THREE

Production of Proteins in Bacteria and Yeast

The human body functions properly only when thousands of bioactive peptides and proteins – hormones, lymphokines, interferons, various enzymes – are produced in precisely regulated amounts, and serious diseases result whenever any of these macromolecules are in short supply. Until 1982, however, the only available pharmaceutical preparations of these peptides and proteins for the treatment of such diseases were obtained from animal sources, and they were sometimes prohibitively expensive. Bioactive proteins and peptides typically occur at low concentrations in animal tissues, so it was difficult to purify significant amounts for medical use. Some important proteins, such as pituitary growth hormone, differ in animals and humans to the extent that a preparation of animal origin is useless for treating humans. Finally, it was extremely difficult to isolate labile macromolecules from human and animal tissues without running some risk that the products might be contaminated by viral particles and viral nucleic acids.

The introduction of recombinant DNA techniques brought about a revolution in the production of these compounds (Chapter 2). It is now possible to clone a DNA segment coding for a protein and introduce the cloned fragment into a suitable microorganism, such as *Escherichia coli* or the yeast *Saccharomyces cerevisiae*. The "engineered" microorganism then works as a living factory, producing very large amounts of rare peptides and proteins from the inexpensive ingredients of the culture medium. And with such products obtained in this way from pure cultures of microorganisms, there is no chance of contamination by viruses harmful to humans.

PRODUCTION OF PROTEINS IN BACTERIA

For several reasons, bacteria were the first microorganisms to be chosen for use as living factories. To begin with, a great deal was known about their genetics, physiology, and biochemistry. After *Homo sapiens*, the bacterium *E. coli* is the most thoroughly studied and best-understood organism in the living world. Furthermore, it is easy to culture bacteria in large amounts in inexpensive media, and bacteria can multiply very rapidly. For example, *E. coli* doubles its mass every 20 minutes or so in a rich medium. Finally, bacteria are so small that up to a billion cells can fit on a single Petri dish only 10 cm in diameter. This permits us to test very large populations in order to find extremely rare mutants or recombinants – an enormous help at many stages of genetic and recombinant DNA manipulations.

INTRODUCTION OF DNA INTO BACTERIA

The field of bacterial genetics grew explosively in the mid-twentieth century, laying much of the groundwork for the development of procedures that efficiently introduce foreign DNA into bacteria. The three basic approaches take advantage of the three modes by which bacteria are known to exchange genetic information. There are two aspects of a genetic exchange: DNA (1) leaves a donor cell and (2) enters a recipient cell. It is the latter process, the uptake of DNA by a cell, that is all-important to biotechnologists.

Direct Introduction by Transformation

Transformation was the first process of genetic exchange to be discovered in bacteria. In 1928, Frederick Griffith injected living cells of *noncapsulated* pneumococcus (*Streptococcus pneumoniae*) together with heat-killed cells of a *capsulated* pneumococcus strain into mice and found that the noncapsulated strain then acquired, presumably from the capsulated strain, the ability to produce a capsule. These experiments thus showed that genetic information can be transferred into living bacterial cells from a preparation containing no living donor cells. In 1944, the substance that carried the genetic information in the transformation process was identified as DNA in the famous work of Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. This discovery led to the development of modern molecular biology.

We now know of several species of bacteria that, like pneumococcus, have a natural ability to undergo transformation, such as *Bacillus subtilis*, *Neisseria gonorrhoeae*, and *Haemophilus influenzae*. In some of these organisms, DNA is known to be taken up via elaborate machinery produced by the recipient cell, suggesting that the uptake is an active process. The ability to take up DNA, which is called *competence*, is typically developed only under special conditions.

The genetics and physiology of naturally transformable species are not well known, however, with the exception perhaps of *Bacillus subtilis*. Thus it was fortunate for biotechnological applications that the best-studied bacterium, *E. coli*, was found to accept exogenous DNA in an artificial transformation process. In the classical process, *E. coli* cells are first converted into a competent state by resuspension in buffer solutions containing very high concentrations (typically 30 mM) of $CaCl_2$ at 0°C. The effect of Ca^{2+} on a membrane bilayer with a high content of acidic lipids is to "freeze" the hydrocarbon interior, presumably by binding tightly to the negatively charged head groups of the lipids. Because the outer membrane of Gram-negative bacteria such as *E. coli* (see Figure 1.3B) contains a large number of acidic groups (in the form of lipopolysaccharide [LPS]) at a very high density, this membrane becomes frozen and brittle, with cracks through which macromolecules, including DNA, can pass. After DNA is added to the suspension, the cells are heated to 42°C and then chilled. Under these conditions, cells have been found to take up pieces of DNA through the cytoplasmic membrane, but the molecular mechanisms of the process still remain obscure.

Transformation can be achieved by similar means in certain other bacteria, but there are many species for which this method does not work. One method that works with many organisms (also including *E. coli*) is *electroporation*. In this process, we apply short electrical pulses of very high voltage, which is believed to reorient asymmetric membrane components that carry charged groups, thus creating transient holes in the membrane. DNA fragments can then enter through these openings, either by spontaneous diffusion or driven by the electric charge.

Introduction by Conjugation

We have said that it is difficult to introduce DNA directly into certain species of bacteria. In such cases, taking an indirect route sometimes achieves the desired result. First, a piece of DNA is introduced into an organism (such as *E. coli*) that *can* receive DNA by transformation. This piece of DNA is then transferred from the *E. coli* into the species of interest by another form of genetic exchange in bacteria, conjugation.

The *conjugational transfer* of genes in bacteria was discovered by Joshua Lederberg and Edward L. Tatum in 1946. Subsequent work has shown it to be a unidirectional transfer from a cell containing a sex plasmid, or F-plasmid (for "fertility"), into a cell lacking that plasmid. The transfer of chromosomal genes by conjugation occurs only in rare donor cells, in which the sex plasmid has become integrated into the chromosome. A more frequent process, which occurs with nearly 100% efficiency, is the transfer of just the F-plasmid from a donor to a recipient (Figure 3.1). Conjugation requires that the donor and recipient cells join to form a stable pair connected, at least in the beginning, by a filamentous apparatus (sex pilus).

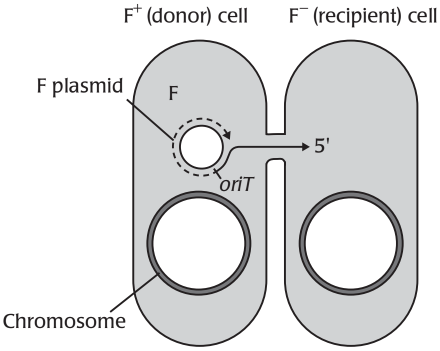


FIGURE 3.1

Conjugational transfer of the F-plasmid. One of the strands of the F-plasmid is cut at a specific position (oriT, for "origin of transfer"). This strand becomes elongated by rolling-circle replication (*broken line*), gradually displacing the old part of this strand, which enters into the F^- cell 5′-end first. A complementary strand is synthesized in the cytoplasm of the recipient cell, and the plasmid is then circularized, converting the recipient cell from F^- to F^+.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

As we shall see, the first step in the cloning of a fragment of DNA is to insert it into a suitable *vector DNA*, and plasmids are the most frequently used vectors. However, the unmodified sex plasmids are *not* used as vectors. If they were, the job of transferring the recombinant plasmids to other strains and species would be easy, because all the proteins needed for such a transfer are encoded on the plasmid itself. But the procedure could also be potentially dangerous, because if a plasmid-containing strain were to escape into the environment, the recombinant plasmid with the foreign DNA could conceivably start to spread into other, naturally occurring bacteria. The current practice, therefore, is to use as vectors only *nonconjugative* or *non–self-transferring* plasmids (plasmids that lack the information for the cell-to-cell transfer). For these plasmids to be transferred by conjugation, the missing information must be supplied from another plasmid. This procedure is called *plasmid mobilization*. It is useful when DNA must be transferred into strains that cannot be made to receive it at high efficiency by transformation.

Injection of Bacteriophage DNA and Transduction

A problem with the transformation process is its low efficiency. With *E. coli* as the recipient, the usual frequency of transformation suggests that only one out of hundreds of thousands of the exogenous DNA molecules enters the cell. In contrast, when bacteriophage (bacterial virus) infects bacterial cells, *every* virus particle adsorbs to a susceptible host cell and injects it with the DNA contained in the virus head at very high efficiency, often close to 100%. (The general features of the bacteriophage replication cycle are described in Figure 3.2.) Scientists have been able to take advantage of this natural process to inject foreign DNA into bacterial cells, thanks to a third type of genetic exchange in bacteria, transduction.

In *generalized transduction*, a piece of bacterial chromosome is transferred into a recipient cell by means of a bacteriophage. The chromosomal

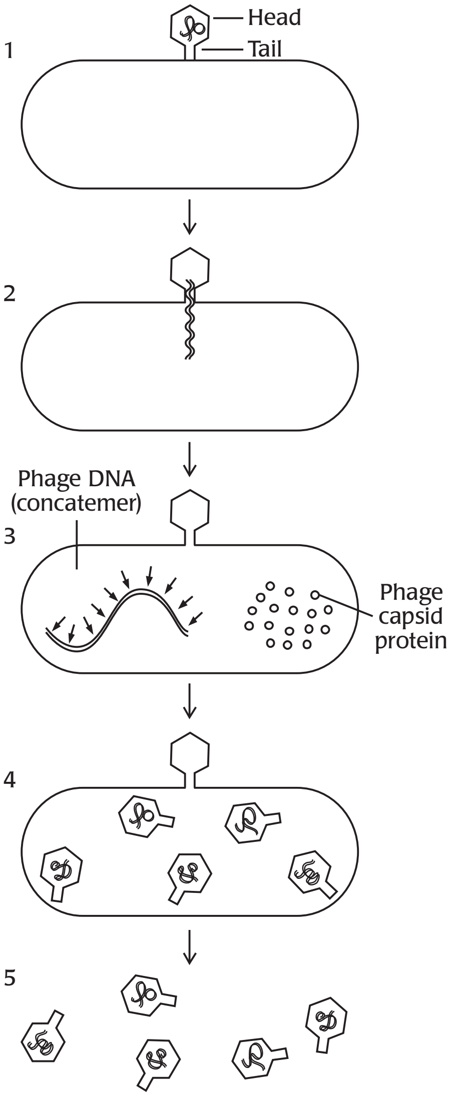


FIGURE 3.2

Multiplication of a virulent bacteriophage (bacterial virus) within a bacterial cell. The bacteriophage first adsorbs to a specific structure on the cell surface (step 1). The phage DNA is then injected into the cytoplasm, in some cases driven by contraction of the tail sheath (step 2). Within the cytoplasm, phage DNA and phage capsid (head as well as tail) proteins are synthesized separately (step 3). With most phages, DNA is synthesized as a concatemer containing many repeats of the genomic sequence. Finally, the DNA is cut to the length that corresponds to one phage genome (*arrows* in step 3) and becomes packaged into phage heads (step 4). The cell is then lysed (step 5). Thus, when a mixture of phages and a larger number of host bacterial cells is spread as a lawn on the surface of a solid medium, phages released by the lysis of one cell infect neighboring cells, causing cycles of lysis and infection and finally producing a small area of clearing (a plaque) where most of the host cells have lysed. This course of events occurs with *virulent* phages, which always cause lytic infections. With *temperate* phages, such as λ or P1, the infection may result in the *lysogenic response*, in which the phage DNA is replicated in step with the host genome without exhibiting the runaway replication of the lytic response. Temperate phages usually produce turbid plaques because some host cells within the plaques survive as lysogenic bacteria.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

DNA gets into the phage head by the mechanism illustrated in Figure 3.3. Once there, the fragment is injected into the cytoplasm of a new host cell in exactly the same way as the phage DNA. The phage head simply injects any DNA it happens to be carrying, regardless of the nature or the source of that DNA. Recombinant DNA technologies utilize this feature of the virus infection process by packaging recombinant DNA into phage heads *in vitro*. The specific vectors used for this type of delivery, phage λ and cosmids, are described in more detail below.

USE OF VECTORS

Let us assume that we have isolated a fragment of DNA coding for a commercially valuable protein and we want to convert *E. coli* into a factory that produces large amounts of this protein. Our first inclination might be to inject this piece of foreign DNA directly into *E. coli* cells by using one of the methods just described. Unfortunately, that approach would not work. A random piece of DNA floating in the cytoplasm would not be replicated. Only DNA that contains a special *replication origin* sequence is recognized and replicated by *E. coli*, and there is almost no chance that a fragment of foreign DNA will contain such a sequence. It is true that the foreign DNA fragment would be replicated if it got inserted into the bacterial chromosome and became a part of it – that is, if it became successfully "integrated" into the chromosome. (We rely on a similar process of integration when we introduce fragments of foreign DNA into higher plants and higher animals to create *transgenic* plants and animals.) In bacteria, however, the chromosomal integration of unrelated pieces of DNA is a rare event. Even if our fragment did become integrated into some part of the bacterial chromosome, the genes in the fragment would exist in the cell as single copies only, so they would not be expressed very strongly. Furthermore, the large size of the chromosome would prevent us from manipulating the fragment further – for example, by cutting it out for *subcloning*.

For these reasons, it is usually necessary to insert a cloned foreign gene into a vector – typically a plasmid or phage DNA that is much smaller in size than the bacterial chromosomes – that replicates autonomously in host microorganisms and acts as a carrier of the inserted foreign DNA sequence. There are hundreds of cloning vectors now available, each with its advantages and disadvantages. However, before we discuss the properties of each type of cloning vector, we must start by drawing a general picture of the cloning process itself.

Strategy for Shotgun Cloning

Say that we are going to clone, in *E. coli*, a gene *X* coding for a protein X from a "foreign" organism (i.e., an organism other than *E. coli*). The coding region of an average prokaryotic gene is only 1 or 2 kilobases (kb) long. In contrast, the genome of a bacterium has a length of thousands of kilobases, and that of a higher eukaryote a total length of millions of kilobases. Thus, gene *X* makes

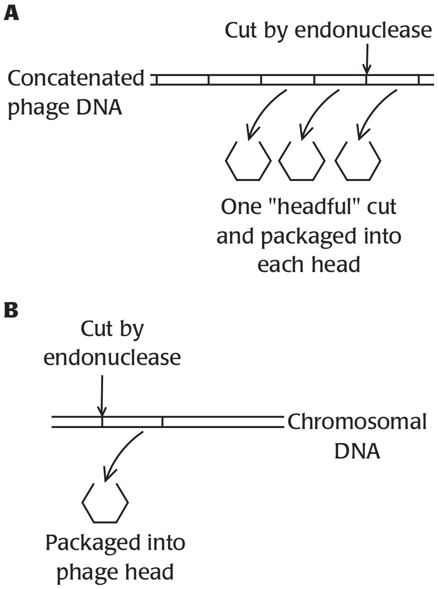


FIGURE 3.3

The generation of transducing phage particles. (**A**) In the normal infection cycle, the DNA of such phages as P22 is synthesized as a long concatemer, which is then cleaved by a phage-coded endonuclease at a specific site, the *pac* site. Then each newly assembled phage head becomes packed with a "headful" length of DNA. (**B**) If the chromosomal DNA is cut by an endonuclease – perhaps because it carries a sequence similar to that of the *pac* site, perhaps for other reasons – the headful packaging mechanism incorporates fragments of chromosomal DNA into newly assembled phage heads. In *generalized transduction*, these transducing particles subsequently inject such fragments of host DNA into other bacteria, and the DNA recombines with the chromosomes of the recipients to generate *transductants*.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

up only a small part (one in thousands to one in millions) of the genome. The usual first step in a cloning effort is therefore to clone random segments of the genome of the source organism (this is often called *shotgun cloning*) so that the subsequent isolation and identification of a clone containing the gene *X*, but not much else, will become possible (Figure 3.4). At this stage, it is advantageous to use vectors that can accommodate large DNA fragments because that dramatically decreases the number of recombinant DNA clones that must be examined in order to find the one containing the gene *X* (Box 3.1).

The large fragment cloned in this first step – the primary cloning – contains many genes in addition to gene *X*. Such complex pieces of DNA are not suitable for use in expression, sequencing, or site-directed mutagenesis. This is why it is necessary to pull out a small portion of the DNA, corresponding to only a little more than gene *X*. This essential step is called *subcloning*, and several different types of vectors are available for the purpose.

Genes from higher eukaryotes usually contain one or more intervening sequences, or *introns*, that do not code for the amino acid sequence of the protein product (Figure 3.5). As a rather extreme example, the gene for thyroglobulin has a size of 300 kb, but that includes 36 introns; the actual coding regions represent only 3% of the total gene length. When RNA transcripts are made from the DNA sequence, they still contain the sequences corresponding to introns. These sequences are then removed from the transcripts by *splicing*, and the mature mRNA molecules that leave the nucleus and enter the cytoplasm do not contain the intervening sequences. The mRNAs are also modified usually at the 3′ terminus by the addition of polyadenylate "tails" (see Figure 3.5).

To determine the nucleotide sequence of a particular gene (say for the purpose of identifying genetic defects in an inherited disease), it is necessary to clone the gene from the genomic DNA so that the intron sequences are included as well. This cloning of intron-containing genes

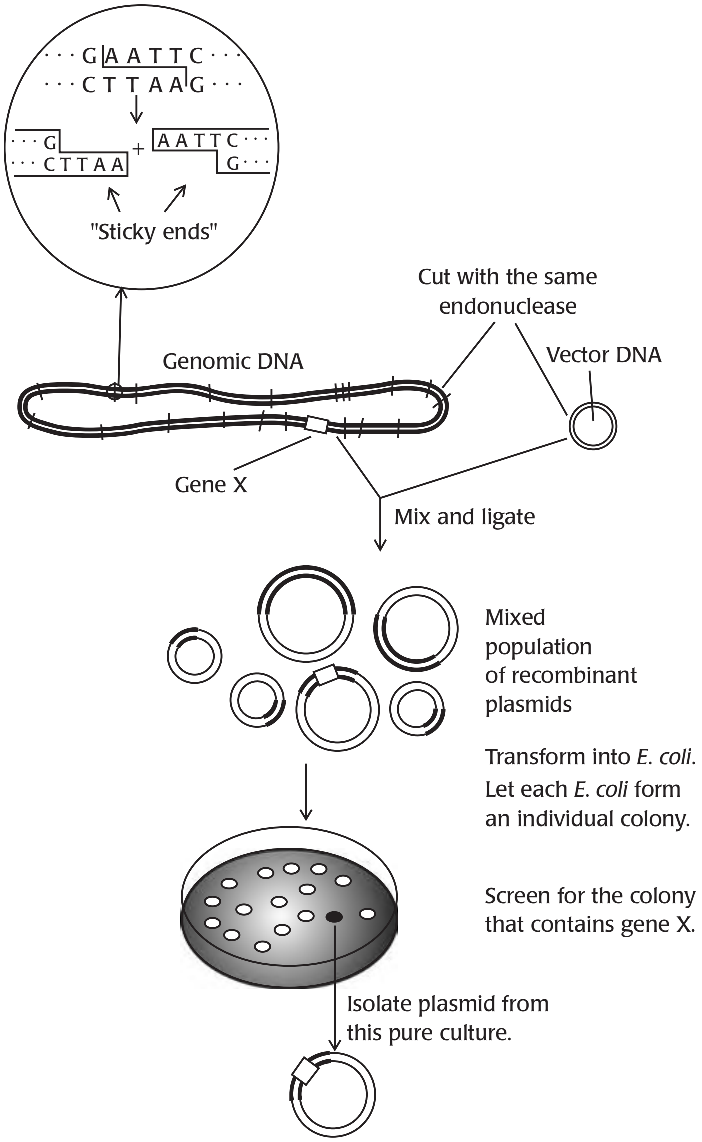

···G AATTC···
···CTTAAG···
···G
···CTTAA
+
AATTC···
G···
"Sticky ends"
Cut with the same endonuclease
Genomic DNA
Vector DNA
Gene X
Mix and ligate
Mixed population of recombinant plasmids
Transform into E. coli. Let each E. coli form an individual colony.
Screen for the colony that contains gene X.
Isolate plasmid from this pure culture.

FIGURE 3.4

Shotgun cloning of genomic DNA in *E. coli*. In the first step, one restriction endonuclease is used both to cut and open the vector plasmid DNA and to create fragments of genomic DNA. With most endonucleases, this procedure creates complementary "sticky ends" (see enlargement, here illustrating the ends created by restriction endonuclease EcoRI), which facilitate the end-to-end attachment of fragments by the complementary annealing of hanging protrusions. In the second step, the opened vector DNA is mixed with the fragments of the donor DNA. Many of the ends of the donor fragments will then anneal to the open ends of the vector DNA because of the complementary overhanging sequences. Addition of DNA ligase results in the covalent connection between the ends of DNA strands, producing a library of recombinant DNA. In the next step, the recombinant DNA pieces are introduced into *E. coli*, and the bacteria are spread on an agar plate containing a suitable growth medium so that each bacterium will produce a colony – a pure clone – well separated from other colonies. When the vector contains an antibiotic resistance gene, the antibiotic is added to the medium so that only those *E. coli* cells that have received the recombinant plasmid (or the resealed vector plasmid) will grow to produce colonies. Because transformation is a rare event, each clone will contain only one plasmid species. The colony containing the desired gene can then be identified by one of the methods discussed in the text. A pure preparation of the recombinant plasmid, amplified to billions of copies, can now be isolated from this *E. coli* strain, and the fragment can be "subcloned" further in different vectors for the purpose of expression, sequencing, or mutagenesis.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

requires specialized vectors, as described later in this chapter; luckily, for most biotechnological applications, it is also undesirable. Bacterial DNAs do not contain introns, and bacteria cannot carry out the splicing reactions. We will see later that even a eukaryotic microorganism such as yeast cannot be relied on to recognize all the splicing signals to be found in the RNA transcripts of genes of higher animals and plants. These eukaryotic genes, therefore, may not be expressed properly in microorganisms. Consequently, in these cases a better template for cloning is usually the mature mRNA, which does not contain the intervening sequences. In such a procedure, the mRNA is first converted to a double-stranded DNA through use of the enzyme reverse transcriptase, which was originally found as a product of RNA viruses (Figure 3.6). Because each eukaryotic mRNA usually contains coding information for only one protein, each of these DNAs, called cDNAs (for "complementary DNA"), also codes for one protein. For this reason, cDNA molecules can then be inserted directly into specialized vectors, such as expression vectors, often circumventing the need for subcloning. Importantly,

Fragment Size and the Probability of Finding a Desired Gene in a Set of DNA Fragments

Let us assume that we use a vector that can accommodate up to 40 kb DNA to clone fragments of a 4000-kb bacterial genome. We use a restriction endonuclease with rare recognition sites so that the genomic DNA is cut into about 100 distinct fragments with an average size of 40 kb. Among these, only one fragment (say, fragment 29) contains the gene *X*. So when we randomly examine clones to find the one containing gene *X*, how certain can we be of success? If we had the 100 fragments from the single chromosome of one bacterium in a box, then that set of 100 would certainly contain fragment 29. In actual practice, however, we will be using fragments generated from a mixture of many DNA molecules obtained from billions (or even more) of bacteria. Thus, when we pick just 100 fragments of these molecules (or 100 clones) at random, we are likely to have gathered multiple copies of some fragments and no copies of others (possibly including fragment 29). Statistical calculation shows that in order to have a probability P of finding the fragment containing *X*, one has to examine N clones, which is expressed by

$$\mathrm{N} = \ln(1 - P) / \ln(1 - R),$$

where R is the ratio of the fragment size (here 40 kb) to the genome size (4000 kb). If one wants a 99% probability ($\mathrm{P} = 0.99$) of fragment 29 being included in the collection, one has to examine 465 clones. This equation shows that if the size of the fragments cloned into vectors is 10 times smaller (4 kb), then the number of clones that must be examined increases to 4500. It is thus advantageous in a *primary cloning* (i.e., in the production of a "genomic library") to use a vector with a large insert size.

BOX 3.1

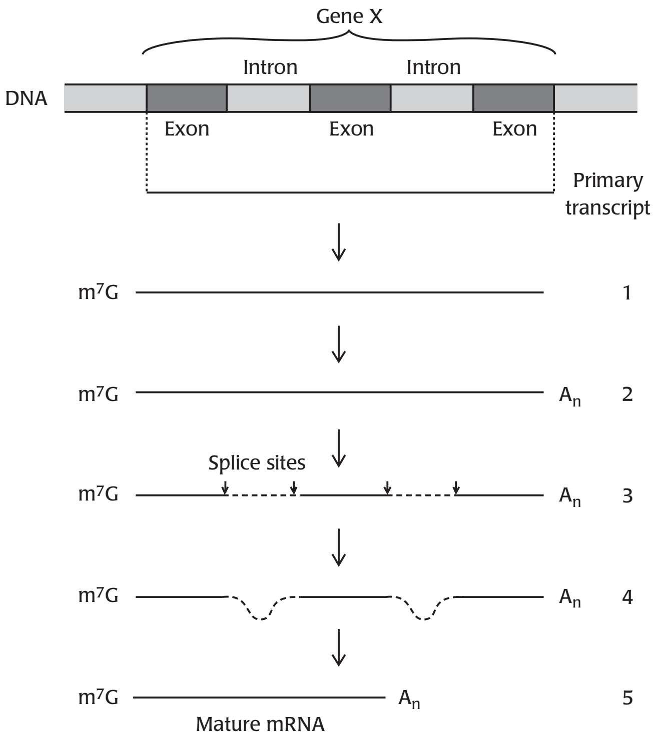


FIGURE 3.5

The processing of RNA transcripts in eukaryotes. A eukaryotic gene, especially one from a higher animal, is likely to contain many intervening sequences, or *introns*. Primary RNA transcripts of eukaryotic genes are processed first by "capping" – that is, by the addition of 7-methyl-guanosine monophosphate units at the $5'$-end through $5'$-$5'$ linkage – and by the shortening of the $3'$-end (stage 1). A polyA tail is then added to the $3'$-end (stage 2). Finally, the RNA sequences that correspond to the introns in the DNA are spliced out (stage 3), producing the mature mRNA.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

PCR-based amplification of individual genes, now made possible because of the presence of enormous amounts of information on gene sequences in thousands of organisms, also allows us to bypass the classical shotgun cloning method.

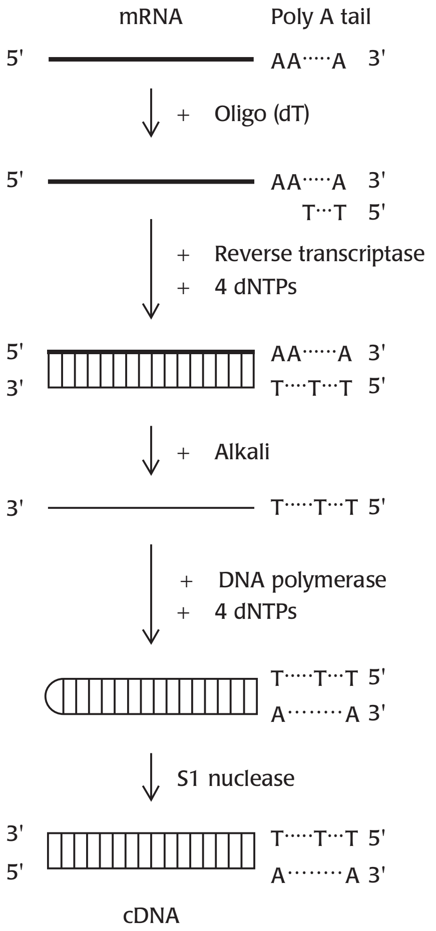


FIGURE 3.6

Production of cDNA from mRNA. With oligo(dT) as the primer, reverse transcriptase is used to synthesize a single strand of DNA. The template mRNA is then degraded with alkali, and DNA polymerase is used to synthesize a complementary DNA sequence on the first strand. Finally, treatment with S1 nuclease cuts the looped end of the DNA, generating a double-stranded cDNA.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

Cloning Vectors

Some cloning vectors are used only for general-purpose cloning, such as the primary cloning and identification of the coding segments. Plasmids are used most commonly for such purposes, although phage λ–derived vectors and cosmids are advantageous in situations that require the cloning of large segments of DNA. Vectors derived from single-stranded DNA phages are used for some special purposes. We discuss below some features of these vectors. Expression vectors, which are used for the high-level expression of cloned genes, are addressed later in this chapter (pages 115).

Plasmids. One of the first generation of plasmid vectors is pBR322 (Figure 3.7). It is still very frequently used, and many other plasmid vectors have been derived from it by the introduction of additional desirable properties. In the following description, we shall use pBR322 as an example and shall examine the various features that make it a good, general-purpose cloning vector.

The first important feature of this and any cloning vector is the presence of an origin of replication (*ori* in Figure 3.7), obtained for pBR322 from a naturally occurring colicin plasmid (Box 3.2). This origin of replication is recognized by the *E. coli* DNA replication machinery, which then initiates replication of the vector (and its foreign DNA inserts). A second feature of pBR322, and indeed of practically all the plasmid vectors, is the presence of an antibiotic resistance gene. In fact, pBR322 contains two such genes: *bla*, coding for *β*-lactamase, which degrades penicillins (including ampicillin) and cephalosporins and thereby produces resistance against these compounds, and *tet*, which codes for a membrane protein that acts as an exit pump for tetracycline, thus producing resistance to tetracycline and its relatives. These resistance markers are needed because, when plasmid DNA is introduced into *E. coli* cells by transformation, only one out of tens of thousands of cells receives a plasmid. Isolating this extremely rare cell would be practically impossible if there were no genetic markers to facilitate its selection (Box 3.3) out of the large excess of cells that failed to acquire the plasmid. Antibiotic resistance is an ideal positive selection marker, because all one has to do after transformation is to spread a large population of cells onto plates containing adequate concentrations of the antibiotics (Figure 3.8). The only cells to survive will be those that have acquired the plasmid, with its resistance genes.

The antibiotic resistance genes also serve a second purpose in pBR322. During the attempt to insert a piece of foreign DNA into a vector DNA that has been opened up by a restriction enzyme (see Figure 3.8), the vector DNA very often recircularizes (closes up again) without incorporating the foreign DNA. This is because unimolecular reactions, which are required for recircularization, occur much more frequently than the bimolecular reactions that are needed for the insertion of another piece of DNA. Reclosure of the vector DNA can be minimized by treating the opened vector with phosphatase (Figure 3.9), but it is difficult to prevent recircularization entirely. Thus, it is important to have a quick way of telling, from the phenotype of the transformed strains (transformants), whether the plasmids contain any inserted foreign DNA. Again, the resistance markers in pBR322 provide the needed information. For example, if one opens up the vector DNA by using BamHI or SalI restriction endonuclease (the cleavage sites for which lie within the *tet* tetracycline resistance gene), then the successful insertion of the cloned DNA will interrupt that gene and create transformants that are susceptible to tetracycline (see Figure 3.8). Screening for such transformants can be achieved conveniently by replica plating (Box 3.4). (By selecting for ampicillin resistance, we can still select for transformants that have successfully acquired plasmids.)

The third characteristic of pBR322 that makes it so useful as a cloning vector is that it contains only one site of cleavage for many commonly used restriction enzymes. (The precursor plasmid to pBR322 did contain multiple restriction sites for some of these enzymes, and the extra sites were eliminated.) This feature is found in most of the widely used cloning vectors and is very important. If the vector contained, say, three sites for the restriction enzyme EcoRI, religation of a mixture containing the three fragments produced from the vector and one fragment of foreign DNA will create many species of recombinant products (Figure 3.10). In contrast, with pBR322 containing a single EcoRI site, a large proportion of the product will be the desired recombinant plasmid, containing complete sequences of the vector and the foreign DNA (see Figure 3.10). Commonly, the foreign DNA is cut using the same restriction enzyme that is used in cutting the vector. Then all the ends of DNA will have the same hanging protrusions ("sticky ends"), which base-pair exactly with each other, increasing the chance of insertion of the foreign DNA (see Figure 3.9).

With plasmid vectors, specially constructed host strains of *E. coli* are often used. One feature of such strains is the defect in the restriction system (e.g., through mutations in the *hsdR* or *hsdS* gene), so that foreign DNA is not destroyed by the restriction enzyme of *E. coli*. Another feature is the defect in the homologous recombination system (e.g., through mutations in the *recA* gene), so as to prevent the alteration in the recombinant plasmids in the host strain. Examples of such strains are DH5α, HB101, and JM109.

λ Phage Vectors. As we have seen already, plasmids are convenient vectors. However, they are not ideal for every application. For example, when very large (>20 kb) pieces of DNA are inserted into the common plasmid vectors, it becomes difficult to introduce the large, recombinant plasmid into a host by transformation and to maintain such plasmids in successive generations of host cells. This is a problem when one wants to clone random fragments of genomic DNA in search of a particular gene, because the odds that the gene of interest will appear in any given fragment plummet when the average size of the cloned fragment decreases (see Box 3.1). The need to clone large fragments becomes especially acute when one is working

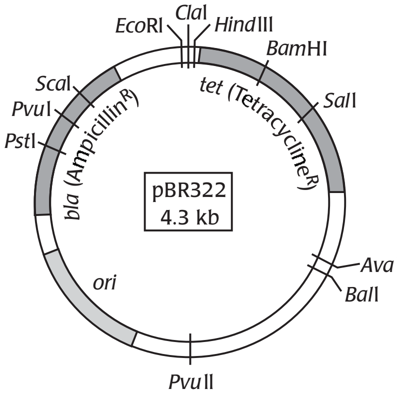


FIGURE 3.7

Structure of a plasmid vector, pBR322. Note that the vector has only a single susceptible site for each of the commonly used restriction endonucleases, such as EcoRI, BamHI, and so on.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

Colicin Plasmids

Many *E. coli* strains produce extracellular proteins, called colicins, that are able to kill a range of other bacteria. In most of these colicin-producing strains, the gene coding for the colicin protein is found on a plasmid (colicin plasmid), often along with genes that endow the colicin-producing strain with immunity against the colicin.

BOX 3.2

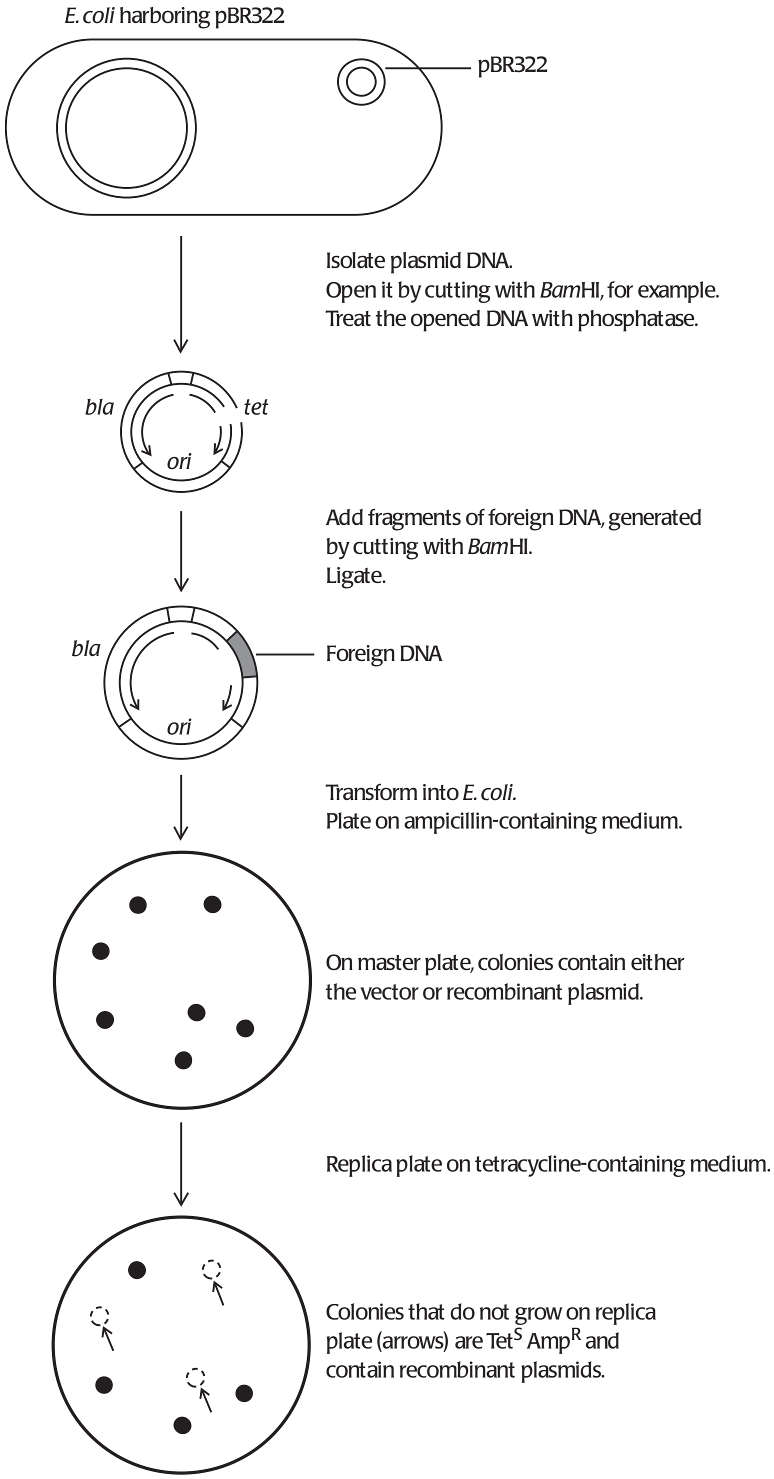
with the genomic DNA of higher animals and plants, because such eukaryotic genes are interrupted frequently by introns, and so only very large pieces of DNA can contain a complete gene. Some of the λ-derived vectors are more useful than plasmids for this type of situation.

Phage λ is a well-known temperate bacteriophage (Box 3.5) containing linear, double-stranded DNA. It was originally discovered in some strains of *E. coli* K-12, the standard strain used in bacterial genetics. The entire λ phage

Selection and Screening

In bacterial genetics, these terms have distinctly different meanings. A procedure in which a number of colonies are examined one at a time for a specific property – say, for their ability to hydrolyze certain substrates – is considered a *screening* (or sometimes scoring) procedure. In contrast, a procedure in which, out of many millions of initial population, only cells with certain properties are allowed to grow and form colonies is called a *selection* procedure. An example is the spreading of large numbers of bacteria on plates containing a particular antibiotic to select for the rare, antibiotic-resistant cells. Selection procedures are much more efficient than screening procedures because they enable us to examine much larger numbers of cells in a single experiment.

BOX 3.3



Isolate plasmid DNA.
Open it by cutting with *Bam*HI, for example.
Treat the opened DNA with phosphatase.

Add fragments of foreign DNA, generated by cutting with *Bam*HI.
Ligate.

Transform into *E. coli*.
Plate on ampicillin-containing medium.

On master plate, colonies contain either the vector or recombinant plasmid.

Replica plate on tetracycline-containing medium.

Colonies that do not grow on replica plate (arrows) are Tet^S Amp^R and contain recombinant plasmids.

FIGURE 3.8

Cloning of foreign DNA segments in pBR322. The vector DNA is cut open by a restriction endonuclease and then treated with phosphatase (see Figure 3.10) in order to prevent its religation. The addition of foreign DNA cut with the same restriction endonuclease results in the annealing of the foreign DNA to the complementary ends of the cut vector. After ligation and transformation into *E. coli*, the cells are plated on a suitable selective medium. In the example shown, the insert was cloned into the BamHI site, thus destroying the *tet* gene. The plasmid-containing cells were therefore selected on ampicillin-containing plates (by using their ampicillin-resistant – Amp^R – phenotype), and the presence of inserts in the plasmids was detected by the inability of certain colonies to grow on tetracycline-containing plates (by using their tetracycline-susceptible – Tet^S – phenotype). This screening can be conveniently accomplished by the replica-plating technique (see Box 3.4). When sites within *bla* genes (such as PstI or PvuI) are used in cloning, the tetracycline-resistant (Tet^R) cells that contain the recombinant plasmids are selected on tetracycline-containing plates, and the presence of inserts is scored on ampicillin-containing plates.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

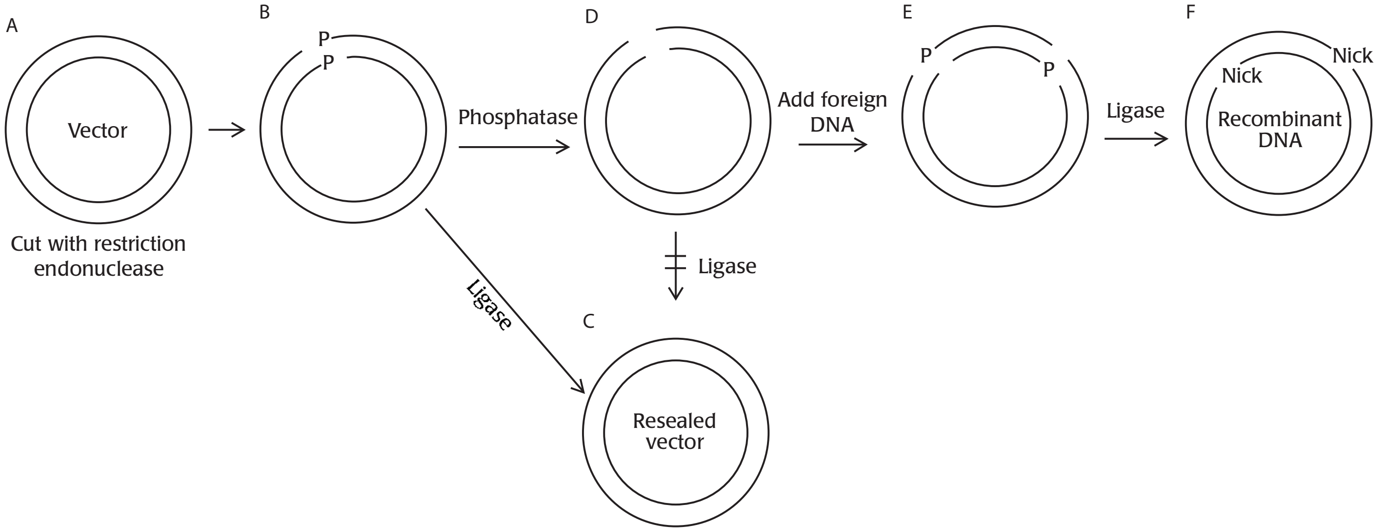


FIGURE 3.9

Preventing the religation of opened vector DNA. When the vector DNA (**A**) is opened by cutting with a restriction endonuclease, the open ends are usually staggered, with the 5′-phosphate groups still in place (**B**). These 5′-phosphate groups can react with 3′-OH ends of other DNA strands in the presence of DNA ligase, producing closed strands linked with phosphodiester bonds. Without further treatments, it is difficult to use this DNA in the construction of recombinant DNA, because ligation will cause much of the vector DNA simply to reseal (**C**). To prevent this, the opened vector DNA is treated with phosphatase. The treated vector DNA (**D**) cannot reseal on itself, because it lacks the 5′-phosphate groups needed for the formation of phosphodiester bonds. If foreign DNA cut with the same endonuclease is added, its staggered ends, with phosphate groups attached, become annealed with the staggered ends of the vector DNA (**E**). Finally, ligase connects the foreign DNA strands at the end containing the 5′-phosphate (**F**). Although the recombinant DNA created still contains nicks, these are readily repaired once it is transformed into the host cell.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

genome is 50 kb long, but an 8-kb region of this DNA (the *b*-region, Figure 3.11) has no known function. Another, adjoining region about 7 kb long and containing *att*, *int*, and *xis* (see Figure 3.11) is not needed for the lytic growth (see Box 3.5) of the phage. These two segments can be removed and replaced with a segment of foreign DNA without affecting the phage multiplication. In a λ-derived vector such as EMBL3 (see Figure 3.11), the insert can be significantly longer (up to 20 kb or even slightly more) than the length of these two deleted λ fragments (15 kb) for two reasons. (1) Two additional short segments (KH54 and *nin5*), totaling about 5 kb and representing regions not needed for lytic growth, were deleted to create EMBL3. (2) The head of the λ phage can package a piece of DNA that is slightly longer (by about 2 kb) than the length of the normal λ DNA.

As with most bacteriophages, λ phage particles are produced during the last stage of infection: The phage DNA is packaged into proteinaceous phage capsids that have been assembled in the cytoplasm of the infected cell (see Figure 3.2). We take advantage of this packaging reaction in using λ-derived cloning vectors. In practice, the fragment of foreign DNA is inserted into the vector DNA by cutting with restriction endonuclease, annealing the ends,

Replica Plating

This method, developed by Joshua and Esther Lederberg, permits the screening of many colonies in one operation. For example, if we want to screen a population of *E. coli* for their susceptibility to tetracycline, we first spread the population on an agar medium without tetracycline so that a few hundred colonies arise, after incubation, on a single plate (the master plate). The surface of the master plate is lightly "stamped" with a flat, sterile piece of velvet, and then the velvet is momentarily placed on the surface of a new plate that contains tetracycline (the replica plate). From each colony on the master plate, a few cells are transferred onto the replica plate by this operation. After incubation of the replica plate, colonies that exist on the master plate but do not develop at corresponding locations on the replica plate are noted: They correspond to tetracycline-susceptible clones.

BOX 3.4

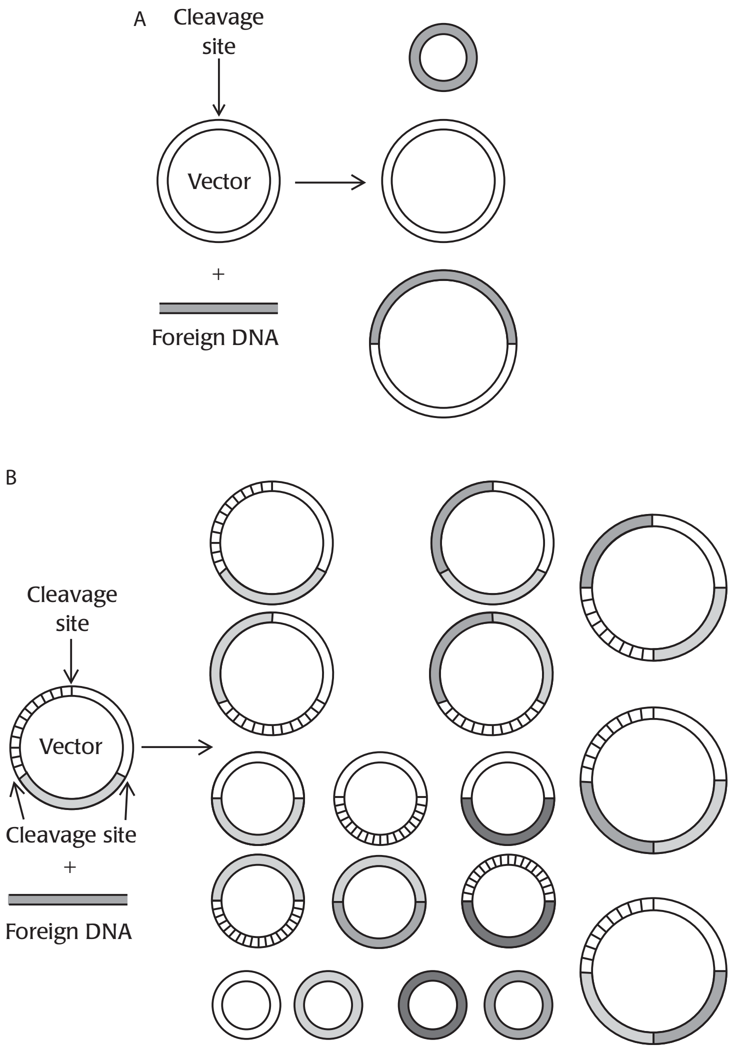


FIGURE 3.10

Vectors containing single or multiple cleavage sites for a restriction endonuclease. If a vector contains a single cleavage site for an endonuclease (**A**), then annealing and ligation with a segment of foreign DNA produce only three species of circular DNA, one of which is the desired recombinant containing the foreign DNA and the vector sequence. In contrast, if a vector is cut at three places by an endonuclease, annealing and ligation with foreign DNA produce many species of circular DNA (**B**), only a small fraction of which is the desired recombinant species. The situation is far worse in reality because for simplicity, the figure does not show the species in which multiple copies of one fragment are present within a single molecule. Clearly, it is a major disadvantage for a vector to have more than one cleavage site for each of the commonly used endonucleases.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

and then ligating with DNA ligase. When one mixes the recombinant DNA thus produced with a mixture of the proteins that form the phage capsids, the capsid is assembled and the DNA is packaged spontaneously into λ particles *in vitro*. After packaging, the new phages containing recombinant DNA are used to infect the host bacteria, a process in which DNA enters the bacteria with an efficiency of nearly 100% (rather than 0.001% or less, which is typical

Lytic and Lysogenic Responses in Phage Infection

Bacteriophages are classified as either virulent (such as T4 and T5) or temperate (such as P1, P22, and λ). When a bacterial host is infected by a *virulent* phage, a lytic response is inevitable: The phage multiplies extensively within the cell, which ultimately bursts (lyses) and dies. Infection by a *temperate* phage brings either a lytic or a lysogenic response. In the latter, replication of the phage genome is limited, and the phage genome continues to coexist within the host as a "prophage," either a separate, plasmidlike piece of DNA (as in the case of P1) or a part of the host chromosome (as in the case of phage λ). Many prophages can be "induced" to initiate a lytic cycle by inactivation of repressor proteins.

BOX 3.5

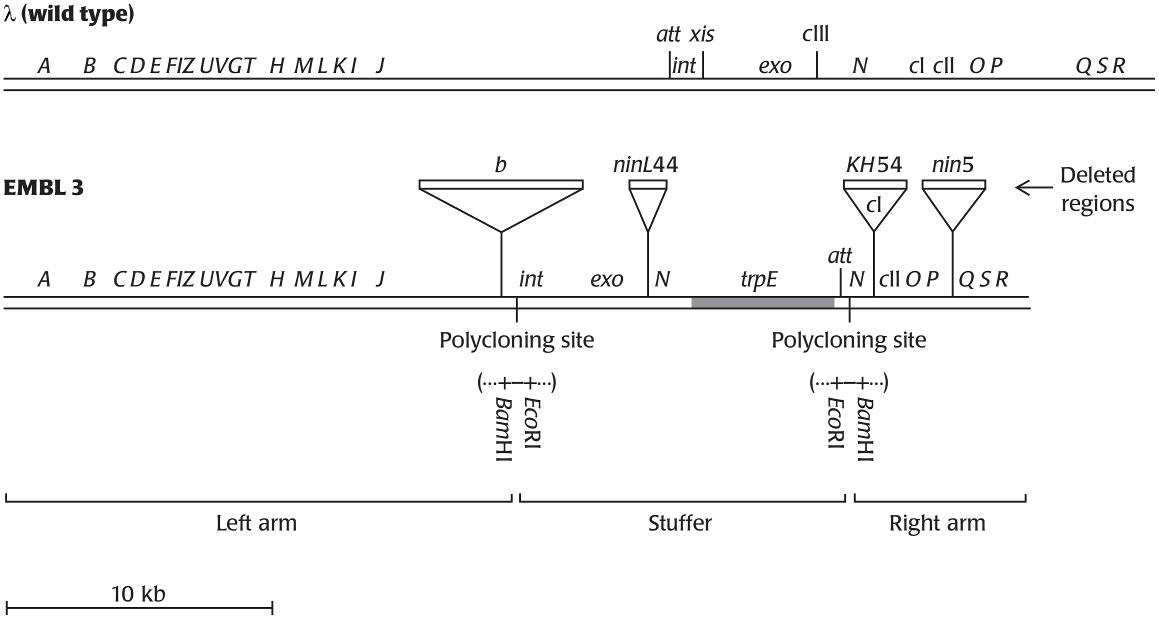


FIGURE 3.11

Phage λ and the EMBL3 vector. The λ-based vectors require that both the vector DNA itself and the recombinant DNA be packed efficiently into the λ phage heads. Packaging demands that the DNA have a length between 78% and 105% of the length of the normal λ phage DNA, so *replacement vectors* such as EMBL3 contain a *stuffer* segment, which is replaced by a foreign DNA segment of the same or somewhat larger size in the recombinant DNA constructs. More specifically, to create EMBL3, several deletions and one insertion (the *trpE* gene) were made in the λ genome. The stuffer sequence between the two polycloning sites increases the size of the vector DNA itself so that it is packaged efficiently into phage heads, allowing workers to prepare sufficient quantities of vector DNA by propagating the vector as a phage. The cloning is performed by cutting the vector DNA at the two polycloning sites, preferably with BamHI, removing the stuffer fragment, and then ligating the insert DNA (cut partially with Sau3A, which generates the same overhanging ends as BamHI) in between the two polycloning sites. The inserted fragment thus replaces the stuffer fragment in the vector, producing recombinant DNA large enough to be packaged into phage heads. Instead of physically removing the stuffer fragment by electrophoresis, one can cut the mixture of the three fragments of vector DNA further with EcoRI (there is no other EcoRI site in the vector), thereby preventing the stuffer, now with EcoRI ends, from becoming religated in the middle of the vector. In the recombinant DNA, the sequences necessary for lysogenic integration into chromosomal DNA (*att* and *int*) are deleted with the stuffer segment. Thus, the "phage" particle containing the recombinant DNA can cause only lytic infection of the host. [Modified from Sambrook, J., Fritsch, F. F., and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, 2nd Edition, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.]

of the transformation process). With some vectors, the recombinant DNA may become integrated into the host chromosome as a prophage and can be stably maintained as such until the prophage is induced to initiate the lytic cycle. With others, however, the part of the phage genome that is required for integration has been deleted (for an example in EMBL3, see Figure 3.11), and all infection events result in extensive multiplication of the phage, followed by cell lysis.

Some products of foreign genes are very toxic to the host, and it is difficult to clone such genes by using a plasmid vector, even when the plasmid exists in small numbers of copies per cell ("has a low copy number"). This is because we can isolate and identify plasmid-containing bacterial strains

Termination of mRNA Transcription in *E. coli*

The elongation of mRNA is terminated by two major mechanisms. In *rho-dependent termination*, a protein factor, rho, recognizes a complex nucleotide sequence and releases the RNA polymerase from the DNA helix. In *rho-independent termination*, a simpler nucleotide sequence, forming a short loop followed by a succession of U in the mRNA, is recognized by the RNA polymerase itself as the termination signal.

BOX 3.6

only when the plasmids coexist with the host bacteria for many generations. For the cloning of such deleterious genes, the λ phage vectors of the nonintegrating type are ideal; with such vectors the infected host cells are soon killed anyway, and the toxicity of the cloned protein does not make much difference. Lambda-based vectors are also very effective at expressing foreign genes, because some promoters in the lambda genome are quite powerful, and because lambda produces an antiterminator protein N, so that rho-dependent termination of transcription (Box 3.6) can be suppressed. Phage λgt11 is an example of a vector that is useful when the screening of the clones is dependent on the expression of foreign genes.

Cosmids. λ DNA is synthesized in the cytoplasm of the infected cells as a polysequence, or concatemer, containing several repeats of the λ genome. A λ-coded enzyme recognizes the *cos* (or *co*hesive *s*ite) sequences that correspond to the proper ends of the genome and cuts the DNA at these points, preparing it to be packaged into the head (Figure 3.12). *Cosmid vectors* are vectors that contain λ *cos* sites but little other material derived from the λ genome. Foreign DNA inserts are cloned between the two *cos* sequences, which then initiate the *in vitro* packaging of the recombinant DNA, composed of the cosmid and its insert, into λ phage heads. Cosmids also contain a plasmid origin of replication, so that they can be replicated as plasmids, and an antibiotic resistance marker, so that cosmid-containing cells can be selected for (Figure 3.13). Because the cosmid vector is so small (typically only several kilobases), it is possible to clone up to 40 kb of foreign DNA into cosmids and deliver the recombinant DNA very efficiently via phagelike particles assembled *in vitro*. Because of their ability to incorporate larger pieces of foreign DNA, cosmids are significantly better than λ vectors for cloning genomic DNA of higher eukaryotes. However, because cosmids have to be propagated as plasmids, it is difficult to use them for cloning genes (or cDNAs) that code for proteins that are toxic for the *E. coli* hosts.

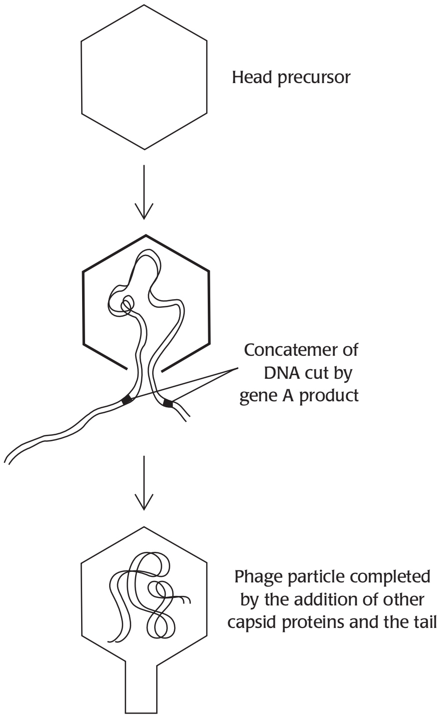

Head precursor
Concatemer of DNA cut by gene A product
Phage particle completed by the addition of other capsid proteins and the tail

FIGURE 3.12

Packaging of DNA into phage head. Normally, λ DNA is produced as concatemers. An enzyme associated with the phage head (gene *A* product) cuts the DNA at each *cos* site, and the linear DNA is then packaged into the phage particle, together with the tail that has been assembled separately.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

Bacterial Artificial Chromosome. Cosmids allow the cloning of DNA sequences up to about 40 kb. However, some genes of higher eukaryotes, containing many introns, are larger. Furthermore, in sequencing the genomes of higher animals and plants, it is necessary to begin with clones of very large segments of DNA, containing hundreds of kilobases. For such purposes, yeast artificial chromosomes, or YACs (described later in this chapter), were the standard vector. However, more recently, bacterial artificial chromosomes (BACs) are the vectors that are most often used. BACs are plasmid vectors, with the F-factor origin of replication and with genes that ensure the partition of the plasmid into both of the daughter cells. BACs are maintained at a very low copy number (1 or 2 per cell), just like the

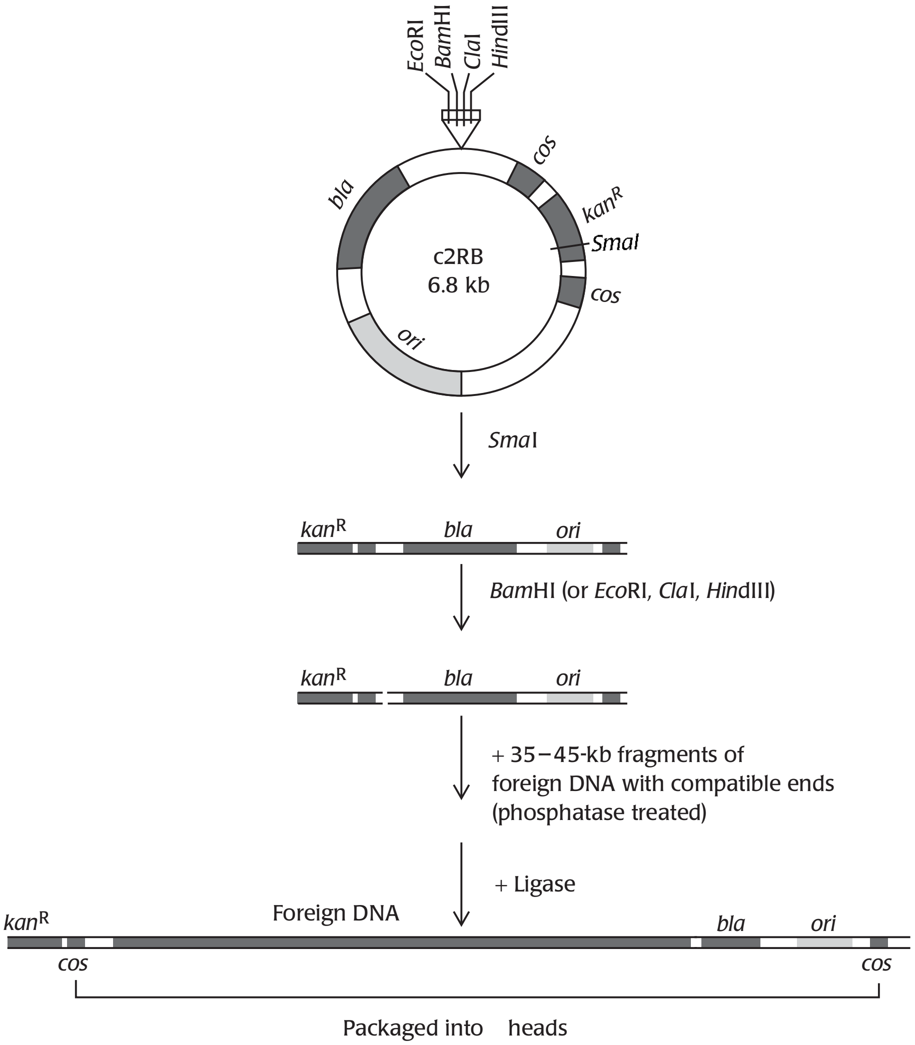


FIGURE 3.13

Cloning with a cosmid vector. The cosmid vector c2RB, shown as an example here, contains two *cos* sites, a plasmid origin of replication, a polycloning site (a short stretch of DNA containing cleavage sites for several restriction endonucleases), and two antibiotic resistance markers (Amp^{R} and Kan^{R}). Cutting the cosmid with, say, SmaI and BamHI produces two cosmid halves. Ligation with 40-kb fragments of foreign DNA partially digested with MboI or Sau3A (which produce the ends complementary to those produced by BamHI) creates the construct shown. This is then packaged *in vitro* and introduced into *E. coli*. The strains containing recombinant DNA should be both ampicillin resistant (because of the *bla* gene) and kanamycin sensitive, because packaging eliminates the kanamycin resistance gene. The latter feature is useful for eliminating plasmids made of multiple copies of the fragments of the vector. [Modified from Sambrook, J., Fritsch, F. F., and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, 2nd Edition, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.]

F-factor, and this also helps the stable maintenance of BAC-based constructs in *E. coli*. The large portion of F-factor, coding for the cell-to-cell transfer of this DNA through conjugation, has been removed so that it will not spread to other cells. YAC DNA is difficult to separate from yeast chromosomal DNA because it behaves almost exactly like other yeast chromosomes. In contrast, the BAC plasmid is easy to isolate away from the bacterial chromosome. Another major advantage of the BAC system is that it is rare for a BAC-based recombinant plasmid to contain more than one piece of cloned DNA, in contrast to YAC-based constructs that tend to contain chimeric pieces of foreign DNA at a very high frequency.

Derivatives of Single-Stranded DNA Phages. One closely related family of phages (fl, fd, M13) infects only those *E. coli* cells that contain the F sex factor. A remarkable feature of these phages is that they continuously produce progeny phages within the growing cells without causing the lysis and death of the host. These phages contain a circular, single-stranded DNA about 6.4 kb long that is replicated as a double-stranded, plasmid-like entity in the *E. coli* cell. The phage particle itself is filamentous, so insertion of foreign DNA at an intergenic site within the phage DNA simply results in an elongation of the phage particle.

These vectors, when they were developed, were essential for DNA sequencing by the dideoxy chain termination method (Box 3.7). However, sequencing is now carried out nearly entirely by using double-stranded DNA. The single-stranded DNA phage vectors were also the vectors of choice for site-directed mutagenesis, but now this can be carried out also by using double-stranded DNA constructs (Figure 3.14). One area where such vectors are still useful is the "phage display" of mutated proteins (Figure 3.15). In this strategy, a DNA sequence coding for a foreign protein of interest is inserted into the 5′-terminal domain of the phage gene coding for protein III. This protein is located at the tip of the filamentous phage, and its N-terminal domain extends into the medium. When the foreign gene is mutated by a site-directed, or random, mutagenesis procedure, each phage particle will express one specific mutated version of the protein. These phages can then be selected out by their affinity to a target. Thus, if the foreign gene codes for an antibody, then the phage expressing a higher affinity antibody can be "fished out" of a mixture of millions of phages, and because the gene coding for this desired mutant is located within the phage, it can be easily recovered. This physical connection between the mutated protein and the coding gene makes this approach extremely useful, especially in an effort to "evolve" proteins of interest through a random mutagenesis approach. More recently, the *ribosome display* method, which exploits the physical connection between the translating ribosomes and the mRNA, has been introduced.

Sequencing of DNA by the Dideoxy Chain Termination Method

This ingenious method, developed by Frederick Sanger, makes possible the sequencing of fairly long stretches of DNA. The first step is to anneal an oligonucleotide primer to the single-stranded DNA one wishes to sequence. DNA polymerase then synthesizes the complementary strand as a 3′-extension of the primer. To each of four such reaction mixtures, one adds low concentrations of an unnatural nucleoside triphosphate containing 2,3-dideoxyribose rather than 2-deoxyribose. This causes chain elongation to stop on those occasions when the unnatural nucleotide is incorporated into the DNA strand. If the template strand contains, for example, C at positions 50, 55, and 60, then the newly made complementary strand becomes truncated when dideoxyguanosine phosphate is incorporated at the corresponding positions. Thus, DNA of 50, 55, and 60 nucleotides in length will be made only in the reaction mixture to which dideoxyguanosine triphosphate was added. Analysis of the products by gel electrophoresis, on the basis of their length, thus permits unequivocal sequencing of the DNA.

BOX 3.7

Two convenient features that were first introduced into M13 vectors (Figure 3.16) are now present in many vectors of other types. The first is a system for distinguishing between recombinant clones and the original vectors. It consists of a fragment of the *lacZ* gene that contains the portion coding for the N-terminal fifth of the LacZ protein. When this truncated LacZ fragment is expressed in a host cell that contains a *lacZ* gene lacking the 5′-terminal part of the gene, both fragments can assemble together spontaneously to produce a functioning enzyme (alpha-complementation). Thus, when a cell harboring this vector phage is placed on a plate containing 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal), hydrolysis of X-gal by β-galactosidase (LacZ protein) produces indoxyl, which is oxidized to indigo that stains the colony blue. When a segment of foreign DNA becomes inserted into the cloning site, the coding sequence of the truncated *lacZ* gene is interrupted, the functional N-terminal LacZ fragment is

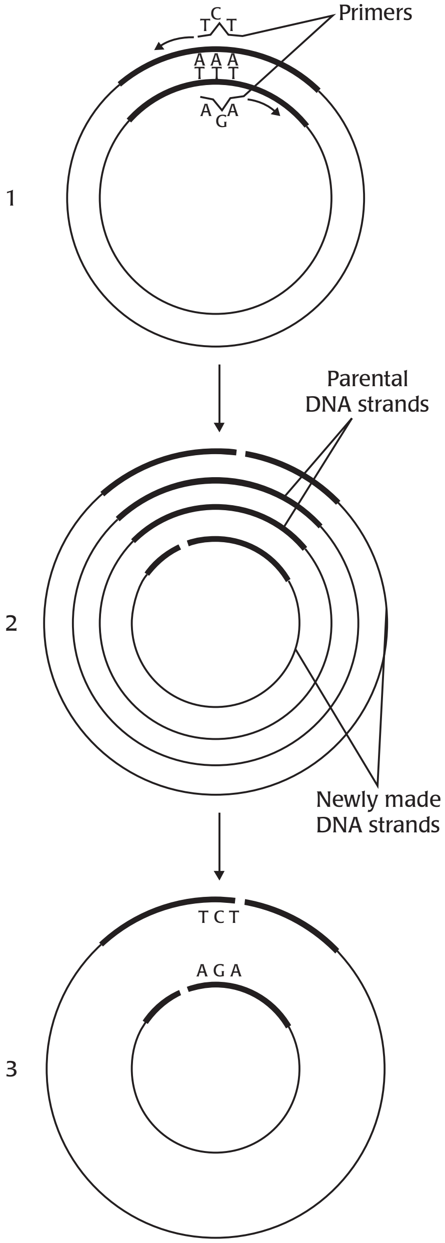


FIGURE 3.14

Site-directed mutagenesis. This figure illustrates the QuickChange method developed by Stratagene. Starting from a recombinant plasmid containing the target gene insert (*thicker line*), two primers covering the same overlapping area are used. If one wants to change a lysine residue in the protein (coded by the AAA codon) into arginine (coded by the AGA codon), the primers will contain a single mismatched nucleotide residue (AGA and TCT, respectively). These are indicated by protrusions in the figure, step 1. PCR is used to elongate the primer and cover the entire plasmid, as well as to amplify the DNA, resulting in the structure shown as step 2, which contains both the newly synthesized strands containing the desired mutation as well as the parental strands not containing the mutations. The latter were made in *E. coli* cells, and therefore some of the bases are methylated. Treatment with the restriction endonuclease DpnI, which specifically cleaves DNA containing 6-methylated guanine, cleaves the parental strands, leaving behind the *in vitro* synthesized, and therefore unmethylated, strands that contain the desired mutation (step 3).

not produced, and the colony stays white. (In principle, the same effect can be achieved by inserting the whole *lacZ* gene in the vector. However, *lacZ* is a large gene, and introducing large DNA fragments makes the recombinant M13 construction rather unstable.) The second feature is the insertion, close to the beginning of the *lacZ* gene, of a short sequence called *polylinker*, or *multiple cloning site*, designed to contain cleavage sites for many popular restriction enzymes. This sequence serves as a convenient site of insertion of foreign DNA. Its proximity to the efficient *lac* promoter ensures good expression of the cloned gene, as long as the gene is in the correct orientation. (The advantage of this construction for gene expression is further discussed in the section dealing with expression vectors).

Phagemids are a variant on these vectors. These chimeric vectors contain two origins of replication, one from a plasmid and the other from fl or some

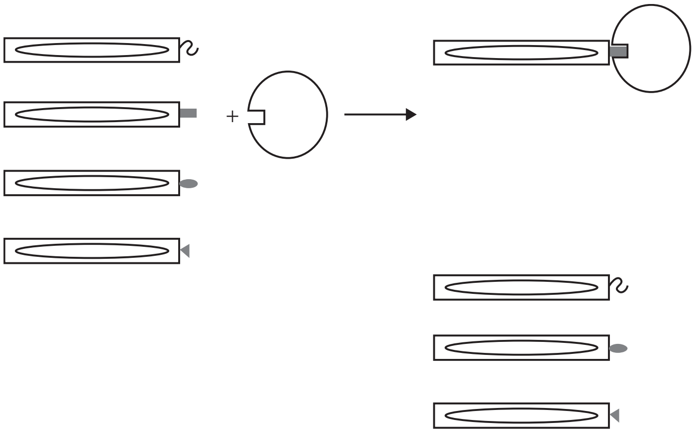


FIGURE 3.15

Phage display technology. In this approach, the DNA sequence coding for the protein to be mutated is inserted in the exposed domain of the protein III (PIII), which is located at the tip of the filamentous phage such as M13. This sequence is subjected to random changes (e.g., by using, in the chemical synthesis of DNA, a mixture of four deoxynucleoside triphosphates, rather than the single one, at given positions in the sequence), and the replicative form DNA is transformed into *E. coli*. An assembly of phages producing various mutated forms of the protein will emerge from the host cells. This mixture is then subjected to an affinity selection, for example, with a protein that may be expected to interact with the protein at the phage tip. Only the phage with the tip protein domain that "fits" with the protein used for selection will be retained. The precise mutation in the gene can be recovered from the genome of the phage that has been retained.

other phage. These vectors multiply as plasmids in the host cell because they lack genes needed for replication as phage DNA and for the assembly of phage particles. However, once the missing phage functions are supplied by superinfecting the host with helper phages, they are replicated as phage DNA, packaged, and released into the medium as phagelike particles.

DETECTION OF THE CLONE CONTAINING THE DESIRED FRAGMENT

Cloning fragments of genomic DNA is not a difficult task. Many restriction endonucleases are commercially available, as are numerous vectors of sophisticated design, such as those we have described. Usually the most challenging step in the shotgun cloning is the detection, among many clones, of the ones that contain the fragment of interest. The magnitude of this task becomes clear when we realize that even with vectors that can accept a 20-kb piece of DNA, and even when the source is a bacterial genome (5000 kb), to have a 99% probability of finding one clone with the desired gene, we have to examine about 1000 clones (see Box 3.1). The thought of attempting the same task with the genome of a higher eukaryote, which could be almost three orders of magnitude larger than that of *E. coli*, is daunting to say the least. One would have to examine almost a million recombinant clones in order to be 99% certain of recovering the gene of interest. Thus, careful strategic planning becomes necessary for the identification of desired clones.

Importance of Using a Better Template

If one wants to express eukaryotic proteins in bacteria, which cannot carry out the splicing reaction, it is best to use mRNA as the template, because the sequences corresponding to the intron sequences are already spliced out, as described earlier. In such cases, the usual procedure is to obtain the specific types of cells in which the target gene is expressed strongly and then to use the mRNA from those cells as the source of genetic information. This approach exploits a source in which the sequence of interest has been very strongly amplified. In such cases, the recombinant DNA constructs will be highly enriched in the target sequence, so a minimal amount of screening will be needed to isolate the desired construct. Large amounts of stable RNA (ribosomal RNAs, transfer RNAs, and so on) are also present in any cell, but

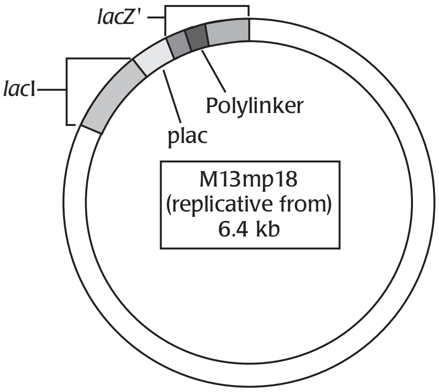


FIGURE 3.16

An example of M13-derived vectors. The M13mpl8 vector contains a polylinker sequence near the 5′-terminus of the sequence coding for a fragment of the *lacZ* gene called *lacZ′*. Precise sequences of the promoter region and one version of the polylinker of this type are given in Figure 3.21.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

they can be removed easily by taking advantage of the fact that eukaryotic mRNA molecules have a polyA "tail" at the 3′-terminus. Once the mRNA fraction is isolated, it can be purified according to size in order to obtain a fraction enriched in the sequence of interest. The mRNA is then converted into double-stranded cDNA as described above (see Figure 3.6) for insertion into a cloning vector. Most of the sequences coding for the production of animal and human peptides and proteins have been cloned by using mRNA preparations.

All of the recombinant constructs should contain the desired piece of DNA, if it was created by the PCR amplification (see below); this allows us to totally circumvent the clone identification step as well as all the other steps of the shotgun cloning.

Clone Identification Based on Protein Products

When a cloned gene is expected to be transcribed and translated in the host bacterium (see the discussion of expression vectors that follows), the task of identification, and perhaps even selection (see Box 3.3), of the cells containing the right clone is fairly straightforward. In the simplest case, one can test for the function of the protein coded by the cloned gene. Let us assume that we want to clone from some organism (call it Organism A) the gene for anthranilate synthase, an enzyme involved in the synthesis of tryptophan, for the purpose of improving the commercial production of this important amino acid (see Chapter 9). *E. coli*, like most bacteria, can synthesize all the usual amino acids from simple carbon sources and ammonia, and it contains a gene, *trpE*, that codes for anthranilate synthase. We first introduce a mutation into the *E. coli trpE* gene. The mutant strain cannot synthesize tryptophan and thus cannot grow unless we add tryptophan to the growth medium. We now introduce into this strain, by transformation, recombinant plasmids containing fragments of the DNA of Organism A and spread a large number of transformant cells on a solid medium that does not contain tryptophan. Most of the cells contain either no plasmid or plasmids with irrelevant pieces of DNA and are unable to grow. The only cells that grow and form visible colonies are those that contain the rare recombinant plasmid with the *trpE* homolog from Organism A. In this manner, we achieve an efficient selection of these rare plasmids.

In the foregoing example, the desired gene had a function required in many microorganisms. In some cases, however, the desired gene would have a significant function in the source organism, Organism A, but not in *E. coli*. An example is an attempt to clone a gene coding for one of the enzymes of the xylene degradation pathway from *Pseudomonas putida*. Many strains of this organism contain a series of enzymes that lead to the complete oxidation of an aromatic hydrocarbon, xylene, but one of these enzymes can perform no useful function in *E. coli*, which does not contain any other enzymes of this series. *Shuttle vectors*, which contain origins of replication of both *E. coli* and some other microorganism, are useful in such situations. We can then screen for the clone that expresses the desired function in a mutant of Organism A that lacks this function, because the recombinant plasmids will be replicated in this organism. At the same time, we can propagate the plasmids in *E. coli*, in which subcloning and other procedures can be carried out more easily.

In many cases, though, a *complementation assay* such as the one described would be difficult or even impossible to perform. For example, if we are trying to clone a eukaryotic gene coding for a hormone that has no homologs in unicellular bacteria, complementation cannot be used as a method of detection. A frequently used approach in these cases is to detect production of the desired protein by its reactivity with specific antibodies. Unfortunately, this usually involves screening, rather than selection, of the recombinant clones. However, if the screening can be carried out on plates with hundreds of colonies on each, it is not so difficult to test tens of thousands of recombinant clones in a single experiment. λ Phage vectors are especially convenient for this method, because within each plaque (see the legend of Figure 3.2) generated by lytic infection by a recombinant phage or by induced lysis of an *E. coli* strain lysogenic for a recombinant phage, the cells will have been lysed already, releasing into the medium the proteins expressed from the recombinant fragment. Furthermore, because a single lysing cell contains hundreds of copies of the phage genome, each including the cloned piece, the expression of the cloned genes is strongly enhanced. The λ gt11 expression vector was especially constructed for screening of this type.

Clone Identification Based on DNA Sequence

The methods we have examined depend on successful expression of the cloned genes. But this is not always assured, especially when the cloned DNA comes from a source phylogenetically distant from the bacterial host. The RNA polymerase of the host bacteria does not recognize the promoter and other regulatory elements of eukaryotic genes, or even those of remotely related bacteria. Pieces of cDNA lack such regulatory "upstream" sequences altogether, and genomic DNA from eukaryotes will not result in the production of whole proteins because of the presence of introns.

Because of these problems, it often becomes necessary to identify the clone containing the desired fragment by its DNA sequence. Scoring for such clones can be done by hybridization with suitable DNA probes, labeled either with a radioactive isotope or with chemical substituents that can be detected by nonradioactive methods, such as fluorescence. The major hurdle in this procedure is finding the requisite DNA probe, especially when the exact sequence of the clone is not yet known. This is not an insurmountable problem, however. If the sought-after gene has homologs in related organisms, and if their sequences are known, it is possible to design probes that correspond to the most conserved regions of the aligned sequences and use conditions of low stringency for hybridization. In fact, this is probably the most frequently used method for cloning genes and cDNAs from eukaryotes, because the evolutionary divergence between higher eukaryotes tends to be quite small in comparison with that between prokaryotic groups. Alternatively, if at least a partial sequence of the protein is known, one can deduce the DNA sequences that would code for such an amino acid sequence and use a “degenerate” probe that contains a mixture of these possible DNA sequences. Using this approach is particularly advantageous when the peptide sequence does not contain amino acids that, like leucine or arginine, are coded by many codons.

In practice, the cells containing recombinant plasmids are spread on plates so that there will be a few hundred colonies per plate. These colonies are replica-plated onto a filter and placed on a fresh plate. After the cells have grown, the filter is lifted out and treated with an NaOH solution to lyse the cells and denature the DNA. The proteins are digested by a protease, and the DNA is fixed onto the filter by “baking” at 80°C. The filter is then incubated with the labeled probe DNA, and any probe that anneals to the DNA on the filter is detected after suitable washing (Figure 3.17). Although this is only a

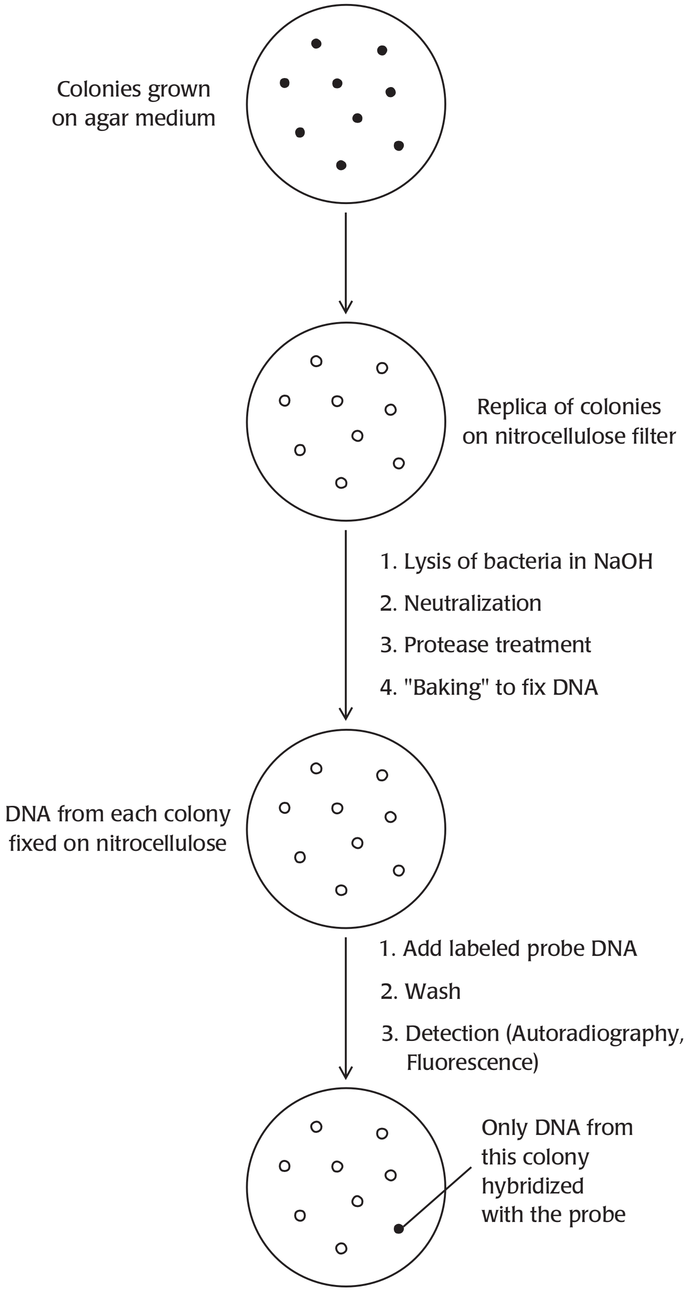

Colonies grown on agar medium
Replica of colonies on nitrocellulose filter
1. Lysis of bacteria in NaOH
2. Neutralization
3. Protease treatment
4. "Baking" to fix DNA
DNA from each colony fixed on nitrocellulose
1. Add labeled probe DNA
2. Wash
3. Detection (Autoradiography, Fluorescence)
Only DNA from this colony hybridized with the probe

FIGURE 3.17

Use of a DNA probe to detect the desired recombinant clone.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

screening method, in this way one can test a fairly large number of colonies in a short time.

Combined Detection of the DNA Sequence and the Protein Product

In some cases, it is possible to combine the two methods we have outlined. For example, to clone gene *X* from a bacterium very distantly related to *E. coli*, one might begin by making random transposon insertions (Box 3.8) in the chromosomes of the bacterium containing gene *X*. If a transposon inserts into gene *X*, it will disrupt this gene, generally resulting in a recognizable phenotype. Genomic DNA from this mutant organism is then cloned into a plasmid vector, and the recombinant plasmids are introduced into an *E. coli* host strain by transformation. Most transposons contain a resistance marker that codes for antibiotic resistance or resistance to other toxic compounds, such as mercury (Figure 3.18). Moreover, because a transposon is a piece of "selfish" DNA that propagates itself in diverse species of bacteria (see Box 3.8), its resistance genes are designed to be expressed efficiently in many bacterial species. Thus the resistance gene, located within the transposon in the recombinant plasmid, will be expressed in the *E. coli* host, and it should therefore be possible to select for this plasmid.

A clone of gene *X* still exists in two pieces within the plasmid, flanking the transposon. The cloned DNA can be cut out of the plasmid, and the fragments of gene *X* DNA can be used as probes in the next phase. One then clones random fragments from the wild-type genome (which does not contain the transposon) into a plasmid or other vector. Screening of these recombinant clones with the DNA probes of gene *X* described above will lead to identification of the clone that contains the wild-type version of the gene, uninterrupted by the transposon sequence. Alternatively, the sequence of the wild-type gene can be retrieved by the procedure called inverse PCR (see below).

POLYMERASE CHAIN REACTION AND THE UTILITY OF GENOMIC DATABASES

In cases in which we know at least short stretches of nucleotide sequence either within the gene of interest or in an area flanking that gene, it is possible to isolate the desired clone without going through the painstaking shotgun cloning procedures we have described. This is done with a technique known as polymerase chain reaction (PCR), and in 1993 its inventor, Kary Mullis, received a Nobel Prize in chemistry for devising it. In this method, we first synthesize oligonucleotide primers that are complementary to opposite strands of these short stretches of DNA (Figure 3.19). We then add a large excess of these primers to a denatured preparation of genomic DNA or cDNA and let the primers anneal to the complementary sequences. Adding a heat-resistant DNA polymerase and a mixture of deoxyribonucleoside triphosphates results in the elongation of primers into complementary strands of DNA, as shown in Figure 3.19, step 1. The mixture is next heated to denature the DNA again, and then it is rapidly cooled. Under these conditions, annealing occurs predominantly between primers and DNA strands (some

Transposons

A transposon is a segment of DNA that has the ability to insert its copy at random sites in the genome. Thus, a transposon has a natural tendency to increase the number of its copies under favorable conditions and can be considered an example of a piece of "selfish DNA." It is bounded by inverted repeat sequences, and it usually contains a drug resistance gene. It also contains gene(s) for enzymes that catalyze the transpositional insertion. Because of these properties, transposons played a major part in the natural formation of "R plasmids," which code for resistance to many antibacterial agents. Transposons are also very useful as genetic tools. For example, introduction of a transposon into a new cell results in the insertion of copies of the transposon at various places on the chromosome. Insertion of such large fragments in the middle of a gene totally inactivates the gene, so *transposon mutagenesis* is a convenient method for generating null mutants (mutants in which the gene function is utterly obliterated). Furthermore, the mutations are easy to analyze genetically, and the alleles can be easily cloned because of the presence of antibiotic resistance markers within the transposon sequences.

BOX 3.8

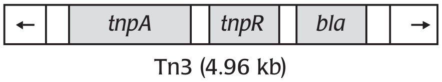


FIGURE 3.18

An example of a transposon, Tn3. The genes *tnpA* and *tnpR* code for transposase (cointegrase) and resolvase, two enzymes needed for the insertion of a copy of the transposon into a new site on the DNA duplex. [For the mechanism, see Grindley, N. D. F., and Reed, R. R. (1985). Transpositional recombination in prokaryotes. *Annual Review of Biochemistry*, 54, 863–896.] The gene *bla* codes for a β-lactamase, which produces resistance to β-lactam antibiotics such as ampicillin and cephalothin. The ends of the transposon contain inverted repeats (*arrows*).

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

of which are newly synthesized), rather than between long DNA strands, because the latter takes place more slowly. Because the polymerase is heat resistant, DNA synthesis begins again by utilizing the primers (Figure 3.19, step 2). In the first round of DNA synthesis, the newly made DNA strands have random ends. In the second round, the mixture becomes enriched for strands that begin and end at sequences corresponding to the two primers, because some primers have annealed to the strands synthesized in the first round (Figure 3.19, step 2). After the DNA synthesis and DNA denaturation/annealing steps are continued for many cycles, most of the newly made strands will have a finite length and will correspond only to the limited region of the DNA between the two primers – that is, to the gene of interest if the primers corresponding to the regions flanking the gene were used.

The crucial factor for the success of PCR was the discovery of a thermostable DNA polymerase that can withstand many cycles of heating and cooling. In theory, the usual heat-labile enzyme should suffice if it is added fresh at the beginning of every cycle; however, a large number of cycles are required to achieve a high degree of amplification, and impurities brought in each time the enzyme is added eventually inhibit the reaction. The heat-stable enzyme commonly used (Taq polymerase) is derived from a thermophilic Gram-negative eubacterium, *Thermus aquaticus*, which grows optimally at around 70°C to 80°C. Lately, even more thermostable DNA polymerases, isolated from archaebacteria living in marine thermal vents at temperatures of 98°C to 104°C, are being used in PCR procedures.

There are several ways to clone the PCR amplification product into vectors. The Taq polymerase creates products with a one-residue overhang of deoxyadenosine at the 3′-end. Thus, the product can be cloned into a cleaved site of a vector, which contains the complementary one-residue overhang of deoxythymidine at the 5′-end. Alternatively, the 3′-5′ exonuclease activity of the Klenow fragment of DNA polymerase can be used to remove the 3′-overhang of the PCR product, to create flush or "blunt" ends. The product can then be cloned into vectors, which were cleaved with endonucleases known to create blunt ends, by a process called "blunt end ligation." Perhaps the most efficient procedure is to use primers that contain 5′-extensions corresponding to the restriction sites of endonucleases to be used. The presence of such extra sequences does not inhibit the PCR process. The product is then cleaved with the restriction endonuclease(s), to generate sticky ends that will anneal with the complementary sticky ends of the vector, created by cleavage by the same enzymes.

One potential problem with the use of Taq polymerase is that it lacks the 3′-5′ exonuclease activity that is used in "proofreading" the newly made strand. Thus, errors occur during DNA synthesis, and if they occur in the early cycles, the amplified DNA may differ in sequence from that of the original template. However, the frequency of error is strongly reduced with the newer archaebacterial enzymes, some of which contain the 3′-5′ exonuclease activity (Chapter 11).

The PCR procedure offers important advantages. As we have seen, one can totally circumvent the complicated cloning steps as well as the steps

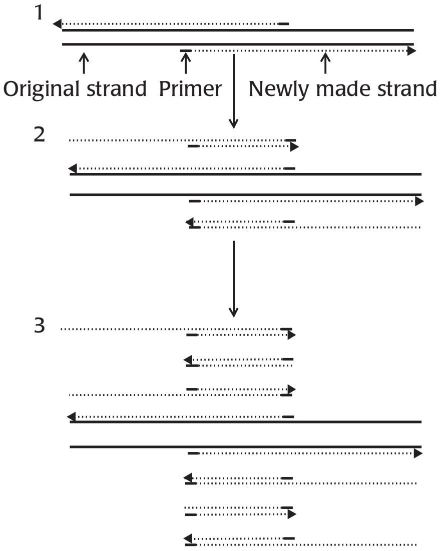


FIGURE 3.19

Amplification of a defined segment of genome by PCR. In step 1, primers (*short lines*) are annealed to complementary sequences of genomic DNA (*continuous lines*). Addition of DNA polymerase and deoxyribonucleoside triphosphates results in elongation of the primer (*dotted lines*). In step 2, the reaction mixture is heat denatured and then renatured, causing some of the primers to anneal to newly made strands (*dotted lines*). Elongation produces new strands, two of which now have a limited length, terminating at positions corresponding to the primer sequences. In the third cycle, after denaturation, renaturation, and elongation again, eight out of the total of 16 strands have this limited length. After further cycles, practically all of the newly made DNA will have the finite, short length.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

involved in identifying the clone that contains the desired gene. Moreover, because the degree of amplification is very large, the process is extremely sensitive. In theory, this procedure could amplify a single copy of a gene, and in many experiments the results have approached this limit. This has led to the development of many diagnostic tools: whereas it took many days (or even weeks) to culture and identify pathogenic bacteria infecting patients, it now takes only a few hours to show the presence of such pathogens by amplifying specific DNA sequences of each pathogen. There are several commercial systems that were approved by the Food and Drug Administration (FDA) for detection of *Mycobacterium tuberculosis*, which causes tuberculosis and grows extremely slowly. Diagnostic PCR is not limited to detection of pathogens. Some types of cancer cells are marked by characteristic changes in the genome, and these can be detected by PCR in a sensitive way.

PCR depends on knowing the sequence either within or around the gene of interest. One way to satisfy this requirement is to isolate the protein product and to determine the amino acid sequences of the N-terminus and of internal peptides generated by the enzymatic or chemical cleavage of the protein. Primers are then made on the basis of these amino acid sequences. These primers are mixtures containing various degenerate codons.

In recent years, however, there has been an explosive growth in our knowledge of genome sequences. The Comprehensive Microbial Resource webpage at The Institute for Genomic Research (http://www.tigr.org) lists, at this writing, more than 400 complete or nearly complete genome sequences of microorganisms. The sequences of individual genes and fragments deposited in GenBank and other databases exceed, by far, the sequences in complete genomes, and the sum of both types of sequences reached 100 gigabases (Gb), or 10^{11} bases, in August 2005. This huge size of information on genes of diverse organisms, coding for almost any imaginable function, now allows us to construct primers on almost any gene. As a hypothetical example, let us suppose that you are interested in the biological oxidation of MTBE (methyl *tert*-butylether), a gasoline additive, which is not easily biodegraded and is polluting the environment. In your search for organisms and enzymes that degrade this compound, you find an article reporting that propane monooxygenase of *Mycobacterium vaccae* rapidly oxidizes MTBE to convert it into innocuous products. However, all *Mycobacterium* species are classified as potential human pathogens, and therefore there is no chance that you can use this bacterial species directly for environmental cleanup. You will thus have to clone the gene coding for this enzyme from *M. vaccae*. However, the sequence of this gene is not known. You know, however, that a propane monooxygenase gene has been sequenced from a related genus, *Gordonia*. In such a situation, you can take advantage of the enormous amount of our knowledge on DNA sequences by first pulling out the *protein* sequences that are most related to the *Gordonia* protein sequence. (Nucleotide sequences change too rapidly during evolution, and it is more useful to rely on protein sequences in order to find homologs). This can be done by using the program BLAST, at the website of National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov). You can then align the sequence of the *Gordonia* enzyme with those of several homologous enzymes, preferably coming from different genera. This will show two internal segments in which stretches of at least five amino acids are completely conserved. You can design primers from these sequences, taking the codon degeneracy into account, and amplify an internal fragment of the desired gene from *M. vaccae*.

In the example above, the PCR procedure amplified only a fragment of the desired gene. A procedure called inverse PCR is a convenient starting point for the cloning of the entire gene. As shown in Figure 3.20, one cuts the chromosome with a restriction enzyme, and then self-ligates the fragments to make them circular. Use of primers from the already cloned small fragment, going in divergent directions, will result in the amplification of the entire sequence of the larger chromosomal fragment. This can be sequenced to elucidate the exact sequences of the 5′- and 3′-termini of the complete gene, and these can then be used to design primers for the amplification of the complete gene.

Finally, error-free chemical synthesis of long (up to 40 kb) DNA is now possible. Thus the entire DNA sequence including promoter, operator, RBS, coding sequence (with optimal codon usage), and terminator at optimal locations and sometimes even including many genes, can be synthesized. Such an approach may soon replace most of the cloning and PCR methods described, if the cost of synthesis becomes competitive.

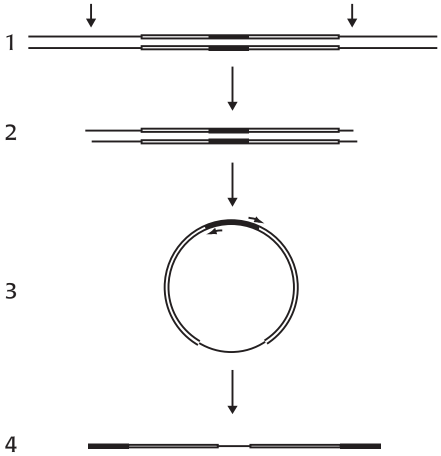


FIGURE 3.20

Inverse PCR. If we know the sequence of only a segment (represented by a *black section* in step 1) of a gene of interest (the rest of the gene is represented by an *empty box* in step 1), it is possible to recover the rest of the gene from the genomic DNA. (The known sequence could be that of a transposon that has inserted into the gene, see p. 112). We first cut the genomic DNA by using a restriction endonuclease (*vertical arrows* in step 1), generating fragments with complementary overhanging ends (step 2). Annealing of the ends and ligation produces circular DNA fragments (step 3). These DNA circles cannot replicate in intact cells as they lack the origin of replication. However, they can be replicated *in vitro* as linear pieces of DNA by using PCR. For this purpose, we use a set of primers that are directed outward from the known segment of the gene (see the *small arrows* in step 3). Because the normal PCR procedures use a forward primer and a reverse primer that face each other, this procedure is called *inverse* PCR, which results in the recovery of the flanking parts of the gene of interest (step 4). Many modifications have been devised for this approach.

EXPRESSION OF CLONED GENES

The usual reason for cloning a gene is to obtain the protein product in substantial quantities. Even when that is not the case, identification of the correct clone often requires expression of the cloned gene. However, many of the general-purpose cloning vectors are not designed for strong expression of cloned genes. There are many reasons why genes in the fragments cloned into pBR322, for instance, are often expressed only at a low level. Frequently, the foreign promoter in a fragment from another organism is not efficiently recognized by *E. coli* RNA polymerase. In such a case, successful transcription must start from promoters recognized well by *E. coli* – those for the *tet* and *bla* genes – and continue onto the cloned segment of the recombinant plasmid. The problem is that there may be sequences in between that act as transcription terminators. Even if the mRNA is successfully produced, it may not contain the proper ribosome-binding sequence (see below) at a proper place. These difficulties indicate that a different arrangement is needed to ensure a high level of expression of foreign genes in a reproducible manner.

Vectors of a special class called *expression vectors* are designed for this purpose. Most expression vectors are plasmids because multiple copies of plasmids can exist stably in the cell. More plasmids carrying a given gene result in a higher production of specific mRNA because each copy of the gene is transcribed independently. This principle is sometimes called the *gene dosage effect*.

Consensus Sequence

Promoters of various genes share a common stretch of nucleotide sequence, with only minor variations. Such homologous sequences, and often their idealized versions, are called consensus sequences. Actual sequences may deviate significantly from idealized sequences: for example, the sequence 10 bases upstream from the transcription initiation site (−10 sequences) for *E. coli trp* operon (TTAACT) and *lac* operon (TATGTT) are similar but not completely identical to the consensus TATAAT.

BOX 3.9

Expression vectors must have a strong promoter. *E. coli* promoters contain two “consensus” sequences (Box 3.9): TTGACA, about 35 nucleotides upstream from the transcription start site, and TATAAT, about 10 nucleotides upstream. These two sequences are therefore separated by a 16- to 18-bp intervening region. When the vector’s promoter has sequences that closely resemble the host bacteria’s, the genes downstream of it tend to be expressed strongly. Strong promoters are also found in phage genomes, because phage life cycle depends on its proteins being produced in very large amounts during the short period of phage infection. The system using phage T7 promoter is described later in this chapter.

Strong promoters introduce a problem, however. When *E. coli* cells produce a very large amount of a protein that does not contribute to cell growth, such a situation tends to be deleterious. Thus, cells that have lost the plasmid, and cells whose plasmids have been altered and have ceased to produce the protein, have a competitive advantage and will eventually become the predominant members of the population. This instability can be a severe problem in industrial-scale production, because extensive scale-up means that *E. coli* must go through a proportionately larger number of generations, significantly increasing the likelihood that nonproducing cells will appear. For this reason, it is preferable, and in most cases necessary, to use promoters whose expression can be regulated so that the production of the foreign protein can be delayed until the culture has reached a high density. Common regulatable promoters used in *E. coli* include pLac (from the lactose operon), pTrp (from the tryptophan operon), pTac (a man-made hybrid between pLac and pTrp, used to produce a much higher level expression than that of its parents) and pAra (from the arabinose operon). The lactose promoter is easy to induce, but its uninduced (basal) level of transcription is often significant and may create problems when the foreign gene products being expressed are strongly toxic to *E. coli*. Accordingly, it is a common practice to use this promoter in the presence of the $lacI^q$ allele, which leads to the increased production of LacI repressor, thanks to a mutation in the *lacI* promoter, to suppress efficiently the uninduced level of transcription. Furthermore, the lactose promoter commonly used contains a mutation called UV5, which abrogates the catabolite repression so that the cloned gene can be expressed in a rich medium.

Good expression vectors need to have a Shine–Dalgarno sequence, or ribosome-binding sequence (RBS), typically AAGGA, a sequence complementary to a part of the 3′-terminal segment of 16S rRNA. This complementarity allows the mRNA to associate with the 30S ribosomal subunit of *E. coli*. The proper distance between the RBS and the first codon, ATG, of the gene is critical for an efficient initiation of translation: in one example, decreasing the distance from the optimal one (seven nucleotides in between these sequences) by only two nucleotides decreased the expression level by more than 90%. The Shine–Dalgarno sequence is absent in eukaryotic mRNA. If the cloned fragment comes from such an organism, it is necessary to insert the *E. coli* RBS into the vector and to place the 5′-terminus of the cloned gene close to this RBS to ensure the efficient translation of the mRNA.

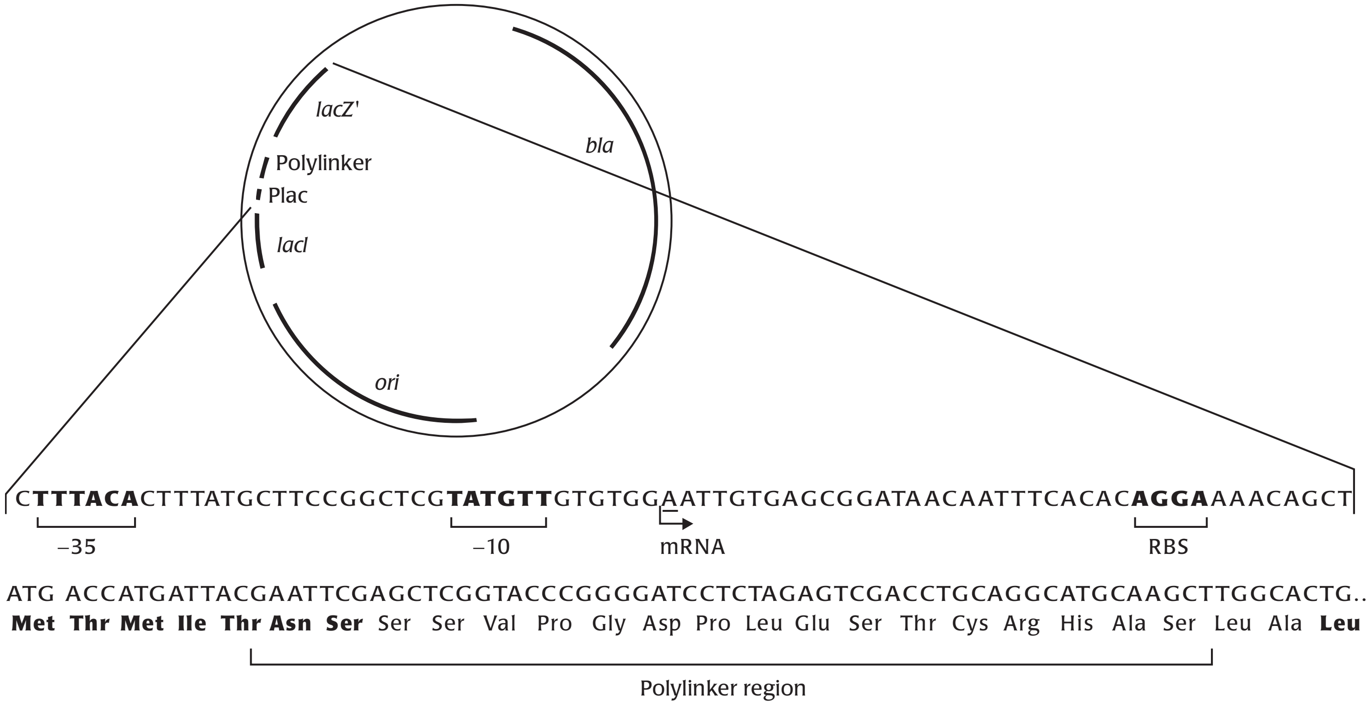

lacZ'
bla
Polylinker
Plac
lacI
ori
CTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAAACAGCT
−35
−10
mRNA
RBS
ATG ACCATGATTACGAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTGGCACTG...
Met Thr Met Ile Thr Asn Ser Ser Ser Val Pro Gly Asp Pro Leu Glu Ser Thr Cys Arg His Ala Ser Leu Ala Leu
Polylinker region

FIGURE 3.21

Structure of the *E. coli* expression vector pUC18. In addition to the origin of replication (*ori*) and an antibiotic resistance marker (*bla*), this vector contains a portion of the *E. coli lac* operon. The latter includes the repressor gene (*lacI*), the promoter region (Plac), and the 5′-terminal portion of the *lacZ* gene, coding for about 60 amino acid residues. As shown at the bottom, a polylinker region (containing restriction sites for more than 10 endonucleases) is inserted inside the *lacZ* gene. The amino acids present in the LacZ protein are shown in boldface type, those coded by the polylinker sequence in standard type. The polylinker does not contain any nonsense codons and is inserted in phase, so the vector codes for a complete N-terminal fragment of LacZ with an 18–amino acid insert. The -35 and -10 regions of the promoter, as well as the RBS, are indicated. Note that these sequences deviate somewhat from the consensus sequences. The catabolite-activator-protein (or cAMP-binding protein) (CAP)-binding sequence of the *lac* promoter is located upstream of the sequence shown here.

Some of these features are illustrated by the vectors of the pUC series (Figure 3.21). The segment that contains the regulatable promoter, pLac, and the 5′-terminal portion of the *lacZ* gene comes intact from the lactose operon of *E. coli*. Thus, the promoter is at a proper (natural) distance from the transcription initiation site. The Shine–Dalgarno sequence is located at its natural distance from the initiation codon of the *lacZ* gene. At the very beginning of the *lacZ* gene, the polylinker developed earlier for the M13 series of vectors (see page 107) provides many cloning sites within a very short stretch of DNA. Because of this arrangement, it is possible to express the cloned protein very efficiently as a fusion protein containing only a few of the N-terminal amino acids of LacZ. A final convenient feature is that the disruption by the cloned fragment of the 5′-terminal fragment of the *lacZ* gene makes it possible to distinguish cells containing recombinant clones from those containing resealed vectors only, as explained in connection with the M13 vectors.

Another example of widely used expression vectors is the pET series (Figure 3.22), developed by Novagen on the basis of studies on T7 phage biology by F. William Studier. The gene to be expressed is cloned within a polylinker behind a very strong T7 promoter. Because this promoter is recognized by the T7 RNA polymerase but not at all by the *E. coli* polymerase, the uninduced level of expression can be kept exceptionally low. When the culture reaches a high density and the cloned gene is ready to be expressed, the T7 RNA polymerase gene, cloned behind pLacUV5 promoter in the host strain, is induced by using IPTG. If the protein to be expressed is exceptionally toxic to the host cell, host strains expressing T7 lysozyme at a low level are used. Because lysozymes become necessary to lyse the host cells only at the last stage of phage infection, in which transcription of other phage genes is

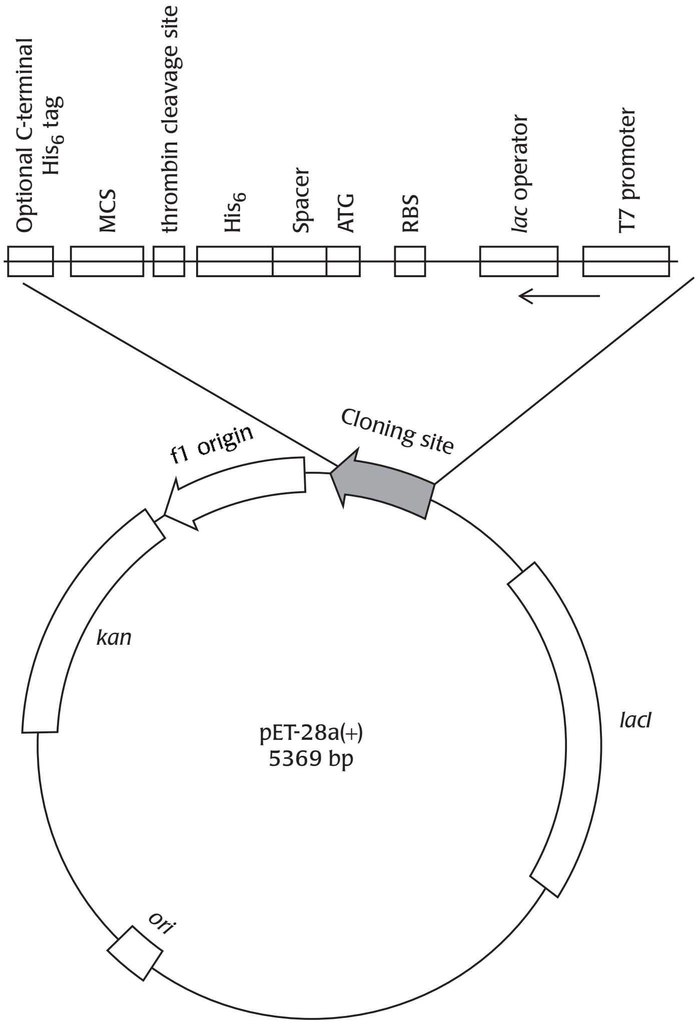


FIGURE 3.22

Structure of the *E. coli* expression vector pET-28a(+). The cloning site (shown by a *thick black arrow*) contains a T7 promoter, followed by the upstream region of the *lac* operon containing the RBS as well as the *lac* operator (similar to the pUC18, Figure 3.21), by the initiation codon ATG, a three-codon spacer, and a hexahistidine tag sequence, then by a cleavage site by thrombin, and finally by the multiple cloning site. Thus, the target protein will be produced with an N-terminal hexahistidine tag (see p. 119), which can be removed after affinity purification by treatment with thrombin. (The cloning site even contains an additional hexahistidine sequence, which can be used to attach a C-terminal tag to the protein). The plasmid, in addition to the antibiotic resistance marker *kan*, which produces kanamycin resistance, also contains the origin of replication for phage f1. Thus, this is a *phagemid* (see text), and the construct can be recovered as a single-stranded DNA by infection with a helper phage. Note also that this multicopy plasmid contains the *lacI* gene, which produces lactose repressor that prevents, in the absence of IPTG, the "leaky" expression of the target gene by binding to the *lac* operator site in the cloning site.

not needed, T7 lysozyme acts as a natural inhibitor of T7 RNA polymerase. Thus, the transcription of the cloned gene by the low levels of T7 polymerase, which could be produced by the baseline level transcription of its gene in the absence of IPTG, becomes nearly completely inhibited. With the pET series, an expression level of a cloned protein approaching 40% to 50% of the total cellular protein is sometimes reported.

Finally, when the cloned gene comes from an organism that is not closely related to *E. coli*, one should pay close attention to the codon usage. For example, arginine codons AGA and AGG are rarely found in *E. coli* genes but are frequent in eukaryotes. Expression of such genes in *E. coli* often leads to translational arrest, with subsequent degradation of mRNA. Such codons should be altered by the site-directed mutagenesis to enhance translation in *E. coli*.

RECOVERY AND PURIFICATION OF EXPRESSED PROTEINS

Even when the cloned gene is successfully expressed in a bacterial host, product recovery is not always a simple matter. Potential problems and some approaches to solving them are discussed below.

TABLE 3.1 Specific cleavage reactions

Cleavage effector	Cleavage site
Acidic pH	–Asp↓Pro–
Hydroxylamine	–Asn↓Gly–
CNBr	–Met↓Xaa–
Trypsin	–Arg (or Lys)↓Xaa–
Clostripain	–Arg↓Xaa–
Collagenase	–Pro–Xaa↓Gly–Pro–Yaa↓
Factor Xa	Ile–Glu–Gly–Arg↓Xaa–
Enterokinase	–Asp–Asp–Asp–Asp–Lys↓Gly–
Tobacco etch virus (TEV) protease	–Glu–Xaa–Xaa–Tyr–Xaa–Gln↓Ser(or Gly)

Xaa and Yaa indicate any amino acid residue, and the vertical arrow indicates position of cleaved peptide bond.

Expression of Fusion Proteins

When short peptides are expressed in *E. coli*, they are likely to be rapidly degraded by the various and plentiful peptidases in the bacterial cytoplasm. To protect these products, the DNA sequences coding for them are usually fused to genes that code for proteins endogenous to *E. coli*. On expression of the resulting fusion protein, the small foreign peptide is folded as a portion of the large endogenous protein and generally escapes proteolytic degradation.

Selective site-specific cleavage of the fusion protein is required to separate these peptides from the "carrier" proteins. Some of the conditions and reagents that cleave proteins at specific sites are listed in Table 3.1. When the peptides do not contain internal bonds that would be cleaved by trypsin, CNBr, or acid, it is safe to generate a cleavage site for one of these agents at the peptide–carrier junction by altering DNA sequence. Peptides that are fairly large are likely to contain sites susceptible to such simple agents; in these cases a protease, such as factor Xa, with its very stringent amino acid sequence specificity, is used to cleave at the desired site.

Another advantage of expressing peptides and proteins as fusion products is that it facilitates product purification. For example, if the foreign gene is fused with a sequence coding for an immunoglobulin G (IgG) antibody–binding domain of protein A from *Staphylococcus aureus*, the fusion protein can be recovered by simply passing the cellular extract through a column of immobilized IgG. Other schemes fuse products to glutathione S-transferase (GST), which allows purification of the product with an affinity column (Box 3.10) of immobilized glutathione, or fuse them to a stretch of histidine residues, and then purify them by exploiting the metal complexation of histidine. Creation of fusion proteins is also an important strategy for avoiding the aggregation of the expressed protein, discussed below.

Formation of Inclusion Bodies

When expressed at high levels in *E. coli* cytoplasm, many foreign proteins, especially those of eukaryotic origin, form insoluble aggregates called *inclusion bodies*. They are presumed to form where high concentrations of the overproduced, nascent proteins favor intermolecular interactions between the hydrophobic stretches of incompletely folded polypeptide chains, and they lead to aggregation and misfolding of these proteins (Figure 3.23).

These high, localized concentrations of nascent proteins are partly a consequence of the use of overexpression systems with their high gene dosage and powerful promoters. They are also partly the result of the prokaryotic structure of *E. coli*. Under the typical eukaryotic conditions of synthesis, many nascent human and animal proteins would be sequestered into compartments separated from the cytosol, such as the lumen of the endoplasmic reticulum. In *E. coli*, however, the newly synthesized proteins must remain at large in the undifferentiated bacterial cytoplasm. Furthermore, several factors tend to retard the folding of foreign proteins in *E. coli*, thus increasing the chances of intermolecular association and aggregation: (1) The conditions in the *E. coli* cytoplasm – for example, pH, ionic strength, and redox potential – are different from the normal environment in which these proteins are folded into their final conformations. Many secreted proteins of eukaryotic origin cannot fold in the highly reducing cytoplasm *of E. coli* because disulfide bonds, which are normally formed in the oxidizing environment of the endoplasmic reticulum and help the folding process, are not produced. (2) The correct folding of many polypeptides is facilitated by various helper proteins (Box 3.11). These include peptidyl-proline *cis/trans* isomerase, which facilitates the interconversion of two forms of proline groups, protein disulfide isomerase, which catalyzes the exchange of disulfide linkages in the substrate protein, thereby facilitating the production of the form with correct disulfide pairs, and a group of *molecular chaperones*, which also enhance the folding process or at least prevent the premature formation of aggregates of denatured proteins. The nature and the concentration of such helper proteins in *E. coli* obviously differ from those in the various compartments of the eukaryotic cells.

In some cases, the formation of inclusion bodies can be avoided, as described in the next section. Even when this is difficult, however, we may be able to use inclusion bodies to advantage in the purification of recombinant proteins. The cells are broken, the extracts centrifuged, and the inclusion bodies recovered as a sediment. Because the sediment also contains membrane fragments, it is customary to wash it by resuspension in detergent solutions (to dissolve and remove membrane components) and by recentrifugation. In this manner, the complex and tedious process of protein purification can be almost completely bypassed. Finally, the inclusion bodies are solubilized with protein denaturants, such as 6 M urea or 8 M guanidinium

Affinity Columns

Traditional methods of protein purification rely on gross, physicochemical properties of the proteins, such as electrical charge (in ion exchange chromatography), size (gel filtration chromatography), and hydrophobicity (hydrophobic chromatography). A single fraction obtained after such purification procedures tends to contain many proteins, if the starting material is a complex mixture. In contrast, affinity chromatography relies on *specific* interaction of the proteins with specific ligand molecules. For example, to purify an enzyme, either the substrate of the enzyme or a substrate analog is covalently linked to a granular matrix material, and the granules are packed into a column. When a crude mixture of hundreds of proteins is passed through this affinity column, only the enzyme is bound specifically to its substrate immobilized in the column; all the other proteins pass through the column unretarded. Elution of the column by procedures that decrease the affinity of the enzyme to the substrate, perhaps by altering the enzyme conformation, results in the one-step purification of the desired enzyme.

Recombinant DNA technology now allows us to put any of the many available "tags" to the protein of interest. For example, the hexahistidine tag allows purification by Ni^{2+} columns, the maltose-binding protein tag by amylose columns, the glutathione *S*-transferase (GST) tag by immobilized glutathione columns, and so forth.

BOX 3.10

Foldases and Molecular Chaperones

Christian B. Anfinsen's group showed in 1957 that a completely denatured ribonuclease A can be spontaneously renatured *in vitro* into its native conformation with the concomitant formation of its four disulfide bonds with correctly paired cysteine residues. This famous discovery was interpreted by many to imply that all proteins become folded spontaneously, without assistance from any other cellular component. However, the fact that certain reactions occur spontaneously does not mean that cells do not use helper proteins to facilitate those processes. Indeed, recent years have witnessed the discovery of two classes of proteins that assist the folding of newly made proteins.

The members of one class of such proteins possess enzyme activities in the classical sense, and are sometimes called "foldases." These include peptidyl prolyl *cis-trans* isomerase and enzymes involved in the formation and isomerization of disulfide bonds. The former enzyme helps in the folding process by facilitating the interconversion between *cis*- and *trans*-configurations of the peptide bond linking the nitrogen atom of proline and the carboxyl group of the preceding amino acid residue. Practically all of the peptide bonds in proteins have the *trans* configuration, but bonds involving proline are the exception. Spontaneous *cis-trans* isomerization of such bonds occurs slowly. Enzymes that facilitate the formation of protein disulfide bonds, and their isomerization, are important in the folding of proteins that contain such bonds. If, in the course of folding, disulfide bonds are not formed, or formed between incorrect pairs of cysteine residues, the protein is likely to become misfolded.

Proteins of the second class that assist in the folding process are molecular chaperones, which appear to perform more subtle and complex functions. The structure of these chaperones has been conserved strongly during evolution, and most of them belong to the *heat-shock proteins* – either to the Hsp70 or the Hsp60 class. (They are called heat-shock proteins because in many organisms they are overproduced when the organism experiences high temperatures; these proteins are thought to facilitate the unfolding and proper refolding of heat-denatured proteins.) In *E. coli*, most of the nascent polypeptides fold with the help of a ribosome-associated protein called Trigger Factor, which is a chaperone that also has prolyl *cis-trans* isomerase activity. The slowly folding proteins, with their exposed hydrophobic patches, are then bound by DnaK, a representative of the Hsp70 class. DnaK shields the hydrophobic patches of the nascent proteins so that they do not interact with other nascent proteins and form insoluble aggregates, or inclusion bodies. DnaK works with two other proteins, DnaJ and GrpE, so that the release of the nascent protein is timed by ATP hydrolysis. If the protein is still incompletely folded, it will be bound to DnaK again, and the cycle will be repeated. The proteins that are most difficult to fold are handled by GroEL, a representative of the Hsp60 class, also called *chaperonin*. In *E. coli*, GroEL occurs as a 14-mer in a double-doughnut configuration with two large cavities. The incompletely folded protein enters one of the cavities, which is closed by a 7-mer of an associated protein GroES. This allows the slow folding of the protein without deleterious interaction with other nascent proteins. The process is again timed by the hydrolysis of many ATP molecules.

BOX 3.11

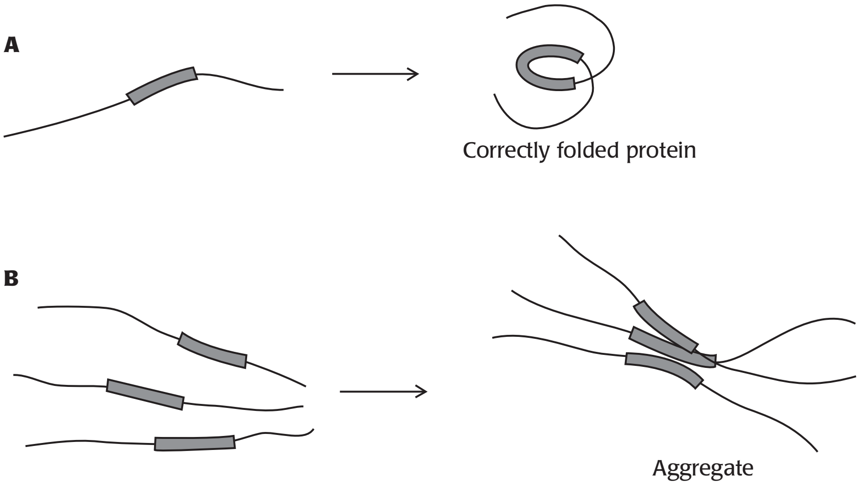


FIGURE 3.23

Presumed mechanism for the aggregation of overexpressed proteins. (**A**) Normally, nascent polypeptides fold into a globular conformation, with hydrophobic stretches (*thick line*) hidden in the interior. (**B**) However, when concentrations of nascent polypeptides are very high, there is increased likelihood that an exposed hydrophobic region on one molecule will interact with that on another molecule before the individual chains have a chance to fold properly. These intermolecular interactions between nascent chains result in aggregation and in an irreversible misfolding of the protein, producing inclusion bodies.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

hydrochloride, and the proteins are renatured by the gradual removal of denaturants (Box 3.12). Procedures of this type have been successfully used for the purification of many proteins. Although theoretical considerations indicate that *in vitro* renaturation should be done at low concentrations of the protein to minimize intermolecular interactions, in practice some systems tolerate fairly high concentrations, presumably because even these concentrations are quite low compared with those reached in overproducing cells.

Renaturation of Proteins Containing Disulfide Bonds

If the protein contains disulfide bonds, its complete unfolding requires their cleavage. This can be done by reduction (with either dithiothreitol or mercaptoethanol). In this case, the solubilized denatured protein must be purified, always in the presence of reducing agents. Alternatively, the disulfide bond can be cleaved by converting cysteine sulfur atoms into S-sulfonates with the addition of sodium sulfite. S-sulfonates are stable at neutral or acidic pH, and thus the solubilized proteins can be conveniently purified if alkalinization of the samples is avoided. After purification, S-sulfonates can be reconverted to sulfhydryl groups by the addition of mercaptoethanol. In both procedures, the renaturation of the protein is accomplished by removal of the reducing agent and of the denaturing agent, and the oxidation of the cysteine residues into disulfides is accomplished by exposure to air. More recently, successful attempts have been made to facilitate renaturation by adding foldases and even chaperones; however, the high cost of the additional proteins would constitute a major problem.

BOX 3.12

Preventing the Formation of Inclusion Bodies

Although inclusion bodies are a convenient starting material for purification, the denaturation and the controlled renaturation steps are costly. It is especially problematic that the renaturation process usually works best at low protein concentrations, a requirement that increases cost and decreases yield. Careful cost analysis shows that the expense of the renaturation step is the main reason why the commercial production of large proteins such as tissue plasminogen activator and factor VIII by recombinant DNA technology is carried out in animal cell cultures rather than in *E. coli*. In animal cell cultures, the recombinant protein folds spontaneously into the native conformation, and inclusion bodies are not formed.

Production costs for these proteins would be further reduced if they could be produced in the native conformation in a microbial host. Much effort has therefore been devoted to finding conditions that would decrease the extent of inclusion body formation in *E. coli*. So far, a technique that universally and drastically decreases inclusion body formation has not been discovered. Among the approaches tried, lowering of the growth temperature was effective in many cases. Attempts have also been made, with success, to co-express chaperones and foldases to improve the correct protein folding in *E. coli*. In some cases, folding of foreign proteins was improved if *E. coli* was grown in the presence of low concentrations of ethanol (usually 2% to 3%). A plausible explanation of this result is that ethanol induces the “heat shock response” in *E. coli*, which leads to increased production of foldases and chaperones. Another approach is to fuse the coding sequences of foreign proteins to the $3'$-terminus of genes coding for “solubilizer” proteins, such as *E. coli* thioredoxin or the mature form of maltose-binding protein. In many cases, the fused proteins were found to be produced in a totally soluble form (Box 3.13).

Secretion Vectors

Whether the recombinant proteins form inclusion bodies in the cytoplasm of the host bacterium or remain in soluble form, their purification is always a challenge. One way to simplify the task of separating the recombinant protein from the myriad of host proteins, at least in principle, is to cause the recombinant proteins to be secreted into the culture medium. After that, purification would become quite straightforward, because bacteria are usually grown in simple, protein-free media. For this reason, much effort has been spent on the construction of *secretion vectors*.

In both prokaryotes and eukaryotes, proteins destined to be secreted from the cell are synthesized with an extra sequence, a *leader* (or *signal*) *sequence*, of about two dozen residues at the N-terminus. This sequence guides the nascent protein to the secretory apparatus in the cytoplasmic membrane and is split off by leader peptidase after the polypeptide is translocated across the membrane. The presence of the leader sequence is a necessary, but not always a sufficient, condition for secretion: Some artificial constructs composed of leader sequences fused to soluble, cytosolic proteins fail to be secreted, presumably because the mature part of the protein folds quickly to a stable, globular conformation and cannot be translocated in that condition. This suggests that the secretion vector strategy will work best when the products are unlikely to fold rapidly into a tight, stable conformation.

Indeed, this strategy proved useful in the production of insulin-like growth factor I (IGF-1), a peptide composed of about 70 amino acid residues, in the early days of biotechnology. The cDNA for IGF-1 was cloned behind the sequence coding for the leader sequence of protein A, a secreted, IgG-binding protein from *S. aureus*, and the plasmid was transformed into *E. coli* HB101. In addition, two copies of the sequence coding for the IgG-binding domain of protein A were inserted between the leader sequence and the cDNA for IGF-1 to facilitate the purification and to inhibit proteolytic degradation (see page 119). An “affinity handle” such as this is important because when proteins are secreted, they must be purified from the culture supernatant, with its very large volume. The affinity handle provides a way of rapidly and efficiently concentrating the desired product (see Box 3.10). In this case, the culture supernatant was passed through a column of IgG-Sepharose, which adsorbed all of the secreted proteins containing the IgG-binding protein A sequence. The fusion protein was then cleaved with hydroxylamine by taking advantage of the hydroxylamine-sensitive Asn-Gly sequence introduced just in front of the IGF-1 sequence. The IGF-1 peptide was then purified by conventional column chromatography methods.

Although secretion vectors have often proved useful, secretion is not yet a universally applicable approach. Some proteins fail to be secreted even when fused to a leader sequence, as mentioned earlier. Unfortunately, *E. coli* cells are surrounded by the outer membrane, and thus the export from the cytoplasm results in secretion into another cellular compartment, the periplasm between outer and inner membranes. In the case mentioned above, apparently a raised temperature (44°C) needed to induce the protein A promoter also permeabilized the outer membrane, causing a large fraction of the periplasmic protein to leak out into the medium. However, this is a rare phenomenon, and *E. coli* strains that are leaky and at the same time grow in a robust manner have not yet been developed. This fact led to attempts to use Gram-positive bacteria, such as *B. subtilis*, as the host for production of recombinant proteins; however, secretion of powerful proteases by such bacteria has so far hampered this effort. In any case, periplasm has several features that are attractive for the correct folding of foreign proteins. It is

Thioredoxin and Maltose-Binding Protein Fusions

E. coli thioredoxin is a small protein (molecular weight 11,675) with two cysteine residues in close proximity. It appears to fold efficiently, since its expression at a very high level (up to 40% of the total *E. coli* protein) still does not cause the formation of inclusion bodies. When foreign genes are fused to the 3′-terminus of the thioredoxin gene, the fusion protein indeed seemed to fold much better than the foreign protein expressed alone, presumably because the initial folding of the thioredoxin domain facilitates the subsequent folding of the following foreign protein. Similarly, fusion of foreign protein with the mature sequence of *E. coli* maltose-binding protein has been used with many examples of success. In this case, it is speculated that the ligand-binding groove of the binding protein may act as a chaperone that holds the incompletely folded foreign protein.

BOX 3.13

a more oxidizing environment and contains enzymatic systems that catalyze the formation and isomerization of disulfide bonds, in contrast to the cytosol of *E. coli*, which is very strongly reducing. It also contains several proteins that function both as chaperones and peptidyl prolyl isomerases, although the classical chaperones requiring ATP, such as Hsp60 and Hsp70, are absent. Thus, it is a common observation that foreign proteins, especially secreted proteins of mammalian origin such as hormones, form inclusion bodies much less frequently when they are secreted into the periplasmic space.

AN EXAMPLE: PRODUCTION OF CHYMOSIN (RENNIN) IN *E. COLI*

Chymosin is the major protease produced in the fourth stomach (abomasum) of calves. Its production is limited to the few weeks during which the calves are nourished by milk. Chymosin is synthesized in the mucosal cells as preprochymosin (containing the "pre" signal sequence, the "pro" sequence removed at the time of activation, and the mature chymosin sequence). The signal sequence of 16 amino acid residues is removed, the protein is secreted as prochymosin (molecular weight 41,000), and this inactive zymogen becomes converted under acidic conditions into the active enzyme chymosin (molecular weight 35,600) by autocatalytic cleavage of the N-terminal "pro" sequence of 27 amino acid residues.

Chymosin is an aspartyl protease. It coagulates milk very efficiently through the limited hydrolysis of κ-casein and is used extensively in the manufacture of cheese. Because the production of cheese has increased rapidly in recent decades and the supply of suckling calves has declined, the availability of chymosin or chymosin substitutes has become an important issue in the dairy industry.

One major solution has been the commercialization of fungal enzymes from *Mucor* and *Endothia* as substitutes for chymosin. These enzymes are less expensive, but they do not quite attain the high coagulation/proteolysis ratio of calf chymosin, and this results in subtle but real differences in the flavor of the cheese. A more satisfactory solution, therefore, would be to produce chymosin by cloning, if it can be done in a cost-effective manner.

Several laboratories succeeded in cloning chymosin cDNA in the early 1980s. In every case, the original template was mRNA from the mucosa of the calf abomasum. cDNA was prepared from this mRNA, in some cases after size fractionation in order to further enrich for (pre)prochymosin mRNA. In the primary cloning step, the cDNA was inserted into *E. coli* plasmid vectors, and the recombinant plasmids were screened, for example, by probe hybridization (see Figure 3.18).

The next step was the cloning in a suitable expression vector. In the calf abomasum, chymosin is made as a preproprotein. Researchers had to decide in which form it should be expressed in *E. coli*. No attempt was made to express chymosin in its mature, processed form because the production of such an active protease in the *E. coli* cytoplasm was expected to be harmful to host cells. Nor, in the initial efforts, was an attempt made to express the entire preprochymosin sequence because of concern that the eukaryotic signal sequence might not lead to efficient secretion in *E. coli*. In several laboratories, therefore, prochymosin was chosen as the form to be expressed.

Two methods were used. In one, the prochymosin sequence was fused to the N-terminal portion of LacZ or TrpE, and the protein was expressed as the fusion protein. In this case, the prokaryotic promoters and the RBS present in front of these highly expressed prokaryotic genes were used to initiate transcription and translation efficiently. In another approach, the sequence coding for prochymosin was inserted directly behind a sequence containing a suitable prokaryotic promoter, RBS, and ATG codon. Some adjustment of distance between the RBS and ATG, as well as of the actual base sequence, was needed to optimize the expression in this case. Both approaches led to the production of prochymosin at a level corresponding to up to 5% of total *E. coli* protein.

The overproduced prochymosin, however, accumulated in a denatured form as inclusion bodies; in retrospect, this is not surprising as prochymosin contains several disulfide bonds. When attempts were made to purify the inclusion bodies and to renature prochymosin from this material by the procedure already described (pages 120), the yield of active prochymosin was disappointingly low, owing primarily to difficulties in the renaturation step. The increased production cost that would result might be tolerated if the product were a human therapeutic compound, but for agricultural products such as prochymosin, the cost was clearly prohibitive. Attempt to express prochymosin in secretion vectors also resulted in failure because parts of prochymosin apparently folded rapidly to prevent its secretion. Prochymosin has since been expressed more efficiently in yeasts (see below).

PRODUCTION OF PROTEINS IN YEAST

We have already described the cloning of foreign genes in bacteria, mostly in *E. coli*. In passing, we touched on the difficulties encountered when bacteria are used to clone and express genes from eukaryotes. For example, many eukaryotic proteins normally undergo one or more posttranslational modifications that are important to their functions or stability. Yeast has often been referred to as a model eukaryote, and in this section, we show how yeast cells are able to carry out many of the posttranslational modifications necessary to produce accurately synthesized proteins using the genes or cDNA of higher organisms.

Glycosylation – the addition of oligosaccharide units to a protein – is one of the most important posttranslational modifications that occur to the gene products of eukaryotic cells (Box 3.14). Indeed, most secreted eukaryotic proteins are glycosylated. Glycosylation often helps ensure the correct folding of proteins and protects them from proteolytic enzymes. In some cases, specific receptors on animal cells recognize serum proteins whose N-linked oligosaccharides lack certain sugars and remove these proteins (usually "old" proteins) from circulation. Thus, the presence of the correct

Posttranslational Modification of Eukaryotic Proteins

Many eukaryotic proteins, especially secreted proteins (including hormones), are glycosylated. They acquire oligosaccharide substituents at asparagine residues via an *N*-glycosidic bond during the secretion process (see Box 3.15). In higher animals, these "*N*-linked" oligosaccharides are typically of the complex, branched type, containing *N*-acetylglucosamine, mannose, galactose, and sialic acid residues. Yeast glycoproteins characteristically carry oligosaccharides containing very large numbers of mannose residues. Other oligosaccharides can be linked to serine or threonine residues via an *O*-glycosidic bond. These "*O*-linked" oligosaccharides are generally less branched than the *N*-linked ones and typically contain *N*-acetylgalactosamine, galactose, and sialic acid.

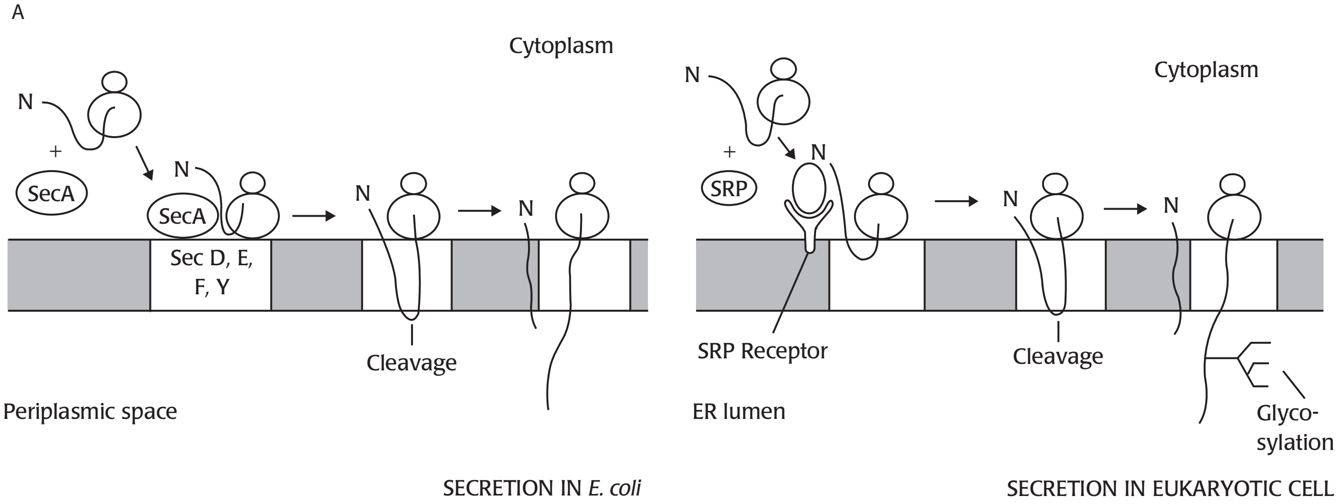
In many eukaryotic proteins, the amino group of the N-terminal amino acid residue is modified by acylation – that is, by the formation of an acyl amide linkage. *N*-acetylation interferes with recognition of the protein by the intracellular proteolytic degradation machinery and thus preserves the proteins for a longer period within the animal or human body. Another characteristic modification of the N-terminal residue is *N*-myristylation, which adds the 14-carbon, saturated fatty acid known as myristic acid onto the amino group. The myristylated proteins can bind to membranes at the fatty acid, thus becoming peripheral membrane proteins. Similar targeting of certain other proteins occurs by the covalent attachment of palmitic acid, a 16-carbon, saturated fatty acid to the sulfhydryl groups of internal (not N-terminal) cysteine residues.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

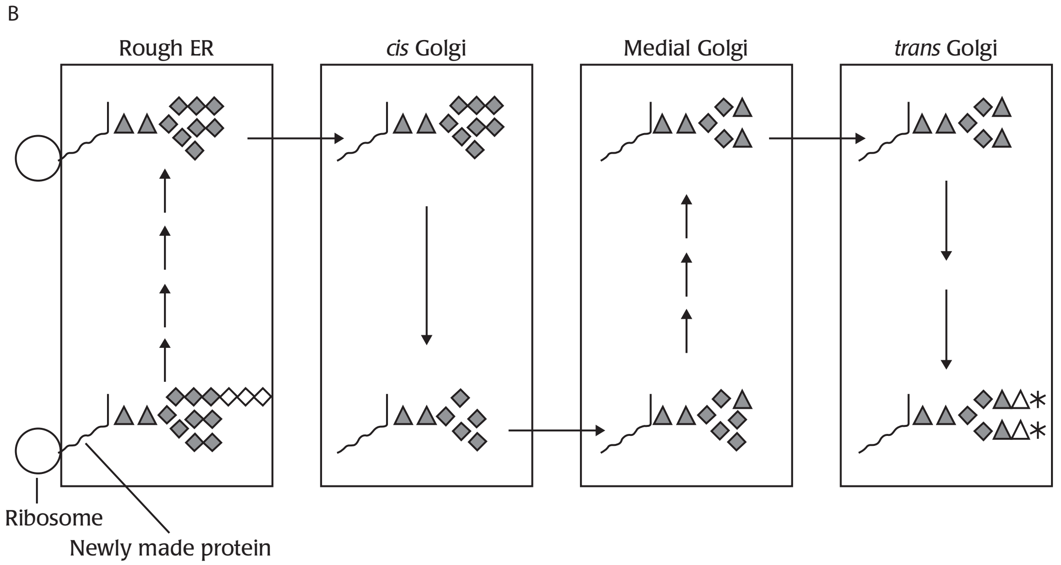
BOX 3.14

Protein-Secretion Pathways in Prokaryotic and Eukaryotic Cells

In prokaryotes, secretory proteins are made with an N-terminal signal sequence and are secreted via the SecYEG(DF) protein complex found in the plasma membrane. SecA protein is thought to help bring the signal sequence to the export machinery. The signal sequence is cleaved when the junction between it and the mature sequence appears on the outer side of the cytoplasmic membrane (see Figure A).



Secreted proteins made by eukaryotic cells also have an N-terminal signal sequence. However, the signal sequence is recognized by a complex structure, the signal-recognition particle (SRP), which contains six proteins held together by a small RNA. SRP then binds to the membrane-associated SRP receptor, thus guiding the nascent protein to the export apparatus located specifically within the membrane of the *rough endoplasmic reticulum*. One of the proteins in SRP also arrests translation until this "docking" at the export apparatus takes place, thus preventing the protein's misfolding in the cytosolic environment. The protein passes across the membrane, presumably in an extended form, and enters the lumen of the rough endoplasmic reticulum. The signal sequence is split off soon after a partial translocation of the protein across the membrane. The environment in the lumen is less reducing than the cytosol, and folding of the protein is often followed by the formation of disulfide bonds. Because the creation of disulfide bonds between "wrong" pairs of cysteine residues might produce a misfolded protein, the lumen contains disulfide isomerase, which splits and reforms disulfide bonds so as to allow the protein to reach the native conformation (see Box 3.11).



Symbols: ▲ , *N*-acetylglucosamine; ◆ , mannose; ◇ , glucose; △ , galactose; ∗ , sialic acid.

Even while the polypeptide is being extruded through the membrane, some sites within the secreted protein become glycosylated. Figure B shows the formation of a complex type of *N*-linked oligosaccharide of the simplest structure in animal cells (there can be many variations on the details of the pathway). Within the endoplasmic reticulum, a "core" oligosaccharide containing two proximal *N*-acetylglucosamine residues, nine mannose residues, and three glucose residues is attached to an appropriate site on the protein; subsequently, all of the glucose residues and one mannose residue are "trimmed off." The glycoprotein is then transported, via small membrane vesicles, into the Golgi apparatus, another complex, membrane-bounded organelle: First, it enters *cis* Golgi vesicles, where three more of the mannose residues are removed. In the next compartment, the lumen of the medial Golgi vesicles, more mannose residues are trimmed off, and two *N*-acetylglucosamine residues are added on. Finally, in the *trans* Golgi compartment, two galactose residues are added to the *N*-acetylglucosamine residues, and sialic acid residues are added onto the galactose residues. The completed glycoprotein is then secreted from the cell by the fusion of glycoprotein-containing vesicles with the plasma membrane.

BOX 3.15

oligosaccharides is very important in producing recombinant human proteins that work well and last for a long time *in vivo*. Glycosylation and other modifications described in Box 3.14 do not occur if eukaryotic genes are expressed in bacteria such as *E. coli*.

In eukaryotic cells, secretory proteins are synthesized by ribosomes associated with the membrane of the endoplasmic reticulum and are translocated across that membrane cotranslationally by a mechanism involving a *signal-recognition particle*. On entering the lumen of the endoplasmic reticulum, these proteins are immediately glycosylated (Box 3.15). By contrast, the prokaryotes' homologs of signal-recognition particles do not play a major role in the export of proteins but are involved in the insertion of cytoplasmic membrane proteins, and bacterial secretory proteins are almost never glycosylated.

Because glycosylation is of such importance in eukaryotes, it follows that eukaryotic microbes, such as yeasts, may be better hosts for the production of proteins of higher eukaryotes. Yeast cells do export many proteins using the endoplasmic reticulum–Golgi pathway, apparently with the participation of a signal-recognition particle, and glycosylate those proteins in the process. Yeast cells also carry out posttranslational *N*-acetylation and myristylation of proteins (see Box 3.14).

One might even expect that yeasts, as eukaryotes, would be able to carry out the correct splicing of nascent RNA transcripts of mammalian genes. However, yeasts contain few introns and therefore may fail to process mammalian intron sequences. The safest procedure when expression of a mammalian protein is desired is to utilize an intron-free gene – that is, to generate a cDNA copy of the mature mRNA for the gene of interest and use that as a template.

Yeasts can be grown to very high densities in simple, inexpensive media. More important, the components and metabolic products of yeast cells are not toxic to humans (remember that LPS, an integral component of the *E. coli* outer membrane, is a very toxic molecule also known as endotoxin). In the following pages, we discuss how foreign DNA is expressed in yeast cells (most often *S. cerevisiae*, or baker's yeast).

INTRODUCTION OF DNA INTO YEAST CELLS

DNA can be introduced into bacteria in a variety of ways. With yeasts, however, transformation is the only practical means of introducing DNA. In one method, the yeast cell wall is removed by enzyme digestion and the resulting "spheroplasts" (cells bounded essentially only by the cytoplasmic membrane) are incubated with DNA in the presence of Ca^{2+} and polyethyleneglycol. Both Ca^{2+} and polyethyleneglycol are agents that stimulate the membrane fusion process and thereby enhance fusion between spheroplasts. Possibly DNA is taken up by yeast cells in the process of spheroplast fusion, but it is not yet clear whether the fusion is necessary for this uptake to occur. In another method, intact yeast cells (with cell wall in place) are treated with Li^{+} ions and then incubated with DNA and polyethyleneglycol. The mechanism of DNA uptake remains obscure in this case, too. A third method is to apply transient high voltages to a suspension of cells. This process, called *electroporation*, creates transient holes in the walls and membranes, as described earlier in this chapter.

YEAST CLONING VECTORS

Several types of cloning vectors are used to manipulate recombinant DNA constructs in yeast. Most are "shuttle" vectors: vectors that can multiply in yeast as well as in *E. coli*. The reason shuttle vectors are preferred is that the basic recombinant DNA manipulations are more easily carried out in *E. coli*, but the resulting DNA constructs must be transferred to yeast to take advantage of the superior properties of this host, such as the expression of glycosylated proteins. Shuttle vectors can be moved between the two hosts because they also contain the origin of replication recognized by *E. coli* and selection markers useful in *E. coli*, as well as features that enable them to survive in yeast cells. There are five major types of yeast cloning vectors: yeast integrative plasmids (YIps), yeast replicating plasmids (YRps), yeast episomal plasmids (YEps), yeast centromeric plasmids (YCps), and yeast artificial chromosomes (YACs).

Yeast Integrative Plasmids

YIps (Figure 3.24) are essentially bacterial plasmid vectors with an added marker that makes possible their genetic selection in yeast. As we have seen already, antibiotic resistance genes are commonly used as selection markers in bacterial vectors. However, not many antibiotics are effective against yeasts. Thus, selection procedures in yeast are commonly designed to utilize a host strain that is defective in the biosynthesis of amino acids, purines, or pyrimidines and a vector that contains a yeast gene for the missing function. Some commonly used nutritional markers are the yeast genes *LEU2* (a gene involved in leucine biosynthesis), *URA3* (a gene involved in uracil biosynthesis), and *HIS3* (a gene involved in histidine biosynthesis). For example, the *leu2* host strain (in yeast genetics, a mutated and, hence, usually functionally defective allele is denoted in lower-case letters) will not grow in a minimal medium – that is, a medium containing only the requisite minerals

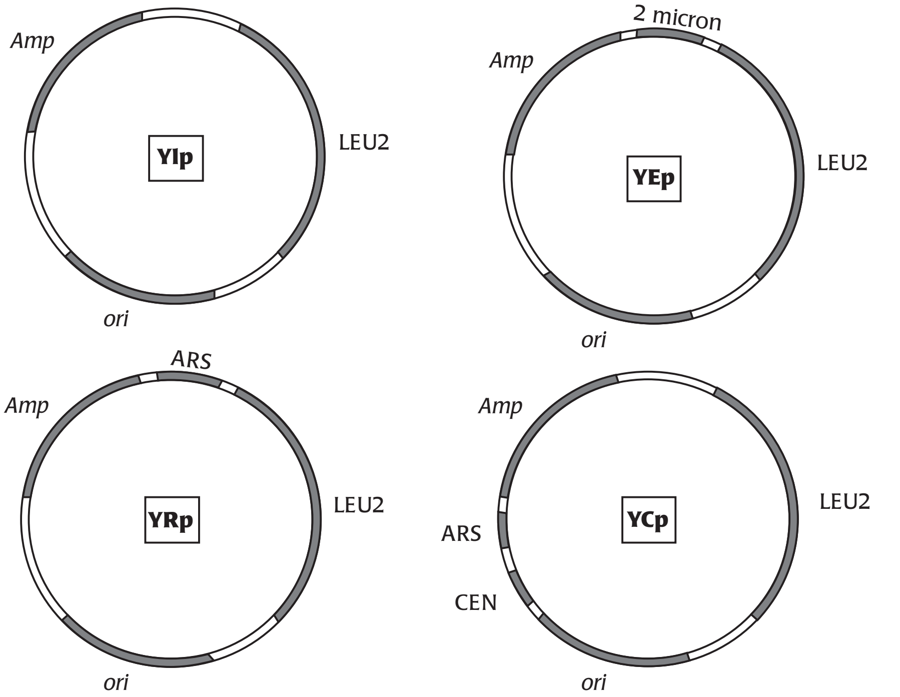


FIGURE 3.24

Plasmid vectors useful for cloning in yeast. Examples of four types of vectors are shown. Abbreviations: *ori, E. coli* origin of replication; Amp, ampicillin resistance gene for selection in *E. coli*; *LEU2*, a gene involved in leucine biosynthesis, for selection in yeast. Regions controlling replication and segregation in yeast, such as ARS, CEN, and 2-μm plasmid sequence, are described in the text.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

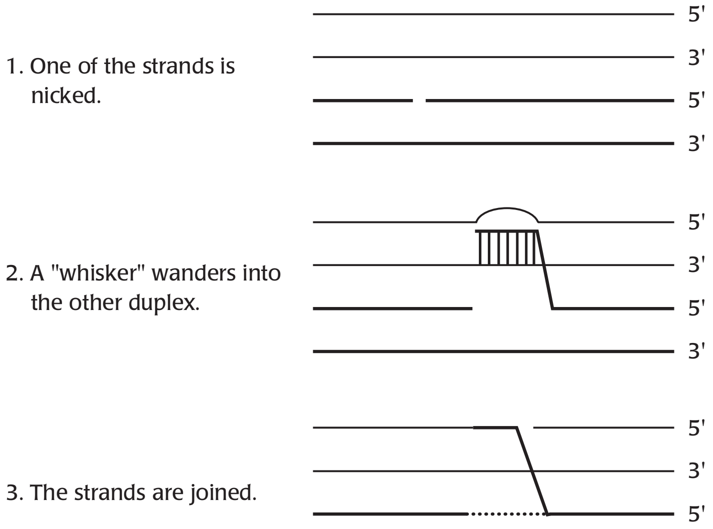
and a carbon source – but the same strain harboring an *LEU2*-containing plasmid will grow because it is able to synthesize leucine.

YIps lack an origin for replication that can be recognized by the yeast DNA synthesis machinery. Therefore, they can be maintained in yeast cells only when they become integrated into a yeast chromosome (usually by homologous recombination at the site of the yeast marker gene or one of the other yeast sequences present in the vector). Once integrated, they are inherited quite stably as a part of the yeast genome. However, integration is a rare event, so the frequency of transformation with plasmids of this type is extremely low (one to 100 transformants/μg DNA compared with the 100,000 transformants/μg that can be obtained with *E. coli*). The frequency of integration can be enhanced somewhat by cutting the plasmid within the region of yeast homology, a procedure that promotes homologous recombination to some degree (Box 3.16). Another drawback of these plasmids is their low copy number. Usually only one copy at most is integrated in one haploid yeast cell, effectively limiting the level of expression of the cloned gene. One way to circumvent this problem is to design the plasmid to integrate into genes that exist in multiple copies in the yeast chromosomes. For example, there are more than 100 copies of the genes coding for rRNA in a yeast cell, so multiple integrations into these sites could create a cell genome with many copies of the cloned genes. Alternatively, one can use as the selectable marker on the YIp vector a gene that has to exist in a large number of copies for the yeast to survive under certain conditions. For example, if the plasmid contains the *CUP1* gene, which codes for metallothionein, a protein that protects yeast cells by binding to heavy metals, yeast cells will survive in a medium containing Cu^{2+} only when a large number of copies of the *CUP1* have been integrated into the genome – that is, when the gene has become “amplified.” YIps are quite useful in spite of their typically low copy

Homologous Recombination Process

Homologous recombination begins with alignment of the homologous regions in two parental DNA duplexes. This is followed by "nicking" (single-stranded cleavage) of one parent-DNA helix and generation of single-stranded "whiskers" (step 1). A whisker wanders into the other duplex and forms Watson–Crick pairs with the complementary strand of the other DNA duplex (step 2). Finally, the end of the whisker is joined covalently to one of the strands of the other duplex, completing the process of crossing over (step 3). This mechanism requires an initial cut in one or both strands. Consequently, using plasmids that are already cut in the region of homology increases the frequency of recombination. More comprehensive schemes for the entire recombination pathway have been proposed [Orr-Weaver, T. L, Szostak, J. W., and Rothstein, R. J. (1981). Yeast transformation: a model system for the study of recombination. *Proceedings of the National Academy of Sciences U.S.A.*, 78, 6354–6358].

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.


1. One of the strands is nicked.
5'
3'
5'
3'
2. A "whisker" wanders into the other duplex.
5'
3'
5'
3'
3. The strands are joined.
5'
3'
5'

BOX 3.16

number because plasmid stability can often become a major problem with other kinds of yeast vectors (see below).

Yeast Replicating Plasmids

In addition to selection markers useful in yeast, YRps (Figure 3.24) contain an origin of replication derived from the yeast chromosome and termed *ARS* (autonomously replicating sequence). With this origin, the plasmids can replicate without having to be integrated into the chromosome. However, yeast cells divide unequally by budding. In the process, only a disproportionately small fraction of the plasmids that were present in the mother cell are partitioned off into the buds, and many of the progeny cells are likely to lack plasmids entirely. Thus, YRp plasmids are lost rapidly unless constant selection pressures are applied. Consequently, they are not very useful for the reproducible expression of cloned genes.

Yeast Episomal Plasmids

Some strains of *S. cerevisiae* contain an endogenous, autonomously replicating, high-copy-number plasmid called 2-μm plasmid. The origin of this plasmid is added to YIps to produce YEps (Figure 3.24), which can exist in high copy numbers (30 to 50 copies/cell). (An *episome* is a genetic element that can exist either free – as a plasmid – or as a part of the cellular chromosome.) Like YRps, YEps are poorly segregated into daughter cells, but they are maintained more stably because of their higher copy number. If the entire 2-μm DNA (6.3 kb) is inserted into a YIp plasmid and introduced into yeast cells that lack an endogenous 2-μm plasmid, copy numbers in excess of 200 per cell can be achieved under certain conditions. Plasmids of this type are obviously most suitable when high-level expression of a foreign gene is desired. One vector of this type, pJDB219, contains an intact *LEU2* gene but not its promoter. Because of the lack of promoter, this construction, called *leu2-d*, does not produce the full-scale expression of the LEU2 protein. Nevertheless, an extremely low level of expression does occur, presumably because of the nonspecific binding of the RNA polymerase or a very weak "readthrough" from upstream genes. This low-level expression produces a detectable phenotype because even a small amount of the enzyme is enough to produce some leucine. But because the amount of the enzyme produced by a single plasmid is far from sufficient, the plasmid-carrying cells cannot grow at a reasonable rate in minimal medium unless the plasmid is present in very large numbers (200 to 300 per cell) in order to complement the completely defective *leu2*$^{-}$ allele of the host. In other words, this plasmid is designed so that its presence in high copy numbers will be favored.

Yeast Centromeric Plasmids

YCps (Figure 3.24) are YRps, or sometimes YEps, in which the sequence of a yeast centromere has been inserted. The centromeric sequence allows these plasmids to behave like regular chromosomes during the mitotic cell division, so YCps are faithfully distributed to daughter cells and are highly stable even without maintenance by selection. However, the "chromosomelike" behavior of these plasmids also means that their copy number is kept very low (one to three per haploid cell). This is a potential disadvantage when the plasmids are used for the expression of cloned genes, although the expression can be increased by the use of highly inducible promoters.

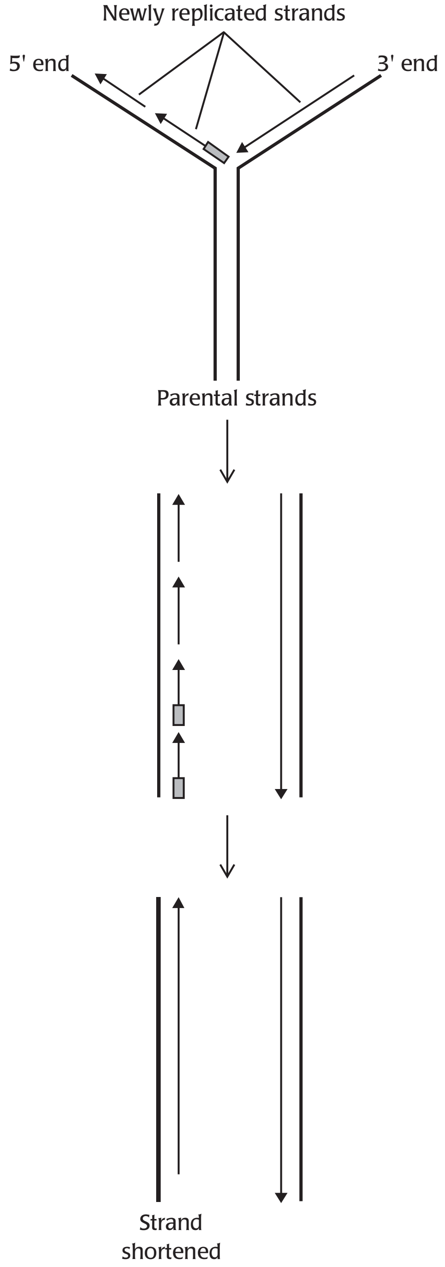
Yeast Artificial Chromosomes

YACs are linear plasmids containing an ARS, a centromeric sequence, and, most important, a telomere (Box 3.17) at each end (Figure 3.25). These features allow the plasmids to behave exactly like chromosomes. Because the plasmid is linear, there is no limit to the amount of foreign DNA that can be cloned into it. This is the most important feature of YACs. Animal genes contain many introns and can exceed 100 kb in size. Such genes cannot be cloned in a single vector except in YACs (and BACs, which we

Telomeres

The ends of linear DNA duplexes, such as chromosomes and YACs, cannot be replicated faithfully because DNA synthesis always occurs in the 5′-to-3′ direction: One of the strands is thus replicated in short segments (Okazaki fragments) formed by the elongation of RNA primers (rectangles). When the RNA primer that is complementary to the very end of the DNA is degraded, there is no mechanism for synthesizing DNA to replace it. Consequently, such linear DNAs become shorter with each replication cycle. Eukaryotic chromosomes solve the problem by having repeated oligonucleotide sequences (telomeres) at their ends (e.g., telomeres in human chromosomes have the sequence $[\mathrm{TTAGGG}]_n$). When telomeres become too short, they are elongated by an enzymatic mechanism that does not require a template.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.



BOX 3.17

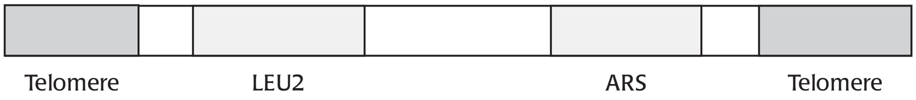


FIGURE 3.25

An example of a YAC vector. *LEU2* gene is needed for selection in yeast, ARS for replication in yeast, and telomeres for stability in yeast.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

discussed earlier in this chapter). YACs, however, are not the first choice when the main objective is a high level of expression of foreign genes.

ENHANCING THE EXPRESSION OF FOREIGN GENES IN YEAST

There are several points to consider when designing a system for expressing foreign-gene products in yeast cells.

Plasmid Copy Number

A high-copy-number plasmid of the YEp class is the best choice for maximal expression of any cloned gene. However, the expression of foreign proteins is often toxic for the yeast cells. Probably one of the main reasons is that some foreign proteins are likely to misfold in the cytoplasm, sequestering many of the chaperone molecules needed for the correct folding and functioning of the yeast's own proteins. In this situation, low-copy-number YEps and, paradoxically, even YIps may produce higher sustainable yields than high-copy-number plasmids.

Another problem that may be important in commercial production runs is the instability of some of the plasmids. In commercial fermentation, the organism must last through a far higher number of generations than is usually necessary in a small, laboratory-scale experiment. Thus, even a moderate degree of plasmid instability can cause a major problem.

Promoter Sequence

Because promoters of foreign origin are unlikely to be expressed efficiently in yeast cells, the coding sequence of a foreign gene is usually inserted behind a strong yeast promoter. Yeast promoters are quite different from bacterial promoters. Although both contain AT-rich recognition sequences for RNA polymerase (typically TATATAA for yeast, in contrast to the TATAAT consensus sequence [see Box 3.9] for *E. coli*), the "TATA sequence" in yeast is located much further upstream (40 to 120 bases) from the mRNA initiation site than it is in *E. coli*, in which the "TATAAT," or Pribnow box, is typically located only 10 bases upstream of the transcription initiation site. In addition, yeast promoters usually require an upstream activator sequence (UAS), an enhancer-like sequence located very far upstream (100 to 1000 bases) from the transcription initiation site. Because of the location of the UAS, most yeast expression vectors contain a long, native "promoter sequence" (typically around 1 kb). Two frequently used promoters are the upstream sequences for an alcohol dehydrogenase gene (*ADH1*) and for a triose phosphate dehydrogenase gene (*TDH3*). *ADH1* was thought to be expressed constitutively at a high level, and its use was popular at one time. However, we now know that this particular isozyme of alcohol dehydrogenase becomes repressed when the culture reaches a high density, so its use has fallen off. (In contrast, another isozyme of alcohol dehydrogenase, *ADH2*, becomes derepressed when glucose in the medium becomes exhausted. The promoter for this enzyme is often used as a regulatable promoter, as we shall see.)

If the expression of the foreign protein inhibits the growth of the yeast cells, it becomes necessary to use regulatable promoters and to initiate expression of the foreign genes only when the culture has reached a high density. For example, the genes involved in galactose catabolism, *GAL1*, *GAL7*, and *GAL10*, have been extensively used as sources of regulatable promoters for cloned genes because they are repressed in the presence of glucose but are induced by the addition of galactose to the medium. The regulation of these genes involves the binding of a positive activator, *GAL4*, to upstream sequences of *GAL1*, *GAL7*, and *GAL10*. Thus, if the recombinant DNA containing the latter genes exists in multiple copies in a cell and *GAL4* is expressed from a single copy of the gene on the chromosome, the *GAL4* protein in the cell might become exhausted by binding before all the recombinant genes are activated. However, this limitation can be removed if *GAL4* is also introduced into the vector so that multiple copies of *GAL4* are present in a cell. A drawback of this system is that it tends to increase the expression of the cloned gene even in the absence of galactose, so it is dangerous if the product is toxic to yeast cells. Other regulatable promoters that have been used include ADH2 (alcohol dehydrogenase regulated by ethanol and glucose) and PHO5 (acid phosphatase, regulated by phosphate). Another attractive regulatable promoter is the one for CUP-1, which codes for metallothionein, a Cu^{2+}-binding protein, and which is induced by the addition of metal ions such as Cu^{2+} or Zn^{2+} to the medium. Several systems that are induced by elevated temperatures have been used successfully in the laboratory, but it may be difficult to get the temperature to change fast enough in a large fermentation tank.

Several hybrid promoters have also been used. These contain (1) a UAS from a regulatable promoter for controlling the level of expression of the gene and (2) the TATA box region from a strong, constitutive promoter for increasing the maximal level of expression. For example, a hybrid promoter containing the UAS sequence of ADH2 and the downstream sequences (containing the TATA box) from the TDH3 promoter has been effective in producing some foreign proteins at levels sometimes exceeding 10% of the total yeast protein.

Transcription Termination and Polyadenylation of mRNA

In higher animals, termination of transcription usually occurs very far downstream from the coding sequence. Following termination, the nascent RNA transcript is cleaved at or near the cleavage signal, AAUAAA, present hundreds of nucleotides upstream from the transcript's 3′-terminus. This newly exposed 3′-terminus is then polyadenylated – that is, a stretch of A is added. In yeast, this processing follows a rather different pattern, with polyadenylation apparently occurring quite close to the 3′-end of the transcript. Unfortunately, the precise structure of a yeast terminator sequence is still not very clear. Because of this uncertainty, when recombinant DNA is constructed for gene expression in yeast, a large segment of “terminator” sequence, taken from the downstream sequence of yeast genes whose transcription is terminated efficiently, is usually placed downstream of the gene to be expressed.

Stability of mRNA

Yeast mRNAs differ greatly in their stability. The sequences that determine their degradation rates have been located in the 3′ untranslated regions and the coding regions of mRNA, but this knowledge is difficult to use for increasing the stability of foreign-gene transcripts in yeast.

If the gene is not followed by an effective terminator sequence, this usually produces unstable mRNA, presumably because it lacks the proper 3′-end that could become protected by polyadenylation. Repeated experiments have shown that such a situation drastically decreases the yield of foreign proteins in yeast. Thus, it is desirable to clone an efficient yeast terminator sequence downstream from the coding sequence of the foreign gene to be expressed, as described above.

Recognition of the AUG Initiation Codon

Efficient synthesis of mRNA, though necessary, is not sufficient in itself to ensure high-level production of a protein. Another basic requirement is efficient translation, for which the correct AUG codon must be readily recognized by initiation factors and by the ribosome machinery. In bacteria, this involves pairing of the Shine–Dalgarno sequence with the complementary sequence in 16S rRNA. There is no equivalent recognition sequence in eukaryotes, but the efficiency of translation initiation is known to depend on the sequences surrounding the AUG codon (such surrounding sequences are often called *context*). Analysis of gene sequences in yeast has shown that the consensus sequence (see Box 3.9) of the context is AxxAUGG (this is called Kozak’s rule).

If the 5′ untranslated segment of the mRNA tends to form base-paired loops, translation can be inhibited quite significantly. G occurs infrequently (taking about 5% of the positions) among the 20 to 40 bases immediately preceding the AUG codon; the presence of a large number of G residues in this region is known to inhibit translation initiation. These facts should be taken into consideration when designing the part of the DNA sequence that codes for the 5′-terminal portion of the mRNA.

Elongation of the Polypeptide

Foreign genes often contain codons that are rarely used by yeast, and this may slow the translation process. Strings of rare codons occurring close together are especially detrimental. To enhance the expression of foreign genes in such cases, codons that yeast prefers have been substituted for those rarely used by yeast, usually by the site-directed mutagenesis procedure. The preferred yeast codons can be determined by analyzing the codons used in endogenous genes that are continuously expressed at high levels, such as genes for yeast glycolytic enzymes.

Folding of the Foreign Protein

Many foreign proteins have been expressed at a high level in yeast cells and have been shown to fold correctly. For example, the hepatitis B virus core protein, the P-28–1 protective antigen of the schistosome, and human superoxide dismutase have all been expressed to a level that corresponds to 20% to 40% of total yeast protein, and yet the proteins do not appear to misfold or to form intracellular aggregates. This is in a striking contrast to the nearly ubiquitous formation of aggregates or inclusion bodies when foreign proteins are expressed in the cytoplasm of bacteria (see the earlier part of this chapter). Although the formation of such aggregates has been reported in yeast, they are not found nearly so frequently. This could be the result of the presence in the yeast cell cytoplasm of many kinds of molecular chaperones (the so-called heat-shock proteins).

Proteolysis

Many proteins in eukaryotic cells are subject to degradation by the ubiquitin pathway. These proteins have at their N-terminus certain amino acids that are recognized by a small protein, called ubiquitin, that tags them for proteolytic degradation. All eukaryotic proteins are translated with methionine at the N-terminus, but subsequent removal of the N-terminal amino acid residues may expose one of the “destabilizing” amino acids and lead to destruction of the protein. If this problem exists in a cloning situation, one way of solving it is to alter the N-terminal amino acid sequence. Another recourse is to fuse the protein to the N-terminal segment of another protein, preferably of yeast origin, that is known not to be degraded by this pathway.

Glycosylation

Animal proteins secreted through the endoplasmic reticulum–Golgi pathway (see Box 3.15) are usually glycosylated in the process. As we have noted, this may help the proteins fold correctly and make them more resistant to proteases. Such posttranslational modifications do occur to foreign proteins cloned in yeast cells (if the proteins successfully enter the secretion pathway; see also below), but the yeast system can add only the high-mannose type of oligosaccharides, not the complex type (see Figure B in Box 3.15) most common in the glycoproteins of higher animals. Sometimes this may affect the folding and protease sensitivity of the protein and, more important, the half-life of the protein *in vivo*. However, genes involved in the production of mammalian-type complex oligosaccharides have been successfully expressed in *Pichia pastoris*, and one can now produce glycoproteins with complex type side chains in yeasts.

EXAMPLE: HEPATITIS B VIRUS SURFACE ANTIGEN

The commercial production of the hepatitis B virus surface antigen (HBsAg) in yeast, a process that led to the first recombinant DNA vaccine licensed in the United States (Chapter 5), illustrates several of the features we have discussed.

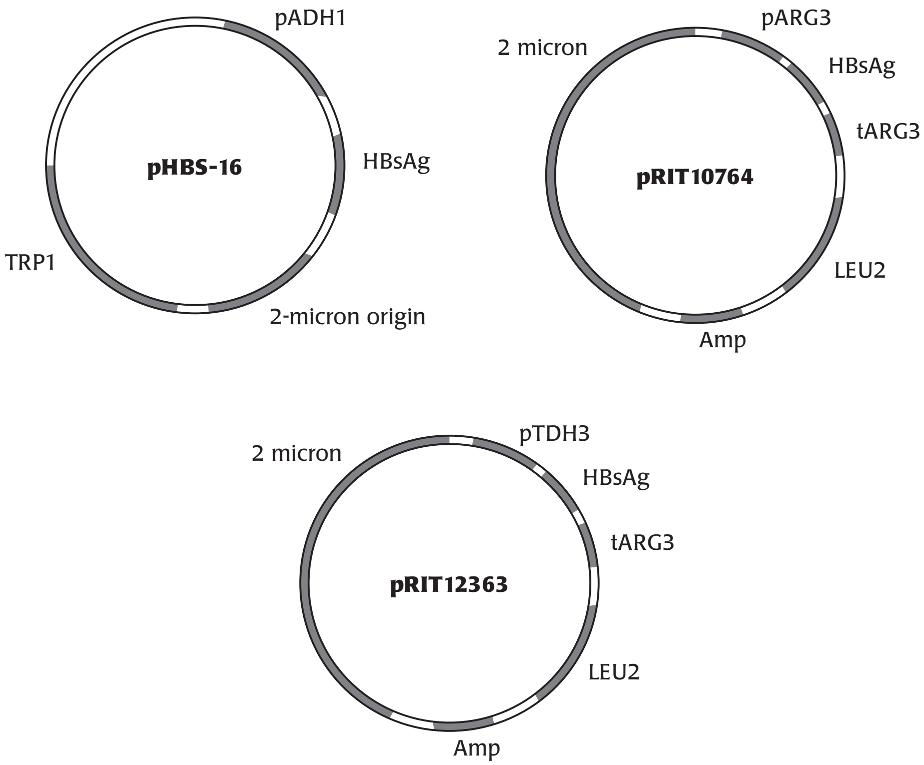


FIGURE 3.26

Recombinant plasmids used for the production of HBsAg. The "first-generation" plasmids include pHBS16 and pRIT10764. These were further developed for commercial use by Merck, Sharpe & Dohme and Smith Kline–RIT, respectively. pRIT12363 is an improved expression plasmid said to be in use at Smith Kline–RIT. Here, *p* indicates a promoter sequence, and *t* denotes a terminator sequence. *TRP1* in pHBS16 is a yeast gene coding for an enzyme of the tryptophan biosynthetic pathway and serves as the selective marker in yeast.

Redrawn based on artwork from the first edition (1995), published by W.H. Freeman.

HBsAg is a major component of the envelope of the hepatitis B virus, and immunization with this protein was known to confer good protection against viral infection (such a substance is called a protective antigen). The coding sequence for this 226-residue protein was identified on the virus genome, and it was successfully inserted into YEp-type yeast cloning vectors in several laboratories in the early 1980s (Figure 3.26). Remarkably, HBsAg folded correctly in yeast and became assembled in the form of empty envelopes, or "22-nm particles," making the subsequent purification somewhat easier. Several years later, the production of HBsAg was commercialized by two companies, Merck, Sharpe & Dohme in the United States and Smith Kline–RIT in Belgium.

S. cerevisiae strains transformed with the first-generation recombinant plasmids produced only small amounts of HBsAg. For example, pHBS-16 (see Figure 3.26), the first plasmid reported to produce HBsAg in yeast, made no more than 25 μg of HBsAg per liter of culture. The subsequent development that led from this plasmid to establishment of the commercial production process at Merck, Sharpe & Dohme is unfortunately not documented in detail in the open literature. However, some of the improvements carried out at Smith Kline–RIT have been documented, so we can get a glimpse of what they entailed.

Plasmid Copy Number. YEp vectors appear to be the most suitable for high-level expression because of their high copy numbers, and they were used in the production of HBsAg. As we have said, it is possible to increase the copy number of the plasmids by replacing LEU2 with the promoterless *leu2-d* so that only cells containing hundreds of copies of the plasmid can make enough leucine to survive. The Smith Kline–RIT group tried such an "improved" vector for HBsAg production but found that the cells rapidly lost the capacity to make HBsAg. It seems likely that the cells were losing the portion of the plasmid that coded for HBsAg. After all, when HBsAg is constitutively expressed (see below), a high level of this foreign protein is likely to be deleterious to the growth of the host cell. Therefore, progeny cells that inherit the *leu2-d*–containing part of the plasmid (which is essential for growth) but fail to inherit the gene for HBsAg are more likely to flourish. This example shows that one cannot blindly apply methods that are supposed to work better without testing and taking numerous factors into account. Production of foreign proteins is rarely neutral for the host, and one should always be alert to their possible toxic effects.

Promoter Sequence. In the first-generation plasmids pHBS16 (Merck) and pRIT10764 (Smith Kline–RIT), the promoter sequences came from the alcohol dehydrogenase (*ADH1*) and ornithine carbamoyl-transferase (*ARG3*) genes, respectively (see Figure 3.26). In the improved production strains used at both Merck and Smith Kline–RIT, the promoter comes from the gene for glyceraldehyde 3-phosphate dehydrogenase (*TDH3*). The TDH3 promoter is especially powerful, as one might have predicted from the fact that the dehydrogenase expressed from this promoter constitutes 5% of the total yeast protein. Clearly, use of the TDH3 promoter was advantageous.

As mentioned above, ADH1 was later found to become repressed toward the end of the exponential growth phase, so TDH3 remained preferable. In pRIT10764, the scientists chose a host strain with a leaky mutation in arginine biosynthesis so that the cells would be starved for arginine and so that the expression of *ARG3*, a gene involved in arginine synthesis, could be sustained at a high level (for regulation of amino acid biosynthetic genes; see Chapter 9). However, the paucity of arginine slowed the growth of the culture, again creating a less favorable situation for commercial fermentation. In these cases, there are rational explanations why the use of TDH3 promoter was preferable, but we must emphasize that in general, it is difficult to predict the levels of expression of foreign genes from the levels of expression of the endogenous yeast genes. There are many reasons why foreign genes may not be expressed as efficiently as host genes: instability of the mRNA, possible effects of the untranslated 5′ sequences of mRNA on the efficiency of translation initiation (see the next section), and the possibility that the coding sequences of the yeast genes contained enhancerlike sequences that were absent in the cloned foreign gene.

Transcription Termination and Polyadenylation of mRNA. Of the first-generation plasmids pHBS16 and pRIT10764, the latter was reported to produce a higher yield of HBsAg – about 200 μg/L of culture compared to the reported yield of less than 25 μg/L for pHBS16. Although much of this difference could be due to trivial factors such as the different quantitation methods used in different laboratories, there is an obvious difference between the two plasmids that could have contributed significantly to the higher yield of

HBsAg in strains containing pRIT10764. In this plasmid, the HBsAg sequence is followed by a terminator sequence taken from the downstream sequence of the *ARG3* gene, whereas no special terminator sequence is present in pHBS16.

Recognition of the AUG Initiation Codon. Another difference between pHBS16 and pRIT10764 is the relative content of G residues directly upstream of the AUG codon. We have noted that large numbers of G residues in this region inhibit the initiation of translation. Of 25 bases in this region in pHBS16, nine are G residues (36%), far more than the proportion found in native yeast promoter sequences. In contrast, pRIT10764, which produced a higher reported yield, contains only three G residues, well within the range found in native yeast promoters.

Glycosylation, Folding, and Acetylation. HBsAg made in human cells is *N*-glycosylated. This suggests that it is exported to the cell surface via the endoplasmic reticulum–Golgi pathway, in spite of the fact that there is no typical, cleaved, signal sequence at its N-terminus. When HBsAg is made in yeast cells, it is not glycosylated, and the protein accumulates in the cytoplasm without entering the endoplasmic–Golgi pathway. Perhaps the cloned sequence is incomplete. In the hepatitis B virus, the HBsAg sequence is preceded by an upstream "preS" sequence. Transcription in human cells may start at the preS sequence, which may contain the export signal. (Analysis of RNA transcripts is difficult with hepatitis B virus, because it cannot be grown in cultured cells.) When the HBsAg sequence is cloned and expressed together with the upstream extension, the product *is* glycosylated in yeast, a result that is consistent with the hypothesis that the preS sequence contains the export signal.

Despite the lack of glycosylation, HBsAg obviously folds correctly. This and the assembly of the protein into 22-nm particles presumably are important in achieving the desired overproduction of the antigen; if HBsAg were folded incorrectly to produce inclusion bodies in the cytoplasm, this would tie up foldases and chaperones that are needed for the folding of essential proteins of yeast, thereby interfering with the growth of the host cells.

The N-terminus of HBsAg becomes acetylated when produced in human cells; in yeast, at least a fraction of the HBsAg molecules become acetylated.

Fermentation Conditions. Some seemingly minor improvements in the fermentation conditions can have major effects on the yield. With the Smith Kline–RIT strains, the initial recombinant plasmid pRIT10764 was reported to produce HBsAg to a level of 0.06% of total yeast protein. Two years later, investigators in the same company reported a yield of 0.4% with the identical plasmid – an improvement presumably caused by a fine-tuning of culture conditions. However, the use of the ARG3 promoter still limited the growth of yeast cells to about 1 g/L. In pRIT12363, which was used for the production strain, use of the TDH3 promoter increased expression to about 1% of the yeast cell protein. However, the major improvement seems to have resulted from the fact that whereas pRIT10764 necessitated the use of leaky arginine biosynthesis mutants as the host, with pRIT12363, prototrophic strains could be used as the host, resulting in a much higher final density of yeast cells: about 60 to 70 g/L.

EXPRESSION OF FOREIGN-GENE PRODUCTS IN A SECRETED FORM

As with bacterial hosts, it is advantageous in many ways if the yeast cell secretes the foreign-gene products into the medium. First, because *S. cerevisiae* does not naturally produce many extracellular proteins, purification of the products is much simpler: One does not have to start from a mixture containing thousands of other cytoplasmic proteins. Second, the secreted protein goes through the endoplasmic reticulum–Golgi pathway, where disulfide bonds – and hence, a stable protein – may be formed under optimal conditions with the help of protein disulfide isomerase, which is present in the lumen of the endoplasmic reticulum. Indeed, α-interferon secreted by yeast cells has been shown to have disulfide bonds at the same positions as α-interferon made by human cells. In contrast, when the same protein is made in the cytoplasm, a large fraction of it appears to become misfolded. Third, the proteins may become glycosylated during their passage through the endoplasmic reticulum and Golgi apparatus. Fourth, hormone precursors may be made into mature products by processing proteases during their secretion by yeast.

Proteins are brought into the endoplasmic reticulum–Golgi pathway when the components of the secretory pathway recognize their signal sequence (see Box 3.15). It is possible to design a recombinant plasmid so that the protein will enter this pathway, by fusing the DNA coding for an effective signal sequence to the coding sequence for the protein. Secretion vectors, which already contain DNA segments coding for the signal sequence, are useful in producing such recombinant plasmids. Signal sequences for secreted invertase (SUC2) and for secreted acid phosphatase (PHO5) have been used in this way and have resulted in the successful secretion of several animal proteins of interest. In some cases, however, a large fraction of the secreted foreign protein remains trapped within the yeast cell wall. This is reminiscent of certain secreted yeast proteins that do not seem to become freely dispersed in the culture medium. We can release these proteins into the medium by digestion of the cell wall, so it is clear that they are not anchored to the plasma membrane; rather, they appear to be trapped in the space between the cytoplasmic membrane and the cell wall. This problem has led to the exploration of yeast mechanisms that produce the genuine secretion of peptides into the surrounding medium.

In *S. cerevisiae*, one such system produces and excretes the mating factor α, a 13-residue peptide. The immediate product of its structural gene, MFα1, is a 165-residue polypeptide containing an N-terminal signal sequence and four copies of the α-factor sequence. The α-factor sequences are separated by a spacer with the sequence Lys-Arg-(Glu-Ala)$_n$, ($n = 2$ or 3) (Figure 3.27). After cleavage of the signal sequence in the lumen of the endoplasmic reticulum, the polypeptide undergoes further proteolytic processing in the later

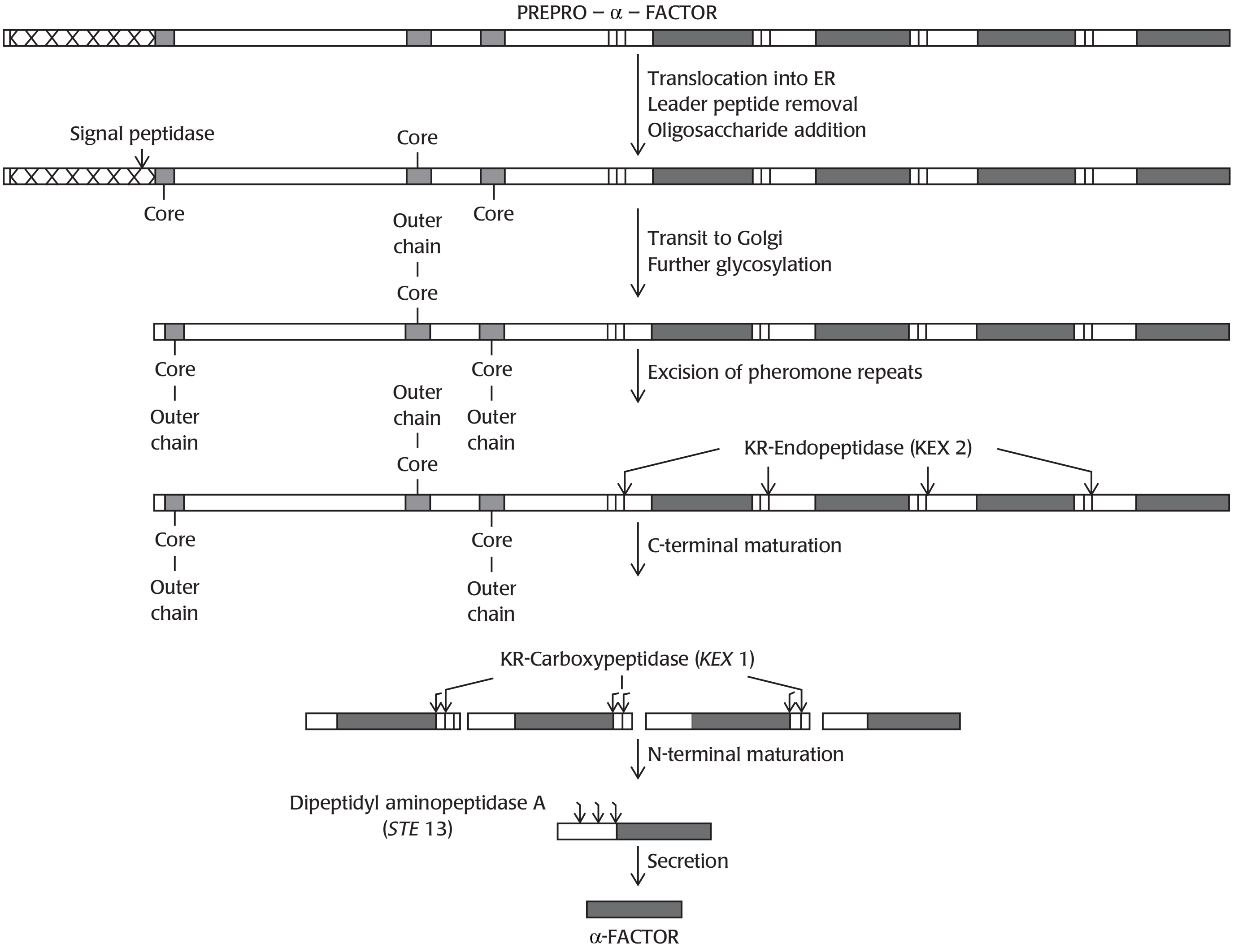


FIGURE 3.27

The processing of α-factor within the secretion pathway. [From Fuller, R. S., Sterne, R. E., and Thorner, J. (1988). Enzymes required for yeast prohormone processing. *Annual Review of Physiology*, 50, 345–362; with permission from the Annual Reviews, Inc.]

stages of the secretion process. First, KEX2 protease cleaves the bond after the Lys-Arg sequence. Then the peptide is shortened from both ends, KEX1 carboxypeptidase removing the Arg and Lys residues from the C-termini and STE13 dipeptidyl aminopeptidase removing Glu-Ala units from the N-termini. This complex processing scheme may prove useful to biotechnologists because it may afford them some flexibility in the design of fusion joints. Genes coding for animal and plant proteins have been fused to the N-terminal "prepro" portion of the *MFα1* gene, and successful secretion of products was observed in many cases. Furthermore, this method has now become a standard approach in the secretion of massive amounts of proteins by the use of "nonconventional" yeast species, and secretion of human proinsulin up to the level of 1.5 g/L has been reported.

In yeast, the Asn-linked core oligosaccharides that are attached to the *N*-glycosylation sites of foreign proteins sometimes become extended into large outer chains, producing high-mannose-type oligosaccharides. These enormous oligosaccharides may impair the proper folding and functioning of the animal-derived proteins. It may therefore be advantageous to use mutants, such as *mnn9*, that are defective in the addition of outer-chain mannose residues. For example, human α_1-antitrypsin, secreted from a *S. cerevisiae mnn9* mutant, carries three N-linked oligosaccharides similar in size to those attached to the human protein. Finally, there is recent progress in efforts to produce mammalian-type glycosylation in yeasts (p. 136).

One general problem with yeast secretion systems is low yield. However, screening of mutagenized yeast cells that contain secretion plasmids has produced high-secretion mutants. In one case, the combination of two mutations produced a strain that secreted 80% of the prochymosin synthesized. An alternative strategy is the use of nonconventional yeast species that are highly efficient in protein secretion (see below).

EXPRESSION OF PROCHYMOSIN IN YEAST

As described earlier in this chapter, the expression of calf prochymosin in the cytoplasm of *E. coli* resulted in the formation of inclusion bodies. It was thought that this problem might be overcome by expression of the protein in yeast cells because inclusion bodies are formed less frequently in yeast. The prochymosin gene was cloned into several YEp-type expression plasmids behind effective yeast promoters. However, the accumulated product was largely insoluble when overproduced.

Better results were expected with the yeast secretion vectors because the protein would then be secreted through the endoplasmic reticulum–Golgi pathway, which is similar to that in animal cells. The recombinant plasmid that was tested contained (1) a strong yeast promoter, such as the one for the phosphoglycerate kinase gene, (2) the DNA sequence coding for the signal sequence and several of the N-terminal amino acid residues of the mature invertase, a secreted yeast protein, and (3) the sequence for prochymosin fused to the invertase fragment. These YEp-type plasmids directed the secretion of prochymosin, but the fraction secreted was quite low, usually less than 5%. Mutagenesis and screening of the host strain have refined the system to the point where up to 80% of the synthesized fused protein is secreted from the yeast cell, as described above. However, the reported yield is still rather low – around 1 mg/g of total yeast protein – even in the best combinations of host with plasmid. A possible explanation is that *S. cerevisiae* normally secretes only a very small fraction of its cellular proteins across its cytoplasmic membrane; in wild-type strains, secreted invertase corresponds to much less than 0.1% of the total cellular protein. For this reason, the recent trend has been to explore other, more secretion-competent species of yeast.

In one experiment, a recombinant plasmid was made in which the sequence coding for prochymosin was placed between a strong LAC4 promoter and the LAC4 terminator sequence from *Kluyveromyces lactis*, a lactose-utilizing yeast species. When this plasmid was linearized and integrated into the *Kluyveromyces* genome, there was only a low-level expression of prochymosin. But most of the prochymosin was secreted into the medium, even though the cloned DNA lacked the sequence coding for the signal sequence. When the prochymosin gene was cloned together with the sequence coding for its own signal sequence, the prochymosin production increased 50- to 70-fold, and 95% of the product appeared in the medium in a correctly processed form. The yield was reported as about 100 enzyme units/ml, which corresponds roughly to 1 g/L, or about 10% of the total cellular protein. Other yeast species that have been shown to produce (and secrete) foreign proteins at a higher level than *S. cerevisiae* include *Pichia pastoris* and *Hansenula polymorpha*, both methylotrophic yeasts (yeasts capable of using methanol as the carbon source), and *Yarrowia lipolytica*, which can grow on alkanes. *P. pastoris*, for example, produces HBsAg to a level of about 50% of the total cellular protein.

There are also non-yeast fungal species that are known to secrete very large amounts of proteins; for example, *Trichoderma reesei* and *Aspergillus awamori* naturally secrete more than 20 g of protein per liter. These species were hypothesized to be even more proficient in catalyzing the export of large amounts of foreign proteins. Initial yields, obtained after cloning of the prochymosin gene in an expression vector, were not exceptional, but several optimization steps increased the yield significantly. These procedures included inactivation of the fungal gene that encodes a prochymosin-inactivating protease and fusion of the prochymosin sequence to the 3′-end of a complete sequence coding for a glucoamylase, an enzyme secreted in very large amounts by *A. awamori*. Such modifications increased the yield to the range of 100 mg/L. Finally, random mutagenesis and screening of the host *A. awamori* strain increased the yield to about 1 g/L, apparently a level that would make production commercially profitable. Importantly, the high-secretion mutant strain selected on the basis of prochymosin production also secretes other foreign proteins at a higher efficiency.

With a variety of tools for solving the production problem – chief among them the approaches described above – several laboratories are now attempting to modify, by site-directed mutagenesis, the structure of the prochymosin molecule itself. Recently, modifying the residues surrounding the glycosylation site to improve glycosylation efficiency resulted in the doubling of yield of prochymosin secreted from *A. awamori*.

SUMMARY

Some proteins and peptides of therapeutic value are difficult to purify in sufficient amounts from their human and animal sources. Recombinant DNA methods have had a revolutionary impact in the production of these compounds. Once the DNA sequences coding for these proteins and peptides have been cloned and amplified in microorganisms, the latter can continue to function as living factories for the inexpensive production of such compounds. Bacteria, especially *E. coli*, are used extensively as the host microorganism. Segments of “foreign DNA” coding for these products are first obtained either by cutting the genomic DNA or by synthesizing a DNA sequence (cDNA) complementary to the mRNA with reverse transcriptase. Such segments must first be inserted into vector DNA, which contains information that makes possible its replication in the bacterial host. In addition to plasmids, which are widely used as general-purpose cloning vectors, there are several types of vectors for cloning in bacteria. λ Phage vectors, cosmids, and BAC vectors are useful for the cloning of large segments of DNA, and single-strand DNA phage vectors are especially well suited for phage display technology that allows the isolation of mutants producing proteins with desired properties. The recombinant DNA – that is, the vector DNA containing the foreign DNA insert – is then introduced into the host cell by transformation or by injection from phagelike particles after it has been packaged into phage heads. The clone that contains the desired gene sequence is identified, sometimes among a vast majority of clones not containing this sequence, by using either the DNA sequence itself or the protein product of the gene as the marker. In some cases, however, PCR enables one to bypass all the steps of primary cloning and screening by direct amplification of the DNA sequence *in vitro*.

Regardless of the source, the sequence coding for the desired product can then be inserted into expression vectors to maximize the synthesis of the product in bacteria. The overproduction of foreign proteins in bacteria, however, frequently results in misfolding and aggregation of these proteins. Several strategies for avoiding aggregation are available, but none of them appears to be universally applicable to all proteins. However, aggregation does not necessarily mean a total failure, because protein aggregates can be easily purified, totally denatured, and then renatured under controlled conditions.

S. cerevisiae and other yeast species have considerable potential as host organisms for the production of foreign proteins, especially proteins of animal origin. Many different vectors are available, and most are shuttle vectors, which allow the recombinant DNA manipulations to be conveniently carried out in *E. coli* before the final recombinant product is introduced into yeast.

One major advantage of expression in yeasts is that foreign proteins appear to have less tendency to become misfolded in yeast than in bacterial hosts, partly because yeast cells presumably contain more efficient chaperones and foldases. Furthermore, proteins can become glycosylated in yeast cells if they can be introduced into the endoplasmic reticulum–Golgi apparatus protein-secretion pathway. Glycosylation not only facilitates the correct folding of some proteins but also makes them less susceptible to degradation in the animal body, thus prolonging their half-life when they are administered as therapeutic agents. Because *S. cerevisiae* is not well-equipped to secrete a large amount of proteins, nonconventional yeast species such as *P. pastoris* and *K. lactis* are increasingly used as hosts of secretion vectors, often making use of the "prepro" sequence of *S. cerevisiae* mating factor α precursor.

HBsAg and prochymosin are two proteins of animal origin that have been successfully produced in yeasts. HBsAg did not enter the secretion pathway and was not glycosylated, however, apparently because the cloned DNA fragment lacked the segment coding for the signal sequence. Nevertheless, it was folded correctly and assembled into a structure resembling the envelope of the virus. When the cDNA for prochymosin was expressed in *E. coli*, it produced inclusion bodies that were difficult to renature. In contrast, when it was expressed from a secretion vector in an *S. cerevisiae* host, the protein entered the endoplasmic reticulum–Golgi pathway, was folded correctly and glycosylated, and was secreted, although the yield remained low. When non-*Saccharomyces* yeasts and non-yeast fungi that physiologically secrete very large amounts of proteins were used as hosts, commercially acceptable yields of secreted prochymosin were achieved.

SELECTED REFERENCES

General References on Recombinant DNA Methods

Primrose, S. B., and Twyman, R. M. (2006). *Principles of Gene Manipulation and Genomics*, 7th Edition, Oxford, UK: Blackwell Science.

Sambrook, J., and Russell, D. W. (2001). *Molecular Cloning: A Laboratory Manual*, 3rd Edition, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Vectors

Balbas, P., Soberon, X., Merino, E., Zurita, M., Lomeli, H., Valle, F., Flores, N., and Bolivar, F. (1986). Plasmid vector pBR322 and its special-purpose derivatives – a review. *Gene*, 50, 3–40.

Casali, N., and Preston, A. (eds.) (2003). *E. coli Plasmid Vectors: Methods and Applications* (Vol. 235, Methods in Molecular Biology), Clifton NJ: Humana Press.

Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiri, Y., and Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proceedings of the National Academy of Sciences U.S.A.*, 89, 8794–8797.

Kehoe, J. W., and Kay, B. K. (2005). Filamentous phage display in the new millennium. *Chemical Reviews*, 105, 4056–4072.

Lipovsek, D., and Plückthun, A. (2004). In-vitro protein evolution by ribosome display and mRNA display. *Journal of Immunological Methods*, 290, 51–67.

PCR

Shamputa, I. C., Rigouts, L., and Portaels, F. (2004). Molecular genetic methods for diagnosis and antibiotic resistance detection of mycobacteria from clinical specimens. *APMIS*, 112, 728–752.

Expression of Cloned Genes

Makrides, S. C. (1996). Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiological Reviews*, 60, 512–538.

Baneyx, F. (1999). Recombinant protein expression in *Escherichia coli*. *Current Opinion in Biotechnology*, 10, 411–421.

Baneyx, F. (ed.) (2004). *Protein Expression Technologies: Current Status and Future Trends*, Norfolk, U.K.: Horizon Bioscience.

Proteolysis

Enfors, S.-O. (1992). Control *of in vivo* proteolysis in the production of recombinant proteins. *Trends in Biotechnology*, 10, 310–315.

Protein Folding, Foldases, and Molecular Chaperones

Baneyx, F., and Mujacic, M. (2004). Recombinant protein folding and misfolding in *Escherichia coli*. *Nature Biotechnology*, 22, 1399–1408.

Thomas, J. G., Ayling, A., and Baneyx, F. (1997). Molecular chaperones, folding catalysts, and the recovery of active recombinant proteins from *E. coli:* to fold or refold. *Applied Biochemistry and Biotechnology*, 66, 197–238.

Schmid, F. X. (2002). Prolyl isomerases. *Advances in Protein Chemistry*, 59, 243–282.

Bader, M. W., and Bardwell, J. C. A. (2002). Catalysis of disulfide bond formation and isomerization in *Escherichia coli*. *Advances in Protein Chemistry*, 59, 283–291.

Hartl, F. U., and Hayer-Hartl, M. (2002). Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, 295, 1852–1858.

Kapust, R. B., and Waugh, D. S. (1999). *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Science*, 8, 1668–1674.

Terpe, K. (2003). Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. *Applied Microbiology and Biotechnology*, 60, 523–533.

Protein Secretion

Georgiou, G., and Segatori, L. (2005). Preparative expression of secreted proteins in bacteria: status report and prospects. *Current Opinion in Biotechnology*, 16, 538–545.

Mergulhão, F. J. M., Summers, D. K., and Monteiro, G. A. (2005). Recombinant protein secretion in *Escherichia coli*. *Biotechnology Advances*, 23, 177–202.

Miot, M., and Betton, J.-M. (2004). Protein quality control in the bacterial periplasm. *Microbial Cell Factories*, 3, 4.

Prochymosin

Beppu, T. (1988). Production of chymosin (rennin) by recombinant DNA technology. In *Recombinant DNA and Bacterial Fermentation*, J. A. Thomson (ed.), pp. 11–21, Boca Raton, FL: CRC Press.

Cloning in Yeast

Guthrie, C., and Fink, G. R. (2002). *Guide to Yeast Genetics and Molecular and Cell Biology, Parts B and C* (Methods in Enzymology, volumes 350 and 351), New York: Academic Press.

Goffeau, A., Barrell, B. G., Bussey, H., et al. (1996). Life with 6000 genes. *Science*, 274, 546–567.

Kumar, A., and Snyder, M. (2001). Emerging technologies in yeast genomics, *Nature Reviews Genetics*, 2, 302–312.

Spencer, J. F. T., Ragout de Spencer, A. L., and Laluce, C. (2002). Non-conventional yeasts. *Applied Microbiology and Biotechnology*, 58, 147–156.

Liti, G., and Louis, E. J. (2005) Yeast evolution and comparative genomics. *Annual Review of Microbiology*, 59, 135–153.

Cereghino, G. P. L., Cereghino, J. L., Ilgen, C., and Cregg, J. M. (2002). Production of recombinant proteins in fermenter cultures of the yeast *Pichia pastoris*. *Current Opinion in Biotechnology*, 13, 329–332.

Gerngross, T. U. (2004). Advances in the production of human therapeutic proteins in yeasts and filamentous fungi. *Nature Biotechnology*, 22, 1409–1414.

Li, H., Sethuraman, N., Stadheim, T. A., et al. (2006). Optimization of humanized IgGs in glyco-engineered *Pichia pastoris*. Nature Biotechnology, 24, 210–215.

Kjeldsen, T. (2000). Yeast secretory expression of insulin precursors. *Applied Microbiology and Biotechnology*, 54, 277–286.

Mohanty, A. K., Mukhopadhyay, U. K., Grover, S., and Batish, V. K. (1999). Bovine chymosin: Production by rDNA technology and application to cheese manufacture. *Biotechnology Advances*, 17, 205–217.

van den Brink, H. M., Petersen, S. G., Rahbek-Nielsen, H., Hellmuth, K., and Harboe, M. (2006). Increased production of chymosin by glycosylation. *Journal of Biotechnology*, 125, 304–310.