

Chapter 14

DNA cloning

FYI 14.1

CsCl gradients

In traditional preparations of highly purified plasmid DNA, the lysate containing the plasmid DNA was mixed with the dye, ethidium bromide, that slips between the bases of the double helix (intercalation) and unwinds the DNA. The amount of ethidium bromide that can intercalate depends on the topology of the DNA molecule. Covalently closed supercoiled molecules bind less ethidium bromide than linear DNA because the DNA is constrained. Linear molecules, such as broken chromosomes, can bind more ethidium bromide because they can unwind as the ethidium bromide binds. Linear molecules can be saturated with ethidium bromide to the point of approximately one molecule of ethidium bromide for every two base pairs of DNA. The ethidium saturated DNA is centrifuged in a cesium chloride gradient. The differential binding of the dye leads to different buoyant densities for chromosomal and plasmid DNA in the cesium chloride gradient. This allows the plasmid DNA to be separated from the chromosome. While ethidium bromide–cesium chloride gradients are time consuming, for extremely pure plasmid DNA, they are still very useful.

Cloning is the process of moving a gene from the chromosome it occurs in naturally to an autonomously replicating vector. In the cloning process, the DNA is removed from cells, manipulations of the DNA are carried out in a test-tube, and the DNA is subsequently put back into cells. Because *E. coli* is so well characterized, it is usually the cell of choice for manipulating DNA molecules. Once the appropriate combination of vector and cloned DNA or construct has been made in *E. coli*, the construct can be put into other cell types. This chapter is concerned with the details of the individual steps in the cloning process:

- 1 How is the DNA removed from the cells?
- 2 How is the DNA cut into pieces?
- 3 How are the pieces of DNA put back together?
- 4 How do we monitor each of these steps?

Isolating DNA from cells

Plasmid DNA isolation

The first step in cloning is to isolate a large amount of the vector and chromosomal DNAs. Isolation of plasmid DNA will be examined first. In the general scheme, cells containing the plasmid are grown to a high cell density, gently lysed, and the plasmid DNA is isolated and concentrated. When the cells are growing, the antibiotic corresponding to the antibiotic resistance determinant on the plasmid is included in the growth media. This ensures that the majority of cells contain plasmid DNA. Without the antibiotic selection, an unstable plasmid (i.e. one without a *par* function) can be lost from the cell population in a few generations.

Cells can be lysed by several different methods depending on the size of the plasmid molecule, the specific strain of *E. coli* the plasmid will be isolated from, and how the plasmid DNA will be purified. Most procedures use EDTA to chelate the Mg^{++} associated the outer membrane and destabilize the outer membrane. Lysozyme is added to digest the peptidoglycan and detergents are frequently used to solubilize the membranes. RNases are added to degrade the large amount of RNA found in actively growing *E. coli* cells. The RNase gains access to the RNA after the EDTA and lysozyme treatments. This mixture is centrifuged to pellet intact cells and large pieces of cell

debris. The supernatant contains a mixture of soluble cell components, including the plasmid, and is known as a lysate.

The methods used to purify the plasmid DNA from the cell lysate rely on the small size and abundance of the plasmid DNA relative to the chromosome, and the covalently closed circular nature of plasmid DNA. Most plasmids exist in the cytoplasm of the cell as circular DNA molecules that are highly supercoiled. The lysate is treated with sodium hydroxide to denature all of the DNA, and with detergent, SDS. The pH is then abruptly lowered, causing the SDS to precipitate and bring with it denatured chromosomal DNA, membrane fragments, and other cell debris. Most of the plasmid DNA renneals to form dsDNA because each strand is a covalently closed molecule and the two strands are not physically separated from each other. The small size of the plasmid allows the plasmid molecules to remain in suspension. The supernatant, which contains plasmid DNA, proteins, and other small molecules can be treated in a number of different ways to purify the plasmid. The most common protocol relies on a column resin that binds DNA. A small amount of the resin is mixed with the plasmid-containing supernatant and the plasmid-bound resin is collected in a small column. The remaining cell components are washed away and the plasmid is eluted from the resin. This procedure is quick, simple, and reliable and can be easily carried out on a large number of samples. Many modifications of this procedure have been devised.

Chromosomal DNA isolation

To isolate chromosomal DNA, cells are lysed in much the same way as for plasmid DNA isolation. The cell lysate is extracted with phenol or otherwise treated to remove all of the proteins. The chromosomal DNA is precipitated as long threads. The chromosomal DNA is very fragile and breaks easily. For these reasons, the chromosomal DNA is not usually purified using columns. Rather, the precipitated threads are collected by centrifugation.

Cutting DNA molecules

Once DNA has been purified, it must be cut into pieces before the chromosomal DNA and the plasmid DNA can be joined. The problem is to cut the DNA so that it will be easy to join the cut ends of the chromosomal DNA to the cut ends of the plasmid DNA. A group of enzymes, called **restriction enzymes**, are used for this purpose. Restriction enzymes are isolated from different bacterial species.

Bacteria use restriction enzymes and modification enzymes to identify their own DNA from any foreign DNA that enters their cytoplasm. The restriction part of the system is an enzyme that recognizes a specific DNA sequence or **restriction site** and cleaves the DNA by catalyzing breaks in specific phosphodiester bonds. The cleavage is on both strands of the DNA so that a double-stranded break is made. The modification part of the system is a protein that recognizes the same DNA sequence as the restriction enzyme. The modification enzyme methylates the DNA sequence so that the restriction enzyme no longer recognizes the sequence. Thus, the bacteria can protect its own DNA from the restriction enzyme. Any DNA that enters the bacteria and contains the unmethylated restriction site is cut and degraded. There are three types of restriction–modification systems (Table 14.1). The types are distinguished based on the

FYI 14.2

The discovery of restriction enzymes

Restriction enzymes were discovered in *E. coli* in the 1950s by scientists studying bacteriophage. Bacteriophage λ can be grown on an *E. coli* K12 strain and titered on *E. coli* K12 to determine the number of phage per milliliter. A high-titer phage lysate will contain approximately 10^{10} plaque-forming units per ml (pfu/ml). If this phage lysate is titered on an *E. coli* B strain, the titer will drop to 10^6 pfu/ml. One of the phage that forms a plaque on *E. coli* B can be used to make a high-titer lysate on *E. coli* B. The lysate grown on *E. coli* B will titer on *E. coli* B at approximately 10^{10} pfu/ml but will titer on *E. coli* K12 at 10^6 pfu/ml. This four-log drop in plating efficiency can be traced to genes encoded by the bacterial chromosome of each strain. The system is known as host restriction and modification. Host restriction is carried out by a restriction endonuclease and modification is carried out by the protein that modifies the restriction site.

Table 14.1 Characteristics of the three types of restriction–modification systems.

Type	No. proteins	Complex	Cofactors	Recognition sequence
I	3	yes	SAM ATP Mg ⁺⁺	3 bp (spacer) 4–5 bp Cleavage occurs 400–700 bp away
II	2	no	SAM	4 to 9 bp palindrome Cleavage is symmetrical within palindrome
III	2	yes	SAM ATP Mg ⁺⁺	Asymmetric 5–6 bp Cleavage on 3' side 25–27 bp away Cleavage on one strand only

SAM, S-adenosylmethionine.

number of proteins in the system, the cofactors for these proteins, and if the proteins form a complex.

Type I restriction–modification systems

Type I systems are the most intricate and very few of them have been described. Three different proteins form a complex that carries out both restriction and modification of the DNA. The complex must interact with a cofactor, S-adenosylmethionine, before it is capable of recognizing DNA. The S-adenosylmethionine is the methyl donor for the modification reaction and all known Type I systems methylate adenine residues on both strands of the DNA. The restriction reaction requires ATP and Mg⁺⁺ for cleavage of the DNA. The complex also has topoisomerase activity. The DNA sequence recognized by Type I enzymes is also complex. The sequence is asymmetric and split into two parts (Fig. 14.1a). The first part is a 3 bp sequence, next is a 6–8 bp spacer of nonspecific sequence, and finally there is a 4–5 bp sequence. Cleavage of the DNA occurs randomly, usually no closer than 400 bp from the recognition sequence and sometimes as far away as 7000 bp.

Type II restriction–modification systems

Type II systems are composed of two independent proteins. One protein is responsible for modifying the DNA and one for restricting the DNA. Modification of the DNA uses S-adenosylmethionine as the methyl donor. The Type II modification enzymes methylate the DNA at one of three places, with each specific modification enzyme methylating the same residue every time. The modifications that have been found are 5-methylcytosine, 4-methylcytosine, or 6-methyladenosine. The DNA sequence recognized by Type II restriction enzymes is symmetric and usually palindromic (Fig. 14.1b). The DNA sequence is between 4 and 8 bp in length, with most restriction enzymes recognizing 4 or 6 bp. Both the cleavage of the DNA and modification of the DNA occur symmetrically on both strands of the DNA within the recognition sequence. Restriction enzymes function as dimers of a single protein so that each protein monomer can interact with one strand of the DNA. Thus, both strands of the DNA are cleaved at the same time, generating a double-stranded break. Several

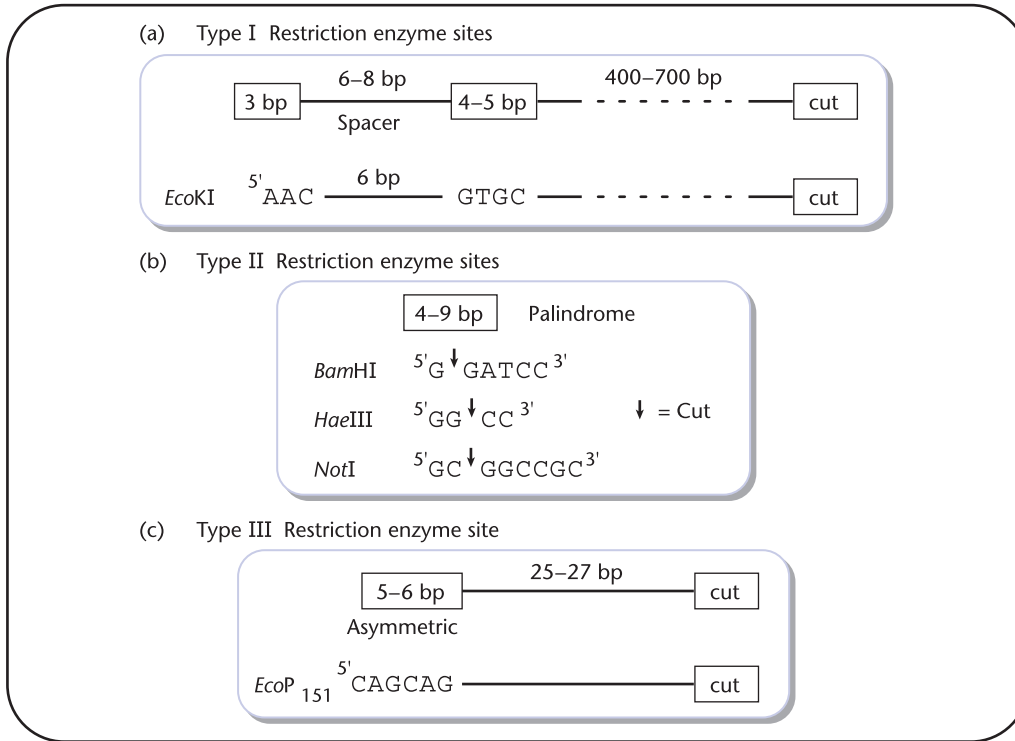


Fig. 14.1 The sequences recognized by restriction enzymes. (a) Type I restriction enzyme sites. (b) Type II restriction enzyme sites. (c) Type III restriction enzyme sites.

thousand Type II systems have been identified. Type II restriction enzymes are the most useful for cloning because they generate DNA molecules with a specific sequence on the ends (Fig. 14.2).

Type III restriction–modification systems

Type III systems are composed of two different proteins in a complex. The complex is responsible for both restriction and modification. Modification requires S-adenosylmethionine, is stimulated by ATP and Mg⁺⁺, and occurs as 6-methyladenine. Type III modification enzymes only modify one strand of the DNA helix. Restriction requires Mg⁺⁺ and is stimulated by ATP and S-adenosylmethionine. The recognition sites for Type III enzymes are asymmetric and 5–6 bp in length. The DNA is cleaved on the 3' side of the recognition sequence, 25–27 bp away from the recognition sequence. Type III restriction enzymes require two recognition sites in inverted orientation in order to cleave the DNA (Fig. 14.1c).

Restriction–modification as a molecular tool

Type II restriction enzymes have several useful properties that make them suitable for cutting DNA molecules into pieces for cloning experiments. First, most cloning ex-

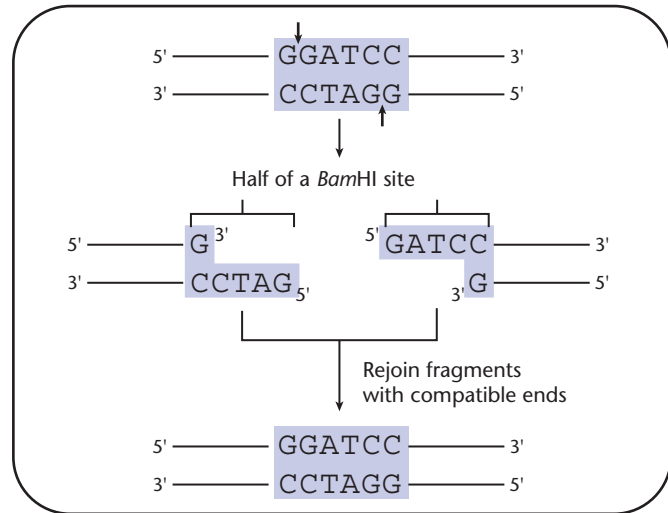


Fig. 14.2 Type II restriction enzymes generate DNA molecules with specific sequences on both ends. These ends can be rejoined to regenerate the restriction site.

FYI 14.3

Naming restriction enzymes

Restriction enzymes are named for the species and strain in which they are first identified. For example, *Bam*HI was the first enzyme found in *Bacillus amyloliquefaciens* H. *Cl*I was the first restriction enzyme found in *Caryophanon latum*. Because the enzymes are named after the species they come from, the first three letters in the restriction enzyme are always italicized. Some species encode more than one restriction enzyme in their genome. Hence, the names, *Dra*I and *Dra*III or *Dpn*I and *Dpn*II.

periments require manipulation of DNA molecules in a test-tube. The fact that the Type II restriction enzymes are a single polypeptide aids in the purification of the enzyme for in vitro work. Second, Type II restriction enzymes recognize and cleave DNA at a specific sequence. The cleavage is on both strands of the DNA and results in a double-stranded break. Cleavage of the DNA leaves one of three types of ends, depending upon the specific restriction enzyme (Fig. 14.3a). Some enzymes leave a **5' overhang**, some a **3' overhang**, and some leave **blunt ends**. The ends with either a 5' or 3' overhang are known as **sticky ends**. Any blunt end can be joined to any other

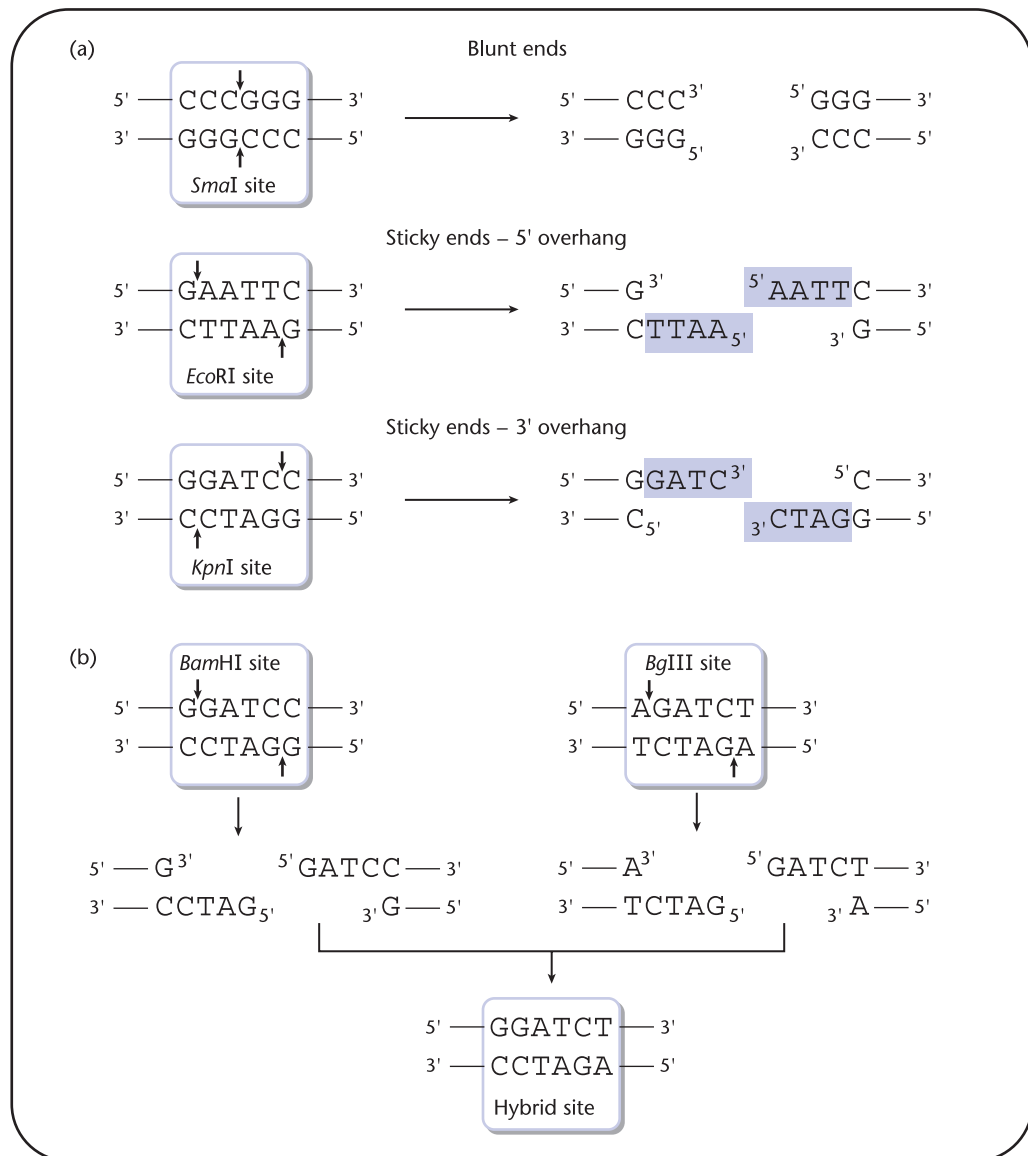


Fig. 14.3 Cleavage of DNA by a Type II restriction enzyme leaves one of three types of ends, depending on the enzyme used. (a) The ends generated can be blunt ends, sticky ends with a 5' overhang, or sticky ends with a 3' overhang. (b) The sticky ends from two molecules cut with two different restriction enzymes can be joined if the overhangs can hybridize. In this example, the hybrid site formed is no longer a substrate for either enzyme.

blunt end regardless of how the blunt end was generated. Sticky ends can be joined to other sticky ends, provided that either the same Type II restriction enzyme was used to generate both sticky ends that are to be joined or that the bases in the overhang are identical and have the correct overhang (Fig. 14.3b).

Generate double-stranded breaks in DNA by shearing the DNA

Usually restriction enzymes are used to cut the chromosomal DNA for cloning. The drawback of this approach is uncovered when the gene contains a restriction site for the enzyme being used to construct the library. One way to solve this problem is to only partially digest the chromosomal DNA with the restriction enzyme. Another way is to shear the chromosomal DNA and clone the randomly sheared DNA into a blunt end restriction enzyme site in the vector. One way to shear chromosomal DNA is by passing it quickly through a small needle attached to a syringe.

Joining DNA molecules

As described above, both plasmid and chromosomal DNA can be independently isolated from cells and digested with restriction enzymes. If, however, DNA with double-stranded ends is simply transformed back into *E. coli*, *E. coli* will degrade it. The double-stranded ends must be covalently attached. A version of this reaction is normally carried out in the cell by an enzyme known as **DNA ligase**. During DNA replication, the RNA primers are replaced by DNA (see Chapter 2). At the end of this process, there is a nick in the DNA that is sealed by DNA ligase. The double-stranded break formed by the restriction enzyme can be thought of as two nicks, each of which is a substrate for ligase.

If a plasmid molecule that has been digested with a restriction enzyme is subsequently treated with ligase, the plasmid molecule ends can be covalently closed by ligase (Fig. 14.4). Ligation is an energy-requiring reaction that occurs in three distinct steps. In the first step, the adenylyl group from ATP is covalently attached to ligase and inorganic phosphate is released. Next, the adenylyl group is transferred from ligase to the 5' phosphate of the DNA in the nick. Lastly, the phosphodiester bond is formed when the 3' OH in the nick attacks the activated 5' phosphate. AMP is released in the process.

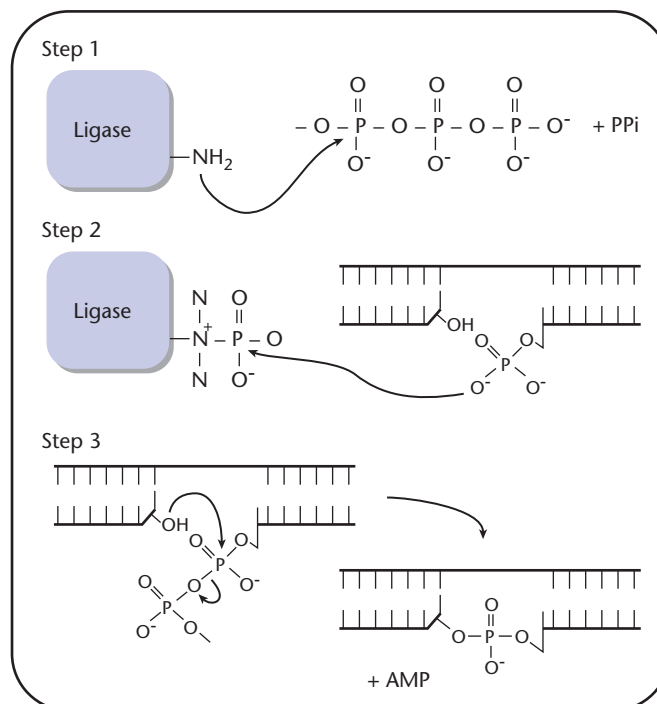


Fig. 14.4 The ends of DNA molecules can be joined and the phosphate backbone of the DNA reformed by an enzyme called DNA ligase. Ligase uses three steps to reform the backbone. (Step 1) A ligase molecule and an ATP molecule interact and the adenylyl group of ATP is covalently attached to the amine group of a specific lysine residue in the ligase protein. (Step 2) The adenylyl group is transferred to the 5' phosphate in the nicked DNA. (Step 3) The 3' OH attacks the activated 5' phosphate, reforming the backbone and releasing AMP.

Because of this mechanism, ligase requires both a 3' OH and a 5' phosphate. If the 5' phosphate is missing (see below), the nick cannot be sealed by ligase.

Manipulating the ends of molecules

While the goal of cloning is to construct a vector carrying the piece of DNA of interest, the reactions of the cloning process are concerned with the DNA ends. How the DNA ends are formed is very important because it dictates if the ends can be ligated to form the desired construct. The physical state of the DNA ends can be manipulated *in vitro* to influence the ligation reaction (Fig. 14.5a). For example, blunt ends cannot be ligated to sticky ends. The sticky ends can be made into blunt ends by the reaction of sev-

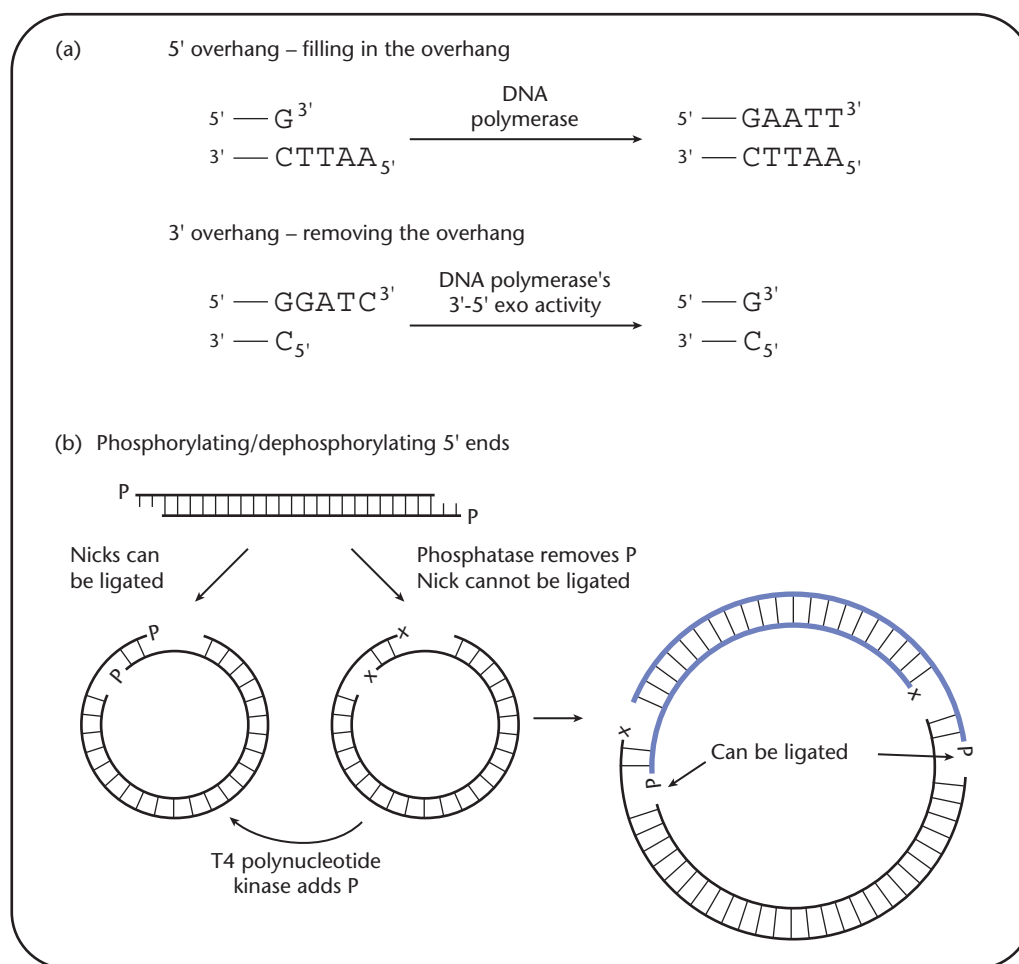


Fig. 14.5 The ends of DNA molecules can be manipulated *in vitro* to meet the requirements of the experiment. (a) 5' overhangs can be filled in by DNA polymerase and 3' overhangs can be removed by the 3' to 5' exonuclease activity of DNA polymerase. (b) 5' phosphates are required for ligase to function. If the 5' phosphate is missing, ligase cannot seal the nick. Phosphates can be removed by phosphatase and added by T4 polynucleotide kinase. Some cloning strategies take advantage of this by removing the 5' phosphates from the vector so that it cannot re-ligate without an insert. The insert allows two of the four nicks to be sealed. This molecule is stable enough to be transformed into cells where the other two nicks will be sealed.

eral different enzymes. If the sticky end contains a 5' overhang, then any one of several different DNA polymerases can be used to add the missing bases to the 3' OH using the 5' overhang as a template. A 3' overhang cannot be filled in, rather the overhang must be removed. Many DNA polymerases have a 3' to 5' exonuclease activity and this activity can be used to remove 3' overhangs.

The 5' phosphate can also be manipulated (Fig. 14.5b). If DNA molecules are missing the 5' phosphate, the phosphate can be added by an enzyme called T4 polynucleotide kinase. T4 polynucleotide kinase is an ATP-requiring enzyme that was originally identified in the bacteriophage T4. Molecules that have been phosphorylated by T4 polynucleotide kinase can be ligated to other molecules by ligase.

When pieces of chromosomal DNA are mixed with cut vector DNA, ligation of several different molecules can take place. The vector DNA ends can be ligated to reform the vector, the ends of a piece of chromosomal DNA can be ligated to each other, the ends of several vector or several chromosomal molecules can be ligated, or the ends of a piece of chromosomal DNA can be ligated to the ends of a piece of vector DNA. The ligation mix is usually put back into cells by transformation (see Chapter 11) and the antibiotic marker on the vector is selected for. Only cells transformed by molecules that contain vector DNA will form colonies. Of the molecules in the ligation mix, only religated vector DNA or vector DNA with a chromosomal insert are a possibility in the transformants. To reduce the number of vector molecules that are religated without a chromosomal insert, the 5' phosphates on the vector can be removed by an enzyme known as a **phosphatase**. The ends of vector molecules that have been dephosphorylated (the 5' phosphate has been removed) can only be ligated to chromosomal DNA molecules that have 5' phosphates (Fig. 14.5b). These molecules still have a nick on each strand but this nick can be sealed inside the cell.

Visualizing the cloning process

At each step of the cloning process, what is happening to the DNA molecules in the test-tube can be monitored using a technique called **gel electrophoresis**. In this technique, a gel (Fig. 14.6) containing small indentations or wells is cast. The DNA is loaded into the wells and the gel is placed in an electric current. Because DNA is negatively charged, it will move in the gel towards the positive pole. The DNA migrates or moves in the electric current based on size and shape. The larger a DNA molecule, the slower it moves. The more compact, or supercoiled a piece of DNA, the faster it moves.

The gel can be made from several different polymers, depending on the specifics of the experiment. Agarose forms a matrix that will separate DNA molecules from ~500 bp up to entire chromosomes (several million base pairs). If an electric current is constantly applied to an agarose gel from only one direction, agarose gels will separate DNA from ~500 bp to ~25,000 bp. If the direction and the timing of the current are varied over the electrophoresis time, then entire chromosomes can be separated in agarose. An alternative polymer, polyacrylamide, can be used to separate molecules a few base pairs in length to approximately 1000 bp.

Once the DNA has been separated in the gel, the gel is immersed in a solution containing ethidium bromide. If ultraviolet light is used to illuminate the gel, the ethidium bromide that is bound to the DNA will fluoresce, indicating the presence of bands of DNA (Fig. 14.6). Each band is composed of DNA molecules that are similar in size and shape. For example, when plasmid DNA is extracted from the cell, the majority of

FYI14.4

What is agarose?

Agarose is a polysaccharide composed of modified galactose residues that form long chains. These chains form the matrix that the DNA must move through. Smaller molecules pass through the matrix faster and larger molecules get caught up in the matrix. Agarose is particularly useful for making gels because it has a high gel strength at low concentrations of agarose. Practically, this means that the gels are strong enough to be handled even when the agarose is present at less than 1% (weight per volume). Agarose is isolated from algae such as seaweed. Different seaweeds have different modifications on the repeating galactose residues. The biological function of agarose is to protect the seaweed from drying out at low tide. Agarose has been used for many years as a stabilizer in the preparation of ice cream and other foods.

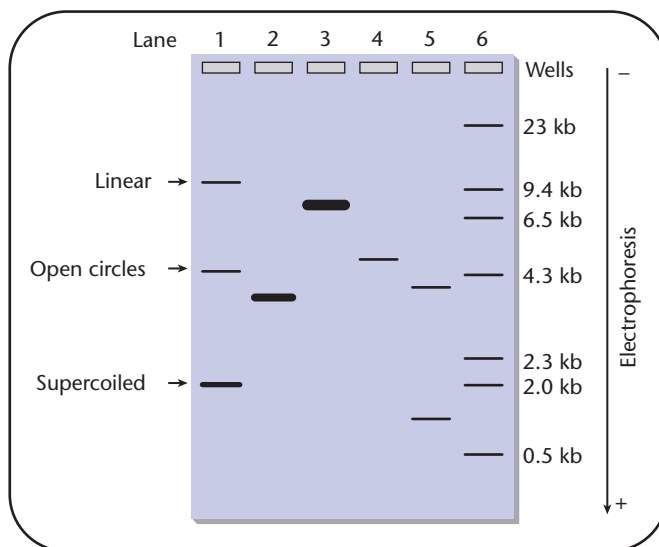


Fig. 14.6 A diagram of an agarose gel after electrophoresis and staining of the DNA with a fluorescent dye such as ethidium bromide. Lane 1 contains supercoiled, open circular and linear DNA forms of the same plasmid. Supercoiled DNA runs faster because of its topology. Lane 2 contains a 5 kb supercoiled plasmid and lane 3 contains a 10 kb supercoiled plasmid. Lanes 4 and 5 contain DNA fragments of different sizes. The band in lane 4 is larger than the bands in lane 5, meaning that the larger band contains more base pairs. Lane 6 contains DNA fragments of known molecular weights (molecular weight standards). Note that the molecular weight standards cannot be used to predict the sizes of supercoiled DNA.

it is supercoiled. Supercoiled DNA migrates very fast in an agarose gel (Fig. 14.6, lane 1). Some of the DNA will get nicked in the process of being extracted. The nick allows all of the supercoils to be removed, resulting in an open circle. Open circles migrate slower than supercoiled DNA. Some of the DNA will have a double-stranded break after isolation and the resulting molecules are linear. Linear DNA migrates the slowest of the three forms. If two different plasmids, one 5 kb and one 10 kb are isolated and run in an agarose gel, the supercoiled 5 kb plasmid will migrate faster than the 10 kb supercoiled plasmid (Fig. 14.6, lanes 2, 3). Likewise, the open circle 5 kb plasmid species migrates faster than open circle 10 kb plasmid species and the 5 kb linear species will migrate faster than the 10 kb linear species.

If a circular plasmid DNA is digested with a restriction enzyme that has two recognition sites in the plasmid, two linear pieces of DNA will result (Fig. 14.7). The shorter piece will migrate faster than the longer piece. If the DNA starts as a linear molecule and is digested with a restriction enzyme that recognizes the DNA in two places, then the DNA will be cut into three pieces (Fig. 14.7). Once all of the molecules are linear, they will migrate in the agarose gel based mainly on size.

Constructing libraries of clones

Sometimes it is desirable to make clones of all of the genes from an organism and subsequently to fish out the clone of interest. A large group of clones that contains all of the pieces of a chromosome on individual vector molecules is known as a library. To construct a library, vector DNA is isolated and digested with a restriction enzyme that

only cuts the vector once. Chromosomal DNA is isolated and digested with the same restriction enzyme. The cut chromosomal DNA and the cut vector DNA are mixed together and treated with ligase. This mixture is transformed into *E. coli* and plated on agar containing the antibiotic that corresponds to the antibiotic resistance determinant on the vector. *E. coli* cells that are capable of forming colonies must either contain the vector without a chromosomal DNA insert or a vector with a chromosomal DNA insert. If the correct ratio of chromosomal DNA to vector DNA is used (usually two chromosomal

molecules for every one vector molecule), rarely does a vector have two distinct pieces of chromosomal DNA inserted into it. Every clone should carry a unique piece of chromosomal DNA. If enough independent colonies are isolated, the entire sequence of the chromosome should be represented by the population of colonies. Libraries can be made from any kind of DNA, regardless of species.

How many independent clones are needed so that the library contains the entire genome? This depends on several factors and can be calculated using the formula:

$$P = 1 - (1 - F)^N \text{ or } N = \frac{\ln(1 - P)}{\ln(1 - F)}$$

where:

P = probability of any unique sequence being present in the library.

N = number of independent clones in the library.

F = fraction of the total genome in each clone (size of average insert/total genome size).

For example, a library with ~10 kb inserts is made from *E. coli*, which has a genome size of 4639 kb. To ensure that any given *E. coli* gene has a 99% probability of being on a clone in the library would require 2302 independent clones.

$$N = \frac{\ln(1 - 0.99)}{\ln(1 - 10/4639)} = 2302$$

DNA detection — Southern blotting

In 1975, E.M. Southern described a technique to detect sequence homology between two molecules, without determining the exact base sequence of the molecules (Fig.

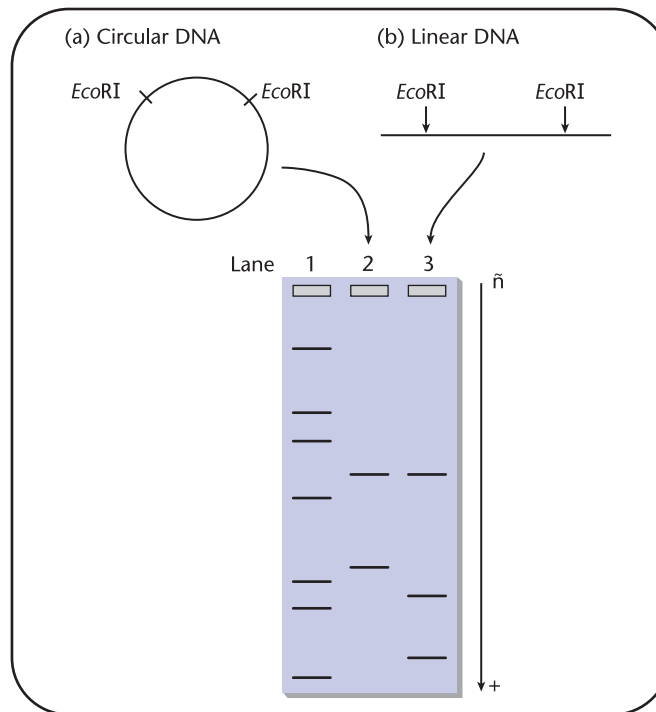


Fig. 14.7 The number of bands a molecule is cut into depends on if the starting molecule is linear or circular. Lane 1, molecular weight standards. Lane 2, bands from a cut circular molecule. Lane 3, bands form the same molecule as in lane 2 except the starting molecule was linear and not circular.

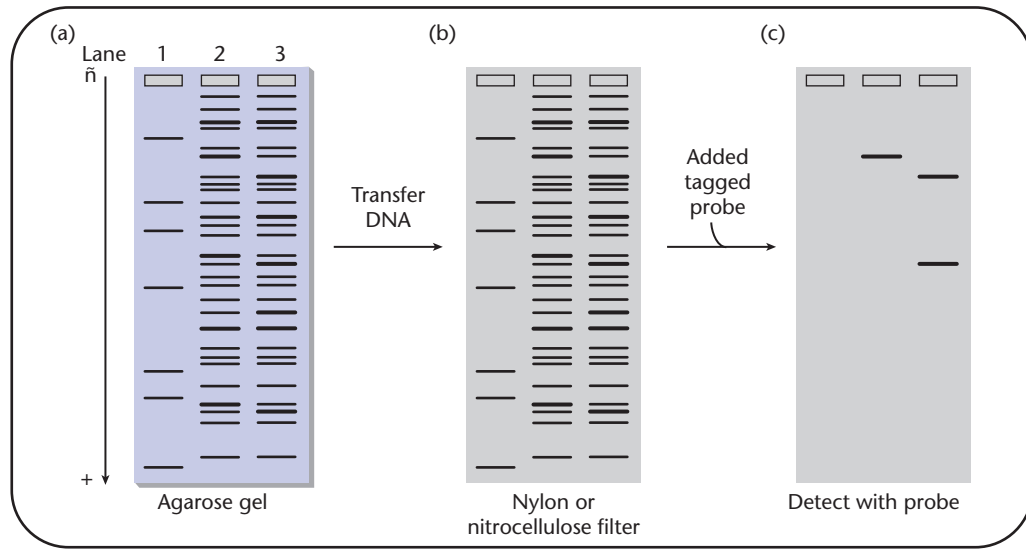


Fig. 14.8 The steps in a Southern blot. (a) DNA (usually chromosomal DNA) is digested with different restriction enzymes and run on a gel. The DNA is cut in many different places and leads to many fragments of different sizes. (b) The DNA is transferred from the gel to a nylon or nitrocellulose filter. (c) A fragment of DNA that contains the gene of interest is tagged and used as the probe. The probe is added to the membrane containing DNA and allowed to hybridize to any DNA fragment on the membrane to which it has homology. Excess probe is washed away and the probe is detected. If any fragments with homology are present, the size of the fragments can be determined based on where probe is detected.

14.8). The technique relies on fractionating the DNA on an agarose gel and denaturing the fractionated DNA in the agarose. The denatured DNA is transferred to a solid support, such as a nylon or nitrocellulose filter. A second DNA, called the **probe**, is labeled with a tag, denatured, and applied to the filter. Probes can be tagged with radioactivity and detected with X-ray film. They can also be labeled with fluorescent nucleotides or enzymes such as alkaline phosphatase or horseradish peroxidase. The enzymes are then detected with special substrate molecules that change color or emit light when cleaved by the enzyme. The probe will hybridize with any DNA on the filter that has complementary base sequences. Once the excess, non-hybridized probe is washed away, the tag attached to the probe can be detected.

Southern blotting can be used in many types of experiments. For example, if you have a clone of your favorite gene and you want to know if your gene exists in other species. You can isolate the chromosomal DNA from all of the species you want to test, digest the DNA with one or several restriction enzymes, and prepare a Southern blot. The clone of your favorite gene is used as the probe. If another species has sequence homology to your gene, then the band corresponding to the fragment containing the homology will be detected. The different species can be as diverse as bacteria, yeast, mice, rats, plants, humans, or as simple as several different types of bacteria. Blots with many different species of DNA included have become known as zoo blots!

A version of Southern blots can be used to identify clones from other species that are related to any gene of interest (Fig. 14.9). This technique is known as **colony blotting**. A population of cells containing a library from another species is plated on several agar plates and the cells are allowed to grow into colonies. The colonies are

transferred to either a nylon or nitrocellulose filter and lysed on the membrane. The DNA from the lysed colonies is attached to the membrane by either UV crosslinking for nylon or heat for nitrocellulose. A plasmid carrying the gene of interest is labeled with a tag and used as the probe. The labeled probe will hybridize with DNA from a lysed colony only if the colony carries a clone that contains complementary sequences. Once the appropriate clone is identified, that colony from the agar plate is purified and used as a source of the cloned gene of interest. For example, a cloned yeast gene can be used to fish out of a human library a related human clone.

DNA amplification—polymerase chain reaction

In 1993, the Nobel Prize in Chemistry was awarded for a novel and extremely important development called polymerase chain reaction (PCR, Fig. 14.10a). PCR allows almost any piece of DNA to be amplified *in vitro*. Normally DNA replication requires an RNA primer, a DNA template, and DNA polymerase. PCR uses the DNA template, two DNA primers, and a unique DNA polymerase isolated from a bacterium that grows at 70°C. The unique properties of this polymerase are that, unlike most DNA polymerases, it is capable of synthesizing DNA at 70°C and it is stable at even higher temperatures.

In PCR, many cycles of DNA synthesis are carried out and these cycles are staged by controlling the temperature that the reaction takes place. For example, the DNA template is double-stranded DNA and the two strands must be separated before any DNA synthesis can take place. The PCR reaction mix is first placed at 90–95°C to melt the DNA. The temperature is lowered to ~40–55°C to allow the primer to anneal to the single-stranded template. Finally, the temperature is raised to 70°C to allow the DNA to be synthesized from the primer. This cycle of melting the template, annealing the primers, and synthesizing the DNA is repeated between 25 and 35 times for each PCR reaction.



Fig. 14.9 A diagram of a colony blot. (a) The cells to be tested are plated on agar plates and incubated until they form colonies. (b) The colonies are transferred to a filter and lysed, releasing their DNA. The DNA is attached to the filter. (c) A tagged probe is added to the filters and the colonies that carry DNA that is homologous to the probe can be identified.

What happens to a template molecule in each cycle of DNA synthesis? In the first cycle, the two strands of the template separate, one primer anneals to each strand and two dsDNA molecules are produced (Fig. 14.10b). One strand of each dsDNA molecule is synthesized only from the primer to the end of the template. In the second cycle, all four strands are used as templates. Four dsDNA molecules are produced, two are similar to the dsDNA molecules produced in the first cycle and the other two have three out of the four DNA ends delineated by the primers. In the third cycle, the eight strands are used as templates and eight dsDNA molecules are produced. Two of the dsDNA molecules are similar to the products produced in cycle one, four of the

dsDNA molecules have three out of four ends delineated by the primers, and the other two molecules now have all four ends delineated by the primer. In the remaining cycles, the number of dsDNA molecules continues to increase linearly. In cycle 4, 8/16 molecules have all four ends delineated by the two primers, in cycle 5, 24/32, cycle 6, 56/64, cycle 7, 120/128, and cycle 8, 248/256. After 35 cycles, the dsDNA molecule with all four ends delineated by the two primers is the predominant molecule in the PCR reaction mix.

The DNA primers used in PCR are chosen so that the piece of DNA of interest is amplified. The DNA primers are synthesized in vitro. By carefully choosing primers, the exact base pairs at either end of the amplified fragment can be predetermined. The template DNA can be from any source. Only a small amount of template is needed. Once a fragment of DNA has been amplified, it can be cloned into an appropriate vector, used as a probe, restriction mapped, or used in a number of other techniques.

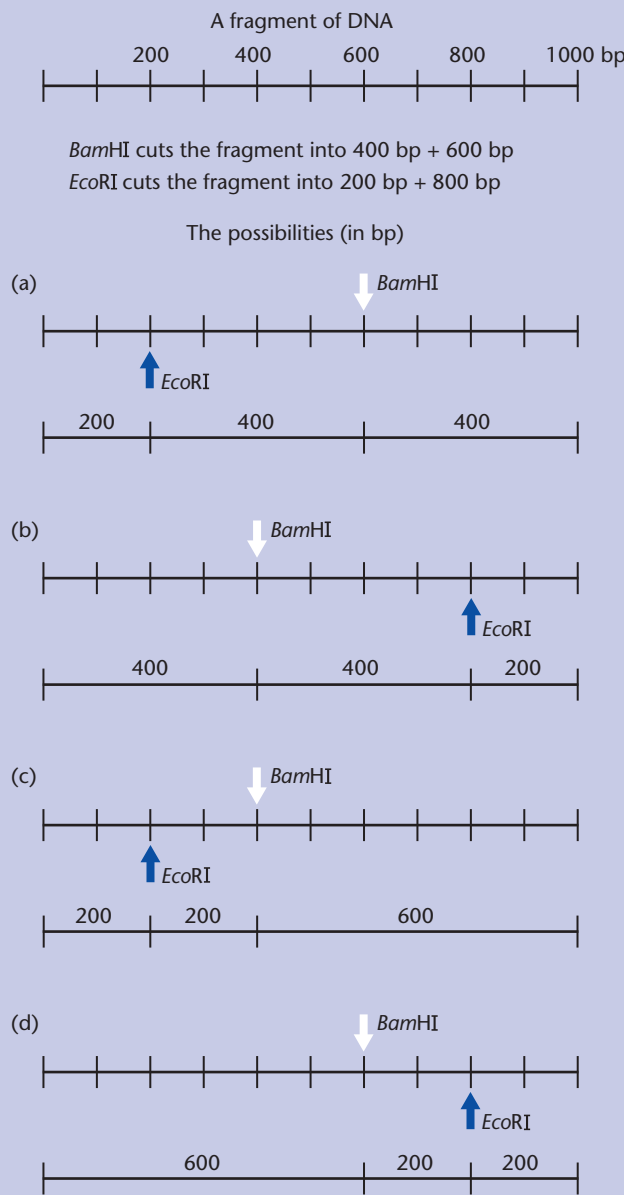
The DNA replication that

FYI 14.5

Restriction mapping

Different restriction enzymes recognize different DNA sequences. For example, *Bam*HI recognizes GGATCC and *Eco*RI recognizes GAATTC. If a linear DNA fragment is cut once by *Bam*HI, and once by *Eco*RI, what will happen if you cut the DNA fragment with both *Bam*HI and *Eco*RI? In this example, three fragments would be generated. The sizes of the individual fragments depend on where the enzymes cut in relation to each other. Determining where different enzymes cut is known as restriction mapping. The restriction maps in A and D can be distinguished from those in B and C. Digesting the fragment with both *Eco*RI and *Bam*HI will give the relative order of *Bam*HI and *Eco*RI sites.

FYI 14.5



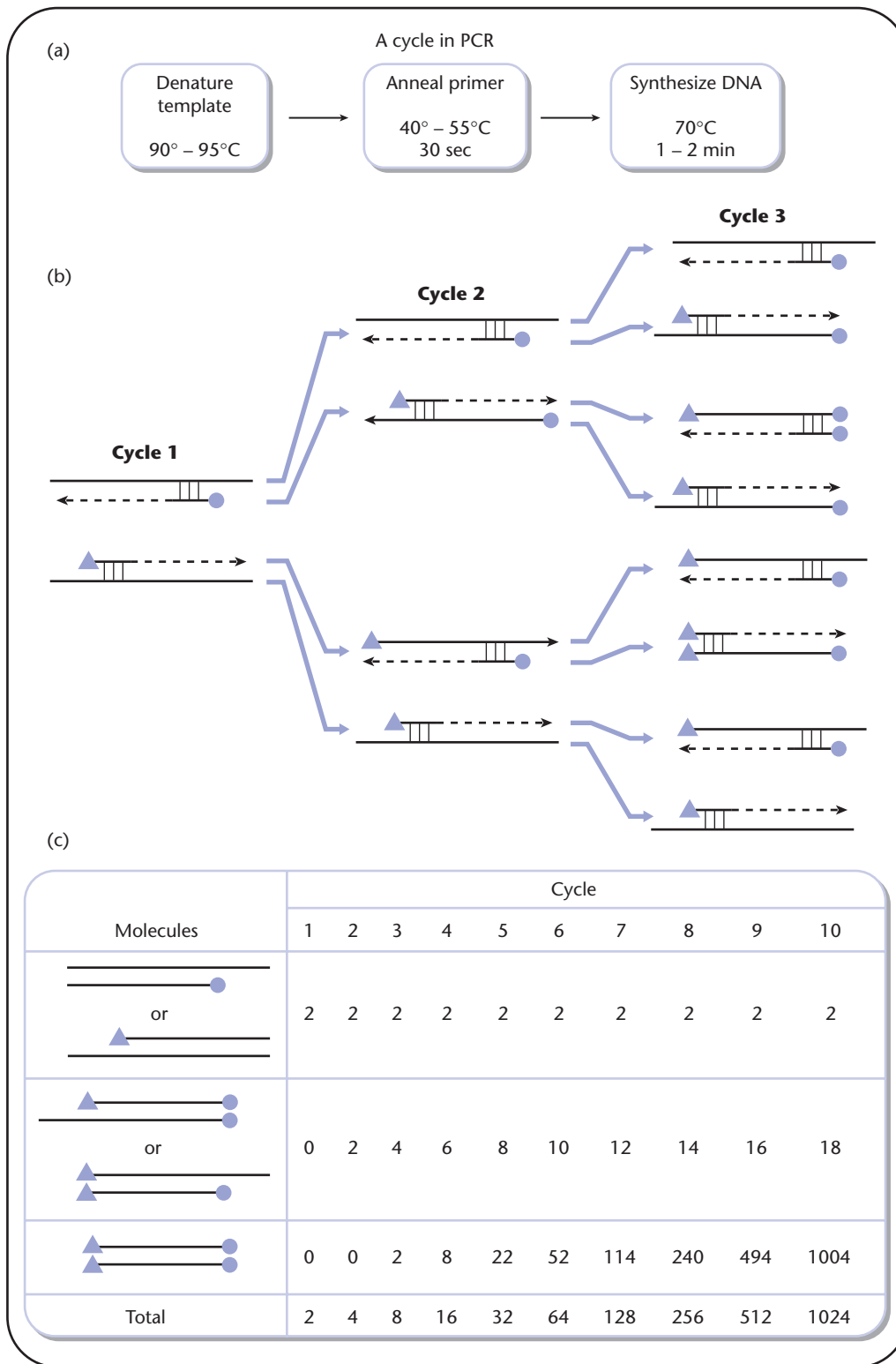


Fig. 14.10 A diagram of the polymerase chain reaction. (a) The general reaction taking place in each cycle. The template must be denatured, the primers annealed, and the DNA synthesized. (b) The fate of the template molecules in the first three cycles. The circle and triangle attached to the ends of the molecules represent the DNA ends that are delineated by the primers. (c) How many molecules of each type are formed in successive cycles. There will always be two molecules that match the ones formed in cycle 1. The number of molecules with three of the four ends matching the primer ends increases by two in each successive cycle. The number of molecules with all four ends determined by the primers is amplified dramatically. At the end of 35 cycles, >99% of the DNA molecules will have all four ends specified by the primers.

FYI 14.6**Amplifying DNA using PCR: functional uses of PCR**

At its most basic level, PCR is used to amplify a small amount of DNA into a large amount of DNA. The large amount of DNA is then analyzed by restriction mapping, cloning, sequencing, or some other technique. PCR is not limited to laboratory uses; it has had profound impacts on many other disciplines. PCR is used to amplify DNA from crime scenes. The small amounts of DNA found in unlikely places can now be analyzed to help prove guilt or innocence. DNA extracted from Egyptian mummies can be amplified by PCR. The relationships between individuals buried in the pyramids have been studied using PCR and restriction mapping. Many disease-causing microbes cannot be grown in the laboratory or take a long time to grow. Not knowing the microbes present in a sick person can prevent them from getting the proper course of treatment in time. PCR can identify the presence of specific bacterial species simply by using species-specific primers in a PCR reaction and a sample (sputum, blood, etc.) from the sick person. A PCR test can be conducted in a few hours or in time to drastically alter the course of treatment. In any discipline where the amount of DNA has traditionally been limiting, PCR can be used to circumvent that problem and help provide information.

takes place in PCR, like in vivo DNA replication, is not 100% accurate. Occasionally, a mistake is made. If the amplified fragment is to be cloned, the resulting clones must be sequenced to ensure that they carry a wild-type copy of the gene. If the amplified fragment is to be used as a probe, a few mutant copies in a mixture that contains a large number of wild-types copies will not present a problem. Thus, depending on the use of the PCR fragment, these contaminating mutant copies of the fragment must be accounted for.

Adding novel DNA sequences to the ends of a PCR amplified sequence

The primers used for PCR amplification can be synthesized with unique DNA sequences on the 5' ends (Fig. 14.11). These unique sequences can contain restriction enzymes sites, transcription factor binding sites, or any other sequence of interest. The unique sequences can be a few bases in length or up to 75–100 bases. If the added sequences are restriction enzyme sites, the resulting PCR fragment can be digested with the corresponding restriction enzyme for cloning into an appropriate vector. By predetermining the use of the PCR fragment, it can be customized in a variety of different ways.

Site-directed mutagenesis using PCR

PCR can be used to introduce a specific mutation into a specific base pair in a cloned

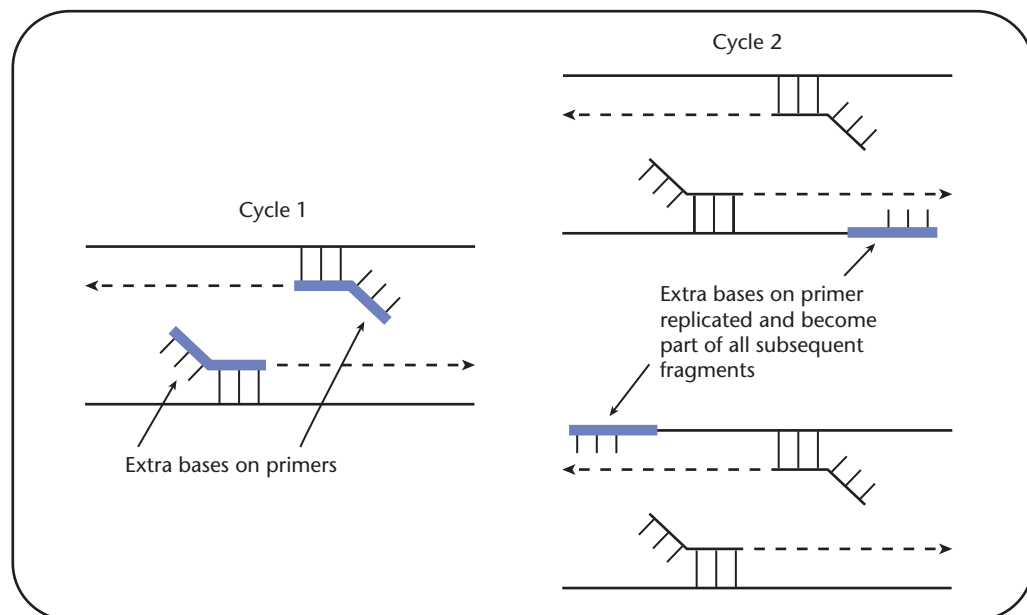


Fig. 14.11 PCR can be used to add specific sequences to the end of a DNA molecule. These specific sequences can be restriction enzymes sites or any other sequence of interest. In cycle 1, the added bases simply do not hybridize to anything. In cycle 2, the added bases are replicated, effectively making them a part of the PCR products. In subsequent cycles, the added sequences will be replicated as part of the DNA molecules.

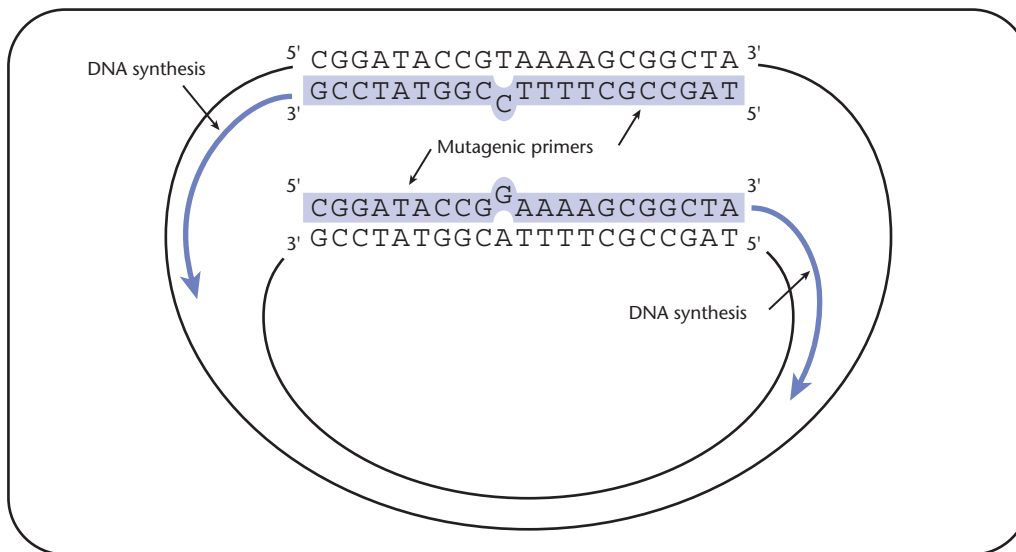


Fig. 14.12 One strategy for site-directed mutagenesis using PCR. Primers are designed to synthesize both strands of the plasmid containing the sequence to be mutagenized. Included in the primers are the base changes to be incorporated into the final mutant product. These primers are used in a PCR reaction to synthesize both strands of the plasmid with the incorporated change.

gene (Fig. 14.12). A primer is designed that contains the mutant base pair in place of the wild-type base pair. This mutant primer is used in a PCR reaction with the plasmid containing the cloned gene. PCR is used to synthesize the entire plasmid and after several rounds of replication, the large majority of plasmid molecules contain the mutation. This mixture of mutant and wild-type molecules is transformed into *E. coli* and the plasmids in individual colonies are tested for the presence of the mutation by determining the DNA sequence of the cloned gene.

Cloning and expressing a gene

In many cases, the goal is to not only clone the gene of interest but it is also to have the cloned gene expressed. When cloning a gene behind a promoter, the spacing of the elements needed for transcription and translation is critical. The closer the spacing of the elements to the optimum spacing for each element, the better the regulation and the better the expression of the cloned gene. The spacing can be manipulated during the cloning process. If a PCR fragment is used as a source of the gene to be cloned, base pairs can be inserted as needed by the design of the primers. For example, many vectors contain a promoter, mRNA start site, ribosome binding site (Shine-Dalgarno sequence), and ATG start codon (Fig. 14.13a). Each of these elements are in the correct order with the correct spacing. Following the ATG start codon are the multiple cloning sites (MCS). A gene is cloned into one of the restriction sites in the MCS. In this case, the reading frame must be maintained from the ATG start codon through the MCS and into the cloned gene (Fig. 14.13b). If the cloned gene is out of frame, this can be corrected by changing the primers used to amplify the fragment (Fig. 14.14).

Fig. 14.13 Cloning vectors can incorporate many different features including transcription and translation signals. (a) In this example, the cloning vector pBAD24 is shown. It contains a regulated promoter, mRNA start site, ribosome binding site, and a start codon. The multiple cloning sites are located after the start codon. (b) To clone into this vector and have your protein of interest be expressed from the regulated promoter, your DNA sequence must be cloned so that the reading frame of the gene is maintained.

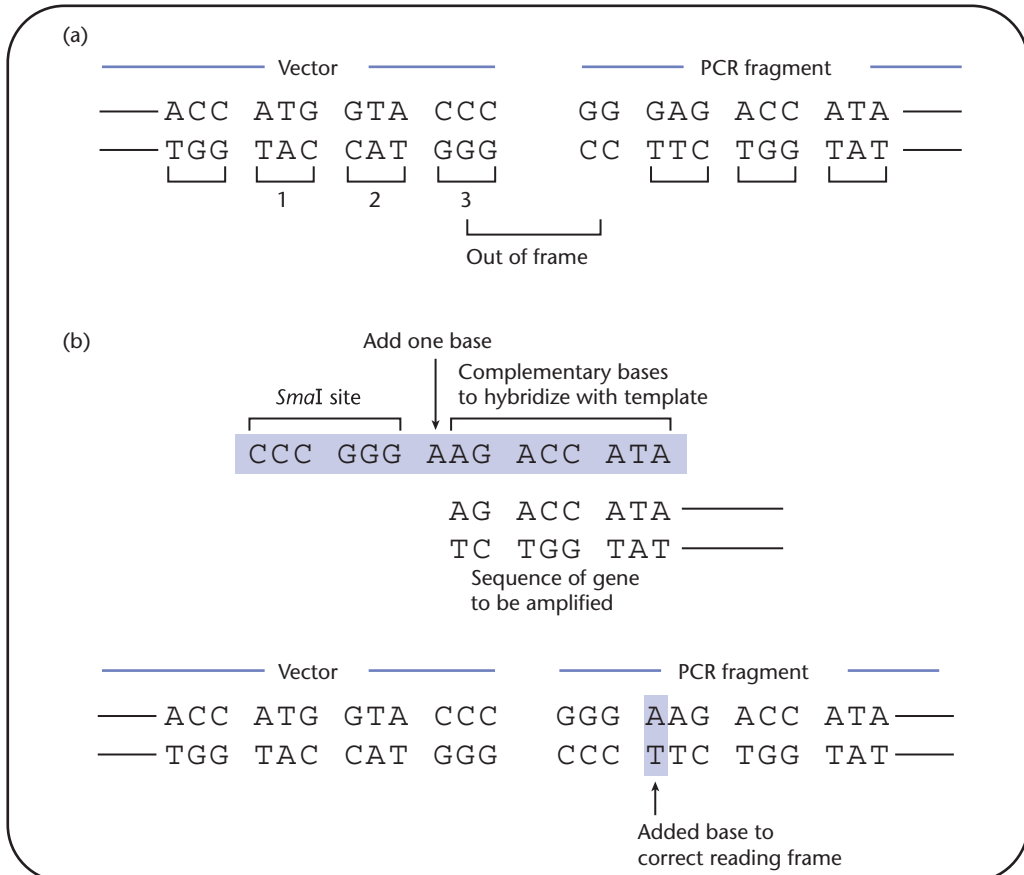
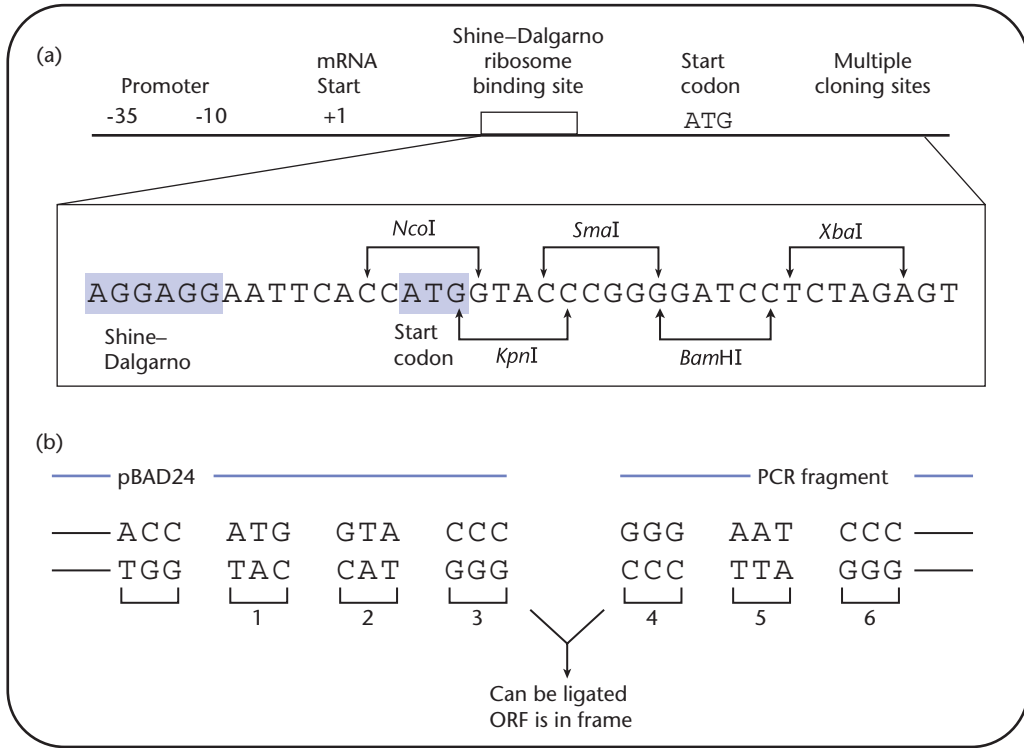


Fig. 14.14 If there are no restriction sites that leave the reading frame intact, then the fragment to be cloned can be amplified by PCR and the correct number of bases added to the primer.

DNA sequencing using dideoxy sequencing

DNA sequencing is the determination of the exact sequence of bases in a given DNA fragment. The start of the DNA sequence is determined by the placement of a DNA primer. Normal deoxyribonucleoside triphosphate precursors (dNTPs or dATP, dTTP, dCTP, and dGTP) and DNA polymerase are added to carry out DNA synthesis (Fig. 14.15a). In addition to the normal dNTPs, four special dideoxynucleosides (ddATP, ddTTP, ddCTP, and ddGTP) are included in small amounts. The special ddNTPs have a fluorescent tag attached to them. ddGTP has a fluorescent tag that fluoresces at one wavelength, ddATP a tag that fluoresces at a different wavelength,

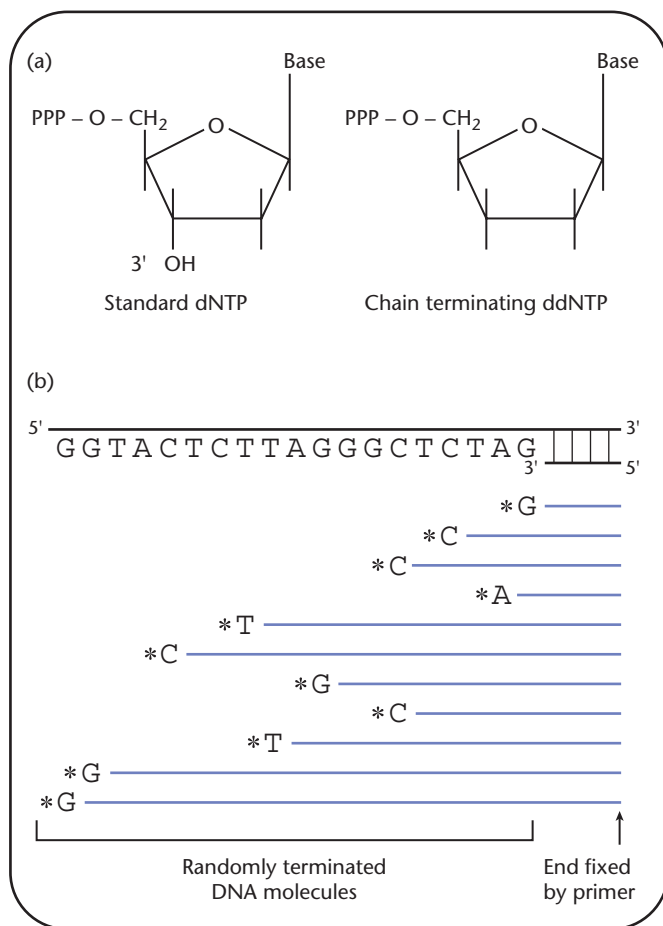


Fig. 14.15 Dideoxy DNA sequencing using fluorescently tagged ddNTPs. (a) A dideoxynucleoside triphosphate does not contain a 3' OH and as such terminates DNA replication. (b) The fluorescently tagged ddNTPs are randomly incorporated into growing DNA molecules. This leads to a collection of molecules and in this population are molecules that are only a few bases long all the way up to about 500 bases. The fragments will differ in size by only one base pair. The population is separated in a gel or column using an electric current. The smallest molecules migrate fastest and will reach the bottom of the gel or column first. A laser is positioned at the bottom of the gel or column and detects the fluorescent bases as they migrate through. The different bases are tagged with different colored dyes. The signals detected by the laser are relayed to a computer that records what color of dye was attached to each sized fragment.

ddTTP a tag with a third wavelength, and ddCTP a tag with a fourth wavelength. ddNTPs do not have a 3' OH and therefore block further synthesis of DNA. The fluorescently tagged ddNTPs are randomly incorporated into the growing DNA molecules (Fig. 14.15b). The results of DNA synthesis in the presence of tagged ddNTPs are a collection of DNA molecules different from each other by one base.

The fluorescently tagged ddNTP at the end of each molecule is dictated by the sequence of the template DNA. The tagged fragments are subsequently separated on either a polyacrylamide gel or on a very thin column. At the base of the column or polyacrylamide gel is located a laser. As the DNA fragments run off the column or gel, they pass through the laser beam, fluoresce, and the wavelength of the fluorescence is recorded and sent to a computer. The order of the fluorescently tagged molecules coming off the column reflects the sequence of the template DNA.

The automation of DNA sequencing has greatly simplified the process and made it much faster. Approximately 350 to 500bp of DNA sequence can be read from one



Fig. 14.16 Sequences that are very similar over the entire length of the protein indicate that the two proteins may have a similar

function. The greater the similarity, the greater the chance the functions are the same.

primer. By carrying out separate sequencing reactions using primers located every 350 to 450 bp on the template, the sequence of the entire template can be determined.

DNA sequence searches

To date, many millions of base pairs of DNA from many species have been sequenced. For example, the chromosomes of at least 50 bacterial species, several yeasts, and the large majority of all human chromosomes have been determined. These sequences contain an incredible amount of information. So much in fact that special computer programs had to be designed to help interpret just a fraction of the data.

When a DNA sequence is published in a scientific journal, it is also deposited in a computer database known as GenBank. When a sequence is placed in GenBank, the known and predicted features of the sequence are also indicated. These include promoters, open reading frames, and transcription factor binding sites. Just a listing of As, Cs, Gs, and Ts are known as a **raw sequence** and the sequence with all of the features indicated is known as an **annotated sequence**.

It is possible to search the sequences in GenBank using several different programs. You can search by the name of an interesting gene very easily. If you have the sequence of a gene of interest, it is possible to search for related sequences. GenBank can be searched using a DNA sequence or using that DNA sequence translated into the

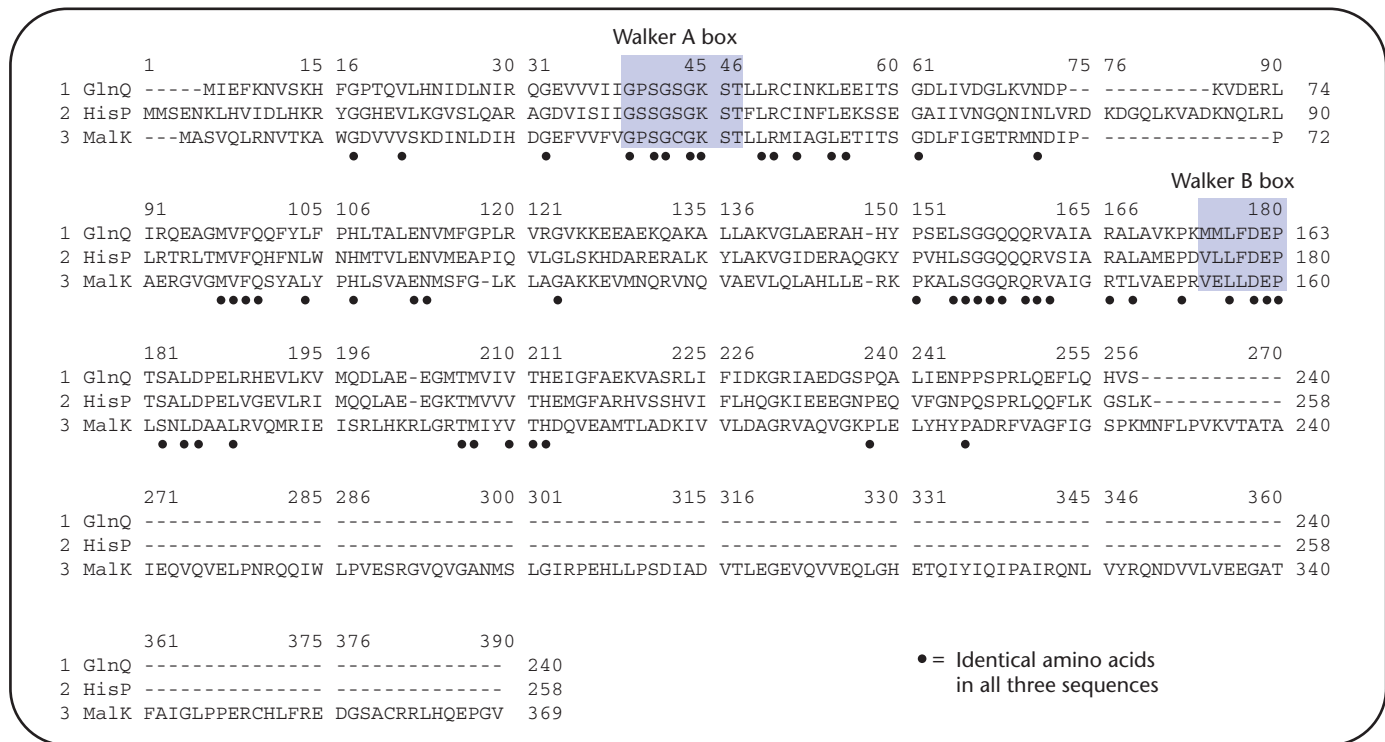


Fig. 14.17 Limited regions of homology can indicate a specific feature such as a nucleotide binding site or a metal binding domain. In this example, the highest regions of homology surround the

Walker A and Walker B boxes. These amino acids form a part of the protein that binds ATP.

protein sequence. The programs used for these searches are capable of identifying not only exact matches but also sequences that have differing degrees of similarity.

What can be learned from sequence searches? First, DNA sequence searches are more stringent than protein sequences. Two DNA sequences either have an adenine in the same position or they do not. Protein sequences can have the same amino acid in the same place and are, thus, **identical** at that position. Proteins can also have similar amino acids in one position, such as valine in one protein and alanine in the other. Because both amino acids are hydrophobic, they can frequently carry out the same functions. In this case, the proteins are said to be **similar** in a given position.

If two proteins have similarity over a large segment of their sequences, they may have similar functions (Fig. 14.16). This kind of analysis is especially useful if the function of one of the proteins has been identified. Knowing the function of one of the proteins suggests that the other protein should also be checked for this function. More limited regions of sequence similarity or identity can indicate the presence of a cofactor binding site. An example of this is the Walker box, which is an ATP binding site (Fig. 14.17). Sequence similarities can provide very valuable information about an unknown sequence and dramatically influence the direction of experiments on the novel gene or protein.

Summary

In 1962, the Nobel Prize in Medicine and Physiology was awarded to Watson and Crick for the discovery of the structure of DNA. The technology developed in the 40 years since has revolutionized how biological research is conducted. The ability to manipulate genes in vitro has greatly increased not only the

experiments that are now possible but also how scientists think about biological problems. Each of the techniques described in this chapter allows scientists to manipulate a novel gene in many different ways with the goal of uncovering its unique role in the cell.

Study questions

- 1 How is plasmid DNA purified away from chromosomal DNA?
- 2 How does ethidium bromide interact with DNA and how does it help in visualizing DNA on an agarose gel?
- 3 What are the distinguishing features between the three types of restriction enzymes?
- 4 Which class of enzymes is used for cloning and what characteristic(s) of this class make it suitable for the job?
- 5 Could you clone the gene for a restriction enzyme into *E. coli* in the absence of the modification enzyme? What would happen to the cellular DNA if you tried this experiment?
- 6 Can a blunt end be ligated to a sticky end?
- 7 Can a clone library be constructed from human DNA? Broccoli DNA? *E. coli* proteins? Human membranes? Why or why not?
- 8 What kind of gel would you use to separate a 100 bp DNA fragment from a 35 bp DNA fragment? A 750 bp fragment of DNA from a 6 kb fragment of DNA?
- 9 How would you determine if a novel bacterium contains a gene for degrading lactose?
- 10 In a PCR reaction, the reactions are cycled between 90–95°C, 40–55°C and 70°C. What biochemical reactions are occurring at each temperature?
- 11 You have the DNA sequence 5' TGCGCTAGGCTCATGGCCTTATAGACTCAGTCAAACGTCGTAGT 3' in your gene. Design an 18 bp primer to change the two Cs to As in the sequence: 5' GGCCTTA 3'.
- 12 Two protein sequences are 55% identical and 83% similar. What does this mean? Can two sequences have a greater identity than similarity?

Further reading

- Grunstein, M. and Hogness, D. 1975. Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. *Proceedings of the National Academy of Science USA*, **72**: 3961.
- Lobban, P. and Kaiser, A.D. 1973. Enzymatic end-to-end joining of DNA molecules. *Journal of Molecular Biology*, **78**: 453.
- Mertz, J. and Davies, R. 1972. Cleavage of DNA: RI restriction enzyme generates cohesive ends. *Proceedings of the National Academy of Science USA*, **69**: 3370.
- Sambrook, J., Fritsch, E.F., and Maniatis, T. 1989. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.