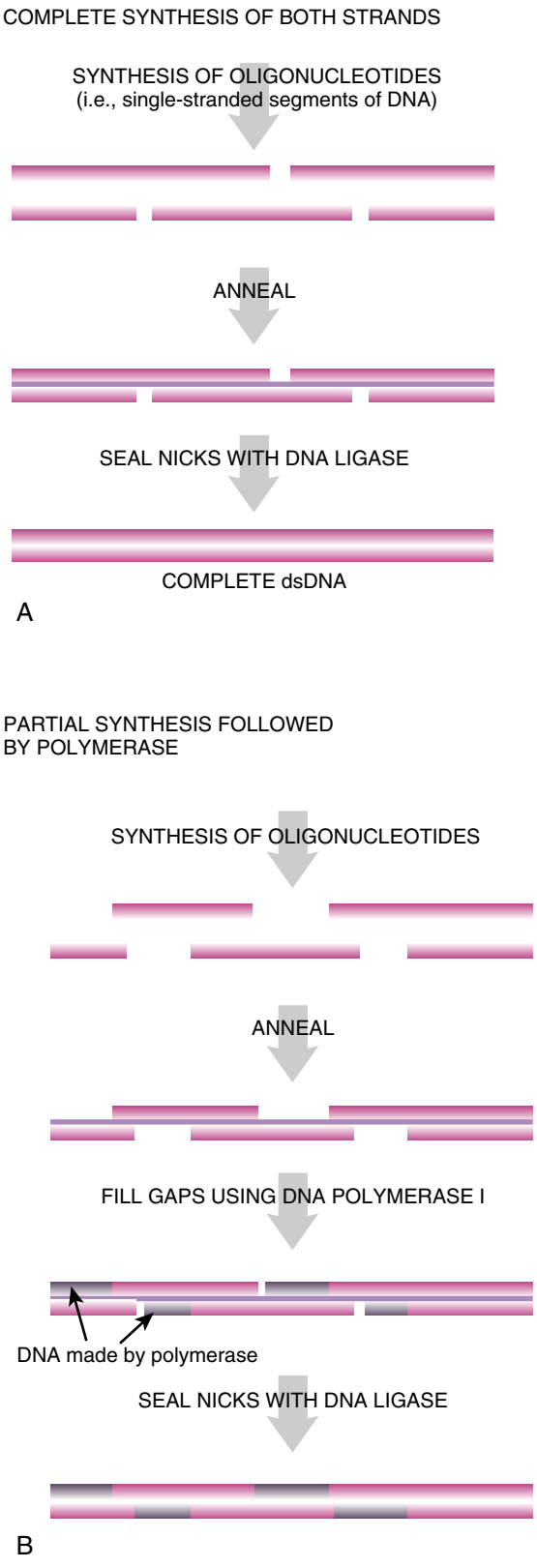

FIGURE 4.18
Synthesis and Assembly of a Gene
(A) Complete synthesis of both strands. Small genes can be chemically synthesized by making overlapping oligonucleotides. The complete sequence of the gene, both coding and noncoding strands, is made from small oligonucleotides that anneal to each other, forming a double-stranded piece of DNA with nicks along the phosphate backbone. The nicks are then sealed by DNA ligase. (B) Partial synthesis followed by polymerase. To manufacture longer pieces of DNA, oligonucleotides are synthesized so that a small portion of each oligonucleotide overlaps with the next. The entire sequence is manufactured, but gaps exist in both the coding and noncoding strands. These gaps are filled using DNA polymerase I, and the remaining nicks are sealed with DNA ligase.

nucleoside triphosphates, and the final reagent is ***Taq* DNA polymerase** from *Thermus aquaticus*, which actually makes the copies.

The basic mechanism of PCR includes heat denaturation of the template, annealing of the primers, and making a complementary copy using DNA polymerase, each step found in DNA replication. The three steps are repeated over and over until one template strand generates millions of identical copies. An amount of DNA too small to be seen can be copied so that it can be cloned into a vector, or visualized on an agarose gel (see Chapter 3). The process requires changing the temperature in a cyclic manner. Changing temperatures is accomplished by a **thermocycler**, a machine designed to change the temperature of its heat block rapidly so that each cycle can be completed in minutes. The temperature cycles between 94°C to denature the template; 50°C–60°C to anneal the primer (depending on the length and sequence of the primer); and 72°C for *Taq* polymerase to make new DNA. Before thermocyclers were developed, PCR was accomplished by moving the mixture among three different water baths at different temperatures every few minutes, which was very tedious.

In principle, the PCR cycle resembles DNA replication with a few modifications (Fig. 4.19). Like other *in vitro* DNA synthesis reactions, the double-stranded template is denatured with high heat rather than enzymes. Then the temperature is lowered so that the primers anneal to their binding sites. The primers are made so that each binds to opposite strands of the template, one at the beginning and one at the end of the gene. Then DNA polymerase elongates both primers and converts both single template strands to double-stranded DNA. (Note: During sequencing, only one primer is used and only one strand of the template is replicated, but during PCR both strands are copied.) *Taq* polymerase is the most widely used polymerase for PCR because it is very stable at high temperatures and does not denature at the high temperatures needed to separate the strands of the template DNA. *Taq* polymerase comes from *Thermus aquaticus*, a bacterium that grows in the hot springs of Yellowstone Park, USA. After the first replication cycle, the whole process is repeated. The two DNA strands are denatured at high heat, and then the temperature drops to allow the primers

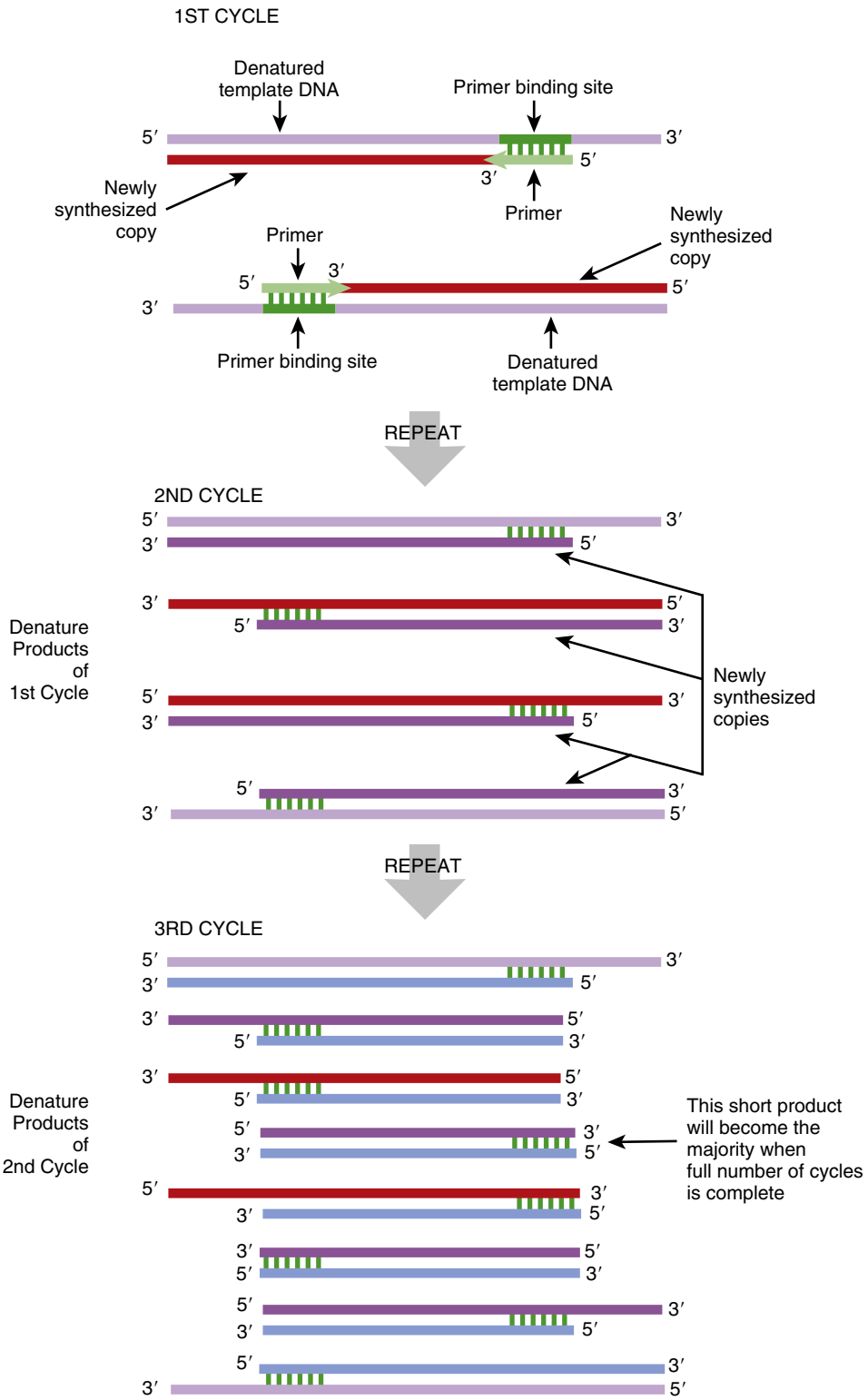

FIGURE 4.19 PCR, the First Three Cycles

In the first cycle, double-stranded template DNA (light purple) is denatured, complementary primers are annealed to the primer binding sites, and a new copy of the template is generated by *Taq* polymerase (red). In the second cycle, the two double-stranded products from the first cycle are denatured to form four single-stranded templates. The same set of primers anneals to the four template strands, and *Taq* polymerase makes each of the four double-stranded (dark purple). In the third cycle, the four double-stranded products from the second cycle are denatured, the primers anneal, and the four products from the second cycle become eight (light blue). Each subsequent round of denaturation, primer annealing, and extension doubles the number of copies, turning a small amount of template into a large amount of PCR product.

to anneal to their target sequences. *Taq* polymerase synthesizes the next four strands, and now there are four double-stranded copies of the target sequence. Early in the process, some longer strands are generated; however, eventually only the segment flanked by the two primers is amplified. Ultimately, the template strands and early PCR products become the minority. The shorter products become the majority.

The primers are key to the process of PCR. If the primers do not anneal in the correct location, if the span between the primers is too large, or if the primers form hairpin regions rather than annealing to the target, then *Taq* polymerase will not be able to amplify the segment. Also, if both primers anneal to the same strand, the reaction will not work. If the template has a known sequence, primers are synthesized based on the sequences upstream and downstream of the region to be amplified. Modifications exist that allow researchers to analyze unknown sequences by PCR (see later discussion).

PCR is a process that uses DNA polymerase in an *in vitro* sequencing reaction. Here, a double-stranded template is replicated to make two copies. Each of these products is replicated to make four, and the process continues exponentially.

MODIFICATIONS OF BASIC PCR

Many different permutations of PCR have been devised since Kary Mullis developed the basic procedure. All rely on the same basic PCR reaction, which takes a small amount of DNA and amplifies it by *in vitro* replication. Many of these variant protocols are essential tools for recombinant DNA research.

Several strategies allow amplifying a DNA segment by PCR even if its sequence is unknown. For example, the unknown sequence may be cloned into a vector (whose sequence is known). The primers are then designed to anneal to the regions of the vector just outside the insert.

In another scenario, the sequence of an encoded protein is used to generate PCR primers. Remember that most amino acids are encoded by more than one codon. Thus, during translation of a gene, one or more codons are used for the same amino acid. Therefore, if a protein sequence is converted backwards into nucleotide sequence, the sequence is not unique. For example, two different codons exist for histidine and glutamine, and four codons exist for serine. Consequently, the nucleotide sequence encoding the amino acid sequence histidine–glutamine–valine can be one of 16 different combinations.

If primers are made that depend on protein sequence, they will be **degenerate primers** and they will have a mixture of two or three different bases at the wobble positions in the triplet codon. During oligonucleotide synthesis, more than one phosphoramidite nucleotide can be added to the column at a particular step. Some of the primers will have one of the nucleotides, whereas other primers will have the other nucleotide. If many different wobble bases are added, a population of primers is created, each with a slightly different sequence. Within this population, some will bind to the target DNA perfectly, some will bind with only a few mismatches, and some won't bind at all. Of course, the annealing temperature for degenerate primers is adjusted to allow for some mismatches.

Inverse PCR is a trick used when sequence information is known only on one side of the target region (Fig. 4.20). First, a restriction enzyme is chosen that does not cut within the stretch of known DNA. The length of the recognition sequence should be six or more base pairs in order to generate reasonably long DNA segments for amplification by PCR. The target DNA is then cut with this restriction enzyme to yield a piece of DNA that has compatible sticky ends, one upstream of the known sequence and one downstream. The two ends are ligated to form a circle. The PCR primers are designed to recognize the end regions of the known sequence. Each primer binds to a different strand of the circular DNA, and they both point "outward" into the unknown DNA. PCR then amplifies the unknown DNA to give linear molecules with short stretches of known DNA at the ends, and the restriction enzyme site in the middle.

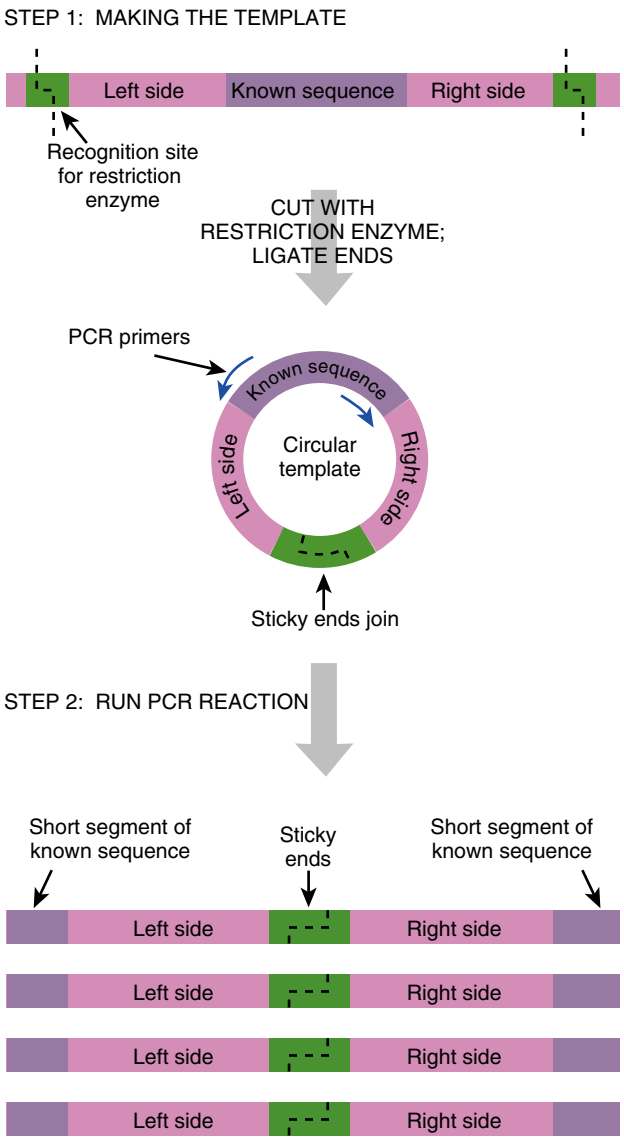

FIGURE 4.20 Inverse PCR

Inverse PCR allows unknown sequences to be amplified by PCR provided that they are located near a known sequence. The DNA is cut with a restriction enzyme that cuts upstream and downstream of the known region but not within it. The linear piece of DNA is circularized and then amplified with primers that anneal in the known region. The PCR products have the unknown DNA from the left and right of the known sequence. These can be cloned and sequenced.

Degenerate primers are designed based on amino acid sequences and contain different nucleotides at the wobble position.

Inverse PCR sequences DNA near a known sequence by finding a restriction enzyme recognition sequence away in the unknown region, cutting out this template, and amplifying the entire piece with *Taq* polymerase.

REVERSE TRANSCRIPTASE PCR

Reverse transcriptase PCR (RT-PCR) uses the enzyme reverse transcriptase to make a cDNA copy of mRNA from an organism and then uses PCR to amplify the cDNA (Fig. 4.21). The advantage of this technique is evident when trying to use PCR to amplify a gene from eukaryotic DNA. Eukaryotes have introns, some extremely long, which interrupt the coding segments. After transcription, the primary RNA transcript is processed to remove all the introns, hence becoming mRNA. Using mRNA as the source of the target

DNA relies on the cell removing the introns. In practice, RT-PCR has two steps. First, reverse transcriptase recognizes the 3′ end of primers containing repeated thymines and synthesizes a DNA strand that is complementary to the mRNA. (The thymines base-pair with the poly(A) tail of mRNA.) Then the RNA strand is replaced with another DNA strand, leaving a double-stranded DNA (i.e., the cDNA). Next, the cDNA is amplified using a normal PCR reaction containing appropriate primers (one usually recognizes the poly(A) tail), *Taq* polymerase, and nucleotides.

RT-PCR uses reverse transcriptase to convert mRNA into double-stranded DNA, and then the gene without any introns can be amplified by regular PCR.

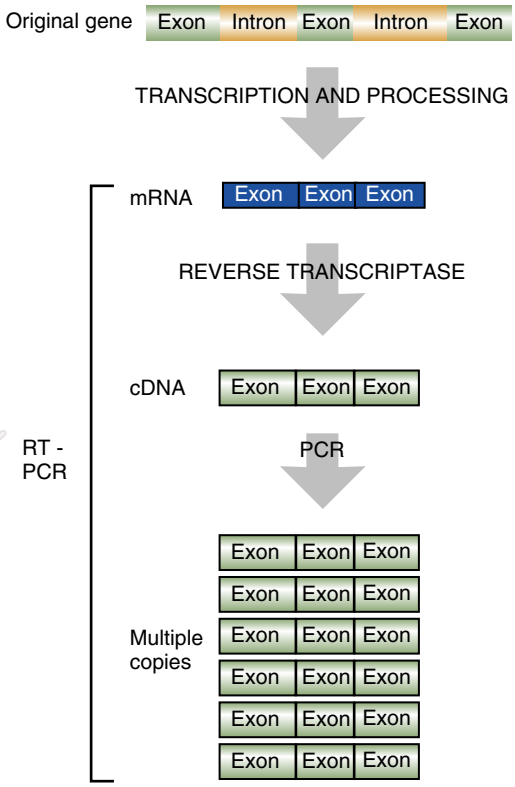

FIGURE 4.21
Reverse Transcriptase PCR
RT-PCR is a two-step procedure that involves making a cDNA copy of the mRNA and then using PCR to amplify the cDNA.

PCR IN GENETIC ENGINEERING

PCR allows scientists to clone genes or segments of genes for identification and analysis. PCR also allows scientists to manipulate a gene that has already been identified. Various modified PCR techniques allow scientists to hybridize two separate genes or genes segments into one, delete or invert regions of DNA, and alter single nucleotides to change the gene and its encoded protein in a more subtle way.

PCR can make cloning a foreign piece of DNA easier. Special PCR primers can generate new restriction enzyme sites at the ends of the target sequence (Fig. 4.22). The primer is synthesized so that its 5′ end has the desired restriction enzyme site, and the 3′ end has sequence complementary to the target. Obviously, the 5′ end of the primer does not bind to the target DNA, but as long as the 3′ end has enough matches to the target, then the primer will still anneal. *Taq* polymerase primes synthesis from the 3′ end; therefore, the enzyme is not bothered by mismatched 5′ sequences. The resulting PCR product can easily be digested with the corresponding restriction enzyme and ligated into the appropriate vector.

Rather than incorporating restriction enzyme sites into the ends of the PCR product, **TA cloning** will clone any PCR product directly (Fig. 4.23). *Taq* polymerase has terminal transferase activity that generates a single adenine overhang on the ends of the PCR products it makes. Special vectors containing a single thymine overhang have been developed, and simply mixing the PCR product with the TA cloning vector plus DNA ligase clones the PCR product into the vector without any special modifications.

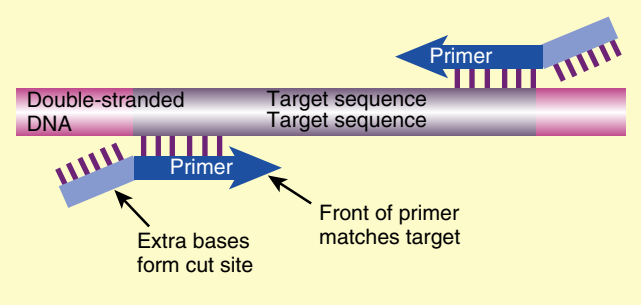

FIGURE 4.22 **Incorporation of Artificial Restriction Enzyme Sites**
Primers for PCR can be designed to have nonhomologous regions at the 5′ end that contain the recognition sequence for a particular restriction enzyme. After PCR, the amplified product has the restriction enzyme sites at both ends. If the PCR product is digested with the restriction enzyme, this generates sticky ends that are compatible with a chosen vector.

PCR can be used to manipulate cloned genes also. Two different gene segments can be hybridized into one using **overlap PCR** (Fig. 4.24). Here, PCR amplification occurs with three primers: one is complementary to the beginning of the first gene segment, one is complementary to the end of the second gene segment, and a third is half complementary to the end of gene segment 1 and half complementary to the beginning of gene segment 2. During PCR, the two gene segments become fused into one by a mechanism that is hard to visualize, but probably involves looping of some of the early PCR products.

PCR can be used to create large deletions or insertions into a gene (Fig. 4.25). Once again the design of the PCR primers is key to the construction. For example, primers to generate insertions have two regions: the first half is homologous to the sequence around the insertion point; the second half has sequences complementary to the insert sequence. For example, suppose an antibiotic resistance gene such as *npt* (confers resistance to neomycin) is to be inserted into a cloning vector. The primers would have their 5′ ends complementary to the sequence flanking the insertion point on the vector and their 3′ ends complementary to the ends of the *npt* gene. First, the primers are used to amplify the *npt* gene and give a product with sequences homologous to the vector flanking both ends. Next, the PCR product is transformed into bacteria harboring the vector. The *npt* gene recombines with the insertion point by homologous recombination, resulting in insertion of the *npt* gene into the vector.

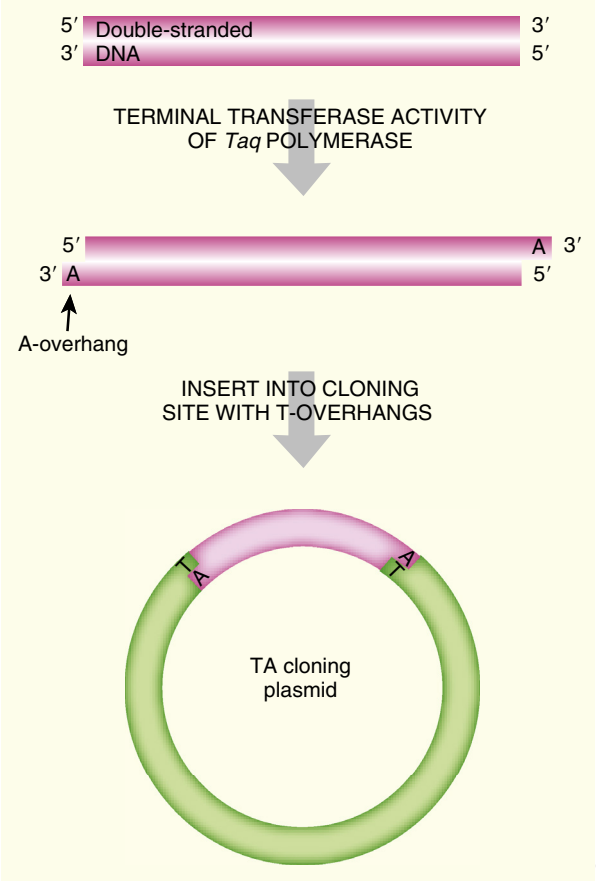

FIGURE 4.23 TA Cloning of PCR Products
When *Taq* polymerase amplifies a piece of DNA during PCR, the terminal transferase activity adds an extra adenine at the 3′ ends. The TA cloning vector was designed so that when linearized, it has a single 5′-thymine overhang. The PCR product can be ligated into this vector without the need for special restriction enzyme sites.

The insertion point(s) will determine whether the antibiotic cassette causes just an insertion or both an insertion plus a deletion. If the two PCR primers recognize separate homologous recombination sites, then the incoming PCR segment will recombine at these two sites. Homologous recombination then results in the *npt* gene replacing a piece of the vector rather than merely inserting at one particular location.

PCR can also generate nucleotide changes in a gene by **directed mutagenesis** (Fig. 4.26). Usually, only one or a few adjacent nucleotides are changed. First, a mutagenic PCR primer is synthesized that has nucleotide mismatches in the middle region of the primer. The primer

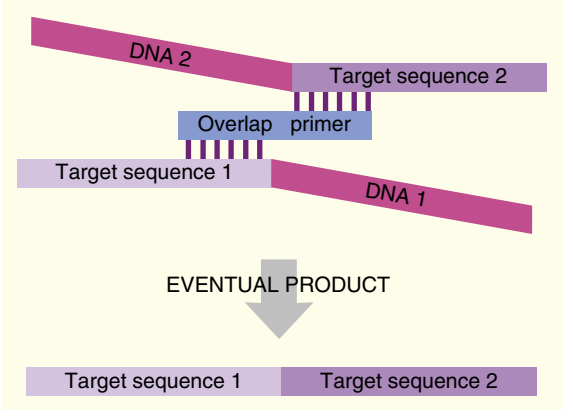

FIGURE 4.24 Overlap PCR
Overlapping primers can be used to link two different gene segments. In this scheme, the overlapping primer has one end with sequences complementary to target sequence 1, and the other half similar to target sequence 2. The PCR reaction will create a product with these two regions linked together.

will anneal to the target site with the mismatch in the center. The primer needs to have enough matching nucleotides on both sides of the mismatch so that binding is stable during the PCR reaction. The mutagenic primer is paired with a normal primer. The PCR reaction then amplifies the target DNA incorporating the changes at the end with the mutagenic primer. These changes may be relatively subtle, but if the right nucleotides are changed, then a critical amino acid may be changed. One amino acid change can alter the entire function of a protein. Such an approach is often used to assess the importance of particular amino acids within a protein.

The 5′ end of PCR primers does not need to be complementary to the template DNA, and can be designed to add restriction enzyme sites to the PCR product. The terminal transferase activity by *Taq* polymerase adds a single adenine onto the 3′ end of the PCR product. These traits allow the PCR product to be cloned into a vector.

PCR can be used to delete, insert, and even fuse different gene segments.

PCR can be used to make small changes in nucleotide sequences by directed mutagenesis.

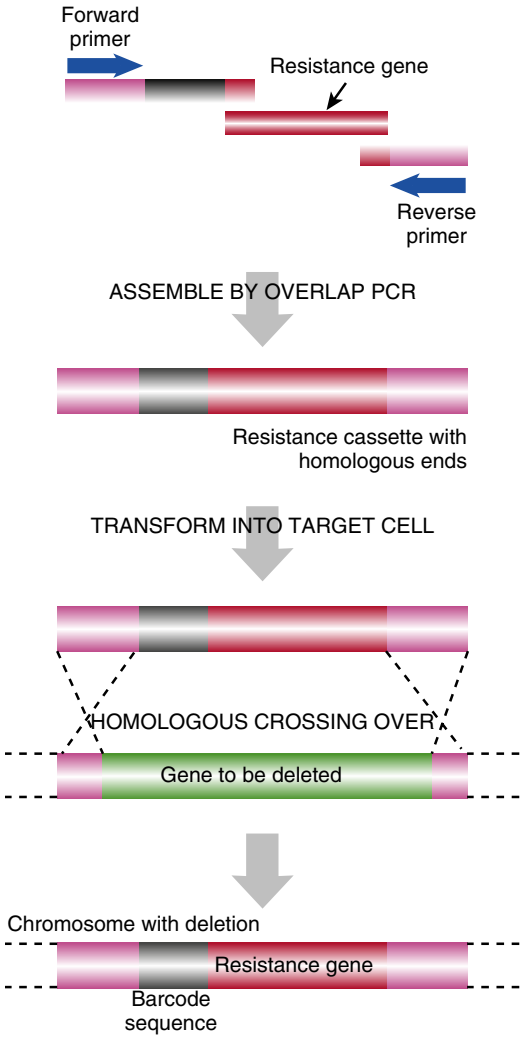

FIGURE 4.25 Generation of Insertions or Deletions by PCR In the first step, a specifically targeted cassette is constructed by PCR. This contains both a suitable marker gene and upstream and downstream sequences homologous to the target site. The engineered cassette is transformed into the host cell, and homologous crossing over occurs. Recombinants are selected by the antibiotic resistance carried on the cassette. The barcode sequence is a unique DNA sequence only found in the cassette used to identify the location of the cassette.

PCR OF DNA CAN DETERMINE THE SEQUENCE OF BASES

Being able to quickly and easily determine the sequence of any gene has been the driving force for the recent advances made in biotechnology. Frederick Sanger developed a method for sequencing a gene *in vitro* in 1974. He was interested in the amino acid sequence of insulin and decided to deduce the sequence of the protein from the nucleotide sequence. He invented the **chain termination sequencing** method, which is still used today (Fig. 4.27). Much like DNA replication, chain termination sequencing requires a primer, DNA polymerase, a single-stranded DNA template, and deoxynucleotides. During *in vitro* sequencing reactions, these components are mixed and DNA polymerase makes many copies of the original template. The first trick needed to deduce the sequence is to stop synthesis of the newly synthesized DNA chains at each base pair. Consequently, the fragments generated will differ in size by one base pair and, when separated by gel electrophoresis, create a ladder of fragments. The next step is to figure out the identity of the last nucleotide. If the final base pair for each fragment is known, the sequence may be directly read from the gel (reading from bottom to top).

But how do we know what the final base is for each fragment on the sequencing ladder? DNA polymerase synthesizes a new strand of DNA based on the template sequence. The chain consists of deoxynucleotides, each with a hydroxyl group at the 3′ position on the deoxyribose ring. DNA polymerase adds the next nucleotide by linking the phosphate of the incoming nucleotide to the 3′-hydroxyl of the previous nucleotide. If a nucleotide lacks

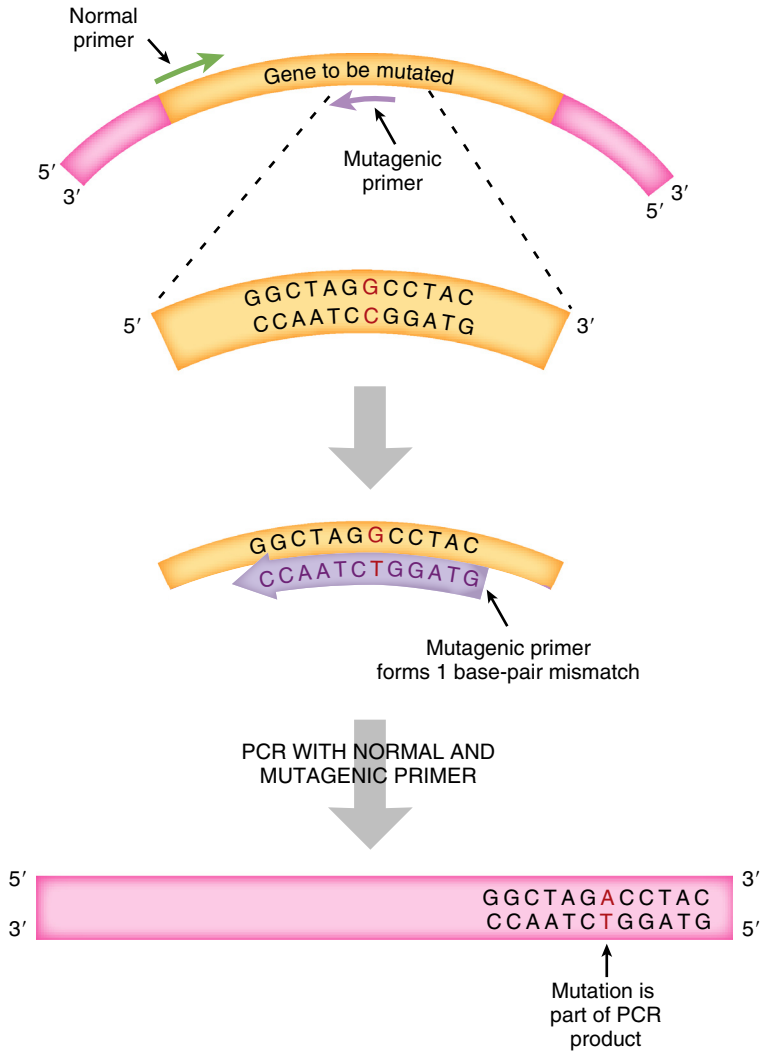

FIGURE 4.26 Direct Mutagenesis Using PCR

The gene to be mutated is cloned, and the entire sequence is known. To alter one specific nucleotide, normal and mutagenic primers are combined in a PCR reaction. The mutagenic primer will have a mismatch in the middle, but the remaining sequences will be complementary. The PCR product will incorporate the sequence of the mutagenic primer.

the 3′-hydroxyl, no further nucleotides can be added and the chain is terminated (Fig. 4.28). During a sequencing reaction, a certain percentage of nucleotides with no 3′-hydroxyl, called **dideoxynucleotides**, are mixed with the normal deoxynucleotides. Such reactions typically have a maximum length of about 800 nucleotides.

The fragments are relatively small for DNA and vary in length by only one nucleotide; therefore, they must be separated by size using polyacrylamide gel electrophoresis (see Chapter 3). The principle is the same as for agarose gel electrophoresis, but polyacrylamide has smaller pores, and so smaller fragments can be separated with higher resolution. The sequence is actually read from the bottom of the gel to the top, because the fragments terminated closest to the primer are smaller (hence run faster) than the ones further from the primer. The bands appear as a ladder, each separated by one nucleotide; therefore, each band represents the fragments ending with the dideoxynucleotide complementary to the template strand.

Automated DNA sequencing uses a PCR-type reaction to sequence DNA. In PCR sequencing, or **cycle sequencing**, the template DNA with unknown sequence is amplified by *Taq* polymerase as any normal PCR reaction. *Taq* DNA polymerase was modified to remove its proofreading ability and increase the speed at which it incorporates nucleotides. Cycle sequencing

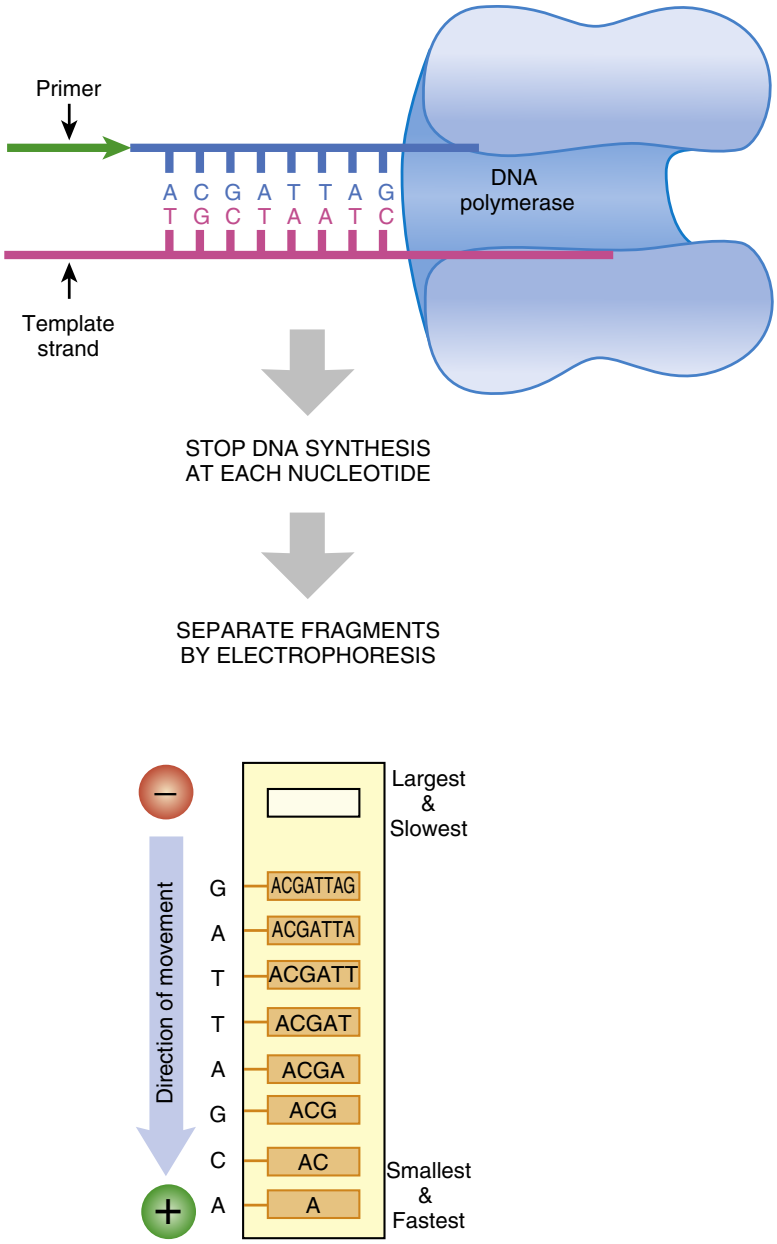

FIGURE 4.27 Chain Termination Method of Sequencing

During chain termination, DNA polymerase synthesizes many different strands of DNA from the single-stranded template. DNA polymerase will stop at each nucleotide, such that strands of all possible lengths are made. They are separated by size using electrophoresis. The smallest fragments are at the bottom and represent the primer plus only the first one or two nucleotides of the template DNA. Longer fragments contain the primer plus longer stretches of synthesized DNA complementary to to the template DNA.

reaction mixtures include all four deoxynucleotides, all four dideoxynucleotides, a single primer, template DNA, and *Taq* polymerase. To discern the identity of the dideoxynucleotide, they are linked to a unique fluorophore for each of the four nucleotides.

The samples are amplified in a thermocycler. First, the template DNA is denatured at a high temperature; then the temperature is lowered to anneal the primer; and finally, the temperature is raised to 72°C, the optimal temperature for *Taq* polymerase to make DNA copies of the template. During polymerization, dideoxynucleotides are incorporated and cause chain termination. The ratio of dideoxynucleotides to deoxynucleotides is adjusted to ensure that some fragments stop at each G, A, T, or C of the template strand. After *Taq* polymerase makes thousands of copies of the template, each stopping at a different nucleotide, the entire

mixture is separated in one lane of a sequencing gel (Fig. 4.29). Bands of four different colors are seen, corresponding to the four fluorescently labeled dideoxynucleotides and hence the four bases.

Cycle sequencing has many advantages. During cycle sequencing, each round brings the temperature to 95°C, which destroys any secondary structures or double-stranded regions. Another advantage of cycle sequencing is to control primer hybridization. Some primers do not work well with regular sequencing reactions because they bind to closely related sequences. During cycle sequencing, the primer annealing temperature is controlled and can be set quite high in order to combat nonspecific binding. Finally, cycle sequencing requires very little template DNA; therefore, sequencing can be done from smaller samples.

Another advance in sequencing has been the detection system. **Automatic DNA sequencers** detect each of the fluorescent tags and record the sequence of bases (Fig. 4.30). Some automatic DNA sequencers can read up to 384 different DNA samples using capillary tubes filled with gel matrix to separate the DNA fragments. At the bottom of each capillary tube is a fluorescent activator, which emits light to excite the fluorescent dyes. On the other side is the detector, which reads the wavelength of light that the fluorescent dye emits. As each fragment passes the detector, it measures the wavelength and records the data as a peak on a graph. For each fluorescent dye, a peak is recorded and assigned to the appropriate base. An attached computer records and compiles the data into the DNA sequence.

Automated sequencing has a large startup cost because the sequence analyzer is quite expensive, but they run multiple samples at one time, and thus the cost per sample is quite low. Many universities and companies have a centralized facility that does the sequencing for all the researchers. In fact, sequencing has become so automated that many researchers just send their template DNA and primers to a company that specializes in sequencing.

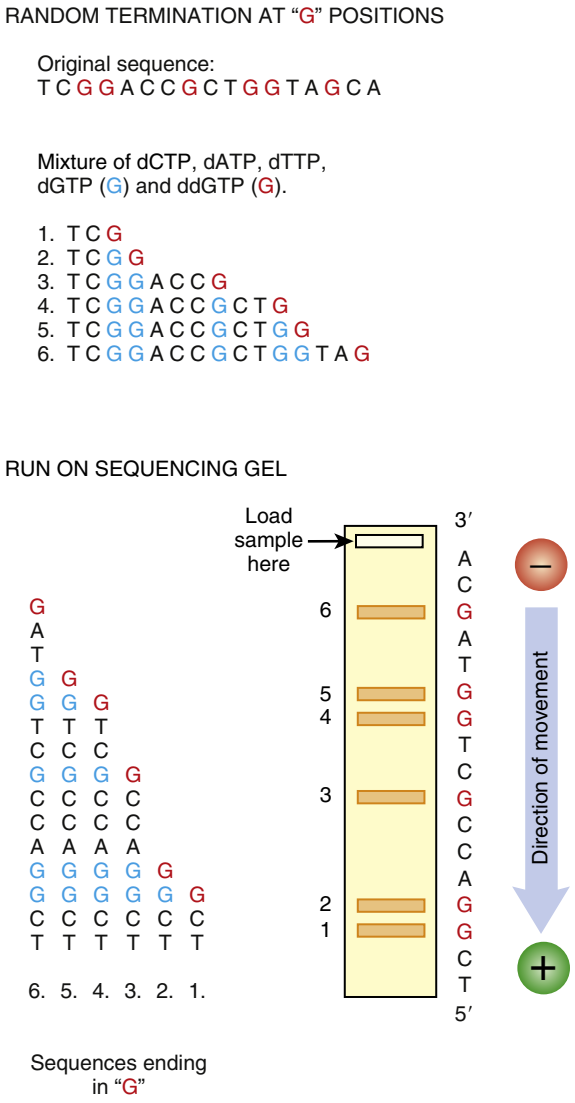

FIGURE 4.28 Chain Termination by Dideoxynucleotides
During the sequencing reaction, DNA polymerase makes multiple copies of the original sequence. Sequencing reaction mixtures contain dideoxynucleotides that terminate growing DNA chains. The example here shows a sample reaction, which includes triphosphates of both deoxyguanosine (dG) and dideoxyguanosine (ddG). Whenever ddG is incorporated (shown in red), it causes termination of the growing chain. If dG (blue) is incorporated, the chain will continue to grow. When the sequencing reaction containing the ddG is separated on a polyacrylamide gel, the fragments are separated by size. Each band directly represents the fragment ending in G from the original sequence.

Chain-terminating dideoxynucleotides are the key to determining DNA sequence. When these are incorporated into an *in vitro* replication reaction, DNA polymerase cannot add any more nucleotides and the synthesis reaction ends. In cycle sequencing, a PCR reaction includes a controlled amount of fluorescently labeled dideoxynucleotides. *Taq* polymerase stops adding nucleotides when a dideoxynucleotide is incorporated. The fluorescent tag is used to identify the ending base of each fragment using an automated sequencer.

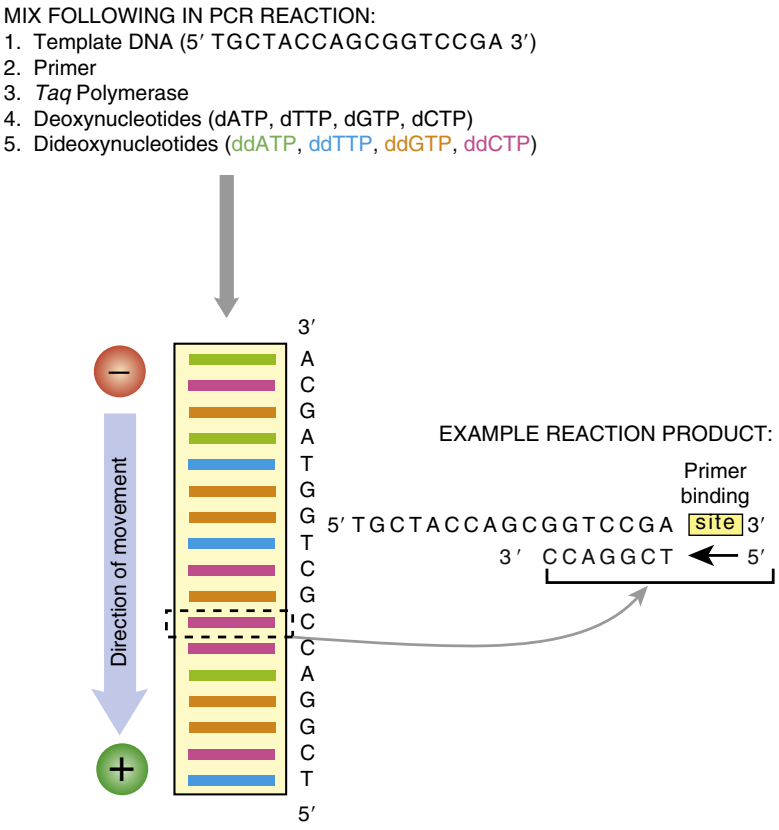

FIGURE 4.29 Cycle Sequencing
During cycle sequencing, the reaction contains template DNA, primer, *Taq* polymerase, deoxynucleotides, and dideoxynucleotides. Each of the different dideoxynucleotides has a different fluorescent label attached. The automated sequencer detects the color and compiles the sequence data.

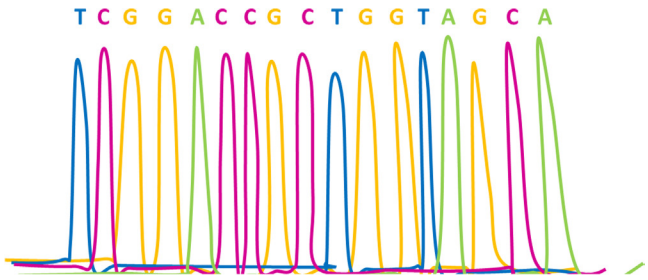

FIGURE 4.30 Data from an Automated Sequencer
A representative set of data from an automated sequencer. The fluorescent peaks for the individual bases are shown. The computer compiles the information into a sequence file for the researcher.

NEXT-GENERATION SEQUENCING TECHNOLOGIES

Sequencing DNA using chain termination was the workhorse for the initial sequencing of the first human genome. Throughout the human genome project, the cost for each base of DNA dropped by making advances in the capillary electrophoresis chain–termination method. The cost for sequencing one million bases of DNA in September 2001, the end of the initial sequence, was $5292, and so for the whole human genome, over $95 million. Because of the advances in chain termination sequencing, the human genome project was done early and under budget. As of October 2013, the cost to sequence one million base pairs of DNA dropped less than 6 cents. The cost to sequence an entire human genome, therefore, is a mere $5096. The incredible decrease in cost stems from the advent of **massively parallel sequencing**, which is a descriptive name for **next-generation sequencing**. These

technologies use a type of platform that can hold millions of DNA fragments in separate locations. There are many different chemistries used in next-generation sequencing, and they are rapidly changing. Two sequencing platforms, 454 sequencing and Illumina, are outlined here.

The first step of any next-generation technology is to prepare the DNA for amplification by PCR (Fig. 4.31). Genomic DNA is isolated from the organism of interest according to a standard DNA isolation protocol. The pure DNA is then sheared into small fragments using sonication. To amplify each of the fragments, the end of each piece of DNA must have known sequence. This is impossible, especially for genomes that have never been sequenced. And even if the genome sequence is known, sonication creates random breaks in the DNA, so there is no way to truly know the sequence at each end. The trick to circumvent this problem is to add **linkers** or **adaptors**, which are short DNA pieces with a known sequence. They are added to the ends using the TA cloning technology. The linker or adaptor sequence depends on which of the next-generation sequencing technologies are employed. A **barcode sequence** or an **index sequence** is a key feature of the adaptor. The barcode or index sequence is much like a zip code in your address: the sequence is unique to the sample of DNA, and it allows multiple samples of DNA to be analyzed at the same time, a procedure called **multiplexing**.

Once the DNA sample is fragmented and adaptors are added onto each end, the DNA is attached to a solid surface so that individual DNA fragments are separated from each other. In 454 sequencing, the DNA fragments are attached to beads via the adaptors. The set of beads with small DNA oligonucleotides complementary to the adaptor is mixed with the DNA at a ratio such that one DNA fragment will attach to a single bead. Ensuring that a single DNA from the genome attaches to a single bead is a critical step for sequencing. In the Illumina sequencing methodology, the same principle applies, but the DNA fragments are added to the surface of a flow cell. The surface has DNA primers complementary to the adaptor scattered on the surface. These must also be of sufficient distance from each other to appear as a separate location by sensors at the bottom of the flow cell.

The next step for next-generation sequencing is to create multiple copies of the single piece of DNA using PCR. For 454 sequencing, **emulsion PCR** creates multiple copies of the single piece of DNA that attaches to the bead. The process begins by creating an emulsion of oil and water such that only one bead is found in each of the water droplets. In addition, the water droplets contain free deoxynucleotides, primers complementary to the adaptors, and *Taq* DNA polymerase. Within the droplets, the DNA fragment is amplified using the traditional denaturation, annealing, and elongations steps. The final result is a bead coated with identical copies of the DNA fragment. The emulsion prevents the DNA from one bead diffusing to a different bead.

In a similar fashion, DNA fragments attached to the surface of a flow cell for Illumina sequencing are amplified by incubating the flow cell with deoxynucleotides and DNA polymerase in a process called **bridge amplification**. The primers used for amplifying the DNA fragment are attached to the flow cell, so the DNA anneals to another primer on the surface, forming a bridge. These are amplified and released to form a cluster of identical DNA fragments.

Once a cluster of identical DNA pieces is produced on the bead or flow cell, these pieces are denatured into single-stranded DNAs competent for sequencing. Sequencing for 454 and Illumina occurs as the single-stranded DNA is replicated. Each technology uses a different detection method for identifying the sequence, but in both 454 and Illumina sequencing, each nucleotide is identified one by one; that is, as a nucleotide is added to the complementary strand, the identity is recorded by a sensor and stored by an attached computer. This method of sequencing is called **sequencing by synthesis**. In 454 sequencing, the beads

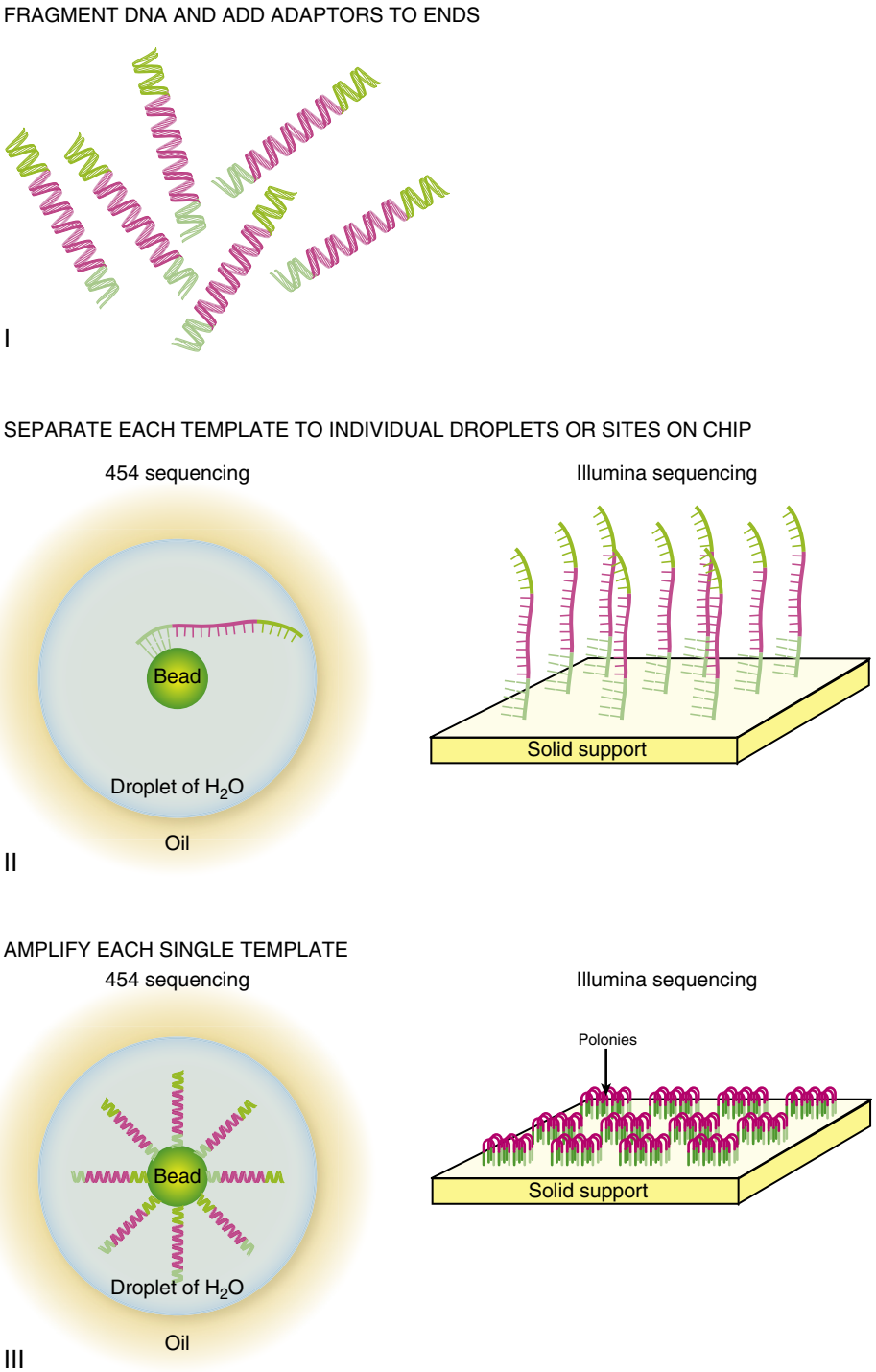

FIGURE 4.31 Next-Generation Sequencing
During next-generation sequencing, the DNA is prepared for amplification by isolating and fragmenting the sample. Adaptors are added to the ends of the fragments and then annealed to complementary oligonucleotides on the surface of the bead for 454 sequencing (left) or flow cell for Illumina sequencing (right). These are added such that one unique fragment attaches to one bead (left), or the spacing of the attached DNA is sufficient for recognition by the detector below the flow cell. Each single DNA template is amplified and denatured to be single-stranded. Sequence is determined after annealing primer to one end of the template and determining the identity of the each individually added nucleotide (see text for details). The sequence is detected by sensors below the well of the picotiter plate or below the flow cell and recorded by an attached computer.

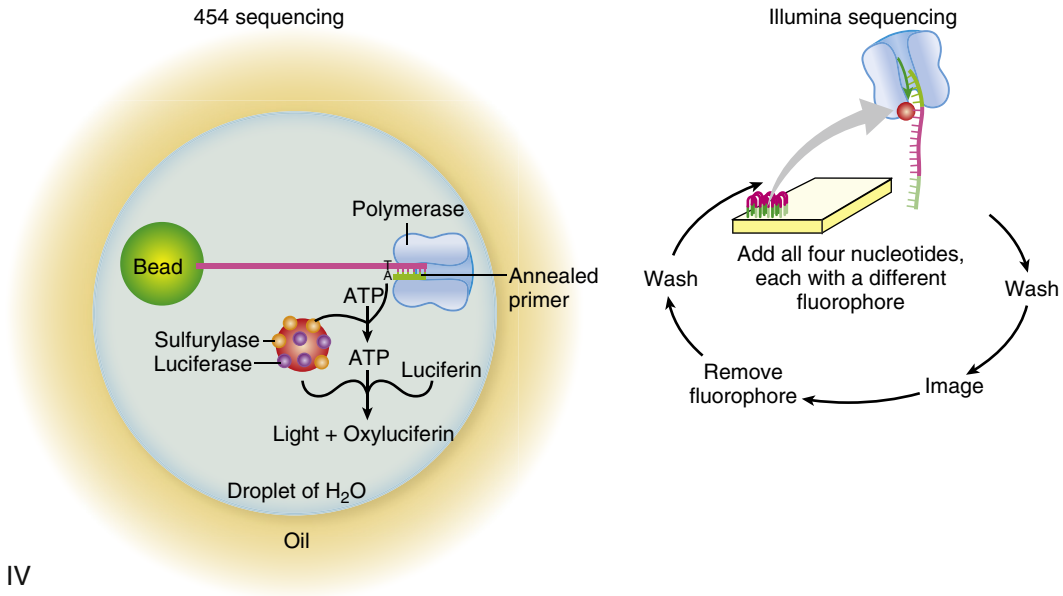

FIGURE 4.31 Cont'd

coated with copies of DNA are separated into a picotiter plate such that only one bead is within each well. A picotiter plate has over a million individual wells or holes in the surface that hold 75 picoliters. The lower surface of the well is optically clear to allow the light to be visualized by the detector. After a single bead enters the individual well or hole, a primer is annealed to the adaptor sequence on the fragment. Just like pyrosequencing, one of the four deoxynucleotides and DNA polymerase are floated across the picotiter plates, and if the template in the well has the complementary base pair, DNA polymerase adds the nucleotide to the primer, releasing a molecule of pyrophosphate. The well of the picotiter plate also contains luciferase and sulfurylase, which react with the released pyrophosphate and release a flash of light. At the bottom of the well, the attached sensor records the flash of light and sends the information to the computer. The flash of light appears only in the cells where the added nucleotide was incorporated, and the other cells remain dark. Each of the four nucleotides is added separately and washed away before adding another.

Illumina sequencing also determines each nucleotide as it is incorporated but uses a unique reversible fluorescent dye for guanine, cytosine, adenine, and thymine. After a primer is annealed to the adaptor sequence on the DNA template, all four nucleotides are added to the flow cell simultaneously. The wavelength of light released at each DNA cluster is recorded by the computer. The fluorescent dye terminator is removed, and another batch of four fluorescently labeled nucleotides is added. Again, the signal for each spot is recorded and stored. As each nucleotide is added, the computer compiles a sequence for each spot on the flow cell.

For both methods, the recording of data from each well of the picotiter plate or spot on the flow cell is compiled as a sequence for each separate DNA. Each of these pieces of sequence information is called a **read**. The technology for each method limits the number of nucleotides that can be determined with certainty, ranging from 50 to 400 base pairs depending on the sequencing machine and sequencing technology. To enhance the quality sequencing data results, **paired end reads** confirm the sequence information by repeating the entire sequence reaction, but using a primer to the opposite side of the fragment, thus essentially sequencing from both ends of the DNA fragment.

The amount of sequence data compiled by next-generation sequencing is tremendous, and without the increase in computer power and storage, there would be no way possible to compile the

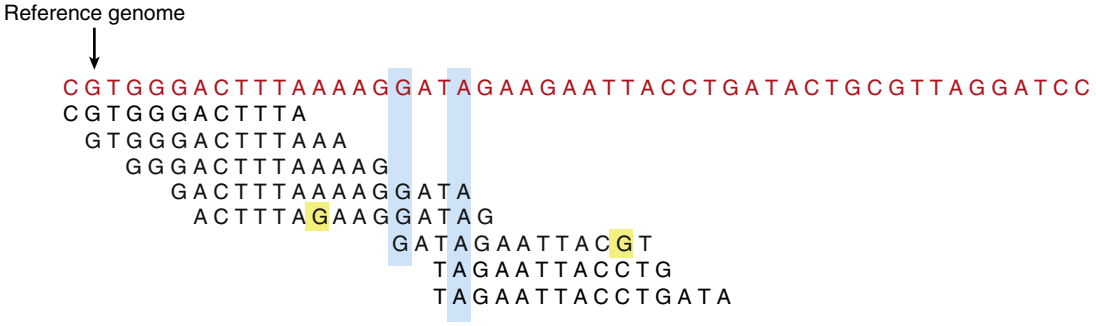

FIGURE 4.32 Data from Illumina Sequencing

The reference sequence is listed across the top. The sequence for the individual reads are shown and aligned under the identical sequence in the reference genome. The yellow boxes represent the location where the read and the reference genome differ. If these differences are found in every read, then they are most likely a true difference. If the difference is seen in only one read, then the change is most likely a sequencing error. Read depth varies from one nucleotide to the next (see blue rectangles). The higher the read depth, the more confidence the researcher has in the data.

data into a linear DNA sequence. Each read— that is, each set of sequence information from each well of the 1.7 million wells of the picotiter plate in 454 sequencing or each of the approximately 2 million DNA clusters on the flow cell in Illumina sequencing—represents a small, unique piece of an entire genome. Computers compare each piece of sequence back to a reference genome, if available, and the final output has the reference sequence across the top with each of the reads aligned below (Fig. 4.32). The number of sequences that map to a region in the genome is called **read depth**, and is used to ensure that a change is not an error in the sequencing process. If, for example, a single nucleotide substitution was found in a single read, but the other 29 reads for that region were identical to the reference genome, this substitution would most likely be considered an error. If, on the other hand, there was a single nucleotide substitution for 29 of 30 reads for that region, then this is most likely a true difference between the genome sequenced and the reference genome. If, for some reason, a region of the genome had a read depth of only 2, and a mutation was identified in this region, the validity would be highly suspect. On the other hand, if there was a mutation in 98% of the reads in a region where the read depth was 200, then the mutation is most likely a real substitution.

The ability to sequence an entire human genome has gone from a multiyear, multibillion dollar project to a simple procedure done in a few hours to few days, depending on the machine and technology used. In fact, the goal is to be able to reduce the cost of a human genome sequence to $1000. The ability to quickly assess the genomic sequence is going to change many disciplines. Understanding one genome can be misleading, but now we have the ability to compare genomes among different organisms and even among different people. In fact, a consortium of different universities and companies has compiled an integrated map of genetic variations by sequencing over 1000 different human genomes. Another major goal for this new technology is to compare cancer genomes to the normal genome. Early studies are identifying what mutations are common to cancers, and identifying mutations in genes that determine whether or not the patient will respond to a particular therapy. The ability to ascertain so much information so fast is bound to have applications that have yet to be discovered.

Summary

This chapter outlines the process of DNA replication. First, to replicate the DNA, DNA gyrase and DNA helicase relax the coiling in the DNA. The relaxed DNA is open and ready for the replisome to assemble at the origin. Single-stranded binding protein coats or binds to the open DNA, which keeps the DNA stable. Then PriA prepares an RNA primer at the origin to provide a 3′-OH group for DNA polymerase to attach the complementary bases during replication. DNA polymerase makes new DNA only in a 5′ to 3′ direction, so on the leading strand, the whole strand is made in one piece. Because the lagging strand is antiparallel, DNA polymerase has to make the strand in smaller segments called Okazaki fragments.

In vitro DNA synthesis can be made by purified DNA polymerase or by chemically linking nucleotides. In reactions done with chemical reagents, the DNA is single-stranded and is short because the process is not very efficient. Chemical synthesis of DNA is primarily used for making short primers or oligonucleotides.

In vitro DNA synthesis by DNA polymerase is very versatile and can be used to amplify a piece of DNA from a few copies to millions using PCR. Modifications of PCR include inverse PCR to amplify unknown regions of DNA, and RT-PCR of mRNA rather than DNA creates copies of genes without any introns. Additionally, PCR can be used to clone copies of genomic DNA into a vector using TA cloning or by adding novel restriction enzyme sites at the end of the PCR product. Finally, PCR can mutate template DNA by inserting or deleting regions, linking two separate regions together, or by mutating single nucleotides.

In vitro DNA synthesis is the basis for determining the sequence of DNA. In cycle sequencing, a single reaction contains four different fluorescently labeled dideoxynucleotides and unlabeled deoxynucleotides at a ratio that ensures one dideoxynucleotide incorporates at each nucleotide position of the template. The final reaction creates a tube filled with DNA fragments that end at each possible position, and that end nucleotide is fluorescently labeled. As the fragments are separated by size in a capillary tube-filled gel matrix, the smallest fragments exit the bottom of the tube first. As each subsequent fragment passes a detector, the identity of the fluorescent tag is determined and recorded. In contrast, next-generation sequencing reads nucleotides one by one as they are added to a primer; 454 sequencing employs pyrosequencing in order to determine what nucleotide is added. Since the release of pyrophosphate is the same for each of the four nucleotides, only one nucleotide is added at a time. The flash of light is recorded for each DNA fragment template where the nucleotide was incorporated. Illumina next-generation sequencing uses reversible 3′-fluorescent dye-linked nucleotides, which are added to the DNA template. Thus, all four nucleotides are added simultaneously to a flow cell containing the DNA templates. After the identity of the nucleotide that is added at each cluster is recorded, the fluorescent dye is removed from the nucleotide and washed away. The main difference between typical chain-termination sequencing and next-generation sequencing is the scale. Chain-termination sequencing occurs on one single template DNA. In contrast, 454 sequencing uses a picotiter plate with 1.7 million wells and Illumina's flow cell with several million DNA clusters, each well or cluster representing a single unique piece of DNA from the genome.

End-of-Chapter Questions

1. Which of the following enzymes aid in uncoiling DNA?
   - **a.** DNA gyrase
   - **b.** DNA helicase
   - **c.** topoisomerase IV
   - **d.** single-stranded binding protein
   - **e.** all of the above

2. Why is an RNA primer necessary during replication?
   - **a.** DNA polymerase III requires a 3′-OH to elongate DNA.
   - **b.** An RNA primer is not needed for elongation.
   - **c.** DNA polymerase requires a 5′-phosphate before it can elongate the DNA.
   - **d.** A DNA primer is needed for replication instead of an RNA primer.
   - **e.** An RNA primer is only needed once the DNA has been elongated and DNA polymerase is trying to fill in the gaps.

(*Continued*)

3. What are the functions of the two essential subunits of DNA polymerase III?
- **a.** Both subunits synthesize the lagging strand only.
- **b.** One subunit links nucleotides and the other ensures accuracy.
- **c.** They both function as a clamp to hold the complex to the DNA.
- **d.** The subunits function to break apart the bonds in the DNA strand.
- **e.** One subunit removes the RNA primer and the other synthesizes DNA.

4. Which of the following statements about mismatch repair is incorrect?
- **a.** MutSHL excise the mismatched nucleotides from the DNA.
- **b.** Mismatch repair proteins identify a mistake in DNA replication.
- **c.** The mismatch proteins recruit DNA polymerase III to synthesize new DNA after the proteins have excised the mismatched nucleotides.
- **d.** MutSHL can synthesize new DNA after a mismatch has been excised.
- **e.** MutSHL monitors the methylation state of the DNA to determine which strand contains the correct base when there is a mismatch.

5. Which of the following statements is incorrect regarding DNA replication?
- **a.** Rolling circle and theta replication are common for prokaryotes and viruses.
- **b.** Each round of replication for linear chromosomes, such as in eukaryotes, shortens the length of the chromosome.
- **c.** Prokaryotic chromosomes have multiple origins of replication.
- **d.** Eukaryotic replication only occurs during the S-phase of the cell cycle.
- **e.** Eukaryotic chromosomes have multiple origins of replication.

6. During *in vitro* DNA replication, which of the following components is not required?
- **a.** single-stranded DNA
- **b.** a primer containing a 3′-OH
- **c.** DNA helicase to separate the strands
- **d.** DNA polymerase to catalyze the reaction
- **e.** nucleotide precursors

7. Which of the following is not a step in the chemical synthesis of DNA?
- **a.** The 3′ phosphate group is added using phosphorylase.
- **b.** The addition of a blocking compound to protect the 3′ phosphite from reacting improperly.
- **c.** The 5′-OH is phosphorylated by bacteriophage T4 kinase.
- **d.** The addition of acetic anhydride and dimethylaminopyridine to cap the 5′-OH group of unreacted nucleotides.
- **e.** The amino groups on the bases are modified by other chemical groups to prevent the bases from reacting during the elongation process.

8. During chemical synthesis of DNA, a portion of the nucleotides does not react. How can the efficiency of such reactions be increased?
- **a.** The unreacted nucleosides are not acetylated so that more can be added in subsequent reactions.
- **b.** The efficiency of the reaction is not critical. Instead, the quality of the final product is more important than the quantity.
- **c.** The desired oligonucleotide can be separated from the truncated oligos by electrophoresis.
- **d.** Oligonucleotides should be made using DNA polymerase III instead of *in vitro* chemical synthesis.

e. The reaction times can be increased to allow the reaction to be more efficient.

9. Which of the following components terminates the chain in a sequencing reaction?
- **a.** dideoxynucleotides
- **b.** Klenow polymerase
- **c.** DNA polymerase III
- **d.** deoxynucleotides
- **e.** DNA primers

10. Which of the following statements about PCR is incorrect?
- **a.** The DNA template is denatured using helicase.
- **b.** PCR is used to obtain millions of copies of a specific region of DNA.
- **c.** A thermostable DNA polymerase is used because of the high temperatures required in PCR.
- **d.** Template DNA, a set of primers, deoxynucleotides, a thermostable DNA polymerase, and a thermocycler are the important components in PCR.
- **e.** Primers are needed because DNA polymerase cannot initiate synthesis, but can only elongate from an existing 3′-OH.

11. Which of the following is not an advantage of automated cycle sequencing over the chain termination method of sequencing?
- **a.** The reactions in an automated sequencer can be performed faster.
- **b.** The reactions performed in an automated sequencer can be read by a computer rather than a human.
- **c.** Higher temperatures are used during cycle sequencing, which prevent secondary structures from forming in the DNA and early termination of the reaction.
- **d.** In cycle sequencing, nonspecific interactions by the primer can be controlled by raising the annealing temperature.
- **e.** All of the above are advantages of cycle sequencing.

12. Which of the following statements about degenerate primers is not correct?
- **a.** Degenerate primers have a mixture of two or three bases at the wobble position in the codon.
- **b.** Because of the nature of degenerate primers, the annealing temperature during PCR using these primers must be lowered to account for the mismatches.
- **c.** Degenerate primers are often designed by working backwards from a known amino acid sequence.
- **d.** Degenerate primers are used even when the sequence of DNA is known.
- **e.** Within a population of degenerate primers, some will bind perfectly, some will bind with mismatches, and others will not bind.

13. Which of the following techniques would allow a researcher to determine the genetic relatedness between two samples of DNA?
- **a.** inverse PCR
- **b.** reverse transcriptase PCR
- **c.** TA cloning
- **d.** overlap PCR
- **e.** randomly amplified polymorphic DNA

(*Continued*)

14. Why would a researcher want to use RT-PCR?
- **a.** RT-PCR is used to compare two different samples of DNA for relatedness.
- **b.** RT-PCR creates an mRNA molecule from a known DNA sequence.
- **c.** RT-PCR generates a protein sequence from mRNA.
- **d.** RT-PCR generates a DNA molecule without the noncoding introns from eukaryotic mRNA.
- **e.** All of the above are applications for RT-PCR.

15. Which of the following is an application for PCR?
- **a.** site-directed mutagenesis
- **b.** creation of insertions, deletions, and fusions of different gene segments
- **c.** amplification of specific segments of DNA
- **d.** for cloning into vectors
- **e.** all of the above

16. In _____ sequencing, the DNA fragments are bound to a solid surface via a flow cell.
- **a.** Illumina
- **b.** 454
- **c.** chain termination
- **d.** Sanger
- **e.** cycle

17. Flashes of light are emitted whenever a base is added in _____ sequencing.
- **a.** Illumina
- **b.** 454
- **c.** chain termination
- **d.** Sanger
- **e.** cycle

Further Reading

Agarwal, K. L., Büchi, H., Caruthers, M. H., Gupta, N., Khorana, H. G., Kleppe, K., et al. (1970). Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast. *Nature*, *227*, 27–34.

Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., et al. (2008). Whole-genome sequencing and variant discovery in C. elegans. *Nature Methods*, *5*, 183–188.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics: TIG*, *24*, 133–141.

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, *470*, 198–203.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature Reviews. Genetics*, *11*, 31–46.

Taft-Benz, S. A., & Schaaper, R. M. (2004). The theta subunit of Escherichia coli DNA polymerase III: a role in stabilizing the epsilon proofreading subunit. *Journal of Bacteriology*, *186*, 2774–2780.

Yoo, B., Kavishwar, A., Ghosh, S. K., Barteneva, N., Yigit, M. V., Moore, A., et al. (2014). Detection of miRNA expression in intact cells using activatable sensor oligonucleotides. *Chemical Biology*, *21*, 199–204.