

Chapter 11

Studying Gene Expression and Function

Chapter contents

CHAPTER CONTENTS

- 11.1 Studying the RNA transcript of a gene
- 11.2 Studying the regulation of gene expression
- 11.3 Identifying and studying the translation product of a cloned gene

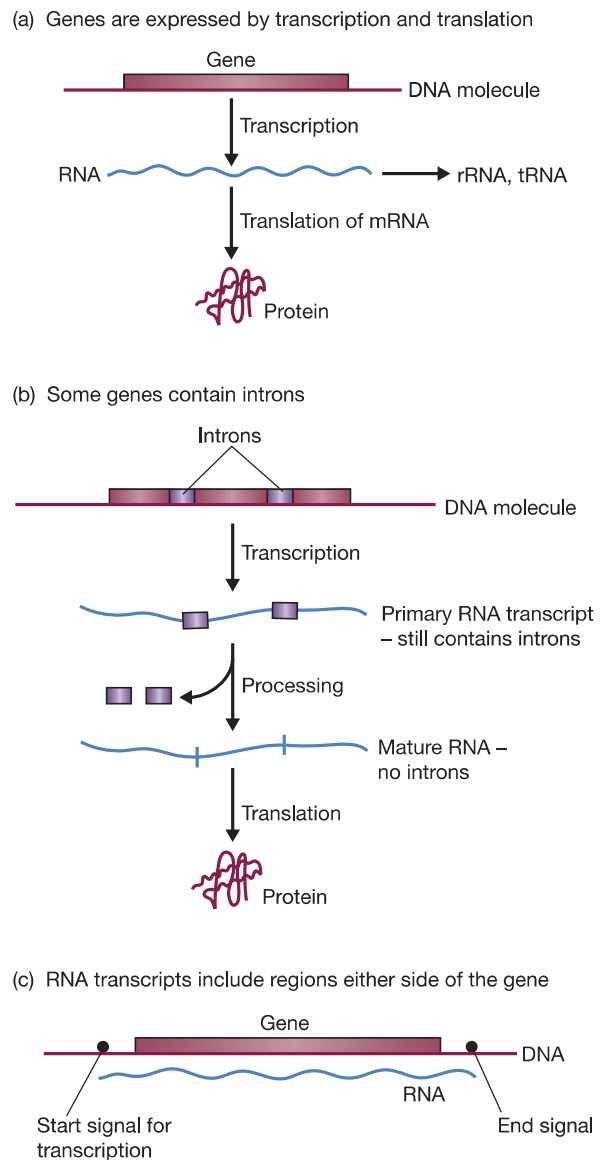
All genes have to be expressed in order to function. The first step in expression is transcription of the gene into a complementary RNA strand (Figure 11.1a). For some genes—for example, those coding for transfer RNA (tRNA) and ribosomal RNA (rRNA) molecules—the transcript itself is the functionally important molecule. For other genes, the transcript is translated into a protein molecule.

To understand how a gene is expressed, the RNA transcript must be studied. In particular, the molecular biologist will want to know whether the transcript is a faithful copy of the gene, or whether segments of the gene are missing from the transcript (Figure 11.1b). These missing pieces are called introns and considerable interest centers on their structure and possible function. In addition to introns, the exact locations of the start and end points of transcription are important. Most transcripts are copies not only of the gene itself, but also of the nucleotide regions either side of it (Figure 11.1c). The signals that determine the start and finish of the transcription process are only partly understood, and their positions must be located if the expression of a gene is to be studied.

In this chapter we will begin by looking at the methods used for **transcript analysis**. These methods can be used to map the positions of the start and end points for transcription and also to determine if a gene contains introns. Then we will briefly consider a few of the numerous techniques developed in recent years for examining how expression of a gene is regulated. These techniques are important as aberrations in gene regulation underlie many clinical disorders. Finally, we will tackle the difficult problem of how to identify the translation product of a gene.

Figure 11.1

Some fundamentals of gene expression.
 mRNA = messenger RNA, tRNA = transfer RNA,
 rRNA = ribosomal RNA.

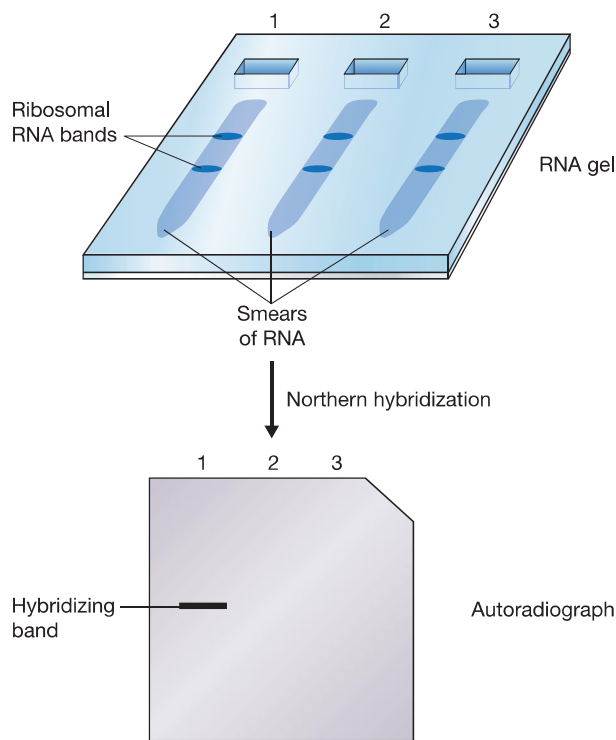


11.1 Studying the RNA transcript of a gene

Over the years a variety of techniques have been devised for studying RNA transcripts. Some of these techniques merely detect the presence of a transcript and give some indication of its length, others enable the start and end of the transcript to be mapped and the positions of introns to be located.

11.1.1 Detecting the presence of a transcript and determining its nucleotide sequence

Before studying the more sophisticated techniques for RNA analysis, we must consider the methods used to obtain basic information about a transcript. The first of these methods is **northern hybridization**, the RNA equivalent of Southern hybridization

**Figure 11.2**

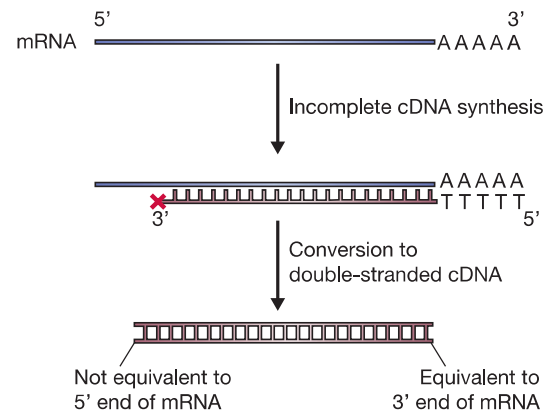
Northern hybridization. Three RNA extracts from different tissues have been electrophoresed in an agarose gel. The extracts are made up of many RNAs of different lengths so each gives a smear of RNA, but two distinct bands are seen, one for each of the abundant ribosomal RNAs. The sizes of these rRNAs are known (e.g. 4718 and 1874 nucleotides in mammals), so they can be used as internal size markers. The gel is transferred to a membrane, probed with a cloned gene, and the results visualized, for example by autoradiography if the probe has been radioactively labeled. Only lane 1 gives a band, showing that the cloned gene is expressed only in the tissue from which this RNA extract was obtained.

(p. 142), which is used to measure the length of a transcript. An RNA extract is electrophoresed in an agarose gel, using a denaturing electrophoresis buffer (e.g., one containing formaldehyde) to ensure that the RNAs do not form inter- or intramolecular base pairs, as base pairing would affect the rate at which the molecules migrate through the gel. After electrophoresis, the gel is blotted onto a nylon or nitrocellulose membrane, and hybridized with a labeled probe (Figure 11.2). If the probe is a cloned gene, the band that appears in the autoradiograph is the transcript of that gene. The size of the transcript can be determined from its position within the gel, and if RNA from different tissues is run in different lanes of the gel, then the possibility that the gene is differentially expressed can be examined.

Once a transcript has been identified, cDNA synthesis (p. 133) can be used to convert it into a double-stranded DNA copy, which can be cloned and sequenced. Comparison between the sequence of the cDNA and the sequence of its gene will reveal the positions of introns and possibly the start and end points of the transcript. For this to be possible, the cDNA must be a full length copy of the mRNA from which it is derived. The 3' end of the transcript will usually be represented in the cDNA, because most methods for cDNA synthesis begin with a primer that anneals to the poly(A) tail of the mRNA, which means that the 3' end is the first part to be copied (see Figure 8.7). The cDNA synthesis might not, however, continue all the way to the 5' end of the transcript, especially if the RNA is more than a few hundred nucleotides in length. Premature termination of cDNA synthesis will result in a cDNA that is not a full length copy of its transcript, and whose 3' end does not correspond to the true 5' end of the mRNA (Figure 11.3). If this is the case, then the start point for transcription cannot be identified from the sequence of the cDNA. It will also be impossible to map the positions of any introns that are located near the start of the gene.

Figure 11.3

The 5' end of an mRNA cannot be identified from the sequence of an incomplete cDNA.



11.1.2 Transcript mapping by hybridization between gene and RNA

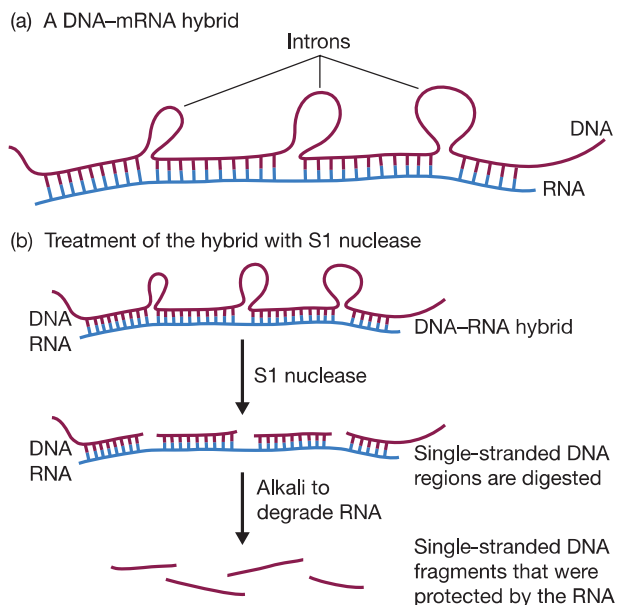
The problems with incomplete cDNA synthesis mean that other, more direct methods are needed to identify the start point of a transcript. The most informative of these methods are based on examination of the hybridization product formed between a cloned gene and its RNA.

Nucleic acid hybridization occurs just as readily between complementary DNA and RNA strands as between single-stranded DNA molecules. If a hybrid is formed between a DNA strand, containing a gene, and its mRNA, then the boundaries between the double- and single-stranded regions will mark the start and end points of the mRNA (Figure 11.4a). Introns, which are present in the DNA but not in the mRNA, will “loop out” as additional single-stranded regions.

Now consider the result of treating the DNA–RNA hybrid with a single-strand-specific nuclease such as S1 (p. 46). S1 nuclease degrades single-stranded DNA or RNA polynucleotides, including single-stranded regions at the ends of predominantly double-stranded molecules, but has no effect on double-stranded DNA or on DNA–RNA hybrids. S1 nuclease will therefore digest the non-hybridized single-stranded DNA

Figure 11.4

A DNA–mRNA hybrid and the effect of treating this hybrid with a single-strand-specific nuclease such as S1.



regions at each end of the DNA–RNA hybrid, along with any looped-out introns (Figure 11.4b). The single-stranded DNA fragments protected from S1 nuclease digestion can be recovered if the RNA strand is degraded by treatment with alkali.

Unfortunately, the manipulations shown in Figure 11.4 are not very informative. The sizes of the protected DNA fragments could be measured by gel electrophoresis, but this does not allow their order or relative positions in the DNA sequence to be determined. However, a few subtle modifications to the technique allow the precise start and end points of the transcript and of any introns it contains to be mapped onto the DNA sequence.

An example of the way in which **S1 nuclease mapping** is used to locate the start point of a transcript is shown in Figure 11.5. Here, a *Sau3A* fragment that contains 100 bp of coding region, along with 300 bp of the leader sequence preceding the gene, has been cloned into an M13 vector and obtained as a single-stranded molecule. A sample of the RNA transcript is added and allowed to anneal to the DNA molecule. The DNA molecule is still primarily single-stranded but now has a small region protected by the RNA transcript. All but this protected region is digested by S1 nuclease and the RNA is degraded with alkali, leaving a short single-stranded DNA fragment. If these

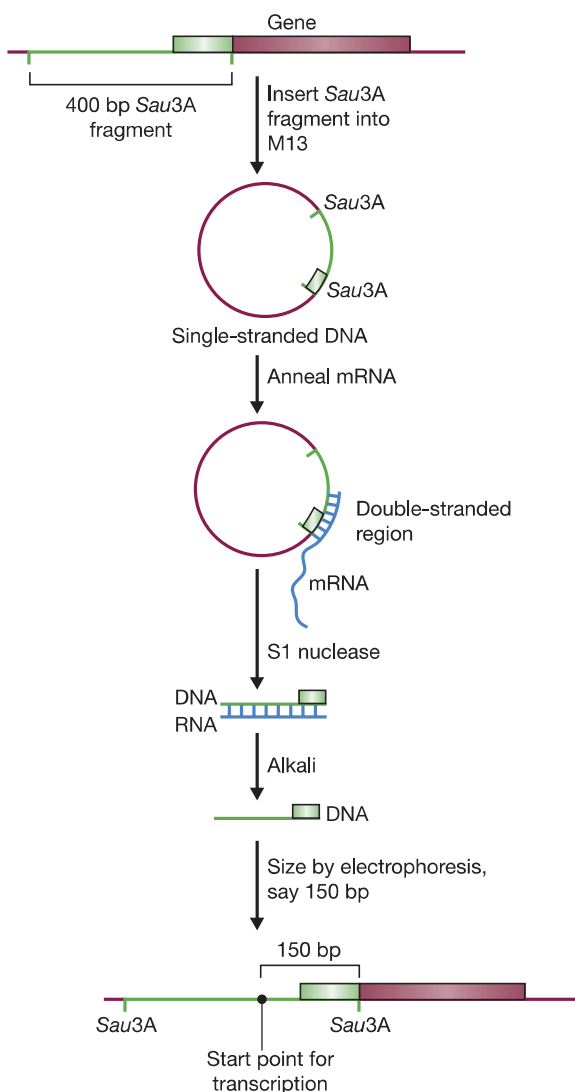


Figure 11.5

Locating a transcription start point by S1 nuclease mapping.

manipulations are examined closely it will become clear that the size of this single-stranded fragment corresponds to the distance between the transcription start point and the right-hand *Sau3A* site. The size of the single-stranded fragment is therefore determined by gel electrophoresis and this information is used to locate the transcription start point on the DNA sequence. Exactly the same strategy could locate the end point of transcription and the junction points between introns and exons. The only difference would be the position of the restriction site chosen to delimit one end of the protected single-stranded DNA fragment.

11.1.3 Transcript analysis by primer extension

S1 nuclease analysis is a powerful technique that allows both the 5' and 3' termini of a transcript, as well as the positions of intron–exon boundaries, to be identified. The second method of transcript analysis that we will consider—**primer extension**—is less adaptable, because it can only identify the 5' end of an RNA. It is, nonetheless, an important technique that is frequently used to check the results of S1 analyses.

Primer extension can only be used if at least part of the sequence of the transcript is known. This is because a short oligonucleotide primer must be annealed to the RNA at a known position, ideally within 100–200 nucleotides of the 5' end of the transcript. Once annealed, the primer is extended by reverse transcriptase (p. 49). This is a cDNA synthesis reaction, but one that is very likely to proceed to completion as only a short segment of RNA has to be copied (Figure 11.6). The 3' end of the newly synthesized strand of DNA will therefore correspond with the 5' terminus of the transcript. Locating

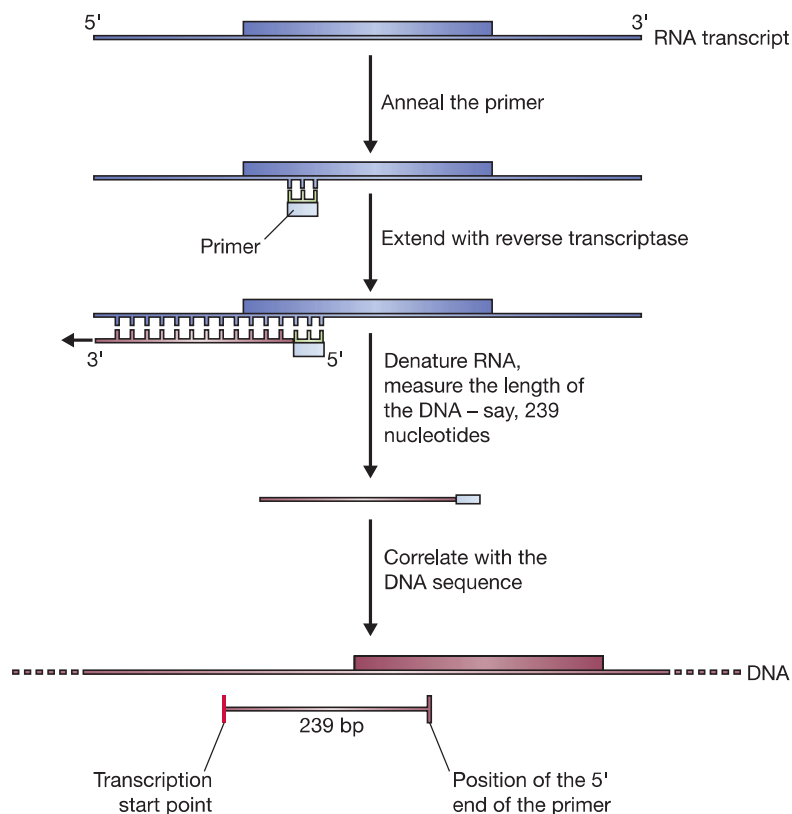


Figure 11.6

Locating a transcription start point by primer extension.

the position of this terminus on the DNA sequence is achieved simply by determining the length of the single-stranded DNA molecule and correlating this information with the annealing position of the primer.

11.1.4 Transcript analysis by PCR

We have already studied the way in which reverse transcriptase PCR is used to quantify the amount of a particular RNA in an extract (p. 160). The standard reverse transcriptase PCR procedure provides a copy of the internal region of an RNA molecule but does not give any information about the ends of the molecule. A modified version called **rapid amplification of cDNA ends (RACE)** can be used to identify the 5' and 3' termini of RNA molecules and so, like S1 analysis, can be used to map the ends of transcripts.

There are several variations to the RACE method. Here we will consider how the 5' end of an RNA molecule can be mapped (Figure 11.7). This procedure uses a primer

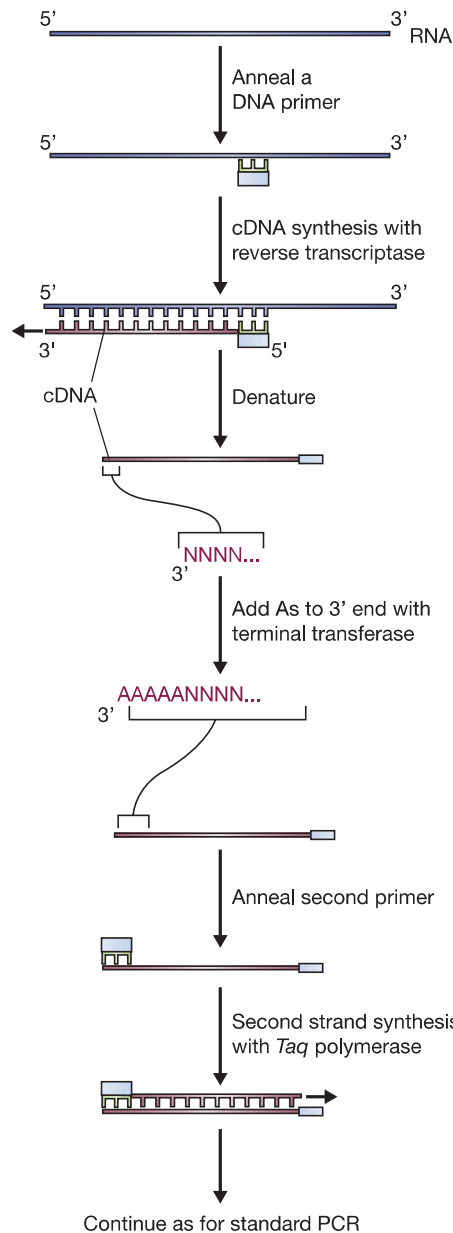


Figure 11.7

One version of RACE.

that is specific for an internal region of the RNA molecule. The primer attaches to the RNA and directs the first, reverse transcriptase catalyzed, stage of the process, during which a single-stranded cDNA is made. As in the primer extension method, the 3' end of the cDNA corresponds with the 5' end of the RNA. Terminal deoxynucleotidyl transferase (p. 50) is now used to attach a series of A nucleotides to the 3' end of the cDNA, forming the priming site for a second PCR primer, which is made up entirely of Ts and hence anneals to the poly(A) tail created by terminal transferase. Now the standard PCR begins, first converting the single-stranded cDNA into a double-stranded molecule, and then amplifying this molecule as the PCR proceeds. The PCR product is then sequenced to reveal the precise position of the start of the transcript.

11.2 Studying the regulation of gene expression

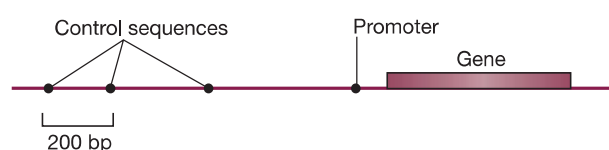
Few genes are expressed all the time. Most are subject to regulation and are switched on only when their gene product is required by the cell. The simplest gene regulation systems are found in bacteria such as *E. coli*, which can regulate expression of genes for biosynthetic and metabolic processes, so that gene products that are not needed are not synthesized. For instance, the genes coding for the enzymes involved in tryptophan biosynthesis can be switched off when there are abundant amounts of tryptophan in the cell, and switched on again when tryptophan levels drop. Similarly, genes for the utilization of sugars such as lactose are activated only when the relevant sugar is there to be metabolized. In higher organisms, gene regulation is more complex because there are many more genes to control. Differentiation of cells involves wholesale changes in gene expression patterns, and the process of development from fertilized egg cell to adult requires coordination between different cells as well as time-dependent changes in gene expression.

Many of the problems in gene regulation require a classical genetic approach: genetics enables genes that control regulation to be distinguished, allows the biochemical signals that influence gene expression to be identified, and can explore the interactions between different genes and gene families. It is for this reason that most of the breakthroughs in understanding development in higher organisms have started with studies of the fruit fly *Drosophila melanogaster*. Gene cloning and DNA analysis complement classical genetics as they provide much more detailed information on the molecular events involved in regulating the expression of a single gene.

We now know that a gene subject to regulation has one or more control sequences in its upstream region (Figure 11.8) and that the gene is switched on and off by the attachment of regulatory proteins to these sequences. A regulatory protein may repress gene expression, in which case the gene is switched off when the protein is bound to the control sequence, or alternatively the protein may have a positive or enhancing role, switching on or increasing expression of the target gene. In this section we will examine methods for locating control sequences and determining their roles in regulating gene expression.

Figure 11.8

Possible positions for control sequences in the region upstream of a gene.



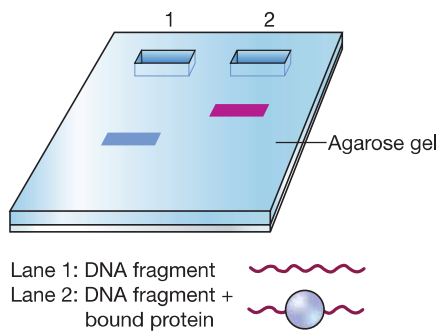


Figure 11.9

A bound protein decreases the mobility of a DNA fragment during gel electrophoresis.

11.2.1 Identifying protein binding sites on a DNA molecule

A control sequence is a region of DNA that can bind a regulatory protein. It should therefore be possible to identify control sequences upstream of a gene by searching the relevant region for protein binding sites. There are three different ways of doing this.

Gel retardation of DNA–protein complexes

Proteins are quite substantial structures and a protein attached to a DNA molecule results in a large increase in overall molecular mass. If this increase can be detected, a DNA fragment containing a protein binding site will have been identified. In practice a DNA fragment carrying a bound protein is identified by gel electrophoresis, as it has a lower mobility than the uncomplexed DNA molecule (Figure 11.9). The procedure is referred to as **gel retardation**.

In a gel retardation experiment (Figure 11.10), the region of DNA upstream of the gene being studied is digested with a restriction endonuclease and then mixed with the regulatory protein or, if the protein has not yet been purified, with an unfractionated extract of nuclear protein (remember that gene regulation occurs in the nucleus). The

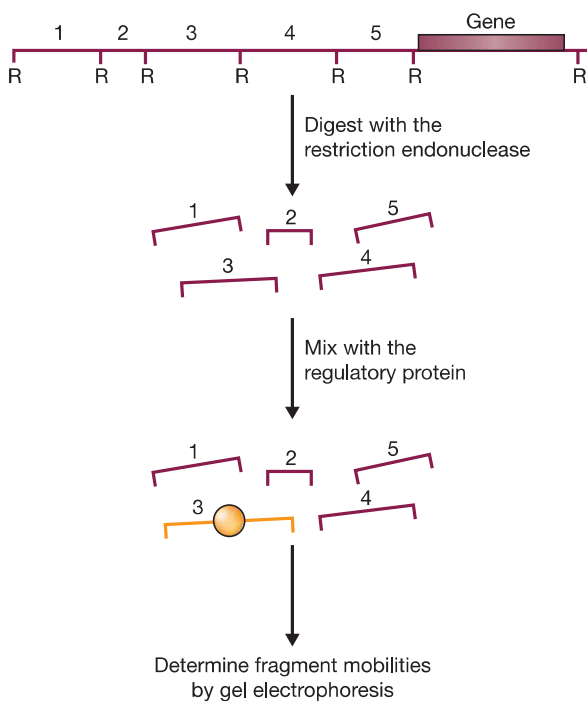
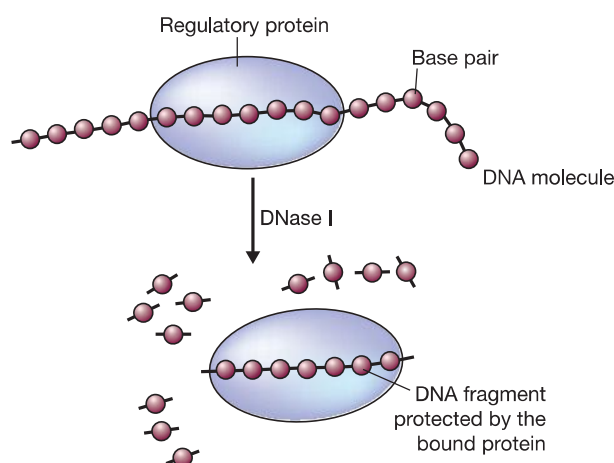


Figure 11.10

Carrying out a gel retardation experiment.

Figure 11.11

A bound protein protects a region of a DNA molecule from degradation by a nuclease such as DNase I.



restriction fragment containing the control sequence forms a complex with the regulatory protein: all the other fragments remain as “naked” DNA. The location of the control sequence is then determined by finding the position on the restriction map of the fragment that is retarded during gel electrophoresis. The precision with which the control sequence can be located depends on how detailed the restriction map is and how conveniently placed the restriction sites are. A single control sequence may be less than 10 bp in size, so gel retardation is rarely able to pinpoint it exactly. More precise techniques are therefore needed to delineate the position of the control sequence within the fragment identified by gel retardation.

Footprinting with DNase I

The procedure generally called **footprinting** enables a control region to be positioned within a restriction fragment that has been identified by gel retardation. Footprinting works on the basis that the interaction with a regulatory protein protects the DNA in the region of a control sequence from the degradative action of an endonuclease such as deoxyribonuclease (DNase) I (Figure 11.11). This phenomenon can be used to locate the protein binding site on the DNA molecule.

The DNA fragment being studied is first labeled at one end, and then complexed with the regulatory protein (Figure 11.12a). Then DNase I is added, but the amount used is limited so that complete degradation of the DNA fragment does not occur. Instead the aim is to cut each molecule at just a single phosphodiester bond (Figure 11.12b). If the DNA fragment has no protein attached to it, the result of this treatment is a family of labeled fragments, differing in size by just one nucleotide each.

After removal of the bound protein and separation on a polyacrylamide gel, the family of labeled fragments appears as a ladder of bands (Figure 11.12c). However, the bound protein protected certain phosphodiester bonds from being cut by DNase I, meaning that in this case the family of fragments is not complete, as the fragments resulting from cleavage within the control sequence are absent. Their absence shows up as a “footprint”, clearly seen in Figure 11.12c. The region of the DNA molecule containing the control sequence can now be worked out from the sizes of the fragments on either side of the footprint.

Modification interference assays

Gel retardation analysis and footprinting enable control sequences to be located, but do not give information about the interaction between the binding protein and the DNA

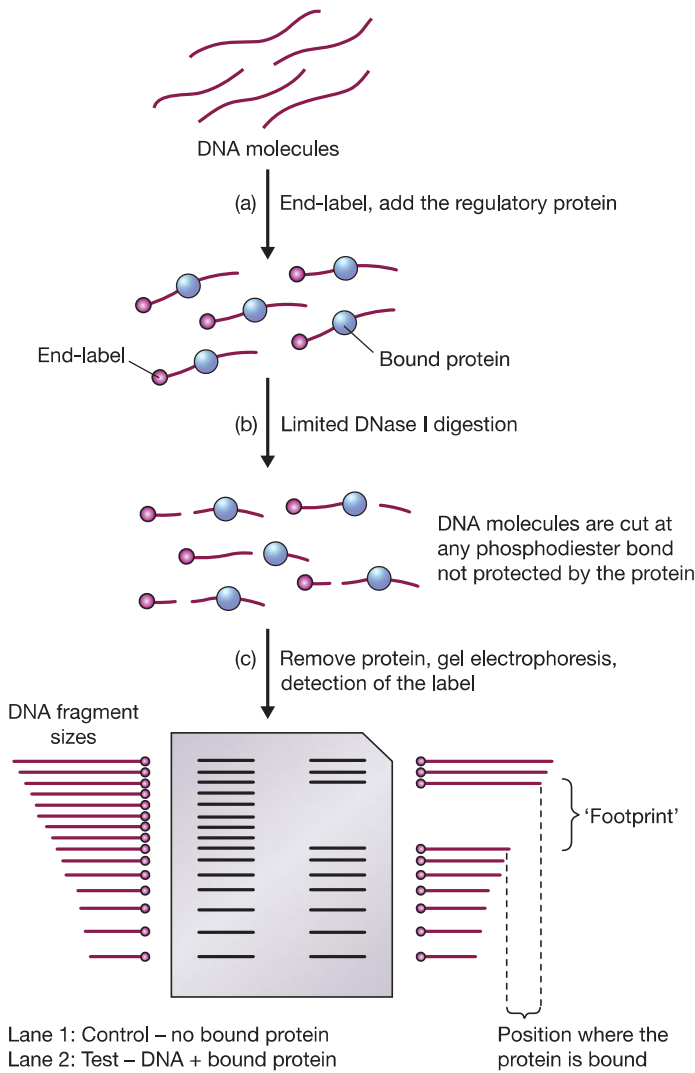


Figure 11.12
DNase I footprinting.

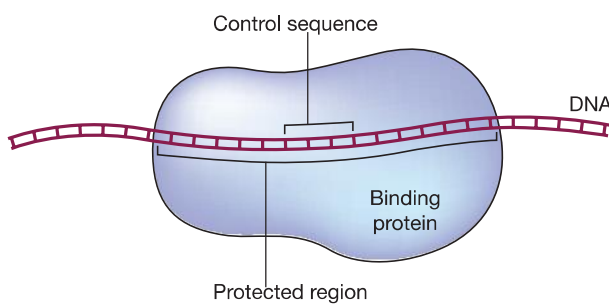
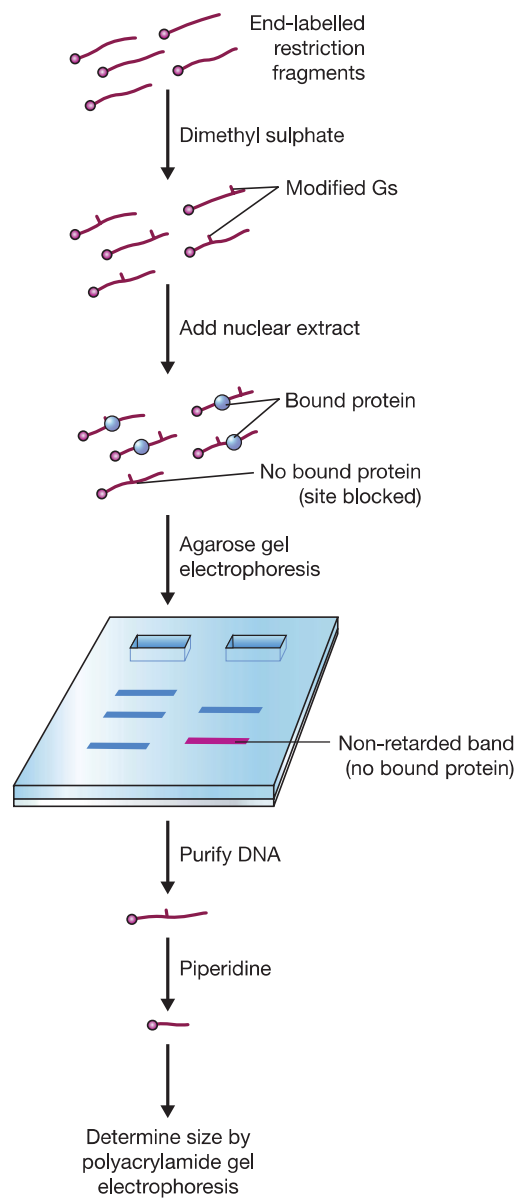


Figure 11.13
A bound protein can protect a region of DNA that is much longer than the control sequence.

molecule. The more precise of these two techniques—footprinting—only reveals the region of DNA that is protected by the bound protein. Proteins are relatively large compared with a DNA double helix, and can protect several tens of base pairs when bound to a control sequence that is just a few base pairs in length (Figure 11.13). Footprinting therefore does not delineate the control region itself, only the region within which it is located.

Figure 11.14

A modification interference assay.



Nucleotides that actually form attachments with a bound protein can be identified by the **modification interference assay**. As in footprinting, the DNA fragments must first be labeled at one end. Then they are treated with a chemical that modifies specific nucleotides, an example being dimethyl sulphate, which attaches methyl groups to guanine nucleotides (Figure 11.14). This modification is carried out under limiting conditions so an average of just one nucleotide per DNA fragment is modified. Now the DNA is mixed with the protein extract. The key to the assay is that the binding protein will probably not attach to the DNA if one of the guanines within the control region is modified, because methylation of a nucleotide interferes with the specific chemical reaction that enables it to form an attachment with a protein.

How is the absence of protein binding monitored? If the DNA and protein mixture is examined by agarose gel electrophoresis two bands will be seen, one comprising the DNA–protein complex and one containing DNA with no bound protein – in essence, this part of the procedure is a gel retardation assay (Figure 11.14). The band made up of unbound DNA is purified from the gel and treated with piperidine, a chemical which

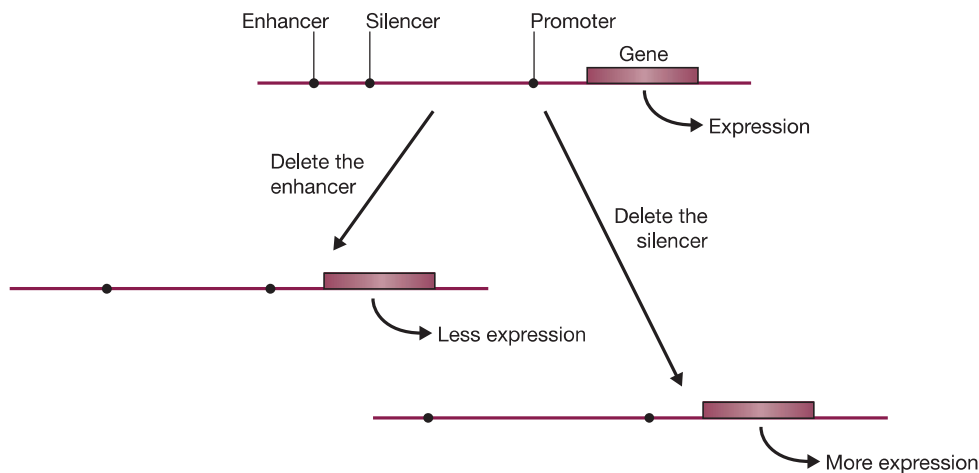


Figure 11.15

The principle behind deletion analysis.

cleaves DNA molecules at methylated nucleotides. The products of piperidine treatment are now separated in a polyacrylamide gel and the labeled bands visualized. The sizes of the bands that are seen indicate the position in the DNA fragment of guanines whose methylation prevented protein binding. These guanines lie within the control sequence. The modification assay can now be repeated with chemicals that target A, T, or C nucleotides to determine the precise position of the control sequence.

11.2.2 Identifying control sequences by deletion analysis

Gel retardation, footprinting, and modification interference assays are able to locate possible control sequences upstream of a gene, but they provide no information on the function of the individual sequences. **Deletion analysis** is a totally different approach that not only can locate control sequences (though only with the precision of gel retardation), but importantly also can indicate the function of each sequence.

The technique depends on the assumption that deletion of the control sequence will result in a change in the way in which expression of a cloned gene is regulated (Figure 11.15). For instance, deletion of a sequence that represses expression of a gene should result in that gene being expressed at a higher level. Similarly, tissue-specific control sequences can be identified as their deletion results in the target gene being expressed in tissues other than the correct one.

Reporter genes

Before carrying out a deletion analysis, a way must be found to assay the effect of a deletion on expression of the cloned gene. The effect will probably only be observed when the gene is cloned into the species from which it was originally obtained: it will be no good assaying for light regulation of a plant gene if the gene is cloned in a bacterium.

Cloning vectors have now been developed for most organisms (Chapter 7), so cloning the gene under study back into its host should not cause a problem. The difficulty is that in most cases the host already possesses a copy of the gene within its chromosomes. How can changes in the expression pattern of the cloned gene be distinguished from the normal pattern of expression displayed by the chromosomal copy of the gene? The answer is to use a **reporter gene**. This is a test gene that is fused to the upstream region

Figure 11.16

A reporter gene.

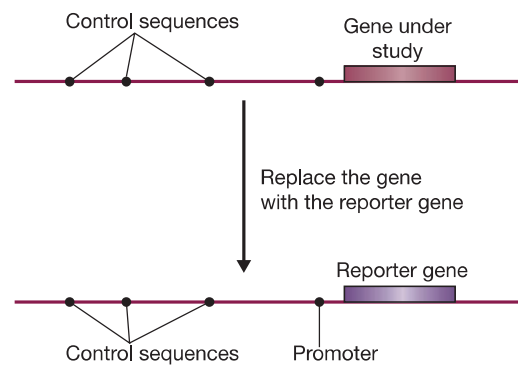


Table 11.1

A few examples of reporter genes used in studies of gene regulation in higher organisms.

GENE*	GENE PRODUCT	ASSAY
<i>lacZ</i>	β -Galactosidase	Histochemical test
<i>neo</i>	Neomycin phosphotransferase	Kanamycin resistance
<i>cat</i>	Chloramphenicol acetyltransferase	Chloramphenicol resistance
<i>dhfr</i>	Dihydrofolate reductase	Methotrexate resistance
<i>aphIV</i>	Hygromycin phosphotransferase	Hygromycin resistance
<i>lux</i>	Luciferase	Bioluminescence
GFP	Green fluorescent protein	Fluorescence
<i>uidA</i>	β -Glucuronidase	Histochemical test

*All of these genes are obtained from *E. coli*, except for: *lux* which has three sources: the luminescent bacteria *Vibrio harveyii* and *V. fischeri*, and the firefly *Photinus pyralis*; and GFP, which is obtained from the jellyfish *Aequorea victoria*.

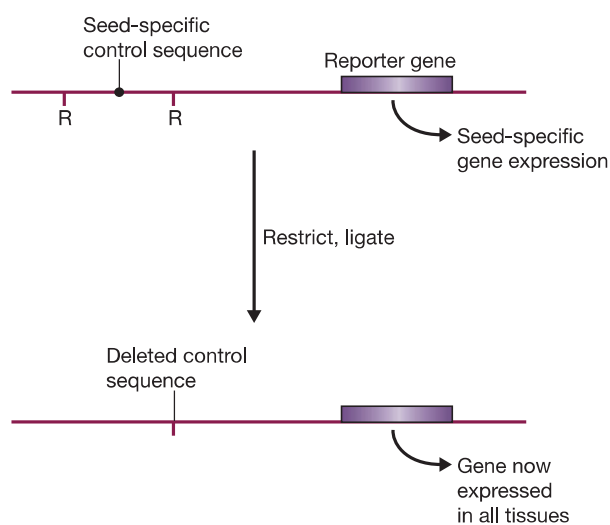
of the cloned gene (Figure 11.16), replacing this gene. When cloned into the host organism the expression pattern of the reporter gene should exactly mimic that of the original gene, as the reporter gene is under the influence of exactly the same control sequences as the original gene.

The reporter gene must be chosen with care. The first criterion is that the reporter gene must code for a phenotype not already displayed by the host organism. The phenotype of the reporter gene must be relatively easy to detect after it has been cloned into the host, and ideally it should be possible to assay the phenotype quantitatively. These criteria have not proved difficult to meet and a variety of different reporter genes have been used in studies of gene regulation. A few examples are listed in Table 11.1.

Carrying out a deletion analysis

Once a reporter gene has been chosen and the necessary construction made, carrying out a deletion analysis is fairly straightforward. Deletions can be made in the upstream region of the construct by any one of several strategies, a simple example being shown in Figure 11.17. The effect of the deletion is then assessed by cloning the deleted construct into the host organism and determining the pattern and extent of expression of the reporter gene. An increase in expression will imply that a repressing or silencing sequence has been removed, a decrease will indicate removal of an activator or enhancer, and a change in tissue specificity (as shown in Figure 11.17) will pinpoint a tissue-responsive control sequence.

The results of a deletion analysis project have to be interpreted very carefully. Complications may arise if a single deletion removes two closely linked control

**Figure 11.17**

Deletion analysis. A reporter gene has been attached to the upstream region of a seed-specific gene from a plant. Removal of the restriction fragment bounded by the sites R deletes the control sequence that mediates seed-specific gene expression, so that the reporter gene is now expressed in all tissues of the plant.

sequences or, as is fairly common, two distinct control sequences cooperate to produce a single response. Despite these potential difficulties, deletion analyses, in combination with studies of protein binding sites, have provided important information about how the expression of individual genes is regulated, and have supplemented and extended the more broadly based genetic analyses of differentiation and development.

11.3 Identifying and studying the translation product of a cloned gene

Over the years gene cloning has become increasingly useful in the study not only of genes themselves but also of the proteins coded by cloned genes. Investigations into protein structure and function have benefited greatly from the development of techniques that allow mutations to be introduced at specific points in a cloned gene, resulting in directed changes in the structure of the encoded protein.

Before considering these procedures we should first look at the more mundane problem of how to isolate the protein coded by a cloned gene. In many cases this analysis will not be necessary, as the protein will have been characterized long before the gene cloning experiment is performed and pure samples of the protein will already be available. On the other hand, there are occasions when the translation product of a cloned gene has not been identified. A method for isolating the protein is then needed.

11.3.1 HRT and HART can identify the translation product of a cloned gene

Two related techniques, **hybrid-release translation (HRT)** and **hybrid-arrest translation (HART)**, are used to identify the translation product encoded by a cloned gene. Both depend on the ability of purified mRNA to direct synthesis of proteins in **cell-free translation systems**. These are cell extracts, usually prepared from germinating wheat seeds or from rabbit reticulocyte cells (both of which are exceptionally active in protein synthesis) and containing ribosomes, tRNAs, and all the other molecules needed for protein synthesis. The mRNA sample is added to the cell-free translation system, along with a mixture of the 20 amino acids found in proteins, one of which is labeled (often

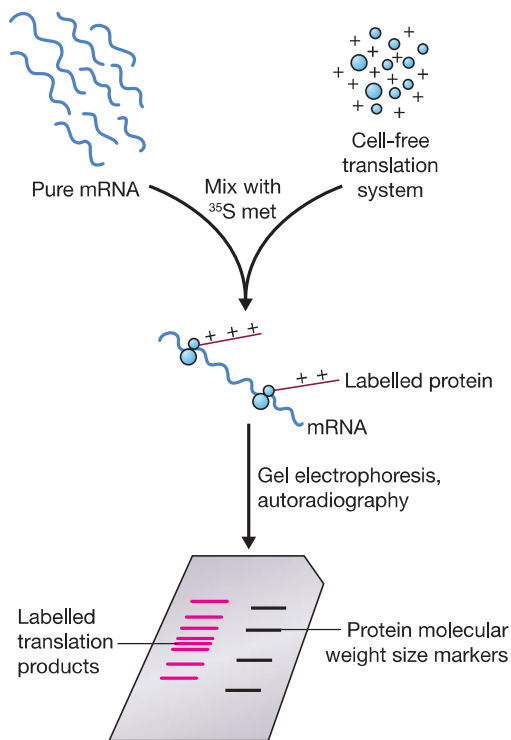


Figure 11.18

Cell-free translation.

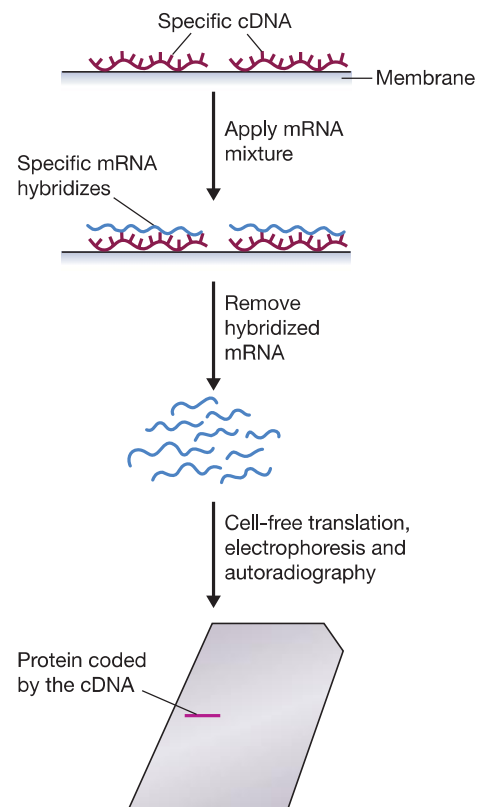


Figure 11.19

Hybrid-release translation.

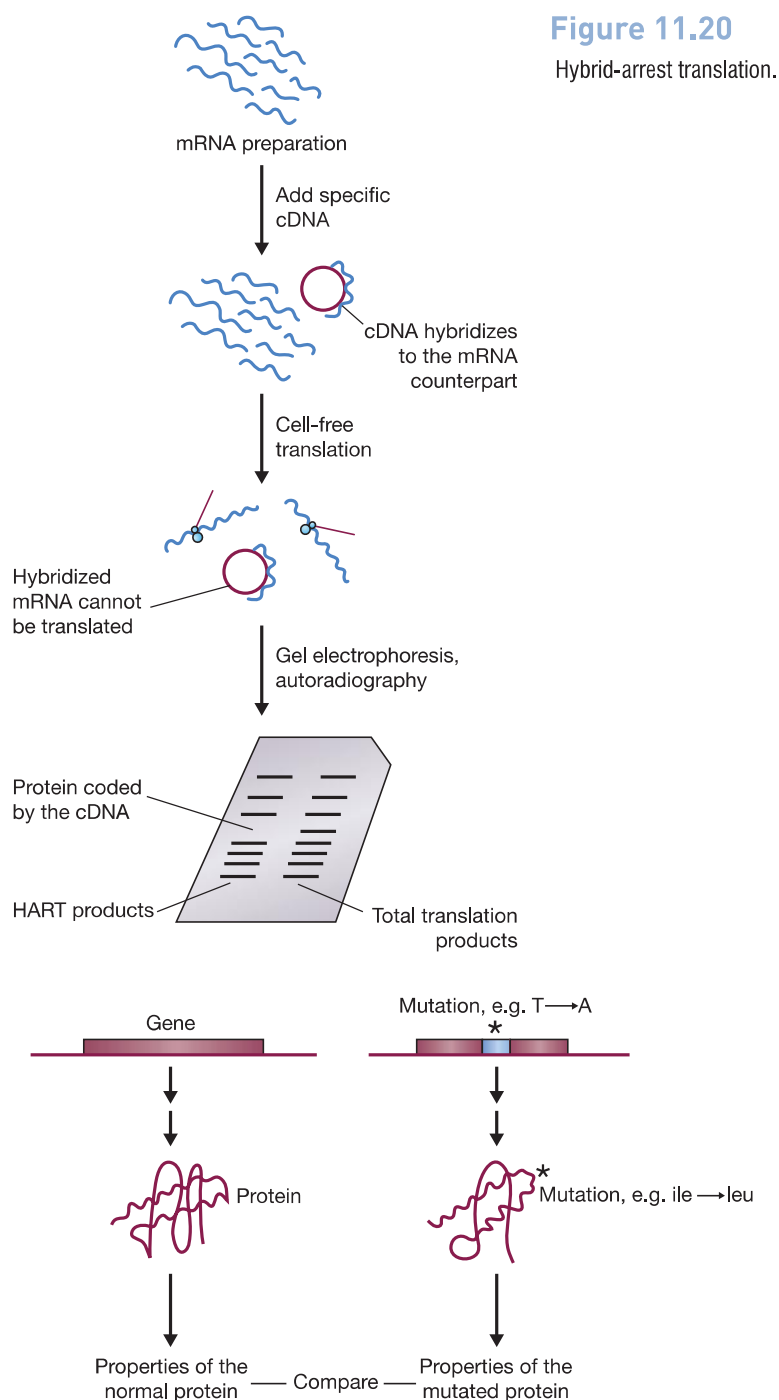
^{35}S -methionine is used). The mRNA molecules are translated into a mixture of radioactive proteins (Figure 11.18), which can be separated by gel electrophoresis and visualized by autoradiography. Each band represents a single protein coded by one of the mRNA molecules present in the sample.

Both HRT and HART work best when the gene being studied has been obtained as a cDNA clone. For HRT the cDNA is denatured, immobilized on a nitrocellulose or nylon membrane, and incubated with the mRNA sample (Figure 11.19). The specific mRNA counterpart of the cDNA hybridizes and remains attached to the membrane. After discarding the unbound molecules, the hybridized mRNA is recovered and translated in a cell-free system. This provides a pure sample of the protein coded by the cDNA.

Hybrid-arrest translation is slightly different in that the denatured cDNA is added directly to the mRNA sample (Figure 11.20). Hybridization again occurs between the cDNA and its mRNA counterpart, but in this case the unbound mRNA is not discarded. Instead the entire sample is translated in the cell-free system. The hybridized mRNA is unable to direct translation, so all the proteins except the one coded by the cloned gene are synthesized. The cloned gene's translation product is therefore identified as the protein that is absent from the autoradiograph.

11.3.2 Analysis of proteins by *in vitro* mutagenesis

Although HRT and HART can identify the translation product of a cloned gene, these techniques tell us little about the protein itself. Some of the major questions asked by the molecular biologist today center on the relationship between the structure of a

**Figure 11.21**

A mutation may change the amino acid sequence of a protein, possibly affecting its properties.

protein and its mode of activity. The best way of tackling these problems is to induce a mutation in the gene coding for the protein and then to determine what effect the change in amino acid sequence has on the properties of the translation product (Figure 11.21). Under normal circumstances mutations occur randomly and a large number may have to be screened before one that gives useful information is found. A solution to this problem is provided by *in vitro mutagenesis*, a technique that enables a directed mutation to be made at a specific point in a cloned gene.

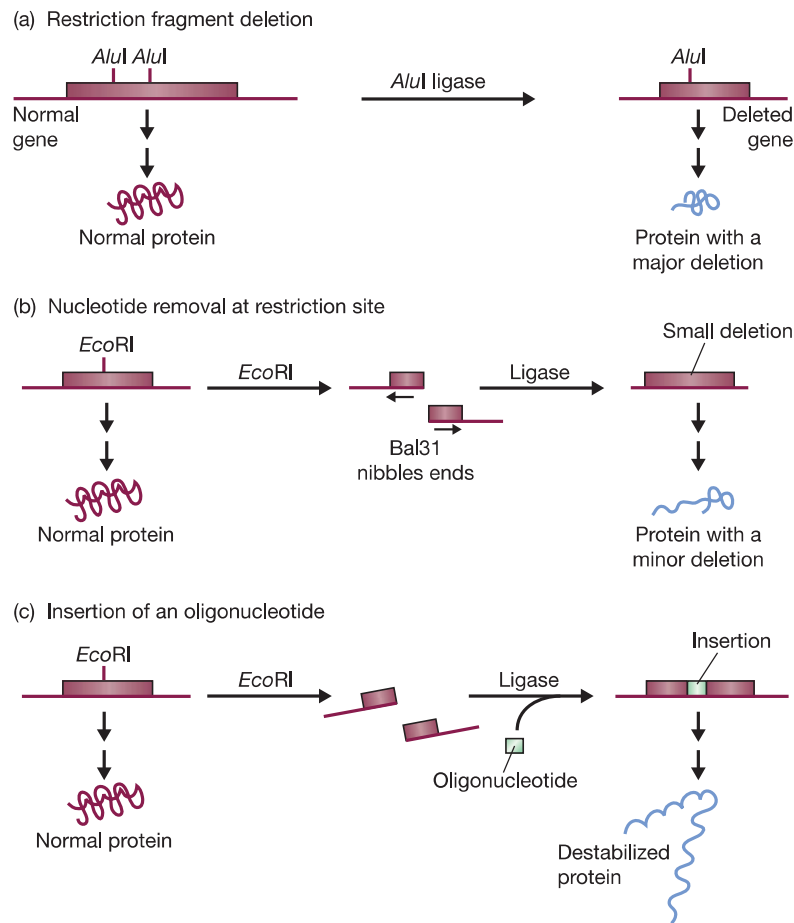


Figure 11.22

Various *in vitro* mutagenesis techniques.

Different types of in vitro mutagenesis techniques

An almost unlimited variety of DNA manipulations can be used to introduce mutations into cloned genes. The following are the simplest:

- A restriction fragment can be deleted (Figure 11.22a).
- The gene can be opened at a unique restriction site, a few nucleotides removed with a double-strand-specific endonuclease such as Bal31 (p. 46), and the gene religated (Figure 11.22b).
- A short, double-stranded oligonucleotide can be inserted at a restriction site (Figure 11.22c). The sequence of the oligonucleotide can be such that the additional stretch of amino acids inserted into the protein produces, for example, a new structure such as an α -helix, or destabilizes an existing structure.

Although potentially useful, these manipulations depend on the fortuitous occurrence of a restriction site at the area of interest in the cloned gene. **Oligonucleotide-directed mutagenesis** is a more versatile technique that can create a mutation at any point in the gene.

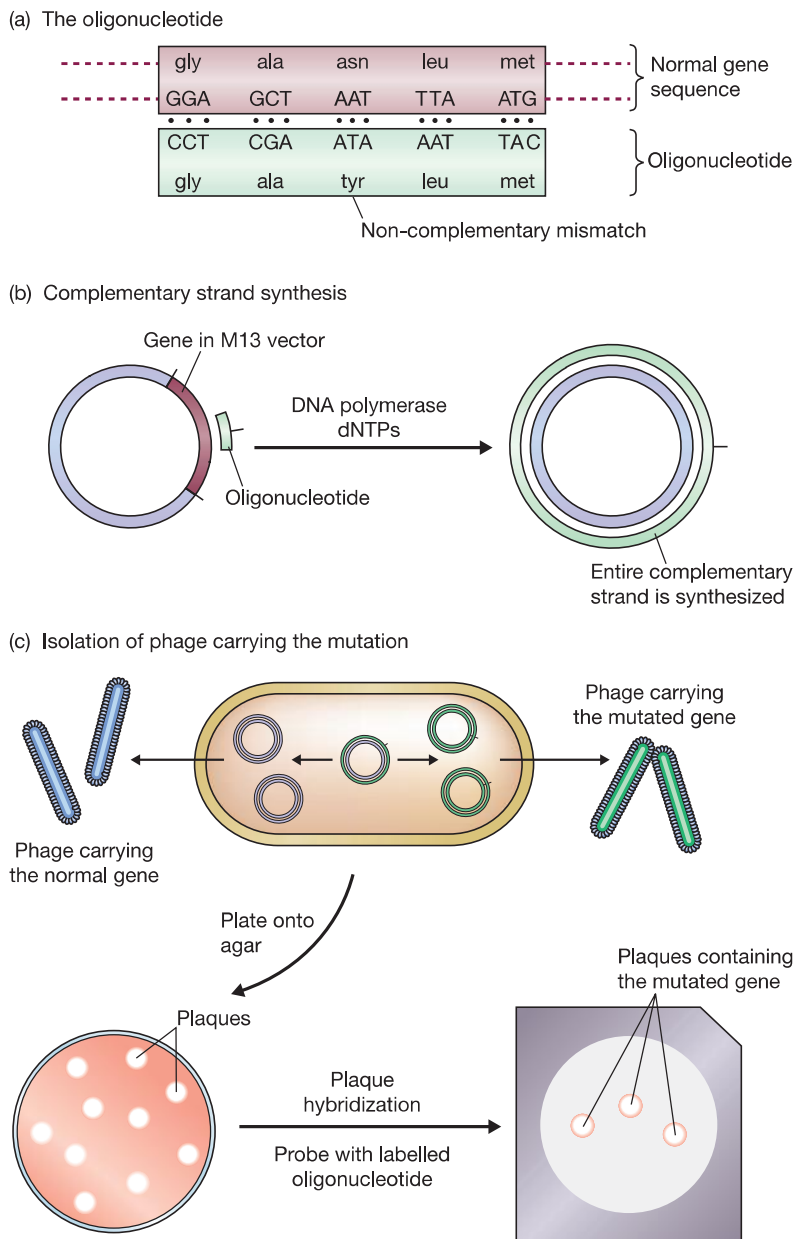


Figure 11.23

One method for oligonucleotide-directed mutagenesis.

Using an oligonucleotide to create a point mutation in a cloned gene

There are a number of different ways of carrying out oligonucleotide-directed mutagenesis; we will consider one of the simplest of these methods. The gene to be mutated must be obtained in a single-stranded form and so is generally cloned into an M13 vector. The single-stranded DNA is purified and the region to be mutated identified by DNA sequencing. A short oligonucleotide is then synthesized, complementary to the relevant region, but containing the desired nucleotide alteration (Figure 11.23a). Despite this mismatch the oligonucleotide will anneal to the single-stranded DNA and act as a primer for complementary strand synthesis by a DNA polymerase (Figure 11.23b). This

strand synthesis reaction is continued until an entire new strand is made and the recombinant molecule is completely double-stranded.

After introduction, by transfection, into competent *E. coli* cells, DNA replication produces numerous copies of the recombinant DNA molecule. The semi-conservative nature of DNA replication means that half the double-stranded molecules that are produced are unmutated in both strands, whereas half are mutated in both strands. Similarly, half the resulting phage progeny carry copies of the unmutated molecule and half carry the mutation. The phages produced by the transfected cells are plated onto solid agar so that plaques are produced. Half the plaques should contain the original recombinant molecule, and half the mutated version. Which are which is determined by plaque hybridization, using the oligonucleotide as the probe, and employing very strict conditions so that only the completely base-paired hybrid is stable.

Cells infected with M13 vectors do not lyse, but instead continue to divide (p. 19). The mutated gene can therefore be expressed in the host *E. coli* cells, resulting in production of recombinant protein. The protein coded by the mutated gene can be purified from the recombinant cells and its properties studied. The effect of a single base pair mutation on the activity of the protein can therefore be assessed.

Other methods of creating a point mutation in a cloned gene

The oligonucleotide-directed procedure illustrated in Figure 11.23 is an effective way of creating a single point mutation in a cloned gene. But what if the objective is to make a number of changes to the sequence of the gene? The oligonucleotide procedure could, of course, be repeated several times, introducing a new mutation at each stage, but this would be a very lengthy process.

An alternative for short genes (up to about 1 kb) would be to construct the gene in the test tube, placing mutations at all the desired positions. This approach is feasible now that oligonucleotides of 150 units and longer can be made by chemical synthesis (p. 140). The gene is constructed by synthesizing a series of partially overlapping oligonucleotides, the sequences of these oligonucleotides making up the desired sequence for the gene. The overlaps can be partial, rather than complete, because the gaps can be filled in with a DNA polymerase, and the final phosphodiester bonds synthesized with DNA ligase, to create the completed, double-stranded gene (Figure 11.24). If restriction sites are included in the end-sequences of the gene, then treatment with the appropriate enzyme will produce sticky ends that allow the gene to be ligated into a cloning vector. This procedure is called **artificial gene synthesis**.

PCR can also be used to create mutations in cloned genes, though like oligonucleotide-directed mutagenesis, only one mutation can be created per experiment. Various procedures have been devised, one of which is shown in Figure 11.25. In this example, the starting DNA molecule is amplified by two PCRs. In each of these, one primer is normal and forms a fully base-paired hybrid with the template DNA, but the second is mutagenic, as it contains a single base-pair mismatch corresponding to the mutation that we wish to introduce into the DNA sequence. This mutation is therefore present in the two PCR products, each of which represents one half of the starting DNA molecule. The two PCR products are mixed together and a final PCR cycle carried out. In this cycle, complementary strands from the two products anneal to one another and are then extended by the polymerase, producing the full-length, mutated DNA molecule. This technique, and related ones using PCR, is very quick and easy to carry out but a major problem is caused by the high error rate of the *Taq* polymerase used in PCR (p. 157). This error rate makes it likely that not only the desired mutation, but also

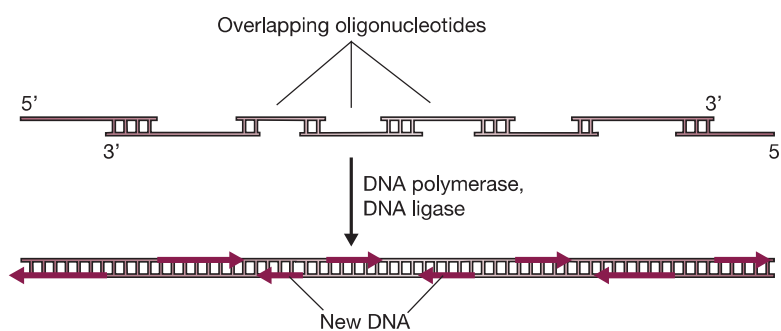


Figure 11.24

Artificial gene synthesis.

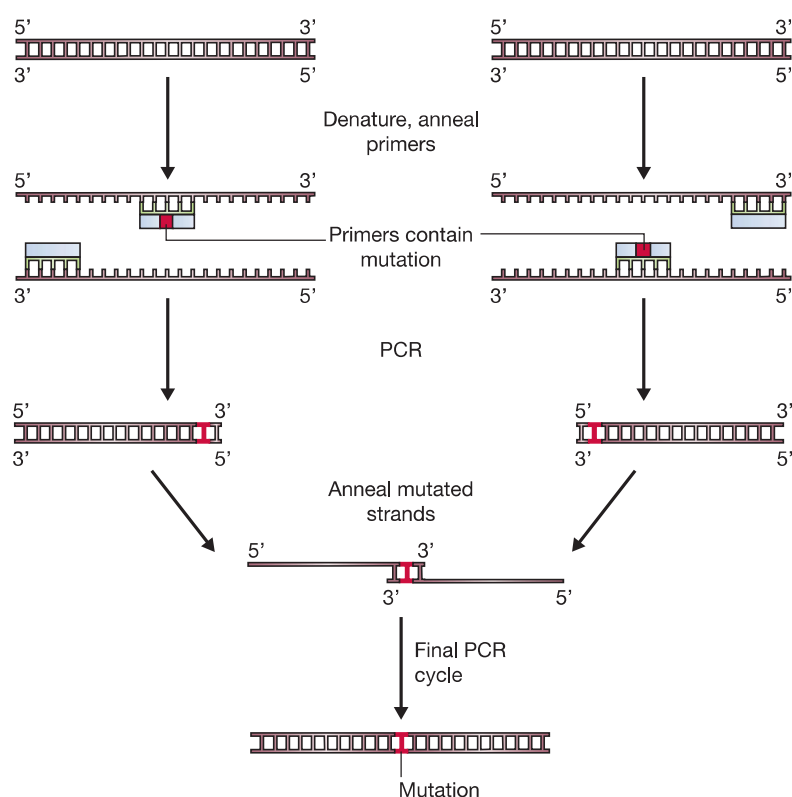


Figure 11.25

One method for using PCR to create a directed mutation.

random ones, will be present at the end of the experiment. The PCR product therefore has to be cloned and the sequences of individual clones checked to find one that has just the desired mutation.

The potential of in vitro mutagenesis

In vitro mutagenesis has remarkable potential, both for pure research and for applied biotechnology. For example, the biochemist can now ask very specific questions about the way that protein structure affects the action of an enzyme. In the past, it has been possible through biochemical analysis to gain some idea of the identity of the amino

acids that provide the substrate binding and catalytic functions of an enzyme molecule. Mutagenesis techniques provide a much more detailed picture by enabling the role of each individual amino acid to be assessed by replacing it with an alternative residue and determining the effect this has on the enzymatic activity.

The ability to manipulate enzymes in this way has resulted in dramatic advances in our understanding of biological catalysis and has led to the new field of **protein engineering**, in which mutagenesis techniques are used to develop new enzymes for biotechnological purposes. For example, careful alterations to the amino acid sequence of subtilisin, an enzyme used in biological washing powders, have resulted in engineered versions with greater resistances to the thermal and bleaching (oxidative) stresses encountered in washing machines.

Further reading

FURTHER READING

- Berk, A.J. (1989) Characterization of RNA molecules by S1 nuclease analysis. *Methods in Enzymology*, 180, 334–347.
- Fried, M. & Crothers, D.M. (1981) Equilibria and kinetics of *lac* repressor–operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Research*, 9, 6505–6525. [Gel retardation.]
- Frohman, M.A., Dush, D.K. & Martin, G.R. (1988) Rapid production of full length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences of the USA*, 85, 8998–9002. [An example of RACE.]
- Galas, D.J. & Schmitz, A. (1978) DNase footprinting: a simple method for the detection of protein–DNA binding specificity. *Nucleic Acids Research*, 5, 3157–3170.
- Garner, M.M. & Rezin, A. (1981) A gel electrophoretic method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Research*, 9, 3047–3060. [Gel retardation.]
- Hendrickson, W. & Schleif, R. (1985) A dimer of AraC protein contacts three adjacent major groove regions at the Ara I DNA site. *Proceedings of the National Academy of Sciences of the USA*, 82, 3129–3133. [An example of the modification interference assay.]
- Kunkel, T.A. (1985) Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proceedings of the National Academy of Sciences of the USA*, 82, 488–492. [Oligonucleotide-directed mutagenesis.]
- Matzke, A.J.M., Stöger, E.M., Scherthaner, J.P. & Matzke, M.A. (1990) Deletion analysis of a zein gene promoter in transgenic tobacco plants. *Plant Molecular Biology*, 14, 323–332.
- Paterson, B.M., Roberts, B.E. & Kuff, E.L. (1977) Structural gene identification and mapping by DNA.mRNA hybrid-arrested cell-free translation. *Proceedings of the National Academy of Sciences of the USA*, 74, 4370–4374.
- Pellé, R. & Murphy, N.B. (1993) Northern hybridization: rapid and simple electrophoretic conditions. *Nucleic Acids Research*, 21, 2783–2784.
- Tsien, R. (1998) The green fluorescent protein. *Annual Reviews of Biochemistry*, 67, 509–544. [A reporter gene system.]