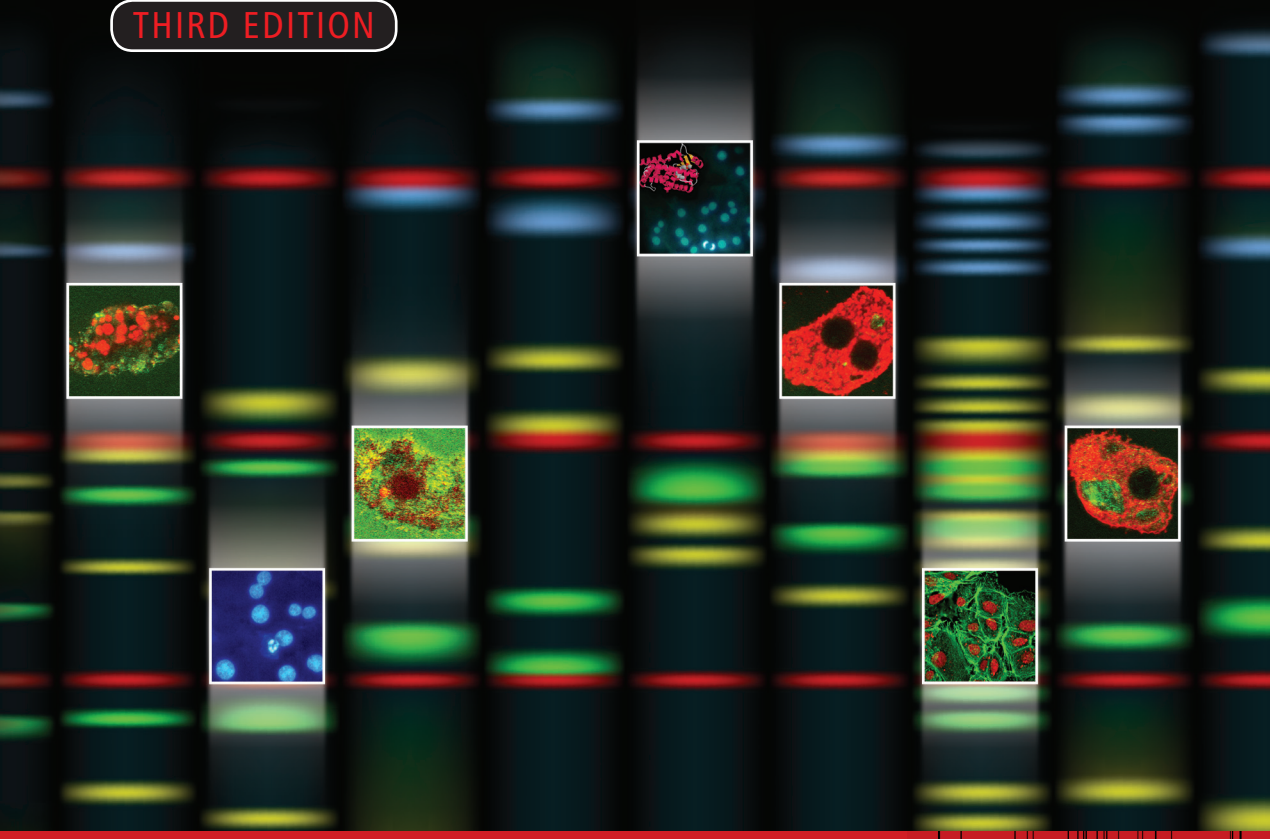
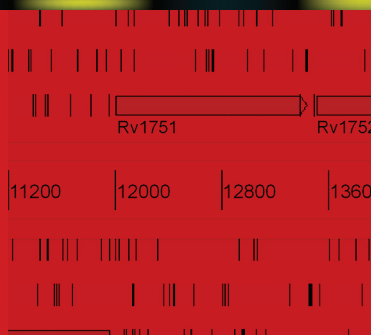
JEREMY W. DALE | MALCOLM VON SCHANTZ | NICK PLANT

FROM GENES TO GENOMES

CONCEPTS AND APPLICATIONS OF DNA TECHNOLOGY

THIRD EDITION





WILEY-BLACKWELL

From Genes to Genomes

Third Edition

From Genes to Genomes

Third Edition

Concepts and Applications of DNA Technology

Jeremy W. Dale, Malcolm von Schantz and **Nick Plant**

University of Surrey, UK

WILEY-BLACKWELL

A John Wiley & Sons, Ltd., Publication

This edition first published 2012

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

Registered office
John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial offices
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data

Dale, Jeremy, Professor.
From genes to genomes : concepts and applications of DNA technology / Jeremy W. Dale, Malcolm von Schantz, and Nick Plant. – 3rd ed.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-68386-6 (cloth) – ISBN 978-0-470-68385-9 (pbk.)
I. Schantz, Malcolm von. II. Plant, Nick. III. Title.
[DNLM: 1. Genetic Engineering. 2. Cloning, Molecular. 3. DNA, Recombinant. QU 450]
LC classification not assigned
660.6′5–dc23

2011030219

A catalogue record for this book is available from the British Library.

This book is published in the following electronic formats: ePDF 9781119953159; ePub 9781119954279; Mobi 9781119954286

Set in 10.5/13pt Times by Aptara Inc., New Delhi, India.

First Impression 2012

Contents

Preface

The first edition of this book was published in 2002. By the time of the second edition (2007) the emphasis had moved away from just cloning genes, to embrace a wider range of technologies, especially genome sequencing, the polymerase chain reaction and microarray technology. The revolution has continued unabated, indeed even accelerating, not least with the advent of high-throughput genome sequencing. In this edition we have tried to introduce readers to the excitement engendered by the latest developments – but this poses a considerable challenge. Our aim has been to keep the book to an accessible size, so including newer technologies inevitably means discarding some of the older ones. Some might maintain that we could have gone further in that direction. Some methods that have been kept are no longer as important as they once were, and maybe there is an element of sentimentality in keeping them – but there is some virtue in retaining a balance so that we can maintain a degree of historical perspective. There is a need to understand, to some extent, how we got to the position we are now in, as well as trying to see where we are going.

The main title of the book, *From Genes to Genomes*, is derived from the progress of this revolution. It also indicates a recurrent theme within the book, in that the earlier chapters deal with analysis and investigation at the level of individual genes, and then later on we move towards genome-wide studies – ending up with a chapter directed at the whole organism.

Dealing only with the techniques, without the applications, would be rather dry. Some of the applications are obvious – recombinant product formation, genetic diagnosis, transgenic plants and animals, and so on – and we have attempted to introduce these to give you a flavour of the advances that continue to be made, but at the same time without burdening you with excessive detail. Equally important, possibly more so, are the contributions made to the advance of fundamental knowledge in areas such as developmental studies and molecular phylogeny.

The purpose of this book is to provide an introduction to the concepts and applications of this rapidly moving and fascinating field. In writing it, we had in mind its usefulness for undergraduate students in the biological and biomedical sciences (who we assume will have a basic grounding in molecular biology). However, it will also be relevant for many others, ranging from research workers and teachers who want to update their knowledge of related areas to anyone who would like to understand rather more of the background to current controversies about the applications of some of these techniques.

Jeremy W. Dale
Malcolm von Schantz
Nick Plant

1
From Genes to Genomes

1.1 Introduction

The classical approach to genetics starts with the identification of variants that have a specific *phenotype*, i.e., they differ from the *wildtype* in some way that can be seen (or detected in other ways) and defined. For Gregor Mendel, the father of modern genetics, this was the appearance of his peas (e.g., green versus yellow, or round versus wrinkled). One of the postulates he arrived at was that these characteristics assorted independently of one another. For example, when crossing one type of pea that produces yellow, wrinkled peas with another that produces green, round peas, the first generation (F_1) are all round and yellow (because round is dominant over wrinkled, and yellow is dominant over green). In the second (F_2) generation, there is a 3 : 1 mixture of round versus wrinkled peas, and independently a 3 : 1 mixture of yellow to green peas.

Of course Mendel did not know why this happened. We now know that if two genes are located on different chromosomes, which will segregate independently during meiosis, the genes will be distributed independently amongst the progeny. Independent assortment can also happen if the two genes are on the same chromosome, but only if they are so far apart that any recombination between the homologous chromosomes will be sufficient to reassort them independently. However, if they are quite close together, recombination is less likely, and they will therefore tend to remain associated during meiosis. They will therefore be inherited together. We refer to genes that do *not* segregate independently as *linked*; the closer they are, the greater the degree of linkage, i.e., the more likely they are to stay together during meiosis. Measuring the degree of linkage (*linkage analysis*) is a central tool in classical genetics, in that it provides a way of mapping genes, i.e., determining their relative position on the chromosome.

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

Bacteria and yeasts provide much more convenient systems for genetic analysis, because they grow quickly, as unicellular organisms, on defined media. You can therefore use chemical or physical mutagens (such as ultraviolet irradiation) to produce a wide range of mutations, and can select specific mutations from very large pools of organisms – remembering that an overnight culture of *Escherichia coli* will contain some 10^9 bacteria per millilitre. So we can use genetic techniques to investigate detailed aspects of the physiology of such cells, including identifying the relevant genes by mapping the position of the mutations.

For multicellular organisms, the range of phenotypes is even greater, as there are then questions concerning the development of different parts of the organism, and how each individual part influences the development of others. However, animals have much longer generation times than bacteria, and using millions of animals (especially mammals) to identify the mutations you are interested in is logistically impossible, and ethically indefensible. Human genetics is even more difficult as you cannot use selected breeding to map genes; you have to rely on the analysis of real families, who have chosen to breed with no consideration for the needs of science. Nevertheless, classical genetics has contributed extensively to the study of developmental processes, notably in the fruit fly *Drosophila melanogaster*, where it is possible to study quite large numbers of animals, due to their relative ease of housing and short generation times, and to use mutagenic agents to enhance the rate of variation.

However, these methods suffered from a number of limitations. In particular, they could only be applied, in general, to mutations that gave rise to a phenotype that could be defined in some way, including shape, physiology, biochemical properties or behaviour. Furthermore, there was no easy way of characterizing the nature of the mutation. The situation changed radically in the 1970s with the development of techniques that enabled DNA to be cut precisely into specific fragments, and to be joined together, enzymatically –techniques that became known variously as genetic manipulation, genetic modification, genetic engineering or recombinant DNA technology. The term ‘gene cloning’ is also used, since joining a fragment of DNA with a vector such as a plasmid that can replicate in bacterial cells enabled the production of a bacterial strain (a clone) in which all the cells contained a copy of this specific piece of DNA. For the first time, it was possible to isolate and study specific genes. Since such techniques could be applied equally to human genes, the impact on human genetics was particularly marked.

The revolution also depended on the development of a variety of other molecular techniques. The earliest of these (actually predating gene cloning) was *hybridization*, which enabled the identification of specific DNA sequences on the basis of their sequence similarity. Later on came methods for determining the sequence of these DNA fragments, and the polymerase chain

reaction (PCR), which provided a powerful way of amplifying specific DNA sequences. Combining those advances with automation, plus the concurrent advance in computer power, led to the determination of the full genome sequence of many organisms, including the human genome, and thence to enormous advances in understanding the roles of genes and their products. In recent years, sequencing technology has advanced to a stage where it is now a routine matter to sequence the full genome of many individuals, and thus attempt to pinpoint the causes of the differences between them, including some genetic diseases.

Furthermore, since these techniques enabled the cloning and expression of genes from any one organism (including humans) into a more amenable host, such as a bacterium, they allowed the use of genetically modified bacteria (or other hosts) for the production of human gene products, such as hormones, for therapeutic use. This principle was subsequently extended to the genetic modification of plants and animals – both by inserting foreign genes and by knocking out existing ones – to produce plants and animals with novel properties.

As is well known, the construction and use of genetically modified organisms (GMOs) is not without controversy. In the early days, there was a lot of concern that the introduction of foreign DNA into *E. coli* would generate bacteria with dangerous properties. Fortunately, this is one fear that has been shown to be unfounded. Due to careful design, genetically modified bacteria are, generally, not well able to cope with life outside the laboratory, and hence any GM bacterium released into the environment (deliberately or accidentally) is unlikely to survive for long. In addition, one must recognize that nature is quite capable of producing pathogenic organisms without our assistance – which history, unfortunately, has repeatedly demonstrated through disease outbreaks.

The debate on GMOs has now largely moved on to issues relating to genetically modified plants and animals. It is important to distinguish the *genetic modification* of plants and animals from *cloning* of plants and animals. The latter simply involves the production of genetically identical individuals; it does not involve any genetic modification whatsoever. (The two technologies can be used in tandem, but that is another matter.) There are ethical issues to be considered, but cloning plants and animals is not the subject of this book.

Currently, the debate on genetic modification can be envisaged as largely revolving around two factors: food safety and environmental impact. The first thing to be clear about is that there is no imaginable reason why genetic modification, per se, should make a foodstuff hazardous in any way. There is no reason to suppose that cheese made with rennet from a genetically modified bacterium is any more dangerous than similar cheese made with ‘natural’ rennet. It is possible to imagine a risk associated with some genetically modified foodstuffs, due to unintended stimulation of the production of natural

toxins – remembering, for example, that potatoes are related to deadly nightshade. But this can happen equally well (or perhaps is even more likely) with conventional cross-breeding procedures for developing new strains, which are not always subject to the same degree of rigorous safety testing as GM plants.

The potential environmental impact is more difficult to assess. The main issue here is the use of genetic modification to make plants resistant to herbicides or to insect attack. When such plants are grown on a large scale, it is difficult to be certain that the gene in question will not spread to related wild plants in the vicinity (although measures can be taken to reduce this possibility), or the knock-on effect that such resistance may have on the ecosystem – if all the insects are killed, what will small birds and animals eat? But these concerns may be exaggerated. As with the bacterial example above, these genes will not spread significantly unless there is an evolutionary pressure favouring them. So we would not expect widespread resistance to weedkillers unless the plants are being sprayed with those weedkillers. There might be an advantage in becoming resistant to insect attack, but the insects concerned have been around for a long time, so the wild plants have had plenty of time to develop natural resistance anyway. In addition, targeted resistance in a group of plants may arguably have less environmental impact than the less targeted spraying of insecticides. We have to balance the use of genetically modified plants against the use of chemicals. If genetic modification of the plants means a reduction in the use of environmentally damaging chemicals, then that is a tangible benefit that could outweigh any theoretical risk.

The purpose of this book is to provide an introduction to the exciting developments that have resulted in an explosion of our knowledge of the genetics and molecular biology of all forms of life, from viruses and bacteria to plants and mammals, including of course ourselves – developments that continue as we write. We hope that it will convey some of the wonder and intellectual stimulation that this science brings to its practitioners.

1.2 Basic molecular biology

In this book, we assume you already have a working knowledge of the basic concepts of molecular and cellular biology. This section serves as a reminder of the key aspects that are especially relevant to this book.

1.2.1 The DNA backbone

Manipulation of nucleic acids in the laboratory is based on their physical and chemical properties, which in turn are reflected in their biological function. Intrinsically, DNA is a remarkably stable molecule. Indeed, DNA of sufficiently high quality to be analysed has been recovered from frozen

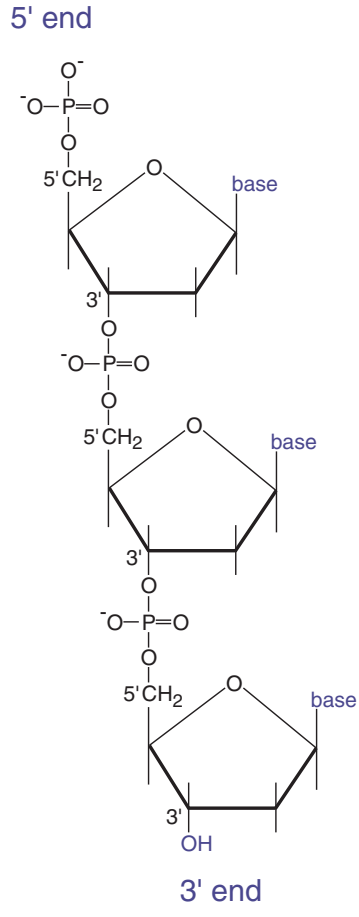


Figure 1.1 DNA backbone.

mammoths thousands of years old. This stability is provided by the robust phosphate–sugar backbone in each DNA strand, in which the phosphate links the 5′ position of one sugar to the 3′ position of the next (Figure 1.1). The bonds between these phosphorus, oxygen and carbon atoms are all *covalent bonds*, meaning they are strong interactions that require significant energy to break. Hence, the controlled degradation of DNA requires enzymes (nucleases) that catalyse the breaking of these covalent bonds. These enzymes are divided into *endonucleases*, which attack internal sites in a DNA strand, and *exonucleases*, which nibble away at the ends. (We can for the moment ignore other enzymes that attack, for example, the bonds linking the bases to the sugar residues.) Some of these enzymes are non-specific, and lead to a generalized destruction of DNA. It was the discovery of *restriction endonucleases* (or *restriction enzymes*), which cut DNA strands at specific positions,

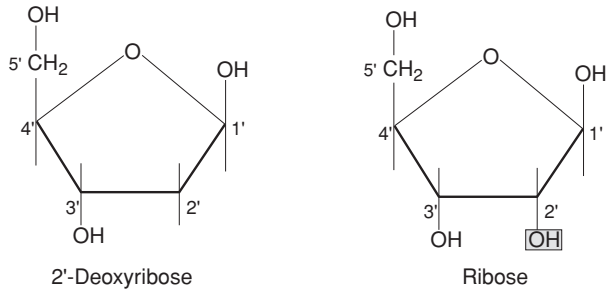


Figure 1.2 Nucleic acid sugars.

coupled with *DNA ligases*, which can join two double-stranded DNA molecules together, that opened up the possibility of *recombinant DNA technology* (‘*genetic engineering*’).

RNA molecules, which contain the sugar ribose (Figure 1.2), rather than the deoxyribose found in DNA, are less stable than DNA, often surviving only minutes within the cell. They show greater susceptibility to attack by nucleases (*ribonucleases*), and are also more susceptible to chemical degradation, especially by alkaline conditions.

1.2.2 The base pairs

In addition to the sugar (2′-deoxyribose) and phosphate, DNA molecules contain four nitrogen-containing bases (Figure 1.3): two pyrimidines, thymine (T) and cytosine (C), and two purines, guanine (G) and adenine (A). It should be noted that other bases can be incorporated into synthetic DNA in the laboratory, and sometimes others occur naturally, but T, C, G and A are the major DNA bases. Because the purines are bigger than the pyrimidines, a regular double helix requires a purine in one strand to be matched by a pyrimidine in the other. Furthermore, the regularity of the double helix requires specific hydrogen bonding between the bases so that they fit together, with an A opposite a T, and a G opposite a C (Figure 1.4). We refer to these pairs of bases as *complementary*, and hence to each strand as the *complement* of the other.

Note that the two DNA strands run in opposite directions. In a conventional representation of a double-stranded sequence the ‘top’ strand has a 5′ hydroxyl group at the left-hand end (and is said to be written in the 5′ to 3′ direction), while the ‘bottom’ strand has its 5′ end at the right-hand end. Because the two strands are complementary, there is no information in one strand that cannot be deduced from the other one. Therefore, to save space, the convention is to represent a double-stranded DNA sequence by showing the sequence of only one strand, in the 5′ to 3′ direction. The sequence of the

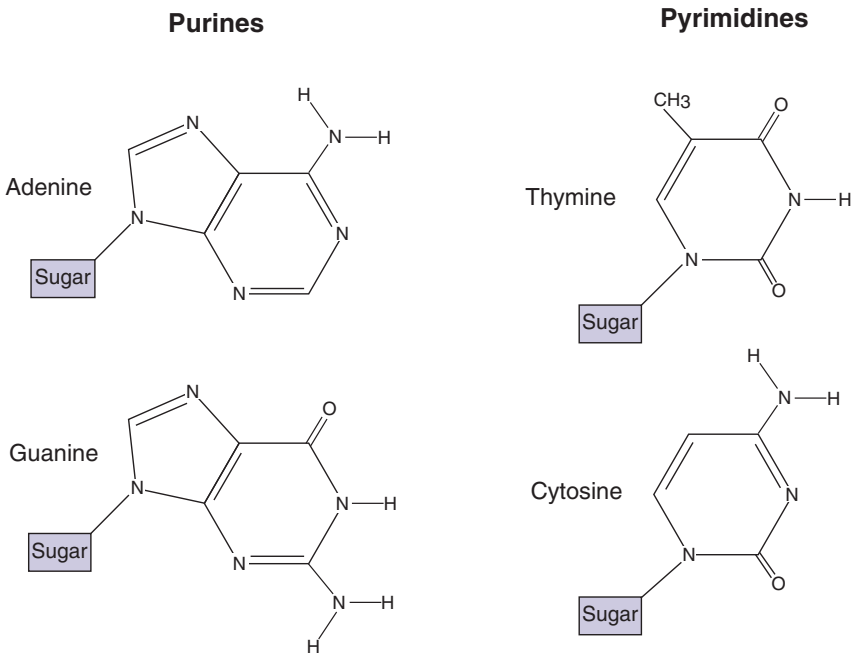


Figure 1.3 Nucleic acid bases.

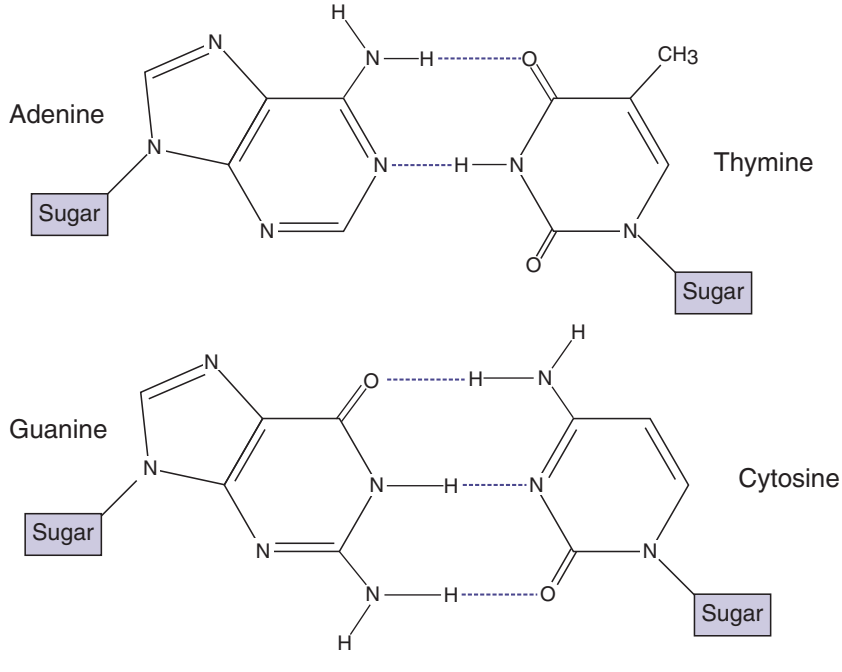


Figure 1.4 Base-pairing in DNA.

Box 1.1 Complementary sequences

DNA sequences are often represented as the sequence of just one of the two strands, in the 5′ to 3′ direction, reading from left to right. Thus the double-stranded DNA sequence

$$5' - \mathrm{AGGCTG} - 3'$$
$$3' - \mathrm{TCCGAC} - 5'$$

would be shown as AGGCTG, with the orientation (i.e., the position of the 5′ and 3′ ends) being implied.

To get the sequence of the other (complementary) strand, you must not only change the A and G residues to T and C (and vice versa), but you must also reverse the order. So in this example, the complement of AGGCTG is CAGCCT, reading the lower strand from right to left (again in the 5′ to 3′ direction).

second strand is inferred from that, and you must remember that the second strand runs in the opposite direction. Thus a single strand sequence written as AGGCTG (or more fully 5′AGGCTG3′) would have as its complement CAGCCT (5′CAGCCT3′) (see Box 1.1).

Thanks to this base-pairing arrangement, the two strands can be separated intact – both in the cell and in the test tube – under conditions that, although disrupting the relatively weak hydrogen bonds that exist between the bases on complementary strands, are much too mild to pose any threat to the covalent bonds that join nucleotides within a single strand. Such a separation is referred to as *denaturation* of DNA, and unlike the denaturation of many proteins it is reversible. Because of the complementarity of the base pairs, the strands will easily join together again and *renature* by reforming their hydrogen bonds. In the test tube, DNA is readily denatured by heating, and the denaturation process is therefore often referred to as *melting*, even when it is accomplished by means other than heat (e.g. by NaOH). Denaturation of a double-stranded DNA molecule occurs over a short, specific temperature range, and the midpoint of that range is defined as the *melting temperature* (T_m). This is influenced by the base composition of the DNA. Since guanine:cytosine (GC) base pairs have three hydrogen bonds, they are stronger (i.e., melt less easily) than adenine:thymine (AT) pairs, which have only two. It is therefore possible to estimate the melting temperature of a DNA fragment if you know the sequence. These considerations are important in understanding the technique known as *hybridization*, in which *gene probes* are

used to detect specific nucleic acid sequences. We will look at hybridization in more detail in Chapter 3.

In addition to the hydrogen bonds, the double-stranded DNA structure is maintained by *hydrophobic interactions* between the bases. The hydrophobic nature of the bases means that a single-stranded structure, in which the bases are exposed to the aqueous environment, is unstable. Pairing of the bases enables them to be removed from interaction with the surrounding water, introducing significant stability to the DNA helix. In contrast to hydrogen bonding, hydrophobic interactions are relatively non-specific. Thus, nucleic acid strands will tend to stick together even in the absence of specific base-pairing, although the specific interactions make the association stronger. The specificity of the interaction can therefore be increased by the use of chemicals (such as formamide) that reduce the hydrophobic interactions.

What happens if there is only a single nucleic acid strand? This is normally the case with RNA, but single-stranded forms of DNA (ssDNA) also exist, in some viruses, for example. A single-stranded nucleic acid molecule will tend to fold up on itself to form localized double-stranded regions, producing structures referred to as hairpins or stem-loop structures (Figure 1.5). This has the effect of removing the bases from interaction with the surrounding water, again increasing stability. At room temperature, in the absence of denaturing agents, a single-stranded nucleic acid will normally consist of a complex set of such localized secondary structure elements; this is especially evident with RNA molecules.

A further factor is the negative charge on the phosphate groups in the nucleic acid backbone. This works in the opposite direction to the hydrogen bonds and hydrophobic interactions; the strong negative charge on the DNA strands causes electrostatic repulsion that tends to make the two strands

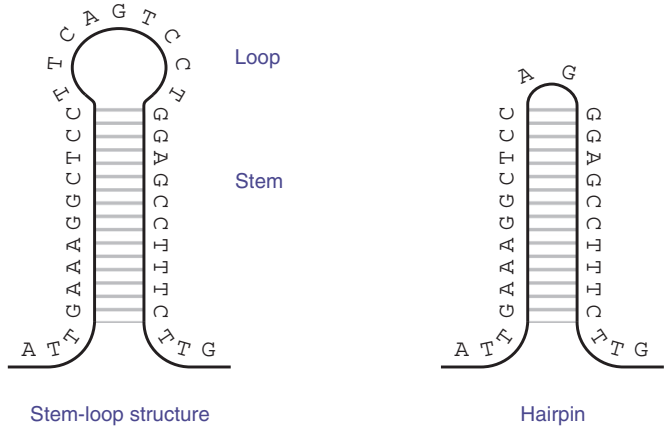


Figure 1.5 Stem-loop and hairpin structures.

repel each other. In the presence of salt, this effect is counteracted by a cloud of counterions surrounding the molecule, neutralizing the negative charge on the phosphate groups. However, if you reduce the salt concentration, any weak interactions between the strands will be disrupted by electrostatic repulsion. Hence, at low ionic strength, the strands will only remain together if the hydrogen bonding is strong enough, and therefore we can use low salt conditions to increase the specificity of hybridization (see Chapter 3). Of course, within the cell the salt concentration is such that the double-stranded DNA is quite stable.

1.2.3 RNA structure

Chemically, RNA is very similar to DNA. The fundamental chemical difference is that whereas DNA contains 2′-deoxyribose (i.e., ribose without the hydroxyl group at the 2′ position) in its backbone, the RNA backbone contains ribose (Figure 1.6). This slight difference has a powerful effect on some properties of the RNA molecule, especially on its stability. For example, RNA is destroyed under alkaline conditions while DNA is stable. Although the DNA strands will separate, they will remain intact and capable of renaturation when the pH is lowered again. However, under such conditions, RNA will quickly be destroyed. A further difference between RNA and DNA is that the former contains uracil rather than thymine (Figure 1.6).

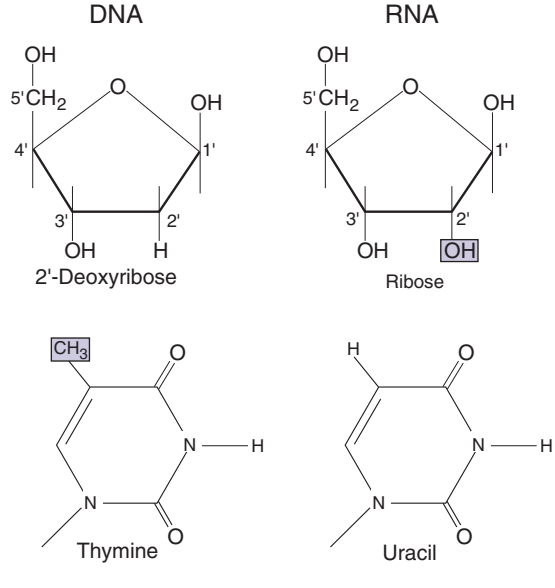


Figure 1.6 Differences between DNA and RNA.

Generally, while most of the DNA we encounter is double-stranded, most of the RNA we meet consists of a single polynucleotide strand. However, DNA can also exist as a single-stranded molecule, and RNA is able to form double-stranded molecules. Thus, this distinction between RNA and DNA is not an inherent property of the nucleic acids themselves, but is a reflection of the natural roles of RNA and DNA in the cell, and of the method of production. In all *cellular* organisms (i.e., excluding viruses), DNA is the inherited material responsible for the genetic composition of the cell, and the replication process that has evolved is based on a double-stranded molecule. By contrast, the roles of RNA in the cell do not require a second strand, and indeed the presence of a second, complementary, strand would preclude its role in protein synthesis. However, there are some viruses that have double-stranded RNA (dsRNA) as their genetic material, as well as some viruses with single-stranded RNA. In addition, some viruses (as well as some plasmids) replicate via single-stranded DNA forms. Double-stranded RNA is also important in the phenomenon known as RNA interference, which we will come to in later chapters.

1.2.4 Nucleic acid synthesis

We do not need to consider here all the details of how nucleic acids are synthesized. The fundamental features that we need to remember are summarized in Figure 1.7, which shows the addition of a nucleotide to the growing end (3′-OH) of a DNA strand. The substrate for this reaction is the relevant deoxynucleotide triphosphate (dNTP), i.e., the one that makes the correct base-pair with the corresponding residue on the template strand. The DNA strand is always extended at the 3′-OH end, thus the nucleotide strand grows in the 5′ to 3′ direction. For this reaction to occur it is essential that the existing residue at the 3′-OH end, to which the new nucleotide is to be added, is accurately base-paired with its partner on the other strand.

RNA synthesis occurs in much the same way, as far as this simplistic description goes, except that of course the substrates are nucleotide triphosphates (NTPs) rather than the deoxynucleotide triphosphates (dNTPs). There is one very important difference though. DNA synthesis only occurs by extension of an existing strand – it always needs a *primer* to get it started. In contrast, RNA polymerases are capable of starting a new RNA strand, complementary to its template, from scratch, given the appropriate signals.

1.2.5 Coiling and supercoiling

DNA can be denatured and renatured, deformed and reformed, and still retain unaltered function. This is a necessary feature, because such a large molecule as DNA will need to be packaged if it is to fit within the cell that

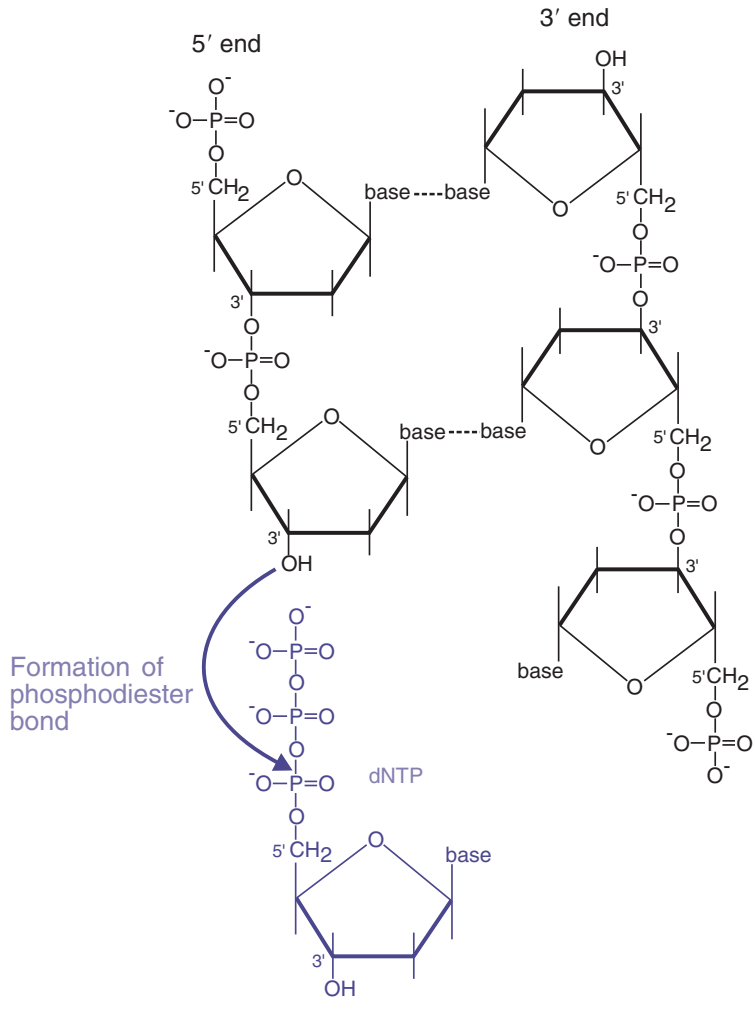


Figure 1.7 DNA synthesis.

it controls. The DNA of a human chromosome, if it were stretched into an unpackaged double helix, would be several centimetres long. Thus, cells are dependent on the packaging of DNA into modified configurations for their very existence.

Double-stranded DNA, in its relaxed state, normally exists as a right-handed double helix with one complete turn per 10 base pairs; this is known as the *B form* of DNA. Hydrophobic interactions between consecutive bases on the same strand contribute to this winding of the helix, as the bases are brought closer together enabling a more effective exclusion of water from interaction with the hydrophobic bases.

The DNA double helix can exist in other forms, notably the *A form* (also right-handed but more compact, with 11 bases per turn) and *Z-DNA*, which is a left-handed double helix with a more irregular appearance (a zigzag structure, hence its designation). However, that is not the complete story. Higher orders of conformation are known to exist. The double helix is in turn coiled on itself – an effect known as *supercoiling*. There is an interaction between the coiling of the helix and the degree of supercoiling. As long as the ends are fixed, changing the degree of coiling will alter the amount of supercoiling, and vice versa. DNA *in vivo* is constrained; the ends are not free to rotate. This is most obviously true of circular DNA structures such as (most) bacterial plasmids, but the rotation of linear molecules (other than very short oligonucleotides) is also constrained within the cell. The net effect of coiling and supercoiling (a property known as the *linking number*) is therefore fixed, and cannot be changed without breaking one of the DNA strands. In nature, there are enzymes known as topoisomerases (including DNA gyrase in bacteria) that do just that: they break the DNA strands, and then in effect rotate the ends and reseal them. This alters the degree of winding of the helix, and thus affects the supercoiling of the DNA. Topoisomerases also have an ingenious use in the laboratory, which we will consider in Chapter 2.

The plasmids that we will be referring to frequently in later pages are naturally supercoiled when they are isolated from the cell. However, if one of the strands is broken at any point, the DNA is then free to rotate at that point and can therefore relax into a non-supercoiled form. This is known as an *open circular* form (in contrast to the *covalently closed circular* form of the native plasmid).

1.3 What is a gene?

The definition of a ‘gene’ is rather imprecise. Its origins go back to the early days of genetics, when it was used to describe the unit of inheritance of a phenotype. This meaning persists in non-scientific usage, rather loosely, as the ‘gene for blue eyes’, or ‘the gene for red hair’. As it became realized that many characteristics were determined by the presence or properties of individual proteins, the definition became refined to relate to the chromosomal region that carried the information for that protein, leading to the concept of ‘one gene, one protein’. As the study of genetics and biochemistry progressed further, it was realized that many proteins consist of several distinct polypeptides, and that the chromosomal regions coding for the different polypeptides could be distinguished genetically. So the definition was refined further to mean a piece of DNA containing the information for a single specific polypeptide (‘one gene, one polypeptide’). With the advent of

DNA sequencing, it became possible to consider a gene in molecular terms. So we now often use the term 'gene' as being synonymous with 'open reading frame' (ORF), i.e., the region between the start and stop codons. In bacteria, this is (usually) a simple uninterrupted sequence, but in eukaryotes, the presence of introns (see below) makes this definition more difficult, since the region of the chromosome that contains the information for a specific polypeptide may be many times longer than the actual coding sequence. We also have to be careful as we may want to refer to the whole transcribed region, which will be longer than the translated open reading frame, or indeed we may want to include the control regions that are necessary for the start of transcription.

Furthermore, this definition, by focusing solely on the regions that code for proteins (or polypeptides), is too limited in its scope. It ignores many regions of DNA that, although not coding for proteins, are nevertheless important for the viability of the cell, or influence the phenotype in other ways. The most obvious of these are DNA sequences that are templates for so-called non-coding RNA molecules. The most well known of these are ribosomal and transfer RNA, although we will encounter other RNA molecules that play significant roles in gene regulation and other activities. Other DNA regions are important in gene regulation because they act as binding sites for regulatory proteins.

In organisms with small genomes, such as bacteria, a high proportion of the genome is accounted for by these coding and regulatory regions (together with elements that may be described as 'parasitic' such as integrated viruses and insertion sequences). There is relatively little DNA to which we can ascribe no likely function, compared to eukaryotic cells, and especially animal and plant cells with much larger genomes, where there is a much higher proportion of non-coding DNA. But we should not be too hasty in writing these sequences off as 'junk'. Increasingly, much of it is recognized as having important functions within the cell. These include enabling the DNA to be folded correctly and in ensuring that the coding regions are available for expression under the appropriate conditions, as well as coding for small (non-translated) RNA molecules that play a major role in modulating gene expression.

Thus, we have to accept that it is not possible to produce an entirely satisfactory definition of the word 'gene'. However, this is rarely a serious problem. We just have to be careful how we use it depending on whether we are discussing only the coding region (ORF), or the length of sequence that is transcribed into mRNA (including untranslated regions), or whether we wish to include DNA regions with regulatory functions as well as coding sequences. In this context, we will also encounter the words *allele* and *locus*. A locus is used in the same way as 'gene' in the broad meaning – i.e., it could be a coding sequence, a regulatory region, or any other region we wish to

consider. An allele is one version of that locus. So variation of a genetic characteristic between individuals would be due to different alleles at one locus (or several loci).

1.4 Information flow: gene expression

The way in which genes are expressed is so central to the subsequent material in this book that it is worth reviewing briefly the salient features. The basic dogma (Figure 1.8) is that while DNA is the fundamental genetic material (ignoring RNA viruses) that carries information from one generation to the next, its effect on the characteristics of the cell requires firstly its copying into RNA (*transcription*), and then the *translation* of the mRNA into a polypeptide by ribosomes. Further processes are required before its proper activity can be manifested: these include the folding of the polypeptide, possibly in association with other subunits to form a multisubunit protein, and in some cases modification, for example by glycosylation or phosphorylation. It should be remembered that in some cases, RNA rather than protein is the final product of a gene (e.g., ribosomal and transfer RNA molecules).

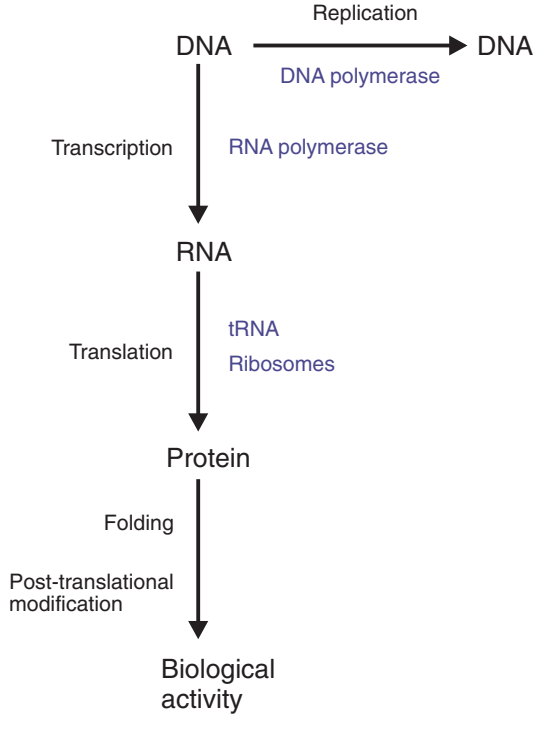


Figure 1.8 Information flow.

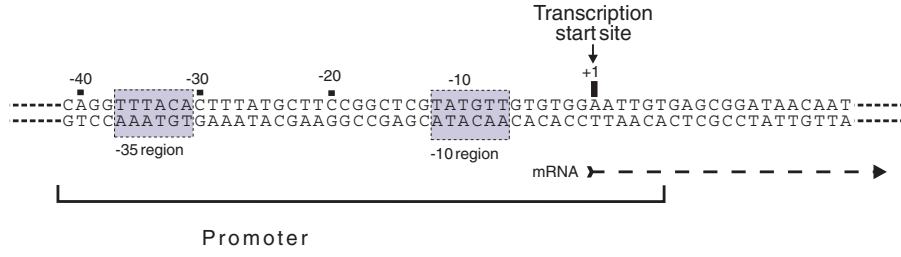


Figure 1.9 Structure of the promoter region of the *lac* operon. Note that the −35 and −10 regions of the *lac* promoter do not correspond exactly to the consensus sequences TTGACA and TATAAT respectively.

1.4.1 Transcription

Transcription is carried out by RNA polymerase. RNA polymerase recognizes and binds to a specific sequence (the *promoter*), and initiates the synthesis of mRNA from an adjacent position.

A typical bacterial promoter carries two *consensus* sequences (i.e., sequences that are closely related in all genes): TTGACA centred at position −35 (i.e., 35 bases before the transcription start site), and TATAAT at −10 (Figure 1.9). It is important to understand the nature of a consensus: few bacterial promoters have exactly the sequences shown, but if you line up a large number of promoters you will see that at any one position a large number of them have the same base (Box 1.2). The RNA polymerase has higher affinity for some promoters than others – depending not only on the exact nature of the two consensus sequences but also, to a lesser extent, on the sequence of a longer region of DNA. The nature and regulation of bacterial promoters, including the existence of alternative types of promoters, is considered further in Chapter 5.

In eukaryotes, by contrast, the promoter is a considerably larger area around the transcription start site, where a number of *trans*-acting transcription factors (i.e., DNA-binding proteins encoded by genes in other parts of the genome) bind to various *cis*-acting elements (i.e., elements that affect the expression of the gene next to them) in a considerably more complex scenario. The need for this added complexity can easily be imagined; if cells carrying the same genome are differentiated into a multitude of cell types fulfilling very different functions, a very sophisticated control system is needed to provide each cell type with its specific repertoire of proteins, and to fine-tune the degree of expression for each one of them. Nonetheless, the promoter region, however simple or complex, gives rise to different levels of transcription of various genes. A further complexity in eukaryotes is that there are commonly additional regulatory elements known as *enhancers*,

Box 1.2 Examples of E. coli *promoters*

```
        -35                                 -10           1
TGGCGGTGTTGACATAAATA CCACTGGCGGTGATACTGAGCA CA   Lambda P_L
CGTGCGTGTTGACTATTTTA CCTCTGGCGGTGATAATGGTTG CA   Lambda P_R
TGCCGAAGTTGAGTATTTTT GCTGTATTTGTCATAATGACTCCTG   Lambda P_O
ATGAGCTGTTGACAATTAAT CATCGAACTAGTTAACTAGTACGCA   trp
CATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCG   lacI
CCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGG A   lacZ
CGTAACACTTTACAGCGGCG  CGTCATTTGATATGATGCGCC CG   tyr tRNA
        TTGACA                  TATAAT           consensus
```

Bases matching the –10 and –35 consensus sequences are boxed. Spaces are inserted to optimise the alignment. Note that the consensus is derived from a much larger collection of characterized promoters. Position 1 is the transcription start site.

which may further alter the level of transcription of one, or several, genes. By definition, enhancers are position and orientation independent, and are often remote from the actual start site of transcription by several thousand base pairs.

Eukaryotes have three different RNA polymerases. Only one of these, RNA polymerase II, is involved in the transcription of protein-coding transcripts, plus the transcription of a group of small non-coding RNAs called micro-RNAs, which we will encounter in Chapter 11. RNA polymerase I is responsible for the synthesis of large ribosomal RNAs, whilst RNA polymerase III makes small RNAs such as transfer RNA (tRNA) and 5S ribosomal RNA.

In eukaryotes, the primary transcript from a protein-coding gene, produced by RNA polymerase II, is called a *heterogeneous* or *heteronuclear RNA* (hnRNA). It is very short-lived as such, being rapidly processed in a number of steps called *maturation*. A specialized nucleotide *cap* is added to the 5′ end; this is the site recognized by the ribosomes in protein synthesis (see below). The precursor mRNA is cleaved at a specific site towards the 3′ end and a *poly-A tail,* consisting of a long sequence of adenosine residues, is added to the cut end. This is a specific process, governed by polyadenylation recognition sequences in the 3′ untranslated region. Nature's 'tagging' of mRNA molecules comes in very useful in the laboratory for the isolation of eukaryotic mRNA (see Chapter 3). This transcript contains intervening sequences (*introns*) between the *exons* that carry the coding information (see below). The final step is the process of *splicing*, by which the introns are removed and the exons are joined together. This process is quite complex; some introns are removed by specific proteins known as splicing factors, whereas other introns are removed independently through autocatalysis. To complicate things further, in some cases the transcript is edited, leading, for example, to the introduction of a tissue-specific earlier stop codon that is not encoded in the corresponding DNA.

In bacteria, the processes of transcription and translation take place in the same compartment and simultaneously. In eukaryotes, by contrast, the mature mRNA molecule is transported out of the nucleus to the cytoplasm, where translation takes place.

The resulting level of protein production is dependent on the amount of the specific mRNA available, rather than just the rate of production. The level of an mRNA species will be affected by its rate of degradation as well as by its rate of synthesis. In bacteria, most mRNA molecules are degraded quite quickly (with a half-life of only a few minutes), although some are much more stable. The instability of the majority of bacterial mRNA molecules means that bacteria can rapidly alter their profile of gene expression by changing the transcription of specific genes. By contrast, the lifespans of most eukaryotic mRNA molecules are measured in hours rather than minutes, although

some short-lived mRNA molecules do exist. This greater persistence of eukaryote mRNA molecules is, again, a reflection of the fact that an organism (or a cell) that is able to control its own environment to a substantial extent is subjected to less radical environmental changes. Consequently, mRNA molecules tend to be more stable in multicellular organisms than in, for example, yeast. Nonetheless, the principle remains – the level of an mRNA in a cell is a function of its production and degradation rates. We will discuss how to study and disentangle these parameters in Chapter 10.

1.4.2 Translation

In bacteria, translation starts when ribosomes bind to a specific site (the *ribosome binding site, RBS*), which is adjacent to the start codon. The sequence of the ribosome binding site (also known as the *Shine–Dalgarno sequence*) has been recognized as being complementary to the 3′ end of the 16S rRNA (Figure 1.10). The precise sequence of this site and its distance from the start codon affect the efficiency of translation, although in nature this is less important than transcriptional efficiency in determining the level of gene expression. Translation efficiency will also depend on the codon usage, i.e., the match between synonymous codons and the availability of tRNA that will recognize each codon. This concept is explored more fully in Chapter 7.

In bacterial systems, where transcription and translation occur in the same compartment of the cell, ribosomes will bind to the mRNA, a process known as *initiation*, as soon as the RBS has been synthesized. Thus, there will be a procession of ribosomes following close behind the RNA polymerase, translating the mRNA in the process of *elongation* as and when it is being produced. So, although the mRNA may be very short-lived, the bacteria are capable of producing substantial amounts of the corresponding polypeptide in a short time. Translation stops when a termination codon is reached.

In eukaryotes, the mechanism is much more complicated. Instead of binding just upstream of the initiation codon, the ribosome binds to the cap at

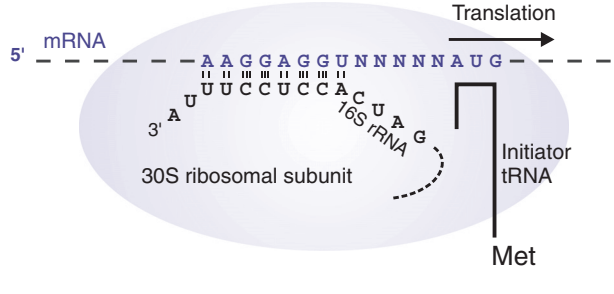


Figure 1.10 Bacterial ribosome binding site.

the 5′ end of the mRNA, and reads along the 5′ untranslated region (UTR) until it reaches an initiation codon. The sequence AUG may be encountered on the way without initiation, because the surrounding sequence is also important to define the start of protein synthesis. The fact that the 5′ UTR is scanned in its full length by the ribosome makes it an important region for specifying translation efficiency, and different secondary structures can have either a positive or a negative effect on the amount of protein that is produced. Sometimes, translation may be initiated at an internal ribosome entry site (IRES); the best studied of these occur in some viruses, but they also occur in some transcripts encoded naturally by the cell.

1.5 Gene structure and organisation

1.5.1 Operons

In bacteria, it is quite common for a group of genes to be transcribed from a single promoter into one long RNA molecule; this group of genes is known as an *operon* (Figure 1.11). If we are considering protein-coding genes, the transcription product, messenger RNA (mRNA), is then translated into a number of separate polypeptides. This can occur by the ribosomes reaching the stop codon at the end of one polypeptide-coding sequence, terminating translation and releasing the product before reinitiating (without dissociation from the mRNA). Alternatively, the ribosomes may attach independently to internal ribosome binding sites within the mRNA sequence. Generally, the genes involved are responsible for different steps in the same pathway, and this arrangement facilitates the coordinate regulation of those genes, i.e.,

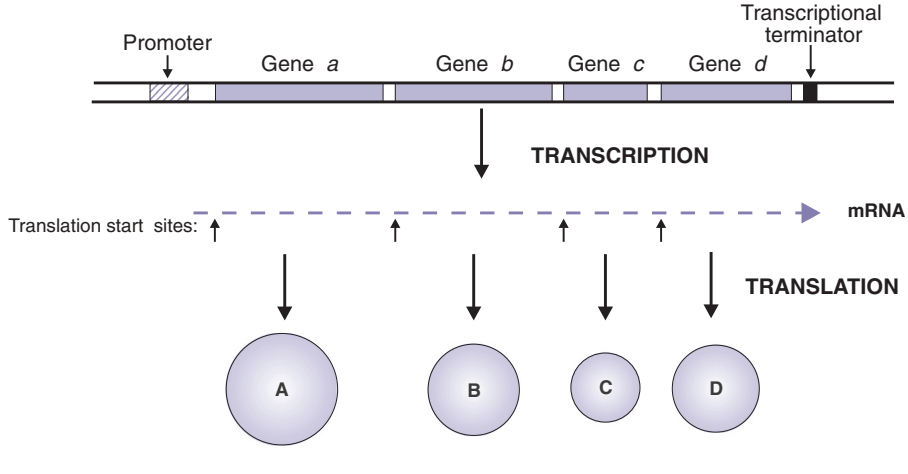


Figure 1.11 Structure of a bacterial operon.

expression of each gene in the operon goes up (or down) together in response to changing conditions.

In eukaryotes, by contrast, the way in which ribosomes initiate translation is different, which means that they cannot usually produce separate proteins from a single mRNA in this way. Although there are some examples where polygenic transcripts analogous to bacterial operons are produced, these are very much the exception rather than the rule. Generally, when a single mRNA gives rise to different proteins, this is due to alternative processing of the mRNA (see below) or by producing one long polyprotein or precursor, which is then cleaved into different proteins (as occurs in some viruses). However, in eukaryotes, there is one type of RNA that is usually produced in a polycistronic fashion, namely microRNA (miRNA). We will consider this in more detail in Chapter 11.

1.5.2 Exons and introns

In bacteria there is generally a simple one-for-one relationship between the coding sequence of the DNA, the transcribed mRNA and the translated protein. This is usually not true for eukaryotic cells, where the initial transcript is many times longer than that needed for translation into the final protein. This pre-mRNA contains blocks of sequence (introns) that are removed by processing to generate the final mature mRNA for translation (Figure 1.12).

Introns do occur in bacteria, but quite infrequently. This is partly due to the need for economy in a bacterial cell. The smaller genome and generally more rapid growth create an evolutionary pressure to remove unnecessary material

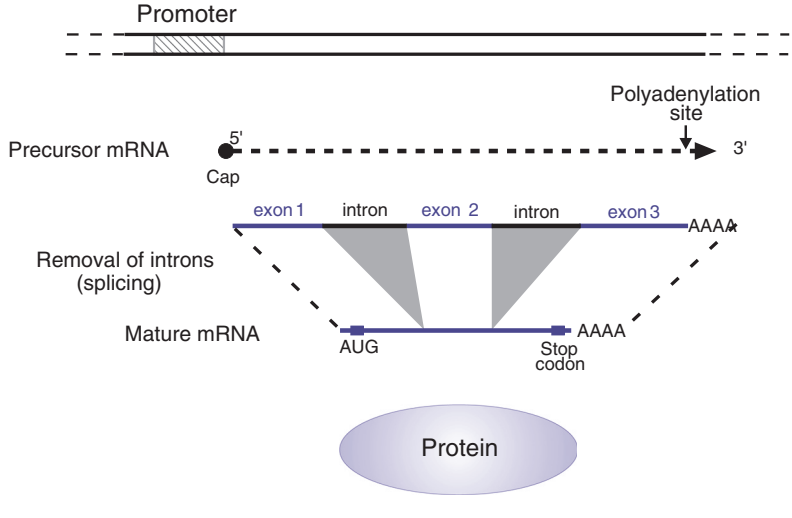


Figure 1.12 Exons and introns.

from the genome. A further factor arises from the nature of transcription and translation in a bacterial cell. Since the ribosomes translate the mRNA while it is being made, there is less opportunity for sections of the RNA to be removed before translation.

1.6 Refinements of the model

The simple model described earlier in this chapter is that the inherited properties of a cell are determined by the sequence of bases in the DNA. These properties are manifested by transcription of the DNA into RNA, which (in eukaryotes) is processed by removal of introns to produce a messenger RNA that is then translated into protein. The levels of the protein are regulated by control of the frequency and rate of transcription and/or translation, as well as by the stability of the RNA and protein molecules.

We now know that several aspects of this model are inadequate. Firstly, there can be differences between cells that have identical DNA sequences, and these differences can be passed on from one cell to its progeny. This phenomenon is known as epigenetic inheritance. The clearest example of how this can happen is due to methylation of specific bases in the DNA. The extent and position of this methylation can be passed on from one cell to its progeny, hence it is inherited, but you will not see that difference by conventional DNA sequencing. Later on (in Chapter 10) we will consider ways in which this can be investigated, together with other differences such as variation in the number of copies of regions of the DNA (copy number variation).

Secondly, we now know that much of the DNA that does not code for proteins is not actually junk at all. It plays an important role in the regulation of the activity of the cell through the action of small RNA molecules. This includes the process of *RNA interference*, which is the action of miRNAs to alter mRNA stability and translational efficiency, with a net effect of reducing protein production. RNA interference probably developed to protect the cell from infection by RNA viruses, which replicate via a double-stranded RNA intermediate. It may also be exploited as a way of silencing gene expression in the laboratory, one that demands much less resources and is much more amenable to high-throughput screening than other methods (see Chapters 8 and 11).

Thirdly, sequencing of eukaryotic genomes (see Chapter 8), especially those of humans and other mammals, showed that the genome appeared to code for far fewer proteins than expected. Human DNA is thought to contain only about 21 000 distinct protein-coding genes, which can be compared to a bacterium such as *Mycobacterium tuberculosis*, with over 4000 such genes. However, the number of potential protein species in a human cell is much larger than 21 000. The main reason for this is that the splicing mechanism, by which introns are removed to produce the mRNA, is capable

of variation, so that in some cells two exons are adjacent in the final product while in others they are separated by an intron that has not been removed. This alternative splicing, which can result in quite different proteins arising from a single protein-coding region, is a significant cause of the physiological variation between cells in different tissues, which have identical DNA structures.

2
How to Clone a Gene

2.1 What is cloning?

Cloning is defined as the use of asexual reproduction to obtain organisms that are genetically identical to one another, and to the 'parent'. This contrasts with sexual reproduction, where the offspring are not usually genetically identical. It is worth stressing that clones are only identical *genetically*; the actual appearance and behaviour of the clones will be influenced by other factors such as their environment. This applies to all organisms, from bacteria to humans.

Despite the emotive language surrounding the word 'cloning', this is a surprisingly familiar concept. In particular, anyone with an interest in gardening will know that it is possible to propagate plants by taking cuttings. These are clones. Similarly, the routine bacteriological procedure of purifying a bacterial strain by picking a single colony for inoculating a series of fresh cultures is also a form of cloning.

The term *cloning* is applied to genes, by extension of the concept. If you introduce a foreign gene into a bacterium, or into any other type of cell, in such a way that it will be copied when the cell replicates, then you will produce a large number of cells all with identical copies of that piece of DNA – you have *cloned* the gene (Figure 2.1). By producing numerous copies in this way, you can sequence it or label it as a probe to study its expression in the organism it came from. You can express its protein product in bacterial or eukaryotic cells. You can mutate it and study what difference that mutation makes to the properties of the gene, its protein product, or the cell that carries it. You can even purify the gene from the bacterial clone and inject it into a mouse egg, and produce a line of transgenic mice that express it. Behind all these applications lies a cloning process with the same basic steps.

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

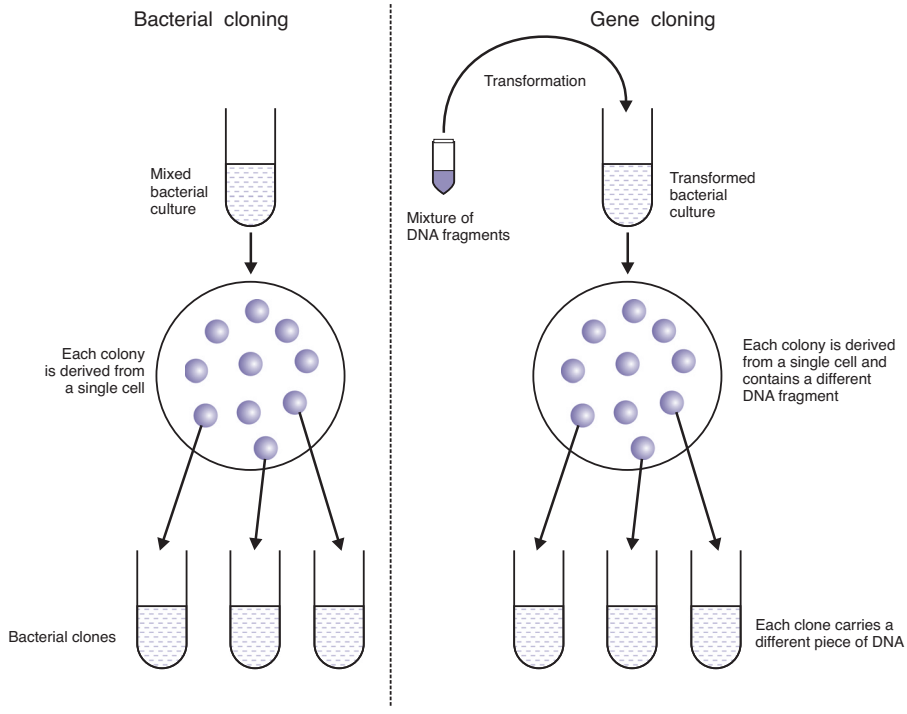


Figure 2.1 (a) Bacterial cloning and (b) gene cloning.

We will start with an overview of the process, before considering the various steps in more detail. In this chapter, we will focus on using bacterial cells (mainly *E. coli*) as the host. Later in the book (Chapter 7) we will extend the discussion to alternative host cells, as well as looking at vectors that are designed for optimising expression of cloned genes.

2.2 Overview of the procedures

Some bacterial species naturally take up DNA by a process known as *transformation*. However, most bacteria have to be subjected to chemical or physical treatments before DNA will enter the cells. In all cases, the DNA will not be replicated by the host cell unless it either recombines with (i.e., is inserted into) the host chromosome, or alternatively is incorporated into a molecule that is recognized by the host cell as a substrate for replication. For most purposes the latter strategy is the relevant one. We use *vectors* to carry the DNA and allow it to be replicated. There are many types of vectors for use with bacteria, but, to start with, we will consider *plasmids*, which are naturally occurring pieces of DNA that are replicated independently of the chromosome, and are inherited by the two daughter cells when the cell divides. In later

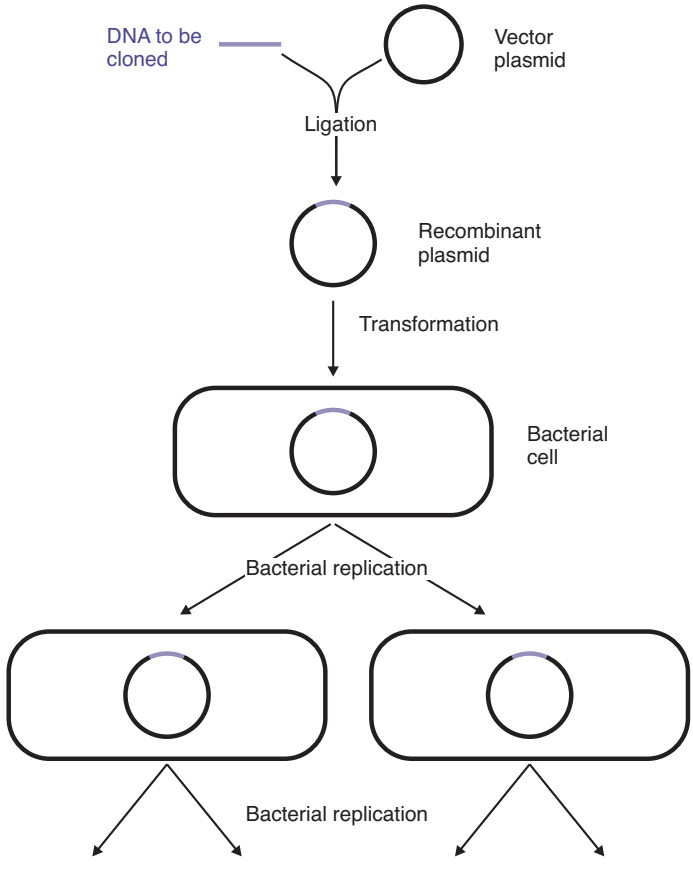


Figure 2.2 Basic outline of gene cloning.

sections we will encounter other types of vectors, including viruses that infect bacteria; these are known as *bacteriophages,* or phages for short.

The DNA that we want to clone is inserted into a suitable vector, producing a *recombinant molecule* consisting of vector plus insert (Figure 2.2). This recombinant molecule will be replicated by the bacterial cell, so that all the cells descended from that initial transformant will contain a copy of this piece of recombinant DNA. A bacterium like *E. coli* can replicate very rapidly under laboratory conditions, doubling its population size every 20 minutes or so. This *exponential growth* gives rise to very large numbers of cells. After 30 generations (10 hours), there will, in theory, be 1×10^9 (one billion) descendants of the initial transformant. Each of these cells carries a copy of the recombinant DNA molecule, so we will have produced a very large number of copies of the cloned DNA.

Of course, exponential growth does not continue indefinitely; after a while, the bacteria will start to run out of the nutrient required for growth, and stop multiplying. With *E. coli*, this commonly occurs with about 10^9 bacteria per millilitre of culture. But if we take a small sample and add it to fresh medium, exponential growth will resume. The clone can thus be propagated, and so we can effectively produce unlimited quantities of the cloned DNA. Similarly, if we can get the bacteria to express the cloned gene, we can also get very large amounts of the product of that gene relatively easily (see Chapter 7 for further details).

In order to insert a piece of DNA into a vector, we require a method for joining pieces of DNA together, as well as a way of cutting the vector to provide an opportunity for joining between vector and insert DNA to take place. The key to the development of gene cloning technology was the discovery of enzymes that would carry out these reactions in a very precise way. The main enzymes needed are *restriction endonucleases*, which break the sugar–phosphate backbone of DNA molecules at precise sites, and *DNA ligases*, which are able to join together the fragments of DNA that are generated in this way (Figure 2.3).

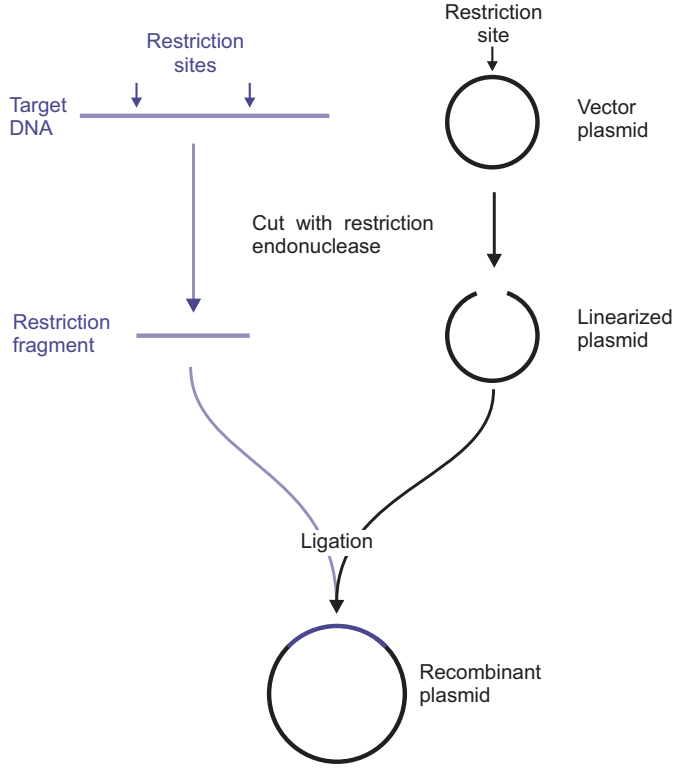


Figure 2.3 Cutting and joining DNA.

Once a piece of DNA has been inserted into a plasmid vector (forming a *recombinant* plasmid) it then has to be introduced into the bacterial host by transformation. Generally, this process is not very efficient so only a small proportion of bacterial cells actually take up the plasmid. However, by using a plasmid vector that carries a gene coding for resistance to a specific antibiotic, we can simply plate out the transformed bacterial culture onto agar plates containing that antibiotic, and only the cells that have received the plasmid will be able to grow and form colonies.

This description does not consider how we get hold of a piece of DNA carrying the specific gene that we want to clone. This may involve the construction of a *gene library*, a large collection of clones containing random fragments of DNA, which can then be screened to identify the required clone (see Chapter 3), or a specific fragment of DNA may be amplified using the *polymerase chain reaction* (*PCR*), as discussed in Chapter 4.

2.3 Extraction and purification of nucleic acids

The first step for most of the procedures referred to in this book is to extract the DNA (or for some purposes RNA) from the cell and to purify it by separating it from other cellular components. In this chapter, we review the concepts underlying the most commonly used methods of purifying and fractionating nucleic acids; for further experimental details, you will need to consult a laboratory manual (see Bibliography).

2.3.1 Breaking up cells and tissues

If we start with a culture of bacterial or eukaryotic cells, the first step is to separate the cells from the growth medium (e.g., by centrifugation). Wherever possible, the material should be freshly harvested or frozen until ready to use, to avoid degradation by enzymes present in the cell extract.

The cells then need to be lysed to release their components. A typical bacterial cell is enclosed by a cytoplasmic membrane and surrounded by a rigid cell wall. *E. coli* (and similar bacteria) also have an outer membrane, which acts to further restrict access to the cell wall. Thus, to lyse a bacterial cell requires a combination of relatively harsh chemicals; for *E. coli*, cell lysis can be achieved by a combination of EDTA (ethylene diamine tetra-acetate), lysozyme and a detergent such as SDS (sodium dodecyl sulphate, also known as sodium lauryl sulphate). Lysozyme digests the polymers that form the rigid cell wall, while EDTA eliminates divalent cations and thus destabilizes the outer membrane – allowing the lysozyme to access the cell wall structure. Use of EDTA also has the additional benefit that it inhibits DNases that would otherwise tend to degrade the DNA we want to isolate. Finally, the

role of the detergent is to solubilize the membrane lipids, releasing the cell contents.

Plant and fungal cells have cell walls that are different from those in bacteria, and require alternative treatments, either mechanical or enzymatic, while animal cells (which lack a cell wall) can usually be lysed by more gentle treatment, using only a mild detergent solution. If you are starting with a more complex tissue sample, this first needs to be homogenized to disperse the tissue into small groups of cells, so that the individual cells can be lysed by the techniques described above.

The resultant crude extract from this lysis contains a complex mixture of DNA, RNA, proteins, lipids and carbohydrates. It should be noted that the sudden lysis of the cell will usually result in fragmentation of chromosomal DNA. Where it is necessary to obtain very large (even intact) chromosomal DNA, more gentle lysis conditions are necessary (see the description of pulsed-field gel electrophoresis in Chapter 9). Bacterial plasmids, however, are readily obtained in their native, circular state by standard lysis conditions.

The next step is to separate the desired nucleic acid from these other components. Removal of RNA from a DNA preparation is easily achieved by treatment with ribonuclease (RNase). Since RNase is a very heat-stable enzyme, it is easy to ensure that it is free of traces of deoxyribonuclease (DNase) that would otherwise degrade your DNA, simply by heating the enzyme before use. Nowadays, it is routine to buy certified DNase-free RNase for this purpose, as well as RNase-free DNase, which is needed for removal of DNA from RNA preparations.

Removal of proteins is particularly important as the cell contains a number of enzymes that will degrade nucleic acids, as well as other proteins that will interfere with subsequent procedures by binding to the nucleic acids. The most effective way of removing proteins is by extraction with a mixture of liquefied phenol and chloroform. When the mixture is vigorously agitated, the proteins will be denatured and precipitated at the interphase, and the nucleic acids can be recovered from the aqueous layer. If you carry out the extraction with untreated (and therefore acidic) phenol, DNA will partition into the organic phase, allowing you to recover pure RNA from the aqueous phase. However, if (as is more often the case) you wish to recover DNA from the extraction, it is essential that the phenol is first equilibrated with a neutral, or even alkaline, buffer so that the DNA will partition into the aqueous phase. However, phenol is very hazardous to use. It is safer, and more convenient, to use affinity chromatography or digestion with a proteolytic enzyme such as proteinase K. There are commercially available kits for this purpose.

Following this protein extraction phase, you will have a protein-free sample of your nucleic acid(s). However, it will probably be more dilute than you want it to be. The answer is normally to concentrate (and further purify) the solution by precipitating the nucleic acid. This is done by adding either

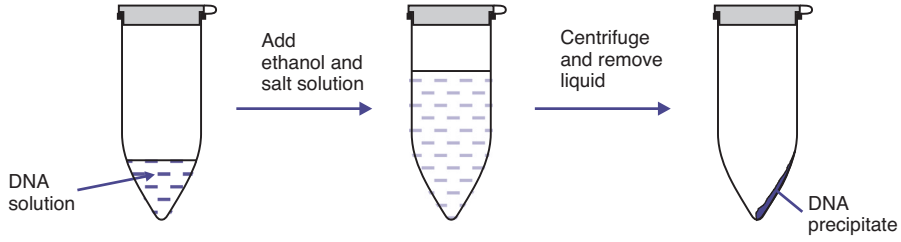


Figure 2.4 Ethanol precipitation. The tubes are spun in an angle rotor, so the precipitate is found at the side of the lower part of the tube.

isopropanol or ethanol to your sample, plus monovalent cations (Na^+, K^+ or NH_4^+), which together will drive the formation of a nucleic acid precipitate; this can be collected at the bottom of the test tube by centrifugation (Figure 2.4). Some of the salt will precipitate as well, but can be easily removed by a subsequent washing of the pellet with 70% ethanol.

2.3.2 Alkaline denaturation

The procedure above provides us with a pure preparation of total DNA from the bacterial cell. The next step is to separate the plasmid from the chromosomal DNA, which can be done conveniently by alkaline denaturation. As stated above, chromosomal DNA is broken into linear fragments during cell lysis. Raising the pH disrupts the hydrogen bonds and allows the linear strands to separate. Plasmids are much less prone to breakage and are not disrupted by cell lysis; they remain as intact supercoiled circular DNA. Although the high pH will disrupt the hydrogen bonds, the two circular strands will not be able to separate physically, and will remain interlinked. When the pH is reduced, the interlinked plasmid strands will snap back to reform the double-stranded plasmid (Figure 2.5). In contrast, the separated linear chromosomal fragments will aggregate into an insoluble network that can be removed by centrifugation, leaving the plasmids in solution. Other cell components, including cell wall debris and many proteins, are also removed by this procedure, further reducing the need for phenol extraction.

2.3.3 Column purification

Two types of column purification are frequently used when purifying nucleic acids. In *size-selection chromatography*, a sample is passed through a matrix of small porous beads. Smaller molecules, such as salts and unincorporated nucleotides, will enter the beads, whereas larger ones such as longer nucleic acid chains will pass right through the column. This type of purification is a fast and simple alternative to purification by alcohol precipitation.

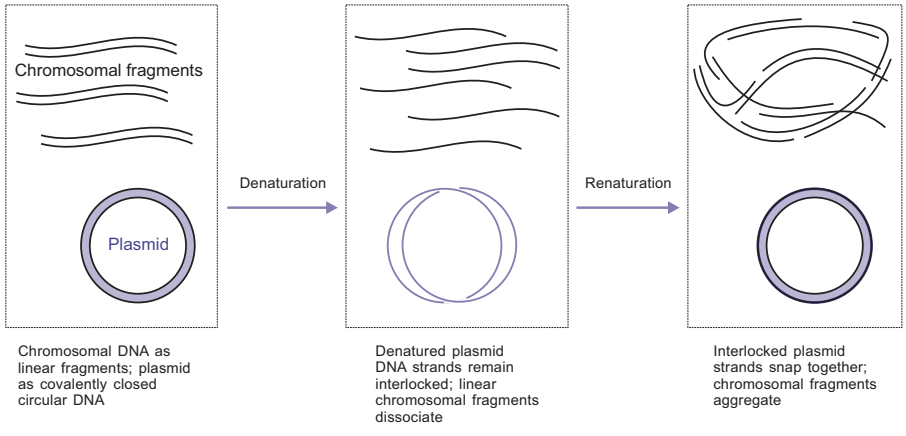


Figure 2.5 Alkaline denaturation procedure for plasmid purification.

In *affinity chromatography* purification, the macromolecules in your sample bind to the resin in the column. This is most usually an anionic resin, which binds to the negatively charged phosphate groups in the nucleic acid backbone; but it may be more sophisticated, such as resins coated with oligo-dT sequences, which specifically bind to the poly(A) tails of eukaryotic mRNA molecules (see Chapter 3). In both cases, undesirable molecules can be washed from the column, after which the stringency conditions are changed and the bound nucleic acids eluted.

2.4 Detection and quantitation of nucleic acids

If your DNA preparation is reasonably pure, then you can estimate the DNA concentration by measuring the absorbance of the solution in an ultraviolet (UV) spectrophotometer at 260 nm. This is convenient, but the dilution required to measure samples in a standard spectrophotometer reduces the sensitivity dramatically. However, equipment is now available for determining the UV absorbance of very small samples (microlitre, or even nanolitre volumes) of nucleic acids. Note that the presence of proteins or phenol will affect this estimate, and the absorbance at 280 nm is often used to assess such contamination, with a 260 : 280 absorbance ratio of between 1.75 and 2 being deemed reasonably pure. Finally, UV absorbance gives you no check on the integrity of your DNA; it can be completely degraded and still give you a reading.

Dyes such as ethidium bromide are commonly used for both detecting and quantitating nucleic acids. Ethidium bromide has a flat ring structure that is able to stack in between the bases in nucleic acids; this is known as *intercalation*. The dye can be detected by its fluorescence when exposed to UV. This

is the most widely used method for staining electrophoresis gels, and can also be used for estimating the amount of DNA (or RNA) in each band on the gel, by comparing the intensity of the fluorescence with a sample of known concentration on the same gel.

Note that ethidium bromide is mutagenic, and precaution must be taken to eliminate health risks. Alternative, less hazardous, dyes are increasingly being used for this purpose.

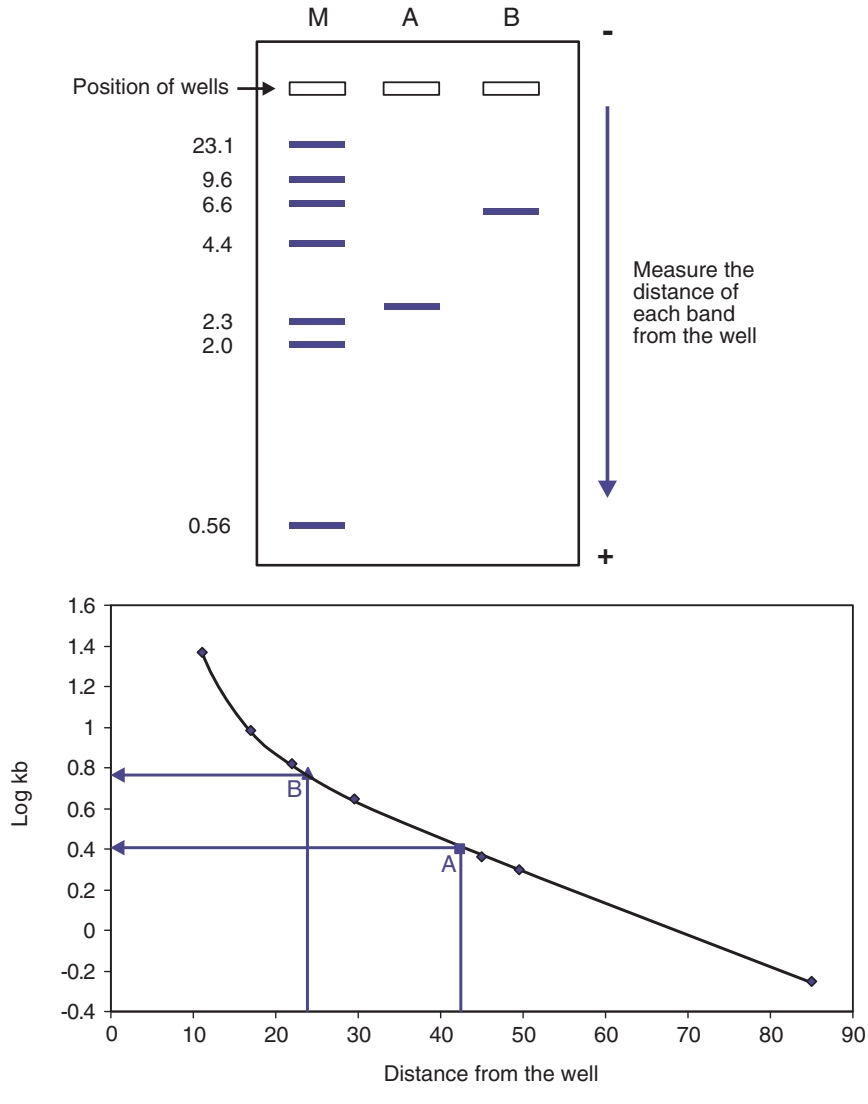
2.5 Gel electrophoresis

Gel electrophoresis is a crucial technique for both the analysis and the purification of nucleic acids. When a charged molecule is placed in an electric field, it will migrate towards the electrode with the opposite charge; DNA and RNA, being negatively charged, will move towards the positive pole (anode). In a gel, which consists of a complex network of pores, the rate at which a nucleic acid molecule moves will be determined by its ability to penetrate through this network. For linear fragments of double-stranded DNA within a certain size range, this will reflect the size of the molecule (i.e., the length of the DNA). We do not have to consider the amount of charge that the molecule carries (unlike some other applications of electrophoresis), since all nucleic acids carry the same amount of charge per unit size.

The effective size range of nucleotides that a gel can separate is determined by the gel's composition. We can use agarose gels for separating nucleic acid molecules greater than a few hundred base pairs, reducing the agarose concentration to obtain effective separation of larger fragments, or increasing it for small fragments. For even smaller molecules, down to only a few tens of base pairs, we would use polyacrylamide gels. These gels are capable of distinguishing DNA chains with only a single base difference in their length, which is important for sequence analysis (see Chapter 5).

2.5.1 Analytical gel electrophoresis

We can use agarose gel electrophoresis for analysing the composition and quality of a nucleic acid sample. In particular, it is invaluable for determining the size of DNA fragments from a restriction digest or the products of a PCR reaction (see Chapter 4). For this purpose it is necessary to calibrate the gel by running a standard marker containing fragments of known sizes; in Figure 2.6, the marker is a *Hin*dIII restriction digest of DNA from the lambda bacteriophage. It can be seen that over much of the size range there is a linear relationship between the logarithm of the fragment size and the distance it has moved. From this calibration graph, you can then estimate the size of your unknown fragment(s) – assuming they are linear double-stranded DNA (see below).



Using the standard marker to provide a calibration curve, the size of fragment A is estimated as 2.5 kb and fragment B as 6.0 kb.

Figure 2.6 Analytical gel electrophoresis.

The reason for the non-linearity of the curve with larger DNA molecules is that these molecules move through the gel in a different way. Although they are large, they are also very thin, and they can in effect slither through the gel end-on. It takes some time for them to become lined up, but once they are, then the rate at which they move is largely independent of their size. So, for a particular gel, all molecules above a certain size will have virtually the

same mobility. In the gel shown, all DNA molecules larger than about 20 kb will not be separated, but the use of gels with a lower agarose concentration will extend the size range. Special techniques, involving frequent switching of the direction of the electric field, are available for separating very large DNA molecules – see the description of pulsed-field gel electrophoresis (PFGE) in Chapter 9.

If a more precise confirmation of the nature of the sample is required, the gel is *blotted* onto a membrane support, and *hybridized* with a nucleic acid probe (see section on Southern blotting in Chapter 3).

Not all DNA molecules are linear. Native plasmids are supercoiled circular molecules, but if a plasmid is nicked (i.e., one of the strands is broken) the loose ends are free to rotate, and it adopts a relaxed, open circular form. Furthermore, a double-strand break will produce a linearized plasmid (see Figure 2.7). It is therefore quite normal for a purified plasmid preparation to show two or three bands in a gel; it does not necessarily mean that there is more than one plasmid. (A further complication is that there may be dimeric or multimeric forms of the plasmid present.) Relaxed and supercoiled plasmids show different relative mobilities, and both are different from linear molecules. Unless plasmids are linearized before electrophoresis by digestion with restriction enzymes, a special supercoiled molecular weight marker has to be used.

Most RNA molecules you will encounter are single-stranded, and in some cases you may come across single-stranded DNA as well. These need special consideration. As described in Chapter 1, single-stranded nucleic acids tend to fold up into complex secondary structures, to remove the hydrophobic bases from the aqueous environment. The migration of the molecule will be greatly influenced by the way it folds. If we want to get a true picture of its size, we have to make sure that it remains in an unfolded state. To do this, we use *denaturing gels*, in which denaturing agents such as urea or formaldehyde are included, preventing the RNA or ssDNA molecule from forming a secondary structure.

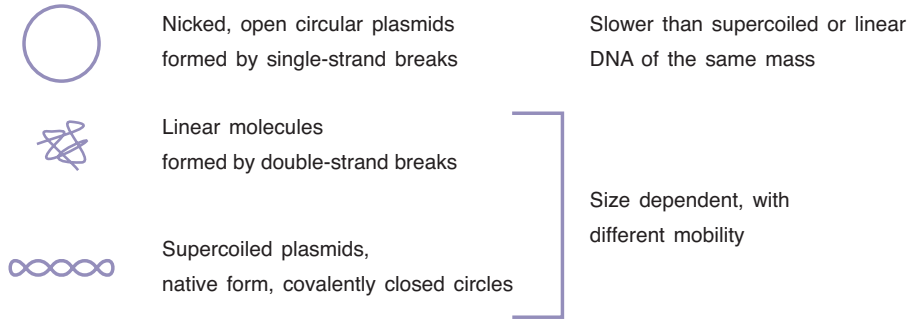


Figure 2.7 Electrophoretic mobility of forms of plasmid DNA.

2.5.2 Preparative gel electrophoresis

Gel electrophoresis is also an important tool in the purification of a specific nucleic acid fragment from a complex mixture. After separating the sample, it is visualized with ethidium bromide. With the gel still on the transilluminator, the band(s) that need to be purified are excised using a razor blade or scalpel. The DNA can then be recovered and purified from the gel fragment, using standard DNA purification procedures. A potential issue with such an approach is that UV irradiation is damaging to DNA, causing the formation of cross-links that may reduce the efficiency of later cloning procedures. To reduce the potential for cross-links to form, several strategies can be used, including the use of longer wavelength UV transilluminators, which are less damaging to DNA, and minimizing the time that your gel is subject to UV illumination. However, an increasingly used strategy is to remove the need for UV illumination at all. Specific dyes such as crystal violet allow the visualization of DNA within an agarose gel under normal illumination, thus reducing the need for potentially damaging UV to be used. A drawback of such dyes is that they are much less sensitive than ethidium bromide, meaning that more DNA must be loaded onto a gel for it to be visualized. However, in the context of gene cloning, the amounts of DNA required are usually within the range that can be easily visualized by these dyes.

2.6 Restriction endonucleases

Restriction endonucleases derive their name from the phenomenon of host-controlled restriction and modification in bacteria. This can occur if a bacteriophage preparation that has been grown using one bacterial host strain is used to infect a different strain. It may be found that infection is extremely inefficient (i.e. phage growth is *restricted* by the new host) compared to infection using the same bacterial strain for both propagation and assay. The reason for this restriction of phage growth is that many bacterial strains produce an endonuclease, which is therefore known as a *restriction endonuclease*, that cuts DNA into pieces, so that the incoming phage DNA is rapidly broken down and only occasionally escapes to produce phage progeny (Figure 2.8). The host DNA is protected against the action of the endonuclease by a second enzyme, which modifies DNA by methylation, so that it is not attacked by the endonuclease. These restriction endonuclease enzymes, often referred to as restriction enzymes for short, are able to cut DNA specifically, providing us with linearized plasmids and insert DNA fragments, which can be joined together. Generally, type II restriction endonucleases are used in gene cloning, as these recognize and cut within (or immediately adjacent to) specific target DNA sequences, generating specific fragments.

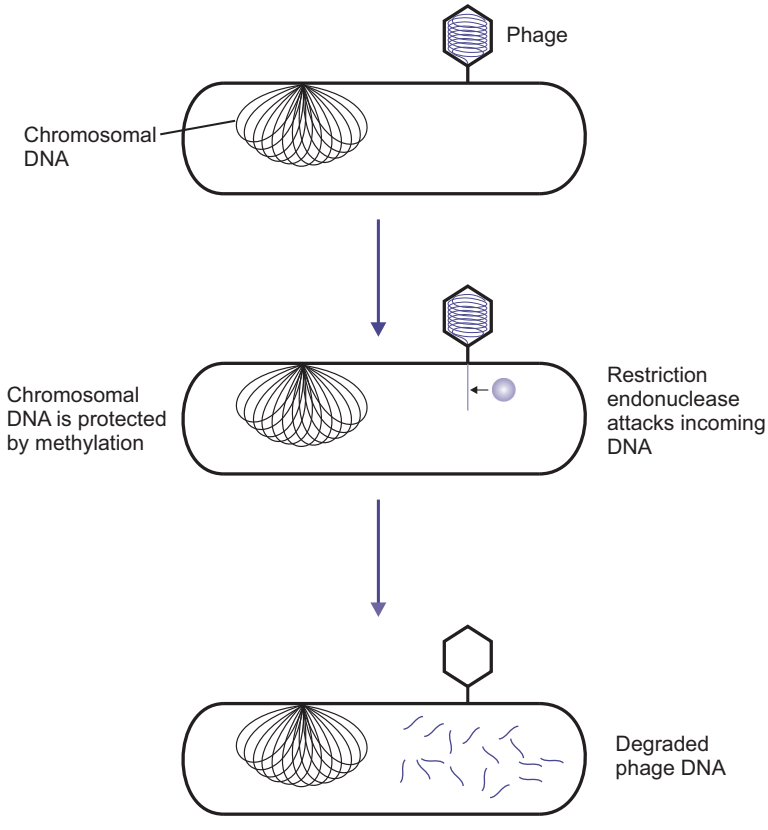


Figure 2.8 Bacteriophage restriction.

2.6.1 Specificity

A very large number of type II restriction endonucleases have been characterized. These are identified by the name of the organism from which they are obtained, using the first letter of the genus and the first two letters of the species name, together with a suffix indicating the specific enzyme from that species. Thus *Pst*I indicates a specific enzyme obtained from the bacterium *Providencia stuartii,* and *Hae*I, *Hae*II and *Hae*III indicate three different enzymes, with different specificities, from *Haemophilus aegyptius*. The convention is to write the first part (derived from the Latin name of the source organism) in italics, just as you would with the species name.

Some examples of restriction enzymes are shown in Box 2.1, together with their recognition sequences and the position of the site at which they break the DNA. One of the key parameters that influences how we use these enzymes is the length of the recognition site, which affects the frequency with which they cut DNA, and hence the average size of the fragments generated.

Box 2.1 Examples of restriction endonucleases

Enzyme	Recognition site	Number of bases	Ends generated	Original source of enzyme
*Eco*RI	G/AATTC	6	5′ sticky	*Escherichia coli* RY13
*Bam*HI	G/GATCC	6	5′ sticky	*Bacillus amyloliquefaciens* H
*Bgl*II	A/GATCT	6	5′ sticky	*Bacillus globigii*
*Pst*I	CTGCA/G	6	3′ sticky	*Providencia stuartii*
*Xma*I*	C/CCGGG	6	5′ sticky	*Xanthomonas malvacearum*
*Sma*I*	CCC/GGG	6	Blunt	*Serratia marcescens*
*Acc*65I*	G/GTACC	6	5′ sticky	*Acinetobacter calcoaceticus* 65
*Kpn*I*	GGTAC/C	6	3′ sticky	*Klebsiella pneumoniae*
*Sau*3A	/GATC	4	5′ sticky	*Staphylococccus aureus* 3A
*Alu*I	AG/CT	4	Blunt	*Arthrobacter luteus*
*Not*I	GC/GGCCGC	8	5′ sticky	*Nocardia otitidis-caviarum*
*Pac*I	TTAAT/TAA	8	3′ sticky	*Pseudomonas alcaligenes*

Only one strand of the recognition site is shown, with a slash (/) showing the position of the cleavage site. All the examples shown are palindromic, so the sequence of the second strand, read as the reverse complement, and the position of the cleavage site, will be the same as that shown. Thus the reverse complement of 5′-GAATTC-3′ is also 5′-GAATTC-3′, and both strands are cut by *Eco*RI between G and A.
**Xma*I is an isoschizomer of *Sma*I, and *Acc*65I is an isoschizomer of *Kpn*I.

For example, take the enzyme *Sau*3A, which has a four-base recognition site, GATC. If there is an equal proportion of all four bases, and if they are randomly distributed, then at any position on the chromosome there is a 1 in 4 chance that it is a G. Then there is a 1 in 4 chance that the next base is an A; the chance of having the sequence GA is the product of the two, or 1 in 4^2 (1 in 16). Extending the argument, the chance of the sequence GAT

occurring is 1 in 4^3, and that of the four-base sequence GATC is 1 in 4^4, or 1 in 256. So this enzyme (and any others with a four-base recognition site) would cut such DNA on average every 4^4 bases, and so would generate a set of fragments with an *average* size of 256 bases. The actual sizes of the fragments will be distributed quite widely on either side of this average value, as in reality the four bases are not present in exactly equal proportions, nor are they randomly distributed.

If a four-base site would occur every 4^4 bases, then a six-base site such as that recognized by *Eco*RI (GAATTC) would occur (on average) every 4^6 bases (or 4096 bases). To give some perspective to this calculation, a moderately sized protein might have about 300 amino acids and would therefore be coded for by a DNA sequence of 900 bases in length (ignoring the possible presence of introns). So a four-base cutting enzyme would often give fragments much smaller than a whole gene. On the other hand a bacterial genome of perhaps 4×10^6 bases (4 Mb) would be expected (given the same assumptions) to be cut into about 1000 fragments by an enzyme such as *Eco*RI.

A further factor that affects the ability of restriction endonucleases to cut DNA is methylation of the DNA. We have already seen that each restriction endonuclease, in its original host, is accompanied by a modifying enzyme that protects the chromosomal DNA against endonuclease cleavage, through methylation of the recognition sequence. In addition, most laboratory strains of *E. coli* contain two site-specific DNA methyltransferases (methylases): The Dam methylase catalyses methylation of the adenine residues in the sequence GATC, while the Dcm methylase modifies the internal cytosine in the sequences CCAGG and CCTGG. Methylation at these positions may make the DNA resistant to attack by restriction enzymes that cut the DNA at sites containing these sequences (see Box 2.1). Methylases also occur in eukaryotes, especially the CpG methylase, which modifies the cytosine in some sites containing the dinucleotide CpG. This can affect the ability of restriction enzymes to cut certain sites in DNA from higher eukaryotes. A further complication is that many *E. coli* strains have other enzymes that attack such methylated DNA. This may be an important factor in the selection of host strain for optimising recovery of cloned DNA from mammals and higher plants.

Although most of the commonly used enzymes recognize either four-base or six-base sites, there are also important roles for enzymes that cut even less frequently, such as *Not*I and *Pac*I (see Box 2.1). These both have eight-base recognition sites, which might be expected to occur (on average) every 4^8 bases (about 65 kb), but the sequences are far from random (being composed entirely of G+C or A+T respectively) so a given genomic DNA sequence may contain very few sites (or even none at all) for such an enzyme.

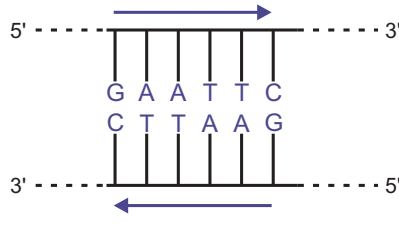
2.6.2 Sticky and blunt ends

Another important parameter is the position of the cut site within the recognition sequence. In this context it should be noted that most (but not all) restriction endonuclease recognition sites are said to be *palindromic*, although the term is not strictly accurate. A verbal palindrome, for example 'radar', reads the same from left to right as from right to left. A restriction site 'palindrome', such as GAATTC, is a bit different. At first glance it does not seem to be the same in the other direction, but you have to remember that the two strands of DNA lie in opposite directions. When we write GAATTC we are looking at the sequence of the 'top' strand, which runs 5′ to 3′ when read from left to right – so we should write the sequence as 5′GAATTC3′. On the 'bottom' (complementary) strand, we have to read from right to left to see the 5′ to 3′ sequence, which would also be 5′GAATTC3′ (Figure 2.9).

Within this sequence, the restriction enzyme *Eco*RI will cut the DNA between the G and the A on each strand, and will hence produce fragments with four bases unpaired at the 5′ end (Figure 2.10). These four bases (AATT) are the same on all fragments generated with this enzyme, and these ends are complementary to one another. They will thus tend to form base pairs, and so help to stick the fragments together. They are therefore referred to as *cohesive*, or *sticky*, ends. Note that the base pairing formed with a sequence of four bases is very weak and would not be stable, so we need to use DNA ligase to finally join the fragments together with a covalent bond to form a recombinant molecule (see below). Nevertheless, the limited cohesiveness of these ends does make the ligation process much more efficient.

Some enzymes, such as *Pst*I, also form cohesive ends, but by cutting asymmetrically within the right-hand part of the recognition site; they thus generate sticky ends with unpaired single-strand sequences at the 3′ ends of the fragment. (Figure 2.11).

Although enzymes that generate cohesive ends are useful for gene cloning because of the increase in efficiency of the ligation process compared to blunt



The top strand, read from left to right (5' to 3'), is the same as the bottom strand when also read 5' to 3' (right to left)

Figure 2.9 Reading a palindromic sequence.

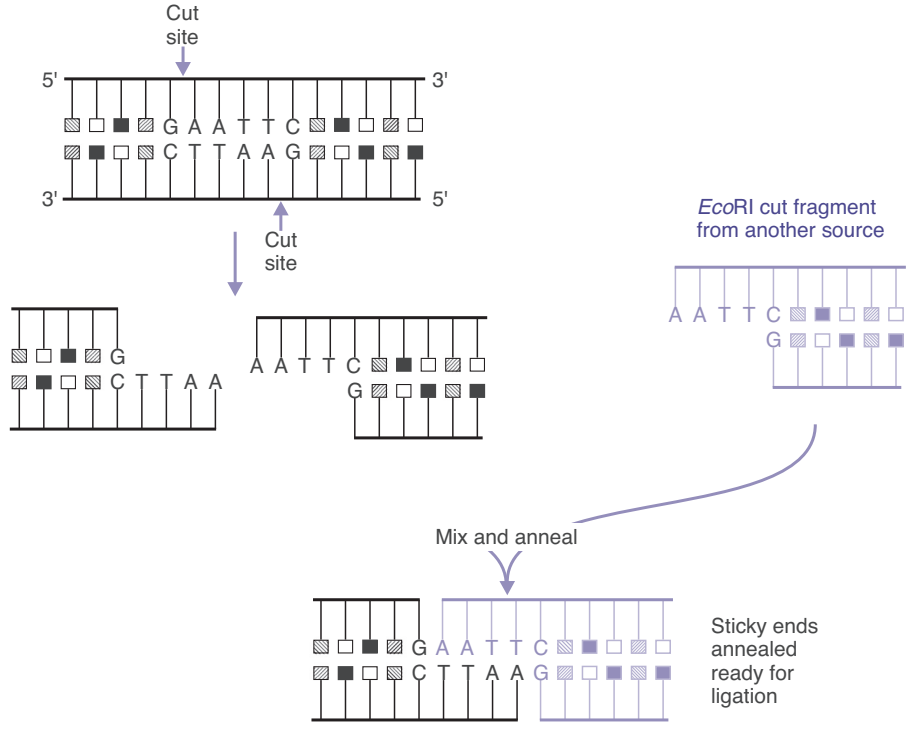


Figure 2.10 Sticky ends generated by *Eco*RI.

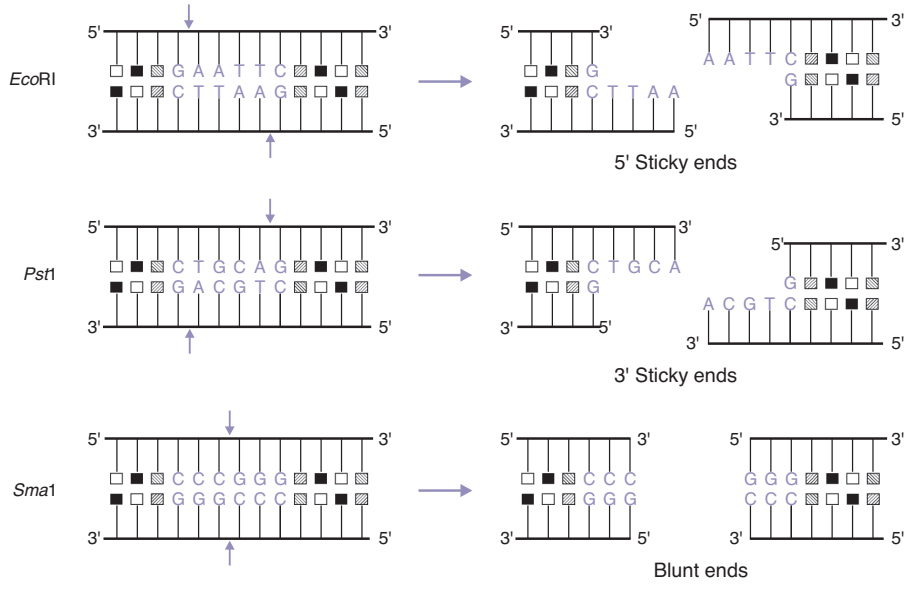


Figure 2.11 Restriction fragment ends.

ends of DNA, there is a limitation. You can only join together fragments with compatible ends. Hence, two *Eco*RI fragments can be ligated to each other, or two *Bam*HI fragments can be joined, but you cannot ligate an *Eco*RI fragment directly to a *Bam*HI fragment. However, there are circumstances in which compatible ends can be generated by different restriction enzymes. For example the restriction enzymes *Bam*HI and *Bgl*II recognize different sequences (see Box 2.1), but they both generate the same sticky ends (with unpaired GATC sequences) and these can be joined. It should be noted that when fragments made with a single restriction enzyme (e.g. *Eco*RI) are ligated, the *Eco*RI site is maintained in the finished DNA molecule. However, when fragments made by two different restriction enzymes that produce compatible ends are ligated (e.g. *Bam*HI and *Bgl*II) neither restriction site is retained, generally, and this can have important implications when designing a cloning strategy.

More flexibility can be achieved by using an enzyme such as *Sma*I, which recognizes the six-base sequence CCCGGG and cuts symmetrically at the centre position (Figure 2.11); it generates blunt ends, which although much less efficiently ligated have the advantage that they can be joined to any other blunt-ended fragment.

Some of the examples shown in Box 2.1 recognize the same restriction site. These are known as *isoschizomers* (Greek *iso*, equal; *skhizo*, to split). Some pairs of isoschizomers cut at different positions within the recognition site, which adds flexibility to our strategy. For example *Acc*65I and *Kpn*I both recognize the sequence GGTACC (Figure 2.12), but cut it at a different place, generating different sticky ends, while for another pair of isoschizomers (*Xma*I and *Sma*I), *Xma*I cuts asymmetrically and produces sticky ends that can be ligated to other *Xma*I fragments, while *Sma*I, as mentioned above, will generate blunt-ended fragments, allowing you to ligate the fragment to other blunt-ended DNA sequences.

2.7 Ligation

The next stage in cloning a gene is to join the DNA fragment to a vector molecule, such as a plasmid or bacteriophage, that can be replicated by the host cell after transformation. The joining, or ligation, of DNA fragments is carried out by an enzyme known as *DNA ligase*.

The natural role of DNA ligase is to repair single-strand breaks (nicks) in the sugar–phosphate backbone of a double-stranded DNA molecule, such as may occur through damage to DNA, as well as the joining of the short fragments produced as a consequence of replication of the ‘lagging strand’ during DNA replication. The action of the ligase requires that the nick should expose a 3′-OH group and a 5′-phosphate (Figure 2.13). Digestion with restriction endonucleases cuts the DNA in this way, i.e., it leaves the phosphate

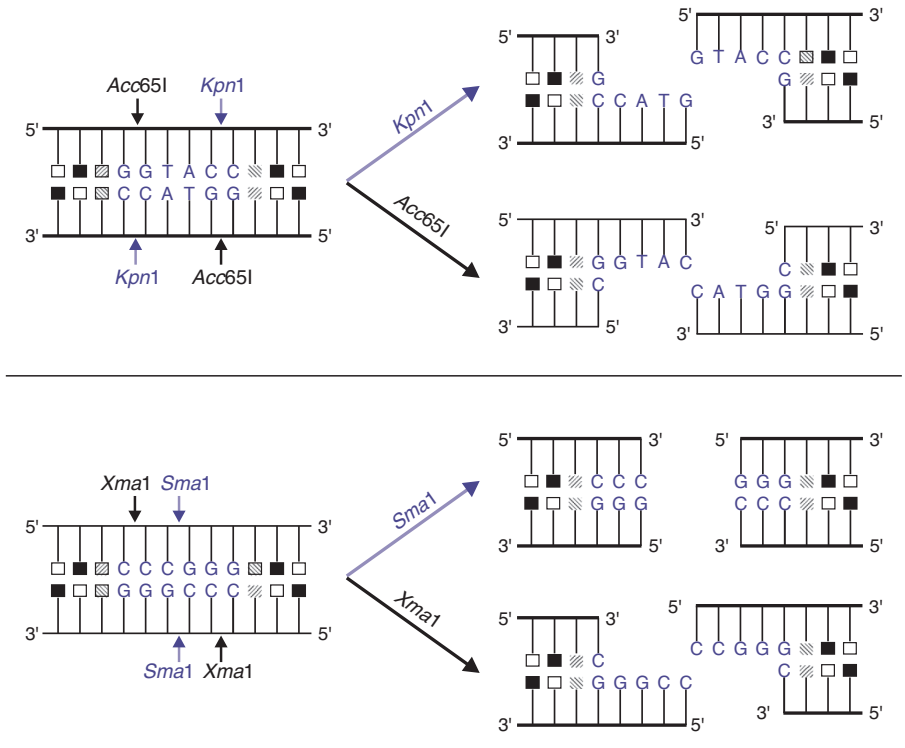


Figure 2.12 Isoschizomers.

on the 5′ position of the deoxyribose, and hence DNA ligase can be used to join together two DNA fragments in a cloning procedure. The unstable pairing of two restriction fragments with compatible sticky ends can therefore be considered as a double-stranded DNA molecule with a nick in each strand, and is therefore a substrate for DNA ligase action.

Some DNA ligases, such as the *E. coli* DNA ligase, are not capable of ligating blunt-ended fragments. They require pairing of overlapping ends. T4 DNA ligase (encoded by the bacteriophage T4), does not have this limitation. As well as joining fragments with sticky ends, it can also ligate blunt-ended fragments, albeit much less efficiently. To keep things simple, we will only consider the action of T4 DNA ligase, which is by far the most extensively used one. We will refer to this enzyme as T4 ligase (or just 'ligase') although there is also a (much less commonly used) T4 RNA ligase.

The T4 ligase requires ATP as a co-substrate. In the first step, the ligase reacts with ATP to form a covalent enzyme–AMP complex, which in turn reacts with the 5′-phosphate on one side of the nick, transferring the AMP to the phosphate group. The final stage is the attack by the 3′-OH group, forming a new covalent phosphodiester bond (thus restoring the integrity of the sugar–phosphate backbone) and releasing AMP. The absolute requirement

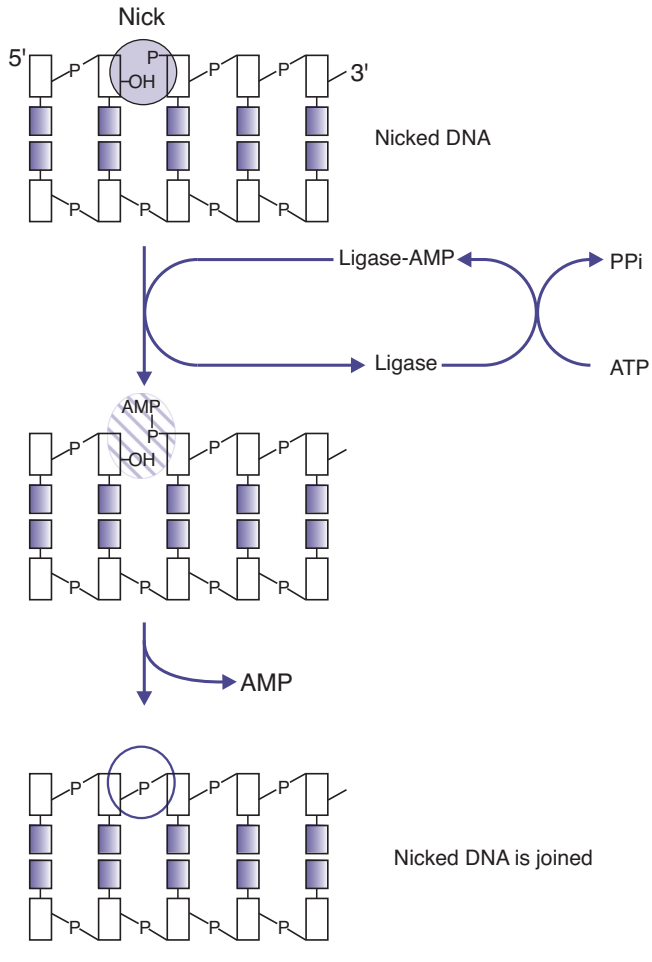


Figure 2.13 Action of T4 DNA ligase.

for the 5′-phosphate is extremely important; by removing the 5′-phosphate we can prevent unwanted ligation, which is an important technique in cloning (see below).

2.7.1 Optimising ligation conditions

Ligation can be one of the most unpredictable steps in the cloning process. Factors that may compromise the success of a ligation reaction include the presence of inhibitory material contaminating our DNA preparations, and degradation of the enzyme or the DNA (including loss of the 5′-phosphate). In addition, the reaction conditions may need to be adjusted to achieve the optimum ligation efficiency.

Since the reaction we want normally involves two different molecules of DNA (*inter*molecular ligation), we would expect it to be extremely sensitive to DNA concentration. It is therefore important to use high concentrations of DNA. However, we have two components: the vector and the insert (ignoring the fact that the insert may itself be a heterogeneous mixture of fragments), and it is therefore important to consider the concentrations required of each component. Indeed, there are a variety of possible reactions that can occur (Figure 2.14), and by adjusting the relative amounts of the two DNA components, as well as the total concentration of DNA, we can influence the likelihood of which of these different products will predominate.

At low concentrations of DNA, we are more likely to get the two ends of the same molecule joining (an *intra*molecular reaction), since the rate of a reaction involving one component will be linearly related to its concentration, whereas the rate with two different components will be proportional to the

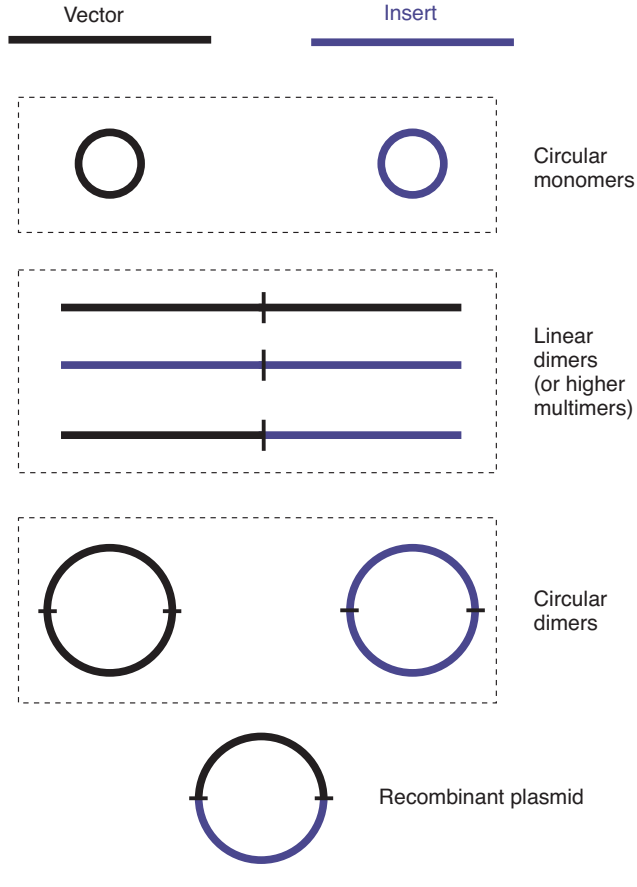


Figure 2.14 Ligation: some of the potential products.

product of the two concentrations. So, increasing the concentration will give a greater increase in the rate of the intermolecular reaction (between two molecules) than of the intramolecular reaction.

If we increase the concentration of the vector but not the insert, we will get an increase in the ligation of two vector molecules together (which we don't want). Conversely, increasing insert concentration will give increased levels of insert–insert dimers (which we also do not want, although they are less of a problem as they will not give rise to transformants).

So we need not only to keep the overall DNA concentration high, but also to create an optimum vector : insert ratio. It is not easy to predict reliably what that ratio should be for any given ligation, but typically it would range from 3 : 1 to 1 : 3. Note that these are molar ratios, referring to the total number of DNA molecules, and thus have to take account of the relative size of the vector and the insert. For example, if the vector is 5 kb and the insert is 500 bases, then a 1 : 1 molar ratio would involve ten times as much vector, by weight, as insert (e.g., 500 ng of vector and 50 ng of insert). In general, to convert the amount of DNA by weight into a value that can be used to calculate the molar ratio, divide the amount used (by weight) by the size of the DNA. Thus, if W_V and W_I are the weights used of vector and insert DNA respectively, and S_V and S_I are the sizes of vector and insert (e.g., in kilobases), then the vector : insert ratio is $(W_V/S_V) : (W_I/S_I)$.

A further complication arises if we are working with a heterogeneous collection of potential insert fragments, as would be the case if we were making a gene library (see Chapter 3). Loading more insert DNA into the ligation mixture will increase the possibility of obtaining multiple inserts. In other words, we may produce recombinant plasmids that carry two or more completely different pieces of DNA. This is not a good idea. It can lead to seriously misleading results when we come to characterize the clones in the library and try to relate them to the structure of the genome of our starting organism.

So adjusting the relative amount of the vector and insert will not only influence the success of ligation but will also affect the nature of the products. Fortunately, we do not have to rely entirely on adjusting the levels of DNA in order to obtain the result that we want. In the next section, we will look at ways of preventing the ligation events that we do not want, and later in this chapter we will consider designs of vector that allow us to distinguish between recombinant and non-recombinant transformants.

2.7.2 Preventing unwanted ligation: alkaline phosphatase and double digests

As described above, the ligation process depends absolutely on the presence of a 5′-phosphate at the nick site. If this phosphate group is removed, that

site cannot be ligated. Removal of 5′-phosphate groups is achieved with an enzyme known as alkaline phosphatase (because of its optimum pH) – one commonly used enzyme is calf intestinal phosphatase (CIP). Treatment of the vector molecule with CIP before ligation will remove the 5′-phosphates, and so make it impossible for self-ligation of the vector to occur. Ligation of vector to insert can occur, however, as the insert still has its 5′-phosphates. It is of course important to remove the CIP enzyme before undertaking the ligation reaction, and this is best done by phenol extraction and subsequent ethanol precipitation, as heat inactivation is not sufficiently reliable as a method to destroy CIP.

However, further inspection of the situation (Figure 2.15) will reveal that it is not quite that simple. At each junction between two DNA fragments there are *two* nicks that need repairing. At one of these, the 5′-phosphate is supplied by the insert, and that can be ligated normally. However, at the other nick, there should be a 5′-phosphate on the *vector*, and that has been removed. So the second nick cannot be mended. How can we deal with this situation?

Fortunately, this is not a real problem. The recombinant plasmid will have unrepaired nicks in its (circular) DNA (one at each end of the inserted fragment) – but it will hold together very stably by virtue of the base pairing all along the inserted DNA fragment. When this nicked molecule is introduced into a bacterial cell, enzymes within the bacterium will rapidly repair the nicks, adding the missing phosphates and ligating the broken ends. The plasmid can then be replicated within the bacterium without any difficulty.

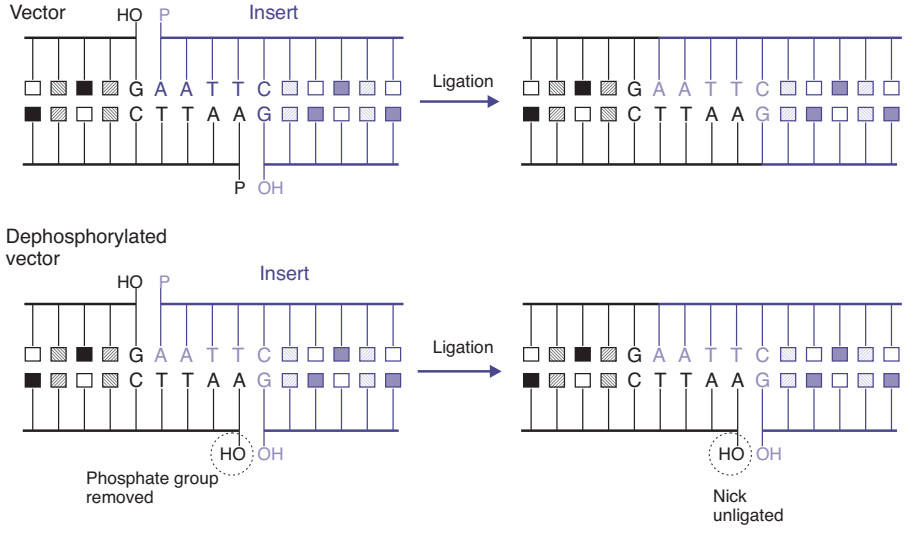


Figure 2.15 Ligation: effect of dephosphorylation of vector.

Some of the other potential problems stated in Section 2.7.1 will still remain, potentially. For example, we could still get insert dimers, or multiple inserts. But since we no longer have to worry about self-ligation of the vector, thanks to CIP, we can increase vector concentration relative to the insert and thus drive the reaction away from multiple inserts and towards the single insert that we want.

Even so, if we are making a gene library, when the possibility of multiple inserts is at its most serious as a problem, we may want to turn this strategy on its head. Phosphatase treatment of the *insert* rather than the vector will prevent multiple inserts, and we can prevent the occurrence of non-recombinant transformants (carrying self-ligated vector rather than vector–insert recombinants) by using special vectors that are unable to produce clones unless they carry an inserted DNA fragment (see later in this chapter).

An alternative strategy is to cut both components (vector and insert) with two different restriction enzymes. Most modern vectors have multiple cloning sites (see Section 2.9.2) so you can cut the vector with, for example, *Eco*RI and *Bam*HI. Provided that both enzymes have cut efficiently (and that you have removed the small fragment between the two sites), then vector religation will be impossible. If your insert fragment has been digested with the same two enzymes, then virtually all the colonies obtained will be recombinant, i.e., they will contain the insert fragment. This is also a useful strategy if you want to ensure that the insert fragment is in a specific orientation, as the insert and vector can only ligate in the orientation where the restriction sites match (i.e. two *Eco*RI sites will ligate, but *Eco*RI and *Bam*HI cohesive ends are incompatible and will not ligate).

2.7.3 Other ways of joining DNA fragments

One modification of the above system is the TA cloning of PCR products, which is described in Chapter 4. In addition, although DNA ligase is the most commonly used enzyme for joining two DNA molecules, it is not the only enzyme that can do this. DNA topoisomerase I, which is involved in controlling the degree of supercoiling of DNA (see Chapter 1), cuts one DNA strand, which is then free to rotate, and subsequently rejoins the cut ends of the DNA. The enzyme remains covalently attached to the phosphate group at the end of the broken strand after cutting it. Vectors are commercially available with Vaccinia virus topoisomerase I covalently attached to phosphate groups at the 3′ ends (Figure 2.16). When the prepared vector is mixed with the DNA fragment to be cloned, the enzyme transfers the phosphate linkages to the 5′ ends of the fragment, thus joining the insert to the vector. Since the topoisomerase is already attached to the vector, the reaction only requires two molecules (the vector and the insert) to come into contact, and is

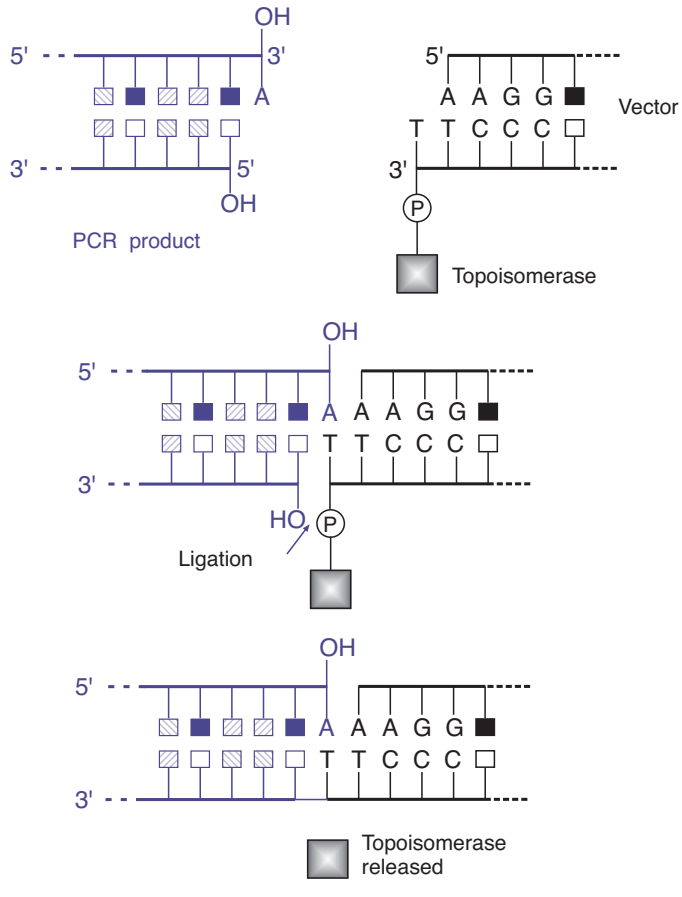


Figure 2.16 Ligation using topoisomerase.

therefore quicker, and often more efficient, than conventional ligation, which needs simultaneous contact between vector, insert and ligase.

Other enzyme systems are available for concerted breakage and rejoining of DNA strands, including the site-specific recombinase of bacteriophage lambda and the Cre-Lox system from the P1 bacteriophage.

2.8 Modification of restriction fragment ends

Restriction fragments with sticky ends are useful as they can be readily ligated. But there is a limitation on their usefulness, as they can only be ligated to another fragment with compatible ends. So an *Eco*RI fragment can be ligated to another *Eco*RI fragment, but not to one generated by *Bam*HI.

This is a potential nuisance, as the vector you are using will only have a limited number of possible sites into which you can insert DNA (and you

may specifically want to put your insert in a particular position), and it may not be possible to generate a suitable insert with the same enzyme. The best strategy in this situation is to add short oligonucleotides (linkers or adaptors) to the ends of your insert fragments.

2.8.1 Linkers and adaptors

Linkers are short synthetic pieces of DNA that contain a restriction site. For example, the sequence CCGGATCCGG contains the *Bam*HI site (GGATCC). Furthermore it is self-complementary, so you only need to synthesize (or buy) one strand; two molecules of it will anneal to produce a double-stranded DNA fragment 10 base pairs long (see Figure 2.17). If this is joined to a blunt-ended potential insert fragment by blunt-end ligation, your

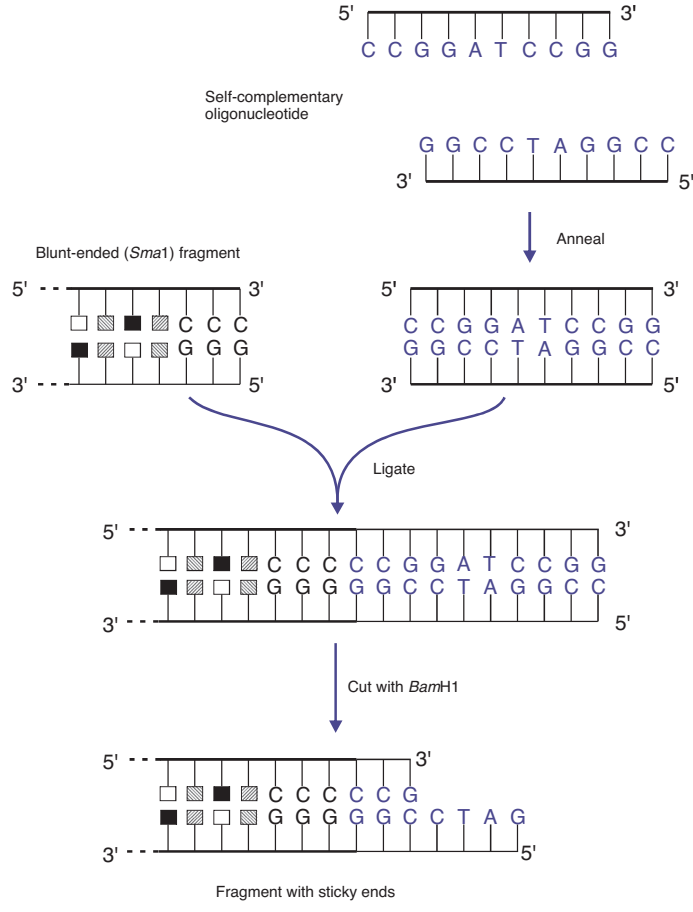


Figure 2.17 Linkers.

fragment now will have a *Bam*HI site near each end. Cutting this with *Bam*HI will generate a fragment with *Bam*HI sticky ends, which can be ligated with a *Bam*HI-cut vector.

You may object that this still requires inefficient blunt-end ligation, to join the linker to the potential insert. However, the efficiency of blunt-end ligation can be markedly improved by using high concentrations of at least one of the components. In this case, you can easily produce and use large amounts of the linker. Furthermore, since the linker is very small (e.g., 10 bases), and it is the *molar* concentration that is important, even modest amounts of the linker by mass will represent an enormous excess of linker in molar terms. For example, if you use 100 ng of a 1 kb insert, then 10 ng of linker will represent a 10 : 1 linker : insert ratio. The high molar concentration of the linker will drive the reaction very effectively. Of course, this efficient ligation is likely to add multiple copies of the linker to the ends of your insert, but this is not a problem – the subsequent restriction digestion will remove them.

Further versatility can be obtained by the use of *adaptors*. These are pairs of short oligonucleotides that are designed to anneal together in such a way as to create a short double-stranded DNA fragment with different sticky ends. For example, the sequences 5′-GATCCCCGGG and 5′-AATTCCCGGG will anneal as shown in Figure 2.18 to produce a fragment with a *Bam*HI sticky end at one end and an *Eco*RI sticky end at the other, without needing

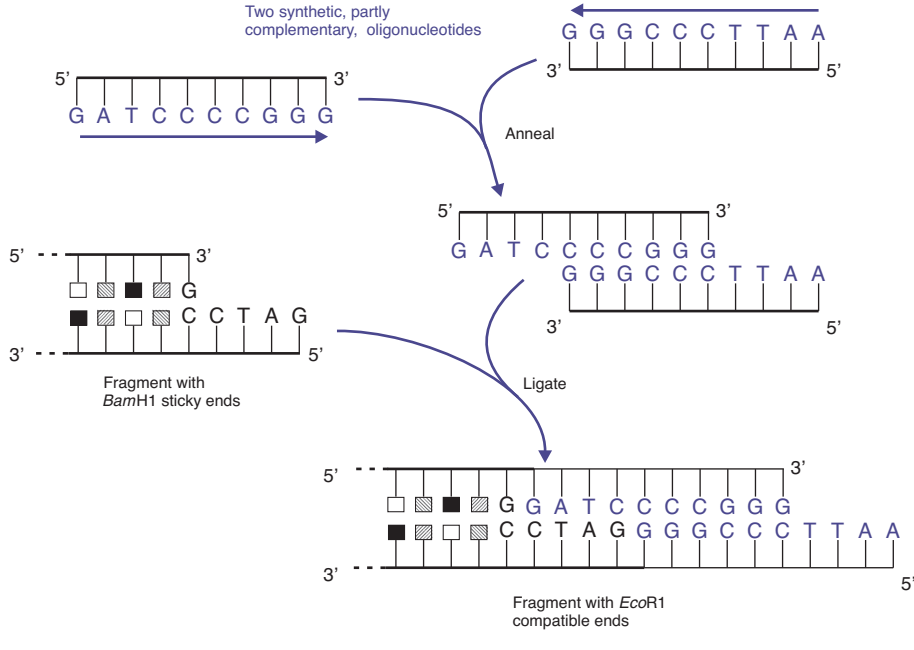


Figure 2.18 Adaptors.

to be cut by a restriction enzyme. Ligation of this adaptor to a restriction fragment generated by *Bam*HI digestion will produce a DNA fragment with *Eco*RI ends that can now be ligated with an *Eco*RI cut vector.

Alternatively, an adaptor with one sticky end and one blunt end can be used to convert blunt-ended DNA fragments, such as those generated by cDNA synthesis (see Chapter 3), into fragments with a sticky end, which increases the cloning efficiency substantially. As with linkers, you can use high molar concentrations of adaptors to drive ligation very efficiently, and the use of adaptors with non-phosphorylated sticky ends ensures that you will not get multiple additions of the adaptor to the end of your DNA fragment.

Linkers and adaptors have applications much wider than adding restriction sites. In particular, you can use the same methods to add, to the ends of all the DNA fragments in a mixture, short oligonucleotides that will act as recognition sites for a pair of PCR primers (see Chapter 4). PCR amplification will then give you an abundant supply of the full range of DNA fragments.

Other methods can also be used to introduce, or remove, restriction sites. In Chapter 4, we describe the use of modified PCR primers to add restriction sites to the ends of a fragment to be cloned, and Chapter 7 includes the use of *in vitro* mutagenesis for adding or removing restriction sites at internal positions within a DNA fragment.

2.8.2 Homopolymer tailing

An alternative way of adding sticky ends to DNA molecules is to use the enzyme *terminal deoxynucleotide transferase* (or *terminal transferase* for short). When supplied with a single deoxynucleotide triphosphate (say dGTP), this enzyme will repetitively add nucleotides to the 3′-OH end of a DNA molecule. (Figure 2.19). This enzyme is different from DNA polymerases in that it does not require a template strand, so this reaction will produce a molecule with a single-stranded run of the nucleotide supplied (G in this example) at each 3′ end, hence the term homopolymer tailing. If the vector is treated in this way, and the insert fragment(s) are treated with terminal transferase and dCTP (generating a tail of C residues), the two tails are complementary and will tend to anneal to one another (Figure 2.20).

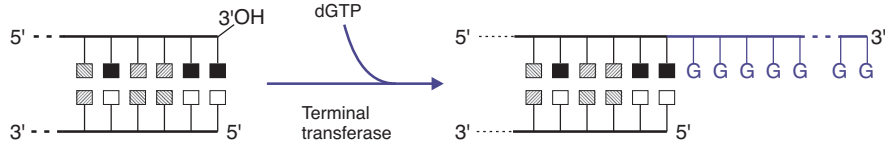


Figure 2.19 Tailing with terminal transferase.

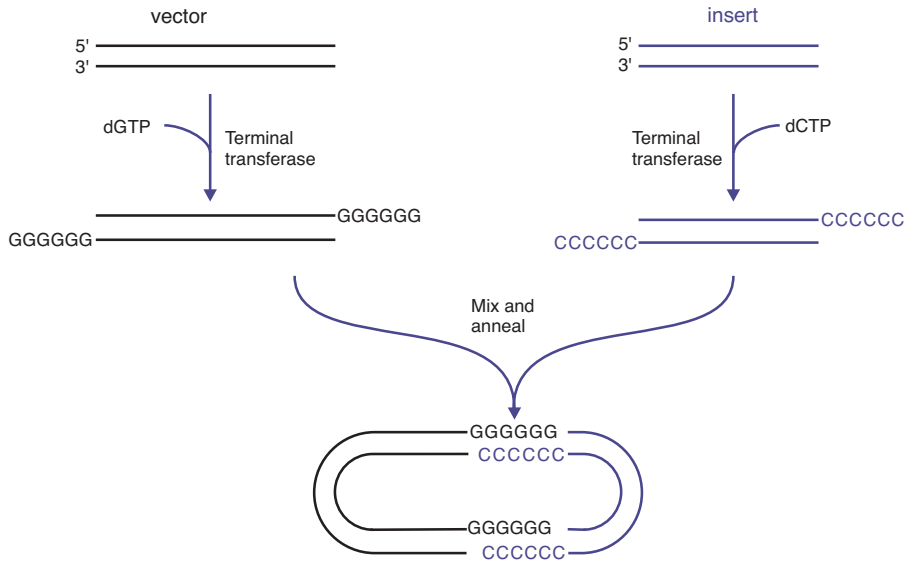


Figure 2.20 Cloning using homopolymer tailing.

It is not possible to ensure that the tails are all exactly the same length, so when the molecules anneal together there will be gaps in one of the sequences. This does not matter. If the tails are longer than about 20 nucleotides, then the pairing of the tails will be strong enough to be stable at room temperature, and the product can be used for transformation without repairing the gaps. Gaps, and nicks, will be repaired within the host cell after transformation.

One advantage of this strategy is that it is not possible for the vector to reform without an insert: the ends of the vector are not complementary to one another. The downside is that you have constructed a recombinant plasmid that contains a variable number of GC base pairs at either end of the insert, so it lacks the precision associated with the other methods described in this chapter. This is a disadvantage if you want to recover the insert from the recombinant vector, for example to reclone it in another vector. However, you can use restriction sites in the flanking region of the vector to release the insert. Or you can sequence the ends of the insert and use that information to design PCR primers (see Chapter 4) to allow you to specifically amplify the insert.

2.9 Plasmid vectors

Having seen how restriction endonucleases can be used to cut DNA, and how the fragments can be joined together, we now need to look at the nature of

the vectors into which the required fragment can be inserted, and how the resulting DNA can be introduced into a bacterial cell.

Plasmids are by far the most widely used, versatile and easily manipulated vectors. Plasmids are found in most bacterial species, as extrachromosomal DNA molecules, usually circular, double-stranded and supercoiled. They vary considerably in size, from a few thousand base pairs up to several hundred kilobases, although the plasmids used as gene cloning vectors are usually small (typically 2–5 kb). Most of the commonly used plasmid vectors are based on (or are closely related to) a naturally occurring *E. coli* plasmid called ColE1. Later in this chapter we will look at other types of vectors such as bacteriophages.

The most notorious property of plasmids lies in their ability to disseminate antibiotic resistance genes. However, it should be noted that the plasmids used for gene cloning are nearly always unable to spread from one bacterium to another, and there are restrictions on experimental protocols to ensure that these experiments do not add new resistance genes to clinically important pathogenic bacteria. Antibiotic resistance is not the limit of the ability of plasmids, nor the reason for their existence. Many naturally occurring plasmids code for other properties, although some appear to code for no obvious property at all, or at least none that we can discern. Plasmids exist because they can replicate within bacteria, and sometimes spread from one bacterium to another. That is all. They are a form of DNA parasite. Any advantage they confer on the host bacterium is a bonus that helps the plasmid to survive.

2.9.1 Plasmid replication

In genetic engineering, this ability of plasmids to be replicated enables us to insert pieces of DNA that are then copied as part of the plasmid, and hence passed on to the progeny when the cell replicates. The most fundamental property of a plasmid is therefore the ability to replicate in the host bacterium. With the simplest plasmids, most or all of the enzymes and other products needed for this replication are already present in the host cell; the amount of information that the plasmid has to supply may be only a few hundred base pairs. This region of the plasmid that is necessary for replication is generally referred to as the origin of replication (*ori*), although literally the origin, or the site at which replication starts, is one specific base.

Plasmids that use the origin of replication from ColE1 or its relatives are known as multi-copy plasmids. Wild-type ColE1 is present at about 15 copies per cell, but most of the engineered vectors used today are present in numbers running into many hundreds of copies per cell. This is convenient in some ways as it makes it easier to purify large amounts of the plasmid, because there are multiple copies within each bacterium. In addition, if you want to express a cloned gene you also get a gene dosage effect, with the presence of

so many copies of the gene in the cell reflected in higher levels of the product of that gene (see also Chapter 7). But this can also be a disadvantage: Even without expression of the cloned gene, the large amount of plasmid DNA that also needs to be replicated by the host bacterium may make the cell grow more slowly. The effect may not be great, but it can lead to instability, as any cells that have lost the plasmid will outgrow those that retain it. This will be exacerbated if the cloned gene or its product is in any way harmful to the bacterium. In extreme cases, it can sometimes be very difficult to isolate the required clone. For some specific purposes therefore it is desirable to use alternative vectors that exist at low copy number (or to use different vectors altogether that do not require continued viability of the cell, such as some types of bacteriophage vector – see below).

Some plasmids are able to replicate in a wide variety of bacterial species (broad host-range plasmids), but most of those that are used for gene cloning are rather more restricted in their host range. In one way this is useful: if there is any question about potential health hazards or environmental consequences associated with cloning a specific fragment of DNA, then using a narrow host range plasmid makes it unlikely that the gene will be transmitted to other organisms.

On the other hand, you may wish to carry out genetic manipulations in a bacterium other than *E. coli*, especially if your interest lies in studying the behaviour of specific bacteria rather than simply using them to clone pieces of DNA. It will then usually be necessary to isolate or construct new vector plasmids, based on a replication origin that is functional in your chosen species. The host range of your new vector will probably also be limited, and it may well be unable to replicate in *E. coli*. This is a disadvantage, because you probably want to use *E. coli* as an intermediate host for the initial cloning and for studying the structure and behaviour of the gene that you have cloned. However, it is possible to insert two origins of replication into your plasmid, so that it will be replicated in *E. coli* using one origin, and in your chosen host using the alternative replication origin. Such a vector is known as a *shuttle plasmid*, because it can be transferred back and forth between the two species. We can also use shuttle vectors to transfer cloned genes between *E. coli* and a eukaryotic organism. We will be coming across various applications of shuttle vectors in subsequent chapters.

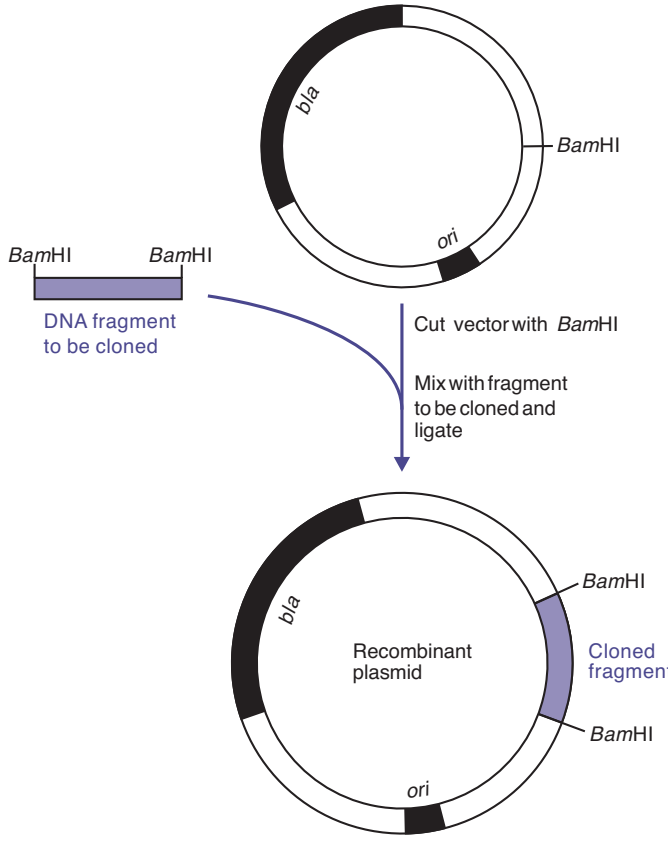
Therefore, the first essential characteristic of a plasmid cloning vector is the origin of replication, usually designated as *ori* in plasmid maps.

2.9.2 Cloning sites

The second necessary characteristic of a vector is a *cloning site*, which is where the new DNA fragment will be inserted. This must of course be located in a region of the plasmid that is not essential for replication or any other

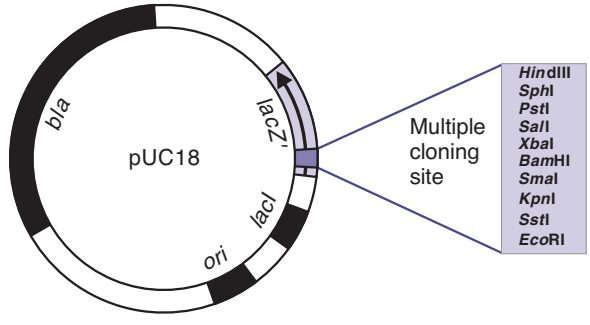
functions that we need. It will contain a unique site at which a specific restriction enzyme will cut, i.e., the enzyme will cut the plasmid once only. If a circular molecule is broken at one position, it is converted into a linear molecule, and it is relatively simple to join the ends together to reform an intact circle. If, however, an enzyme cuts more than once, the plasmid will be cut into two or more pieces, and joining them up again, in the correct order, to remake the intact plasmid will be much more inefficient.

If the vector contains only one cloning site, it would be difficult to insert more than one fragment. If you insert say a *Bam*HI fragment into a site on the vector that has been cut with *Bam*HI, the resulting recombinant plasmid will have two *Bam*HI sites: one at each end of the inserted fragment (Figure 2.21). Since the recombinant plasmid now has two *Bam*HI sites, it would be difficult to clone further *Bam*HI fragments into it.



bla = beta-lactamase (ampicillin resistance); selective marker
ori = origin of replication

Figure 2.21 Cloning with a plasmid vector.



bla = beta-lactamase (ampicillin resistance); selective marker
ori = origin of replication
lacZ' = beta-galactosidase (partial gene)
lacI = repressor of *lac* promoter

Figure 2.22 Structure of the plasmid cloning vector pUC18.

The problem here is that in many cases we do want to insert several fragments into the same plasmid. We may want to combine the expression signals of one gene with the coding region of another, or we may want to insert additional markers that can be used to identify the presence of the plasmid. The best way around this problem is to create a *multiple cloning site* (MCS), i.e., a short DNA region that contains recognition sites for a number of different enzymes. Importantly, these restriction sites will usually only appear once within the entire plasmid, meaning that they are unique to the MCS, increasing their utility as sites for insertion of DNA. An MCS is originally created by synthesizing a short piece of DNA with the required restriction sites, and inserting that into the plasmid in the usual way. Figure 2.22 shows the structure of pUC18, one of a family of similar plasmids that are commonly used as cloning vectors, and you will see that pUC18 contains such a multiple cloning site. Insertion of a fragment into the *Bam*HI site, as in Figure 2.23, will still leave a selection of other sites available for further inserts.

2.9.3 Selectable markers

One further feature that is essential for a functionally useful vector is a selectable marker. The need for this arises from the inefficiency of bacterial transformation. Even with the high-efficiency systems that are now available for *E. coli* (see below), the best yield available, using native plasmid DNA, implies that only about 1% of the bacterial cells actually take up the DNA. In practice the yields are likely to be lower than this – and if you are using a host other than *E. coli*, many orders of magnitude lower. Therefore in order to be able to recover the transformed clones, it is necessary to prevent growth of the non-transformed cells (i.e., those cells that have not taken up the

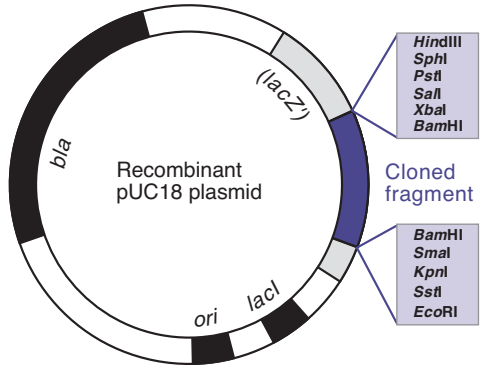


Figure 2.23 A recombinant plasmid formed using the cloning vector pUC18. The *lacZ* gene has been disrupted by insertion of a DNA fragment, resulting in white colonies on X-gal plates. *bla*, beta-lactamase (ampicillin resistance) – selective marker; *ori*, origin of replication; *lacZ′*, beta-galactosidase (partial gene); *lacI*, repressor of lac promoter.

plasmid). The presence of an antibiotic resistance gene on the plasmid vector means that you can simply plate out the transformation mix on an agar plate containing the relevant antibiotic, and only the transformants will be able to grow. In Figure 2.22 you can see that pUC18 carries a beta-lactamase gene (*bla*), coding for an enzyme that hydrolyses beta-lactam (penicillin-like) antibiotics such as ampicillin (and hence often referred to as Amp^R, for ampicillin resistance).

2.9.4 Insertional inactivation

In Figure 2.22 you will see a further useful feature of pUC18. The multiple cloning site is near to the 5′ end of a beta-galactosidase gene (*lacZ′*). The synthetic oligonucleotide that creates the multiple cloning site was designed so that it does not affect the reading frame of the *lacZ* gene. It merely results in the production of beta-galactosidase with some additional amino acids near to the amino terminus of the protein. This does not affect the function of the enzyme; it is still able to hydrolyse lactose. (More accurately, we should say that pUC18 carries a *part* of the *lacZ* gene; we use *E. coli* strains that carry the remainder of the gene. The product of the host gene is unable to hydrolyse lactose by itself, and so the host strain without the plasmid is Lac^-, i.e., it does not ferment lactose. When pUC18 is inserted into the host, the plasmid-encoded polypeptide will associate with the host product to form a functional enzyme. We say that pUC18 is capable of *complementing* the host defect in *lacZ*).

We can easily detect the activity of the beta-galactosidase by plating the organism onto agar containing the chromogenic substrate *X-gal*, together with

the inducing agent isopropyl thiogalactoside (IPTG). The X-gal substrate is colourless but the action of beta-galactosidase releases the dye moiety, resulting in a deep blue colour. Colonies carrying pUC18 are therefore blue when grown on this medium. However, if we are successful in inserting a DNA fragment at the cloning site, the gene will (normally) be disrupted (Figure 2.23) and the resulting *E. coli* will be unable to cleave X-gal, resulting in bacterial colonies referred to as 'white' due to their lack of blue coloration. The advantage of this *insertional inactivation* is that we can tell not only that the cells have been transformed with the plasmid (since they are able to grow in the presence of ampicillin), but also that the plasmid is a recombinant, and not merely the original pUC18 self-ligated. An insertional inactivation marker such as this is not an essential feature of a cloning vector but it does provide a useful way of monitoring the success of the ligation strategy and overcoming some of the problems referred to previously.

One word of caution: insertional inactivation is not 100% reliable. If the insert is relatively small, and if it happens to consist of a multiple of three bases, transcription and translation of the *lacZ* gene may still occur and the enzyme may still have enough activity (despite the addition of still more amino acids at the amino terminus) to produce a detectable blue colour. Conversely, a white colony is not a guarantee of cloning success as the deletion of even a single base at the cloning site, or the insertion of undesirable junk fragments (in other than multiples of three bases), will put the *lacZ* gene in the wrong reading frame and thus inactivate it.

The advantage of plasmid vectors, compared to the other vectors described subsequently in this chapter, is that they are small and easy to manipulate; also they are conceptually simple and universal. You can make and use plasmid vectors for a wide range of organisms without a detailed knowledge of the molecular biology of the host or the vector. However, the basic plasmid vectors that we have been considering so far are limited in their cloning capacity, i.e., the size of the insert they can accommodate. Later in this chapter, we will look at other vectors that will accommodate larger inserts, but first we need to consider the ways in which we can introduce the recombinant plasmids into host bacterial cells.

2.9.5 Transformation

Bacterial transformation was discovered in 1928 with the demonstration by Fred Griffith that cultures of the pneumococcus (*Streptococcus pneumoniae*) that had lost virulence could have their pathogenicity restored by addition of an extract of a killed virulent strain. It was the identification, many years later (by Avery and colleagues, in 1944), that the 'transforming principle' is DNA that resolved the question of the chemical nature of the genetic material.

This experiment rests on the natural ability of the pneumococcus to take up 'naked' DNA from its surroundings. This ability is known as *competence*. Although the range of bacteria that exhibit natural competence is much wider than was thought for many years, it is still too limited in scope (or too inefficient, or too selective) to be of much use for genetic engineering. In particular, *E. coli* does not seem to exhibit natural competence. It was therefore necessary to develop alternative ways of introducing plasmid DNA into bacterial cells. Although these methods are radically different, they are all still referred to as transformation, which is defined as the uptake of naked DNA, to distinguish it from other methods of horizontal gene transfer in bacteria, namely *conjugation* (direct transfer by cell to cell contact) and *transduction* (which is mediated by bacteriophage infection).

The breakthrough came with the demonstration that competence in *E. coli* cells could be induced by washing them with ice-cold calcium chloride, followed by adding the plasmid DNA and subjecting the mixture to a brief, mild heat shock (e.g., at 42°C). However, this process was very inefficient, with yields of perhaps 10^4 transformants per microgram of pure, supercoiled plasmid DNA (and less with ligation mixtures or non-supercoiled DNA). Nowadays, following technical improvements to the system, and the selection of improved *E. coli* strains, it is possible to obtain transformation frequencies in excess of 10^9 transformants per microgram of plasmid DNA. Note that although transformation frequencies are generally quoted as number of transformants per microgram of DNA, you would usually use much less DNA than that – so the real result might be 10^6 transformants with 1 ng of DNA. Transformation works best with low levels of DNA. If you increase the amount of DNA, the number of transformants does not increase in proportion.

Reference back to the discussion of ligation will disclose a quandary here. Ligation works best with high concentrations of DNA. The following step, transformation, works best with small amounts of DNA. The resolution is clear, although unpalatable: use only a small proportion of your ligation mix in the transformation step. If you really need very large numbers of transformants, scaling up the transformation step does not work very well – it is usually much better to carry out several separate small-scale transformations.

Transformation based on induced competence and heat shock can be used for bacterial species other than *E. coli*, but you immediately lose all the advantages that have been gained by optimisation of transformation conditions for selected strains of *E. coli*. At best, therefore, transformation is likely to be very inefficient – and in most cases simply using an *E. coli* procedure will not work at all. Therefore laboratories that are interested in manipulating other bacterial species have had to develop alternative methods of transformation.

Electroporation is the most versatile transformation procedure. Bacterial cells, washed with water to remove electrolytes from the growth medium, are mixed with DNA and subjected to a brief pulse of high-voltage electricity.

This appears to induce temporary holes in the cell envelope through which the DNA can enter. The cells are then diluted into a recovery medium before plating on a selective medium in the same way as above. Although it is comparatively easy to obtain *some* transformants with a wide range of bacteria (or with most other cells), there are many parameters that need to be adjusted to obtain optimum performance, including the conditions under which the cells are grown, the temperature of the suspension, and the duration and voltage of the electric pulse.

Since the added DNA seems to simply diffuse through the holes created (briefly) by the electric pulse, the effect is not specific for DNA; other substances, notably RNA or proteins, can also be introduced into bacterial cells by electroporation. Nor is it directionally specific. Material within the cell can diffuse out as well, and the procedure has been used for isolating plasmid DNA from bacterial cells. It follows from this that, since the plasmid that comes out of one cell can enter another one, electroporation can be used to transfer plasmids from one strain to another, simply by applying it to a mixture of the two strains.

Other methods that are used more commonly with animal and plant cells, including microinjection, biolistics and protoplast transformation, are considered in Chapter 11.

2.10 Vectors based on the lambda bacteriophage

2.10.1 Lambda biology

Plasmid vectors are at their best when cloning relatively small fragments of DNA. Although there is probably no fixed limit to the size of a DNA fragment that can be inserted into a plasmid, the recombinant plasmid may become less stable with larger DNA inserts, and the efficiency of transformation is reduced. Vectors based on bacteriophage lambda allow efficient cloning of larger fragments, which is important in constructing gene libraries. The larger the inserts, the fewer clones you have to screen to find the one you want (see Chapter 3).

In order to understand the nature and use of lambda cloning vectors, some knowledge of the basic biology of bacteriophage lambda is necessary. While we hope you are familiar with this, a recapitulation of the salient features (summarized in Figure 2.24) will be useful.

Lysogeny Lambda is a temperate bacteriophage, i.e., on infection of *E. coli* it may enter a more or less stable relationship with the host known as *lysogeny*. In the lysogenic state, expression of almost all of the phage genes is switched off by the action of a phage-encoded repressor protein, the product of the *cI* gene.

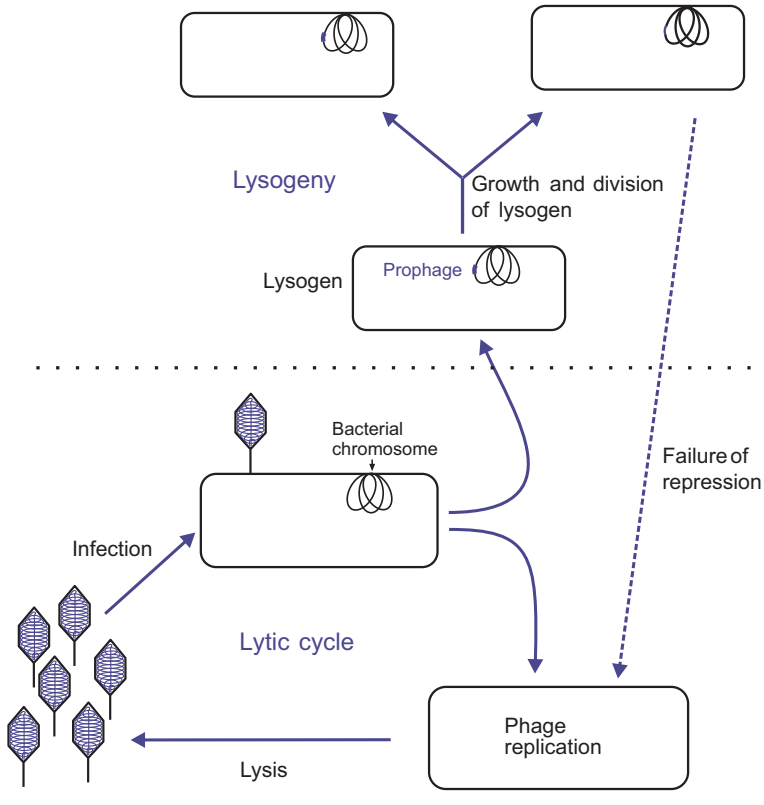


Figure 2.24 Bacteriophage growth: lytic cycle and lysogeny.

When you add a lambda phage preparation to an *E. coli* culture, some of the infected cells will become lysogenic, and some will enter the lytic cycle. The proportion of infected cells going down each route is influenced by environmental conditions, as well as by the genetic composition of the phage and the host. Some phage mutants will only produce lytic infection, and these give rise to clear plaques, while the wild-type phage produces turbid plaques due to the presence of lysogenic cells which are resistant to further attack by lambda phage (known as *superinfection immunity*). So the lysogens continue to grow within the plaque, and the plaque is therefore turbid. However, some bacterial host strains carrying a mutation known as *hfl* (high frequency of lysogenization) produce a much higher proportion of lysogens when infected with wild-type lambda, which can be useful if we want a more stably altered host strain, for example, if we are studying the expression of genes carried by the phage. Generally, when using lambda vectors we are more interested in the recombinant phage carrying the cloned genes, and the lytic cycle is the more relevant one in such cases.

Although the lysogenic state is relatively stable, that stability is not absolute. A culture of a bacterial lysogen will normally contain phage particles in the supernatant, due to a low level of spontaneous failure of the repression mechanism. This rate of breakdown of repression can be increased by treating the culture with agents that damage the DNA, such as UV irradiation; the DNA damage induces the production of repair enzymes, which, amongst other things, destroy the *cI* repressor protein, allowing initiation of the lytic cycle. Some widely used lambda vectors carry a mutation in the *cI* gene that makes the protein more temperature-sensitive (*cI857* mutation). A bacterial strain carrying such a mutant phage can be grown as a lysogen at a reduced temperature and the lytic cycle can be induced by raising the temperature, due to inactivation of the repressor protein.

In the lysogenic state, lambda is normally integrated into the bacterial chromosome, and is therefore replicated as part of the bacterial DNA. However, this integration, although common amongst temperate phages, is not an essential feature of lysogeny. Lambda can continue to replicate in an extrachromosomal, plasmid-like state. With some bacteriophages (including P1, which we will encounter later in this chapter) this is the normal mode of replication in lysogeny.

Particles of wild-type bacteriophage lambda have a double-stranded linear DNA genome of 48 514 base pairs, in which the 12 bases at each end are unpaired but complementary (Figure 2.25). These ends are therefore 'sticky', or 'cohesive', much like the ends of many restriction fragments – but the longer length of these sticky ends makes the pairing much more stable, even at 37°C. The ends can be separated by heating lambda DNA, and if it is then cooled rapidly you will get linear monomeric lambda DNA. At low temperatures, the ends of the molecule will move slowly, and therefore the reannealing of the sticky ends will take a long time. Eventually, however, it will resume a circular (although not covalently joined) structure. When lambda infects a bacterial cell and injects its DNA into the cell, the DNA forms a circular

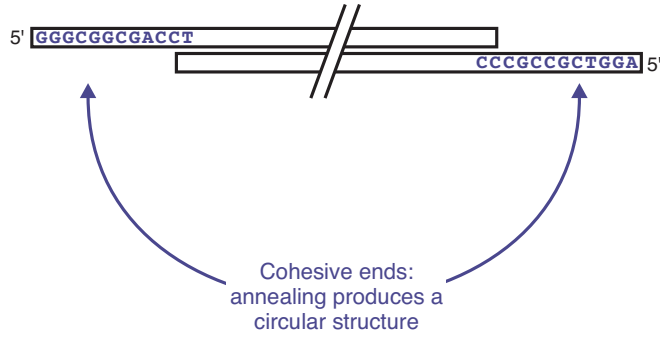


Figure 2.25 Cohesive ends of lambda DNA.

structure, with the nicks being repaired *in vivo* by bacterial DNA ligase. At about this time, a complex series of events occur that affect subsequent gene expression, determining whether the phage enters the lytic cycle or establishes lysogeny. We do not need to consider the details of the *lytic-lysogenic decision*, except to emphasize that it is essentially irreversible so that, once started on one or the other route, the phage is committed to that process. But we do need to consider the events in the lytic cycle.

Lytic cycle In the lytic cycle, this circular DNA structure is initially replicated, in a plasmid-like manner (*theta* replication), to produce more circular DNA. Eventually however, replication switches to an alternative mode (*rolling circle replication*), which generates a long linear DNA molecule containing a large number of copies of the lambda genome joined end to end in a continuous structure (Figure 2.26). While all this is going on, the genes carried by the phage are being expressed to produce the components of the phage particle. These proteins are assembled first of all into two separate structures: the head (as an empty precursor structure into which the DNA will be inserted), and the tail (which will be joined to the head after the DNA has been packaged).

The packaging process involves enzymes recognizing specific sites on the multiple length DNA molecule generated by rolling circle replication and making asymmetric cuts in the DNA at these positions. These staggered breaks in the DNA give rise to the cohesive ends seen in the mature phage DNA; these sites are known as *cohesive end sites* (*cos* sites). Accompanying these cleavages, the region of DNA between two *cos* sites – representing a unit length of the lambda genome – is wound tightly into the phage head. Following successful packaging of the DNA into the phage head, the tail is added to produce the mature phage particle, which is eventually released when the cell lyses.

One of the most important features of this process from our point of view is that the length of DNA that will be packaged into the phage head is determined by the distance between the two *cos* sites. If we insert a piece of DNA into our lambda vector, we will increase that distance, and so the amount of DNA to be packaged will be bigger. But the head is a fixed size, and can only accommodate a certain amount of DNA (up to about 51 kb altogether, which is about 5%, or 2.5 kb, more than wild-type). As one of the reasons for using lambda is to be able to clone large pieces of DNA, this would be a serious limitation. The way round this potential problem is to delete some of the DNA that is normally present. This is possible because the lambda genome contains a number of genes that are not absolutely necessary – especially if we only need lytic growth, meaning we can delete any genes that are required solely for the establishment of lysogeny. But we cannot delete too much. The stability of the phage head requires a certain amount of DNA, so even though

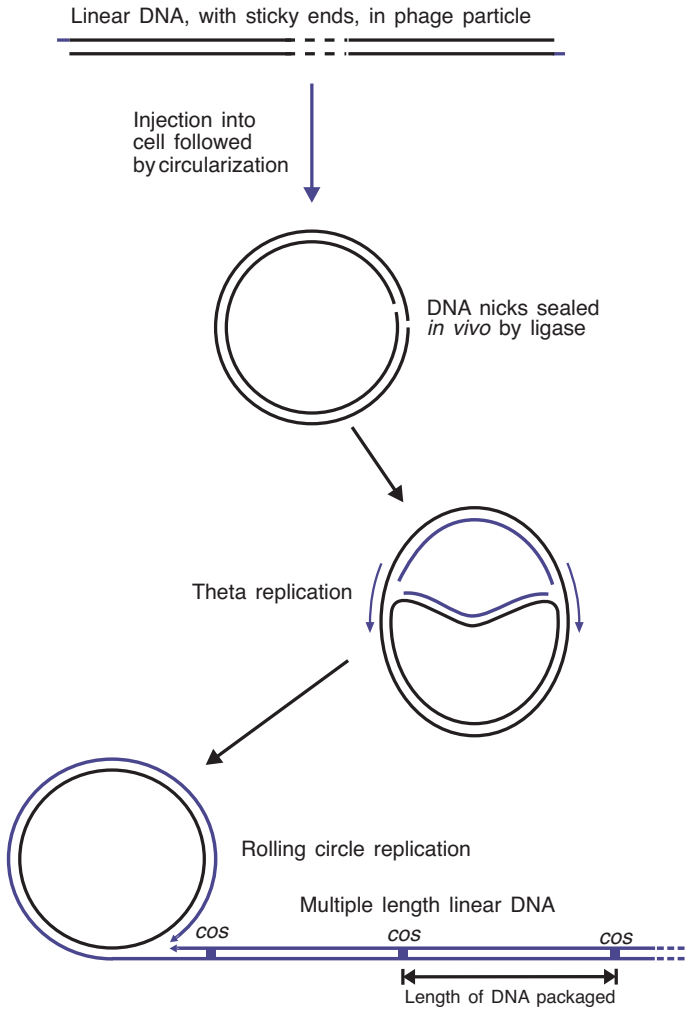


Figure 2.26 Replication of bacteriophage lambda DNA.

there are more genes that are not required, we cannot delete all that DNA. To produce viable phage, there has to be a minimum of 37 kb of DNA (about 75% of wild-type) between the two *cos* sites that are cleaved.

The existence of these *packaging limits* is a very important feature of the design and application of lambda vectors, and also of cosmids, which we will discuss later.

2.10.2 *In vitro* packaging

Naked bacteriophage DNA can be introduced into a host bacterial cell by transformation (often referred to as *transfection* when talking about phage

DNA), in much the same way as we described for a plasmid. The big difference is that in this case instead of plating on a selective agar and counting bacterial colonies, we would mix the transfection mix with a culture of a phage-sensitive indicator bacterium in molten soft agar and look for *plaques* (zones of clearing due to lysis of the bacteria) when overlaid onto an agar plate. Note that in this case we do not need an antibiotic resistance gene as a selective marker.

However, the large size of most bacteriophage DNA molecules, including that of lambda, makes transfection an inefficient process compared to plasmid transformation. But there is a more efficient alternative. Some mutant lambda phages, in an appropriate bacterial host strain, will produce empty phage heads (as they lack a protein needed for packaging the DNA), while others are defective in the production of the head, but contain the proteins needed for packaging. The two extracts are thus complementary to one another. Use of the mixture allows productive packaging of added DNA, which occurs very effectively *in vitro* (including the addition of the tails). The resulting phage particles can then be assayed by addition of a sensitive bacterial culture and plating as an overlay, as above. Since *in vitro* packaging of lambda DNA is much more effective than transfection, it is the method that is almost always used.

One feature of this system that is markedly different from working with plasmid vectors is that the packaging reaction is most efficient with multiple length DNA. The enzyme involved in packaging the DNA normally cuts the DNA at two different *cos* sites on a multiple length molecule; monomeric circular molecules with a single *cos* site are packaged very poorly. So whereas with plasmid vectors the ideal ligation product is a monomeric circular plasmid consisting of one copy of the vector plus insert, for lambda vectors it is advantageous to adjust the ligation conditions so that we *do* get multiple end-to-end ligation of lambda molecules together with the insert fragments. The stickiness of the ends of the linear lambda DNA means this happens very readily.

2.10.3 Insertion vectors

The simplest form of lambda vector, known as an *insertion vector*, is similar in concept to a plasmid vector, containing a single cloning site into which DNA can be inserted. However, wild-type lambda DNA contains many sites for most of the commonly used restriction enzymes; you cannot just cut it with say *Hin*dIII and ligate it with your insert DNA. *Hin*dIII has seven sites in normal lambda DNA, and so will cut it into eight pieces. (Note the difference between a circular DNA molecule such as a plasmid, and a linear molecule like lambda: cut a circular DNA molecule once and you still have one

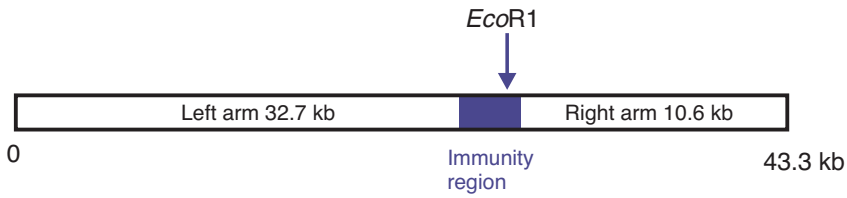


Figure 2.27 Lambda insertion vector gt10.

fragment; cut linear DNA once and you have two fragments.) It would be almost impossible to join all these fragments (and your insert) together in the right order. To circumvent this, all lambda vectors have been genetically manipulated to remove unwanted restriction sites.

In Figure 2.27 we see one example of a lambda vector, known as lambda gt10. In this vector, there is only a single site at which *Eco*RI will cut the DNA. The manipulations that this phage has undergone have removed the unwanted sites, and have also reduced the overall size of the phage DNA to 43.3 kb (which is still large enough to produce viable phage particles), hence allowing the insertion of foreign DNA up to a maximum of 7.6 kb.

Cutting this vector with *Eco*RI will produce two DNA fragments, referred to as the left and right arms. Although it therefore appears that the insert would have to be ligated to two different pieces of vector DNA, in practice this does not complicate the ligation as much as might be imagined. One end of each of the two arms is derived from the cohesive ends of the lambda DNA, and will therefore anneal quite stably at 37°C – so although not covalently joined they can be considered as a single DNA fragment.

Lambda gt10 also provides us with another example of how insertional inactivation can be used to distinguish the parental vector (which may form by re-ligation of the arms without an insert) from the recombinants. The *Eco*RI site is found within the repressor (*cI*) gene, so the recombinant phage, which carry an insert in this position, are unable to make functional repressor. As a consequence, they will be unable to establish lysogeny and will give rise to clear plaques, whereas the parental gt10 phage will give rise to turbid plaques. So, by picking the clear plaques you can select for recombinant phage, as opposed to re-ligated vector, without having to use dephosphorylation.

Another, rather special, example of an insertional vector is known as lambda gt11 (Figure 2.28). This has been engineered to contain a beta-galactosidase gene, and has a single *Eco*RI restriction site within that gene – but in contrast to pUC18, the cloning site is towards the 3′ end of the beta-galactosidase gene. This confers two properties on the vector. Firstly, insertion of DNA at the *Eco*RI site will inactivate the beta-galactosidase gene, so that recombinants will give ‘white’ (actually colourless) plaques on a medium containing X-gal. Secondly, the insert, if in the correct orientation and in

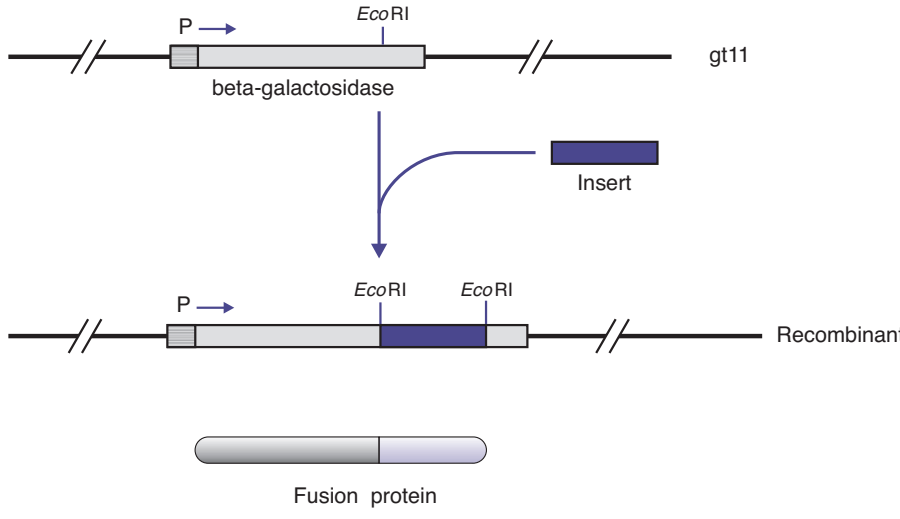


Figure 2.28 Use of lambda gt11 for generation of fusion proteins.

frame, will give rise to a fusion protein containing the product encoded by the insert fused to the beta-galactosidase protein. This fusion protein is unlikely to have the biological functions associated with your cloned gene, but that is not the point. It *is* reasonably likely to react with some antibodies to the natural product, which makes it a useful way of detecting the clone of interest, as we will see in Chapter 3. Lambda gt11 is an example of a translational fusion expression vector, a topic we will deal with more fully in Chapter 7.

Since the packaging limits for lambda DNA are between 37 and 51 kb, we cannot make an insertion vector smaller than 37 kb, or we would be unable to grow it to produce the DNA that we need. And we cannot insert a DNA fragment so big that it would make the product larger than 51 kb; the recombinant DNA would be unable to be packaged into the phage heads. It follows that the maximum cloning capacity for an insertion vector is $(51 - 37) = 14$ kb. This is considerably larger than we could clone comfortably in a plasmid vector, but still smaller than we would like for some purposes. In order to increase the cloning capacity, we have to turn to a different type of lambda vector, known as a *replacement vector*.

2.10.4 Replacement vectors

The packaging limits that restrict the cloning capacity of insertion vectors are imposed by the physical requirements of the phage head rather than by the nature of the genes needed. There are further genes that are not essential for lytic growth and could be deleted, except that it would make the phage DNA

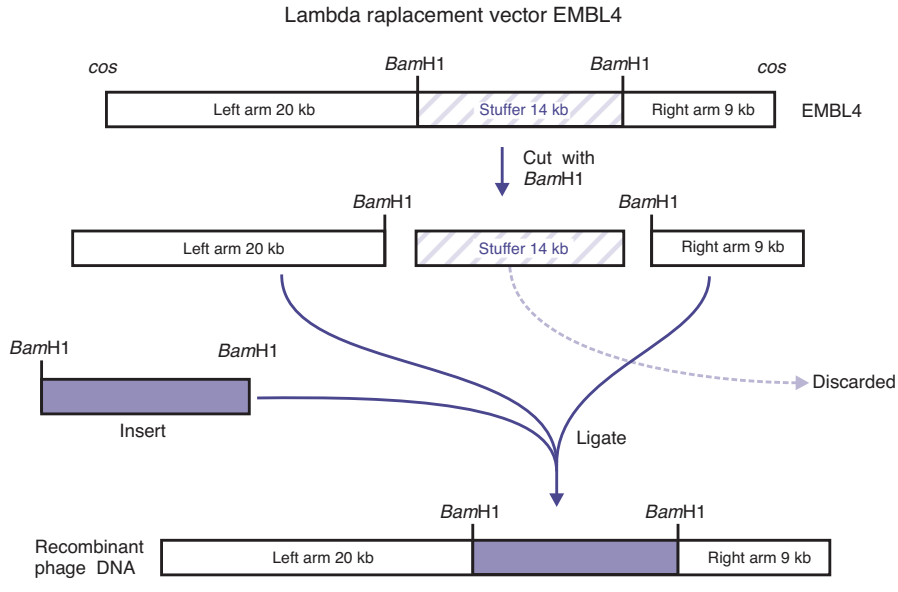


Figure 2.29 Lambda replacement vector EMBL4.

too small to produce viable progeny. That provides a clue to an alternative design of lambda cloning vectors. Instead of merely inserting extra DNA, arrange the vector so that a piece of DNA can be removed and replaced by your insert – hence the term *replacement vector*.

Figure 2.29 shows an example of a lambda replacement vector, EMBL4. Instead of being cut just once by the restriction enzyme of choice (in this case *Bam*HI), there are two sites where the DNA will be cleaved. The vector DNA will therefore be cut into three fragments: the left and right arms (which will anneal by virtue of their cohesive ends) and a third fragment that is not needed (except to maintain the size of the DNA) and can be discarded. Since the only purpose of this fragment is to help fill up the phage head it is known as a *stuffer fragment*.

In use, this vector would be cut with *Bam*HI and the fragments separated, for example, by gel electrophoresis. The stuffer fragment would be thrown away, and the arms mixed with the restriction fragments to be cloned. These could be generated by *Bam*HI digestion, or by cleavage of the target with another enzyme that produces compatible ends, for example, *Sau*3A. Ligation of the mixture produces recombinant phage DNA that can be packaged into phage heads by *in vitro* packaging. The cloning capacity of the vector is thus considerably increased; in this case the size of the arms combined comes to 29 kb and thus you can clone fragments up to 51 – 29 kb, = 22 kb.

There is a further advantage to such a vector. The combined size of the arms is only 29 kb, which is less than the minimum required for packaging.

Any pairs of arms that are ligated without an insert will therefore be too small to produce viable phage particles, as viable particles will only be produced if ligation results in an insert of at least 37 – 29 kb, = 8 kb. The vector thus provides a positive selection for recombinants as opposed to parental phage, and furthermore for recombinants that contain an insert of at least 8 kb. The gene library will therefore be free of non-recombinant phage.

Earlier in this chapter we discussed strategies for ensuring that we obtained recombinant plasmids rather than parental vector molecules, including alkaline phosphatase treatment of the vector to prevent recircularization. As detailed above, any re-ligation of the arms of a replacement vector will result in non-viable phage particles, and hence CIP treatment is not necessary for replacement vectors. Since we do not have to treat the vector with phosphatase, we have another possibility – dephosphorylation of the *insert*. In the production of a gene library, the insertion of more than one fragment into the same vector molecule is a problem that can give rise to anomalies in characterizing the insert in relation to the genome it came from. Phosphatase treatment of the insert will prevent insert-insert ligation, and hence will ensure that all of the recombinants carry only a single insert fragment.

There is yet another useful feature that can be built into a replacement vector. Since the stuffer fragment is not necessary for phage production (apart from filling up the phage head), it does not have to be lambda DNA. It can be anything we want. So we could, for example, put in a fragment carrying a beta-galactosidase gene. Then any plaques formed by phage that still carry the stuffer fragment would be blue (on a medium containing X-gal). Of course ideally there should not be any. But there may be some phage DNA molecules that have not been cut completely, or there may be some stuffer DNA contaminating the preparation of the vector arms, which could then be ligated back into the vector. Any plaques containing the stuffer will be blue – so you have an immediate check that everything has gone according to plan. Or not.

So we see that lambda vectors provide a highly versatile and efficient system for primary cloning of unknown fragments, especially in the construction of genomic and cDNA libraries (see Chapter 3). They extend the cloning capacity over that readily obtainable with plasmid vectors by around five times, and can easily generate the very large numbers of recombinants that are required for a gene library, at least for a bacterial genome.

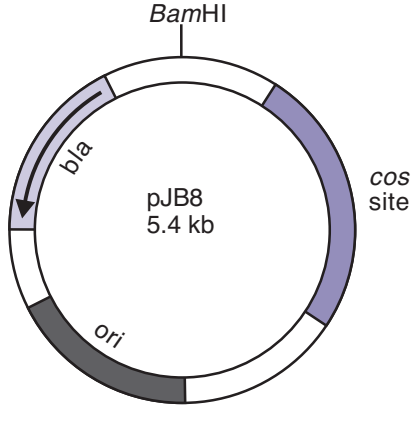
Although lambda phages are the most widely used phage vectors, there are other phage vectors that are used for specific purposes, including the filamentous bacteriophage M13. This was used extensively in the past as it provided a convenient way of obtaining single-stranded copies of the cloned DNA, which was necessary for the early sequencing strategies. However, modern sequencing techniques do not need single-stranded DNA, so we will not discuss M13 in detail. Another bacteriophage, P1, is capable of accommodating

larger inserts, and is referred to later in this chapter. But first we need to look at a special class of vector that combines some of the features of lambda and plasmid vectors. These are the *cosmids*.

2.11 Cosmids

The lambda packaging reaction has two fundamental requirements: the presence of a *cos* site, and the physical size of the DNA. Cosmids exploit this relatively simple set of requirements to provide cloning vectors with a capacity larger than can be achieved with lambda replacement vectors.

Basically, a cosmid is simply a plasmid that contains a lambda *cos* site. As with all plasmid vectors, it has an origin of replication, a selectable marker (usually an antibiotic resistance gene) and a cloning site. Digestion, ligation with the potential insert fragments, and subsequent purification of recombinant clones, are carried out more or less as for a normal plasmid vector (see Section 2.9.5). However, instead of transforming bacterial cells with the ligation mix, as you would with a plasmid, you subject the ligation mixture to *in vitro* packaging as described in Section 2.10.2 for lambda vectors. Since the cosmid carries a *cos* site, it can be a substrate for *in vitro* packaging – but only if it is big enough. The vector itself is quite small – in the example shown in Figure 2.30, it is 5.4 kb, which is much too small for successful packaging. The packaging reaction will only be successful if you have inserted a DNA fragment between about 32 and 45 kb in size. So you not only have an increased cloning capacity, but also a positive selection for an



ori = origin of replication
bla = beta-lactamase (ampicillin resistance)
cos site = region of lambda DNA required for packaging

Figure 2.30 Structure of a cosmid.

insert, and for an insert that can be up to 10 times larger than is common for plasmid clones.

Of course the products of the packaging reaction, although they are phage-like particles, will not give rise to more phages after infection of a host bacterium. They do not carry any of the genes that are needed for production of phage particles, nor for lysis of the cell. So you will not get phage plaques. But the cosmid will replicate as a plasmid, and hence will give rise to more cosmid-containing cells, and these can be selected as colonies on agar containing an antibiotic (in this case ampicillin).

For very large genomes, such as mammalian ones, the cloning capacity of cosmids is still rather too small for convenience, and alternative vectors with even greater capacity are available (see below). But for smaller genomes, cosmids can be extremely valuable. A complete bacterial genome can be covered by only a few hundred cosmids, which was useful in genome mapping and sequencing projects (although now superseded by genome sequencing, as described in Chapter 8).

2.12 Supervectors: YACs and BACs

Although cosmids were the first vectors that made the production and use of mammalian gene libraries feasible, their limited capacity would still not have sufficed for decoding the human genome. This was only made possible by the development of novel supervectors that were able to carry 100 kb or more. The first of these were the yeast artificial chromosome (YAC) vectors (Figure 2.31). As with the shuttle plasmids we referred to earlier, a YAC vector is propagated as a circular plasmid in *E. coli*. Restriction enzyme digestion removes the stuffer fragment between the two telomeres, and cuts the remaining vector molecule into two linear arms, each carrying a selectable marker. The insert is then ligated between these arms, as in the case of phage lambda, and transformed into a yeast cell, with selection for complementation of auxotrophic markers; this ensures that the recombinants contain both arms. Furthermore a successful recombinant must contain the *tel* sequences at each end, so that the yeast transformant can use these sequences to build functional telomeres. The titans amongst vectors, YACs are routinely used to clone 600 kb fragments, and specialized versions are available that can accommodate inserts close to 2 Mb, which is approximately one thousand times more DNA insert than in a plasmid. As such, they will not only easily accommodate any eukaryotic gene in its entirety, but also the gene will be within its framework of three-dimensional structure and distant regulatory sequences. They have therefore been very useful in the production of transgenic organisms (see Chapter 11).

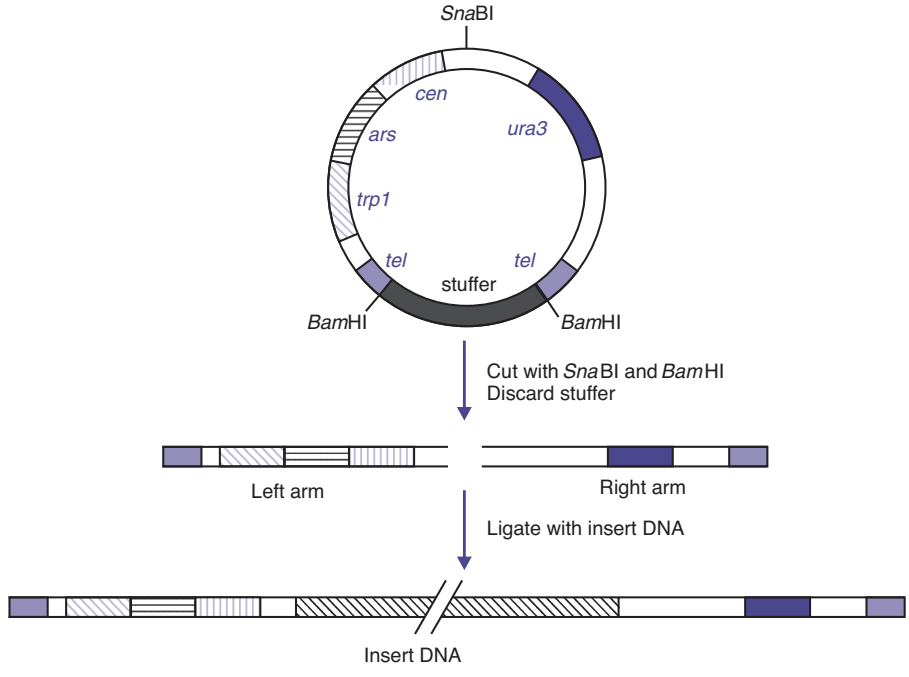


Figure 2.31 Structure and use of a yeast artificial chromosome vector.

However, YACs have problems with the stability of the insert, especially when working with very large fragments as these can be subject to rearrangement by recombination. Furthermore, apart from the fact that many laboratories are not set up for the use of yeast vectors, the recombinant molecules are not easy to recover and purify. Thus, larger bacterial vectors are used more than YACs even though their capacity is lower. These include vectors based on bacteriophage P1, which are able to accommodate inserts in excess of 100 kb, and bacterial artificial chromosomes (BACs), which are based on the F plasmid and can accommodate 300 kb of insert. These vectors are more stable than YACs, and have played an important role in genome sequencing projects (see Chapter 8).

2.13 Summary

In this chapter we have described the basic technology needed for cloning pieces of DNA. Further topics such as the use of expression vectors for optimising product formation, and vectors for eukaryotic cells, are described in Chapter 7. One essential aspect of the basic techology has not been dealt with, namely how do you use these methods to get hold of a clone carrying a

specific gene? In the next chapter, we will look at the construction and screening of gene libraries, which formed the central part of the earlier achievements in this field. Nowadays, with the amount of information readily available from genome sequencing projects, for many purposes it is often possible to bypass the use of gene libraries and use the polymerase chain reaction to amplify the required gene. We will deal with this in Chapter 4.

3
Genomic and cDNA Libraries

In the previous chapter, we described the techniques needed for cloning a fragment of DNA. One essential feature remains to be considered – how do you use these methods to obtain the specific fragment of DNA carrying the gene you are interested in? Even a small and relatively simple organism like a bacterium contains thousands of genes, and they are not discrete packets, but regions of a continuous DNA molecule. If we break this molecule into smaller fragments, we will obtain a very large number of different fragments of DNA with no easy way of reliably purifying an individual fragment, let alone isolating the specific fragment that carries the required gene.

Fortunately, it is not necessary to purify specific DNA fragments. One of the strengths of gene cloning is that it provides a much more powerful way of finding a specific piece of DNA. Rather than attempting to separate the DNA fragments, we take the complete mixture and use DNA ligase to insert the fragments into the prepared vector. Under the right conditions, only one fragment will be inserted into each vector molecule. In this way, we produce a mixture of a large number of different recombinant vector molecules, which is known as a *gene library*. On transforming a bacterial culture with this library, each cell will only take up one molecule. When we then plate the transformed culture, each colony, which arises from a single transformed cell, will contain a large number of bacteria all of which carry the same recombinant plasmid, with a copy of the same fragment of DNA from our starting mixture. So instead of a mixture of different DNA fragments, we have a large number of bacterial colonies, each of which carries one fragment only (Figure 3.1). We still have a very complex mixture, but whereas purifying an individual DNA fragment is extremely difficult, it is simple to isolate

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

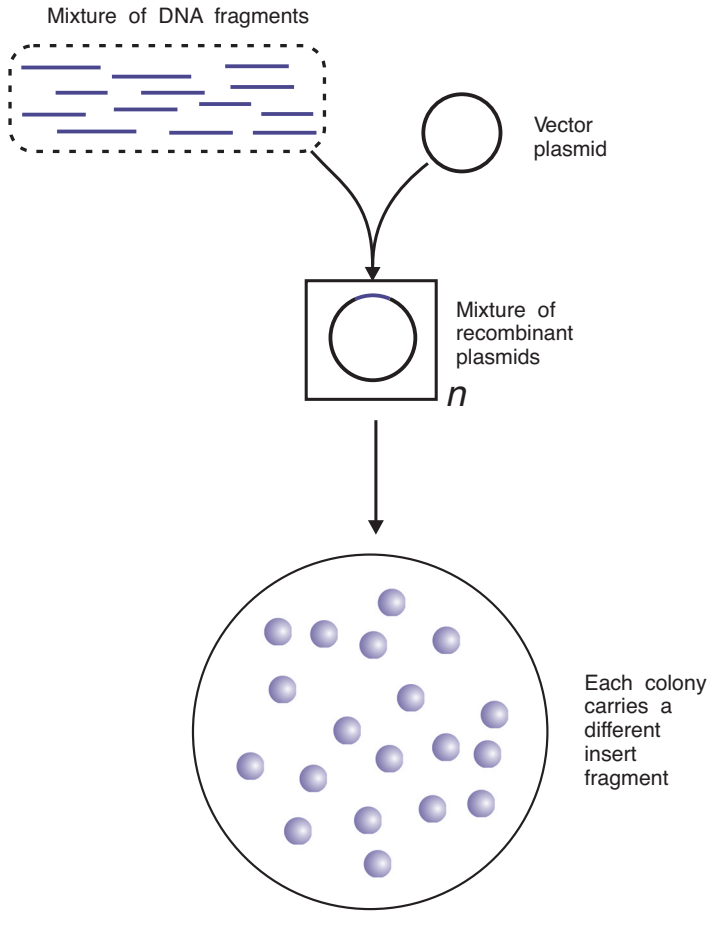


Figure 3.1 Making a genomic library.

individual bacterial colonies from this mixture – we just pick them from a plate with something as simple as a toothpick. As each individual bacterial colony will carry a different piece of DNA from our original complex mixture we can identify which bacterial colony carries the gene that we are interested in, and subsequent purification of it becomes a simple matter. We just have to pick the right colony and inoculate it into fresh medium. However, we still have the problem of knowing which of these thousands/millions of bacterial colonies actually carry the gene that we want. Later in this chapter we will look at ways of screening gene libraries.

A gene library, therefore, is a collection of clones that between them represent the entire genome of an organism. More specifically, we should refer to such a library as a *genomic* library, to distinguish it from a different sort of gene library that is constructed from DNA copies of the mRNA present

in the originating cells at the time of isolation. These DNA copies of mRNA are referred to as *copy* or *complementary DNA* (*cDNA*), and hence such a library is referred to as a *cDNA library*: We will deal with cDNA libraries later in this chapter.

Although the advent of whole genome sequencing has reduced the importance of gene libraries, there are still many important applications of the concepts involved, which justifies their continued inclusion – apart from the need to understand how we arrived at the present position.

3.1 Genomic libraries

The first step in producing a genomic library is to fragment the genomic DNA into pieces of a suitable size for cloning in an appropriate vector. It might be considered that the simplest way to do this is to digest the DNA to completion with a restriction endonuclease such as *Eco*RI. But therein lies a problem: In Chapter 2, we saw that the average fragment size generated by *Eco*RI is about 4 kb (given certain assumptions about DNA composition). But this is only an average. Even if restriction sites are randomly distributed we would expect some fragments to be very much bigger, and some would be very small. Ligation tends to work best with smaller fragments, so these would be overrepresented in the library, while some of the largest fragments may be too big to be cloned efficiently, if at all, and hence would be absent from the library. The end result of this would be that the library would not actually represent the entire genome, and hence would be of only limited use.

And that is only part of the problem. Identifying a clone carrying a specific DNA fragment is often not the end of the story. For many purposes, we are likely to want to be able to isolate the adjacent DNA as well. This would enable us to piece together all the small bits of DNA represented by individual clones so as to build up a bigger picture. If we make a library of, say *Eco*RI fragments, then we have no way of knowing how they fit together. There is no information in the library that connects one clone with another. To provide that information, we need a library of overlapping fragments. Figure 3.2 shows how these overlapping fragments enable the identification of clones on either side of the one that we originally selected. These clones can then be used to identify further overlapping clones, and so we can move along the chromosome in either direction. This is the basis of a technique known as *chromosome walking*, which we will come back to in Chapter 8.

3.1.1 Partial digests

One way to construct a library of overlapping fragments is to use partial digestion. This means using conditions, such as short digestion times, that result in only a small proportion of the available sites being cut. A similar effect

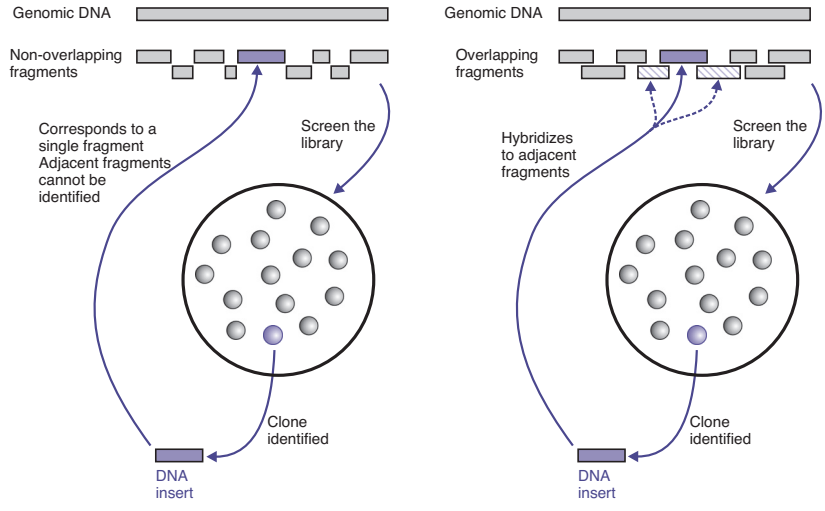


Figure 3.2 Genomic libraries: overlapping and non-overlapping fragments.

can be obtained by using very small amounts of enzyme, or by incubating the digest at a reduced temperature, or a combination of these. The digested material may then be fractionated by electrophoresis to obtain fragments of the required size range before cloning in an appropriate vector. If there is an equal probability of cutting at any site, the result will be a series of overlapping fragments (Figure 3.3), which would overcome the difficulties referred

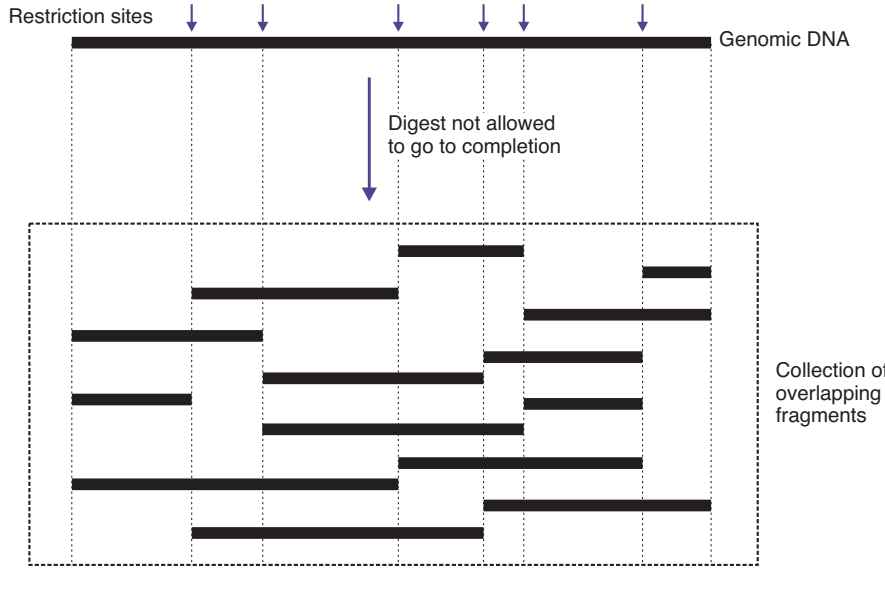


Figure 3.3 Using a partial digest to produce a collection of overlapping fragments.

to above. If we do this with an enzyme, such as *Eco*RI, that recognizes a six-base sequence, the average size of fragments in a partial digest will be too large for cloning in typical plasmid or lambda vectors, but vectors that can accommodate large inserts (such as BAC or YAC vectors; see Chapter 2) can be used for generating a genomic library of such fragments from a large genome. For smaller genomes, such as those of bacteria, we would tend to use partial digests with an enzyme that cuts more frequently, for example, one such as *Sau*3A that has a four-base recognition site (a 'four-base cutter'), and clone the products using a lambda or cosmid vector.

A four-base cutter such as *Sau*3A will produce fragments of 256 bp on average (assuming an even distribution of sites). The fact that the fragments suitable for a lambda replacement vector are 15–20 kb in length, implies that we need to cut only a small percentage of the available sites. There is a further advantage in using such a low degree of digestion: in some places there will be a number of *Sau*3A sites very close together, and if we digest too much we will get only very small fragments from this region, which will be either eliminated by size fractionation, or unsuitable for cloning in vectors such as cosmids or lambda replacement vectors, where only recombinants carrying large inserts are viable. This would result in lack of representation of that region in the library. The converse situation – lack of restriction sites in certain regions – is less likely to be an overwhelming problem in this case. There are unlikely to be any regions where neighbouring *Sau*3A sites are more than 20 kb apart, so *Sau*3A digestion should yield *some* fragments in the appropriate size range from all parts of the genome (although the representation of different parts of the genome may be far from equal).

There is a further potential problem. We made the assumption that, in a partial digest, there is an equal probability of the enzyme cutting at any site. But this is often not true. In a partial digest, some sites may be cut more efficiently than others. The nature of the adjacent sequences, and the formation of secondary structures in the DNA, may cause some sites to be cut more rapidly and/or efficiently than others. If this occurs, then our library of partial digest fragments will not be completely overlapping. When we come to try to fit the clones together we will find discontinuities in the map so that we will be unable to identify the adjacent gene (or the clone carrying it). One way of overcoming this is to use separate partial digests with a number of different enzymes.

Even with a library made in this way, we have not completely guaranteed that all our problems are solved. There is still the possibility that some regions of the genome will be over-represented and other regions will occur less frequently in the library. The only way of avoiding this, and ensuring that all parts of the chromosome are equally represented, is to abandon the strategy of using restriction enzymes altogether, and instead to use a truly random way of fragmenting the genome. One way to achieve this would be by mechanical

shearing, for example, by passing the solution rapidly and repeatedly through a small syringe needle, or by sonication (ultrasound). This is an effective way of generating relatively large fragments, and has the advantage of requiring no special equipment, but the vulnerability of the DNA to shearing diminishes as the average fragment size decreases, meaning that it can be difficult to achieve sufficiently small fragments to clone into, for example, a plasmid vector.

Although the complete randomness of mechanical shearing makes it more attractive in principle than partial digests, most workers continue to use restriction digests. Not only is it easier to control the extent of degradation, but also the restriction digests can be directly ligated with the vector. In contrast, the fragments generated by mechanical shearing have either blunt or 'ragged' ends – i.e., they may contain variable lengths of single-stranded regions at the 5′ or 3′ ends. These are not suitable for ligation, and have first to be converted to blunt ends by trimming back the unpaired ends with a single-strand-specific exonuclease. And then, since blunt end ligation is a relatively inefficient process, addition of linkers or adaptors (see Chapter 2) will be necessary in order to generate enough clones to constitute a representative library.

3.1.2 Choice of vectors

Of the various types of vectors described in Chapter 2, which should we choose for constructing our genomic library? This choice is influenced by three interlocking parameters: the size of insert that these vectors can accommodate; the size of the library necessary to obtain a reasonably complete representation of the entire genome; and the total size of the genome of the target organism.

Superficially, you might think that if we start with a genome of 4 Mb (4×10^6 bases; most bacterial genomes are of this order of magnitude) and produce a library of fragments that are 4 kb (4×10^3 bases) long, then you should be able to cover the entire genome with 1000 clones, since $(4 \times 10^6)/(4 \times 10^3) = 10^3$. But this would only be possible if (i) all the clones are different, and (ii) the clones are non-overlapping. But, as we are considering a library of random fragments, the first condition is not met. Secondly, as discussed above, we want the library to contain overlapping fragments, and thus the second condition does not hold either.

In a random collection of clones, the more clones we look at, the more likely it is that some of them will be completely or partially identical. In other words, the larger and more complete the library is, the more redundancy there will be. As we increase the size of the library it becomes less and less likely that each additional clone adds any new information. Ultimately, it is not possible to produce a library that is *guaranteed* to carry all of the genetic information from the original genome. We have to use probabilities. We

could require a 90% probability ($P = 0.9$) of having the gene that we want, or a 99% probability ($P = 0.99$). The level at which we set this probability will affect the required size of the library.

More specifically, the number of independent clones that are needed can be calculated from the formula:

$$N = \frac{\ln(1 - P)}{\ln(1 - f)}$$

where

N = required number of clones;
P = the probability of the library containing the desired piece of DNA;
f = the fraction of the genome represented by an average clone, which is calculated by dividing the average insert size by the total genome size.

Note that this calculation refers to the number of *independent* clones, i.e., the number of cells that were transformed originally (or the number of phage particles arising from a packaging reaction). This is usually determined from the number of colonies or phage plaques produced. Once you have plated out the library, and resuspended the colonies or plaques, you have *amplified* the library, and each clone is represented by thousands of individual cells or phage in the tube containing your library. You cannot increase the size (or *complexity*) of the library by plating out larger volumes. If your original library contains 1000 clones, plating it out to produce 10 000 plaques will simply mean each clone is present (on average) 10 times. You are *not* screening 10 000 independent clones.

Box 3.1 shows that with the example of a bacterial genome of, say, 4 million bases and a plasmid vector carrying inserts with an average size of 4 kb we will need a library of nearly 5000 clones to have a 99% chance of recovering any specific sequence, which is five times higher than the number we got by merely dividing 4 Mb by 4 kb. It should be noted that even if we make 5000 clones, we may not achieve completely even coverage across the genome: Some fragments of DNA may be lethal, or may be difficult to clone for other reasons, and will always be under-represented in the library.

Screening a library of 5000 plasmid clones is possible, but we can shorten the procedure by using vectors with a greater capacity for insert size. For example, we can obtain a representative bacterial genomic library with about a thousand clones using a lambda replacement vector. If we use a cosmid vector, the required size of the library is even smaller – and we can reduce it still further by using vectors with a higher cloning capacity. But there is a trade-off. One of the main purposes of producing a gene library is to be able to identify clones carrying a specific fragment of DNA, so that we can isolate and characterize that gene. The larger the insert size, the more work we have to do subsequently to find out which bit of that insert carries the

Box 3.1 Estimates of the required size of genomic libraries

Organism	Genome size	Vector type	Average insert size	*P*	Library size
Bacterium	4×10^6 bases	Plasmid	4 kb	0.99	4.6×10^3
		Lambda replacement	18 kb	0.99	1.0×10^3
		Cosmid	40 kb	0.99	458
		BAC	300 kb	0.99	59
Mammal	3×10^9 bases	Plasmid	4 kb	0.99	3.5×10^6
		Lambda replacement	18 kb	0.99	7.7×10^5
		Cosmid	40 kb	0.99	3.5×10^5
		BAC	300 kb	0.99	4.6×10^4

The values shown for the genome sizes of bacteria and mammals are examples for the purpose of this calculation. The actual genome sizes vary quite widely from one organism to another. The insert sizes for specific vectors will also vary. *P* is the probability of the library containing the desired piece of DNA.

gene we are interested in. (This can be illustrated by extending the argument to the absurd limit: the smallest gene library would be represented by a single clone carrying the entire genome, which would get us no nearer to identifying the gene that we want!) Therefore, for bacterial genomic libraries, lambda replacement vectors are usually the best compromise between a reasonable number of clones required, versus the size of the DNA fragment contained within each clone.

Note, however, that genomic libraries can be used for other purposes as well, especially for bridging any gaps that might be present in a genome sequence (see Chapter 8). For these purposes, vectors with a larger cloning capacity, such as cosmids and BACs, can be invaluable.

With larger genomes, such as those of mammals, the situation is rather different. As can be seen from Box 3.1, a library created in a lambda replacement vector would have to consist of nearly a million clones to be reasonably representative of the average mammalian genome, and would thus be far more laborious to screen: Here, the use of vectors with larger cloning capacity is a prerequisite to reduce the required size of the library to more manageable proportions.

A further factor that operates in favour of the use of larger inserts with mammalian (and other eukaryotic) genomes is that genes commonly contain introns. The overall size of the gene may therefore be too large to be contained within even a lambda replacement vector. Therefore, if we need to obtain a clone carrying the entire gene, we will have to use a vector that can accommodate a large enough DNA fragment. More commonly, we would use a cDNA library (see below) for such a purpose, as this allows us to remove the introns from the library, reducing the average insert size required considerably.

3.1.3 Construction and evaluation of a genomic library

The basis of the construction of a genomic library has already been partly covered. But by way of recapitulation, first the genomic DNA is fragmented, as randomly as possible, into suitably sized pieces for insertion into your chosen vector. Next, the vector is prepared by digestion with the appropriate enzyme, and (for a lambda replacement vector) removing the stuffer fragment. Finally, the vector is ligated with the complete mixture of genomic fragments. If you have a chosen a lambda vector, or a cosmid, you will need to mix your ligation products with packaging extracts for the assembly of infectious phage particles (see Chapter 2). If you are using a plasmid vector, you will introduce the mixture of ligated DNA into a bacterial cell by transformation or electroporation. The library will then be obtained as bacterial colonies (if using plasmid or cosmid vectors, or BACs) or phage plaques on a bacterial lawn (with lambda vectors). A library in a yeast artificial chromosome (YAC) vector would similarly be electroporated into yeast cells. For long-term storage of the library, you would make a pooled suspension of the colonies, or of the phage harvested from the plate(s).

After constructing your library, it is important to assess how good it is, including the *complexity* and *quality* of the library. Firstly you count the number of colonies or plaques (at an appropriate dilution so that you get countable colonies/plaques). If you do this using the original plates, it tells you the size (or complexity) of your library. It is important to note that if you determine the titre (the number of colonies or phage plaques) of a stored library recovered from the original plate, you will get a falsely elevated estimate of the size of the library, since this library has already been amplified. This concept of amplification of a gene library is illustrated (Figure 3.4) with a plasmid vector. The concept with a phage vector is the same, except that plaques are produced and the amplified library will consist of phage particles. The library would consist of thousands or millions of clones, rather than the few shown.

Secondly, you will want to determine the quality of the library, i.e., what proportion of the clones actually contain an insert, and how large are those inserts? Both of these questions can be answered by picking a number of

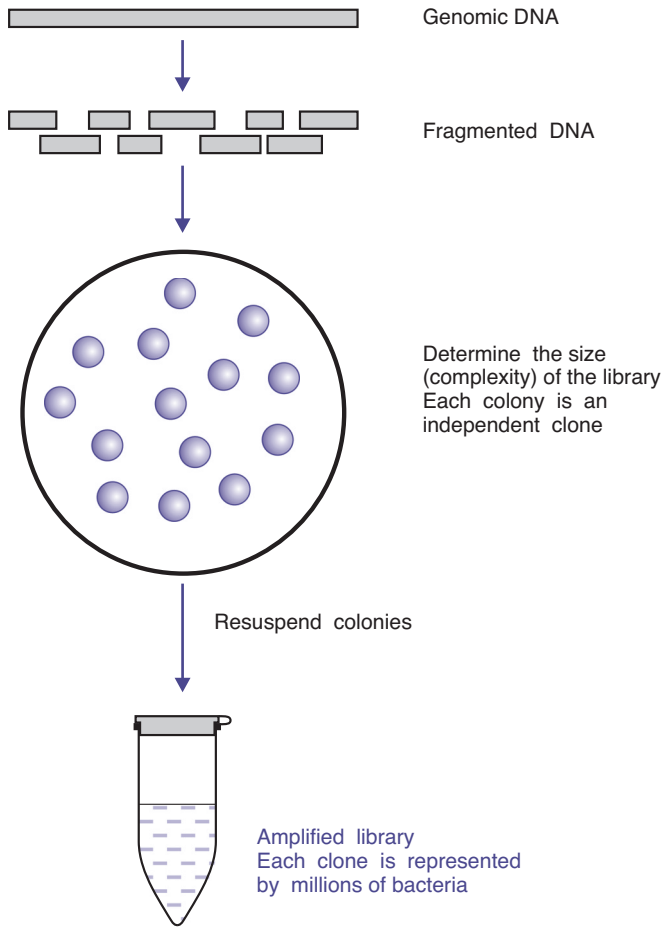


Figure 3.4 Amplification of gene libraries.

clones, growing them up individually, extracting the plasmid or phage DNA, and subjecting it to restriction digestion followed by agarose gel electrophoresis. The details will vary according to the vector and cloning strategy, but at the simplest level (insertion of restriction fragments into a plasmid) you will see that each clone has one fragment of constant size corresponding to the vector, and fragments of various sizes, which are your insert fragments (Figure 3.5). By determining the size of these fragments, you can estimate the average insert size in your library (assuming you have picked a representative sample). Those without an insert band (see track 5) are probably re-ligated vector, so you can estimate what proportion of clones do not have an insert at all. Clearly you do not want too many of these, as it will reduce the efficiency of your screening.

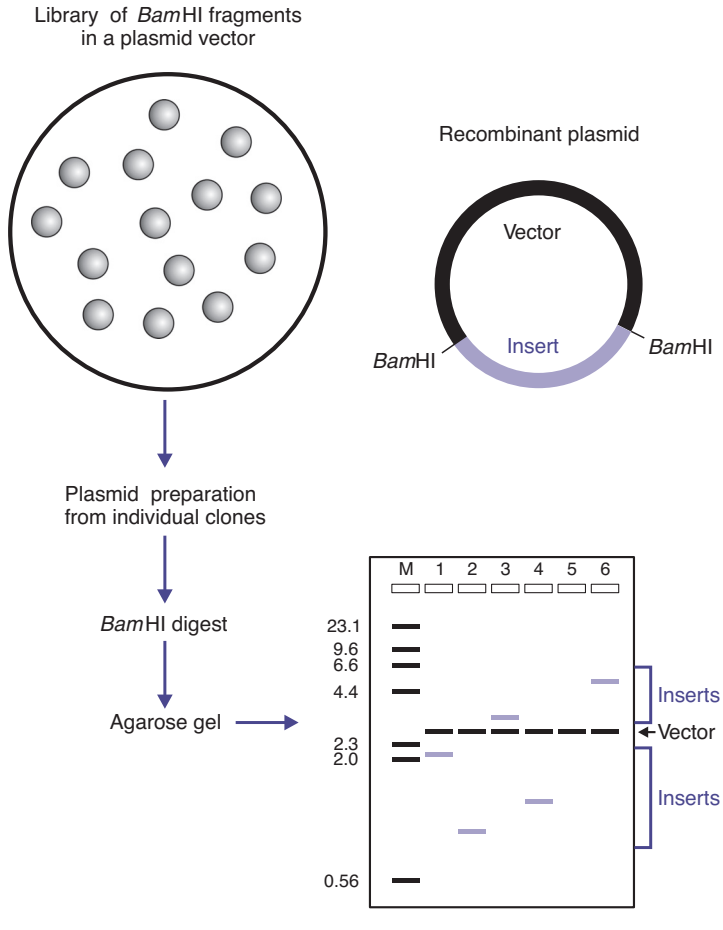


Figure 3.5 Assessing the quality of a gene library.

In Chapter 2, we described how some vectors, such as lambda gt11 or plasmid vectors of the pUC family (such as pUC18), provide you with a more direct estimate of the proportion of clones lacking an insert. With these examples, the non-recombinant vectors (without an insert) will produce blue plaques/colonies on an X-gal/IPTG-containing plate (due to the beta-galactosidase gene in the vector). If you have successfully introduced an insert fragment into the vector, the insert will (usually) disrupt the beta-galactosidase gene, giving 'white' clones. Lambda replacement vectors should not produce any clones lacking an insert (as they would be too small to be packaged) but intact vector may persist, or the stuffer fragment may not have been completely removed. Cosmids also provide positive selection for the presence of inserts, again because the re-ligated empty vector is too small for the packaging reaction.

Ideally, each clone should contain only a single insert fragment. Multiple inserts can cause problems later on, as they will provide misleading information about the relationship between different parts of the genome. With those vectors that provide positive selection for inserts (lambda replacement vectors and cosmids) it is possible to dephosphorylate the insert (using alkaline phosphatase) – rather than the vector – thus virtually eliminating the possibility of multiple inserts. If you do this with other vectors, you run the risk of obtaining an unacceptably high frequency of vector re-ligation. It may be worth running this risk, especially if you are using a vector such as pUC18 or lambda gt11, when you can tell straightaway what your insertion frequency is. If you get too many blue colonies/plaques, then you throw the library away and try again.

3.2 Growing and storing libraries

Once a library has been made, it represents a potentially useful resource for subsequent experiments, as well as for the initial purpose for which it was produced. You will therefore want to store it safely for future use. A random library will consist of a tube containing a suspension of pooled colonies from a plate (if you used a plasmid or cosmid vector), or pooled bacteriophage particles (for a phage vector). This would normally be kept at −80°C; bacterial cells in a plasmid library are protected from the adverse effects of freezing by glycerol, while phage libraries are cryoprotected by dimethyl sulphoxide (DMSO). When the library is to be screened, a small lump of the frozen stock is removed and thawed. Plasmid libraries in cells are simply spread out on agar plates containing the appropriate antibiotic, which ensures that the cells do not shed their plasmids. Phage libraries first need to be mixed with, and infect, prepared bacterial cells before being plated out. It is preferable to divide your library into aliquots before freezing it, so that each aliquot is only thawed once. This avoids the loss of clones that will accompany repeated freezing and thawing.

Following resuscitation of your library, the first task is to determine the titre of the library, which almost inevitably will have dropped since the original stock was frozen. A dilution series is produced and each dilution is spread out on an agar plate (on a lawn of bacteria, in the case of phage libraries) and grown at 37°C overnight. This allows you to calculate the titre of the library, and to determine how much is to be used for each plate in the actual screening. The number of clones required for the screen can be calculated using the formula described above. Thus, the number of clones that need to be screened is dependent on (i) the size of the fragments in the library and (ii) the size of the genome. However, note the earlier comments about library amplification. If your original library contained 1000 clones, you

cannot screen 10 000 clones by growing up more of the library and plating out a larger amount. You are merely screening the same 1000 clones ten times.

3.3 cDNA libraries

A genomic library represents the entire DNA in the genome, whether it is expressed or not. But very often it is really the genes that are being expressed that are our main target for investigation, and this is likely to be quite a small proportion of the total DNA, especially in animal or plant cells. So if we base our library on the mRNA extracted from the target cells, rather than on their DNA, we will be able to focus more closely on the real target, and make the identification of the required clones much more efficient.

The advantages of cDNA libraries, and their contrast with genomic libraries, extend further than that. First of all, since introns are removed during processing of the mRNA, the cDNA clones will reflect only the coding regions of the gene (exons) rather than the much longer sequence contained in the genome. This is especially relevant if we want to try to express the gene in a bacterial host, which would be unable to splice out the introns.

Secondly, in any organism, some or most of the DNA does not appear to code for anything. This is often referred to as *junk DNA* – but this term is misleading as many of these sequences are now known to have regulatory functions (see Chapter 8). The amount of non-coding DNA in the genome is to some extent related to genome size; bacterial genomes carry mobile elements and phages, and some other repetitive elements, but in most cases relatively little non-coding DNA compared to mammalian genomes.

Finally, a library based on mRNA, rather than a genomic library, will reflect only those genes that are actually expressed in a particular cell or tissue sample at a particular time. Any non-transcribed regions will be excluded. But it goes further than that. A cell will use only a part of its genetic capability at any one time. A bacterial cell will switch genes on or off, depending on its environment and its stage of growth. And in a multicellular organism, differentiation of cells into tissues and organs will be reflected in more or less permanent changes in the nature of the genes that are expressed. So a cDNA library from liver, for example, will be different from a kidney cDNA library from the same individual. Furthermore, some genes will be active only at certain times, such as during specific developmental stages, or different times of the day. You may also choose to make a library from a cancerous sample, or from an individual who suffers from a genetic disease. A library of this sort will reflect the nature of the cells from which the mRNA was obtained. As we will see later on, this not only reduces very substantially the number of clones needed for a representative library, but it also provides us with a variety of ways in which we can focus attention on the differences between various cells

or tissues, and thus identify genes that are selectively expressed in different environments or in different tissues.

However, we cannot clone the mRNA directly. We have to produce a complementary DNA (cDNA) copy; hence the designation of such a library as a *cDNA library*. The synthesis of the cDNA is carried out using an enzyme known as *reverse transcriptase*. (Since transcription refers to the production of RNA from a DNA template, the opposite process – RNA-directed DNA synthesis – is known as reverse transcription.) Although this is not a normal process in most cells, some types of viruses, such as leukaemia viruses and HIV, replicate in this fashion; the viral particle contains RNA that is copied into DNA after infection, using a virus-encoded enzyme. Some cellular DNA polymerases also have reverse transcription capability.

3.3.1 Isolation of mRNA

Most of the RNA in a cell is not messenger RNA, but rather ribosomal RNA (rRNA) or transfer RNA (tRNA). Any total RNA preparation will contain substantial amounts of rRNA and tRNA, and for construction of a cDNA library it is highly desirable to purify the mRNA. With eukaryotic cells, we can take advantage of the fact that mRNA carries a tail at the 3′ end – a string of A residues that is added post-transcriptionally. Such polyadenylated mRNA will anneal to synthetic oligo(dT) sequences (i.e., short polymers of deoxythymidine, or in other words short stretches of synthetic DNA containing just T residues), which can be incorporated into affinity columns for mRNA purification. Other RNA species, and non-RNA components, will not anneal and can be washed off (Figure 3.6). Although some mRNA in bacteria does have poly(A) tails, these are much shorter and only a small proportion of mRNA is polyadenylated. Therefore, this purification strategy does not

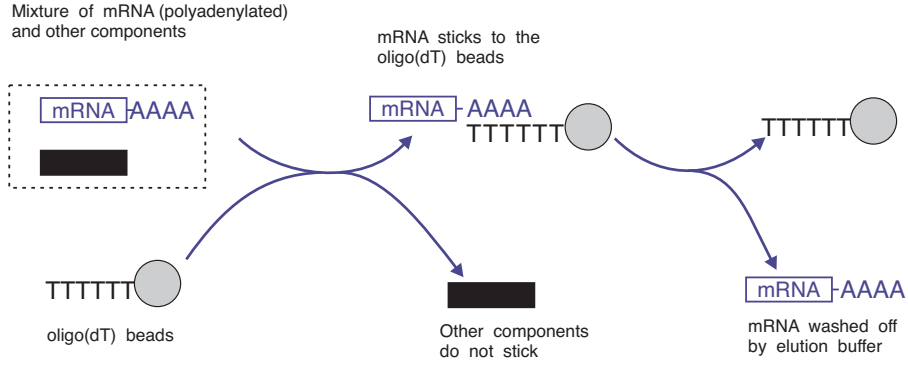


Figure 3.6 Principle of oligo(dT) purification of mRNA.

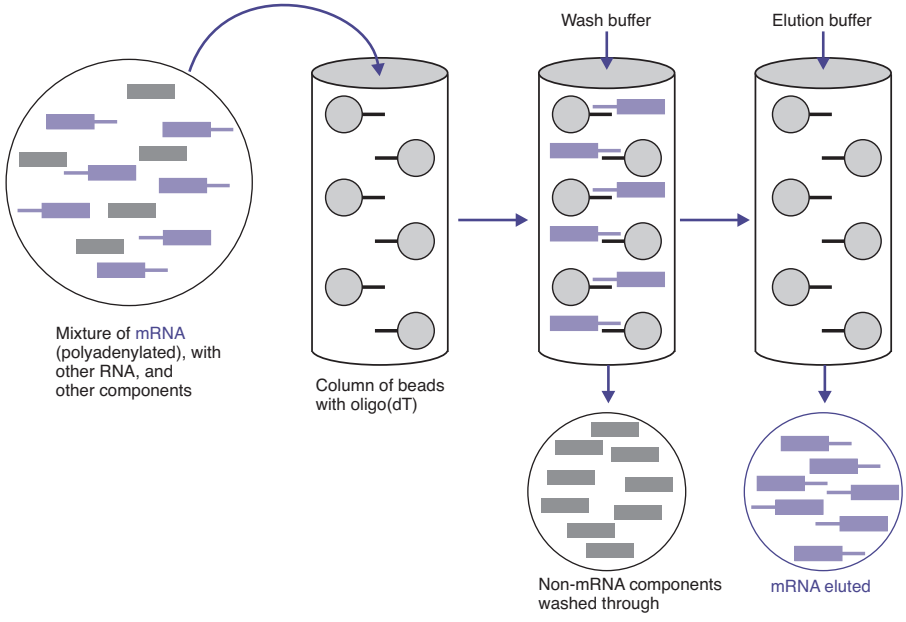


Figure 3.7 Purification of mRNA through an oligo(dT) column.

provide a reliable way of isolating bacterial mRNA, which is discussed in more detail in Section 3.3.3.

When the RNA preparation is passed through an affinity column comprising a polymer coated with synthetic oligo(dT) fragments, the poly(A) tail will anneal to the oligo(dT) residues and will be retained on the column while other RNA species will pass through. This is in effect a hybridization process, and as such the hybrids can be made unstable by lowering the salt concentration, enabling the elution of purified mRNA from the column (Figure 3.7). The eluate will be a complex mixture of all the mRNA species present in the cell at the time of extraction. The relative amounts of the different transcripts will vary substantially, which has major implications for the ease of obtaining certain cDNA clones. This is a further clear distinction from a genomic library. Although the following description is presented in terms of a single mRNA, bear in mind that we would in reality be dealing with a complex mixture.

3.3.2 cDNA synthesis

The presence of a poly(A) tail is also useful in the reverse transcription step (Figure 3.8). Reverse transcriptase, like DNA-directed DNA polymerase, requires a primer for initiation. An oligo(dT) primer will anneal to the poly(A)

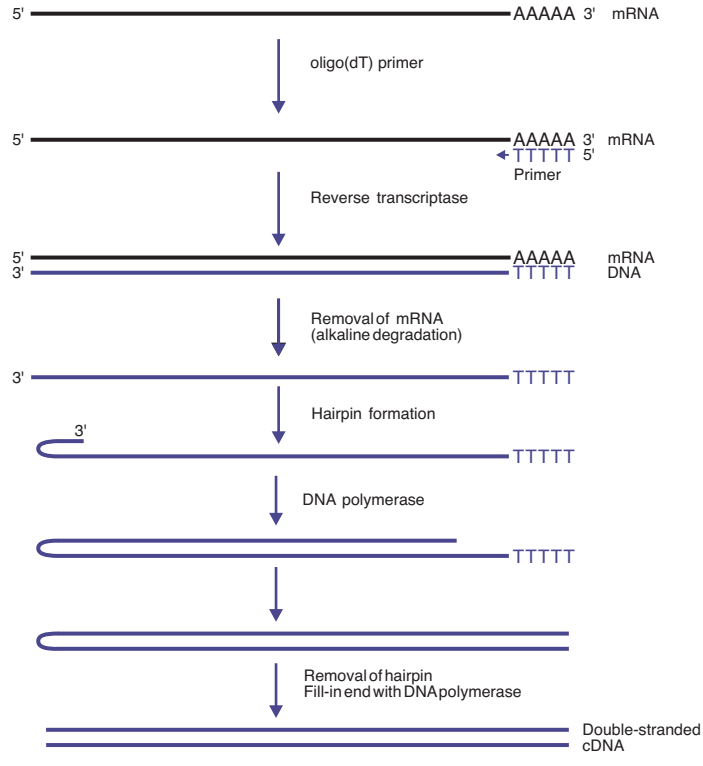


Figure 3.8 Synthesis of cDNA from mRNA.

tail; reverse transcriptase will then extend this primer, using the mRNA as the template, and will produce a single-stranded cDNA copy.

We now have a double-stranded heteroduplex molecule, with one DNA strand and one RNA strand. The next step is to replace the RNA with a DNA strand of the same sequence, producing the final cDNA molecule. One way of doing this is firstly to degrade the RNA strand by alkali treatment. This leaves the cDNA strand largely in single-stranded form. Single-stranded nucleic acid molecules tend to form secondary structures, looping back on themselves, because of the hydrophobicity of the bases. The single-stranded cDNA will therefore tend to form a hairpin loop at the 3′ end. This hairpin loop is used by DNA polymerase I to prime synthesis of the second strand. The product is a double-stranded DNA molecule, with a hairpin loop at one end. That loop is then removed by treatment with S1 nuclease (which will cut single-stranded DNA, including exposed loops).

This method suffers from several disadvantages. In particular, the requirement for hairpin formation to prime second strand synthesis, and the need to cleave the hairpin with S1 nuclease, can cause the loss of sequence at the

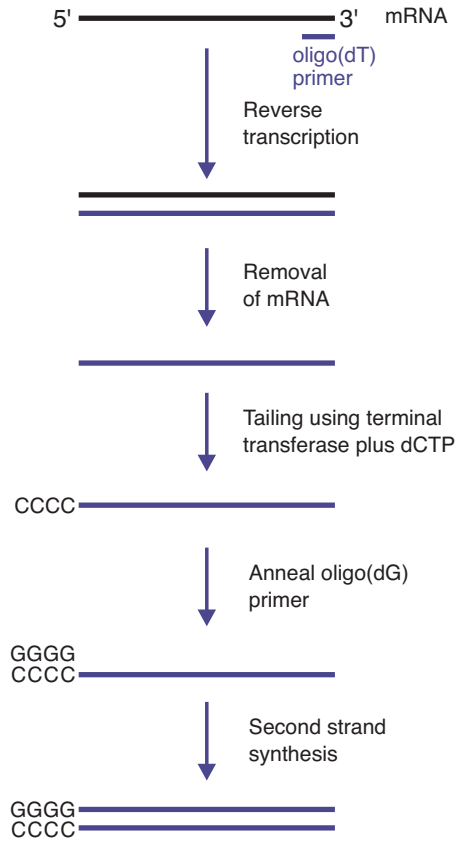


Figure 3.9 cDNA synthesis: using homopolymer tailing.

5′ ends of the mRNA. It has therefore been largely superseded by alternative strategies, such as using homopolymer tailing to add a tail to the 3′ end of the first cDNA strand (see Figure 3.9). As described in Chapter 2, terminal transferase, if provided with, for example, dCTP, will add a string of C residues to the 3′ ends of DNA molecules. This enables the use of an oligo(dG) primer to initiate second strand synthesis, without requiring hairpin formation and cleavage.

Alternatively, degradation of the RNA strand by RNaseH (rather than alkali treatment) will leave small RNA fragments, which act as primers for second strand synthesis. This also avoids the need for cutting the hairpin with S1 nuclease.

For cloning the cDNA, adaptors (see Chapter 2) are added, by blunt-end ligation, to make the cDNA molecules compatible with the chosen vector (Figure 3.10). After size-fractionation, eliminating excess adaptors and small, abortive cDNA fragments, the library is inserted into the vector in a second

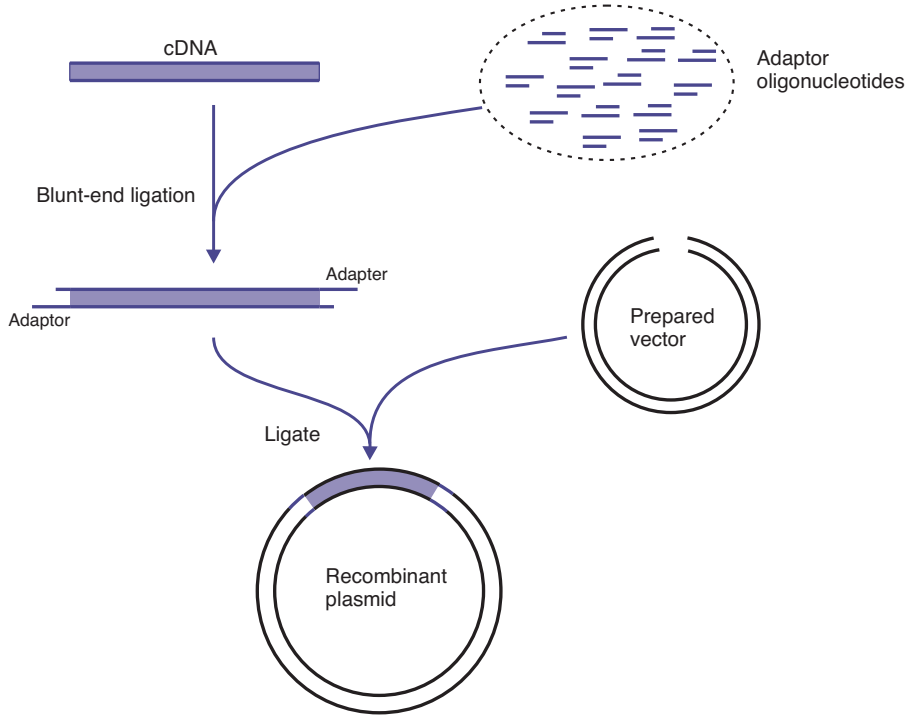


Figure 3.10 Cloning cDNA.

ligation. The size of cDNA molecules makes it unnecessary to consider vectors with a large cloning capacity. The choice is essentially between a plasmid vector, or a phage lambda insertion vector such as gt10 or gt11 (see Chapter 2).

The library that we have produced not only reflects those mRNAs expressed within the target cell at the time of extraction, but also their relative abundance. If a particular mRNA is present in high number (i.e., it is highly expressed), then it will be more often incorporated into the library clones, meaning that it will be more frequent than a lowly expressed mRNA in the final cDNA library. This is important because to determine the number of clones in a cDNA library that you need to screen, you have to take account of the abundance of the relevant mRNA. If you are searching for a cDNA clone for a reasonably abundant mRNA, you might have to screen 20 000 clones, while for a very rare transcript, you might need 200 000 clones.

One limitation of these procedures is that you may not get a full-length cDNA in any individual clone. Constraints such as elements of secondary structure in the mRNA may interfere with reverse transcription, so that the enzyme rarely, if ever, reaches the end of the mRNA. As a result, the regions at the 5′ end of the mRNA may be under-represented in the cDNA library. This can be partially addressed by using random primers rather than

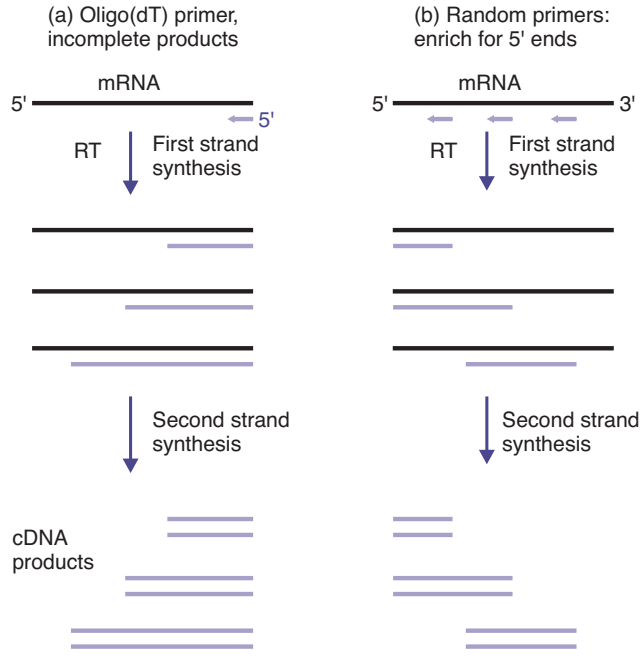


Figure 3.11 cDNA synthesis: enhancing representation of 5′ mRNA ends. RT, reverse transcriptase.

oligo(dT) primers. These random primers will initiate first strand cDNA synthesis at intermediate points, and hence the enzyme will be more likely to reach the 5′ end of the mRNA (see Figure 3.11). You are unlikely to get any clones containing full-length cDNA, but these clones containing the 5′ end can be compared to other clones carrying the missing 3′ portion, making it possible to devise strategies for obtaining full-length molecules. The use of random primers rather than oligo(dT) primers also overcomes the problem that some RNA molecules (e.g., bacterial mRNA and genomic RNA from viruses) are not polyadenylated.

Alternatively, if you can predict the sequence of the ends of the mRNA, for example, from genome sequence data, you can simply use a pair of specific primers for RT-PCR (reverse transcription PCR) amplification (see Chapter 4), which can generate your specific, full-length cDNA directly from even small amounts of mRNA in your starting material. This bypasses completely the need for production and screening of a cDNA library.

3.3.3 Bacterial cDNA

The arguments in favour of cDNA, rather than genomic, libraries carry much less force with bacterial targets. The smaller size of bacterial genomes, and the (virtual) absence of introns, means that a genomic library is usually quite

adequate – and a lot easier to construct. There are also technical difficulties in producing cDNA with bacteria. Not only is the mRNA not consistently polyadenylated, but also it is remarkably unstable: many bacterial mRNA species have a half-life (*in vivo*) of only a few minutes. Furthermore, the organisation of bacterial genes into polycistronic operons (groups of genes that are transcribed into a single long mRNA) means that a bacterial mRNA can be as much as 10–20 kb in length. Not only is it difficult to isolate this mRNA intact, but it would be very difficult to produce a full-length cDNA copy from it.

As a consequence, bacterial cDNA libraries are rarely produced. But for some purposes, such as the analysis of gene expression (see Chapter 6), production of bacterial cDNA can play an important role, for example, in identifying those transcripts that are relatively abundant in the bacterial cells under selected conditions.

3.4 Screening libraries with gene probes

Once we have constructed our library, we still have to find a way of identifying which clone(s) carry the gene/cDNA that we are interested in. This means that we need ways of rapidly screening very large numbers of clones. This is most commonly done using a nucleic acid *probe* (DNA or RNA), which will hybridize to the DNA sequence we are looking for in a specific clone. The principle involved is that the library (in the form of bacterial colonies or phage plaques) is replicated onto a filter, which is then treated to release the DNA and bind it to the filter. The filter then carries a pattern of DNA spots, replicating the position of the colonies or plaques on the original, and can be hybridized with the probe, which has first been labelled so that it can be easily detected. This allows you to detect which DNA spots hybridize to the probe, and recover the corresponding clones from the original plate.

In order to appreciate the power of this technique it is necessary to consider more closely the question of hybridization. Later in this chapter we will consider an alternative strategy, which involves using antibodies to screen an expression library.

3.4.1 Hybridization

The fundamental basis of hybridization was described in Chapter 1. It is based on the difference in stability between the covalent bonds in the nucleic acid backbone of each strand, and the much weaker hydrogen bonds that bind the two strands in the double helix together by base pairing. This means that the two strands can be safely, and reversibly, separated by conditions such as heat or high pH without endangering the covalent bonds in the backbone. This is referred to as *denaturation* of DNA. Lowering the temperature, or

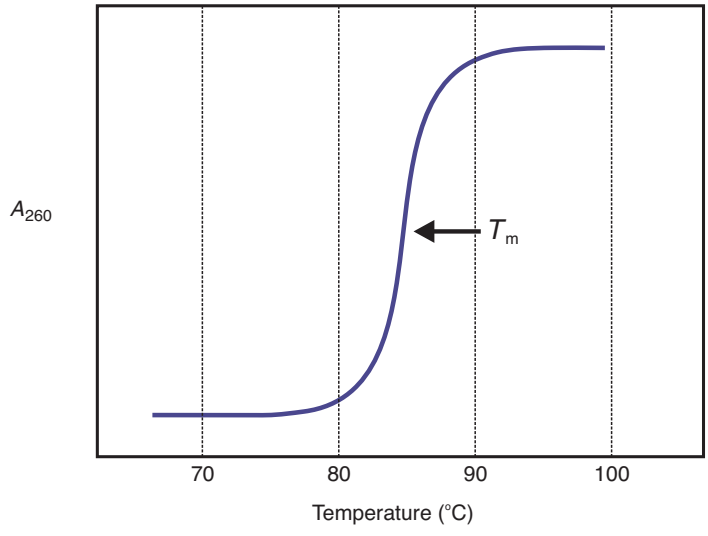


Figure 3.12 Melting (denaturation) of DNA.

reducing the pH, will allow the strands to join together again and *renature*. The separation of the strands during denaturation causes a radical change in the physical properties of DNA, such as optical density (Figure 3.12), which changes dramatically over a short temperature range, and then stabilizes after the strands have separated entirely. The midpoint of this temperature range is denoted the *melting temperature* (T_m). Under physiological conditions, the T_m of DNA is usually in the range 85–95°C (depending on the base composition of the DNA). In the laboratory, we can adjust other factors, such as the salt concentration (see below) to bring the melting temperature down to a more convenient level.

The T_m varies according to the base composition of the DNA because guanine–cytosine base pairs are joined together by three hydrogen bonds, whereas adenine–thymine base pairs have only two. If we take reasonably large DNA molecules, such as might be obtained by isolating total DNA from an organism, we can make an estimate of its base composition by measuring the T_m. Or, if we know the base composition, we can calculate the T_m. For shorter sequences, such as the 20–30-base synthetic oligonucleotides that are commonly used as *primers*, other factors have to be taken into account. The strength of the association between two bases (expressed as ΔG, the energy released on formation of a base pair) depends also on the adjacent bases, because hydrophobic interactions between adjacent bases (*stacking*) also affect the stability of the pairing. Some examples are shown in Figure 3.13, with the negative values for ΔG indicating that energy is released on formation of a base-paired structure. More energy released means greater stability. It can be

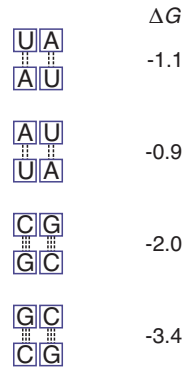


Figure 3.13 Strength of association of base pairs.

seen that the free energy of base pairing (kcal/mol) for the CG/GC doublet is not the same as for the GC/CG doublet. For short oligonucleotides, calculation of the T_m therefore has to take account of the context of each base in the sequence, although computer programs are readily available to do this for you, and hence you do not need to overly concern yourself. This is not likely to be such an important factor in determining the T_m of a longer stretch of DNA because there will be a tendency for these differences to average out.

Although the normal base pairs (A-T and G-C) are the only forms that are fully compatible with the canonical Watson–Crick double helix, pairing of other bases can occur, especially in situations where a regular double helix is less important (such as the folding of single-stranded nucleic acids into secondary structures – see below).

As described in Chapter 1, hydrophobic interactions between the bases on opposite strands are important in maintaining the stability of the double-stranded DNA structure (this is in addition to the hydrophobic stacking of adjacent bases on the same strand, as referred to above). The specificity of the interaction can therefore be increased by the use of chemicals (such as formamide) that reduce the hydrophobic interactions.

In addition, the strong negative charge on the DNA strands causes electrostatic repulsion between the two strands. This effect is counteracted by a cloud of counterions surrounding the molecule. However, if you reduce the salt concentration, any weak interactions between the strands will be disrupted by electrostatic repulsion – so, if we use low salt conditions, we can increase the specificity of a hybridization reaction (see below).

If two similar, but different, double-stranded DNA fragments are mixed, melted, and then left to renature, some of them will anneal with their perfectly complementary halves, but others will form hybrid, imperfectly matched, pairs (Figure 3.14). This is known as a *heteroduplex*. This would happen, for example, if you were to mix the cDNA molecules encoding

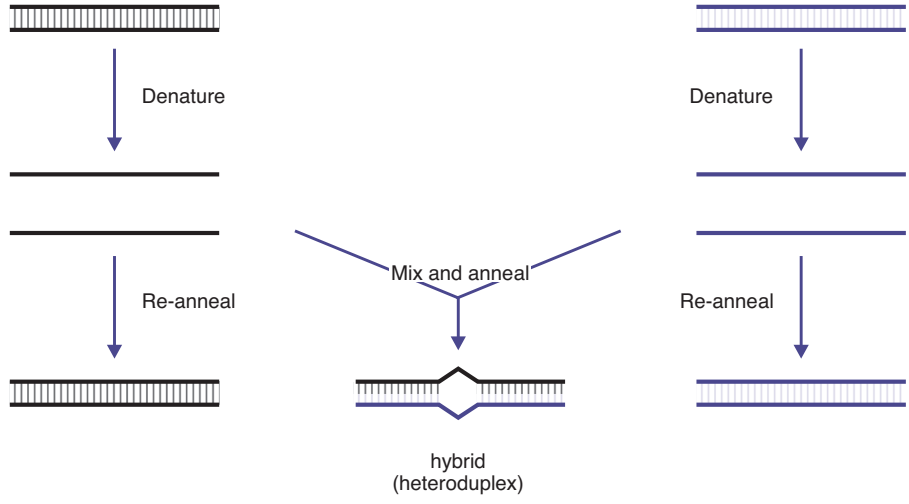


Figure 3.14 Formation of hybrid DNA between similar but non-identical DNA molecules.

the human red- and green-sensitive photopigments, or the actin genes from mouse and rat. Any base pairs that do not match will cause imperfections in the resulting heteroduplex. A higher number of mismatches will lead to a less stable hybrid – in other words one that would have a lower melting temperature.

This is very important when single strands of nucleic acids are hybridized in the laboratory. The investigator can choose conditions that would be more or less forgiving of partial mismatches. This is referred to as varying the *stringency* of the hybridization. The most important and obvious way to increase the stringency is to raise the temperature. Secondly, hybrid DNAs are more stable at higher salt concentrations. At low salt concentrations, electrostatic repulsion between the two strands means they will only stay double-stranded if the two strands are well matched (high stringency). At higher salt concentrations, the presence of positive counterions will relieve this repulsion. In the laboratory, salt concentration is usually denoted as multiples of SSC (Standard Saline Citrate; 1 × SSC is defined as 0.15 M sodium chloride and 0.015 M sodium citrate). Sometimes, formamide is added to the hybridization solution, which lowers the melting temperature and is therefore used in situations where hybridization temperatures need to be kept low, such as when carrying out *in situ* hybridization (see Chapter 6).

The basic types of hybridization that are used in the laboratory are filter hybridization, solution hybridization and *in situ* hybridization. In general, and in this chapter in particular, a specific nucleic acid fragment (DNA or RNA) is labelled and used as a *probe* to detect a corresponding DNA strand, the *target*, in a complex mixture such as a gene library. However, we will also

come across situations where *reverse hybridization* is used. For example, in the use of arrays (see Chapters 9 and 10), a series of specific DNA fragments is arranged on a filter or a slide, which is hybridized with a labelled complex mixture such as fragments of total genomic DNA. In this situation, we will refer to the labelled complex mixture as the probe, rather than the specific fragments, but the terminology is not always consistent.

3.4.2 Labelling probes

A fundamental feature of nucleic acid hybridization is that the probe is labelled in a way that will make it possible to detect it, and thereby the target it has bound to, after the hybridization. The classic labelling method is to incorporate a radioactive isotope into the probe molecule. Isotopes that are used for this include ^{32}P, ^{33}P, ^{35}S and ^{3}H. A more energetic isotope, such as ^{32}P, gives a stronger signal but less good resolution. This makes it useful for experiments where submillimetre resolution is irrelevant, such as Southern blot hybridization (see below). At the other extreme, ^{3}H gives a weaker signal and thus requires a much longer exposure time to be detectable. However, it provides excellent resolution, and can be used in experiments such as *in situ* hybridization (see Chapter 6), where it is important to be able to assign the probe labelling not only to specific cells, but even to specific regions of chromosomes.

The classic way to detect radiolabelled probes is to place the probe-target hybrid on an X-ray film. Radioactive particles will expose the region of the film with which they are in contact just like X-rays or visible light would. A more modern method for detecting binding of radioisotopes is phosphoimaging. This method requires specialized and expensive apparatus, but is faster, and the phosphoimaging plates can be reused, in contrast to X-ray films.

Most investigators have abandoned radioactive labelling methods in favour of non-radioactive ones. These have advantages in terms of worker safety, detection speed and cost. They are also unaffected by the continuous decay of radioisotopes, and therefore labelled probes are more stable, meaning that they can be kept and reused. There are a variety of non-radioactive labels that can be used. Nucleotides substituted with biotin or digoxigenin can be incorporated into the probe, and then detected with specific antibodies (or, in the case of biotin, using avidin, which binds very strongly and specifically to biotin). The antibody (or avidin) that is used is itself labelled with an enzyme, such as horseradish peroxidase (HRP) or alkaline phosphatase. These enzymes can be detected using a chromogenic substrate (i.e., a substrate that yields a coloured product when reacted with the enzyme) or a chemiluminescent substrate (where the initial reaction product is unstable and light is emitted as it breaks down). In the latter case, the emitted light will darken

X-ray film just like a radioisotope would. As with radioactive methods, specialized equipment can be used instead of X-ray film.

Alternatively, the probe can be labelled directly with HRP, or a fluorescent label can be incorporated, which allows direct detection of the labelled probe. Such probes are especially useful for *fluorescent in situ hybridization*, or *FISH* (see Chapter 6).

3.4.3 Steps in a hybridization experiment

Nucleic acid probes have a tendency to bind non-specifically to other materials on the filter, or even to the filter itself. To minimize this non-specific probe binding, the hybridization solution contains various blocking agents, which may include detergents, bovine serum albumin and non-homologous DNA. It is often advantageous to pretreat the filter with the hybridization solution without added probe (*prehybridization*), as this can help reduce non-specific probe binding.

If the probe is double-stranded DNA, it will need to be boiled prior to being used in the hybridization to separate the strands, so that they can bind to the target. After adding the probe to the target, the mixture is incubated in a controlled-temperature chamber overnight. The non-specifically bound probe is then removed by washing. This is the step where you can most conveniently decide how tolerant the experiment should be of partially mismatched probe–target hybrids, by choosing an appropriate combination of temperature and salt concentration. If, for example, the intention is to hybridize a cDNA probe with genomic DNA from the same species, then high stringency conditions (i.e., high temperature and low salt concentration) should be chosen (Figure 3.15). This will ensure that the probe will remain bound to the membrane only where it has annealed to the correct complementary sequences. Where the probe has annealed to DNA that is partially similar (e.g., members of the same gene family), the annealing will be disrupted by these conditions, and the probe will be washed off the filter. If, on the other hand, you want to detect, say, a chicken gene with a probe made from human DNA (a heterologous probe), then low stringency (i.e., low temperature and high salt concentration) should be chosen in order to protect the expected partially mismatched hybrids.

The hybridization of the filter with an appropriate probe is the first example of *filter hybridization* that we will come across in this book. In filter hybridization, as the name implies, the target is immobilized on a filter and the probe is free in the hybridization solution until it binds to the target. Other important applications of filter hybridization include *Southern blots* (see later in this chapter) and *Northern blots* (Chapter 6). In both cases, the target nucleic acids are first size-separated in an electrophoresis gel before being transferred onto the membrane. The difference between these techniques is that

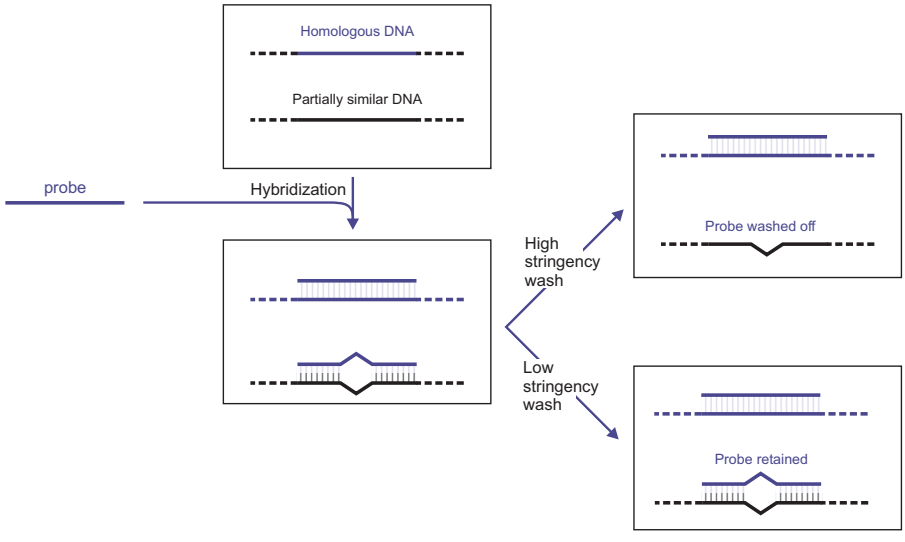


Figure 3.15 High and low stringency washing.

in Southern blots the target nucleic acids are fragments of DNA, whereas Northern blots are used for the identification of RNA. The subsequent hybridization of these filters follows the same principles as outlined above.

3.4.4 Screening procedure

We can now return to the question of screening a gene library, and how you can use hybridization with a gene probe to identify a specific clone. Remember that at this stage your gene library is in the form of a number of colonies (or phage plaques) on an agar plate (or, more probably, a number of such plates). The first step is to produce a replica filter of each plate, using a procedure known as a *colony lift* (or a *plaque lift*). A nitrocellulose or nylon membrane is placed on top of the plate, for a minute or so. During this time, a part of each bacterial colony, or some of the phages from each plaque, will bind to the membrane. The membrane is then removed and soaked in a sodium hydroxide solution, which releases the DNA from the cells or phage, and denatures it. After neutralization with a buffer solution, the single-stranded DNA molecules are fixed to the membrane by heat or UV irradiation. The membrane filter is then hybridized with the appropriate probe. Following hybridization, the filter is washed, under the chosen stringency conditions, before detection of the probe, using procedures appropriate for the nature of the label.

If you are successful, you will now have an X-ray film with one or more black spots on it that mark the position(s) of the desired clone(s) on the

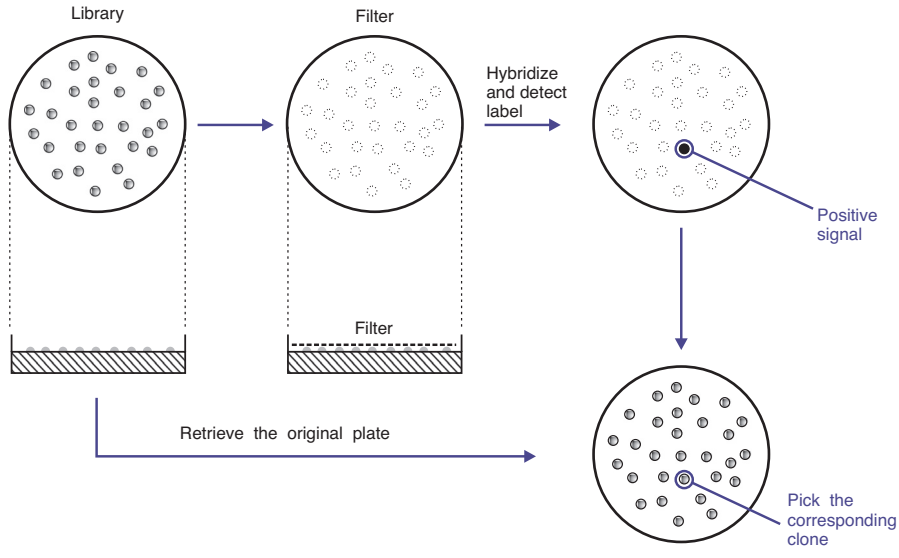


Figure 3.16 Screening a gene library by hybridization with a gene probe.

original plate. You then return to that plate, align it with the X-ray film image, pick that colony and subculture it. This is the clone that you want. This procedure is summarized in Figure 3.16.

In practice, especially if you are screening a library at high density, you are unlikely to be able to pick an individual clone without also collecting some of the neighbouring ones. You will then have to use this mixture of clones for rescreening at a lower density. Furthermore, some of the clones will be false positives (for various reasons), and so it will be necessary to submit them to further testing to verify their identity.

3.4.5 Probe selection and generation

Screening a gene library with a nucleic acid probe implies that you have access to a suitable DNA (or RNA) fragment. Yet the whole purpose of screening a gene library is to isolate a clone carrying a novel piece of DNA. There are a variety of ways around this circular argument. In this section we will describe the use of homologous and heterologous probes, and probes generated by *back translation*. Chapter 4 will describe another approach, which is the use of PCR-based techniques for generating suitable probes.

Homologous probes Using a probe coming from the same gene in the same species is not as strange an idea as it may seem. Firstly, you may already have a clone, but an incomplete one. For example, you may not have the

complete cDNA sequence, or may wish to identify adjacent sequences, such as promoter elements. Alternatively, you may have access to a cDNA clone, but desire to know the genomic sequence that encodes it. Finally, your business may be to study genetic variation between individuals or strains, such as polymorphisms or mutations that cause disease. Under these situations, it is possible to use a homologous probe, although this would usually be generated by PCR (see Chapter 4) or synthetic oligonucleotides would be used, rather than using cloned DNA.

Heterologous probes If the gene we are trying to clone is not completely unknown, and a related one has already been cloned and characterized from another source, then we can use that clone as a *heterologous* probe. So in isolating the human insulin gene, for example, it was possible to use the previously characterized rat insulin gene to probe a human gene library. The availability of genome sequences (see Chapter 8) has made this approach much easier. You can now browse through these sequences to find genes that may be related to the one you are attempting to clone, and use the sequence data to design appropriate probes.

This takes advantage of the fact that genes with the same function are often very similar in different organisms. (We will see examples of this in Chapter 9.) As we would expect, the similarity tends to be greatest for the most closely related species (you might not like to think it but we are closely related to rats in this sense!). The extent of the relationship varies from one gene to another; some genes are very highly conserved while others are more variable. Generally, the closer the relationship between the two species, the more reliable the screening is likely to be.

If we move towards less closely related species, then the similarity is likely to decline to a level at which we can no longer detect hybridization using high stringency conditions. We would therefore need to use low stringency (e.g., lower temperature or higher ionic strength). However, there is a price to be paid. As you lower the stringency in order to allow your heterologous probe to hybridize, so you also allow the probe to hybridize to other genes that are similar to the gene you are trying to clone. In other words, you are more likely to get false positive signals, which will require additional checks to be made.

It is worth noting that the source of your heterologous probe does not necessarily have to be another species. It is possible to use a probe from one member of a gene family to detect closely related family members within the same species. Using this approach, for example, it was possible to identify novel members of the nuclear receptor superfamily by probing with a heterologous probe designed against an already known member of this superfamily. An obvious disadvantage of this approach is that you will also detect the original sequence to which the probe was designed against, but this is a relatively small price to pay for being able to identify new, related sequences.

Back translation An alternative strategy becomes possible if you are able to determine part of the amino acid sequence of the protein of interest. Technological advances, especially using mass spectrometry (see Chapter 10), have made this a much more attractive option, since a partial amino acid sequence can be obtained from a protein spot or band on a gel. This information can then be used to infer the likely sequence of the gene itself. Since this process is the reverse of normal translation it is referred to as *back translation* or *reverse translation*. Once you have done that, you can then synthesize an appropriate nucleic acid probe.

However, you have to remember the redundancy of the genetic code. If you know the DNA or RNA sequence you can predict accurately the amino acid sequence of the protein (subject to a few peripheral assumptions). But in the reverse direction it is a different matter. If you know there is, say, a leucine residue in the protein at a specific position, then you have considerable uncertainty over what the DNA sequence actually is. It could be any one of six codons (see Box 3.2).

You can easily accommodate ambiguity in the sequence by programming the DNA synthesizer to include a mixture of bases at that position. But every such ambiguity reduces the specificity of the probe. If you allow too much ambiguity you will have a probe that will react with an unacceptable number of non-specific clones in the library. One way of reducing the ambiguity of the probe is to take account of the codon usage of the organism. If this shows that there is a very marked preference for one codon over another synonymous codon, then you can use that information in selecting the codon to be included in your probe. Alternatively, you can reduce the required ambiguity by careful selection of the region of the amino acid sequence to be included, so as to avoid amino acids such as leucine, arginine or serine (with six possible codons) in favour of those amino acids with unique codons (methionine, tryptophan) or with only two possible codons (e.g., tyrosine, histidine).

3.5 Screening expression libraries with antibodies

Libraries made with an expression vector such as lambda gt11 (see Chapter 2) allow an alternative method of screening, using antibodies (see Box 3.3). For this, the library is cultured under conditions that permit protein expression from each clone to occur. This means, in this case, growing the cultures at 42°C to inactivate the temperature-sensitive phage repressor, and in the presence of IPTG in order to induce expression from the *lacZ* promoter. Under such conditions, each clone will produce the protein encoded by the insert DNA fragment. The plates are then overlaid with a filter in the same way as with nucleic acid probe screening, with the exception that the filters

Box 3.2 Genetic code: possible codons for each amino acid

Amino acid	Possible codons						Number of codons
Alanine	GCU	GCC	GCA	GCG			4
Arginine	CGU	CGC	CGA	CGG	AGA	AGG	6
Asparagine	AAU	AAC					2
Aspartate	GAU	GAC					2
Cysteine	UGU	UGC					2
Glutamate	GAA	GAG					2
Glutamine	CAA	CAG					2
Glycine	GGU	GGC	GGA	GGG			4
Histidine	CAU	CAC					2
Isoleucine	AUU	AUC	AUA				3
Leucine	UUA	UUG	CUU	CUC	CUA	CUG	6
Lysine	AAA	AAG					2
Methionine	AUG						1
Phenylalanine	UUU	UUC					2
Proline	CCU	CCC	CCA	CCG			4
Serine	UCU	UCC	UCA	UCG	AGU	AGC	6
Threonine	ACU	ACC	ACA	ACG			4
Tryptophan	UGG						1
Tyrosine	UAU	UAC					2
Valine	GUU	GUC	GUA	GUG			4
Stop	UAA	UAG	UGA				3

are not treated with sodium hydroxide. Next, instead of a gene probe, the filters are incubated with a diluted antibody to the protein in question, which will identify which clone contains the DNA sequence that encodes for the target protein. After incubating the filter with the antibody, excess antibody is washed off, and the filter incubated with a labelled *secondary antibody* that

will bind to the first one, such as anti-rabbit immunoglobulin, allowing detection of the clones that reacted with the primary antibody (see Box 3.3). This antibody can be produced by purifying the protein in question and injecting it into an animal such as a rabbit. Alternatively, a synthetic peptide can be produced corresponding to a known or assumed part of the protein, and used similarly for immunization.

Box 3.3 Antibodies

Antibodies are made by animals in response to antigens – proteins or peptides, for our purposes. These will recognize different parts of the protein (*epitopes*). Some bind to short amino acid sequences (*linear* epitopes), irrespective of conformation. If you immunize with a short peptide, you will *only* get antibodies that recognize linear epitopes. Other antibodies bind to *conformational* epitopes, formed by folding the amino acid chain. Such antibodies only react with native, not denatured, protein. As many procedures use denatured protein, it is important that the antibody reacts with a linear epitope. Furthermore, if the immunizing protein is glycosylated (or has other post-translational modifications), some of the antibodies will recognize modified epitopes, and will be no use for screening gene libraries. However, a typical antiserum will contain a mixture of antibodies, and some of these will recognize linear, non-modified epitopes.

Antibodies vary in specificity – some will react with proteins other than your target. The antiserum may also contain antibodies to other antigens to which the animal has been exposed. You have to test the specificity of the antiserum – and sometimes purify it by absorbing out cross-reacting antibodies.

Alternatively, isolation and culture of individual antibody-producing cells will yield a single antibody species rather than the mixture you get in an antiserum. Antibody-producing cells cannot be maintained in culture, but they can be fused with other cells to form a *hybridoma,* which can be grown and used to produce a single antibody. This is a *monoclonal antibody*, made by a single clone of antibody-producing cells. Although monoclonal antibodies are important where specificity is needed (e.g., in diagnosis), they are not *inherently* specific. Their specificity arises through selecting clones that make antibodies that *are* specific. For us, this specificity can be a disadvantage. If the monoclonal antibody recognizes a glycosylated or conformational epitope, it will be useless for many of our applications.

The other advantage of monoclonal antibodies is their constancy. With conventional antisera, once you have used up your supply you have to immunize another animal. The antiserum will differ in both titre and

specificity. (Larger animals, such as goats, can supply antibodies for years – but they are expensive.) But a monoclonal antibody is produced from a permanent cell line, so you can obtain continuous supplies of identical antibody.

The antibody needs to be labelled, for example, with an enzyme, such as horseradish peroxidase, which can be detected using chromogenic or chemiluminescent substrates. It is often more convenient to use a second antibody. For example, mouse anti-rabbit IgG is an antibody raised in a mouse that reacts with rabbit antibodies. A wide range of labelled second antibodies can be bought off the shelf.

It is also possible to use lambda gt11 libraries to identify antigens that have not been characterized, for example, by using antisera from experimentally infected animals, or from human subjects recovering from an infectious disease. This has proved to be a very powerful way of identifying those antigens that are especially important in the natural course of infection, or for protection against infection.

It is important to note the limitations of using such antibody screening methodologies. First, screening a gene library with antibodies obviously needs expression of the cloned gene – limiting vector choice somewhat. Second, any protein that is expressed in *E. coli* may not fold into its correct, natural conformation – especially if you are using a vector like lambda gt11 that generates a fusion protein. You therefore need an antibody that binds to a *linear epitope* (see Box 3.3). Third, it must recognize a non-glycosylated epitope, since the protein you are trying to detect will not be correctly glycosylated in *E. coli*.

An antiserum derived from immunizing an animal with a purified protein will contain a variety of different antibodies, some of which will recognize linear, non-modified epitopes. The problem with conformational or modified (e.g., glycosylated) epitopes is most likely to occur with monoclonal antibodies (see Box 3.3), where it is necessary to confirm the nature of the recognized epitope before using it for screening libraries. Similar considerations arise when using antibodies for Western blotting (see Chapter 6).

3.6 Characterization of plasmid clones

Whether your clone is produced directly or by subcloning from a larger fragment derived from a gene library, at some stage you are certain to want to characterize the plasmid, and the fragment that it contains. The most basic procedure is to run total DNA extracts or crude plasmid preparations on an agarose gel and look for a change in the size of the plasmid compared to the original vector. If the insert is a reasonable size compared to the vector,

you will be able to distinguish recombinant plasmids (those carrying an insert) from the parental vector. However, you need to use a supercoiled size marker rather than a conventional one, as the supercoiling of intact plasmids has a strong effect on the rate of migration through the gel.

More accurate and more reliable estimates of the size of the insert can be obtained by digesting the plasmid with an appropriate restriction enzyme to generate linear fragments that can be accurately sized on an agarose gel. An obvious candidate is the same enzyme that was involved in the cloning step. For example, if you have inserted an *Eco*RI fragment into a unique *Eco*RI site on the cloning vector, then digestion with *Eco*RI will yield two fragments: one corresponding to the linearized vector, and the second being the inserted fragment.

3.6.1 Southern blots

Further confirmation of the nature of your insert comes, once again, from hybridization. You cannot hybridize a probe to DNA fragments while they are in an agarose gel. You need to transfer them to a filter first. The technique for doing this is Southern blotting, named after E.M. Southern who developed it. The basis of the method is illustrated in Figure 3.17. A membrane (nitrocellulose or, more commonly, nylon-based) is placed on the gel and a stack of dry paper towels on top of the filter. Buffer is drawn up through the gel and filter by capillary action, carrying the DNA fragments with it. These fragments

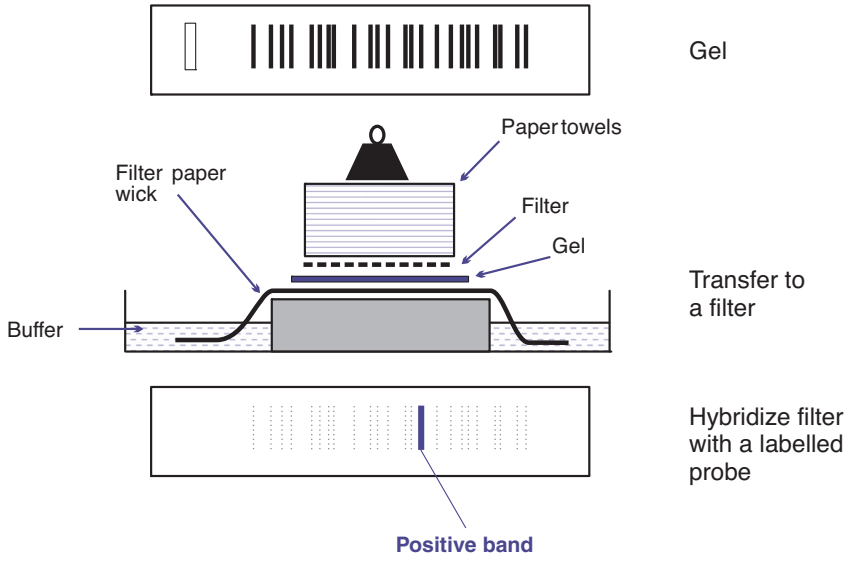


Figure 3.17 Southern blotting.

are trapped on the membrane, which thus acquires a pattern of DNA bands that corresponds to the position of those fragments in the agarose gel. The arrangement shown in the figure looks very crude, and there are much more elegant pieces of equipment available; nevertheless, many molecular biologists use their own home-made apparatus. Following transfer of the DNA to the filter, the relevant DNA fragment can be identified by hybridization in the ways already described.

Southern blotting is a very useful way of verifying that the band you can see on a gel is indeed the insert that you have been seeking. But that is only the beginning of the applications of Southern blotting. We will come across it again later on, particularly in Chapter 9, where it is invaluable for detecting specific gene fragments in a digest of total chromosomal DNA (which just looks like a smear if stained with ethidium bromide), and for comparing the banding patterns obtained with specific probes in DNA from different strains or different individuals.

3.6.2 PCR and sequence analysis

Usually the sequence of the vector is known. It is therefore simple to design a pair of primers that will hybridize to a region at either side of the cloning site, and use these primers to PCR amplify a fragment containing the inserted DNA (see Chapter 4). PCR amplification will then produce a product of a characteristic size, if you have cloned the right piece of DNA. You can screen a substantial number of potential recombinants in this way. Once you have found one that does produce the correct size of product in the PCR, you can determine the sequence of the PCR product to confirm the nature of the insert. Alternatively, if you are using a small vector such as a plasmid, you can skip the PCR step and simply sequence the purified plasmid DNA (see Chapter 5 for the method of doing this).

Sequence analysis is much the best way of characterizing your clone, as, rather than providing estimates of size or similarity to the original probe, it will provide the exact sequence of the DNA fragment. It is also much cheaper and faster than it used to be, and bypasses the laborious procedure of blotting and hybridization. However, the other procedures outlined above are useful if you need rapidly to screen a large number of clones to obtain the right one. Once the ‘correct’ clone has been identified, this will be confirmed beyond doubt through the use of sequencing.

4
Polymerase Chain Reaction (PCR)

On 10 December 1993, Kary B. Mullis received the Nobel Prize in Chemistry from King Carl XVI Gustaf of Sweden for the invention of the polymerase chain reaction (PCR) method. Mullis had published and patented this invention only eight years earlier, in 1985. The same year, the Physiology and Medicine prize was awarded to Richard J. Roberts and Phillip A. Sharp, for their independent discovery that eukaryotic genes are composed of introns and exons. These laureates had waited twice as long, 16 years, since their fundamental discovery made in 1977. When you consider how fundamental the idea of introns and exons is to eukaryotic genetics, you gain some idea of the immense impact PCR has had, when its discovery was so quickly honoured. Although it is (as we shall see) quite a simple method with obvious limitations, the applications have revolutionized both basic and applied biology. This has been particularly dramatically illustrated in forensic science, where old open cases have been solved due to the ability to gain information from small amounts of trace evidence, and in clinical applications, where it has suddenly become possible to make in hours diagnoses that previously took weeks.

In essence, PCR serves but one function, which is to amplify a relatively short segment of DNA many times, even if it initially forms only a minute part of a very complex mixture. As we will see below, this procedure requires subjecting the sample to a cycle of defined temperature steps, including heating to 95°C. The key factor in transforming the initial PCR method into one that could be undertaken in thousands of laboratories worldwide, and that has such an important impact in many areas, was the use of a thermostable DNA polymerase called *Taq* polymerase. It was originally isolated

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

from *Thermophilus aquaticus*, a thermophilic archaean that thrives in hot springs at temperatures close to the boiling point of water. As a result, all enzymes in this organism have evolved to withstand high temperatures at which most proteins from most other organisms would denature immediately and irreversibly. Hence *Taq* polymerase can survive the repeated high temperatures required by the PCR without significant loss of activity. Apart from this distinct feature, *Taq* polymerase (not to be confused with *Taq*I, a restriction enzyme from the same species) is a normal DNA polymerase, which is able to synthesize a new DNA strand complementary to a single-stranded DNA template. Like all other DNA polymerases, it requires a *primer* from which to start its synthesis. In fact, it is not a particularly outstanding DNA polymerase. Although it has high processivity (which means it has a propensity for remaining bound to the DNA and continuing to add successive nucleotides without dissociating from the DNA), it lacks proofreading activity, so it is unable to correct erroneously incorporated nucleotide bases. This is significant for some applications, as the product may not be a completely accurate copy of the original sequence. As we will see later, other thermostable polymerases are available that do have proofreading ability and can be used in these applications.

PCR uses *Taq* polymerase (or other thermostable DNA polymerases) for the exponential amplification of a DNA fragment from a longer initial template, which could be as long as a whole chromosome. The amplified fragment (*amplicon*) is defined by two short synthetic oligonucleotides, *primers*, which are complementary to the opposing DNA strands of the template that is being amplified. This introduces an important limitation to the method: You must know the sequence for at least part of the DNA molecule you wish to amplify – or you must at least be able to make an educated guess.

So how do we go from a small amount of, say, total human genomic DNA to a large amount of one short region that has been exponentially amplified to the extent that it entirely dominates the reaction mixture? This will be clear if we go through the first few cycles in a PCR amplification.

4.1 The PCR reaction

In this example, we will be starting with genomic DNA purified from human buccal cells, collected by swabbing the inside of the mouth with a cotton bud. The preparation will contain sheared chromosome fragments. Only a small amount is required even from such a complex template – in fact, the DNA from a single cell can be sufficient as the starting material for PCR.

We will also need two primers. These can be synthesized to your specifications from a specialist supplier at a very low cost within a couple of days. In this example, we will replicate the work of Mullis and his colleagues and use primers flanking a region within the human beta-globin gene that is mutated

in sickle-cell anaemia. A great excess of primer molecules is added to the reaction (this of course refers to an excess in molar, or molecular, terms, as the primer is very much smaller than the template – see the discussion in Chapter 2).

The binding, *annealing*, of the primer to the template is a typical DNA:DNA hybridization reaction, and follows similar principles to the hybridization of probes as described in Chapter 3. First, the double-stranded template needs to be *denatured*. The temperature used for this in PCR, 95°C, hardly does any damage to the *Taq* polymerase molecule during the minute or so that the PCR reaction is heated to this temperature.

The temperature is then lowered to the optimal annealing temperature, where the two primers can bind to the opposing DNA strands (Figure 4.1). This is the only temperature in a PCR cycle that can be varied widely. It is chosen for the optimum binding of the primers to the correct template, with a minimum binding of them to other, non-specific sequences. If the annealing temperature is too low, the primers will bind at other positions on the template, resulting in false products, or no detectable product at all. If the annealing temperature is too high, the primers may fail to bind at the correct site, resulting in no amplification. The temperature needed will depend on the exact sequence and length of the primers, as discussed in Section 4.2.1. Because the primers are short, and are used at relatively high molar concentrations, annealing is rapid, taking less than a minute.

The temperature is then raised to approximately 72°C, which is normally the optimum *extension* temperature for a PCR reaction. The *Taq* polymerase will now produce complementary DNA strands starting from the primers. The extension proceeds at approximately 1000 bases per minute. Hence, to amplify a region of DNA that is 500 bases in length, under normal conditions you should allow an extension time of at least 30 seconds.

However, it is important to note that because the initial template in this case is many times larger than the length of the desired amplicon, the polymerization will proceed until it is interrupted. This happens when the temperature is yet again raised to 94°C in order to start the next cycle in the PCR reaction, which consists of steps that are identical in temperature and duration to the previous ones. As we finish the first PCR cycle, we have two double-stranded DNA molecules for each one that we started with. Each one contains one strand of the original template, and one novel strand, which is defined at one end, specifically, by the oligonucleotide primer and at the other end, non-specifically, by how far polymerization was able to proceed during the time we allowed for extension.

The advantage of *Taq* DNA polymerase over, say, *E. coli* DNA polymerase will now become apparent. The *E. coli* polymerase could have performed the extension, but it operates at 37°C, which means that you would not be able to increase stringency and would risk non-specific hybridization.

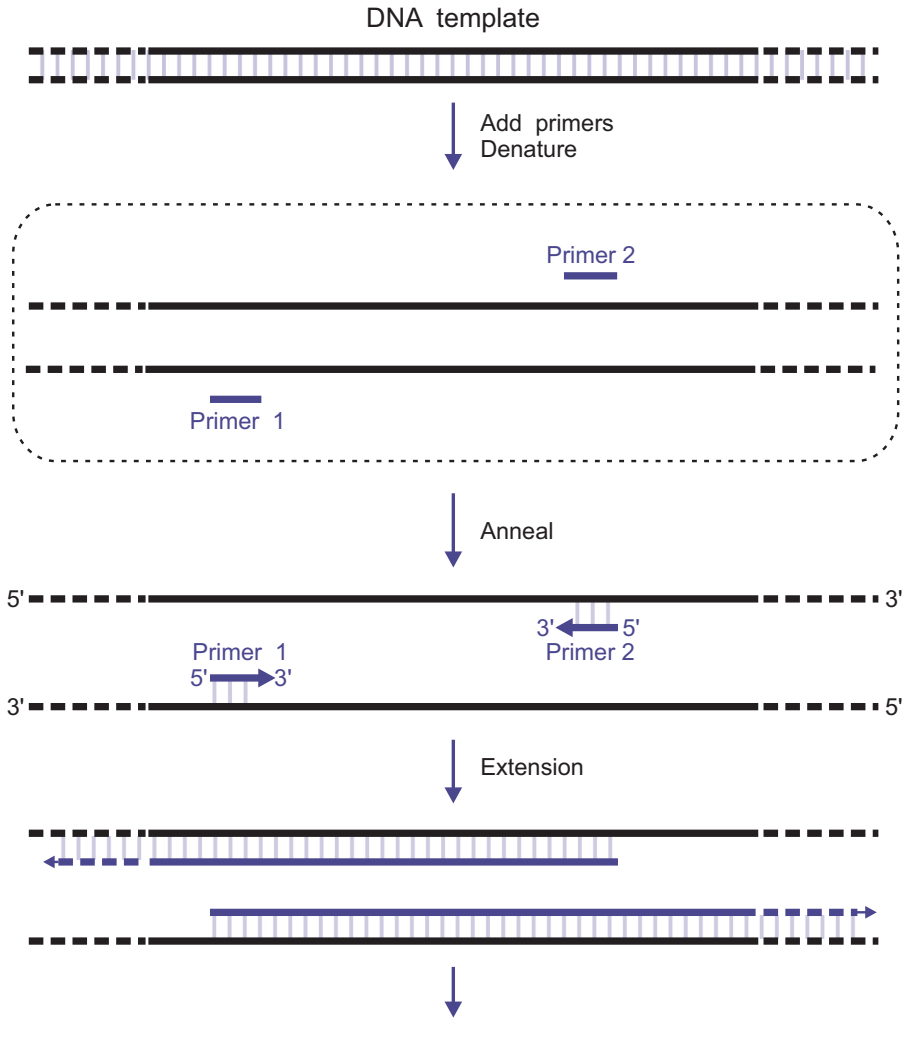


Figure 4.1 Polymerase chain reaction: first cycle.

Above all, however, you would lose all activity of the *E. coli* DNA polymerase in the denaturation step, and would have to add fresh enzyme. *Taq* polymerase, by contrast, survives the denaturation step unscathed.

This second denaturation step creates four single-stranded template molecules, two of which are derived from the original template and two of which are our newly synthesized strands. At the following annealing step, one molecule of the complementary primer will bind to each of these single strands. As the temperature is again raised to 72°C, *Taq* polymerase will

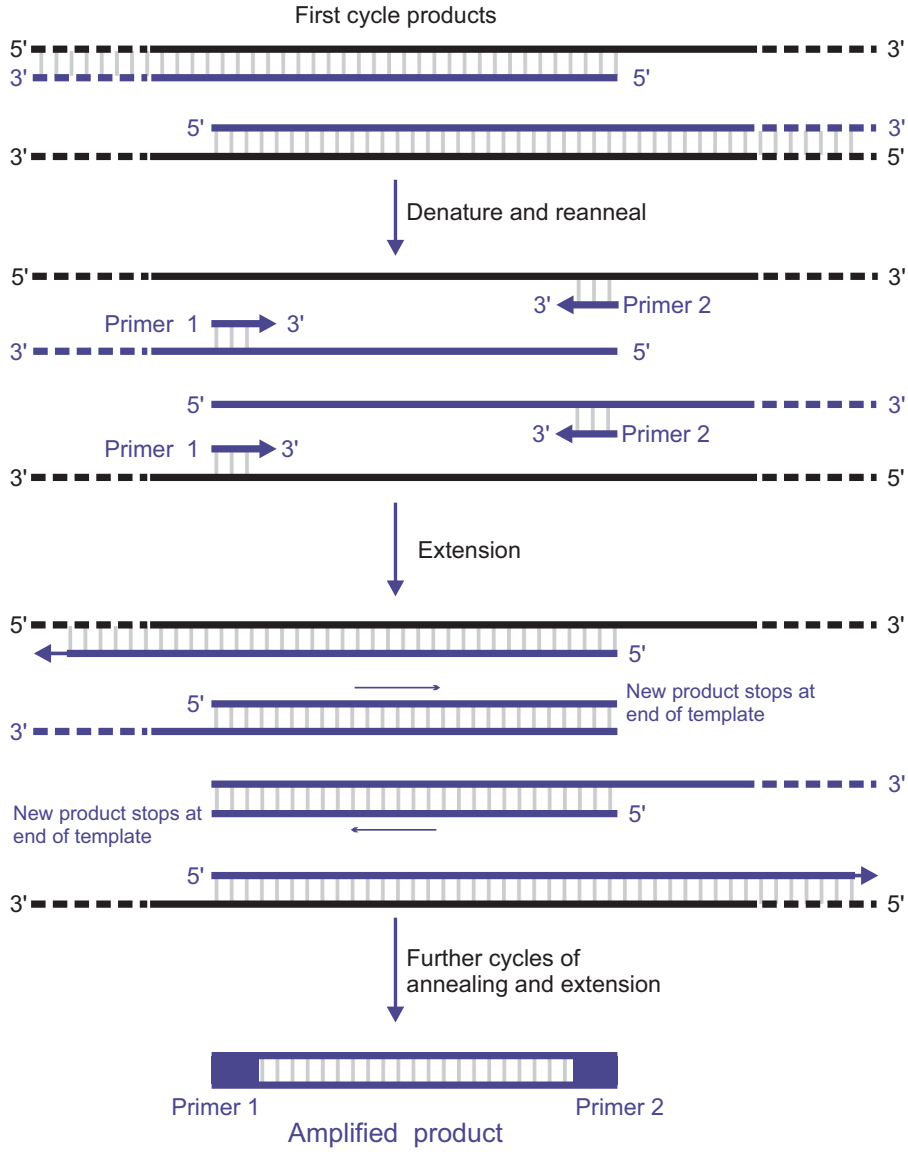


Figure 4.2 Polymerase chain reaction: second cycle.

begin to extend the primers (Figure 4.2). Two of the four extensions, where the primer has again bound to the original chromosomal template, are identical to the first extension in that they terminate only when the temperature is raised. Note, however, that this does not apply to the two strands that are produced with the new strands as templates. These templates end abruptly where the opposite primers had bound.

If we envisage yet another cycle, then again the two original template strands will give rise to a long product, limited only by the duration of the extension reaction. Priming on all the other strands will yield a product that is defined and delimited by the primers at both ends. Each subsequent cycle will produce two new long strands, but the number of new short strands made will increase exponentially, so that eventually the reaction mixture will be completely dominated by the newly formed short DNA strands, our target amplicon, with one primer incorporated at each end. It follows from this that the ends of the new DNA are actually defined by the primers, unlike the intervening regions, which are entirely defined by the original template (apart from any mistakes in the amplification).

4.2 PCR in practice

The launching of PCR as a revolutionary new technology was enabled by the development of the programmable thermocycler. These instruments are based on metal heating blocks with holes for the PCR tubes. The blocks are designed to switch between the programmed series of temperature steps with great speed and precision by a combination of heating and cooling systems. The use of small (0.2–0.5 mL), thin-walled tubes helps to ensure a rapid change of temperature. Alternatively, for larger numbers of samples, microtitre plates are used, allowing 96, 384 or even 1536 reactions to be run simultaneously.

Because the PCR reaction is performed in a small volume (typically less than 50 μL), and because of the high temperatures involved, it can be easily imagined that the water would quickly evaporate and end up on the inside of the lid rather than on the bottom of the tube. There are two ways of preventing this. The original one was to place a drop of mineral oil over the reaction. The less messy approach applied almost universally today is the use of PCR machines with heated lids, meaning that there is an even temperature throughout the tube, thus preventing condensation.

4.2.1 Optimisation of the PCR reaction

The importance of the annealing temperature in PCR has already been discussed. If the temperature is too high, binding of the primers to the target will not be stable enough for amplification to take place. If it is too low, the system will become too tolerant of partial primer–template mismatches, and will therefore be non-specific, so you may get unwanted products. Furthermore, if the primers bind at too many places, you will get so many abortive reactions that you may not see any products at all.

The annealing temperature will be affected by the sequence and length of the primers. Because G-C pairing is stronger than A-T pairing, with three

hydrogen bonds rather than two occurring between the bases, the more Gs and Cs there are in the primer, the stronger it will bind to its complementary sequence in the template DNA, and therefore higher annealing temperatures can (and must) be used. See Chapter 3 for a discussion of other factors that influence the optimal annealing temperature. Although computer algorithms are available to predict the optimal annealing temperature for a primer, in practice some trial and error is often needed. Normally, the annealing temperature is chosen somewhere between 40 and 60°C, although for templates with a high GC content annealing temperatures as high as 72°C (the normal extension temperature) may be used. Note also that we may want to use conditions that will allow priming from sites that are only partially matched to the primers, and for this purpose we would use a lower annealing temperature. This would apply, for example, if we were using primers derived from a sequence that is likely to be only partly related to the (unknown) sequence we are trying to amplify.

Another important factor is the concentration of magnesium ions in the reaction mixture, as these are a necessary cofactor for the *Taq* polymerase. Typically, a magnesium concentration of 1.5 mM is used. A higher magnesium concentration usually gives a higher amplicon yield, but also a lower specificity. Lowering the magnesium concentration increases specificity, but will decrease enzyme activity and, hence, the amount of product produced.

4.2.2 Primer design

The most important PCR parameter by far is the design of the primers. Assuming that we know the sequence of the target gene, we want to pick a region of a suitable length for amplification. For analytical purposes, 200–500 bp is adequate. Much smaller than 200 bp and it will be difficult to detect on an agarose gel; if it is too big, the amplification will be inefficient. (For other purposes, it is possible to amplify much longer sequences, but it requires special conditions – see Section 4.4.) We then select two sequences (usually 20–25 bp) either side of this region for binding of our complementary primers. It is essential to get the orientation of the primers right. The 'left-hand' (forward) primer will be the same as the sequence of the 'top' strand – remember that double-stranded DNA is conventionally represented with the top strand in the 5′-to-3′ direction, reading from left to right; this is the strand that is shown if your DNA is shown in single-stranded format. The 'right-hand' (reverse) primer must be the sequence of the complementary strand, read from right to left (which is the 5′-to-3′ direction for the lower strand). This is illustrated in Figure 4.3 (using unrealistically short primers for clarity).

There are some further factors that affect the design of the primers. If an oligonucleotide contains sequences with complementarity to itself or the other primer, then these can result in base pairing between the primers or

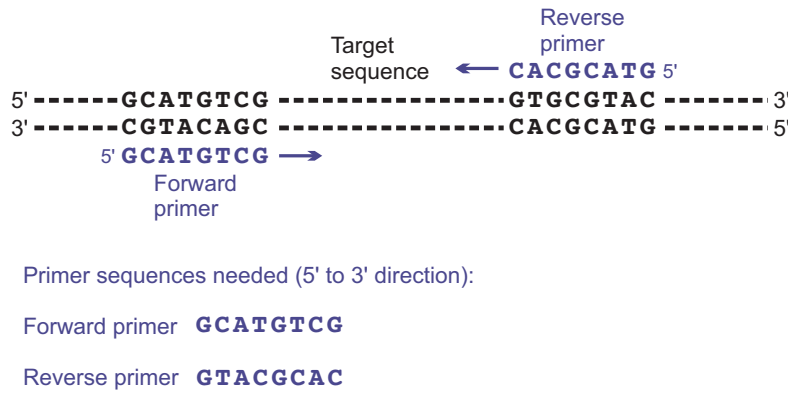


Figure 4.3 Forward and reverse primers.

even within one primer, rather than annealing to the target sequence. Because the primers are present in such great excess, they are much more likely to encounter a partially complementary primer molecule than a perfectly complementary template molecule, forming these redundant pairings. Binding of the primer molecules to each other may cause the production of low-molecular-weight fragments called *primer dimers*. Thus, primer complementarity is to be avoided. It is also desirable for the two primers to have similar annealing temperatures, which means they should be similar in length and base composition. Computer programs for primer design also check that the chosen sequences are suitable in both of these respects.

We also want to make sure, as far as possible, that each primer will anneal only to the chosen sequence, and not to any other region of the template DNA. If we are working with DNA from an organism for which the complete genomic sequence is known, then we can search the database to check for other sequences that are complementary to our primers (see Chapter 5 for a description of how these searches are done). Ultimately, it is in part a matter of trial and error – even the best designed primers do not necessarily work well, in which case we need to change one or both primers.

What if we do not know the sequence of the target region? This problem can often be circumvented by some informed guesswork. For example, we may use sequence information from a more or less closely related organism to design our primers. The fact that the sequences will probably not be a perfect match can be accommodated in two ways. Firstly, we can use a lower annealing temperature, to allow the imperfectly matched primers to anneal, and secondly we can create degenerate pools of primers, i.e., we can incorporate ambiguities into the primers. This is analogous to the use of heterologous probes for library screening (see Chapter 3). When using heterologous primers in this way it is essential to remember that the ends of the amplified

product represent the primer sequence, and are *not* derived from the target genome.

4.2.3 Analysis of PCR products

The normal way of analysing the products of a PCR reaction is to separate the samples in an agarose electrophoresis gel. This allows you to ascertain that only one fragment is obtained in each reaction, which is usually the objective. By comparing the size of the amplified fragment to a molecular weight standard, it is also possible to make sure that the molecular weight is the same as the predicted one (which is usually known). The assumption is that if the fragment is the predicted size, then it probably corresponds to the predicted fragment – an assumption that occasionally leads you down the wrong track, emphasizing the importance of alternative methods for checking the identity of the fragment (see below).

As well as primer dimers (see Section 4.2.2), you may get an amplification that appears to be specific (as shown by the absence of a fragment of that size in control reactions), but that gives a product of a different size than the expected one. This may indicate that one, or both, of the primers is binding at a position other than the intended one, producing an artefactual product. This may also be manifested as a specifically amplified fragment together with a more or less complex mixture of non-specific ones. In these cases, greater specificity may be obtained by adjusting one or more of the PCR parameters (see the previous section).

Alternatively, *nested* PCR may be used to increase the specificity (and/or the sensitivity) of the amplification. In nested PCR (Figure 4.4), a small aliquot of the original reaction is transferred to a second, 'nested' PCR reaction. In the nested PCR reaction, one or both primers are replaced with a second set of primers that will bind specifically within the desired amplification product. In this way, the undesirable products that have typically been amplified because of a coincidental sequence similarity in the region of the original primers, will normally disappear, as it is highly unlikely that they will contain annealing sites for this second set as primers as well.

As mentioned above, production of an amplicon of the correct size suggests that the PCR has produced the desired amplification, but does not prove it. In order to provide a further level of certainty about the identity of a PCR product two approaches may be taken. First, the electrophoresis gel of the PCR products may be blotted and hybridized (see Chapter 3) with a probe complementary to the expected product. Second, because of the remarkable recent progress in sequencing technology, most people find it faster, easier and cheaper to perform direct sequencing on the PCR product (see Chapter 5).

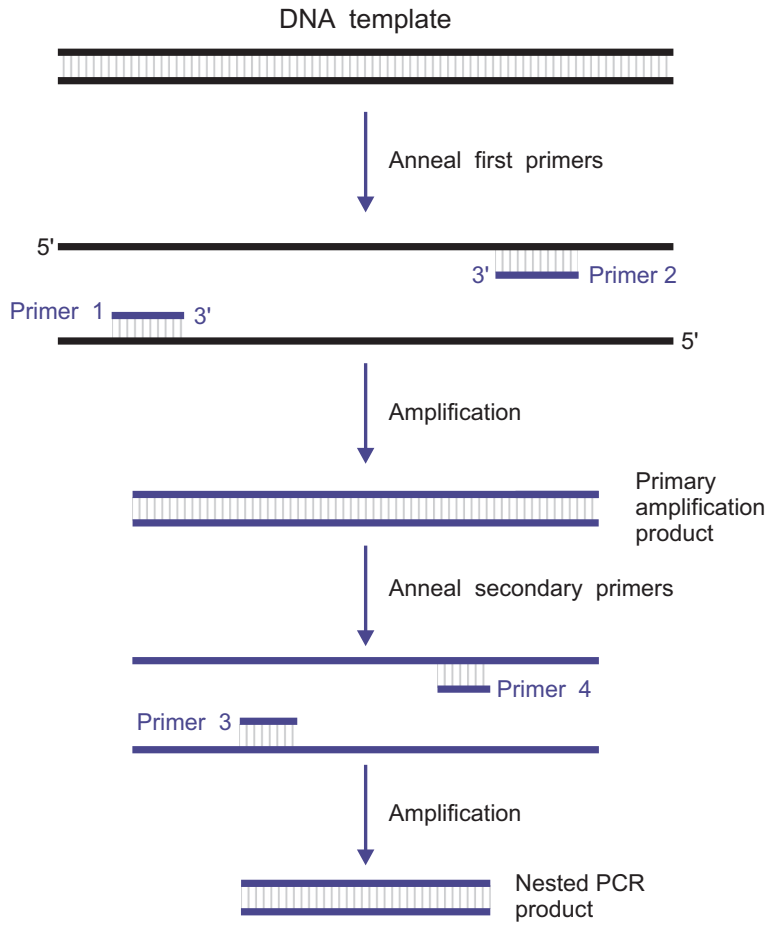


Figure 4.4 Nested PCR.

4.2.4 Contamination

The exquisite sensitivity of PCR, which enables large amounts of DNA to be amplified from (in theory) as little as a single molecule, carries with it a severe risk of contamination. To reduce this risk, it is necessary to prepare the materials and set up the PCR reaction in a clean environment, separate from other activities in the lab, especially the analysis of the PCR products (which contain large amounts of material that can act as template for further amplification). A battery of additional precautions must be taken, including the use of negative controls, which lack template DNA. Even opening the tube after PCR will liberate substantial amounts of product. This is especially a problem with nested PCR, when the tube has to be opened to add the second set of primers.

Contamination can be controlled reasonably easily for many routine lab activities, where abundant amounts of template can be used. But in situations where you are trying to detect minimal amounts of a rare target DNA, it is a serious problem that demands extreme precautions. For example, in trying to amplify DNA from ancient bones, the slightest amount of contamination by modern human DNA can ruin the results.

Similarly, extreme precautions need to be adopted to prevent contamination in forensic and diagnostic applications (see below). For example, if you are trying to detect a pathogen using PCR, the tiniest amount of contamination of a negative sample can lead to a false positive result, with obvious important clinical implications.

4.3 Cloning PCR products

Although PCR is commonly used merely to detect the presence of a specific sequence in different templates, it is also often employed for the amplification of such sequences as a convenient way of obtaining a specific product for cloning. This is especially important when the starting material is very scarce, as in such cases the previously more common routes, such as constructing and screening a gene library, are impossible.

You might expect, from the description so far, that the PCR products would be blunt ended, and could therefore be cloned by normal blunt-ended cloning (see Chapter 2). However, as we heard previously, blunt-ended cloning is rather inefficient. Fortunately, *Taq* polymerase often tends to add, non-specifically, an adenosine residue to the 3′ ends of the product. The product is therefore not blunt-ended, but contains a tiny 3′ overhang, which is exploited in the so-called TA cloning method. This method uses a linearized vector plasmid that has been engineered to contain a thymidine overhang at each end. This produces short, but nonetheless sticky ends, which will anneal to the adenosine overhangs of the PCR products, allowing fairly efficient ligation (Figure 4.5). Note, however, that not all polymerases add adenosine residues to the product; in particular, the 'proofreading' polymerases produce genuinely blunt-ended PCR products, which would require the inefficient blunt cloning method. To increase the efficiency of cloning such products, several approaches exist: Firstly, if TA cloning is desired of such products, *Taq* polymerase may be added at the end of the amplification and allowed to proceed through one amplification cycle. This will add the A residue to the 3′-end of the PCR product that is needed for this procedure.

Secondly, as we saw in Chapter 2, an improvement in ligation efficiency can be gained through the use of *Vaccinia* virus DNA topoisomerase I (TOPO). Briefly, *in vivo*, topoisomerases are involved in the supercoiling/relaxation of DNA. They will cleave DNA at specific sites, leaving a sticky end. The energy released by breaking the phosphodiester bond is stored in a covalent bond

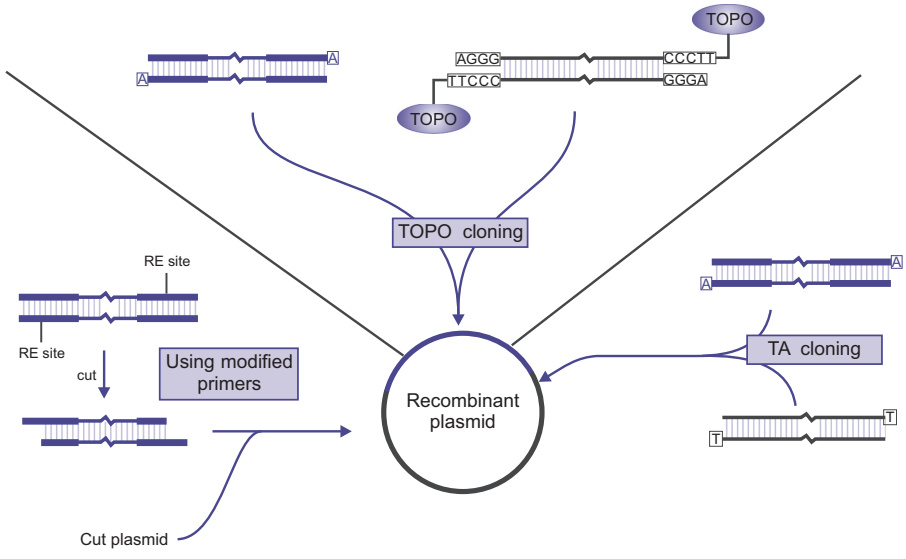


Figure 4.5 Cloning PCR fragments. RE, restriction endonuclease.

between the enzyme and one of the cleaved strands. The enzyme trapped in the sticky end will then rapidly and efficiently release its stored energy into the formation of a new phosphodiester fragment as soon as the sticky end encounters its complementary partner. Thus, TOPO has both endonuclease and ligase activity. Commercially available TOPO vectors offer sticky end overhangs, ranging from TA cloning to more complex sequences, already bound to the TOPO enzyme, ensuring a high ligation efficiency.

Finally, ligation efficiency can be increased through the use of modified primers. Since the primers only have to match at their 3′ regions, you can incorporate additional sequence, such as restriction sites, in the 5′ regions of the primer (Figure 4.6). In the first PCR cycle, the 5′ ends of the primer will not pair with the template, but that does not prevent annealing (provided there is sufficient similarity in the 3′ region of the primer) or extension (which is dependent on a perfect match of the final 3′ nucleotide). In subsequent rounds, the part of the primer carrying the restriction site (which is now incorporated into the product) will be accurately replicated, as the primers are now a perfect match with the template. The end product is a DNA fragment carrying the restriction site near the ends, so it can be cut with the restriction enzyme and ligated with an appropriate vector (Figure 4.5). This is particularly useful when cloning larger PCR fragments. As depicted in Figure 4.5, it is important to note that these incorporated restriction sites cannot be at the extreme end of the primer, usually requiring two or three nucleotides extra to be added on to the 5′ end. The extra nucleotides are important to ensure stable binding of

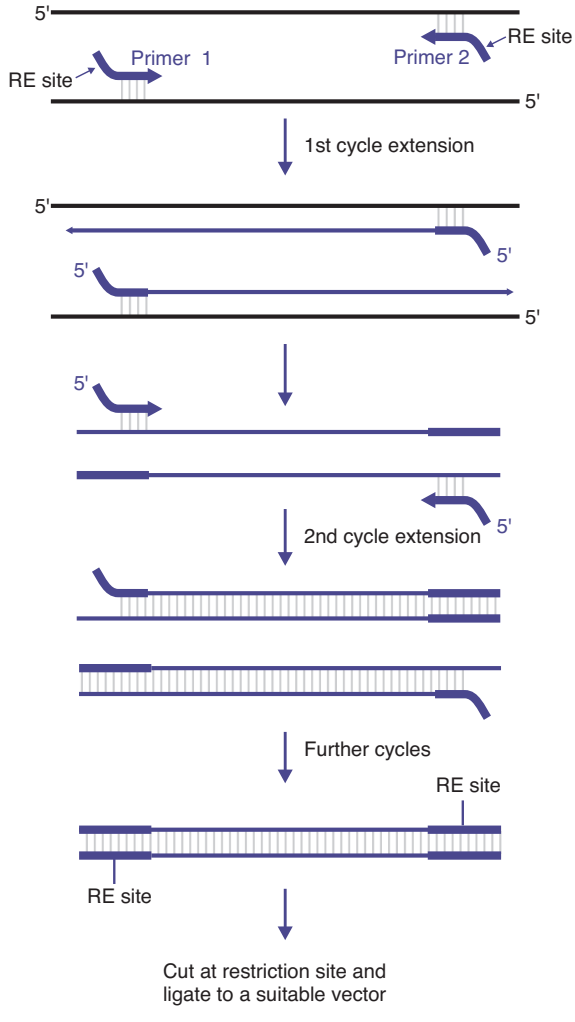


Figure 4.6 Adding restriction sites to a PCR product.

the restriction enzyme to the end of the amplified region of DNA, allowing efficient digestion, and subsequent ligation with the vector.

4.4 Long-range PCR

We have already discussed the fact that *Taq* polymerase lacks proofreading activity and is thus unable to correct its own errors. These errors occur approximately once per 9000 nucleotides, on average. Such mistakes will then

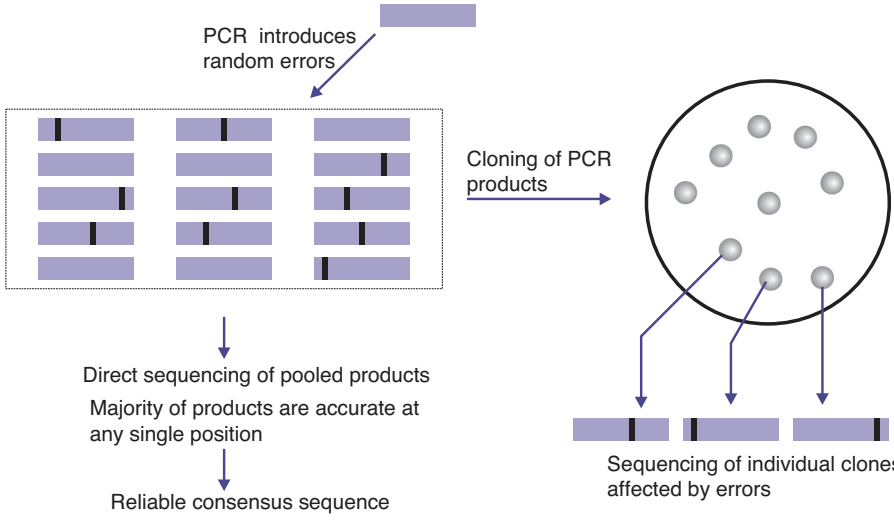


Figure 4.7 Detection of PCR errors.

be perpetuated in all new molecules descended from the one containing the error. If direct sequencing is used, this is not really a problem, because the defects will be randomly distributed and the vast majority of molecules in the reaction tube will be correct in any given position. However, if the PCR products are cloned it will potentially be a problem, because any error in the one molecule that happens to be cloned will be perpetuated in all the offspring of that one clone (Figure 4.7). For this reason, if you wish to clone a region of DNA by PCR, as opposed to merely detect its presence, then *Taq* polymerase is suboptimal as it may introduce errors.

Another consequence of the relative sloppiness of *Taq* polymerase is the fact that *Taq* polymerase can only efficiently amplify fragments of a few thousand base pairs. Both these problems can be solved by the introduction of thermostable DNA polymerases from other thermophilic organisms, such as *Pfu* and *Pwo* DNA polymerases. Unlike *Taq* polymerase, these enzymes do have proofreading activity, meaning that they produce amplification products with much higher fidelity (i.e., fewer, if any mistakes). Unfortunately, these polymerases are not necessarily as efficient as *Taq*. This has been ingeniously overcome by the introduction of proprietary mixtures of *Taq* and proofreading DNA polymerases. By combining the fidelity of amplification from the proofreading enzymes, with the efficiency and high yield of *Taq*, and exploiting the ability for one polymerase to continue the extension where the other one stalls, these mixtures may allow the efficient amplification of DNA fragments as large as 50 kb.

4.5 Reverse-transcription PCR

In Chapter 3, we described the use of reverse transcriptase to obtain a cDNA copy of mRNA, and the construction of cDNA libraries. Combining reverse transcriptase (RT) with PCR, a procedure known as RT-PCR, extends the application of PCR into the analysis of gene expression, either qualitatively or quantitatively, as well as greatly facilitating the construction of cDNA libraries or the cloning of specific cDNAs. The use of RT-PCR for the analysis of gene expression is described in Chapter 6.

One problem with RT-PCR is that the initial mRNA preparation may be contaminated with genomic DNA. The PCR step can then result in amplification of the contaminating DNA. When working with eukaryotic material, this can often be overcome by designing the primers so that the amplicon spans at least one intron. In this way, amplification of any genomic DNA will either be prevented altogether (because the presence of the intervening intron makes the sequence too large to be amplified), or at least it will be readily distinguished from amplified cDNA (because of the different size of the product). Generally, it is preferable to remove all traces of DNA from the mRNA, usually by treatment with RNase-free DNase.

4.6 Quantitative and real-time PCR

It is difficult to derive reliable quantitative information from conventional PCR, for example, by measuring the amount of product formed. There is no simple and reliable relationship between the amount of template you start with and the amount of product obtained, unless you can be certain that the PCR reaction is truly proceeding exponentially throughout. This will usually no longer be the case during the latter phase when the availability of reagents becomes limiting. In practice, this has now been superseded by *real-time PCR,* a method that is much more accurate and also simpler.

4.6.1 SYBR Green

The basis of real-time PCR lies in detecting the product as it is formed, without having to stop the reaction and run the products out on a gel. The simplest method to understand is to add, to the PCR mixture, a dye such as SYBR™ Green (pronounced as 'cybergreen') that fluoresces when it binds to double-stranded DNA. This is the cheapest and simplest way of performing real-time PCR, because the dye is generic (and thus not reaction-specific) and you can use the same primers as in any normal PCR reaction. However, this method may lead to artefacts, as you will be unable to distinguish the signal produced if the wrong product is made. More specific methods are described below.

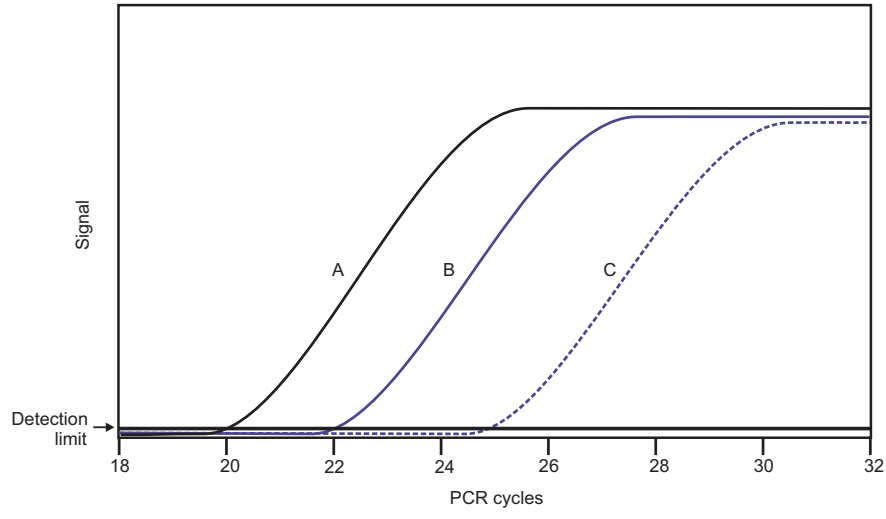


Figure 4.8 Real-time PCR.

If you carry out the PCR reaction in a machine that not only performs the temperature cycling needed for a PCR reaction but also will detect the fluorescence of your samples at the end of each cycle, then you can monitor the progress of your PCR in real time. Initially, the template is single-stranded, and so there is no signal. As amplification proceeds, double-stranded product is formed, and eventually enough of this product is made to allow the resulting fluorescence to be detected by the machine (Figure 4.8). The level of the fluorescence will increase over a number of cycles, and by extrapolation of the resulting curve back to zero you can determine the number of cycles needed for formation of a detectable amount of product. This value, known as the C_T value, is inversely related to the initial amount of template: higher amounts of initial template will result in lower C_T values. In the figure, sample A produces a detectable signal after 20 cycles, so we say it has a C_T value of 20. This represents more template than in samples B ($C_T = 22$) or C ($C_T = 25$). During the exponential phase of a PCR amplification, the amount of product theoretically doubles every cycle, providing that the reaction is working at 100% efficiency. Therefore, if samples A and B have C_T values two cycles apart then there must have been approximately four (2×2) times as much template for A in the initial reaction mixture. Likewise, there must be 32 (2^5) times as much of the A template as of the C template. In reality, most reactions are not 100% efficient, and it is usual to compare the amplification against a standard curve of known template concentrations to ensure accurate quantitation.

It is important to note that since SYBR Green will bind to any double-stranded DNA, there is no guarantee (apart from the specificity of the

primers) that the fluorescence detected is due to formation of a specific product. It is possible to circumvent this by programming the thermocycler to generate a melting temperature curve of the product after the end of the amplification. As discussed in Chapter 3, the melting temperature of a DNA fragment is dependent upon its length and sequence. Hence, melting point analysis can be used to determine the number of amplified products in a reaction; if only a single curve appears in the analysis then only one product has been formed during the PCR, presumably your specific product. However, the presence of multiple curves indicates that non-specific amplification has occurred, diminishing the quantitative nature of the PCR. Under such circumstances, it is usually necessary to further optimise the PCR conditions, or to buy new primers.

4.6.2 TaqMan

The other commonly used methods for quantitative PCR achieve specificity through the use of a probe that hybridizes specifically to the desired product, the probe being labelled in such a way that fluorescence only occurs as a consequence of the PCR reaction. The most generally employed method is the so-called TaqMan method (Figure 4.9). Here, an oligonucleotide probe that is complementary to an internal sequence within one of the amplified strands is labelled with a fluorescent group at its 5′ end, and a *quencher* at its 3′ end, which quenches the fluorescence at the 5′ end so long as both fluorophore and quencher are within the same molecule. Thus, virtually no fluorescence is emitted at the beginning of the reaction. As the reaction proceeds, the oligonucleotide probe, and the primers, will bind to the increasing number of newly synthesized strands. When such a strand is copied, the intrinsic 5′→3′ exonuclease activity of the polymerase will cleave the fluorophore away from the probe, thus liberating it from the quencher and enabling it to fluoresce. The intensity of this fluorescence is a directly proportional measure of the amount of product that has been generated. An additional advantage of this approach is that because the oligonucleotide sequence is complementary to an internal part of the amplicon, independent of the primers, it also helps to increase the specificity of the reaction. When using this method for amplifying cDNA, the oligonucleotide is often designed to span the junction between two exons, controlling for the presence of contaminating DNA in eukaryotic samples (see Section 4.5).

4.6.3 Molecular beacons

Another method, known as ‘molecular beacons’, also uses a probe containing a fluorophore and a quencher, but in this case instead of a linear probe

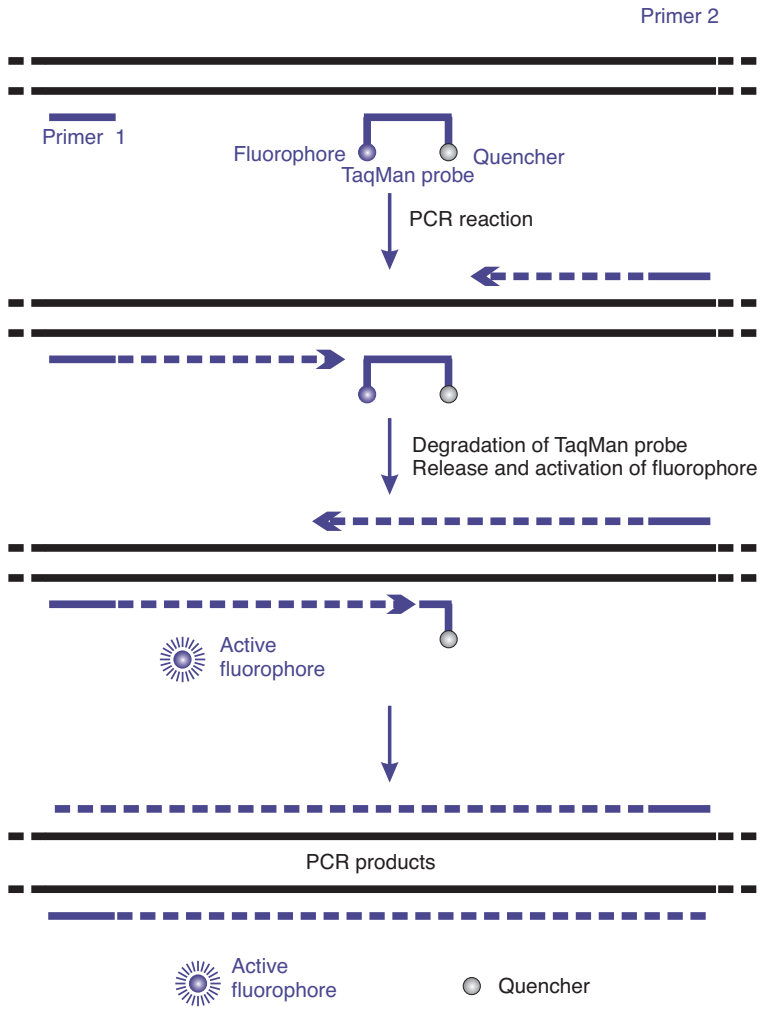


Figure 4.9 Real-time PCR: TaqMan reaction.

(as in the TaqMan method) the probe is designed with complementary ends that form a hairpin structure, with the loop containing a sequence complementary to the target (Figure 4.10). Here, the quencher/fluorophore combination works in a different way from that used for TaqMan, in that fluorescence is only suppressed if the two are in close proximity. When the probe binds to an amplified DNA strand, the hairpin conformation is lost, and the fluorophore is activated, because it is no longer physically close to the quencher. Since the detection is dependent on the relative stability of the hairpin and the target binding, it is more specific than TaqMan probes, and is especially useful for detecting single base changes (i.e., single nucleotide polymorphisms, or SNPs).

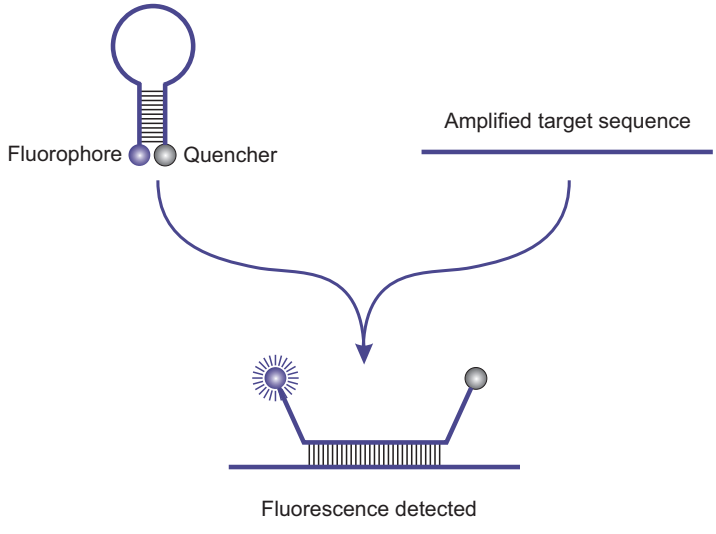


Figure 4.10 Real-time PCR: molecular beacons.

Real-time PCR is an important technique for estimating the level of expression of specific genes (see Chapter 6), when it is used in combination with reverse transcription. Its application for the detection of DNA sequences, including medical diagnostic applications (see Chapter 9), rests more in the convenience of detection of the product than in quantification. When analysing a number of samples, it is quicker than using agarose gels, and can be readily automated.

4.7 Applications of PCR

4.7.1 Probes and other modified products

One application of PCR in the lab is in the generation of nucleic acid probes for hybridization experiments. Not only does PCR readily generate large amounts of a specific probe, but also it provides a convenient way of labelling them at the same time, by including one dNTP modified by the incorporation of a dye, or biotin, which can subsequently be readily detected. Attaching fluorescent dyes, or other labels, to one or both primers provides a good way of obtaining a labelled product. For shorter probes, synthetic oligonucleotides, with similar labels incorporated, are often used instead.

Other modifications can be introduced by using modified primers. We have already described the use of this strategy to add restriction sites to the product, but it goes further than that. We can add sequences to the primer that encode tags that will form part of an expressed protein. Examples include

incorporation of histidine residues that will facilitate purification of the product, or sequences that will be recognized by specific monoclonal antibodies, which enable sensitive and specific detection of the product. Furthermore, by using primers that contain a deliberate mismatch, we can incorporate mutations into the product. These procedures are described more fully in Chapter 7.

4.7.2 PCR cloning strategies

PCR can be used in several ways to provide alternatives to the screening of random libraries as described in Chapter 3. For example, analysis of the known sequences of a set of related genes may show that although the genes are not similar enough to use long regions as reliable hybridization probes, there may be some shorter regions that are relatively highly conserved. It is possible to construct pairs of primers directed at these conserved regions, and use these to amplify the corresponding fragment from the genomic DNA of your target organism (Figure 4.11). Nested PCR is often used to add to the power of this approach. You can then sequence this product, to try to confirm that it is indeed derived from the correct gene, and you can use it as a probe for screening a gene library to isolate clones carrying the complete gene.

This approach needs to be treated carefully. One reason why part of a gene may be highly conserved is that it codes for an essential substrate-binding site, so any gene whose product uses the same substrate is likely to have a similar sequence. For example, many enzymes that use ATP as a substrate have a similar sequence that represents the ATP-binding site.

Alternatively, if you already know the sequence of the gene, then you can devise a reliable PCR system to amplify the whole gene or part of it, whether from genomic DNA or cDNA, and either clone or sequence the product. Cloning and/or sequencing a gene whose sequence you already know is much more useful than it sounds. For example, you may want to know whether the sequence varies at all in different strains (see Chapter 9), or you may want to identify alternatively spliced transcripts of a eukaryotic gene. PCR amplification combined with direct sequencing provides a quick route to answering that question. It is also employed in medical genetics in the search for individual sequence differences that may cause disease.

In Chapter 7, we discuss the introduction of changes into the coding sequence of a gene (*site-directed mutagenesis*); PCR plays a major role in this as well. A further topic in Chapter 7 is the use of a procedure known as *assembly PCR* for assembling synthetic genes. There is virtually no end to the versatility of the procedures to which you can adapt PCR as an aid to cloning and re-cloning DNA fragments in a research laboratory.

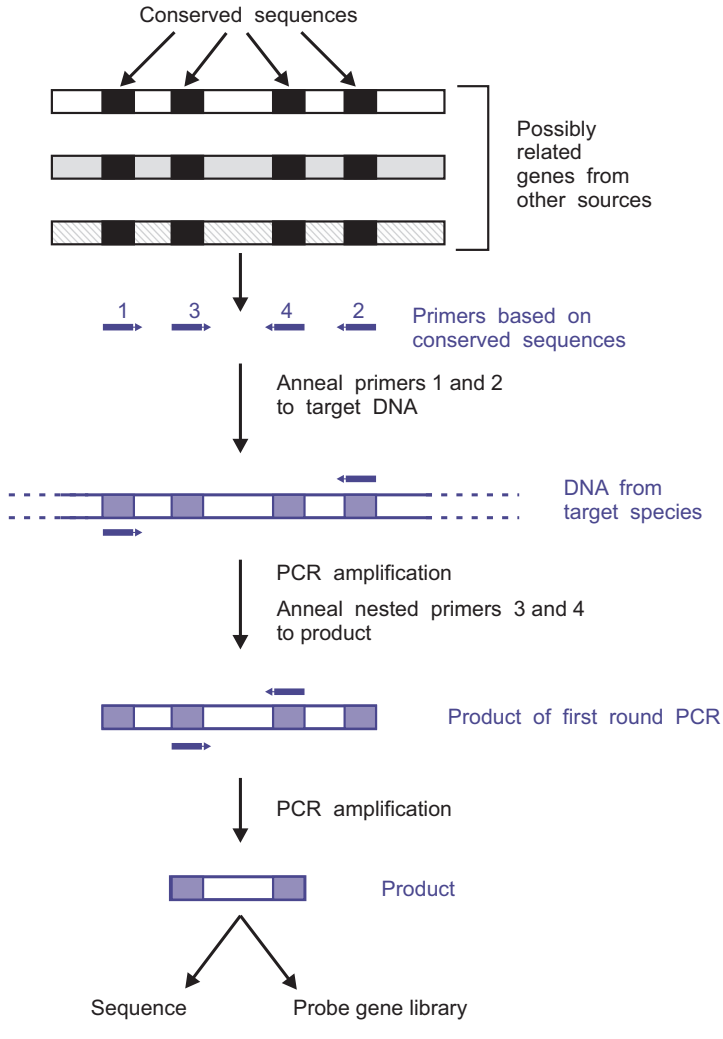


Figure 4.11 Use of conserved sequences to design PCR primers.

4.7.3 Analysis of recombinant clones and rare events

The best way of characterizing a recombinant clone, to make sure you have indeed produced the structure that you intended, is to sequence it – as a minimum, you will need to sequence across the new junctions that have been made. For this, you can either directly sequence the appropriate part of the recombinant plasmid, or you can amplify across the inserted fragment and sequence the PCR product.

PCR can also be used to detect specific events that occur *in vivo*, without having to clone the affected region of the genome. For example, if we want to

detect the transposition of an insertion sequence into a specific site, or if we want to test if our gene knockouts (see Chapter 8) have really inserted a foreign DNA fragment into that site, we can use PCR. In this case, we would use one primer directed at the insert and one at the genomic flanking sequence. Hence, we would only get a product if the event we are looking for has indeed occurred, as this is the only way in which the two primer sites could occur in close proximity. Otherwise the two primers would bind to completely different regions, and no PCR amplification would be possible.

However, such techniques need a word of caution, which applies to some extent in many applications of PCR. The technique is splendid for amplifying DNA where a genuine target exists. However, when there is no genuine target, nor other sequences highly similar to the intended target, then there is the possibility of artefacts being created, due to the power of the technique for amplifying extremely rare events. These may include the possibility of the polymerase 'jumping' from one DNA molecule to another. This may seem highly unlikely (although there are reports of it happening), but it only needs to happen once, in the early stages, and the product will then be amplified effectively, leading to an incorrect conclusion.

A similar effect can occur if the template contains a stable hairpin (or stem-loop) structure. This will cause the polymerase to pause (or even to stop altogether). However, while the polymerase is paused, it may encounter the other side of the hairpin and 'jump' across, yielding a product that lacks the hairpin. Although this rarely happens, the product will then amplify much more effectively than the original template, being shorter, and the final product will be shorter than it should be – leading to the suggestion that your strain contains a deletion at that point.

4.7.4 Diagnostic applications

PCR has many important practical applications in the real world outside molecular biology laboratories. Amongst these are forensic applications – the detection and identification of specific DNA fragments that can be traced to a particular individual – and uses in medical diagnostics, including the detection of mutations causing human genetic diseases. PCR can be employed for the detection of pathogenic microorganisms in a clinical specimen (such as sputum or blood). This is especially valuable for organisms that are difficult to culture, including many viruses and some bacteria. In addition, PCR tests (often in combination with specific gene probes) are available for the identification of pathogens that have been cultured. In these contexts, the problem of contamination, as discussed earlier, has to be taken very seriously.

We will look at some of these applications further in later chapters.

5
Sequencing a Cloned Gene

Having considered in previous chapters the ways in which we can obtain specific genes or fragments of DNA – either by cloning or by PCR amplification – we can now look at how we characterize that gene. In this chapter, we will deal with the sequencing of individual genes, annotation of database entries, and some ways of confirming the function of the gene. In the following chapters, we will look at how we analyse the expression of individual genes, and ways in which genes and their expression can be manipulated. In Chapters 8–10 we will extend these topics to genome-wide studies. While the frontiers of research have largely moved on from cloning and sequencing individual genes, to studies of whole genomes, there is still an important role for sequencing single fragments, There will probably always be a need to confirm that you have the right piece of DNA, or to compare single genes between different individuals and species. We can also use this chapter to introduce some of the analytical methods that will be developed further in relation to genome studies in later chapters.

5.1 DNA sequencing

5.1.1 Principles of DNA sequencing

Sequencing is the primary way of characterizing a macromolecule, whether it be determining the order of amino acids in a protein, or of bases in a nucleic acid. Protein sequencing was a very important tool before genes could be cloned and sequenced. With the advent of recombinant DNA technology, it was largely superseded by the much more efficient and economical method of DNA sequencing, with the sequence of the encoded protein being deduced

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

from the sequence of the gene. However, the technology available for direct protein sequencing has improved dramatically over recent years. Although this is beyond the scope of this book, it is worth noting that protein sequencing is a viable alternative to cloning and sequencing an unknown gene, if you have the purified protein (and the necessary equipment), especially as it provides additional information, such as identifying post-translational modifications, that are not available from the DNA sequence. Despite these potential advantages, protein sequencing remains overshadowed by DNA sequencing, especially as genome sequencing can now give you the predicted sequence of *all* the proteins that are potentially produced, not just one.

The method that was initially responsible for this revolution was developed by Frederick Sanger, for which he was awarded the Nobel Prize in Chemistry in 1980 – having previously (1958) received a Nobel Prize for his work on protein structure, one of only five people to receive more than one such award. There are now several sequencing methods (which will be described in Chapter 8) that are much quicker for determining genome sequences and, for that purpose, have largely supplanted the Sanger method. However, the Sanger method is still routinely used for sequencing individual cloned genes, or PCR-amplified fragments of DNA.

The principle of the Sanger method, also known as dideoxy sequencing, is illustrated in Figure 5.1. To understand the dideoxy procedure it is necessary to remember two fundamental facts about DNA synthesis. Firstly, synthesis of a DNA strand does not start from scratch. It requires a *primer* annealed to a *template* strand. Synthesis of the new strand works by adding bases to the primer that are complementary to the template, extending the sequence. Thus, by using a primer that binds to a specific position on the template, we can ensure that all the new DNA that is made starts from the same point. Secondly, the addition of bases to the growing strand occurs by formation of a covalent phosphodiester bond between the 5′-phosphate on the nucleotide to be added and the 3′-OH group on the existing molecule. The substrate for this reaction is a 5′-dNTP, i.e., a deoxynucleotide with three phosphates at the 5′-position of the deoxyribose sugar; two of these phosphates are eliminated in the reaction.

The sugar part of the natural substrate is more specifically a 2′-deoxyribose. This means that it does not have a hydroxyl group at the 2′-position (which distinguishes it from the ribose sugars that occur in RNA). But it does have a 3′-OH group, which is necessary for the formation of the next phosphodiester bond when a subsequent nucleotide is incorporated into the DNA strand. What happens if we use, instead of the natural substrate, one in which there is no 3′-OH group, i.e., a 2′,3′-dideoxy derivative (ddNTP; see Figure 5.2)? In this case, the ddNTP can be incorporated into the DNA, by formation of a phosphodiester bond between its 5′-phosphate and the 3′-OH on the

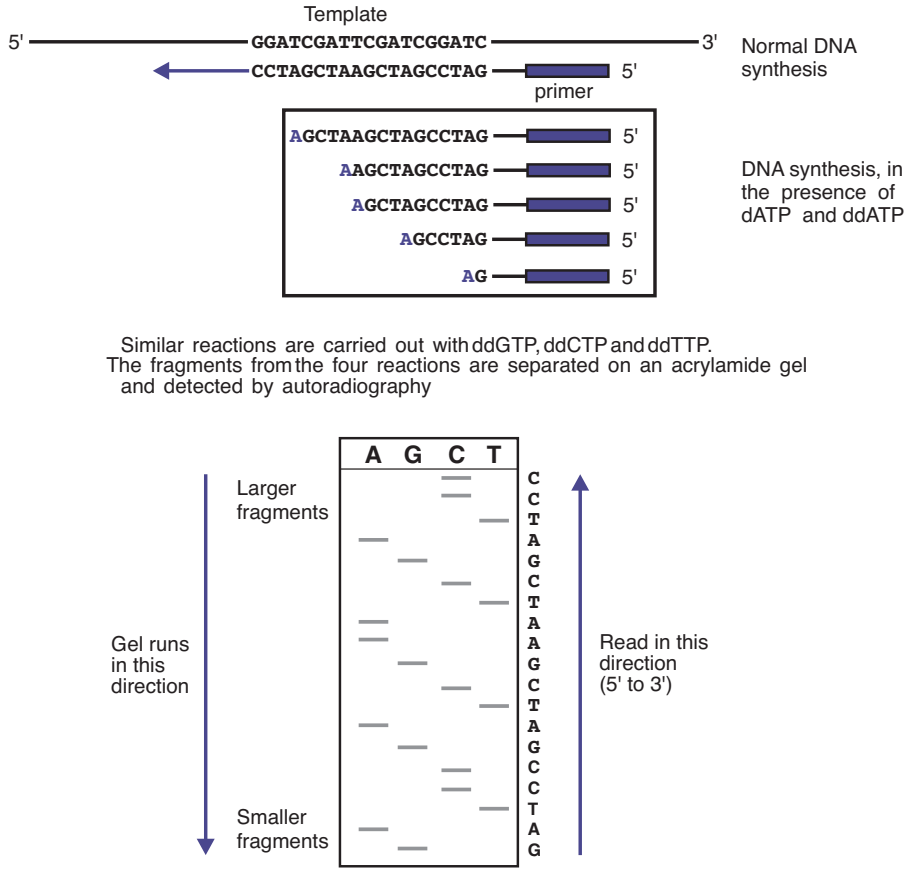


Figure 5.1 Determination of DNA sequence (Sanger method).

previous residue. However, this reaction produces a strand that does not have a 3′-OH at the end, and so no further bases can be added. DNA synthesis will therefore terminate at that point.

So if we replace one of the dNTPs with a dideoxy derivative – for example, if we use a mixture of dGTP, dCTP and dGTP (the normal substrates) but replace dATP with the relevant 2′,3′-dideoxy derivative (ddATP) – then DNA synthesis will proceed only as far as the first A residue, after which it will stop (Figure 5.3). If we carry out a set of four such reactions, in each case replacing one of the dNTPs with its corresponding ddNTP, we would produce four molecules of different lengths, each proceeding just as far as the first occurrence of the relevant nucleotide in the sequence.

Just determining the first occurrence of each base would not be much use. However, instead of completely replacing, say, dATP with ddATP, we can

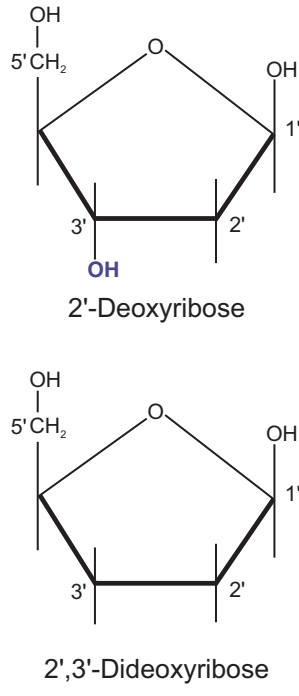


Figure 5.2 2′-Deoxyribose and 2′,3′-dideoxyribose.

use a mixture of the two nucleotides, with most of these bases being the normal dATP. So, at the first T residue in the template, a few molecules of the new strand will have ddA added (and will therefore terminate), but most will have the normal A residue incorporated, and the reaction will be able to proceed. At the next T in the template, some more molecules will terminate, and so on. We will thus have a series of molecules of different chain lengths, each ending with ddA, corresponding to a T in the template (see Figure 5.1). These can be separated by electrophoresis in a polyacrylamide gel, using denaturing conditions to prevent the DNA strands from folding up. The molecules are separated on the basis of their size, with smaller molecules running faster through the gel. We therefore get a series of fragments, corresponding to the positions of A residues in the new strand (T residues in the template). With a set of four reactions, each using a different ddNTP, we get a set of four lanes, from which the sequence can be read as shown in the figure.

For manual DNA sequencing, we would carry out the reaction with one of the dNTPs radioactively labelled, so that exposure of the gel to an X-ray film would result in a pattern of black bands on the film. Because the smallest fragments (which are closest to the primer) will migrate faster, they will be at the bottom of the gel; the autoradiogram is therefore read from

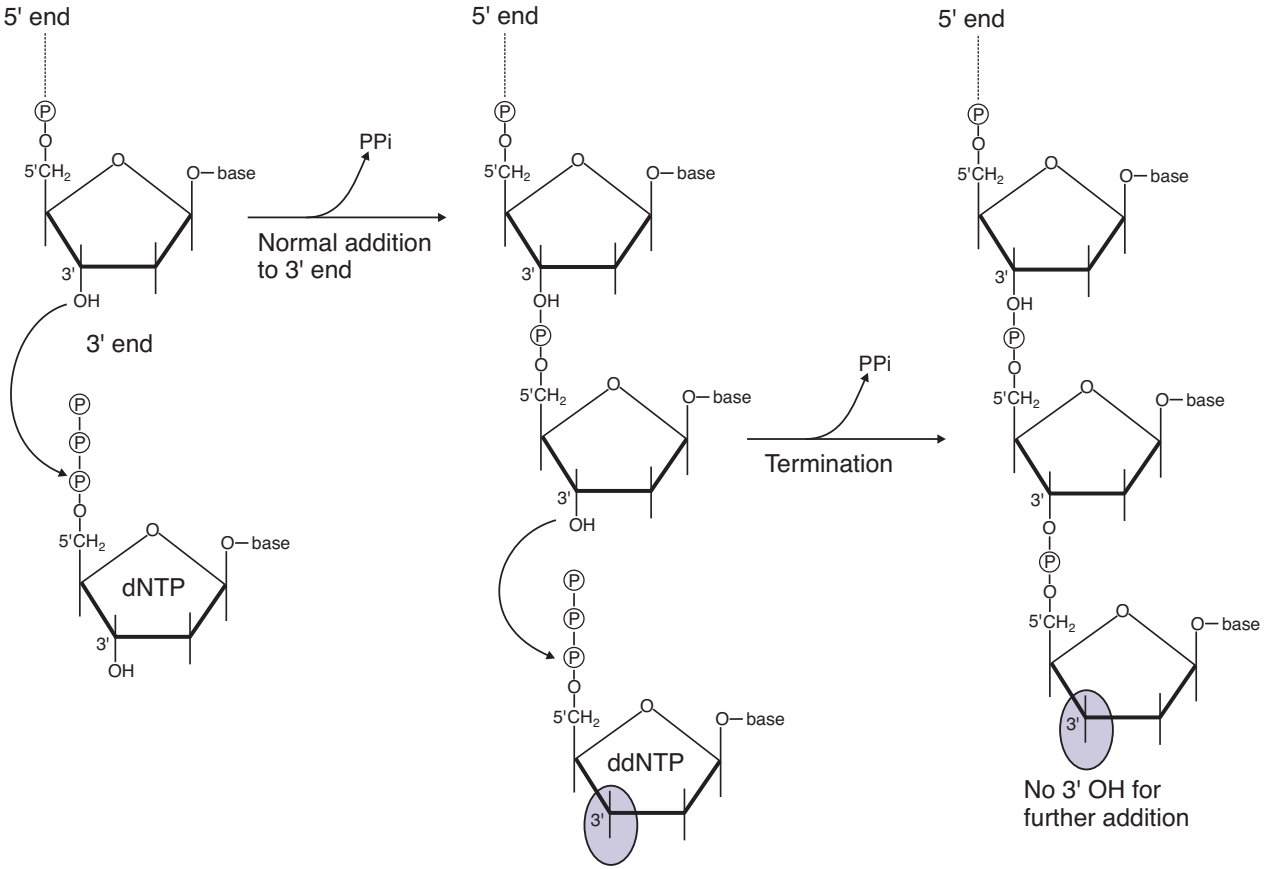


Figure 5.3 Chain termination by dideoxy nucleotide triphosphate (ddNTP).

bottom to top. However, manual sequencing has now largely been replaced by automated sequencing.

5.1.2 Automated sequencing

Although the principle of automated sequencing is the same as for the manual method, the method of detection is different. For automated sequencing, either the primer or the ddNTPs are labelled by incorporation of a fluorescent dye. Thus, rather than running the gel for a finite time, and reading the result, the machine uses a laser to read the fluorescence of the dye as the fragments pass a fixed point. Much longer sequences can be read from each track in this way, because the separation is not stopped at a specific point – instead, each fragment is allowed to proceed sequentially to the bottom of the gel where the resolution is the greatest. Due to this, 1000 base reads are now routine for automated sequencers. A further advantage is that the sequence read by the machine is fed automatically into a computer. This is not only much quicker than reading a gel manually and typing the resulting sequence into a computer, but also avoids the errors that are virtually inescapable with manual data entry. On the other hand, the computerized interpretation algorithm is more or less prone to some errors of its own, and may require manual checking to resolve any inconsistencies.

If the primer is labelled, all the products carry the same dye, and so you still have to use four lanes, one for each of the ddNTPs. However, if the four ddNTPs are each labelled with different dyes that can be distinguished by the photometer, the sequencing reactions can be performed in a single tube and separated in a single lane, thus increasing the capacity of the machine. A further development on this is to replace the polyacrylamide gel, which needs to be cast anew for each run, with a reusable matrix-filled capillary. The machine can run a large number of samples at the same time, increasing throughput and decreasing cost.

Both manual and automated DNA sequencing methods suffer from problems that can give rise to errors in the sequence. These are often associated with the presence of certain combinations of bases in the DNA – for example, runs of identical nucleotides, when it can be difficult to determine exactly how many bases there are in the run, stretches of C and G, which may cause such a high melting temperature that the sequencing is unable to proceed past them, and the presence of secondary structures in the DNA such as hairpin loops, which can interfere either with DNA synthesis (causing premature termination) or with the running of the fragment on the gel. Some of these potential errors can be identified and minimized by altering the reaction conditions, by sequencing a different overlapping fragment covering the problem region, or by determining the sequence of the complementary strand. A good

complete sequence will therefore be derived by reading and assembling several overlapping sequences in each direction.

5.1.3 Extending the sequence

Although automated DNA sequencers allow much longer reads than manual methods, the length of sequence that can be obtained from a single run is still limited. As the fragments get larger, it becomes harder to resolve the small differences in total length caused by a single nucleotide, and the amount of product in each fragment is reduced, making the signal weaker. The sequence thus becomes progressively less reliable. Sometimes this is not a problem: if we are using sequencing to verify the structure of a recombinant plasmid that we have made, or if we are looking at the variation of a specific region between different strains, then the sequence of a few hundred bases may be quite sufficient. But if we want to sequence a whole gene, then we will usually require a longer string of sequence data.

One strategy for extending the length of sequence determined is referred to as *walking* (Figure 5.4). Remember that the sequence depends on DNA synthesis starting from a specific primer. Usually, with an unknown cloned sequence, we would start with a primer directed at the cloning vector close to the point of insertion. The (forward) primer from one side of the insert will read the sequence in one direction, while the sequence of the complementary strand will be obtained using the (reverse) primer from the other side of the insert. We can extend the length of sequence determined by using the results from the first sequencing experiment to design a new primer that would

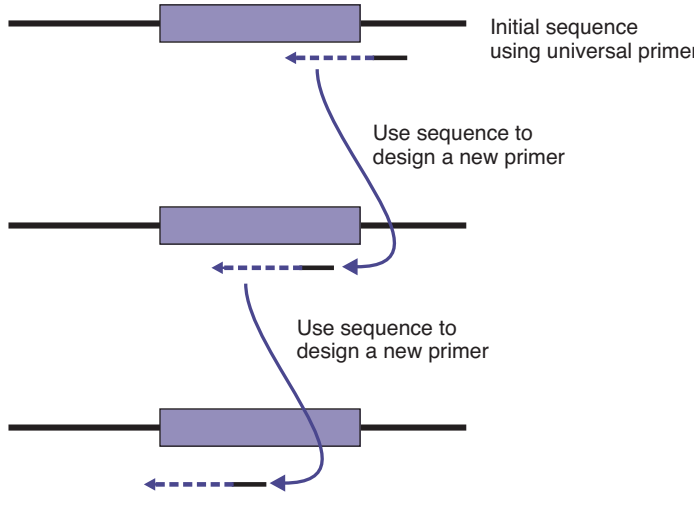


Figure 5.4 Extending a sequence by primer walking.

start synthesis further along the insert. That will produce a further length of sequence, which can then be used to design a third primer, and so on. This procedure is effective for relatively short sequences but becomes excessively tedious if long stretches of DNA are to be sequenced. It can, however, be a useful strategy for finishing larger sequences when other approaches have left short gaps (see below).

5.1.4 Shotgun sequencing; contig assembly

For sequencing a longer fragment – say a cloned fragment of 10 kb – the best procedure is often to split it up into smaller fragments, each of which is a suitable size for sequencing in full. This is like producing a gene library. The insert from your recombinant vector is fragmented and cloned using a suitable vector. The bacteriophage vector M13 was originally useful for this purpose, since it produced single-stranded versions of the fragments, which gave cleaner results. However, improvements in the technology mean that equally good sequence results can be obtained with double-stranded templates.

Note that it is *essential* to have overlapping fragments in this library, so mechanical fragmentation is often chosen (see Chapter 3). You then pick recombinant clones at random from this mini-library and sequence each one – an approach known as *shotgun* sequencing (Figure 5.5). At the start, you will have no idea where each bit of sequence comes from in the original

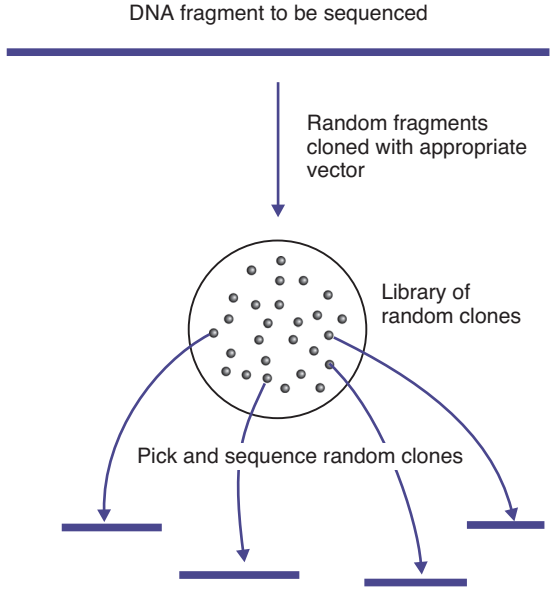


Figure 5.5 Shotgun cloning and sequencing.

fragment – nor even which strand of the original it is derived from. But once you have a number of such fragments sequenced, you can start to use computer algorithms to compare each bit of sequence to all the others. The computer will find any overlaps between the fragments, including comparing the complementary strand in case you have two fragments that overlap but are derived from different strands. Where there is an overlap, the sequences of those fragments will be joined together to form a *contig*. As the project progresses, each contig will grow longer, and will then start to overlap other contigs. So the contigs themselves are joined together until eventually you have one single contig covering the whole of the original piece of DNA, and all the sequenced fragments fit into that single contig (Figure 5.6).

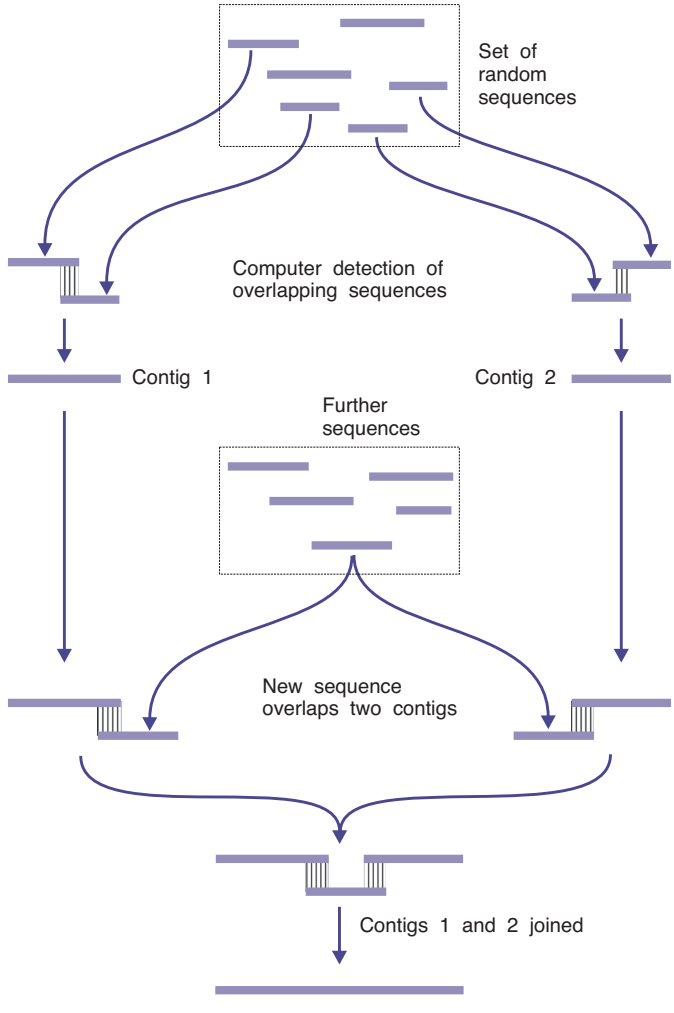


Figure 5.6 Contig formation and joining.

Of course, while initially each piece of sequence is new information, as you proceed, the sequencing becomes less and less productive. Towards the end, nearly all the clones sequenced will be completely contained within one of the existing contigs. Nevertheless, provided all parts of the fragment are equally represented in the library, it is quicker to continue with a shotgun approach rather than try to screen the library for the missing pieces. Sometimes, however, parts of the original fragment are under-represented in the library. For example, they may be hard to clone for some reason. If this is the case, you will have a gap in your sequence that can be difficult to fill with a shotgun approach. If you have reason to believe that the gap is quite small, then you can use primer walking as described above to bridge the gap. Alternatively you can use the sequences you have determined to design PCR primers that will enable you to amplify a DNA fragment that spans the gap, so it can then be sequenced. Other approaches for bridging gaps in longer sequences, such as genome sequences, are considered in Chapter 8.

For any sequencing project, whatever the size of the DNA to be sequenced, and whatever strategy is used, a diminishing returns effect is very marked. You may get 90% of the sequence accurately determined quite quickly; the next 9% may take as long again; and then the next 0.9% a similar length of time. It is usually necessary to set some limits to this: how complete and how accurate do you need the sequence to be? Will 90% do, or must you have 99% or 99.9%? Without some sort of compromise you can be drawn into a costly exercise trying to determine whether one difficult base in a sequence of a million bases is G or C.

5.2 Databank entries and annotation

Once we have determined a DNA sequence, it should be made publicly and freely accessible by submitting it to a databank (EMBL, GenBank or DDBJ; see Box 5.1). This is done electronically via the internet, and the databanks have web-based procedures that make this submission a simple automatic process. Box 5.1 lists the web addresses for the databanks, and for the other databases and tools that are referred to in this and subsequent chapters. In practice, the databanks can be considered as one and the same, as they share their information on a daily basis, so you can just choose the one whose interface you prefer. Due to this data sharing, it is only necessary to submit your sequence to (or to search in) one of them. An example of an EMBL databank entry for an individual sequence (human cDNA coding for thymidylate synthetase) is shown in Figure 5.7 (you may find when you retrieve a sequence that you get a more attractive display than this, but the information is the same). An increasing proportion of databank entries relate to whole genome sequences rather than individual genes like this one. We will consider the annotation and analysis of genome sequence data further in Chapter 8.

Box 5.1 A selection of available on-line resources

Resource	Description	Web address
Nucleotide sequence databanks		
DDBJ (DNA Data Bank of Japan)		http://www.ddbj.nig.ac.jp/
GenBank		http://www.ncbi.nlm.nih.gov/Genbank/
European Molecular Biology Laboratory (EMBL)		http://www.ebi.ac.uk/embl/
Other databases		
UniProt Knowledgebase (UniProtKB)	Merged protein sequence data from Swiss-Prot, TrEMBL and PIR-PSD	http://www.ebi.uniprot.org/
InterPro	Integrated resource for protein families, domains and functional sites	http://www.ebi.ac.uk/interpro/
PROSITE	Database of protein families and domains	http://www.expasy.org/prosite/
Pfam	Multiple sequence alignments and HMMs for protein domains	http://www.sanger.ac.uk/resources/databases/pfam.html
PRINTS	Compendium of protein fingerprints (motifs characterizing protein families)	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/
dbSNP	Single nucleotide polymorphisms (SNPs) and short deletion and insertion polymorphisms	http://www.ncbi.nlm.nih.gov/SNP/
2D protein gels		http://www.mpiib-berlin.mpg.de/2D-PAGE/>
German Resource Center for Genome Research	Clone libraries, expressed sequence tags (ESTs), plus a variety of tools	http://www.imagenes-bio.de/
Online Mendelian Inheritance in Man (OMIM)	Compendium of human genes and genetic phenotypes	http://www.omim.org/
International HapMap Project	Haplotype map of human genome	http://hapmap.ncbi.nlm.nih.gov/

(*Continued*)

Box 5.1 (Continued)

Resource	Description	Web address
Tools		
A wide variety of tools and databases are available at:		
NCBI		http://www.ncbi.nlm.nih.gov/Tools/
EBI (European Bioinformatics Institute)		http://www.ebi.ac.uk/
Sanger Institute		http://www.sanger.ac.uk/
Some specific tools/sites:		
BLAST	Basic Local Alignment Search Tool	http://www.ebi.ac.uk/blast2/ http://www.ncbi.nlm.nih.gov/BLAST/
FASTA	Alternative to BLAST for database comparisons	http://www.ebi.ac.uk/fasta33/
Ensembl Genome Browser	Browse mammalian and other eukaryotic genome sequences	http://www.ebi.ac.uk/ensembl/
Artemis	Genome viewer and annotation tool: visualization of sequence features	http://www.sanger.ac.uk/resources/software/artemis/
ACT (Artemis Comparison Tool)	Downloadable genome sequence comparison viewer	http://www.sanger.ac.uk/Software/ACT/
WebACT	On-line use of ACT for prokaryotic genome sequences	http://www.webact.org/WebACT/home
ProtScale	Computes and represents a variety of profiles on a selected protein	http://www.expasy.org/tools/protscale.html
gMap	Genome sequence comparison	http://www.ncbi.nlm.nih.gov/sutils/gmap.cgi
ClustalX2	Downloadable version of sequence alignment program	http://www.clustal.org/

Entries in GenBank look slightly different, but contain the same information, and you should have little difficulty in switching from one to the other. Both are computer-readable, provided that the software has been set up to recognize the format.

The notes that we have added to the annotation should mainly be self-explanatory, but some aspects need additional comment. The *accession*

Some elements of the annotation have been omitted, and sequence data is truncated

Accession number

```
ID   X02308; SV 1; linear; mRNA; STD; HUM; 1536 BP.
AC   X02308;
```

Definition

```
DT   07-NOV-1985 (Rel. 07, Created)
DT   12-SEP-1993 (Rel. 36, Last updated, Version 2)
DE   Human mRNA for thymidylate synthase (EC 2.1.1.45)
```

Keywords

```
KW   inverted repeat; synthetase; tandem repeat.
```

Species and Classification

```
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae;Homo.
```

Reference

```
RP   1-1536
RX   PUBMED; 2987839.
RA   Takeishi K., Kaneda S., Ayusawa D., Shimizu K., Gotoh O., Seno T.;
RT   "Nucleotide sequence of a functional cDNA for human thymidylate synthase";
RL   Nucleic Acids Res. 13(6):2035-2043(1985).
```

Link to the paper

Features table / Coding sequence / Translated sequence

```
FH   Key             Location/Qualifiers
FT   source          1..1536
FT                   /organism="Homo sapiens"
FT                   /mol_type="mRNA"
FT                   /db_xref="taxon:9606"
FT   misc_feature    14..103
FT                   /note="triple tandemly repeated elements"
FT   misc_feature    35..103
FT                   /note="pot. stem-loop structure"
FT   CDS             106..1047
FT                   /note="thymidylate synthase (aa 1-313)"
FT                   /db_xref="GDB:120465"
FT                   /db_xref="GOA:P04818"
FT                   /db_xref="HGNC:12441"
FT                   /db_xref="InterPro:IPR000398"
FT                   /db_xref="PDB:1HVY"
FT                   /db_xref="PDB:1HW3"
FT                   /db_xref="UniProtKB/Swiss-Prot:P04818"
FT                   /protein_id="CAA26178.1"
FT                   /translation="MPVAGSELPRRPLPPAAQ ..."
FT   misc_feature    1231..1236
FT                   /note="pot. polyadenylation signal"
FT   misc_feature    1519..1524
FT                   /note="pot. polyadenylation signal"
FT   polyA_site      1536..1536
FT                   /note="polyadenylation site"
```

link to taxonomy

links to entries in other databases

DNA sequence

```
SQ   Sequence 1536 BP; 390 A; 369 C; 399 G; 378 T; 0 other;
     gggggggggg ggaccacttg gcctgcctcc gtcccgccgc ...
```

Figure 5.7 Sequence annotation (EMBL).

number (AC) is important as it is the most convenient way of retrieving a specific sequence from the databank, and the one that is referred to in publications; the accession number is unique to that particular sequence, and is the same in each of the different databanks for any given sequence. This is because it refers to a particular sequence submission, rather than to a specific gene or locus. There may therefore be several entries, with different accession numbers, relating to essentially the same sequence (e.g., from different

strains or variants of the same species). If you want to look for different versions of a specific gene you may therefore need to search the databank by gene name, or alternatively use a known sequence and search for similarities (e.g., using BLAST, see below).

In addition to the sequence itself, each entry contains a considerable amount of *annotation* that makes the sequence information much more useful. Some of the annotation is supplied with the submitted sequence, while other annotation is added by the databanks. Obviously, this must include information about the source of the sequenced DNA. In addition, the identification and extent of any open reading frame is a basic requirement (together with the computer prediction of the protein sequence), and intron/exon boundaries (if applicable). Further information on expression signals, motifs, structural elements and so on will enhance the value of the entry, as will identification of the presumed or actual function of the gene. This information goes into the *features table* (lines starting with FT), and can be read by programs such as Artemis (see Chapter 8) to produce a visual display of the features of the sequence. How some of these features are identified is discussed below.

In addition to DNA sequences, there are databanks of protein sequences (see Box 5.1). These databanks, including the most popular one known as SWISS-PROT, have been combined into a single databank, UniProt. You will see in Figure 5.7 that a reference is provided that identifies the corresponding entry in the UniProt database; the link will take you directly to that entry, which is shown in Figure 5.8. In addition to annotation that is similar to that described previously for DNA sequence entries, this entry contains information about the 3D structure of the protein, and cross-references to a number of other databases, especially Pfam and Prosite, which will be considered later.

There are two types of protein sequence information, physical and predicted. The former are derived from direct protein sequencing, but the majority (especially those arising from genome sequencing projects) are of the latter variety, being derived by computer translation of DNA sequences. Some care is needed when using computer-generated protein sequences, especially as there may be no direct evidence that this protein actually exists. Furthermore, if the predicted protein is eukaryotic in origin, then it may be based on incorrectly predicted intron/exon boundaries, or alternative splicing may create different exon combinations, or the identification of the start site may be incorrect, or a variety of other factors (including post-translational modification or cleavage) may result in the protein within the cell being substantially different from the primary translation product. In particular, a base missing or incorrectly inserted will result in a shift of the reading frame, and the protein sequence beyond that point will bear no relationship to the real product.

General information

Entry name	**TYSY_HUMAN**
Accession number	**P04818**
Integrated	13-AUG-1987, UniProtKB/Swiss-Prot.

Description and origin of the Protein

Description	Full=Thymidylate synthase; Short=TS;TSase EC=2.1.1.45;
Organism source	Homo sapiens (Human).
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

References

[1] Takeishi,K., Kaneda,S., Ayusawa,D., Shimizu,K., Gotoh,O., Seno,T., **Nucleotide sequence of a functional cDNA for human thymidylate synthase.** (1985) *Nucleic Acids Res.* **13**:2035-2043

[2] Kaneda,S., Nalbantoglu,J., Takeishi,K., Shimizu,K., Gotoh,O., Seno,T., Ayusawa,D., **Structural and functional analysis of the human thymidylate synthase gene.** (1990) *J. Biol. Chem.* **265**:20277-20284

Comments

CATALYTIC ACTIVITY	5,10-methylenetetrahydrofolate + dUMP = dihydrofolate + dTMP.
PATHWAY	Pyrimidine metabolism; dTTP biosynthesis.
SUBUNIT	Homodimer.
SIMILARITY	Belongs to the thymidylate synthase family.

Database cross-references

EMBL	X02308; CAA26178.1; -; mRNA.
	D00596; BAA00472.1; -; Genomic_DNA.
PIR	A23047; YXHUT.
PDB	1HVY; X-ray; 1.90 A; A/B/C/D=26-313.
	1HW3; X-ray; 2.00 A; A=1-313.
Ensembl	ENST00000323274; ENSP00000315644; ENSG00000176890.
KEGG	hsa:7298; -.
H-InvDB	HIX0017793; -.
MIM	188350; gene.
GO	GO:0004799; F:thymidylate synthase activity; TAS:Reactome.
Pfam	PF00303; Thymidylat_synt; 1.
PRINTS	PR00108; THYMDSNTHASE.
PROSITE	PS00091; THYMIDYLATE_SYNTHASE; 1.

Protein Existence

1: Evidence at protein level;

Features

Key	Begin	End	Length	Description
CHAIN	2	313	312	Thymidylate synthase.
ACT_SITE	195	195	1	
MOD_RES	114	114	1	Phosphoserine.
HELIX	30	43	14	
STRAND	45	47	3	
TURN	75	78	4	

Sequence information

Length: **313 aa**, molecular weight: **35716 Da**

MPVAGSELPRRPLPPAAQERDAEPRPPHGELQYLGQIQHILRCGVRKDDRTGTGTLSVFG.....

Figure 5.8 Protein sequence annotation (UniProt).

5.3 Sequence analysis

5.3.1 Identification of coding region

Assuming we are dealing with a protein-coding gene, the first task is to identify precisely the location of the region that is translated into protein. As we are starting with a cloned fragment, this is likely to be relatively straightforward. If the gene is bacterial in origin, we don't have to worry about introns; if it is a eukaryotic gene, we will assume that the clone in question is cDNA, so again we can ignore introns. The problem of identifying protein-coding regions in genome sequences, including the prediction of intron-exon boundaries in eukaryotes, is more complex, and will be dealt with in Chapter 8.

Since the mRNA can be translated into protein in any of three *reading frames*, we need to ascertain which is actually used. Because at this stage we may not know in which direction the DNA is transcribed, there are actually six possible reading frames altogether – three in one direction, and three others in the other direction. Fortunately, we are provided with a clue. In those reading frames that are not used for translation, there are usually frequent stop codons. The same applies to any regions of the sequence that do not code for proteins. On the other hand, any region of the sequence that does code for a protein must, obviously, be free of stop codons (until the stop codon that signifies the end of the protein). Such a sequence, without any stop codons, is known as an *open reading frame* (ORF). So we can use a computer to search the sequence for stop codons, and an ORF (without stop codons) can be predicted to code for a protein, in that reading frame. Having identified an ORF, it is then straightforward to translate the DNA sequence into a protein sequence.

We would also look for a start codon. We can refine the concept of an ORF to mean the distance between a start codon and the first stop codon in the same reading frame. This is not actually as straightforward as it sounds. Although we regard ATG (using DNA rather than RNA nomenclature) as the 'normal' start codon, many organisms sometimes use other start codons as well, such as GTG, TTG or CTG – in some cases, especially in bacteria, these other start codons can be quite frequent. The uncertainty as to whether one of these codons is a real start codon explains why the above description starts with the region between two stop codons, rather than looking for start codons. Once you have identified an open reading frame, between two stop codons, this narrows down the search for the start codon. The evidence that a potential start codon is actually where protein synthesis starts is considerably strengthened by the presence of an adjacent ribosome-binding site (in bacteria) or a Kozak sequence (in eukaryotes) – see Chapter 7.

5.3.2 Expression signals

An ORF is of no significance unless that region of the DNA is actually transcribed, and in the correct direction. So the identification of potential transcription start (and stop) sites is important information relating to our sequenced gene. Although the identification of such signals by analysing the sequence is not completely reliable, we can often achieve some indication of how the gene is expressed. In most bacteria, a high proportion of genes are transcribed by RNA polymerase recognizing promoter sites that have two relatively conserved consensus regions: the −35 region (i.e., a region centred at 35 bases before the start of transcription) and a −10 region. The consensus sequences for these two regions are TTGACA and TATAAT respectively (see Chapter 1). But these only represent a consensus. Very few promoters have exactly that sequence at either position, and the distance separating them can vary by a few bases as well. However, if we can find sequences resembling a consensus promoter, in the right place, that is sufficient for us to label it as a putative promoter. We may not find one, for several reasons. The gene may be transcribed as part of an operon, in which case the promoter may be a long way from the sequence we are looking at. In addition, the specificity of bacterial RNA polymerase can be changed by substitution of a different sigma factor, enabling it to recognize promoters with a markedly different structure. Some of these are well characterized and can also be identified.

In eukaryotes, as usual, the situation is even more complicated. The binding of RNA polymerase II is mediated by a number of canonical elements (often known as 'boxes' or 'response elements') with similarity of structure, including the TATA-box, GC-box and CAAT-box. However, the spacing of these elements is not always consistent. Moreover, there is an element of species variation, so the parameters for searching for a transcription start site in a fruit fly are different from those needed to make the corresponding prediction in a human sequence. A number of websites are available for the prediction of transcription start sites in model organisms. These take advantage of neural networks that use information from known promoters in the same organism to identify weighted consensus sequences to search for.

In addition to searching for promoter sequences (i.e., sites at which RNA polymerase may bind directly), we can look for sites at which regulatory proteins can attach to the DNA to repress or activate transcription. Many of these proteins have quite well-conserved, and characterized, recognition sequences. For example, in most bacteria iron uptake is regulated by proteins belonging to one of two families, related to either Fur (the Ferric Uptake Regulator protein of *E. coli*) or DtxR (the diphtheria toxin repressor of *Corynebacterium diphtheriae*). These proteins, in the presence of Fe^{2+}, bind to specific DNA sites and repress transcription of the adjacent genes.

Identification of a site such as the so-called 'Fur box', to which Fur binds, therefore provides an indication that the associated gene will be repressed in the presence of an adequate supply of iron, which in turn suggests that the function of that gene may be connected with iron uptake. There are a number of other known 'boxes', to which different regulatory proteins will bind, in both prokaryotes and eukaryotes – for example, the E-box (for enhancer) has been associated with an expression pattern following a circadian (24-hour) rhythm. It is possible to screen DNA sequences for each of these boxes, providing evidence not only relating to the regulation of the relevant genes, but also a clue as to their possible function; the methods for searching for such binding sites, and other structural features of a protein, are considered later. As with putative promoter sites, it is essential to view computer predictions as just a first stage in an investigation. They provide clues that need to be confirmed by more direct, experimental evidence.

It is important to keep in mind that the promoter and its associated regulatory sequences may extend in the 3′-direction beyond the transcription start site. Many of the latter are often found in the first intron, which is often very large, in eukaryotic genes, and exert their function in the double-stranded DNA rather than in the single-stranded RNA transcript.

5.4 Sequence comparisons

5.4.1 DNA sequences

There is a vast and rapidly growing amount of sequence data available in the databanks. How can we find out if our sequence is the same as (or very similar to) some other sequence that has already been determined?

We can start by considering the simplest approach. Suppose we have 1 kb of sequence data (still assuming the absence of introns), and we want to compare it with another specific sequence. We can put our sequence (the *query sequence*) alongside the second sequence, and get the computer to count the number of bases that match; this would determine the *identity* of the two sequences. But the two sequence fragments may be of different lengths, and may not start at the same place, which would considerably complicate our analysis. So we (or rather the computer) must slide our query sequence along the sequence it is being compared with and see where the best match comes. However, there is another problem. Our sequence may have one or more gaps in it, compared to the second sequence (or vice versa). This can happen in several ways: there may be a genuine difference between the sequences (e.g., if we are comparing genes from different species), or one of the sequences might be wrong. Addition or removal of even a single base would render this simplistic approach non-viable. For example, our sequence may read at one point GGACT, while at the corresponding point in the databank

sequence, which otherwise matches perfectly, the sequence is GG**G**ACT. This is a very small difference, but the consequence would be that part of our sequence would line up perfectly, while the rest would not match at all (or more strictly it would match just as well as any two random bits of DNA). In order to accommodate this problem, we must allow the computer to introduce gaps into one or both sequences, so that it can come up with the best possible match between the two sequences.

But we cannot allow the computer indefinite licence to introduce as many gaps as it likes, otherwise it would produce perfect matches between any two sequences, just by repeatedly sliding them apart until it found the next base that matched. The algorithm used by the computer to align the two sequences therefore incorporates a *gap penalty*; every time a gap is introduced, into either sequence, there is a reduction in the score. The gap penalty can be set at different levels, including different penalties for different lengths of gaps, although the software usually has a default value so we do not have to think about it.

There are several methods available for comparing DNA sequences, and some of them employ algorithms that essentially work as described above (although the description is highly simplified!). However, such algorithms require a lot of computer power for longer sequences, if they work literally as described: As the computer slides the two sequences along base by base, at each position it has to calculate a score with all possible combinations of gaps – even for a sequence of say 1 kb, there is a very large number of possible combinations. And usually we do not want to compare our query sequence with just a single other sequence, but all sequences in the databank, which would require even more computational power.

Other programs therefore employ different algorithms that are designed to speed up the process. One approach is basically to split our query sequence into small fragments (*words*), using these words to decide how the two sequences will align best and then computing a score for the optimal alignment, including allowing for gaps. This is more or less the basis for one of the common search programs known as FASTA.

A somewhat different approach is used by one of the most widely used programs, known as BLAST (Basic Local Alignment Search Tool). At a highly simplified level, this works by finding very short matches (segment pairs) and then extending that match outwards until the score falls below a set value. Each matched pair of sequences above a certain length is then stored and reported (as High-scoring Segment Pairs, or HSPs), starting with those with the highest score. Figure 5.9 shows selected results of a database search using BLASTN (the version of BLAST used with DNA sequences), using the human thymidylate synthetase cDNA (see Figure 5.7) as the *query sequence*. (Note that this is an edited version of the output, derived from hundreds of reported matches.) The first column identifies the sequence that matches, by

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
NC_006601.2	Canis familiaris	1380	1380	96%	0.0	80%
NC_006485.2	Pan troglodytes	1079	2659	99%	0.0	100%
AC_000032.1	Mus musculus	1023	1023	56%	0.0	85%
AC_000071.1	Rattus norvegicus	392	392	56%	6e-105	70%
NC_001348.1	Human herpesvirus 3	288	288	51%	2e-73	68%
NT_033779.4	Drosophila melanogaster	277	277	52%	3e-70	68%
NC_007325.4	Bos taurus	253	1300	79%	4e-63	96%
NC_006034.2	Candida glabrata	194	194	40%	3e-45	66%
NC_006089.2	Gallus gallus	181	825	53%	2e-41	88%
NT_078267.5	Anopheles gambiae	161	161	22%	2e-35	71%
NC_007247.1	Leishmania major	147	147	47%	4e-31	65%
NC_001147.5	Saccharomyces cerevisiae	132	132	40%	9e-27	64%
NC_003062.2	Agrobacterium tumefaciens	131	177	15%	3e-26	88%
NC_012850.1	Rhizobium leguminosarum	129	175	16%	1e-25	79%
NC_010742.1	Brucella abortus S19	129	173	15%	1e-25	84%
NC_003075.7	Arabidopsis thaliana	118	166	20%	2e-22	88%
NC_003279.6	Caenorhabditis elegans	118	118	29%	2e-22	67%
NC_011283.1	Klebsiella pneumoniae	102	102	9%	1e-17	75%
NC_009910.1	Plasmodium vivax	100	100	26%	5e-17	65%
NC_014479.1	Bacillus subtilis subsp. spizizenii str. W23 |	96.9	96.9	14%	6e-16	70%
NC_011896.1	Mycobacterium leprae	95.1	95.1	12%	2e-15	71%
NC_013008.1	Escherichia coli O157:H7	93.3	93.3	9%	8e-15	73%
NC_007606.1	Shigella dysenteriae	93.3	141	15%	8e-15	73%
NC_011035.1	Neisseria gonorrhoeae	89.7	89.7	14%	9e-14	68%
NC_013016.1	Neisseria meningitidis	87.8	132	17%	3e-13	84%
NC_012488.1	Listeria monocytogenes	86.0	86.0	16%	1e-12	67%
NT_166529.1	Aspergillus niger	75.2	75.2	7%	2e-09	75%
NC_008405.2	Oryza sativa Japonica	73.4	196	18%	7e-09	82%
NC_014029.1	Yersinia pestis Z176003	68.0	68.0	8%	3e-07	72%
NC_002952.2	Staphylococcus aureus MRSA	62.6	62.6	16%	1e-05	66%

Notes
Score = a measure of the similarity of the target and query sequences
E value = the probability of this match occurring by chance.
The highlighted entry is referred to in subsequent analyses

Figure 5.9 Comparison of DNA sequences: selected output from a BLASTN search of databanks with human thymidylate synthase cDNA as the query sequence.

accession number, and in the on-line output it is an active link enabling you to obtain any of these sequences. The next column shows the description of the sequence (truncated in this example). Although the order of entries does not correspond very closely to taxonomic relationships (you would need much more data to do that), the general trend is for the most similar sequences to be from various mammals, followed by other eukaryotes, with bacteria being further down the list.

The most significant column is that headed ‘E value’, which provides an estimate of the probability of each match occurring by chance; thus a low E number, with a high negative logarithm, indicates a strong match. The entries are arranged in descending order, so that the most significant matches are at the top. Note that while for most biological experiments we are happy with a significance probability of $P < 0.01$ or 0.05, here we see probabilities many

```
>NC_014479.1     Bacillus subtilis subsp. spizizenii str. W23 |

Score = 96.9 bits (106),  Expect = 8e-24
Identities = 156/222 (70%), Gaps = 2/222 (1%)
Strand=Plus/Minus

Query  673  ATGGCGCTGCCTCCATGCCATGCCCTCTGCCAGTTCTATGTGGTGAACAGTGAGCTGTCC  732
            ||||||  ||||  ||  ||||||   |||  |  |||  |||||  |||     ||  |   ||||||||
Sbjct  375  ATGGCGTTGCCGCCGTGCCATTGCCTGTTCCAATTCTACGTGTCTGACGGCAAGCTGTCC  316

Query  733  TGCCAGCTGTACCAGAGATCGGGAGACATGGGCCTCGGTGTGCCTTTCAACATCGCCAGC  792
            ||  ||||||||  ||   |  ||  |   ||   |    ||||||||||||  ||  ||  ||  ||
Sbjct  315  TGTCAGCTGTATCAACGTTCTGCTGATGTTTTCCTCGGTGTGCCGTTTAATATTGCTTCT  256

Query  793  TACGCCCTGCTCACGTACATGATTGCGCACATCAC-GGGCCTGAAGCCAGGTGACTTTAT  851
            ||  ||||||  |  ||     ||||||||  ||   |  ||  ||||  ||||  ||  ||  ||  |||||
Sbjct  255  TATGCCCTGTTAACCATGATGATTGCTCATGTGACTGGGCTTGAA-CCGGGCGAGTTTAT  197

Query  852  ACACACTTTGGGAGATGCACATATTTACCTGAATCACATCGA  893
              ||  ||  ||  ||  ||||   ||||||||  |   |||||  ||  ||
Sbjct  196  TCATACATTTGGCGATGTTCATATTTATCAAAATCATATTGA  155
```

Notes.
Expect = 8e-24. This shows the probability of the score, for this part of the sequences, occurring by chance. This is a low value, suggesting that this match is highly significant, although the sequences are clearly far from identical
Strand=Plus/Minus. This indicates that the match shown contains the plus strand of the query sequence (human cDNA) and the minus (complementary) strand of the subject (*E. coli*)
See Fig 5-9 for additional notes.

Figure 5.10 Results of a database search using BLASTN (DNA search). Example of a reported match, from the search results shown in Figure 5.9.

orders of magnitude more significant than this – this is a result of the large number of sequences in the database that the program compares before it achieves a match.

The BLAST search will also return the pairwise alignment of the query sequence with each of the matches, although this is not necessarily the optimal alignment (see Section 5.4.3 below). As an example, the alignment with the *E. coli thyA* gene is shown in Figure 5.10. Although the E value indicates that this is a highly significant match, you will see that the sequences are far from identical (70% identity). In the next section, we will look at the match between these two sequences at the protein level.

5.4.2 Protein sequence comparisons

DNA sequence comparisons are inherently noisy, that is they are riddled with false positive matches. If you just take two random DNA sequences and put them side by side, there is a 25% chance of a match at any specific position. Sometimes you have to use DNA sequence comparisons, for example, if you are working with a non-coding sequence, but generally for searching databanks with an unknown sequence you will get much cleaner

results with a protein sequence. A further advantage of searching at the protein level is that, owing to the degeneracy of the genetic code, related proteins are generally more conserved than the genes encoding them. Note that all the predicted coding sequences in the DNA databanks are translated and incorporated in the protein sequence databank. Alternatively, using a variant of BLAST (BLASTX), you can input a DNA query sequence and the computer will translate it in all six reading frames (three in each direction) and compare all six to the sequences in a protein databank. Conversely, TBLASTN will compare a protein query to translated DNA sequences from the databank. Or you can even use TBLASTX, which compares a translated DNA sequence with translated sequences from a nucleotide databank. However, in this case, we will use BLASTP, for comparing a protein sequence query to a protein databank.

The initial basis for protein sequence comparisons is just the same as described above for nucleotide sequence comparisons, and the algorithms used are fundamentally the same. The major difference arises from the fact that you have 20 amino acids instead of four bases. This does not have to change anything – you could treat the 20 amino acids as all different from one another, and require a perfect match (*identity*) at any position. But the computer offers you a more sophisticated scoring system, in which some pairs of amino acids are regarded as more different than others. This scoring system is based on a matrix of all the possible amino acid pairings, ranging from some pairs that score almost the same as a perfect match through to other pairs that are regarded as completely different. All of the scores in the matrix are used to calculate the overall score for the match, although the alignment presented will only mark matches above a selected cut-off score.

There are several such matrices to choose from, although here again the package you use will probably have a default in case you do not want to make a choice. But even if you accept the default, you should have some understanding of the basis of the matrix that is used. It is a common misconception that the alignment is calculated solely on the basis of the similarity between the amino acids. For example, a biochemist might group valine, leucine and isoleucine together as they have large non-polar (hydrophobic) side chains; phenylalanine, tyrosine and tryptophan (aromatic); glutamate with aspartate (acidic, negatively charged); lysine and arginine (basic, positively charged); and so on. A change from one amino acid to another within the same group (say lysine to arginine) would be expected to have less effect on the structure and function of the protein than a change between groups (say lysine to phenylalanine) – and so you would expect this similarity to be reflected in the scoring system.

That is indeed true, and such similarities are important, but the matrices are more sophisticated than that. Matrices such as the Dayhoff Point Accepted Mutation (PAM) matrices are derived empirically from

comparisons of related proteins from different sources, and the score is determined from the frequency with which specific changes in amino acid sequence occur. This partially reflects the similarity between the amino acids, but is also influenced by the nature of the genetic code. Where a pair of amino acids can be interchanged by only a single base change in the DNA, this will occur more frequently than a change that requires two or three base changes. There are a series of PAM matrices with different values; one example (the Dayhoff PAM250 matrix) is shown in Figure 5.11. If you are not familiar with the one-letter amino acid notation, see Box 5.2. Values of 0 indicate neutral changes, while increasing positive or negative values indicate increasingly acceptable or unacceptable mutations, respectively. Note that in some cases, a mismatch can give a positive score – and in particular a mismatch between tyrosine (Y) and phenylalanine (F) actually gives a higher score than a perfect match of most other amino acids. (This is because tyrosine and phenylalanine are relatively infrequent amino acids, so a Y-F match, scored 7 in this matrix, is less likely to occur by chance than is the case for, say, an alanine-alanine match, with a score of 2.) There are different versions of the PAM matrices, which vary in their sensitivity. So the PAM250 matrix is suitable for distantly related sequences, while others such as PAM40 are better for more similar sequences.

Although the PAM matrices still perform well, they are no longer regarded as the best matrices for database searching. Most implementations of programs such as BLASTP use, as the default, an alternative family of matrices known as BLOSUM ('blossom'), or BLOcks SUbstitution Matrix. These are derived in a somewhat different way, which makes them more suitable for procedures based on local alignments. One example, BLOSUM45, is shown in Figure 5.11. The numbering of the matrix title works in the opposite direction from the PAM matrices, so the commonly used BLOSUM62 matrix is less sensitive for weak alignments than BLOSUM45. Yet other matrices are available. For example, some implementations use the Gonnet matrix as the default. Generally speaking, at the start the best option is to use the default matrix, and only start to explore different matrices if you do not find the matches that you are looking for.

Figure 5.12 shows selected results from the output from a BLASTP search of the UniProt database, using, as the query sequence, the amino acid sequence of human thymidylate synthetase, whose gene we employed for the nucleic acid search above (Figure 5.9). The general structure of the table is the same as before. Note that although almost all of these proteins are labelled as thymidylate synthetases (or as bifunctional DHFR-TS enzymes), many are derived from genome sequences and are labelled as such solely because of their similarity to other thymidylate synthetases, without direct experimental evidence. With these examples, it is probably correct, but the potential for a circular argument has to be remembered.

Dayhoff PAM250 MATRIX

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

BLOSUM45 matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	-1	4																		
T	-1	2	5																	
P	-4	-1	-1	9																
A	-1	1	0	-1	5															
G	-3	0	-2	-2	0	7														
N	-2	1	0	-2	-1	0	6													
D	-3	0	-1	-1	-2	-1	2	7												
E	-3	0	-1	0	-1	-2	0	2	6											
Q	-3	0	-1	-1	-1	-2	0	0	2	6										
H	-3	-1	-2	-2	-2	-2	1	0	0	1	10									
R	-3	-1	-1	-2	-2	-2	0	-1	0	1	0	7								
K	-3	-1	-1	-1	-1	-2	0	0	1	1	-1	3	5							
M	-2	-2	-1	-2	-1	-2	-2	-3	-2	0	0	-1	-1	6						
I	-3	-2	-1	-2	-1	-4	-2	-4	-3	-2	-3	-3	-3	2	5					
L	-2	-3	-1	-3	-1	-3	-3	-3	-2	-2	-2	-2	-3	2	2	5				
V	-1	-1	0	-3	0	-3	-3	-3	-3	-3	-3	-2	-2	1	3	1	5			
F	-2	-2	-1	-3	-2	-3	-2	-4	-3	-4	-2	-2	-3	0	0	1	0	8		
Y	-3	-2	-1	-3	-2	-3	-2	-2	-2	-1	2	-1	-1	0	0	0	-1	3	8	
W	-5	-4	-3	-3	-2	-2	-4	-4	-3	-2	-3	-2	-2	-2	-2	-2	-3	1	3	15

Figure 5.11 Examples of protein substitution matrices.

Box 5.2 Amino acid notations

Amino acid	Three-letter notation	One-letter notation
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartate	Asp	D
Cysteine	Cys	C
Glutamate	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

As before, the program will also return a pairwise alignment of the sequences. If you compare the protein alignment, using the *B. subtilis* enzyme, shown in Figure 5.13, with the alignment of the corresponding DNA sequences in Figure 5.10, you will see that the proteins match better than the DNA sequences. You may wonder why the percentage identity (53%) is lower than that for the DNA comparison (70%), when we said that amino acid sequences tend to be more conserved. The reason lies in the noisiness of the DNA comparison, where many of the matches will be due to chance.

Accession	Description	Max score	Total score	Query coverage	E value
NP_001182436.1	thymidylate synthase [Macaca mulatta]	613	613	100%	1e-173
AAH20139.1	Tyms protein [Mus musculus]	555	555	100%	3e-156
NP_062052.1	thymidylate synthase [Rattus norvegicus]	551	551	100%	6e-155
XP_533309.1	PREDICTED: similar to Thymidylate synthase (TS) (TSase) [Canis familiaris]	546	546	95%	1e-153
XP_001150304.1	PREDICTED: thymidylate synthetase isoform 2 [Pan troglodytes]	540	540	100%	1e-151
DAA15786.1	thymidylate synthase [Bos taurus]	528	528	100%	4e-148
NP_001096659.1	thymidylate synthase [Xenopus laevis]	500	500	100%	9e-140
AAH65845.1	Tyms protein [Danio rerio]	484	484	91%	7e-135
NP_040136.1	thymidylate synthase [Human herpesvirus 3]	430	430	91%	1e-118
XP_001663452.1	bifunctional dihydrofolate reductase-thymidylate synthase [Aedes aegypti]	402	402	91%	3e-110
AAC27622.1	thymidylate synthase [Drosophila melanogaster]	399	399	98%	4e-109
AAC97508.1	thymidylate synthase [Caenorhabditis elegans]	395	395	91%	5e-108
NP_001104916.1	bifunctional dihydrofolate reductase-thymidylate synthase [Zea mays]	393	393	91%	2e-107
NP_195183.2	dihydrofolate reductase/ thymidylate synthase [Arabidopsis thaliana]	393	393	91%	2e-107
CAD38046.1	dihydrofolate reductase-thymidylate synthase [Pisum sativum]	390	390	91%	1e-106
EEE53150.1	hypothetical protein OsJ_35972 [Oryza sativa Japonica Group]	390	390	91%	1e-106
ABG43014.1	thymidilate synthase [Candida glabrata]	363	363	91%	1e-98
XP_001680857.1	dihydrofolate reductase-thymidylate synthase [Leishmania major]	356	356	91%	3e-96
XP_002493237.1	Thymidylate synthase [Pichia pastoris]	354	354	91%	1e-95
EGA72874.1	Cdc21p [Saccharomyces cerevisiae]	353	353	91%	2e-95
AAX40325	dihydrofolate reductase-thymidylate synthase [Trypanosoma cruzi]	341	341	84%	7e-92
XP_001826568.1	thymidylate synthase [Aspergillus oryzae]	335	335	92%	5e-90
AAA29586.1	dihydrofolate reductase-thymidylate synthase [Plasmodium falciparum]	334	334	95%	9e-90
XP_002559080.1	Pc13g06460 [Penicillium chrysogenum]	322	322	91%	3e-86
YP_634079.1	thymidylate synthase [Myxococcus xanthus]	285	285	90%	6e-75
ZP_06871975.1	thymidylate synthase [Bacillus subtilis subsp. spizizenii str. W23]	283	283	90%	3e-74
YP_001685010.1	thymidylate synthase [Caulobacter sp.]	282	282	91%	5e-74
NP_939189.1	thymidylate synthase [Corynebacterium diphtheriae]	280	280	85%	2e-73
NP_948831.1	thymidylate synthase [Rhodopseudomonas palustris]	280	280	90%	2e-73

Notes
See Figure 5-9 for explanation of headings
The highlighted entry corresponds to the highlighted gene in Figure 5-9

Figure 5.12 BLASTP search of UniProt database for proteins resembling human thymidylate synthase (selected results only).

Comparison of the E values shows that the protein sequence match is much more significant than the DNA alignment.

Note that all the cautionary comments discussed in relation to nucleic acid searches also apply to protein sequence comparisons: the list of matches is not definitive, the order shown is not necessarily a true reflection of the order of similarity, and you may get a different set of matches if you use a

```
> ref|ZP_06871975.1| thymidylate synthase [Bacillus subtilis subsp. spizizenii ATCC
6633]
Length=264
Score =  307 bits (787),  Expect = 1e-81, Method: Compositional matrix adjust.
Identities = 151/283 (53%), Positives = 196/283 (69%), Gaps = 22/283 (8%)

Query  32   QYLGQIQHILRCGVRKDDRTGTGTLSVFGMQARYSLRDEFPLLTTKRVFWKGVLEELLWF  91
            QY    +H+L  G +K DRTGTGT+S FG Q R+ L++ FP+LTTK++ +K +  ELLWF
Sbjct  3    QYKDLCRHVLEHGEKKGDRTGTGTISTFGYQMRFHLQEGFPMLTTKKLHFKSIAHELLWF  62

Query  92   IKGSTNAKELSSKGVKIWDANGSRDFLDSLGFSTREEGDLGPVYGFQWRHF-GAEYRDME  150
            +KG TN + L   GV+IW+                 E G+LGPVYG QWR + GA+
Sbjct  63   LKGDTNVRYLQENGVRIWNEWAD------------ENGELGPVYGSQWRSWRGAD-----  105

Query  151  SDYSGQGVDQLQRVIDTIKTNPDDRRIIMCAWNPRDLPLMALPPCHALCQFYVVNSELSC  210
                G+ +DQ+ R+I+ IKTNP+ RR+I+ AWN  ++  MALPPCH L QFYV + +LSC
Sbjct  106  ----GETIDQISRLIEDIKTNPNSRRLIVSAWNVGEIDKMALPPCHCLFQFYVSDGKLSC  161

Query  211  QLYQRSGDMGLGVPFNIASYALLTYMIAHITGLKPGDFIHTLGDAHIYLNHIEPLKIQLQ  270
            QLYQRS D+ LGVPFNIASYALLT MIAH+TGL+PG+FIHT GD HIY NHIE + +QL
Sbjct  162  QLYQRSADVFLGVPFNIASYALLTMMIAHVTGLEPGEFIHTFGDVHIYQNHIEQVNLQLT  221

Query  271  REPRPFPKLRILRKVEKIDDFKAEDFQIEGYNPHPTIKMEMAV  313
            R+ RP PKLR  R+++ I +F  EDF IE Y+PHP IK  ++V
Sbjct  222  RDVRPLPKLRFAREIDSIFNFAFEDFIIEDYDPHPHIKGAVSV  264
```

Notes
Query sequence is the human thymidylate synthase
The line between the query and subject shows the agreement between the two sequences
Identities = percentage of identical residues in the two sequences
Positives = the percentage of similar residues (shown as + in the consensus)

Figure 5.13 Results of a database search using BLASTP (protein search): example of a reported match, from the results shown in Figure 5.12.

different program, such as FASTA, or if you set the variables (including, especially, the matrix) differently. FASTA and BLAST, although they work in different ways, have one thing in common: they are essentially shortcuts to enable database searches to be carried out rapidly without demanding too much computer power. The price to be paid is the possibility of missing something significant. A more sensitive method is provided by a method referred to as the Smith–Waterman algorithm, which is used by a program known as MPsrch.

The alignment reported by BLAST is not necessarily the best, and this is because BLAST is designed for speed and ease of use. For optimal alignment, and for alignment of multiple sequences, Clustal (see below) should be used.

5.4.3 Sequence alignments: Clustal

As indicated in the above discussion, search methods such as FASTA and BLAST do not necessarily display the optimal alignment between your query sequence and the target, as they are deliberately designed for computational ease and speed. Furthermore, they only show pairwise alignments (i.e., an

alignment between two sequences), whereas you are likely to want to see how your sequence lines up with a collection of other proteins. (Similar arguments apply to nucleic acid comparisons, but we will just look at protein alignments.) In order to produce optimised multiple alignments, the most commonly used program is Clustal.

There are many versions of Clustal available. You can use it on-line, with ClustalW2, or you can download ClustalX2 and run it locally. The essentials are the same. It starts by comparing each one of a number of user-defined sequences to produce a matrix of pairwise alignment scores. The two most similar sequences are then aligned, and a consensus generated. Each of the other sequences is then aligned in turn, in order of similarity, with a new consensus being generated at each step, until a multiple alignment of all the proteins is produced. The comments made above for BLAST alignments concerning gap penalties and amino acid substitution matrices also apply to Clustal alignments. An example of part of the output from Clustal, using a selection of protein matches from Figure 5.12, is shown in Figure 5.14. It can be seen that a substantial number of amino acids (indicated by an asterisk) are conserved in all the sequences shown, from a wide variety of sources – mammals, insects, plants, protozoa, fungi and bacteria. In addition, a number of the differences represent substitutions of similar amino acids, suggesting that they may have limited impact on the biological functionality of the protein. It is important to note that the alignment produced should not be regarded as absolute – indeed, a visual inspection may show places where the alignment could obviously be improved, for example, by introducing an extra gap into one of the sequences. Rather than taking an alignment of sequences as an absolute, it should be used, cautiously, as an aid towards the establishment or confirmation of evolutionary relationships. Some of the conserved amino acids that are identified by the alignment may have real significance, such as indicating a substrate-binding motif, and it is then likely that the aligned amino acids do actually represent an evolutionary relationship. However, in many cases, to be able to draw such a conclusion, you would need to look at the position of this amino acid in the three-dimensional structure of the proteins you are comparing, to see if it occurs in a region important for the function of the protein.

The multiple alignment of sequences from programs such as Clustal is the first step in the production of a putative phylogenetic tree based on the sequence similarities of the proteins. Indeed, Clustal will produce such a tree, with an example shown in Figure 5.15, where it can be seen that the results largely agree with what you would expect from the taxonomic relationships of the organisms. For example, the mouse and rat sequences are very close, while that from humans is less so, but still grouped with them. Similarly the plant sequences cluster together. However, the bacterial and protozoal species are not grouped as one would expect, and are in a different

```
Rat              G-MQARYSLRD-EFPLLTTKRVFWKGVLEELLWFIKGSTN-AKELSSKGVRIWDANGSRD 110
Mouse            G-MQARYSLRD-EFPLLTTKRVFWKGVLEELLWFIKGSTN-AKELSSKGVRIWDANGSRD 110
Human            G-MQARYSLRD-EFPLLTTKRVFWKGVLEELLWFIKGSTN-AKELSSKGVKIWDANGSRD 116
Herpesvirus 3    G-MQARYNLRN-EFPLLTTKRVFWRAVVEELLWFIRGSTD-SKELAAKDIHIWDIYGSSK 104
Arabidopsis      G-CQMKFNLRR-NFPLLTTKRVFWRGVVEELLWFISGSTN-AKVLQEKGIRIWDGNASRA 368
Maize            G-CQMRFNLRK-NFPLLTTKRVFWRGVVEELLWFISGSTN-AKVLQEKGIHIWDGNASRE 324
Trypanosome      G-AQMRFSLRNNRLPLLTTKRVFWRGVCEELLWFLRGETY-AKKLSDKGVHIWDDNGSRA 324
Candida          APPQLRFSLKDDTFPLLTTKKVFTKGIILELLWFIAGCTD-AKKLSDQGVKIWEGNGSRE 96
Drosophila       G-SQMRFDMRN-SFPLLTTKRVFFRAVAEELLWFVAGKTD-AKLLQAKNVHIWDGNSSRE 122
Plasmodium       G-YIMKFDLSQ-YFPLLTTKKLFLRGIIEELLWFIRGETN-GNTLLNKNVRIWEANGTRE 411
C. diphtheriae   G-QQMRFDLTE-AFPLITTKKVYWKGVIGELLWFLQGSSN-VRWLQKRNIHIWDEWAS-- 92
B. subtilis      S-KQMRFDNSE--VPILTTKKVAWKTAIKELLWIWQLKSNDVTELNKMGVHIWDQWKQ-- 99
                 .    ::.     .*::***::  :    ****:    :     *   .::**:

Rat              FLDSLGFSARQEGDLGPVYGFQWRHFGADYKDMDSDYSGQGVDQLQKVIDTIKTNPDDRR 170
Mouse            FLDSLGFSARQEGDLGPVYGFQWRHFGAEYKDMDSDYSGQGVDQLQKVIDTIKTNPDDRR 170
Human            FLDSLGFSTREEGDLGPVYGFQWRHFGAEYRDMESDYSGQGVDQLQRVIDTIKTNPDDRR 176
Herpesvirus 3    FLNRNGFHKRHTGDLGPIYGFQWRHFGAEYKDCQSNYLQQGIDQLQTVIDTIKTNPESRR 164
Arabidopsis      YLDGIGLTEREEGDLGPVYGFQWRHFGAKYTDMHADYTGQGFDQLLDVINKIKNNPDDRR 428
Maize            YLNSVGLAHREEGDLGPIYGFQWRHFGAEYTDMHADYTGKGFDQLMDVIDKIKNDPEDRR 384
Trypanosome      FLDSRGLTEYEEMDLGPVYGFQWRHFGAAYTHHDANYDGQGVDQIKAIVETLKTNPDDRR 384
Candida          YLDKMGFVDRREGDLGPVYGFQWRHFGAEYKTCEDDYTGQGVDQLKEVIHKLKTNPYDRR 156
Drosophila       FLDKMGFTGRAVGDLGPVYGFQWRHFGAQYGTCDDDYSGKGIDQLRQVIDTIRNNPSDRR 182
Plasmodium       FLDNRKLFHREVNDLGPIYGFQWRHFGAEYTNMYDNYENKGVDQLKNIINLIKNDPTSRR 471
C. diphtheriae   EEGELGPVYGVQWRSWPTP--------DGQHVDQIAQALDILKNNPDSRR 134
B. subtilis      EDGTIGHAYGFQLG-------KKNRNLNGEKVDQVDYLLHQLKNNPSSRR 142
                              :*  **.*                 : .**:   :. ::.:* .**

Rat              IIMCAWNPKDLPLMALPPCHALCQFYVVNG-----ELSCQLYQRSGDMGLGVPFNIASYA 225
Mouse            IIMCAWNPKDLPLMALPPCHALCQFYVVNG-----ELSCQLYQRSGDMGLGVPFNIASYA 225
Human            IIMCAWNPRDLPLMALPPCHALCQFYVVNS-----ELSCQLYQRSGDMGLGVPFNIASYA 231
Herpesvirus 3    MIISSWNPKDIPLMVLPPCHTLCQFYVANG-----ELSCQVYQRSGDMGLGVPFNIAGYA 219
Arabidopsis      IIMSAWNPSDLKLMALPPCHMFAQFYVANG-----ELSCQMYQRSADMGLGVPFNIASYS 483
Maize            IILSAWNPSDLKKMALPPCHMFAQFYVENG-----ELSCQMYQRSADMGLGVPFNIASYS 439
Trypanosome      MLFTAWNPSALPRMALPPCHLLAQFYVSNG-----ELSCMLYQRSCDMGLGVPFNIASYA 439
Candida          IIMSAWNPPDFPKMALPPCHVFSQFYVNFPKDGKPQLSCLLYQRSCDMGLGVPFNIASYA 216
Drosophila       IIMSAWNPLDIPKMALPPCHCLAQFYVSEKRG---ELSCQLYQRSADMGLGVPFNIASYA 239
Plasmodium       ILLCAWNVKDLDQMALPPCHILCQFYVFDG-----KLSCIMYQRSCDLGLGVPFNIASYS 526
C. diphtheriae   NIVSAWNVADLNNMALPPCHLLFQLYVADG-----KLSCQLYQRSADMFLGVPFNIASYS 189
B. subtilis      HITMLWNPDELDSMALTPCVYETQWYVKQG-----KLHLEVRARSNDMALGNPFNVFQYN 197
                  :   **   :  *.*.**    * **        :*   :  ** *: ** ***:  *
```

Notes
* = identical residues in all sequences
: and . = similar residues in all sequences

Figure 5.14 Clustal alignments (partial) of selected thymidylate synthases.

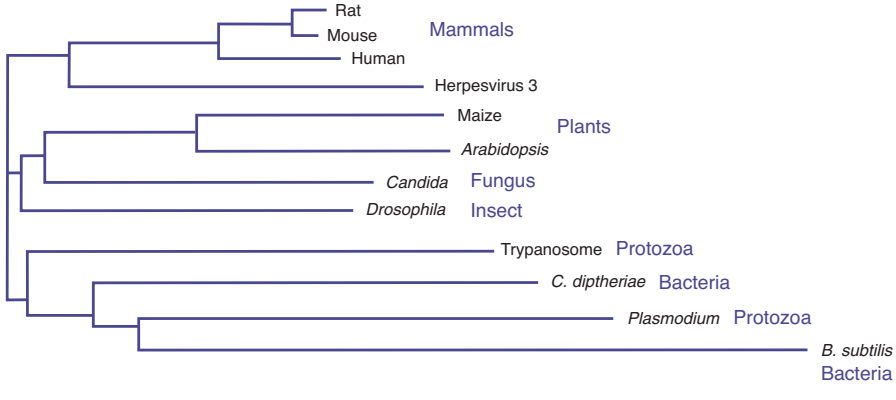


Figure 5.15 Tree based on Clustal alignments of selected thymidylate synthases.

order from that shown in Figure 5.14. This reflects the fact that a different algorithm (known as Neighbour Joining) was used for construction of the tree. We will look further at *molecular phylogeny,* and the ways of constructing trees, in Chapter 9. For the moment, it is enough to say that it would be a mistake to read too much into a tree based on a single protein as in Figure 5.15.

5.5 Protein structure

5.5.1 Structure predictions

We have now identified an ORF, and used the computer to translate it into a protein sequence. The comparison with other known proteins may give us some idea of its structure and function, particularly if our protein is a member of a family of proteins, some of which have been characterized structurally. By aligning the sequences, and looking for specific regions (motifs and domains – see later) that are characteristic of families of proteins, we can therefore make predictions as to the likely conformation, and function, of our unknown protein.

Independently of such alignments, we can investigate a number of aspects of the structure of that protein. This is obviously particularly important where we do not already have clues to the function of our gene through its similarity with others that have already been characterized. A full consideration of all the possibilities is beyond the scope of this book, but it is worth a brief and selective overview, especially as an examination of a protein's structure can provide some leads as to its possible function.

For example, if we look at the amino acid composition of the protein, and especially at the occurrence and distribution of hydrophobic and hydrophilic amino acids, we should be able, just on this evidence, to detect if it is likely to be a membrane protein. Proteins that are embedded in a membrane will normally have substantial stretches of hydrophobic amino acids. By comparison, soluble proteins are more likely to be predominantly hydrophilic, although they can contain hydrophobic regions if the protein folds into a pocket that shields those hydrophobic regions from the aqueous cytosol. If we look more closely at the predicted structure (especially in combination with testing for specific motifs – see below) we may be able to guess whether it is involved in transport across the membrane, or energy generation, or acts as a receptor for signal transduction, or any of the other functions commonly associated with membrane proteins. As an example, Figure 5.16 shows a hydrophobicity plot for the human rhodopsin protein, a member of the so-called seven-transmembrane receptor family. From the plot one can see seven regions of marked hydrophobicity, with intervening regions that are more hydrophilic. Such alternating stretches of hydrophobic and hydrophilic amino acids are

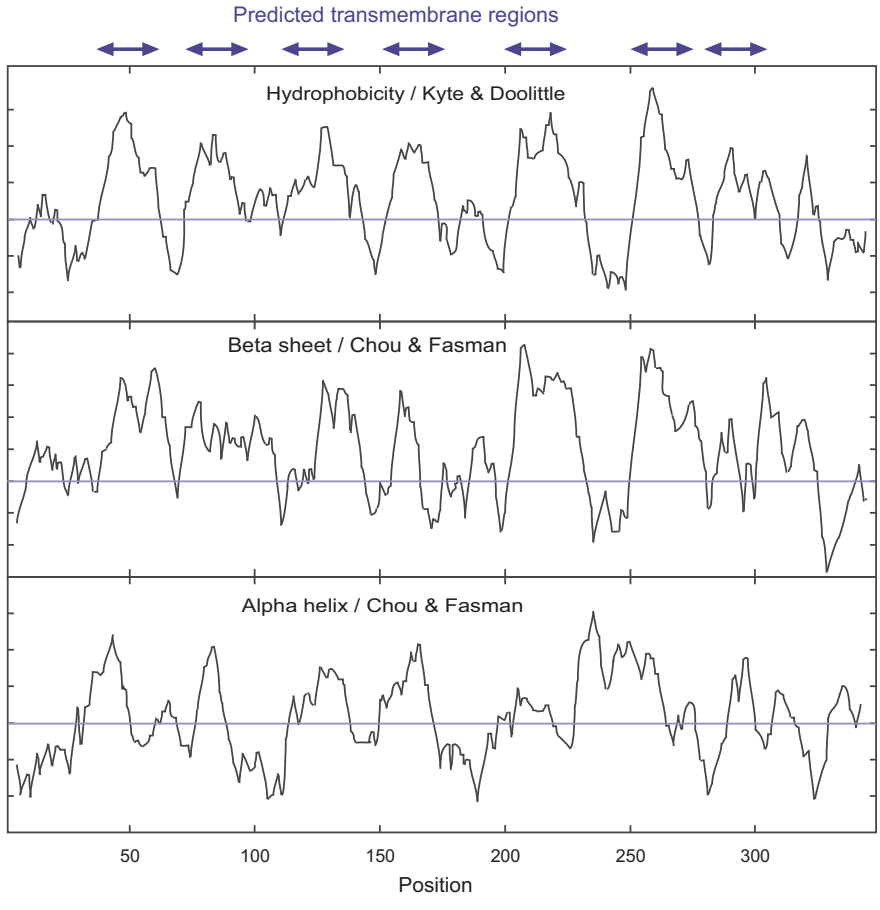


Figure 5.16 Protein structure predictions: Kyte–Doolittle hydrophobicity plot and Chou and Fasman secondary structure predictions for human rhodopsin (UniProt P08100). Analysis performed using ProtScale in the ExPASy server of the Swiss Institute of Bioinformatics (http://expasy.org/tools/protscale.html).

highly characteristic of a membrane-spanning protein, with the hydrophobic regions embedded in, and spanning, the membrane, while the hydrophilic regions protrude into the cytoplasm, or the external environment, and link these membrane-spanning regions. In addition, as there are seven membrane-spanning regions we can conclude that one end of this transmembrane protein is exposed to the cytoplasmic environment, while at the other end the protein protrudes into the external environment. This is a commonly occurring theme in membrane proteins, especially in those proteins that function as transporters (i.e., they ferry material across the membrane) or as signal transducers (i.e., they respond to changes in the external environment and transmit a signal across the membrane to the interior of the cell).

Proteins do not, of course, exist merely as one-dimensional sequences of amino acids, but adopt higher orders of conformation. A variety of programs are available to predict, from the primary sequence of amino acids, which parts will adopt secondary structures such as alpha-helices and beta-sheets, with examples shown in Figure 5.16. However, these predictions are not entirely reliable, as the factors that determine the final structure of the protein are complex, meaning that interpretation is complicated. Predicting how those elements of secondary structure will fold into the tertiary and higher orders of structure that are characteristic of the native protein is even more difficult. For this, in spite of the huge increase in computer capacity, we are still left with the need to produce crystals of the pure protein for X-ray crystallography. While the science of X-ray crystallography is beyond the scope of this text, it is enough to note that in recent years the rate at which proteins are being crystallized, and their three-dimensional structures derived, is increasing rapidly. Even the three-dimensional structures of membrane-bound proteins, which are notoriously difficult to crystallize, are currently being solved with relative rapidity. Web-based resources such as UniProt contain a large, searchable database of all currently solved three-dimensional structures, and the interested reader is directed there for further information.

5.5.2 Protein motifs and domains

Even if our protein does not show much overall similarity to any other characterized protein, we may find, if we look more closely, that it contains short sequences of amino acids that are very similar to parts of a number of other proteins. These conserved regions are known as *motifs*. The recognition of motifs provides yet another clue as to the function of our protein, or other aspects of its structure.

Some motifs occur because a wide range of enzymes, with otherwise disparate properties, may use the same substrate. For example, a wide range of enzymes use ATP as a substrate. In many cases, the region of the enzyme that binds ATP has a similar structure – even though the remainder of the sequence may show no similarity. If we can recognize the presence of a putative ATP-binding site in our protein, then we can infer that it is probably an enzyme that uses ATP as a substrate.

As well as substrate-binding sites, there is a range of other motifs that indicate sites for structural modification such as lipid or sugar attachment, for secretion or for targeting to specific cellular compartments, or in the case of regulatory proteins, for DNA binding. There are libraries of known motifs available, so it is a simple matter to screen a protein sequence for any of these motifs. One of the longest-established libraries of patterns is PROSITE (see Box 5.1). Figure 5.17 shows the output from a web-based search for motifs present in protein Rv0194 (a protein from *Mycobacterium tuberculosis*

The query protein was a probable ABC (ATP-binding cassette) transporter protein from *M. tuberculosis* (UniProt O53645); the analysis was done using ScanProsite, at http://expasy.org/tools/

Edited output:

O53645 **O53645_MYCTU** (1194 aa)
PROBABLE DRUGS-TRANSPORT TRANSMEMBRANE ATP-BINDING PROTEIN ABC TRANSPORTER. *Mycobacterium tuberculosis*

PS50929 ABC_TM1F ABC transporter integral membrane type-1 fused domain profile :
21 - 301: **score** = 47.505
628 - 910: **score** = 45.116

PS50893 ABC_TRANSPORTER_2 ATP-binding cassette, ABC transporter-type domain profile :
334 - 568: **score** = 22.056
Predicted feature:
NP_BIND 367 374 ATP (Potential) [condition: [AG]-x(4)-G-K-[ST]]
942 - 1177: **score** = 21.382

Figure 5.17 ProSite scan.

that was identified as a probable transporter protein). This output demonstrates the presence of two ABC (ATP-binding cassette) transporter transmembrane domains, and two ATP-binding motifs, which are characteristic of the family of proteins known as ABC transporters.

The overall reaction catalysed by some enzymes is actually a series of separate reactions, with different parts of the enzyme responsible for different steps. For example, acetyl-CoA carboxylase (catalysing the first step in fatty acid synthesis) has one component with a covalently attached biotin, a second component with (non-covalent) binding sites for ATP and CO_2 that carboxylates the biotin, and a third component, with an acetyl-CoA binding site, that transfers the carboxyl moiety from the biotin to the acetyl-CoA. In *E. coli* these components are made as separate polypeptides, which associate with one another at the post-translational stage to form the complete holoenzyme. In contrast, in mammalian cells the enzyme is made as a single polypeptide with different parts of the structure responsible for the different activities. We refer to these elements as *domains*. It is not uncommon to find enzymes that consist of several subunits in one organism, but as a single polypeptide with several domains in another source – perhaps due to a fusion of genes that originally evolved separately.

Domains often form more or less structurally independent regions of a protein – in other words each domain folds up into its own secondary structure with little or no structural interaction with the other domains, other than a flexible loop connecting them. You can visualize such a structure, roughly, as a set of balls connected by bits of string.

In a typical protein, some regions, especially those that are essential for the function of the enzyme, show relatively little variation from one species to another. They are referred to as *conserved* domains. However, other regions may be much more variable. These often include the terminal regions, and other sequences that form loops protruding from the main body of the folded protein. The latter are especially important in some viruses (notably HIV) where they are both *hypervariable* and *immunodominant*, i.e., they vary very extensively and rapidly, and constitute the principal antigen 'seen' by the body's immune system. This is significant as it makes it extremely difficult to produce effective vaccines.

The domain structure of proteins can give rise to a high degree of variability through the formation of so-called *mosaic* proteins. This means that a protein in one strain contains a domain that is not related to the sequence found in other strains of the same species, but is closely related to the sequence found in a different organism (Figure 5.18). The inference is that information has been transferred between organisms, so that a part of the gene is replaced with information from a different source. This contributes a further dimension to our overall understanding of the dynamic nature of the genome (sometimes referred to as *genome plasticity*).

A useful tool for the identification and comparison of domains is the Pfam database of protein domain families, available at several sites, including the Sanger Institute (see Box 5.1). This uses a set of multiple sequence alignments for each family. (Technically, these alignments are encoded using a statistical treatment known as Hidden Markov Models (HMMs), which have a variety of applications in bioinformatics. See Chapter 8 for further information.)

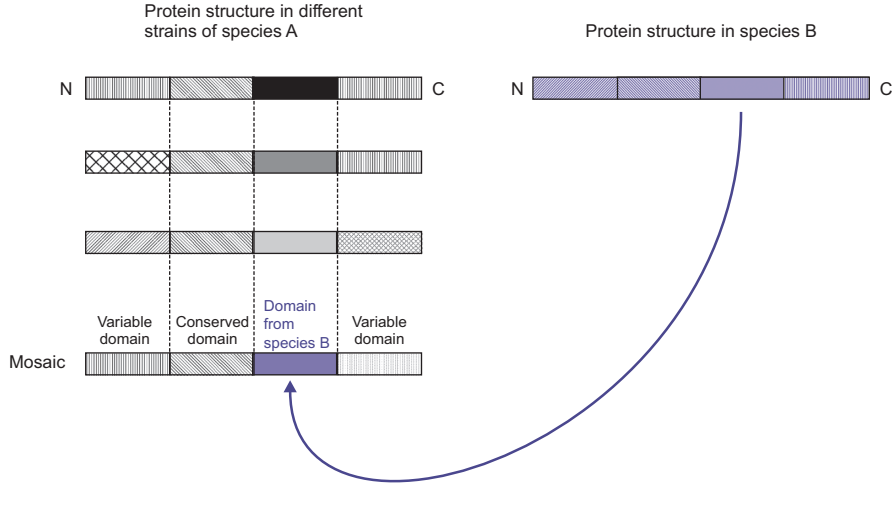


Figure 5.18 Conserved domains and domain shuffling.

Description from UniProt for O53645_MYCTU probable drugs-transport transmembrane ATP-binding protein ABC transporter

Source	Domain	Start	End
Pfam A	ABC_membrane	20	289
Pfam A	ABC_tran	374	499
Pfam A	ABC_membrane	627	898
Pfam A	ABC_tran	983	1108
transmembrane		20	41
transmembrane		56	76
transmembrane		144	173
transmembrane		255	285
transmembrane		627	648
transmembrane		660	681
transmembrane		743	761
transmembrane		767	785
transmembrane		848	868
transmembrane		874	895

ABC transporters belong to the ATP-Binding Cassette superfamily, which uses the hydrolysis of ATP to translocate a variety of compounds across biological membranes. ABC transporters are minimally constituted of two conserved regions: a highly conserved ATP binding cassette (ABC) and a less conserved transmembrane domain (TMD). Most ABC transporters function as a dimer and therefore are constituted of four domains, two ABC modules and two TMDs.

Figure 5.19 Protein families: Pfam database.

Figure 5.19 shows part of the output from an analysis of the same protein as used in Figure 5.17. The graphic display shows the presence of two ABC membrane domains and two ABC transporter domains. The explanation below shows predicted transmembrane regions, associated with the membrane domains, and two ATPase domains, corresponding to the two transporter domains.

PROSITE and Pfam originated in different ways, and are compiled differently, but it will be seen that functionally they now overlap considerably. If you look back to Figure 5.8, showing the annotation of a protein sequence in the UniProt database, you will see that there are links to both PROSITE and Pfam. All UniProt entries have been analysed for the presence of motifs and domains in these databases, so you only need to click on the links to get this information.

5.6 Confirming gene function

From the sequence comparisons and other analyses outlined above, we may be able to reach a conclusion as to the probable biochemical function of the protein encoded by our cloned gene. How do we confirm that this is really its activity? The obvious answer – to express the cloned gene in a suitable

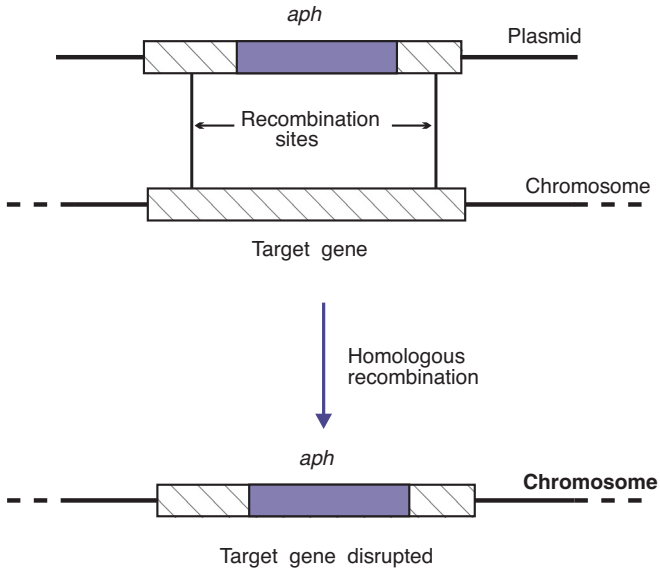
host and assay its activity – is often not possible, as many proteins do not have an enzymic function that is easily assayable *in vitro*. Even if it does, that still does not provide the full answer as to what its role is within the original organism. One of the most powerful strategies in such a situation is to make a specific deletion of that gene and determine the effect on the organism, using a procedure known variously as *allelic replacement, gene replacement* or *gene knockout*.

5.6.1 Allelic replacement and gene knockouts

Allelic replacement relies on the natural process of *homologous recombination*, which means that when there are two identical pieces of DNA in the cell, enzymes within the cell may break the two DNA chains, cross them over and rejoin them. (This is a highly simplistic version of a more complex process, but it will do for our purposes.) We can exploit this process to delete a specific gene (or part of a gene) in the chromosome, or to replace it with a version that we have inactivated *in vitro*, thus destroying the function of that gene. This allows us to test the consequences of inactivating an individual gene, and thus make deductions about its function.

A typical procedure for allelic replacement in a bacterial host would be to manipulate the cloned gene so as to replace the central part of the gene with an antibiotic resistance gene (Figure 5.20). We would do this using a suicide plasmid (a plasmid that is unable to replicate in the chosen organism) so that when we transform the bacteria with the construct, only those cells in which the resistance gene has become incorporated into the chromosome will become antibiotic resistant. We can select these on agar containing the antibiotic. If things go well, incorporation into the chromosome will have occurred by homologous recombination at the required position, thus inactivating the gene concerned.

It should be noted that replacement of the gene actually requires recombination at two positions, one either side of the gene. This is referred to as a double crossover event. A single crossover, in the homologous region at one side or the other of the construct, will produce resistant bacteria by incorporation of the entire plasmid into the chromosome rather than replacing the gene. Single crossovers may result in gene inactivation, depending on the details of the construct, but will usually be unstable, as further recombination may eliminate the plasmid again, restoring the original intact gene. We can select against single crossovers by incorporating a counter-selectable marker into the plasmid. In other words, we put a gene on the plasmid that will, if it is still present, confer a disadvantage on the cell. A gene known as *sacB* is commonly used for this purpose in bacteria, as the presence of *sacB* renders the cells sensitive to sucrose in the medium. Plating the cells on a



aph = aminoglycoside phosphotransferase, causing kanamycin resistance

Figure 5.20 Gene disruption by allelic replacement.

sucrose-containing medium will result in any cells containing *sacB* (which includes the single crossovers but not the doubles) being unable to grow and form colonies.

An obvious limitation of gene replacement technology is that the inactivation of the gene concerned may be a lethal event. Nevertheless, it is a valuable approach for identifying the function of specific genes.

Although we have described gene replacement in terms of identifying gene function, its applications extend far beyond that. It can be used to inactivate genes that are necessary for the virulence of a pathogenic bacterium, thus producing attenuated strains that may be useful vaccine candidates. In addition, essentially identical procedures can be used to knock out genes in other organisms, including experimental animals (especially mice). Many strains of mice, lacking individual genes, have been produced in this way, and are invaluable for research purposes. This is considered further in Chapters 8 and 11, together with *RNA interference*, which provides an often simpler way of silencing a specific gene in eukaryotic hosts.

Other methods that are used for genome-wide identification of genes with specific functions are also considered in Chapter 8.

5.6.2 Complementation

The above discussion makes the implicit assumption that the phenotypic consequences of the mutation are due solely to the effect of the loss of that gene. This is not always true. In particular, some mutations may show an effect known as *polarity*. This means that the mutation affects not only the altered gene, but also those adjacent to it. In bacteria, this can arise from the arrangement of genes into operons, which are transcribed into a single mRNA. Mutation of one gene may interfere with transcription of the operon, and thus affect the expression of the genes downstream from it. Furthermore, genes and their products interact in many complex ways within the cell, so that disruption of one gene may have unexpected effects on the activity of other genes and their products.

It is therefore necessary to interpret the results arising from these experiments with care. The standard way of checking that the altered phenotype is a direct consequence of the inactivation of a specific gene is by *complementation*. This involves introducing into the mutant cell a fully active version of the affected gene (most simply, by a cloned version on a plasmid). If the alteration in the phenotype is indeed due solely to the loss of the affected gene and its product, then the mutation will be complemented by the plasmid, i.e., the wild-type phenotype will be restored. This is not entirely foolproof. For example, if the original mutation disrupts the regulation of other genes, complementation may be successful in restoring the wild-type phenotype, even though the gene product is not directly responsible for the observed characteristics. Nevertheless, complementation does provide an element of confirmation of the consequences of the original mutation.

Complementation can be used in the same way to relate a cloned gene to the phenotype of mutant strains produced by other means, including mutants isolated by screening for specific altered phenotypes. For example, if you have isolated a non-motile bacterial strain that is unable to produce flagella, you can introduce your cloned gene into that strain and test its motility. If motility, and the ability to produce flagella, are restored, then you have good evidence (subject to the caution above) that your cloned gene is the same one that is mutated in the host strain. The equivalent of complementation in eukaryotic model organisms is the phenotypic rescue of a mutation by transgenic delivery (see Chapter 11).

6 Analysis of Gene Expression

Gene expression can be divided into two main phases: *transcription* (copying DNA into RNA), and *translation* (production of a protein or polypeptide to the specifications of an mRNA template). For the purpose of this chapter, we will treat the polypeptide chain as the end product of gene expression; this ignores subsequent post-translational modification, including the folding of the polypeptide into the correct conformation for biological activity, as well as other modifications, such as phosphorylation or the specific addition of carbohydrate groups, which may modify the protein's biological activity. However, it is important to note that there are classes of genes that are transcribed into RNA, but not into proteins. These *non-coding RNAs* include not only the well-characterized rRNAs and tRNAs, but also the recently discovered miRNAs (microRNAs), which have an important role in gene regulation. In this chapter, we will discuss methods for the study of the products of single genes, and in Chapter 10 we will extend this discussion to the study of complex samples on a genome-wide basis.

6.1 Analysing transcription

The first group of techniques for examining gene expression consists of various ways of assessing the amount of a specific transcript in a specific sample at a specific time. Other methods, described later in the chapter, are concerned with studying the transcriptional activity associated with the gene in question. Two points concerning the measurements of mRNA levels should be noted. Firstly, the amount of a specific mRNA in a cell at a given point in time is influenced not only by the level of transcriptional activity, but also

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

by the stability of that mRNA (i.e., it reflects the balance between transcription and degradation). Therefore, a gene that is transcribed at a low level but gives rise to a stable mRNA may result in a greater amount of mRNA than a transcriptionally more active gene that produces an unstable message. Secondly, the amount of mRNA present does not necessarily correlate with the amount of protein made. This is most obvious with bacterial cells, where a single polycistronic message may be translated into many different polypeptides; although these are all translated from the same mRNA template, the levels of the different proteins made can be widely different – translation efficiencies can vary considerably.

Although some of the methods described in this chapter are less important than they used to be – having been supplanted primarily by RT-PCR (for individual genes) and microarrays and transcriptome sequencing (for genome-wide assays – see Chapter 10), they retain some usefulness, and are of more than merely historical interest.

6.1.1 Northern blots

The oldest methods for detecting a specific mRNA rely on hybridization to specific, labelled probes. One of these is known as a *Northern blot*. (The name has nothing to do with geography; it is derived by analogy with a Southern blot, as described in Chapter 3. It is a sort of joke! Later in this chapter, we will see a further attempt at humour in the term 'Western blot'.) A Northern blot involves electrophoretic separation of a purified extract of RNA from living cells, followed by immobilization onto a membrane and hybridization with a specific probe, as shown in Figure 6.1. This enables us to detect the presence of a specific mRNA, assuming that the probe we use is sufficiently

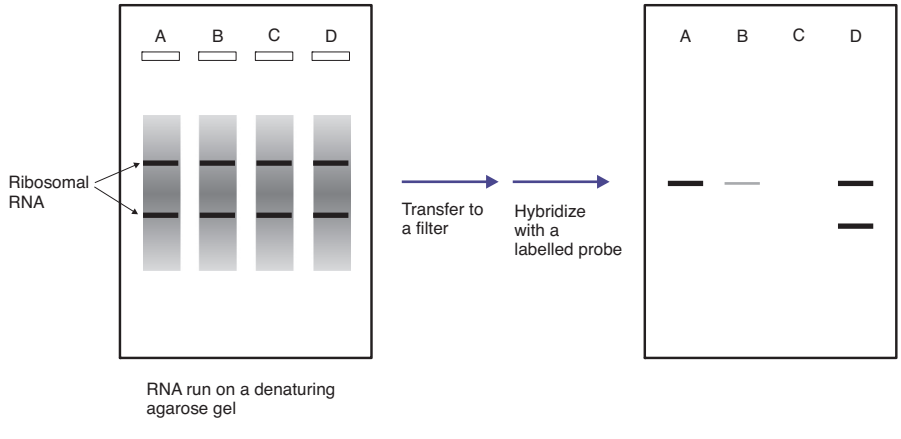


Figure 6.1 Northern blotting.

specific, and to estimate its size. Comparing the strength of the signal in different samples, we can get an idea of the *relative* amount of specific mRNA present in each. In Figure 6.1 we can see that sample B gave a weaker signal than sample A, and therefore contained correspondingly less of the specific mRNA, while in sample C expression of the gene was not detected. This estimate assumes that an equivalent amount of total mRNA is loaded in each track, which can be verified by using a probe that detects an mRNA that is expressed at constant levels.

However, there are some limitations to this technique. Firstly, although we can get a *relative* quantitative estimate of the strengths of the signals, it is generally not possible to obtain an *absolute* measurement of the number of molecules of a specific mRNA. Secondly, the technique is not very sensitive, and requires fairly large amounts of RNA (normally at least a microgram of poly(A)-enriched mRNA or 10 μg of total RNA). It may not be possible to obtain sufficient quantities of starting material. For example, if we want to look at the expression of specific genes in a pathogenic bacterium when it is growing within an infected host (rather than the artificial situation of laboratory culture), or levels of mRNA in material obtained by biopsy, it is likely to be extremely difficult to get large enough quantities of RNA to produce a Northern blot.

For these reasons, PCR-based techniques (specifically quantitative RT-PCR, see next section) are now more popular. Nevertheless, Northern blotting has some lasting advantages. Above all, it remains a major method for determining the size of a specific transcript, and for detecting the presence of different transcripts for a specific gene (e.g., in Figure 6.1, we can see the presence of a second, smaller transcript in sample D). This is becoming more important as the human genome project teaches us that the number of genes is considerably smaller than the number of gene products. An important part of the explanation for this is that many genes are differentially spliced, and Northern blotting is an important tool for detecting this.

6.1.2 Reverse transcription-PCR

A much more sensitive way of detecting specific mRNA is to adapt the polymerase chain reaction (PCR, see Chapter 4) in order to be able to amplify the specific message. Because this requires copying the mRNA into cDNA using the enzyme reverse transcriptase (see Chapter 3), it is described as reverse transcription PCR (RT-PCR). The principle of the method is illustrated in Figure 6.2.

The sensitivity of this procedure is such that it is possible to detect a specific mRNA species in a single cell. This has obvious advantages, such as being able to detect low abundance mRNA (i.e., mRNA that is present at extremely low levels), or to analyse gene expression in cells that are difficult to obtain

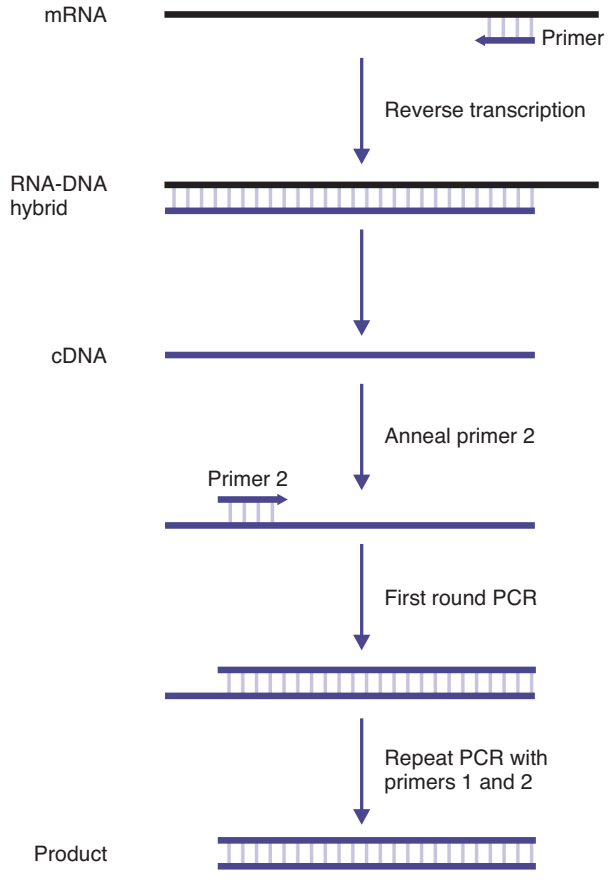


Figure 6.2 Reverse transcription PCR.

in large numbers. Examples are the analysis of gene expression in bacteria from lesions in an infected host, or in cells from tumours, in order to find out which genes are expressed under those circumstances. Culturing the bacteria or tumour cells is no use, as the gene expression would change, whereas RT-PCR enables us to analyse gene expression in the limited number of cells recoverable from the lesion in their native state.

The sensitivity of RT-PCR confers a further, rather less obvious, advantage. If we want to look at tissue-specific expression, or at expression in cells under different physiological conditions (whether bacterial or eukaryotic), we have to try to ensure that all the cells from which we extract the mRNA are indeed expressing the same repertoire of genes. The relatively large amount of mRNA required for a Northern blot means that we have to use RNA extracted from a considerable number of cells, and this population is likely to be heterogeneous. In comparison, RT-PCR requires much less

mRNA, and hence far fewer cells. It is therefore much more likely that all the cells are in the same physiological state, and we will thus get a much more accurate picture of the true profile of gene transcription in those cells. Reporter genes (see below) provide an alternative way of addressing this question.

For analysis of expression of a specific gene, we need to be able to do more than merely detect the presence or absence of the relevant mRNA. We need to be able to assess how much of it is present. In Chapter 4, we saw how quantitative PCR results can be obtained using various techniques collectively known as *real-time PCR*, in which the PCR reaction is adapted so that product formation is accompanied by an increase in fluorescence, which can be monitored continuously. (Do *not* be tempted to abbreviate 'real-time' to RT, which is the accepted abbreviation for reverse transcription; the correct abbreviation is qPCR.) The number of cycles needed for formation of a detectable amount of product, known as the cycle threshold (C_T) value, is related to the initial amount of template; higher amounts of initial template will result in lower C_T values. Because PCR only works with DNA as a template, for quantitative assessment of the amount of a specific mRNA in a sample, you first need to reverse transcribe it into DNA before carrying out the real-time PCR; this is then known as quantitative RT-PCR.

If you are comparing the levels of mRNA in different samples, you have to make sure that you start with the same overall amount of mRNA. Standardizing the amount of *total RNA* (rather than specifically mRNA) is not completely reliable, as the ribosomal content of the cells may vary considerably. The best way of standardizing the template is to carry out a parallel RT-PCR using another gene that is known to be constitutively expressed (or at least expressed at the same level under the different conditions you are testing). Common genes that can be used for this standardization in eukaryotes include the so-called housekeeping genes, such as β-actin. Such standardization enables testing for differences in the specific expression of an individual gene under different conditions. Can you use the same method to compare the expression of different genes? Subject to certain safeguards, yes. The problem to be dealt with is that PCR is not equally efficient with all combinations of primers and templates. Even if the amounts of template are the same, a more efficient PCR will result in earlier detection (a lower C_T value) than for another gene with a less efficient PCR. Some preparatory work is therefore necessary to ensure that the templates are amplified equally efficiently, possibly using different primers to change the efficiency so that they match. (It should be noted that this consideration also applies to the comparison of your gene with the standard template as referred to above.)

Quantitative RT-PCR is much the best, and most commonly used, way of assessing the expression levels of individual genes (or at least the amount of specific mRNA present at the time of sampling). However, there are additional methods that still have some applications.

6.1.3 *In situ* hybridization

Whereas qRT-PCR will tell you how much of a transcript is present within a cell population, it will not provide information on the location of its expression. Such information can be important on two different scales: Firstly, on the larger scale, tissues within multicellular organisms are usually made up of a mixture of cell types; in such a situation it may be of interest to determine not only if a gene is being expressed in a tissue, but also within which cell types of that tissue. Secondly, on the smaller scale, we now understand that RNAs may be localized within a cell in a functionally important manner; for example, the localization of the *bicoid* mRNA at a single point within the single-cell stage of a developing *Drosophila* is critical to which end the head and tail will develop from. *In situ* hybridization (ISH) answers this question.

In this type of hybridization, the target is not purified from the cells that express it; instead, it remains immobilized within each cell on a microscopic slide. The target may be either DNA or RNA. The former was often used for gross chromosomal mapping of genes, prior to the availability of genome sequences. A labelled probe is hybridized to cells with metaphase chromosomes. By using a differentiating counterstain, the investigator can identify the localization of the gene to a specific region of a specific chromosome (see Chapter 9 for a more detailed account).

In the context of gene expression, RNA *in situ* hybridization can be used to identify those cells (e.g., in a tissue section) that produce mRNA for a specific gene. As an alternative to radioactive labelling, the probe can be labelled by attaching a fluorescent dye. This method is known as fluorescent *in situ* hybridization (FISH).

Although it is possible to obtain semi-quantitative information by counting the number of grains (when using radioactive labels detected by dipping in a liquid film), quantification is not the main strength of this method. Rather, its strength is to tell us which particular cell types express a specific transcript. This information is important not least because we still do not know the exact function of many genes. *In situ* hybridization allows us to refine this analysis further than the other methods described. ISH also lends itself to high-throughput automated processing of probes for many genes in the same sample type. Images from a number of such projects are available online – see, e.g., GenePaint (www.genepaint.org) and the Allen Human Brain Atlas (www.brain-map.org).

6.2 Methods for studying the promoter

The approaches described above allow the quantitative or qualitative study of mRNA within a cell; however, they do not allow us to begin to understand the regulatory mechanisms that control gene expression. To achieve this, we

need to use technologies that allow us to examine the function of the regulatory elements of a gene, those elements that control the rate at which RNA polymerase binds to the DNA and initiates transcription. In this section, we will describe firstly, how to locate these regulatory regions, and secondly, the common technologies used to examine their function.

6.2.1 Locating the promoter

As we have previously seen, genome sequence data enable you to identify open reading frames that potentially code for proteins. It is also possible to use such data to find the location of potential regulatory regions that will control the rate of transcription for a gene. However, there is a complication to such an approach. As we saw in Chapter 5, alignment of the protein sequences for the same protein from two different species (so-called orthologues), produces an alignment of high similarity. If the alignment was undertaken at the cDNA level, we would also see high identity between the sequences, although it will be less than that seen at the protein level (due to synonymous codons; see Chapter 7). However, the nucleotide sequence that encodes a given protein is under a strong evolutionary pressure to remain constant, meaning that it will produce a protein of similar biological function in different species, and hence there are fewer than the randomly expected number of changes in this region even over relatively large evolutionary distances. If, finally, the comparison were to be undertaken at the genomic level in a eukaryotic organism, you would find considerable variation in the size of the exons. You might be somewhat more surprised to discover a difference in the number of introns, for introns have been lost (more commonly than added) in some phylogenetic groups.

In comparison, the regulatory regions of genes tend to evolve much faster, allowing each organism to fine tune its gene expression to its particular requirements. It is important to note, however, that these regulatory regions still evolve slower than the intergenic DNA that makes up the rest of our genome. We can use these different rates of evolution to help identify putative regulatory region(s) for a gene by comparing genome sequences between two species that are much more closely related than we would use to examine differences in the coding region. A commonly used suite of programs for such an approach is VISTA Tools (http://genome.lbl.gov/vista/index.shtml), and Figure 6.3 shows an output from this program, aligning the nucleotide sequence upstream of the coding region for the *ABCC2* gene between human and rat. Areas of identity are shown as peaks within the trace, which are highlighted when they reach statistical significance. In panel (a) we can clearly see the identity between the exons for *ABCC2* in rat and human, but there is very little identity upstream of the first exon, which is where we would expect the proximal promoter regions to be located.

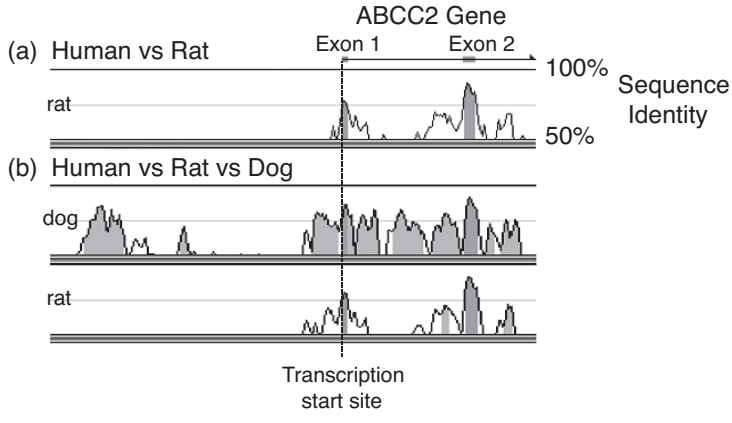


Figure 6.3 Regulatory regions. VISTA output for an alignment upstream of the *ABCC2* gene for (a) human and rat, (b) human, rat and dog.

However, panel (b) shows a similar output, but using a multiple alignment between human, rat and dog sequences. Once again, we can see the high identity between the exons from each species, but this time there are areas of high identity within the upstream regions. Immediately upstream of exon 1 is a region of high identity that stretches for approximately 1 kb; this is likely to be the promoter region. You will also note other regions of identity further upstream, and these may represent further regulatory regions for this gene, which may also be targets for further experimental confirmation.

The approach above has the advantage that is quick and simple to undertake, but the distinct disadvantage that this is only a computer prediction, which must be confirmed by later experimental work. An alternative approach to locate the promoter of a specific gene could therefore be to clone fragments of DNA around and upstream from the transcription start site of a gene, using a vector that carries a promoterless reporter gene (this is known as a *promoter probe vector* – see Figure 6.4). Expression of the reporter gene will only occur if an active promoter is inserted at the cloning site. Or you can perform genome-wide screening for promoters with specific function, for example, by making a library of random DNA fragments in a promoter probe vector, and selecting those clones that exhibit expression of the reporter under the chosen conditions. In Chapter 8 we describe the related technique of *enhancer trapping*, which involves the use of a transposable element carrying a reporter gene with a weak promoter.

This approach will identify a DNA fragment that has promoter activity, but will not tell you which specific regions within that fragment are important for gene expression, nor which transcription factors may interact with these regions. You can extend the approach to provide a more precise

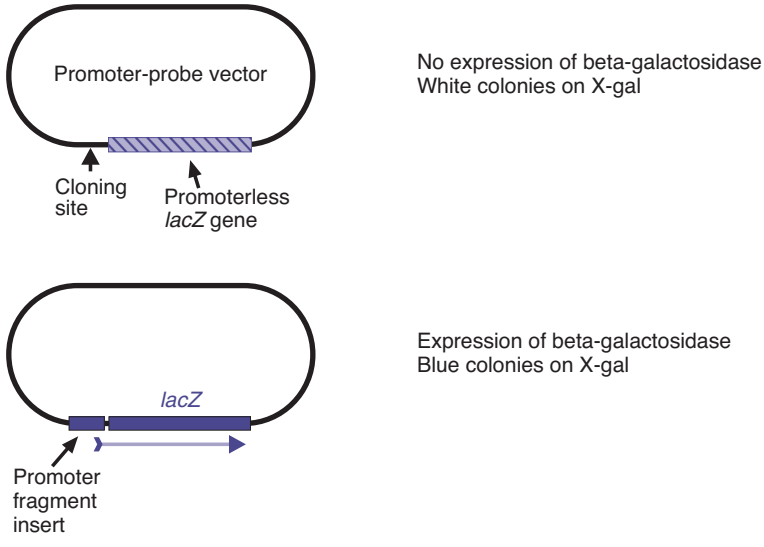


Figure 6.4 Use of a promoter-probe vector.

identification of the sequences that are necessary for promoter activity. At the simplest level, this could involve making a series of deletions from one or both ends of the fragment, and seeing how much you can remove without affecting promoter activity. Other techniques that we will describe later (Chapter 7) can be used to alter specific bases within the promoter in order to assess the sequence requirements for initiation of transcription. This approach can also be used to identify regulatory sequences that affect the activity of the promoter, usually through binding of regulatory proteins. Subsequent sections in this chapter examine other methods for investigating the DNA binding of RNA polymerase, transcription factors and regulatory proteins.

6.2.2 Reporter genes

Reporter gene technology allows us to examine gene expression within a cellular context, in a simple and reliable fashion. This is invaluable, for example, in the study of cellular differentiation in a multicellular organism, where any single promoter element may behave differently in different cells or tissues. In addition, reporter gene studies are very important in the initial analysis of a promoter and the transcription factors that may bind to it (see below), and are a crucial preamble to the construction of a transgenic organism (see Chapter 11).

Reporter gene analysis involves the use of gene cloning technology to make a construct in which a gene that codes for a readily detectable product

(the reporter) is attached to the promoter from the gene being investigated (for simplicity, we will mainly refer only to promoters, but similar methods are used for studying other regulatory elements). Examples of commonly used reporters include beta-galactosidase (detected using a chromogenic or fluorogenic substrate), green fluorescent protein (GFP; a naturally fluorescent protein originating from a jellyfish) or firefly luciferase (where enzymatic activity is detected by the emission of visible light). The simplest procedure is then to introduce this construct, carrying a reporter gene driven by a promoter, into the organism in question. Cells in which the promoter is activated will show detectable and quantifiable expression of the reporter. A distinct advantage of using reporter genes is their enhanced stability. If expression occurs only transiently, detection of the native mRNA may be difficult, but the greater stability of the reporter protein will allow the identification of the time at which expression is activated. (It follows from this, however, that it is less easy to determine when expression is switched *off*.)

One important disadvantage of using reporter genes on a plasmid is that the system is not free from artefacts. This is because DNA sequences do not exist as naked DNA *in vivo*, but are packaged in higher order structures (*chromatin*). These higher order structures are important modulators for the expression of a gene, and fragments of DNA, taken out of context, may show quite different effects in their ability to promote transcription. Some promoters do not work properly (or their regulation is quite different) when located on a plasmid; conversely, some DNA fragments that produce positive signals on a promoter-probe vector are subsequently found not to be 'real' promoters. It is thus important that, while reporter gene assays are a common first step in the analysis of gene expression, complementary approaches are used to ensure that results are relevant to the *in vivo* situation. In prokaryotes, reporter genes often give more reliable results when inserted into a specific chromosomal position rather than in the artificial environment of a plasmid. In eukaryotes, we tend to use more complex methodologies such as chromatin immunoprecipitation (see Section 6.3.4), to examine gene expression within the chromatin context.

A further application of reporter proteins, for ascertaining the localization of a protein within the cell, is described later in this chapter.

Reporter gene assays based on joining a promoter to a reporter gene have been very useful both for dissecting the components of the promoter and for quantifying its activation in different situations. But of course it has limitations. If you join your promoter to a gene that encodes a completely different protein, your readout will be a function of the intrinsic properties of that mRNA and of that protein. These include the stability of the mRNA and the protein, as well as the efficiency of translation. However, as long as you compare like with like – the activation of the promoter, *as measured by its effect on reporter gene levels*, under different circumstances – you can make a

convincing case that your observations are comparable as a measure of promoter activation, because you are keeping all other parameters constant.

However, you may (especially in eukaryotic systems) wish to study other parameters that affect the levels of the transcript encoded by your gene, such as the 5′-untranslated region (UTR) and its effect on translation efficiency, or the 3′-UTR and its effect on mRNA stability. In these cases, you will wish to keep all other factors constant, including the promoter, and vary only the UTR regions. For such studies, you choose a vector that contains your reporter gene under the control of a constitutively active promoter without cellular or organismal specificity, such as that of the large T antigen of the SV40 virus. You can then slot the whole or part of your 5′-/3′-UTR on either side of your reporter gene. By comparing the amount of reporter gene product in different constructs, and in the native vector, you are able to gauge the positive or negative impact of different elements on mRNA *translatability*.

6.3 Regulatory elements and DNA-binding proteins

The most direct, precise and reliable way of locating a promoter, or any other DNA sequence that influences transcription through the binding of specific proteins (which includes many but not all regulatory elements), is to detect the binding of those proteins to specific DNA fragments. Four methods are worth describing briefly: yeast one-hybrid assays, DNase footprinting, gel retardation assays and chromatin immunoprecipitation (ChIP).

6.3.1 Yeast one-hybrid assays

The purpose of the yeast one-hybrid system is to identify proteins that interact with a defined DNA sequence, or conversely to locate DNA regions that will bind to known transcription factors. In the example illustrated in Figure 6.5, tandem copies of the DNA sequence under investigation (known as the *bait*) are inserted upstream from a reporter gene. This construct is then integrated into the yeast genome. In the absence of transcription factors that bind to the bait, the reporter gene will not be expressed. However, if you transform the recombinant yeast strain with a cDNA expression library, any clones that express a transcription factor (the *prey*) that can interact with the bait will activate transcription of the reporter gene (Figure 6.5b).

The versatility of the system can be extended by exploiting the bipartite nature of eukaryotic transcription factors. This means that in general they consist of two domains: one part of the protein (the DNA-binding domain) makes specific contacts with certain DNA sequences, while the other part (the activation domain) is responsible for activating transcription. These

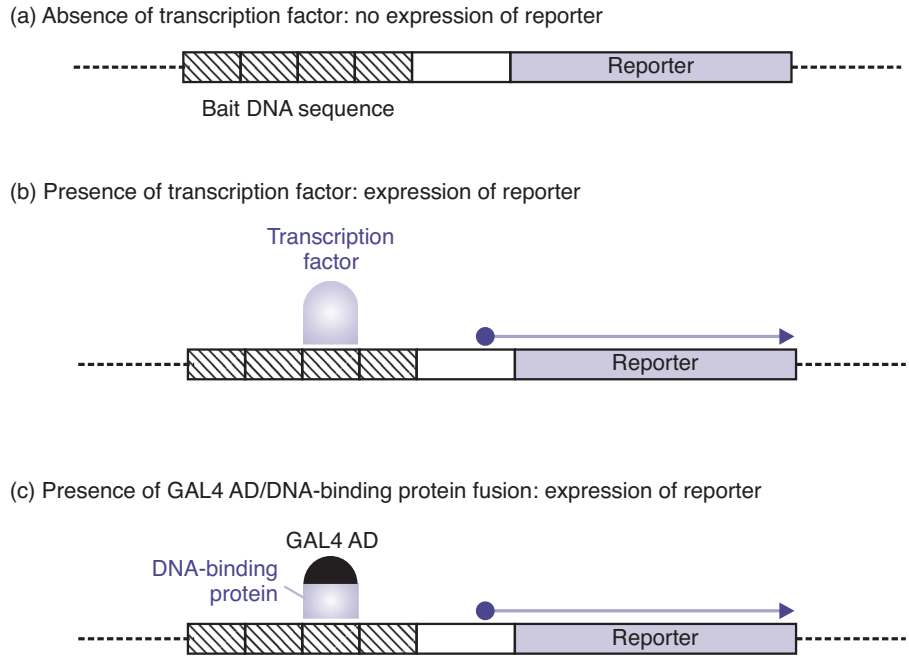


Figure 6.5 Yeast one-hybrid system.

domains can be separated, and function independently, so the activation domain (AD) can be joined to other DNA-binding proteins, when it will activate transcription from other sites according to the specificity of the new DNA-binding domain. The best-known example is the yeast activator GAL4, which normally controls the transcription of genes involved in galactose metabolism. If the DNA-binding domain of GAL4 is removed and replaced with another DNA-binding domain, the hybrid protein will activate transcription of a different set of genes. To take advantage of this, the cDNA library is made using a special vector containing the GAL4 AD so that the products will be expressed as fusion proteins with the GAL4 AD. If the product binds to the bait sequence, the GAL4 AD will activate transcription of the reporter (Figure 6.5c). Since the activation is provided by the GAL4 AD, this system can detect a wider range of DNA-binding proteins, and is not limited to those that are themselves transcription factors.

One potential complication with the yeast one-hybrid system is that we are expressing our prey proteins in a single-celled eukaryotic organism. Whereas many proteins from higher eukaryotes (i.e., mammals) will fold correctly in yeast, not all will, or to interact with the bait sequence these proteins may require specific cofactors that are not present in yeast. Either of these scenarios would give us false negatives in the assay, suggesting no interaction when

one really exists *in vivo*. The obvious solution to this is to perform the assay in mammalian cells, and in particular the cell type that we are interested in; such assays are undertaken in exactly the same way as the yeast one-hybrid assay, but are referred to as, not surprisingly, mammalian one-hybrid assays.

An extension of this concept, the yeast two-hybrid system, is used to study protein–protein interactions, and is considered in Chapter 10.

6.3.2 DNase I footprinting

In *DNase I footprinting* (or just *footprinting*), you identify the part of the suspected promoter sequence to which transcription factor proteins bind. Nuclei are isolated from whole cells (whether a cultured cell line or a whole tissue), and proteins isolated from the nuclei. These proteins, or a specific protein purified from the mixture, are then mixed with the DNA fragment you wish to analyse. A parallel control reaction contains DNA without protein. In both reactions, the DNA is radiolabelled at the 3′ end. You then add the nuclease DNase I to both reactions. DNase I will cleave all the phosphodiester bonds in the DNA chain that it can reach. The reaction is not allowed to go to completion; thus, as a result, the DNA in your control reaction will be fragmented randomly into a continuous ladder of DNA fragments of all possible sizes between full size and one base. However, in the other reaction, DNase I will be unable to access its substrate wherever a *cis*-acting DNA element has been protected by a bound *trans*-acting protein. Thus, if you run these two samples out on a sequencing gel side by side, you will find that the continuous ladder as seen in the control is interrupted in these protected regions (Figure 6.6). Since DNase digestion is random, there will of course be a lot of other products as well as those shown – but as you are using autoradiography to detect the label attached to the DNA, you will only detect those fragments that contain the 3′ end of the original DNA molecule. By running out a sequencing reaction as well, you will be able to tell exactly which regions of DNA were protected. Note, however, that the extent of the DNA that is protected by a specific protein is usually larger than the region that is necessary to make specific contacts with the protein.

6.3.3 Gel retardation assays

Another way of determining if a particular sequence of eukaryotic DNA is involved in transcription factor binding is a *gel retardation* or *electromobility shift assay* (*EMSA*). Here, a specific DNA fragment of interest is labelled and then mixed with nuclear proteins, or a specific protein of interest. If one or more of these bind to the DNA fragment, its mobility in a polyacrylamide or agarose gel will be decreased. This can be easily detected by comparing it to a control sample without added protein (Figure 6.7). However, binding

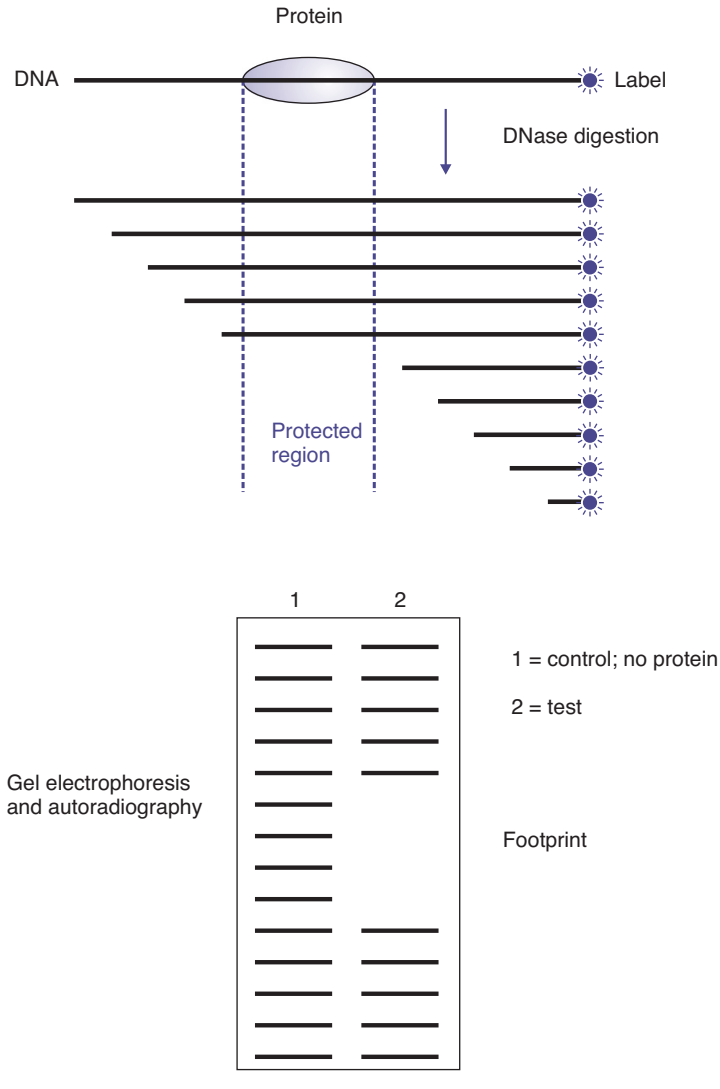


Figure 6.6 DNase I footprinting.

of a protein to DNA may occur in a specific manner (i.e., the targeting of a transcription factor to its appropriate response element), or in a non-specific manner (i.e., the random association of the positively charged DNA-binding domain of the transcription factor with a negatively charged DNA fragment). To distinguish between these two phenomena, it is important to use a competition assay. In this experiment, increasing amounts of unlabelled DNA are added to the reaction, and if the reaction is specific then we will be able to

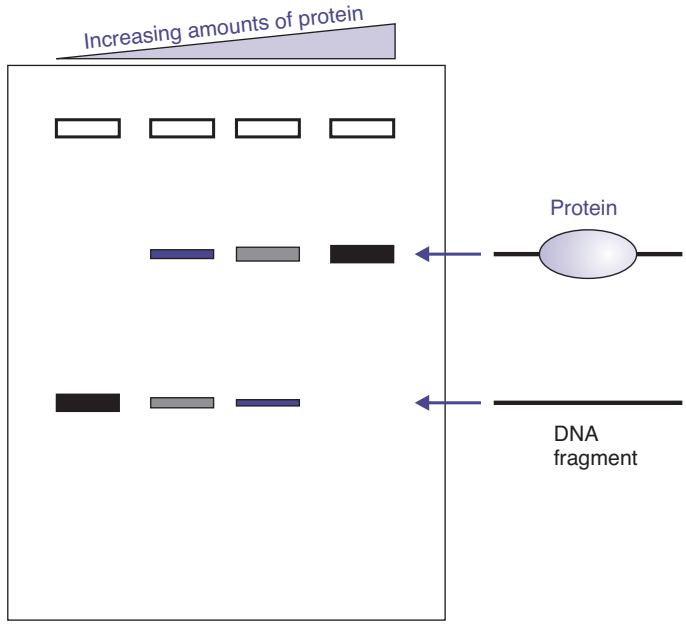


Figure 6.7 Gel retardation assay.

compete away the shifted band. In comparison, non-specific binding will not be removed in such a manner.

An advantage of both footprinting and gel retardation assays is that they are applicable to any protein that binds to specific DNA sequences, and are not limited to those proteins that regulate transcription of the DNA. As we saw with reporter gene assays, we must be careful when interpreting data from naked DNA, not bound in higher order chromatin, and such is the case with both footprinting and gel retardation assays. Although it is possible to undertake footprinting *in vivo*, this is complex and time-consuming. Instead, it is more common to use techniques such as chromatin immunoprecipitation to study these interactions within the chromatin context.

6.3.4 Chromatin immunoprecipitation (ChIP)

Chromatin immunoprecipitation (ChIP; Figure 6.8) starts with breaking the cells open to produce a mixture of genomic DNA, proteins and other cell components. In this mixture, any DNA-binding proteins will remain attached to their target DNA-binding sequences. Treatment with a cross-linking agent such as formaldehyde makes the binding irreversible. The DNA, with the bound proteins, is fragmented, for example, by sonication, so that the average size of the DNA fragments is around 500 bp. To isolate the DNA that

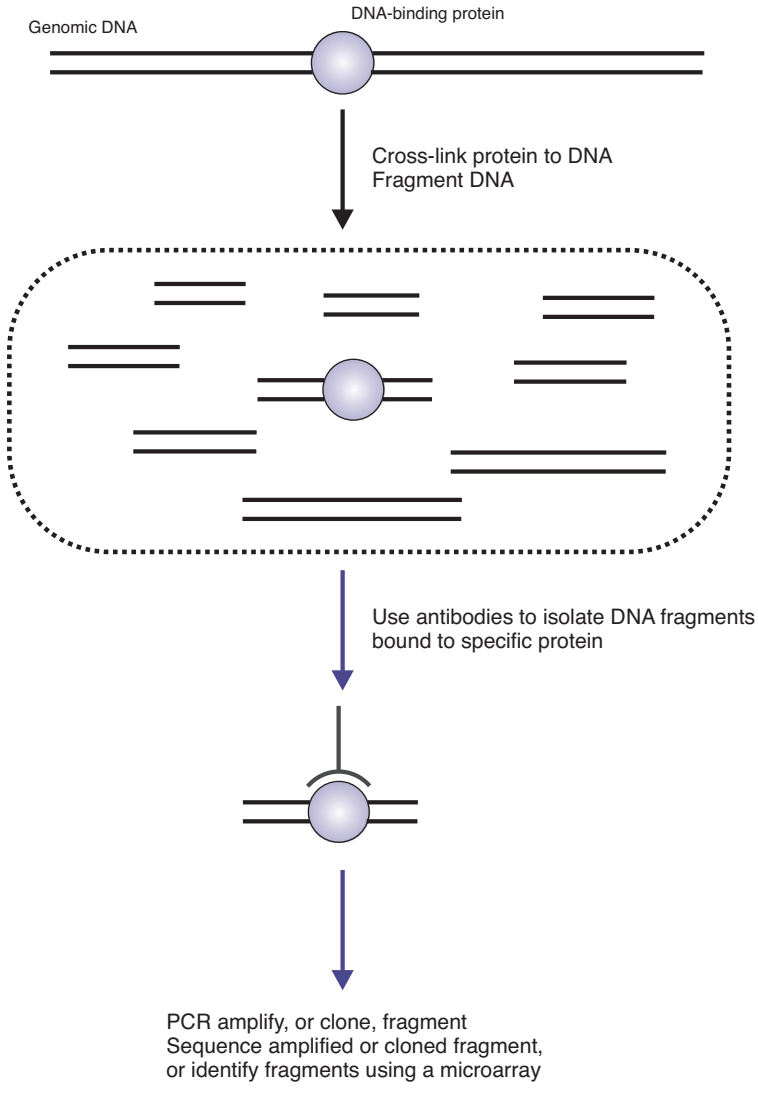


Figure 6.8 Chromatin immunoprecipitation (ChIP).

is bound to the chosen DNA-binding protein, an antibody against the specific protein is added, precipitating the protein–DNA complexes. The purified protein–DNA complexes are then heated to allow the DNA to be separated from the proteins. The identity of the DNA fragments isolated can then be determined by cloning and sequencing them. The precipitated complexes should contain DNA fragments containing all the DNA targets to which the protein was bound at the moment of cross-linking.

The rate-limiting steps for ChIP are the cloning and sequencing steps. To make the process more efficient, ChIP can be combined with microarray techniques (see Chapter 9) to produce a method that eliminates the need for cloning. The process, called *ChIP-on-Chip*, follows the same general procedure but the DNA that is released from the protein complex is amplified by PCR and labelled with a fluorescent tag. The labelled fragments are then washed over a DNA microarray that represents all the genes of a genome sequence, thus identifying the bound DNA fragments.

6.4 Translational analysis

Previously we have discussed the various methodologies for studying gene expression. However, it is important to remember that production of an mRNA does not necessarily mean that an active protein will be produced. It is, therefore, important that any studies undertaken at the level of mRNA are balanced with experiments on protein levels and, preferably, activity. Indeed, the methods for studying protein expression were developed earlier than DNA cloning methods. For example, it was possible to sequence proteins (if they could be purified in large enough amounts) before gene sequencing was developed. When gene cloning techniques became available, the characterization of DNA and mRNA became much easier than sequencing their protein products, with many scientists switching to these methods. However, seasoned protein biochemists have gleefully witnessed the renaissance of their craft, now renamed proteomics (see Chapter 10). Not least the realization that the human genome contains many fewer genes than there are proteins has reminded molecular biologists that gene products must be studied on both levels (to say nothing of the actual effect on the cell or organism as a whole) in order to understand how a gene works.

6.4.1 Western blots

The conventional way of analysing the proteins produced by a cell involves electrophoresis through a polyacrylamide gel in the presence of sodium dodecyl sulphate (SDS polyacrylamide gel electrophoresis, or SDS-PAGE). The presence of SDS denatures the proteins by disrupting non-covalent bonds that are needed for the stability of their three-dimensional structure, and this ensures that separation is purely on size alone and not three-dimensional structure. In general, there are too many proteins present to give a clear picture of the complete protein profile if you simply stain the gel with a general protein stain, although much higher resolution can be obtained by two-dimensional gel electrophoresis (see Chapter 10). In a one-dimensional SDS-PAGE, you will usually only be able to see the major proteins. Even then, it is not always easy to identify any specific ones within this mixture.

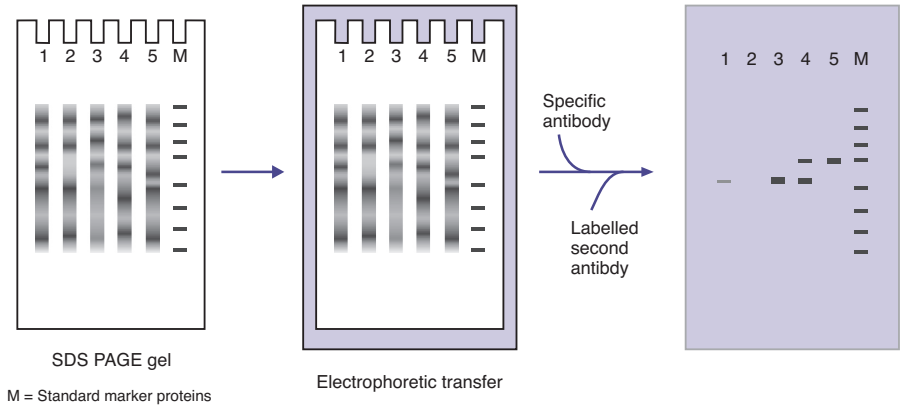


Figure 6.9 Western blotting.

The generally applicable technique for the detection of a specific protein, and analysis of its expression, relies on the use of specific antibodies, in combination with SDS-PAGE. Following size separation of the cell extract by SDS-PAGE, the proteins are transferred to a membrane by running a perpendicular current through the gel into the membrane, and a specific protein or proteins detected using *antibodies* (see Chapter 3). These antibodies can be either labelled themselves, or, much more commonly, detected by a second labelled antibody. For example, if the primary antibody, which recognizes your target protein, was raised in a mouse, then binding of that antibody can be detected by a second, labelled antibody raised in a second species, such as rabbit, that recognizes the non-variant region of the mouse immunoglobulin molecule. The perceived analogy with a Southern blot led to this technique being termed *Western blotting*. The technique will identify not just the presence or absence of a protein that reacts with the antigen, but also its size and an estimate of relative levels of expression. In Figure 6.9, we see that expression was not detected in sample 2, and only weakly in sample 1, while the protein detected in track 5 was a different size, perhaps due to post-translational modification; both proteins were present in track 4. As we saw with qPCR it is important for our samples to be standardized if we wish to undertake comparative quantitation between them. Once again, the measurement of a product, a protein in this case, that is expressed at the same level in all samples is used to standardize. Commonly, the structural protein actin is used.

The major considerations in the use of antibodies for molecular biology were covered in Chapter 3 (see Box 3.3). The main point to remember is that proteins in an SDS-PAGE gel are denatured, so the antibody must be capable of recognizing a linear epitope. For recombinant proteins, you also require an antibody that binds to a non-glycosylated epitope, as recombinant proteins may not be correctly glycosylated. The complex mixture of antibodies

in conventional polyclonal antiserum is virtually guaranteed to contain some suitable antibodies, but an individual monoclonal antibody may fail one or both criteria.

Later on (Chapter 10) we will look at further methods for identifying proteins in one- or two-dimensional gels, especially the use of mass spectrometry.

6.4.2 Immunocytochemistry and immunohistochemistry

This method (differently named depending on whether we are working with cells or tissues) relates to Western blotting in the same way that *in situ* hybridization relates to Northern blotting. The two blotting methods allow us to determine whether the probe (either an antibody or a nucleic acid probe) is specific, and if it is, how many polypeptides or transcripts it recognizes, and what size they are. The cytological or histological method, by contrast, allows us to determine which cell type the gene product is localized to, or in which area within or outside of the cell. Apart from being much easier to perform and less prone to artefacts than *in situ* hybridization, the study of the translated protein product has an added bonus. The *intracellular* localization of an mRNA, as determined by *in situ* hybridization, is rarely particularly revealing. However, the destination of the translated protein shows much greater variation, and this is, generally, more obviously linked to their function. Some proteins are cytoplasmic, some are membrane-bound; some gather at one pole of the cell, and others re-enter the nucleus. The localization of a protein can give us important clues to its function, and indeed the incorrect transport of a protein can give us important information about a disease phenotype. Such localization experiments can be undertaken using antibodies, in a similar way to that which we saw for Western blotting.

In addition, reporters such as GFP can also be used to study the fate of specific proteins within the cell. By fusing the *gfp* gene to a construct coding for the protein in question, we can easily see when the protein is expressed, and where it ends up (e.g., in the cytoplasmic membrane) simply by detecting the location of the fluorescence using a microscope. Another advantage over Western blotting is that whereas such blotting reflects a single point in time (i.e., when the samples were prepared), the localization of GFP fusion proteins can be studied in real time. This means that we can not only see where a protein is located, but also what happens to that protein localization following a stimulus.

7
Products from Native and Manipulated Cloned Genes

In the preceding chapters we looked at ways of characterizing individual genes and analysing their expression. Genetic techniques can be used for modification as well as analysis, and in this chapter we focus on ways in which genes can be manipulated for product formation – whether of the natural product of that gene, modified versions of the product or, ultimately, entirely novel proteins. In Chapter 11 we will extend the discussion to consider the establishment of transgenic plants and animals.

We use proteins, or smaller polypeptides, in many ways – ranging from enzymes that can be added to washing powders, to hormones that are used for treating medical conditions such as diabetes. Some of these proteins can be extracted from starting material that is readily available, such as plant material or microbial cultures, although the ease and cost of these procedures vary. But often the potential source of such proteins is scarce or difficult to obtain, which severely limits the application of this approach – or may even make it impossible. This applies especially to proteins and polypeptides derived from human sources, which may well be the most appropriate for treating specific diseases. The classic example here is of human growth hormone (somatotropin), which is used for treating a condition known as pituitary dwarfism, where a child's growth is affected by a deficiency in the production of growth hormone by the pituitary gland. For a long time, the only source of this hormone was from pituitary glands removed from the bodies of people who had died from a variety of causes. The limitations on supply can be easily imagined, but the safety implications, in terms of the potential transmission of disease from the donor by this route, were only partly appreciated at the time. These safety concerns were highlighted in the 1990s by the discovery

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

that spongiform encephalopathies (e.g., Creutzfeldt–Jakob disease, or CJD) could be transmitted by such material.

Many bacteria and other microorganisms, including unicellular eukaryotes such as yeast, can be grown easily and cheaply in essentially unlimited quantity. They thus represent ideal hosts for the expression of the gene coding for the required protein to enable us to obtain large amounts of product – especially as we can manipulate the gene to maximize its expression and/or characteristics. Making a therapeutic product in this way also removes the possibility of transmitting infectious agents, from human or animal tissues, so it is much safer than using product derived from 'natural' sources. It should also be emphasized that (apart from any post-translational modifications, which will be discussed later in this chapter) the purified product obtained from recombinant bacteria is exactly the same as the pure product from its original source. The only way in which the product differs is that it is likely to be purer and will be free from infectious agents.

The realization that the new genetic technologies would provide a way of making products that had hitherto been expensive or impossible to produce was a major driving force behind the commercial side of the biotechnology revolution. The research implications are even more dramatic. Some proteins, especially those responsible for regulating cellular activities and for communication between cells, are produced naturally in tiny amounts. It is likely to be effectively impossible to extract and purify these proteins in sufficient quantities to be able to characterize their structure or to analyse the way they work. Yet in principle, as long as you can identify and clone the gene responsible, it will be possible to express that gene in a microbial host and obtain substantial quantities of the product.

7.1 Factors affecting expression of cloned genes

Expressing a foreign gene in a bacterial cell is not entirely straightforward. The expression signals that control transcription and translation can be quite different from one organism to another, especially if you are moving a gene from a eukaryotic source into a bacterial host. Expression can also be affected by the base composition of the gene, and by the codon usage. These factors were covered in Chapter 1, but are summarized here for ease of reference (see Figure 7.1). We are generally assuming, to start with, that the host organism for these expression experiments is a bacterium. The expression of foreign genes in eukaryotic cells is considered later on.

7.1.1 Transcription

The main factor to be considered here is the properties of the promoter, i.e., the region where the RNA polymerase binds to initiate RNA synthesis

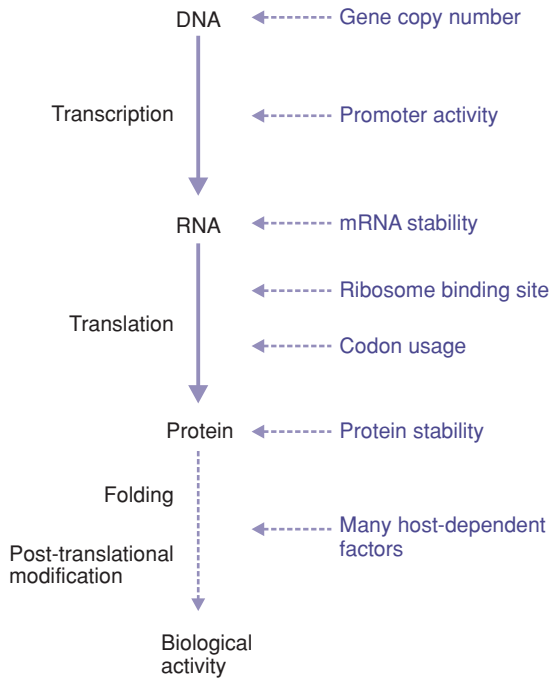


Figure 7.1 Gene expression.

(Figure 7.2). This is the principal step where bacteria control which DNA regions are to be expressed, and how strong that expression should be. The structure of the RNA polymerase (and especially of the sigma factors that determine the specificity) and the sequences of the promoter regions have evolved together to produce this careful regulation of gene expression. It is therefore not surprising that a gene from one species may not be expressed when inserted into a different host.

Transcription signals in eukaryotes are substantially different from bacterial promoters, which would limit expression of any gene under control of a eukaryotic promoter in a bacterium. However, this can usually be easily overcome by replacing these promoters with appropriate bacterial ones. For expression of eukaryotic genes in bacteria, you will usually require cDNA rather than genomic DNA. This removes introns, which are not needed and, most importantly, cannot be removed by the bacterial cell. You also have to provide a bacterial promoter and any other necessary bacterial expression signals upstream of this cDNA sequence. This is typically achieved by cloning your sequence into an *expression vector*. This concept is considered further later in this chapter.

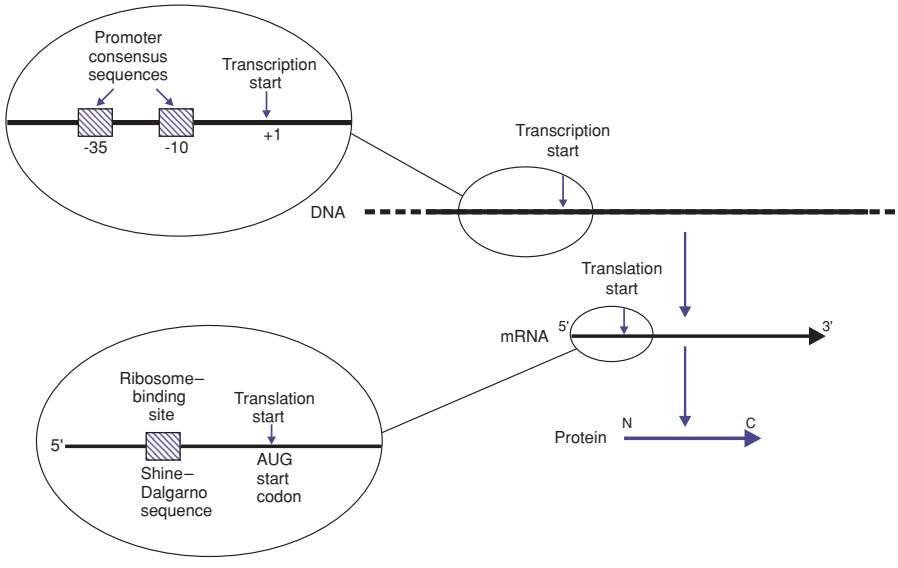


Figure 7.2 Principal factors in bacterial gene expression.

7.1.2 Translation initiation

In bacteria, ribosomes normally bind to a region of the mRNA adjacent to the start codon, facilitated by a sequence (the Shine–Dalgarno sequence) that is (partially) complementary to the 3′ end of the 16S rRNA (Figure 7.2). This structure is reasonably well conserved between bacteria, such that you could expect to get some degree of translation initiation if you move a gene from one bacterium to another (assuming that transcription occurs). However, the exact structure may not be optimal; for example, the distance between the Shine–Dalgarno sequence and the start codon can affect the efficacy of translation initiation. So you may not obtain maximum gene expression unless you manipulate this region.

In eukaryotes, ribosomes bind to the mRNA in a rather different way. In essence, the small ribosome subunit binds to the capped 5′ end of the mRNA and scans along the mRNA until it finds a translation start site. This consists of a nine-base consensus sequence (known as a Kozak sequence) that contains a start codon. Here, the large subunit is added and translation starts. The absence of a bacterial-type ribosome-binding site means that if you choose a bacterial host for expression of eukaryotic genes, you will usually have to provide a ribosome-binding site, and a suitably positioned start codon, as well as a promoter, to get production of your protein.

The nature of the translation start codon may also have to be considered. In the standard genetic code, AUG is used to signal the start of translation.

However, in some organisms, a proportion of genes (sometimes the majority) use alternative start codons such as GUG, UUG or CUG. Trying to express such a gene in an organism with a stricter preference for an AUG start may limit or prevent translation, or give a product with a different N-terminus. The simple solution to this is to engineer your sequence to replace these alternative initiator codons with the more common AUG, which will be efficiently recognized in your bacterial expression system.

7.1.3 Codon usage

In the standard genetic code, which is basically conserved between all living organisms, there are many sets of *synonymous codons*, i.e., several codons that code for the same amino acid. For example, any one of six different codons in the mRNA will result in incorporation of leucine into the polypeptide (see Box 3.2). However, these different codons are not completely equivalent. The cell will use some codons more readily than others, depending on the availability and specificity of tRNA molecules that recognize these codons. One consequence of this is that any occurrence of the less readily used codons will slow down translation, which could result in premature termination of the polypeptide chain. The sequence of the genes in any given organism will have evolved to match the availability of tRNAs (and/or vice versa), with some synonymous codons occurring more frequently than their alternatives – we refer to this as *codon bias*. In a different organism, the tRNA population is different, and the codon usage will also be different, and we must consider this when expressing a protein in an alternative host. The extent to which this is a problem is related to the level of gene expression. The more we have enhanced the level of transcription of our cloned gene, the more likely it is that codon usage will become a limiting factor.

In extreme cases, you may actually fail to get expression at all. The host organism may be virtually unable to recognize certain codons in your cloned gene (and these would be absent from natural genes in that host); in effect, these are acting as additional stop codons, and will cause premature termination of protein synthesis (remember, the standard stop codons terminate translation primarily because there is no tRNA that can bind to them).

The converse can occur, too. In some organisms, a codon that is a stop signal in the standard code is actually used by that organism to code for a specific amino acid. For example, in some bacteria UGA codes for tryptophan rather than being a stop codon. If you try to express such a gene in *E. coli*, you are likely to get premature termination at the UGA codon. It is possible to overcome this by using *E. coli* host strains that have been engineered to produce additional types of tRNA.

So, how do we get around the problem of codon bias? There are several potential fixes: first, we can just ignore it, and hope that any loss of efficiency in translation is not sufficient to prevent expression of our protein product. This is often the first approach tried, and is met with mixed success. Second, we can alter the coding sequence of the cDNA, such that we replace any codons for rare tRNAs with those that represent more common tRNAs in the host cell. While potentially possible, this could be time consuming, especially for longer cDNA sequences. Finally, there are now a number of bacterial strains that have been engineered to express tRNAs that are normally rare in that host. This is achieved by stably integrating expression plasmids for these tRNAs into the bacteria genome. As an example, the Rosetta strains of *E. coli* (named after the Rosetta stone, the discovery of which enabled the translation of messages written in hieroglyphs) have been engineered to express six tRNAs that are rare in wild-type *E. coli*, but commonly used in mammals. Such an approach comes with some complications for maintaining and growing the cells, but can vastly improve one's ability to express a mammalian protein in a bacterial host.

7.1.4 Nature of the protein product

We also have to remember that the nature of the protein product may influence the amount of protein that is recovered. Most obviously, the stability of the protein can have a major effect: an inherently unstable protein will not be recovered in high yield, even if we have maximized the rate of production. Amongst other factors, the location of the product can have an important influence of the levels obtained. If the protein remains in the cytoplasm, it may become insoluble at high rates of synthesis, leading to the production of *inclusion bodies*, or aggregates of insoluble product. Or, if the protein becomes inserted into the cytoplasmic membrane, it is likely to have a deleterious effect on the functions of the membrane of an organism to which it is alien; expression of such proteins may be lethal, even at quite low levels.

For high levels of protein production from bacterial cultures, it is often advantageous to express the product in a form that can be secreted into the culture medium, for example, by attaching *signal sequences*. Unfortunately, attaching secretion signals may not work. Proteins that are not naturally secreted may not pass properly through the cytoplasmic membrane, even when directed to do so by a secretion signal.

We also need to bear in mind that our aim is usually the production of a biologically active protein, rather than a simple polypeptide chain. Biological activity often depends on a variety of post-translational effects, ranging from folding of the polypeptide into the correct three-dimensional structure, to post-translational cleavage or modifications such as glycosylation. This can be a difficult barrier to surmount when attempting to express a protein in a

foreign host, and is often a major reason for using a non-bacterial host for expression of eukaryotic proteins.

7.2 Expression of cloned genes in bacteria

From the above discussion, it can be seen that if you take a DNA fragment from another organism and clone it in *E. coli*, there are many reasons why it may not be expressed. The basic way of encouraging (although not ensuring) expression of the cloned gene is to incorporate expression signals into the vector, adjacent to the cloning site. This is then known as an *expression vector* (Figure 7.3).

Expression vectors are of two main types. On the one hand, if the vector just carries a promoter, and relies on the translation signals in the cloned DNA, it is referred to as a *transcriptional fusion* vector. On the other hand, if the vector supplies the translational signals as well (so you are inserting the cloned fragment into the coding region of a vector gene), then you have a *translational fusion*.

7.2.1 Transcriptional fusions

If you are using a transcriptional fusion vector, which carries a suitable promoter adjacent to the cloning site (Figure 7.4), insertion of the cloned fragment at the cloning site (in the correct orientation) will put that gene under the control of the promoter carried by the vector. Thus, transcription initiated at the promoter site will continue through the cloned gene, resulting in a fusion of the gene and the promoter into a single transcriptional unit – hence it is referred to as a *transcriptional fusion* (cf. translational fusions; see below). This is very similar to the concept of a reporter gene (see Chapter 6), but for a different purpose. A reporter gene is used to study the activity of

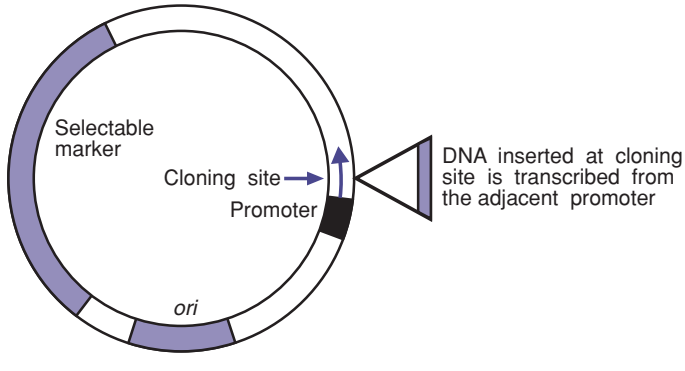


Figure 7.3 Basic features of an expression vector.

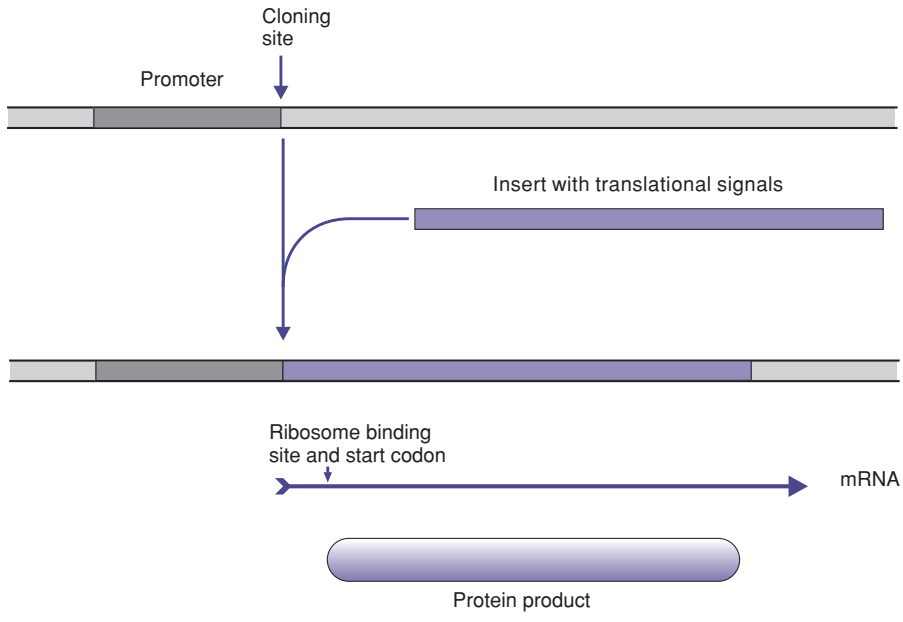
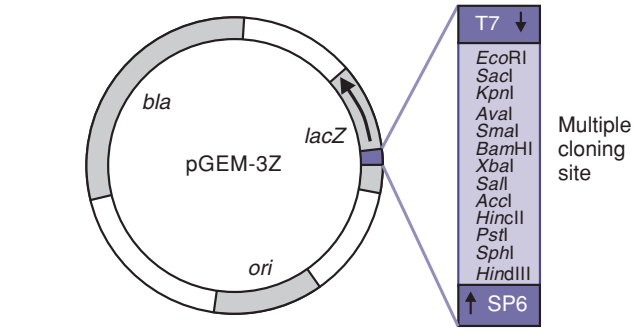


Figure 7.4 Expression vectors: transcriptional fusions.

the promoter. Here we know what the promoter does, and we are using it to stimulate transcription of the cloned gene.

In *E. coli* there are a wide range of suitable promoters that can be used in this way. Amongst naturally occurring promoters, the *lac* and *trp* promoters that normally drive expression of the *lac* (lactose utilization) and *trp* (tryptophan synthesis) operons respectively, and the P_L promoter, from bacteriophage lambda, are examples of commonly used promoters, the P_L promoter being one of the strongest natural promoters in *E. coli*. Alternatively, genetically modified promoters can be used, one of the commonest being known as the *tac* promoter, since it represents a hybrid between the *trp* and *lac* promoters, and it is capable of higher levels of transcription than either of the parent promoters.

To make full use of an expression vector you may want to be able to switch gene expression on or off. This allows you to maintain your bacterial culture under conditions optimal for growth, and then induce expression under different conditions that you have optimised for successful production of your protein of interest. However, the regulation of most inducible bacterial promoters, such as that of the *lac* operon, is rather 'leaky' – i.e., there is still some expression even in the uninduced or repressed state. Such expression may result in reduced efficiency of bacterial growth. More robust control can be achieved by the use of promoters from bacteriophages, notably one from the bacteriophage T7. In T7, this promoter controls the expression of the 'late'



bla = beta-lactamase (ampicillin resistance)
ori = origin of replication
SP6, T7 are specific promoters recognized by the SP6 and T7 RNA polymerases respectively

Figure 7.5 Structure of the expression vector pGEM-3Z.

genes, i.e., the genes that are only switched on at a late stage of infection. This promoter is not recognized by *E. coli* RNA polymerase, but requires the T7 RNA polymerase, a product of genes expressed earlier in the infection cycle. So if we clone our DNA fragment downstream from a T7 promoter, using an 'ordinary' *E. coli* host (lacking a T7 polymerase gene) we will get no expression at all. This can be useful, as the product might be deleterious to the cell. Once we are satisfied that we have made the right construct, we can isolate the plasmid and put it into another *E. coli* strain that has been engineered to contain a T7 polymerase gene, and hence will allow transcription of the cloned gene. If the expression of the T7 polymerase gene is itself regulated, for example, by putting it under the control of a *lac* promoter, then we can regulate the activity of the T7 polymerase (by altering the concentration of the inducer IPTG) gaining control over the *level* of expression rather than just a simple 'on' or 'off' switch.

The pGEM® series of vectors (see Figure 7.5) provides an example. In this case, there is a multiple cloning site adjacent to the T7 promoter, so any DNA inserted will be under the control of the T7 promoter. This is a transcriptional fusion vector, and is often useful for generating substantial amounts of an RNA copy of your cloned fragment, which can then be used as a probe for hybridization. The vector has a second specific promoter, derived from another bacteriophage (SP6), which is another example of a promoter that is not recognized by the *E. coli* RNA polymerase. The SP6 promoter is located at the other side of the multiple cloning site, so if you provide an SP6 polymerase you will get an RNA copy of the other strand, the antisense strand. The usefulness of this will be apparent when we consider applications of antisense RNA.

In addition to the choice of promoter in our expression vector, we need also to consider the plasmid copy number. Most routine cloning vectors for use in *E. coli* are referred to as 'multi-copy' plasmids – but this can conceal a wide variation in the actual copy number. The early cloning vectors (such as pBR322) are normally present in 15–20 copies per cell (although under special conditions they can be amplified to up to 1000 copies per cell). Subsequent development of these vectors removed some control elements, meaning that many of the currently used vectors (such as the pUC series) have hundreds of copies per cell, and are referred to as 'high-copy' plasmids. More copies of the plasmid means more copies of the cloned gene, and hence (usually) more product formation. However, the relationship may not be linear; you will expect to get more product from a 200-copy plasmid compared to a 20-copy plasmid, but not necessarily 10 times as much.

So if you want to maximize gene expression, not only should you optimise the promoter, but you should also use a high-copy-number plasmid. Under these conditions (assuming translation works perfectly), you may get a bacterial clone in which your product represents up to 50% of the total protein of the cell. For commercial production, such high yields not only mean more product per litre of culture (which is of course important) but in addition the proportion of contaminating protein (and other material) that has to be removed is lower – thus reducing the costs of downstream processing. However, expressing large quantities of protein, and maintaining very-high-copy-number plasmids, brings problems, which we need to consider.

7.2.2 Stability: conditional expression

There is a downside to such high levels of product formation. Producing vast quantities of a protein that is useless to the cell is inevitably going to result in a reduction in growth rate, because of the resources that are being diverted in a manner that is non-productive, from the host cell's perspective. This applies even if the protein itself has no damaging effects. Slower growth rates will reduce the efficiency of the process. If the protein is directly damaging to the host cell, the problem becomes much more acute.

A bigger problem is that this slower growth rate for cells producing your product means that there is a very strong selective pressure in favour of any of a wide range of potential mutants that are non-productive. These include cells that have lost the plasmid altogether, as well as any mutation to the plasmid that reduces or prevents product formation. This can be a problem whether you are growing a few millilitres in the laboratory or many thousands of litres in an industrial fermenter. However, on a laboratory scale it is possible to include antibiotic selection in your culture, to ensure that any mutants that have lost the plasmid will be unable to grow. On a large scale, this is not only

an expensive solution, but also the disposal of large volumes of antibiotic-containing waste is a problem.

One way of ameliorating this situation is to use *controllable* promoters, i.e., promoters whose activity can be altered by changes in the culture conditions. The promoters listed above provide examples of such control. For example the *lac* promoter is naturally active only if *E. coli* is growing on lactose as a carbon and energy source. (Normally bacterial geneticists use IPTG – see Chapter 2 – instead of lactose, as it is both more convenient and is not broken down by the cell's beta-galactosidase; it is known as a *gratuitous inducer*.) In the absence of an inducer, a repressor protein binds to a DNA sequence known as the *operator* (which in this case overlaps with the *lac* promoter) and prevents transcription from the *lac* promoter. The inducing agent binds to the repressor protein, altering its conformation so that it no longer binds to the operator (Figure 7.6). We can grow the culture to an appropriate density

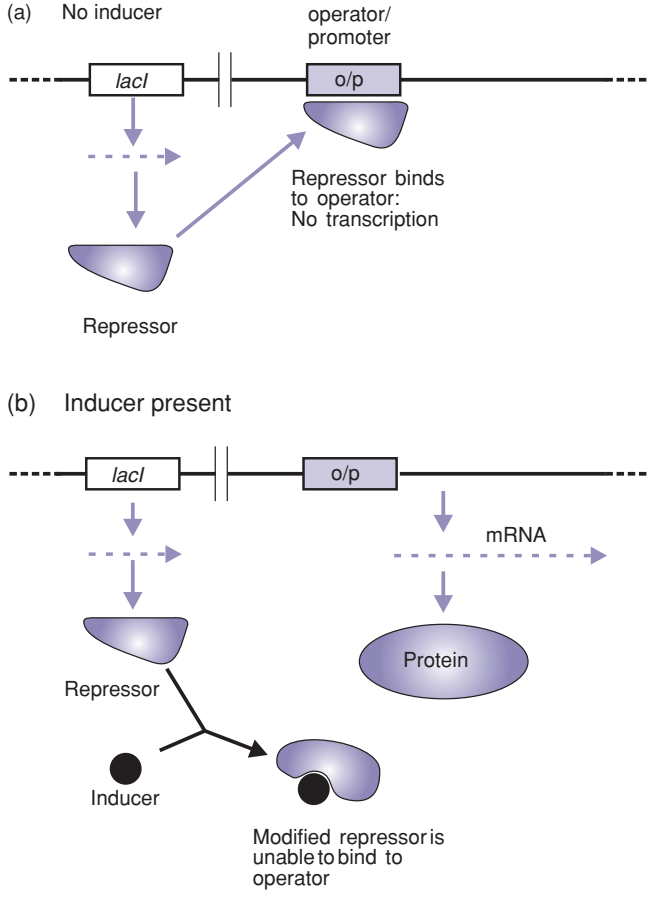


Figure 7.6 Regulation of the *lac* promoter.

in the absence of an inducer, so that we will not get high levels of gene expression and hence the selective pressure will not exist. Then, when we have enough cells, we add the inducing agent and switch on gene expression.

However, we have to remember that we are using a multi-copy vector. *E. coli* produces enough of the repressor protein to switch off the single copy of the promoter that it has in the chromosome, but this is not enough to switch off several hundred copies of this promoter. We refer to the repressor being *titrated out* by the presence of so many copies of the operator. So we have also to increase production of the repressor protein. The gene that codes for the repressor (the *lacI* gene) is not actually part of the *lac* operon; it has its own promoter. So one way of increasing production of the LacI repressor protein is by using a mutated version of the *lacI* gene that has a more active promoter (an *up-promoter* mutant). This altered *lacI* gene is known as $lacI^q$. Or we can put the *lacI* gene onto the plasmid itself, so subjecting it to the same gene dosage effect and therefore increasing the production of LacI. Commonly, we would do both – i.e., put a $lacI^q$ gene onto the plasmid.

Adding IPTG to a laboratory culture is fine, but on a commercial scale it is still not an ideal solution. Adding IPTG to an industrial scale fermenter would be very expensive. An example of an alternative strategy is to use a promoter such as the *trp* promoter, which controls transcription of the tryptophan operon. This is subject to repression by tryptophan; *E. coli* switches off expression of the tryptophan operon when the enzymes encoded by it are not needed, i.e., if there is a plentiful supply of tryptophan. It is possible to monitor and control the availability of tryptophan so that there is an adequate supply during the growth phase, and then limit the supply of tryptophan when expression is required. You may be puzzled by this. If we stop supplying tryptophan, how is the cell going to make the protein that we want, which will probably contain some tryptophan residues? However, it is possible to supply a low level of tryptophan which is not enough to switch off the *trp* promoter, but will still enable production of the required protein. And we can then feed the culture continuously with a low level of tryptophan so that protein production will continue.

This may solve the problem in part, by removing the selective pressure imposed by excessive product formation. But there is still some element of selection imposed by the presence of so many copies of the plasmid, which may also slow growth rates. We can counter this by using a plasmid with a different replication origin, so that replication is tightly controlled at only one or two copies per cell (see Chapter 2). However, the level of expression achieved with a low-copy vector will be less than that achievable with a multi-copy plasmid, other things being equal. We can adopt a similar strategy to that described above for programming gene expression, using a so-called *runaway plasmid*. If the control of plasmid copy number is temperature sensitive, then growing the culture initially at say 30°C will produce cells with only a few

copies of the plasmid. Then, once sufficient growth has been achieved, the culture can be shifted to a higher temperature, say 37°C; control of plasmid replication is lost and the copy number increases dramatically, until it represents perhaps 50% of the DNA of the cell. If we switch on gene expression at the same time, we will get a very substantial amount of product. Eventually the cells will die, but by that time we have enough of our product.

An alternative way of achieving a similar effect is by providing the vector with two origins of replication: one that results in many copies of the plasmid and a second that will produce only one or two copies per cell. If we can control which of the two origins is used, we can again switch conditions at an appropriate stage, so that the culture starts off with only a few copies of the plasmid per cell and then when we want to switch on gene expression we can, as well as inducing the promoter, switch to the other replication origin and increase the copy number of the plasmid.

7.2.3 Expression of lethal genes

So far, we have only considered the expression of proteins that are relatively benign to their *E. coli* host; at worst, their high level of production will slow the rate of growth of the bacterial culture. However, some genes code for products that are very damaging or even lethal to the host. Expression of such genes in *E. coli* poses obvious problems. You might consider that the above approaches would be equally applicable in such a situation, with just an acceptance that as you produce your protein, you kill your host. Unfortunately, it is not quite that simple: The regulation of many promoters is not tight enough. The *lac* promoter, for example, is active at a low level even in the absence of induction. If the gene product is *very* damaging, the cell will not be able to tolerate even this low level of expression. Other promoters with tighter control, such as the T7 promoter referred to earlier, have to be used in these situations. Alternatively, obtaining high levels of production of a protein that is highly damaging to *E. coli* may require the use of an alternative host.

7.2.4 Translational fusions

If we want to provide our gene with translational signals (ribosome binding site and start codon) as well as a promoter that are recognized by the host, we can use a similar approach. For this purpose we need a different type of expression vector: one that will give rise to a *translational fusion.* In this case, part of the translation product (the protein or polypeptide) is derived from the insert and part from the vector (Figure 7.7). When using a translational fusion vector, we have to be much more careful in the design of our construct.

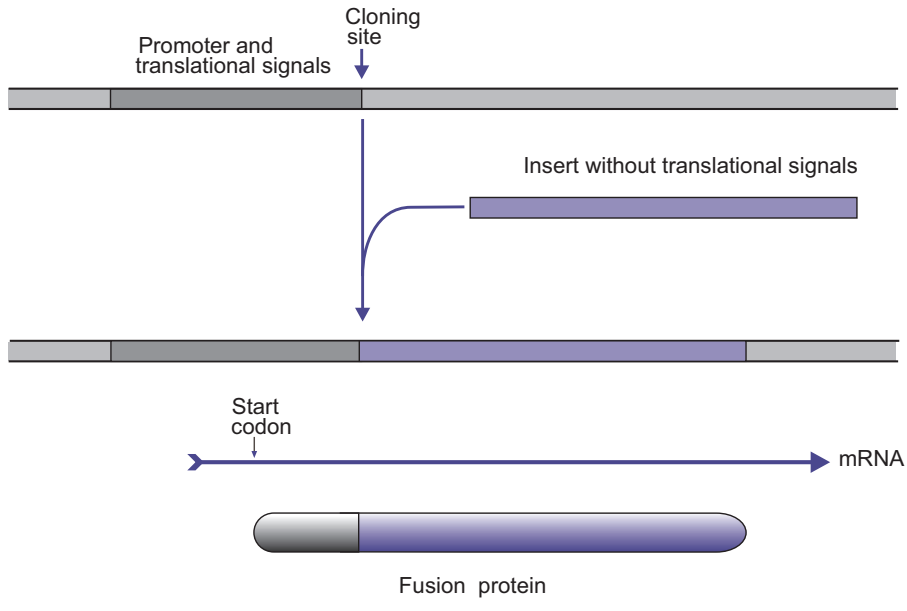


Figure 7.7 Expression vectors: translational fusions.

With a transcriptional fusion vector, as described above, we simply have to make sure the insert is the right way round. Within reason, it does not matter too much where it is – some untranslated leader mRNA can be tolerated, so it does not have to be precisely located with respect to the promoter. However, with a translational fusion, the location is all important. Because translation starts at the initiation codon in the vector sequence and the ribosomes then read the sequence in triplets, you have to make sure that the correct reading frame is maintained at the junction. One or two bases out in either direction, and your insert will be read in the wrong frame, giving rise to a completely different amino acid sequence (and probably resulting in premature termination as the ribosomes will soon come across a stop codon in this frame) – see Figure 7.8.

You can easily achieve this if you know the sequence of your insert, and the sequence of the vector, or at least of that region of the vector between the cloning site and the start codon. (Contrast this with a transcriptional fusion vector, where it is quite possible to insert an uncharacterized DNA fragment into the vector and be successful in obtaining expression.)

Furthermore, you have to work out exactly what will happen at the cloning site when you cut the insert and the vector and then join them together. If you are relying on a restriction site that is naturally present in the insert, there is only a 1 in 3 chance that it will be in the right frame when joined to the vector, even assuming the insert goes in the right way round. You might need

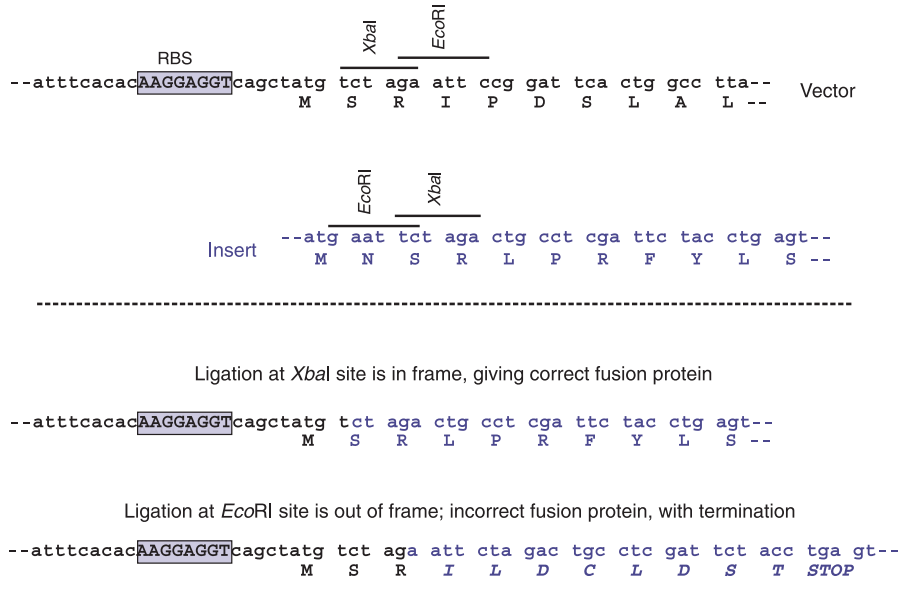


Figure 7.8 Translational fusion: in-frame and out-of-frame fusions.

to alter your choice of vector, or to modify either the vector or the insert, or both (e.g., by using linkers or adaptors – see Chapter 2) to achieve the desired reading frame. One approach is the design of sets of translational fusion plasmids that have the AUG start codon located for each of the three different reading frames relative to the multiple cloning site; this means that for any cloning strategy one of the plasmids will produce your insert into the correct reading frame to produce a functional protein. However, this risk of incorrect reading frame can be dramatically reduced if PCR is used to generate the insert fragment, if you design the primer to contain a restriction site that will produce the correct reading frame (see Chapter 4).

Finally, having done all of this, it is essential to check, by sequencing the recombinant product, that ligation has indeed resulted in the construct that you have designed, and that no mistakes have been introduced (especially if you have used PCR with *Taq* DNA polymerase). In particular, it is important to ascertain that the reading frame is actually correct. Loss of a single base in the ligated product will result in an incorrect reading frame.

In theory, it is possible to use a vector that just has the ribosome binding site and no start codon, and to use the natural start codon of the cloned gene. However, in *E. coli* the optimum distance between the Shine–Dalgarno sequence and the start codon is only 5–7 bases, so there is not a lot of room for manoeuvre. More commonly, the vector will have the start codon as well, and the insert will not have a start codon. In this case, the translational fusion product will have a few amino acids derived from the vector sequence, and

will lack some of the N-terminal amino acids that are normally found in this polypeptide. For many purposes, this is of little significance, as proteins can often tolerate a considerable amount of variation in the N-terminal sequence. In some cases it may be important to have a product that is precisely the same as the naturally occurring one – for example, if you are making a protein for human therapeutic purposes. You will then need to carry out more precise manipulations rather than using off-the-shelf expression vectors; for example, you can ensure that the vector-derived amino acids replace exactly those that are lost from the insert.

Sometimes, it is useful to be able to add a short sequence of amino acids to the N-terminus of your product. We will consider the applications of such *tagged proteins* later in this chapter.

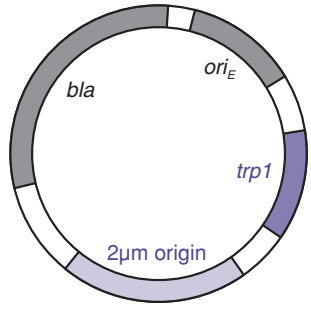
7.3 Yeast systems

Bacteria are convenient hosts for many purposes, with *E. coli* usually the host of choice for initial gene cloning. However, such bacterial systems have many limitations for the *expression* of cloned genes, especially for large-scale production of proteins of eukaryotic origin. In particular, the post-translational modifications needed for conversion of the primary product of translation into the mature, folded, biologically active protein are more likely to occur correctly in a eukaryotic host than in a bacterium. There are a wide variety of systems available to address this requirement; in this chapter we will consider just a few examples, starting with yeast hosts.

7.3.1 Cloning vectors for yeasts

Microbiologically, 'yeasts' are single-celled fungi, as opposed to filamentous fungi, but the term is quite imprecise. Not all 'yeasts' are related taxonomically, and indeed some filamentous fungi can also grow in a unicellular form that is referred to as a yeast form. Although in common usage the term 'yeast' would be taken to mean the baker's yeast *Saccharomyces cerevisiae*, other yeasts are also used as hosts (especially members of the genus *Pichia*). But for the moment we will limit ourselves to *S. cerevisiae*.

The vectors that will be most familiar, after reading about bacterial cloning vectors in Chapter 2, are the yeast episomal plasmids (YEp). These are based (Figure 7.9) on a naturally occurring yeast plasmid known as the 2μm plasmid, and they are able to replicate independently in yeast, at high copy number (25–100 copies per cell). As is usually the case with vectors for eukaryotic cells, these plasmids also have an *E. coli* origin of replication, enabling them to be grown and manipulated in an *E. coli* host (i.e., they are *shuttle vectors*). There is one point of detail in which they differ from bacterial cloning vectors, and that is the nature of the selectable marker. For bacterial vectors, we can



bla = beta-lactamase (ampicillin resistance, selectable marker in *E. coli*)
ori$_E$ = origin of replication in *E. coli*
trp1 = selectable marker in *S. cerevisiae* auxotrophs
2μm origin = origin of replication in *S. cerevisiae*

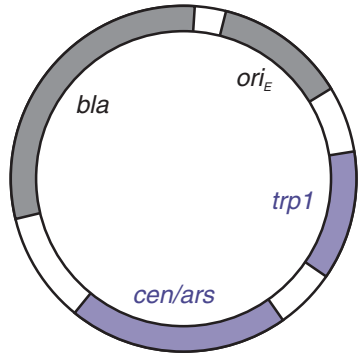
Figure 7.9 Structure of a yeast episomal vector.

exploit the large number of antibacterial antibiotics (and the correspondingly large number of antibiotic resistance genes) to enable us to select our transformants. There are fewer antibiotics available to which yeasts are sensitive (although some fungicides can be used), and therefore selection more commonly makes use of complementation of auxotrophic mutations in the host strain as selectable markers. For example, a host strain of *S. cerevisiae* with a mutation in the *trp1* gene will be unable to grow on a medium lacking tryptophan. If the vector plasmid carries a functional *trp1* gene, then transformants can be selected on a tryptophan-deficient medium. Other markers that are commonly used for selection in a similar manner include *ura3* (uracil), *leu2* (leucine) and *his3* (histidine). These vectors would also be designed as shuttle vectors, carrying an antibiotic resistance marker for selection in *E. coli*.

Vectors that replicate as plasmids in *S. cerevisiae* are often rather unstable, in that they tend to be lost from the culture as plasmid-free daughter cells accumulate. This is due to erratic partitioning of the plasmids during mitosis. Newer versions of YEp vectors, taking advantage of better understanding of the biology of the 2μm plasmid, are more stable.

Autonomously replicating plasmids can also be constructed by inserting a specific sequence from a yeast chromosome; this sequence is known, unsurprisingly, as an *autonomously replicating sequence*, or *ars*. The early versions of these plasmids were very unstable, but newer constructs that also include a centromere are more stable (Figure 7.10). In contrast to the YEp vectors, these yeast centromere plasmids (YCp) are normally maintained at low copy number (1–2 copies per cell), which can be advantageous if your product is in any way harmful to the cell, or if you want to study its regulation.

The vectors described so far are maintained in yeast as circular DNA molecules, much like the bacterial plasmids we considered previously. Two



bla = beta-lactamase (ampicillin resistance, selectable marker in *E. coli*)
ori$_E$ = origin of replication in *E. coli*
trp1 = selectable marker in *S. cerevisiae* auxotrophs
cen/ars = centromere and autonomously replicating sequence, providing an origin of replication in *S. cerevisiae*

Figure 7.10 Structure of a yeast centromere vector.

other classes of vectors deserve a mention. Firstly, there are the yeast integrating plasmids (YIp). These do not replicate independently but integrate into the chromosome by recombination (Figure 7.11). The frequency of transformation is very low, and it is difficult to recover the recombinant vector after transformation. The main advantage is that the transformants are much more stable than those obtained with the autonomously replicating plasmids.

Finally, there are the *yeast artificial chromosomes* (YACs), which carry telomeres that enable their maintenance in *S. cerevisiae* as linear structures resembling a chromosome. The use of these vectors, for cloning very large pieces of DNA, is quite distinct from the uses of the vectors described above. The use of YACs was considered in Chapter 2.

7.3.2 Yeast expression systems

Yeasts have many advantages for expression of cloned genes. They grow rapidly, in simple defined media, and as unicellular organisms they are relatively easy to manipulate and enumerate.

S. cerevisiae has been used as an experimental organism in microbial genetics for many years and there is a wealth of biochemical and genetic information available to support its use for gene cloning – including the complete genome sequence. The principles involved in the design of expression vectors for use in yeast are similar to those described above for bacterial expression vectors, except that of course the expression signals involved are those applicable to *S. cerevisiae* rather than *E. coli*. If required, this can include signals for secretion, or for targeting to the nucleus or other cellular compartments.

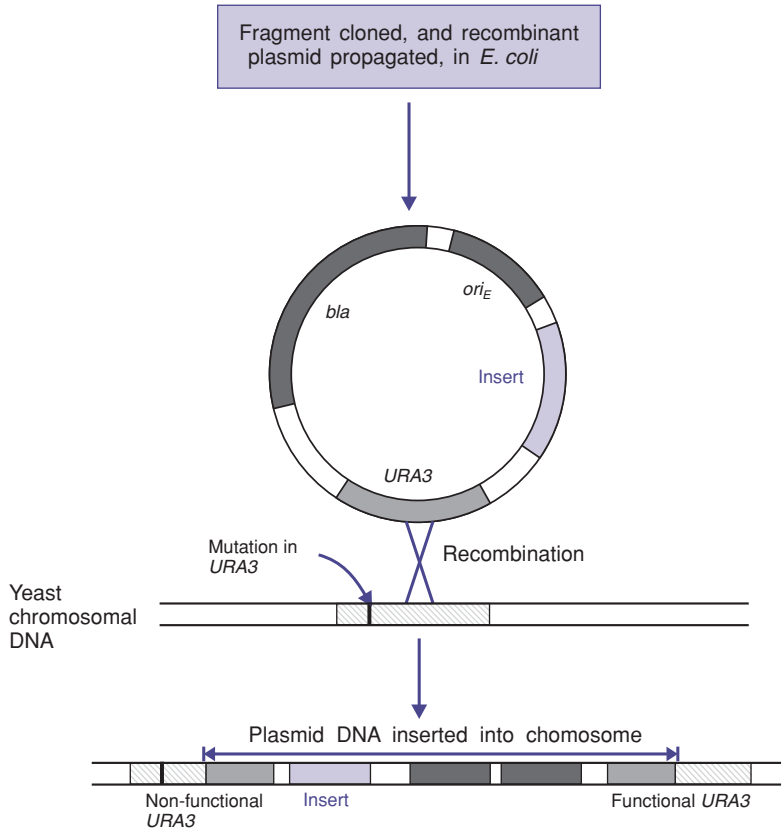


Figure 7.11 Structure and use of a yeast integrative plasmid (YIp).

Note that *S. cerevisiae*, although eukaryotic, has very few introns, and is not the host of choice if you want to ensure correct excision of introns. This means that it is still preferable to use cDNA, rather than genomic, sequences.

In many respects, especially if using episomally replicating vectors, the concepts are similar to those involved in the design of plasmid-based expression vectors in bacteria. There is a choice between vectors carrying the origin of replication from the naturally occurring 2μm yeast plasmid, which are maintained episomally at high copy number (up to 40 copies per cell), and centromere vectors, which are maintained at low copy number (1–2 copies per cell). Both types of vectors can be designed as shuttle vectors, i.e., they also carry an *E. coli* replication origin, which enables the initial construction, verification and amplification of the recombinant plasmid to be carried out in *E. coli* before transferring the finished construct into yeast cells. As with bacterial expression vectors, these *S. cerevisiae* vectors are designed with a controllable promoter adjacent to the cloning site to enable expression of the

cloned gene to be switched on or off. Most commonly, this involves the promoter and enhancer sequences from the *GAL1* (galactokinase) gene, which is strongly induced by the addition of galactose.

Although *S. cerevisiae* is the most commonly used host, powerful and versatile systems are also available for several other yeast species, notably *Pichia pastoris*. *P. pastoris* is able to use methanol as a carbon source, with the first step in this pathway being catalysed by the enzyme alcohol oxidase (the product of the *AOX1* gene). This gene is tightly controlled, so that in the absence of methanol no alcohol oxidase is detectable. On addition of methanol to the culture, the *AOX1* gene is expressed at a very high level. The use of the *AOX1* promoter in the expression vectors, adjacent to the cloning site, therefore provides vectors that are capable of generating substantial levels (up to several grams per litre) of the required product.

The vectors in *P. pastoris* are not maintained episomally, but are designed to be integrated into the yeast chromosome. They do, however, contain an *E. coli* replication origin, so that they can be maintained as plasmids in *E. coli*, enabling formation and verification of the recombinant plasmids in *E. coli* before transformation of *Pichia*.

Not only are the expression levels in *Pichia* usually higher than those obtainable in other systems (prokaryotic or eukaryotic), but scaling up to industrial levels of production is relatively straightforward, and the organism grows readily in simple defined media, leading to lower costs than those involved in insect or mammalian systems (as described below).

7.4 Expression in insect cells: baculovirus systems

Whereas yeast cells offer some advantages over bacterial hosts for the expression of animal proteins, they are not perfect. The fact that yeast are single-celled eukaryote organisms that diverged from the animal kingdom about a billion years ago means that differences have evolved, which may result in suboptimal expression of your gene of interest. A compromise is to use cells from an animal, and in this case insect cells are often used.

Baculoviruses, such as the *Autographa californica* multiple nuclear polyhedrosis virus (AcMNPV), infect insect cells and are exploited as the basis of systems for gene expression in such cells. During normal infection, the cells produce large amounts of a virus-encoded protein called polyhedrin, which forms a matrix within which the virus particles are embedded. The high level of production of polyhedrin is due to a very strong promoter, which can be used to drive expression of the cloned gene, by inserting it downstream from the polyhedrin promoter. Production of polyhedrin itself is not completely essential for virus production, but in its absence the virus does not produce

inclusions; the resultant plaques can be distinguished from those produced by wild-type virus. However, the viral DNA is itself too large (>100 kb) for direct manipulation to be easily carried out. The gene to be expressed is therefore first inserted into the polyhedrin gene in a smaller *transfer vector*. Once again, the transfer vector is a shuttle vector, i.e., it contains an *E. coli* origin of replication and other functions that allow manipulation in an *E. coli* host. For production of recombinant virus particles, the transfer vector must recombine with viral DNA. For this to happen, insect cells are cotransfected with the recombinant transfer vector and with viral DNA. Recombination within the insect cells leads to the production of recombinant viruses, although at a low frequency. These are separated from the wild-type, non-recombinant viruses by picking the plaques that are characteristic of polyhedrin-deficient viruses.

More efficient versions of this system are now available. In the example shown in Figure 7.12, the transfer vector contains a multiple cloning site adjacent to (and downstream from) the polyhedrin promoter so that the cloned gene will be transcribed from that promoter. Flanking this region, on either side, are two DNA regions derived from the virus that allow homologous recombination with viral DNA. For the other component of the system, instead of intact viral DNA, a linearized DNA is used. The linearized viral DNA lacks a portion of an essential gene, thus eliminating the production of non-recombinant viruses. Recombination with the transfer vector, after cotransfection of insect cells, restores this essential gene. Additional features can be built in to this system, including the incorporation of a marker such as beta-galactosidase so that blue recombinant plaques can be readily identified.

Following transfection, viral plaques can be picked and the virus characterized to verify the presence of the cloned gene. The characterized recombinant virus can then be used to infect a large-scale culture of insect cells; high levels of expressed protein are usually obtained before the cells lyse. The levels of production are generally lower than those obtainable with *Pichia* expression systems, but insect cells are more closely related than yeasts to other animals, and this has the advantage that post-translational modification of protein is claimed to more closely resemble that in, for example, mammalian cells. Nevertheless, the yeast cells are much easier and cheaper to grow, and thus you must decide which is the appropriate system to use on a case-by-case basis.

7.5 Mammalian cells

Most research on animal genes pertains to those of humans and other mammals. Production of mammalian proteins in artificial hosts, such as the bacterial, yeast or insect systems described above, will always have the potential confounding issue of incorrect processing of gene expression. This may lead to the production of proteins that have only a partial function or none at all.

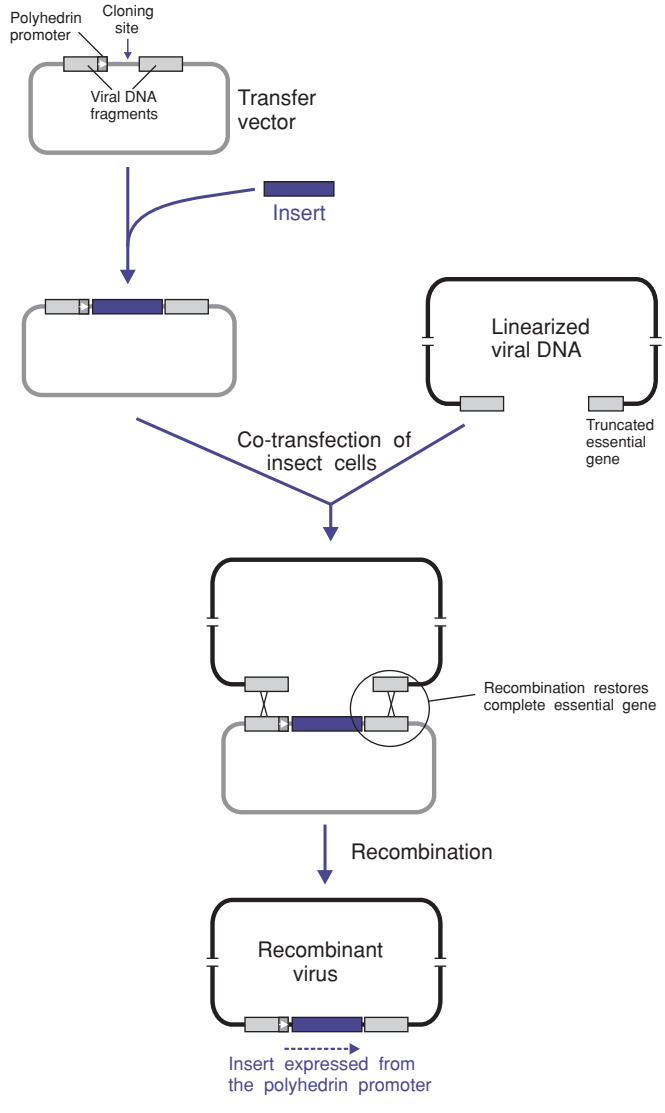


Figure 7.12 Cloning using a baculovirus vector.

Can we not then just use mammalian cells to produce mammalian proteins? The answer is yes, but with a number of important caveats, which will be described in the following sections.

7.5.1 Cloning vectors for mammalian cells

In bacteria, cloning vectors replicate separately from the chromosome, as plasmids or bacteriophages. As we have seen above, the same is true of many

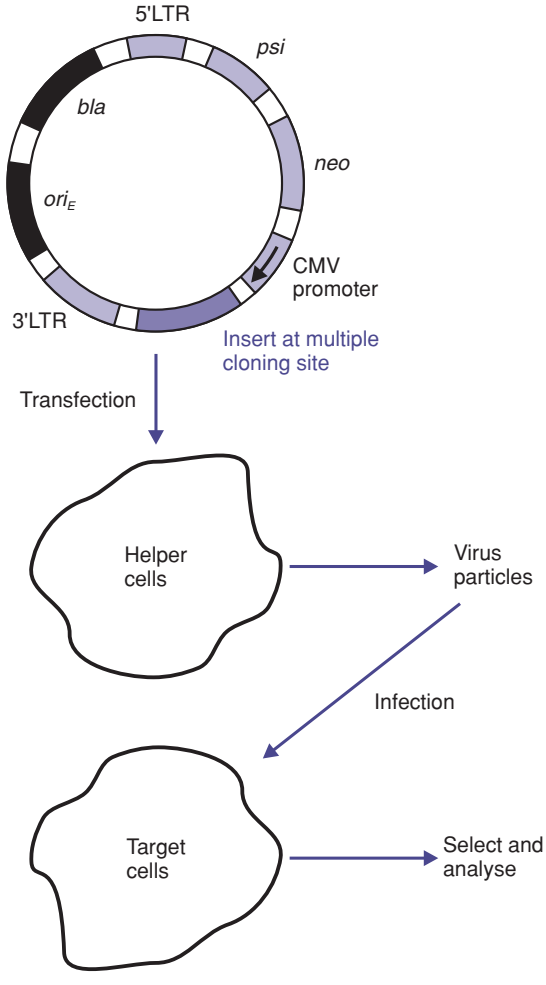
types of vectors used in yeast. The situation with cloning in mammalian cells is somewhat different, in that independent, plasmid-like replication is often not sustained. Some vectors are capable of plasmid-like replication, especially those carrying the origin of replication from the virus SV40 (simian virus 40), which replicate episomally (outside the nucleus) in some mammalian cells (such as COS cells). However, more stable clones can be obtained by inserting the DNA into a chromosome, which happens readily in mammalian cells. In either case, the cloning vector enables you to organise your cloned gene in relation to a set of expression signals, as considered further below.

There are other types of vector available, based on various viruses, which can be used to transmit your cloned gene from one cell to another. Of these, the *retroviral* vectors deserve a special mention, and in order to understand these we need a brief account of retroviral biology. Retroviruses have an RNA genome. When a cell is infected, the RNA is copied into double-stranded DNA by the action of a viral protein, *reverse transcriptase*. This protein is present in the virion and enters the cell along with the RNA. (Reverse transcriptase is formally described as an RNA-directed DNA polymerase, and we encountered it in Chapter 3, where we considered its use in the production of cDNA from mRNA templates.) This DNA then circularizes and is integrated into the host cell DNA by the action of another virion protein known as *integrase*. The efficiency of integration of the DNA into the genome is one of the main attractions of this system for genetic manipulation of animal cells.

The integrated DNA is bounded by sequences known as *long terminal repeats* (LTRs), which include a strong promoter for transcription of the integrated viral genes *gag*, *pol* and *env*. Full-length transcripts from these provide the viral RNA, which can then be assembled into virus particles. One region of the virus, known as the *psi* site, is essential for this process. The packaged virus particles acquire envelope glycoproteins from the host cell membrane as they bud off from the cell, without causing lysis of the host cells. These glycoproteins determine the type of receptors present on the surface of other mammalian cells that will recognize the virus, leading to further rounds of infections.

Development of vectors based on retroviruses rests on the knowledge that most of these functions can be provided *in trans*, for example, by genes from a defective helper virus already integrated into the genome of the host cell. The main features that are *cis*-acting, and therefore need to be located on the vector itself, are the LTR sequences and the *psi* site.

The basic features of the use of such a vector are outlined in Figure 7.13. The vector is a shuttle plasmid, so *E. coli* is used for construction of the recombinant plasmid by inserting the required gene at the multiple cloning site. This construct is then used to transfect a culture of a special cell line (*helper cells*) that contains the *gag*, *pol* and *env* genes required for virus



bla = beta-lactamase (ampicillin resistance), for selection in *E. coli*
ori$_E$ = *E.coli* origin of replication
neo = neomycin phosphotransferase, for selection in infected cells
CMV promoter is for transcription of cloned gene
LTR, *psi* = essential *cis*-acting retroviral sequences (see text)

Figure 7.13 Structure and use of a retroviral vector.

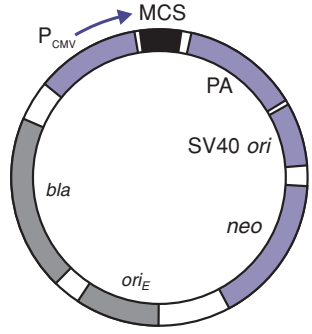
production, integrated into the genome. The transfected cells will therefore be able to produce virus particles containing an RNA copy of your construct. These particles are able to infect other cells that do not contain these essential genes. Since the viral particles carry preformed reverse transcriptase and integrase, the RNA will be copied into DNA in such cells, and the DNA will be efficiently integrated into the genome. However, since these cells do not carry the essential genes, no further production of viral

particles will occur. Nonetheless, your gene is now stably integrated into the chromosome and can be expressed from the adjacent promoter derived from the vector.

The specificity of the viral particles for other cells will be determined by the envelope gene carried by the helper cells. Replacing that gene by other genes for envelope glycoproteins from other viruses, in particular the VSV-G gene from vesicular stomatitis virus, enables a wider range of target cells to be used, not just mammalian cells but extending to, for example, chickens, oysters, zebrafish, mosquitoes – in fact cells from virtually any animal species can be infected. The incorporation of foreign genes into the genome of whole animals (*transgenesis*) is considered further in Chapter 11.

7.5.2 Expression in mammalian cells

There is an enormous variety of expression vectors for mammalian cells, and a full treatment of this field is beyond the scope of this book. Although the details are more complex than the systems described so far, the general principles remain familiar, and rest on incorporation of the foreign gene into a vector that provides expression signals, many of which are derived from viruses such as SV40 or cytomegalovirus (CMV). The general features of such a vector are exemplified in Figure 7.14. A gene inserted at the multiple cloning site (MCS) enables high-level constitutive expression from the CMV promoter, while the presence of a polyadenylation signal increases mRNA stability. The SV40 origin allows episomal replication in COS cells, and the neomycin



P_{CMV} = CMV promoter. High level constitutive expression in mammalian cells
MCS = multiple cloning site
PA = polyadenylation signal
SV40 *ori* = origin of replication; episomal replication in COS cells
neo = neomycin phosphotransferase; resistance to G-418, for selection in mammalian cells
ori_E = *E. coli* replication origin
bla = beta-lactamase (ampicillin resistance), for selection in *E. coli*

Figure 7.14 Structure of a basic episomal vector for gene expression in mammalian cells.

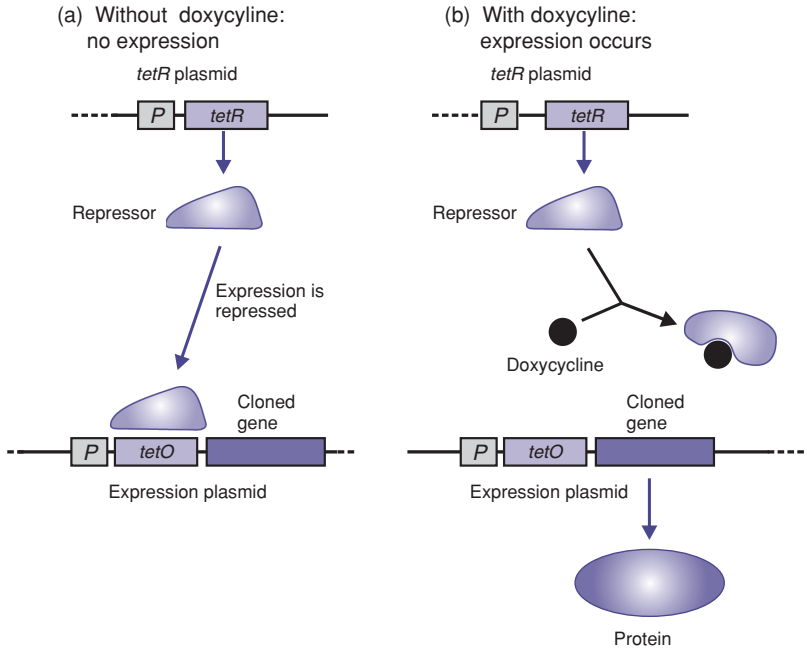


Figure 7.15 Control of gene expression using tetracycline.

phosphotransferase gene permits selection for resistance to the antibiotic G-418 (Geneticin®). Note that this is a shuttle vector, carrying an *E. coli* origin of replication and an ampicillin resistance gene (beta-lactamase), so the construction can be carried out in *E. coli* before transferring the recombinant plasmid to a mammalian cell line.

Many enhancer/promoter systems give high-level, constitutive expression. As in prokaryotic systems, it is desirable to control the onset of expression. One way of achieving this is through the use of the *tet*-on system (Figure 7.15). In this system, there is an operator sequence (*tetO*), from the bacterial tetracycline resistance operon, between the CMV promoter and the cloned gene. If the mammalian cells are cotransfected with a second plasmid containing the tetracycline repressor gene (*tetR*), also expressed using the CMV promoter, the TetR protein will bind to the *tetO* site, thus preventing transcription. When tetracycline, or more commonly its less toxic derivative doxycycline, is added to the culture medium it will bind to the TetR protein, altering its conformation and releasing it from the DNA, thus derepressing transcription of the cloned gene. The advantage of this is that, because this system is of prokaryotic origin, the ligand is not produced naturally in the host, and the activation it produces does not affect the induction of native mammalian genes. Additional sequences can be added to the cDNA to be expressed, thus enabling targeting of the product to specific cellular locations

such as the nucleus, mitochondria, endoplasmic reticulum or cytoplasm, or for secretion into the culture medium.

In contrast to bacterial cells, the introduction of DNA into mammalian cells does not depend on the independent replication of the vector, and this has some important consequences. If the introduced DNA is stably integrated into the nuclear DNA, then it will be copied as the cell divides, providing a *stable* integrant. However, if the introduced vector is left in an unintegrated, episomal, state then it will have a finite lifespan. In such transient systems, as the cells divide, the vector will be diluted out until no protein production is observed. While the latter may seem somewhat inefficient, it has the advantage of being technologically much easier (and faster) than making a stable integrant and is, hence, a popular choice. However, it should be noted that some expression vectors can combine the best of both situations, being stably maintained at high copy extrachromosomally. These include those containing the origin of replication from the Epstein–Barr virus (EBV); with a suitable promoter system these are capable of allowing high levels of protein expression.

The advantage of using mammalian cells for expression of eukaryotic genes, especially those from mammalian sources, lies in the greater likelihood of a functional product being obtained. In addition, there is a much greater chance that any accessory proteins necessary for protein function will also be present within the system. This is especially relevant for studies of structure–function relationships and the physiological effect of the protein on cell function. However, the relative difficulty, and cost, of scaling up production, compared to either *Pichia* or baculovirus systems, makes mammalian cells less attractive if the objective is the large-scale production of recombinant proteins for other uses.

7.6 Adding tags and signals

7.6.1 Tagged proteins

The fact that a translational fusion vector (see above) adds a short stretch of amino acids to the N-terminal end of your polypeptide product can be turned to advantage. Purification of the recombinant product by conventional means can often be tedious and inefficient, and would need to be devised and optimised again for each new product. However, if the vector contains not just a start codon but also a short sequence coding for a few amino acids (known as a *tag*), the resulting fusion protein will carry this tag at the N-terminus. For example, the tag may constitute a recognition site (an *epitope*) for a monoclonal antibody (see Box 3.3). You can then recover the protein from the cell extract in a single step by affinity purification, using the ability of the monoclonal antibody to bind to the epitope tag (Figure 7.16). A further application

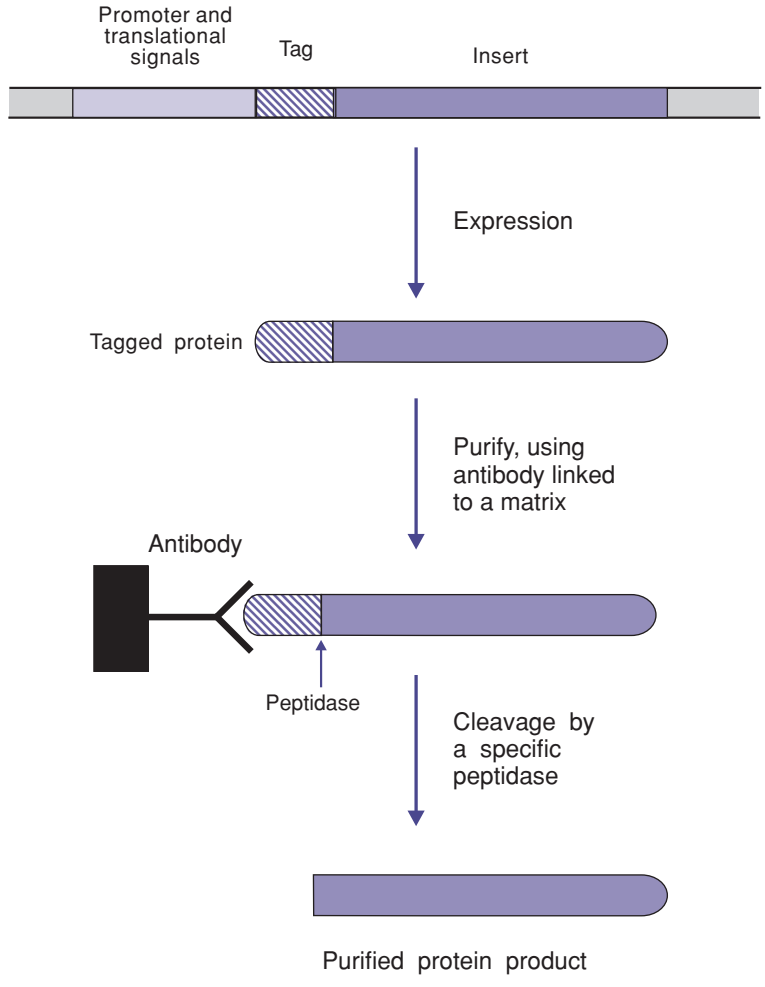


Figure 7.16 Tagged proteins.

of this type of vector is that you can use this specific monoclonal antibody to detect the expression of your product, which can be particularly useful if you are studying a protein for which no robust antibody currently exists. There are a variety of such tags (and corresponding monoclonal antibodies) available, and they can either be included as part of the vector backbone, or be added to either end of your insert (N- or C-termini), using modified PCR primers.

An alternative to epitope tagging is provided by designing the expression vector (or manipulating the insert) such that a sequence coding for a number of histidine residues is added to one end of the sequence to be expressed. This will produce a His-tagged protein, which can be purified using a Ni^{2+} resin.

The affinity of the histidine residues for nickel will result in the tagged protein being retained by the resin, while other proteins will be washed through. A further advantage of the His-tag is that it can also be used as an epitope tag, since there are antibodies available that will specifically recognize this sequence of histidine residues, which facilitates detection of the tagged product.

Of course you may not want to have a tag permanently attached to your product. For this reason, vectors can be designed to contain not only the tag but also a site at which the product can be cleaved by highly specific peptidases. If the peptidase is sufficiently specific it will cleave the product only at this site and not elsewhere. Therefore, you can affinity purify your product using the tag, and then use the peptidase to remove the tag, allowing purification of the untagged protein (see Figure 7.16).

7.6.2 Secretion signals

In a similar way to the incorporation of tags into the product, we can add secretion signals. *E. coli* does not secrete many proteins into the culture supernatant – most of the proteins secreted across the cytoplasmic membrane remain in the periplasm, trapped by the outer membrane. However, in some other bacteria, particularly Gram-positive bacteria such as *Bacillus subtilis*, many enzymes are secreted into the culture supernatant. This can be advantageous for several reasons. Firstly, if a protein is synthesized at high levels and accumulated within the cytoplasm, it will result in a very high concentration of that protein. This can cause aggregation of the protein into insoluble *inclusion bodies*. These may be damaging to the cell, and are often very difficult to resolubilize. Secretion of the protein into the culture medium will prevent this, since the volume of the supernatant is very much greater than that of the total volume of the cytoplasm of all the cells. Even with a very dense culture, the space occupied by the cells is only a small proportion of the culture volume. Furthermore, although *Bacillus* secretes a number of enzymes, the culture supernatant will contain a much simpler mixture of proteins than the cell cytoplasm. The task of purifying the required product is therefore much easier.

Secretion of proteins by bacteria depends normally on the presence of a *signal peptide* at the N-terminus. This labels it as a secreted protein, so it is recognized by the secretion machinery and transported across the cytoplasmic membrane. Therefore, if we incorporate a sequence coding for a signal peptide into our vector (or into the insert) the final product may be secreted. It is not inevitable, as it also depends on the overall structure of the protein. If the protein is naturally secreted in its original host, then we may be successful. If it is normally a cytoplasmic protein, the chances of getting it secreted successfully are very much lower.

As described above, in mammalian systems, signals can be added to direct the product to specific cellular locations, or to obtain a secreted product.

7.7 *In vitro* mutagenesis

In the early part of this chapter (Section 7.1.3) we described how codon usage may affect optimisation of gene expression in *E. coli* – and in other hosts may prevent expression altogether. What can we do about it? One possibility is to use site-directed mutagenesis (see below) to alter the sequence at these specific positions, so as to change a rare codon into a more appropriate one – still maintaining the same coding properties so we are not altering the nature of the final product. (Alternatively, as indicated in Section 7.1.3, we can use *E. coli* host strains that have been engineered to express the rare tRNAs.)

There are many other situations in which we would want to introduce specific changes into the structure of a gene, or the protein that it encodes. For example, if we are dissecting the molecular function of a protein, we can make specific modifications, such as mimicking constitutive phosphorylation by exchanging a specific amino acid residue. If we want to make more extensive changes, then we could use synthetic techniques to remake the complete gene – either by direct DNA synthesis or by PCR-based methods such as assembly PCR. All these techniques are described in the subsequent sections.

7.7.1 Site-directed mutagenesis

This refers to the specific alteration of either a single base or of a short sequence of bases in a cloned gene. If there is a unique restriction site within your cloned DNA fragment, at or near the point where you want to make a change, then there are several simple possibilities, including the use of exonucleases to make a short deletion, and/or the insertion of a short synthetic oligonucleotide (Figure 7.17). Although these methods are limited in their scope, they can be useful, for example, in the removal or addition of restriction sites within a DNA fragment. In addition, the deletion procedure will remove various lengths of DNA, which enables you to identify the limits of regions that have specific functions, such as the ability to bind regulatory proteins (see Chapter 10). Those deletions that extend into the regulatory region will produce clones showing altered regulation of the gene concerned.

The fundamental limitation to such an approach is the requirement for a unique restriction site at the appropriate place. More generally applicable techniques involve the use of oligonucleotide primers to introduce changes in the DNA sequence. There are a number of variations in the technique, so we will just describe a small selection of procedures to illustrate the principle of the method.

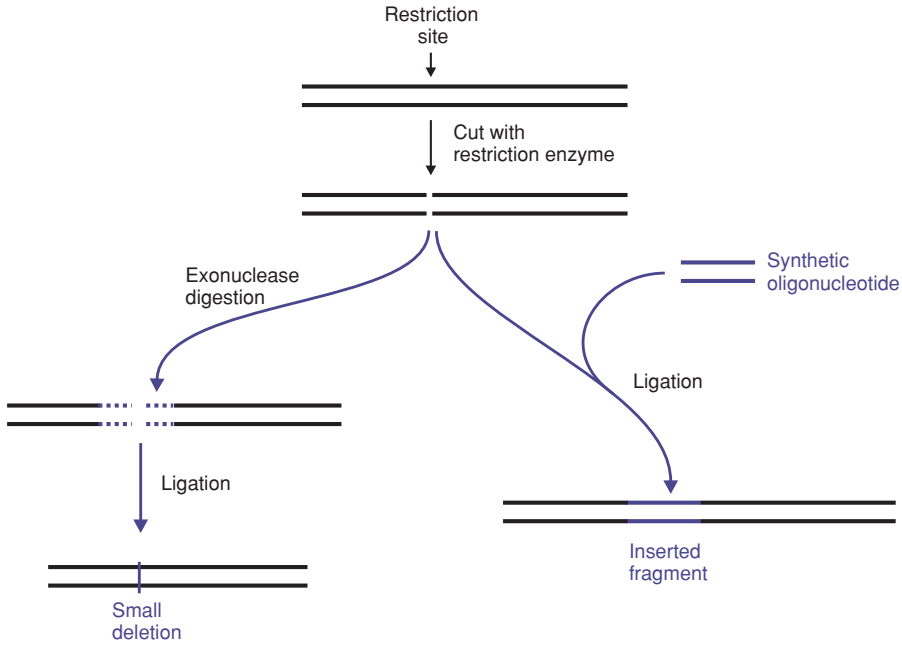


Figure 7.17 *In vitro* mutagenesis: examples of some simple procedures.

Fundamentally, these procedures rely on the ability of single-stranded DNA molecules to anneal to one another even if their sequences do not match perfectly. Therefore, if we design an oligonucleotide that contains the sequence we require, which is only slightly different from the natural sequence, it will anneal to the complementary strand at the corresponding position. This provides a primer that can be used by DNA polymerase, in a PCR reaction (see Chapter 4), as illustrated in Figure 7.18. Using a mismatch primer (primer B) directed at the position where we want a mutation, and a second, perfectly matched, primer (primer A) at the end of the gene, the PCR will produce a DNA fragment containing the changed sequence at one end. But this is only part of the gene, and of course we want the whole thing. So we do a second PCR amplification, using a complementary mismatch primer C (which will anneal to the other strand and amplify in the opposite direction), together with another perfectly matched primer D at the other end of the gene. This PCR will produce a fragment corresponding to the right-hand portion of the gene, but also carrying the mutation at one end. To produce the complete, altered gene, we can mix the products of the two reactions, denature and re-anneal. Some of the single strands from the first amplification will anneal to strands from the second amplification in the region where they overlap, corresponding to the sequences of primers B and C. A subsequent PCR, using the outermost primers (A and D) will amplify the

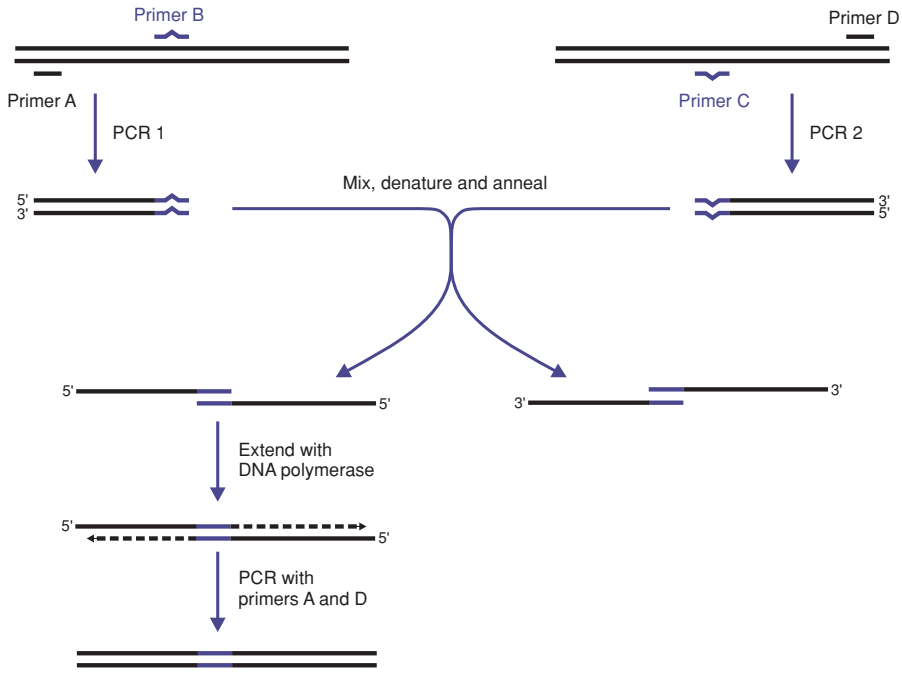


Figure 7.18 PCR-based mutagenesis.

full-length product as shown. This can then be cloned and sequenced to verify the presence of the mutation. You may spot that re-annealing of the products of the initial PCR reactions will produce other structures, not shown in the figure, but these will not be amplified by the subsequent PCR, so we don't need to worry about them.

The need to clone the PCR product can be avoided by using a PCR to amplify the entire plasmid (Figure 7.19). The primers are designed so that the final PCR product can be ligated to reconstitute an intact plasmid for transformation of a host cell. Of course, the PCR product will be contaminated with some of the original plasmid template, and this has to be removed. One way of doing this is to produce the initial plasmid in a specific strain of *E. coli* that will methylate the plasmid DNA. This methylated DNA is sensitive to cleavage by a specific enzyme, the *Dpn*I endonuclease. The PCR product will not be methylated, and so will resist digestion.

A further alternative is shown in Figure 7.20. This involves the use of two primers to introduce changes into the plasmid. In addition to the mutagenic primer, which introduces an alteration in the target gene, there is a second primer, the selection primer, which removes a unique restriction site. Both primers face in the same direction (anticlockwise in the figure), so extension from the two primers, using DNA polymerase, will produce a

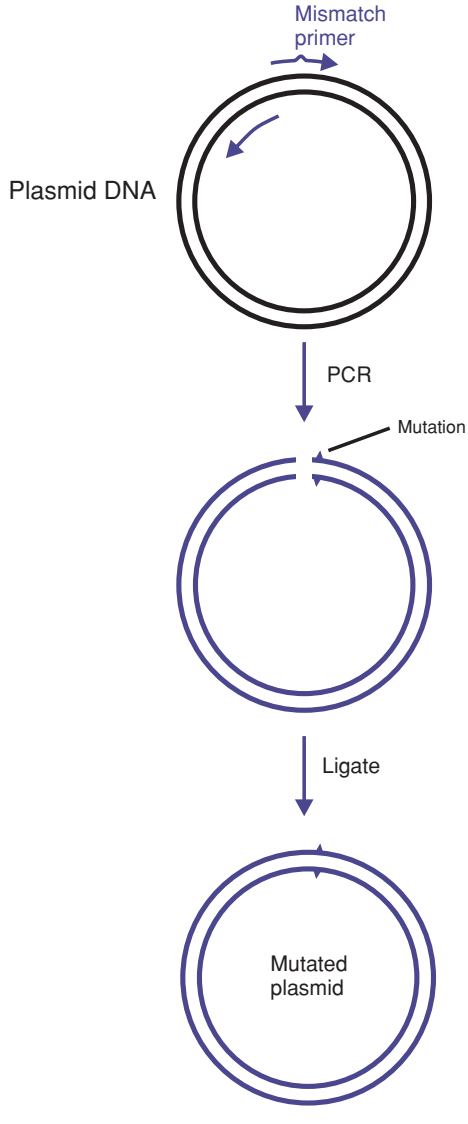


Figure 7.19 *In vitro* mutagenesis: plasmid PCR.

complete complementary strand containing both mutations. The resulting plasmid will be a hybrid molecule, containing one parental strand and one mutated strand, so after transformation, and replication of the plasmid, some of the transformants will have the parental plasmid and some will have the altered one. This is where the change introduced by the selection primer is useful. Digestion with the appropriate restriction enzyme will linearize the parental plasmid, while the mutated plasmid will not be affected. If you then

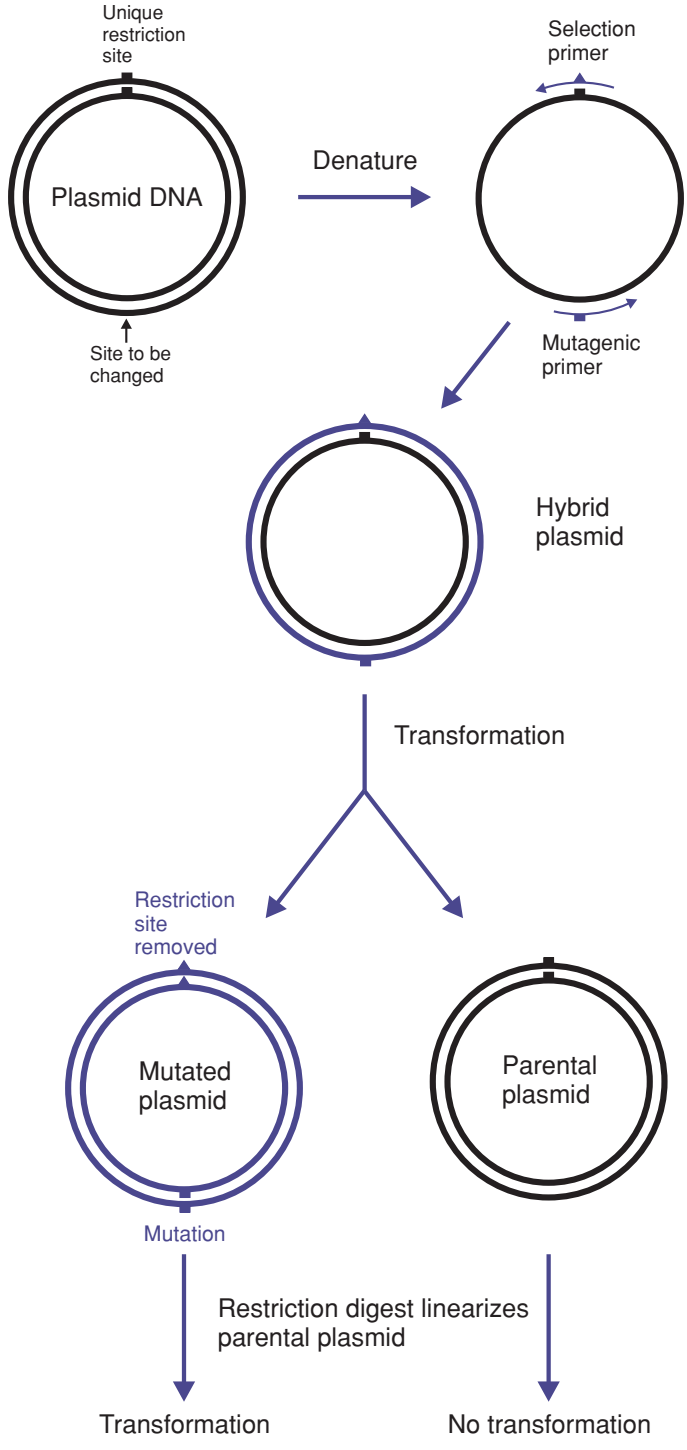


Figure 7.20 *In vitro* mutagenesis: selection for mutated plasmid using restriction site removal.

use the digested mixture of plasmid DNA for bacterial transformation, only the intact mutated plasmid will yield transformants.

The early applications of PCR for site-directed mutagenesis suffered from the risk of introducing unwanted mutations into the product, due to the absence of proofreading ability of *Taq* polymerase. The introduction of alternative heat-stable polymerases that do have proofreading ability, and hence show greater fidelity, has greatly enhanced the application of PCR for this purpose. Nevertheless, it is important to sequence the product to ensure firstly that the required mutation has indeed been achieved, and secondly that no other alterations have been introduced into the sequence.

7.7.2 Synthetic genes

Although it is possible to use site-directed mutagenesis to introduce a number of changes into the sequence of a gene, it becomes a rather laborious procedure if we want to make extensive changes – such as altering the codon usage throughout. One possibility then is to abandon the natural gene and simply synthesize a fresh one from scratch. Taking the amino acid sequence of the protein, we can use a computer to *back-translate* it into a nucleic acid sequence with optimised codon usage (and/or optimised base composition). It is actually rather more complicated than that, as you would want to check that you are not introducing other undesirable features such as secondary structures that might interfere with transcription; you also would want to consider the introduction (or omission) of restriction endonuclease sites at strategic points.

In principle, you would program a DNA synthesizer to make the DNA sequence, which could then be cloned. Here again, it is not quite that simple. DNA synthesizers work by adding bases sequentially, one at a time, to the growing oligonucleotide through a series of chemical reactions. With each base added, there is a very small but finite possibility of the failure of the reaction, the main consequence of which is a gradual reduction in the yield of the product. Synthesis of oligonucleotides of 100–150 bases is reliable enough; although longer sequences can be made, the risk of introducing unwanted errors becomes a problem. One way round this is to make the gene in shorter fragments and ligate them together, bearing in mind that the synthesizer produces single-stranded DNA, so you need to make two complementary strands for each fragment. This would clearly involve a lot of work, but if you really need to do it, it is quite possible.

7.7.3 Assembly PCR

An alternative procedure, which is a lot simpler in practice than that described above, is simply to mix together a complete set of synthetic

fragments (each of which overlaps, and is complementary to, its neighbours on either side). As before, these fragments are designed to optimise codon usage and base composition, and to introduce any other changes you want to make. This complex mixture is then subjected to a number of rounds of PCR, which results in, amongst other products, some full-length DNA. This is further amplified, specifically, by addition of primers directed at the ends of the full-length gene, and further rounds of PCR. Although the synthesis of a complete gene of say 1000 base pairs involves a highly complex mixture of starting oligonucleotides, which might be expected to produce an impossible number of different products, the procedure really does work.

7.7.4 Synthetic genomes

Although the above approaches were described in terms of synthesizing individual genes, there is no reason to stop there. In 2010, a group of researchers working under Craig Venter succeeded in synthesizing the complete genome of a species of *Mycoplasma*, and inserted it into a related bacterium from which the genomic DNA had been removed. Although this was widely publicized as the 'creation of artificial life', in reality it was merely transforming one bacterium into another one, and so is formally an extension, albeit a radical one, of the existing techniques for genetic modification. Nevertheless, it does represent a major step forward, and illustrates the potential of these methods for the development of novel forms of cells.

7.7.5 Protein engineering

The above discussion started off by considering site-directed mutagenesis and synthetic DNA as a means of introducing a limited number of changes into a naturally occurring gene. The product is still the same as the naturally occurring one. But there is no reason why we should limit these techniques to the production of naturally occurring proteins. We can just as easily synthesize a gene that codes for an altered protein. We could introduce specific changes into the sequence and test their effects on, for example, the substrate specificity of the enzyme. Or we could introduce cysteine residues into the sequence at strategic points, so that the protein produced would contain additional disulphide bridges. This would be expected to increase the thermal stability of the enzyme, which could be advantageous. (Unfortunately such a change often results in the loss of enzyme activity as well.)

This concept, often referred to as *protein engineering*, would in principle culminate in the production of totally novel enzymes, tailor-made to carry out specific enzyme reactions. The limiting factor now is not the techniques for producing such engineered proteins, but the inadequacy of our knowledge

of how a specific sequence of amino acids will fold into a three-dimensional structure, and how we can predict the enzymic activity of such a novel protein. However, such approaches have been used to engineer 'optimised proteins' for expression in bacterial hosts, which can then be used for therapeutics, and two examples will be given to demonstrate the potential. Firstly, is the production of human betaferon, where a single cysteine residue was mutated to a serine. As noted above, bacterial hosts do not form disulphide bridges efficiently and this can lead to the incorrect folding of the produced protein. In this case, mutation of a single cysteine residue prevented the formation of an unwanted disulphide bond, ensuring correct folding of the produced protein. Secondly, we can consider Humalog, an 'enhanced' version of insulin, where two engineered amino acid changes were used to effectively swap the final lysine and proline residues. This change resulted in an engineered insulin that was less likely to form dimers when given, meaning that it provided a faster response than traditional insulin.

7.8 Vaccines

One of the best examples of the use of genetic manipulation for product formation is the development of novel vaccines, as an alternative to the conventional use of killed or live (attenuated) pathogens, or toxoids (toxins treated to render them harmless). We will consider two approaches here: the production of an individual antigen from the pathogen (subunit vaccine), and the use of DNA itself as a vaccine. The production of live, genetically modified vaccines will be dealt with in Chapter 11.

7.8.1 Subunit vaccines

The concept behind the use of genetic manipulation to produce a recombinant subunit vaccine is simple. If you can identify the key (protein) antigen that is needed for a protective effect, you can clone the gene responsible using an expression vector, and produce large quantities of the protein. This has many advantages over producing the same component from the pathogen itself: it is cheaper and more effective, since you may get much higher levels of expression; it is easier to produce a safe vaccine, as there is no problem of contamination with any other damaging components of the vaccine; and the production process is safer, since you do not have to grow large-scale cultures of a dangerous pathogen. The best example, illustrating the power of this approach (as well as some of the problems), is the development of a vaccine against the hepatitis B virus. Unfortunately, this virus cannot be grown in cell culture, which makes the development of a conventional vaccine next to impossible. However, we know that one virus protein, the surface antigen (HBsAg, see Figure 7.21) is able to confer protection, so the expression of this

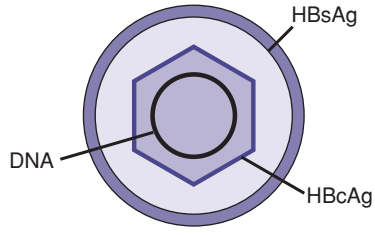


Figure 7.21 Hepatitis B virus: diagrammatic structure.

gene in a suitable host should yield a product that can be used as a vaccine. Unfortunately, when this gene was expressed in *E. coli*, it was found that the product did not give effective protection. The problem here is that the product from *E. coli*, although having the authentic amino acid sequence, does not adopt the correct conformation. However, it was found that the expression of this gene in the yeast *S. cerevisiae* did give rise to HBsAg with a natural conformation, and this product was immunogenic and protective – and this is the basis of the currently used hepatitis B vaccine.

We can also use genetic manipulation as an alternative to chemical treatment of toxins as a way of producing toxoids for immunization. For example, the cholera toxin consists of two types of subunit: the A component is responsible for the toxic effects after it gets into the target cells, while the B component is need for attachment to cell surface receptors. The B component by itself is non-toxic. Expression of the B component alone will give a product that can cause an immune response without the toxicity of the intact toxin.

7.8.2 DNA vaccines

The most radical of the new approaches stems from the surprising observation that injection of DNA, usually as a recombinant plasmid containing the relevant genes, may stimulate an immune response to the product of those genes, and in some cases this will lead to protective immunity. This presumably occurs because the plasmid is taken up by some of the cells in the body, resulting in expression of the genes carried by the plasmid. One advantage of this approach is the ease of administering the vaccine. The plasmid DNA can be coated onto tiny inert particles, which are projected at high velocity into the skin, penetrating a very short distance below the surface of the skin (see also the discussion of *biolistics* in Chapter 11). The particles are small enough, and the velocity is high enough, for some of the particles to penetrate

into cells within the skin. The plasmid is constructed so that the genes to be expressed are located downstream from a strong constitutive promoter (such as the CMV promoter, see earlier in this chapter). Although the plasmid is not replicated within these cells, there is sufficient transient expression of the cloned genes to generate an immune response. DNA vaccines represent an exciting new development in vaccine technology. However, we have to recognize that there are many questions to be answered before it is accepted for widespread use in humans.

At the same time, this technique has achieved extensive use as a means of raising antibodies to recombinant proteins in experimental animals. Instead of the time-consuming procedure of purifying the protein and immunizing animals in the conventional way (see Box 3.3), a recombinant plasmid, carrying the cloned gene together with appropriate expression signals, can be used directly as an immunogen.

8
Genomic Analysis

Up to this point, we have been looking at individual genes (or small groups of genes) – how to clone and characterize them, and how to study their specific expression. In keeping with the title of the book, our theme now moves to methods that are applicable to the study of the whole genome. This includes not only the more recent developments in genome sequencing and the genome-wide study of gene expression, but also longer-established techniques that are invaluable for connecting the data obtained in this way to the properties of the organism as a whole.

8.1 Overview of genome sequencing

In 1975, the first complete DNA genome was sequenced – that of the small bacteriophage ⊘X174 (5 kb). Other complete sequences followed gradually with the sequencing of ever larger viral genomes. A new dimension was opened up in 1995 by the first sequence of the genomes of independently living organisms, those of the bacteria *Haemophilus influenzae* and *Mycoplasma genitalium*. Although these are small genomes (1.83 Mb and 0.58 Mb, respectively), they are orders of magnitude larger than viral genomes, and this, therefore, represented a major scientific advance. Only six years later, in February 2001, two competing entities, the publicly funded Human Genome Project and the private company Celera, published the human genome sequence of no less than three billion base pairs. The human genome sequencing projects provided a major driving force to the development of the technology and capacity required to undertake large-scale genome sequencing projects. This has been put to good use: Sequencing bacterial genomes is now a routine operation – to the extent that for many important organisms the complete genome sequence of many strains is available.

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

Genome sequences are also available for many higher organisms – mammals, plants, insects and many others. These genome sequences provide an invaluable tool for fundamental studies of these organisms, including their phylogenetic relationships, as well as for applied studies such as the nature of inherited disease and drug development. From a scientific point of view, these projects were obviously revolutionary. New genome sequences are released every day – up-to-date information can be obtained on the Genomes On-Line Database (www.genomesonline.org). With regard to the study of human genetics, perhaps one of the most exciting is the 1000 Genomes Project (www.1000genomes.org), which, as the title suggests, has the aim of sequencing the entire genomes of 1000 people, and providing the first deep catalogue of human genetic variation.

In addition to the finished genomes, there are a very large number of ongoing sequencing projects. Some of the genome projects have a policy of releasing all sequence data as they become available. This is raw sequence with no annotation (i.e., nothing to tell you what it codes for, or any other information), but it can be searched for the occurrence of specific sequences.

8.1.1 Strategies

Various strategies were developed for determining genome sequences in the days when the Sanger method was the only one available. One is to extend the concept of shotgun sequencing as described in Chapter 5; i.e., simply to make a library of small random fragments of the whole genome and sequence them (Figure 8.1). This requires the isolation and sequencing of a vast number of

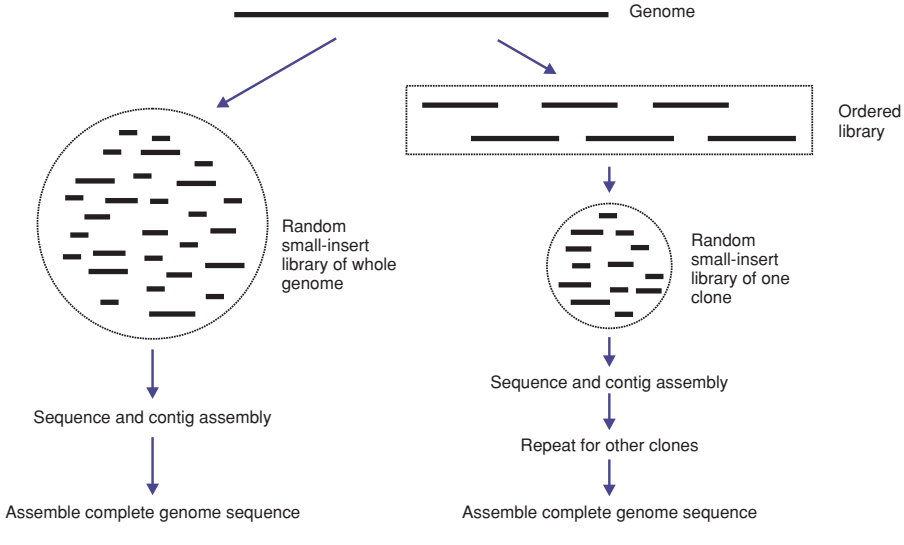


Figure 8.1 Genome sequencing strategies.

clones. This became the backbone of genome sequencing strategies, and remains in use to some extent even with the availability of *de novo* sequencing using next-generation methodologies (as described below). The main alternative was to split the problem up into defined units first. You can make a library representing the whole genome or chromosome, using a vector that can carry large inserts. You would then determine the sequence of the insert in individual clones, and subsequently assemble the complete genomic sequence. Some sequencing projects started with random clones, while others attempted to arrange the clones into an *ordered* library first (i.e., one in which the relative position of all of the clones has been defined – see Section 8.7.2).

One advantage of this 'clone by clone' strategy was that very many laboratories are capable of determining a sequence of the size of individual inserts, so it was possible to distribute them to different laboratories for sequencing. A further advantage of the 'clone by clone' approach was that useful information is generated at intermediate stages in the project – the sequence of an individual clone is a coherent piece of information – while the shotgun approach consists of large numbers of essentially meaningless sequences of small fragments until sufficiently large contigs have been assembled. However, the advent of the newer sequencing technologies, in which no cloning step is required, have largely obviated the need for such considerations.

8.2 Next generation sequencing (NGS)

Until recently, the advances in genome sequencing have been achieved using essentially the same underlying techniques (Sanger, or dideoxy, sequencing) as described in Chapter 5. What distinguished this from the sequencing performed in an ordinary research laboratory was solely the scale of the process. This includes development of a high degree of automation of sample preparation, setting up the sequencing reactions, and loading the products into the sequencer, coupled with extensive computer power for processing and analysing the results. This made possible the sequencing of the human genome, although this project took 10 years, and cost $10 billion. Following this achievement, the concept of high-throughput sequencing has been developed, which, combined with new methodologies, has reduced both the cost and the time of sequencing by several orders of magnitude.

Currently, there are three leading NGS platforms, manufactured and marketed in competition by three different companies. One advantage that all these processes have in common is that they read directly the sequences of individual fragments of DNA, without the cloning step that is required for Sanger sequencing. This, coupled with the ability to read very large numbers of sequences simultaneously, speeds up the process massively. Box 8.1 lists some sources of additional information.

Box 8.1 Further information on next-generation sequencing

Company sources

454 sequencing http://www.454.com/products-solutions/how-it-works/index.asp

SOLiD™ System http://www.appliedbiosystems.com/

Solexa/Illumina http://www.illumina.com/technology/sequencing_technology.ilmn

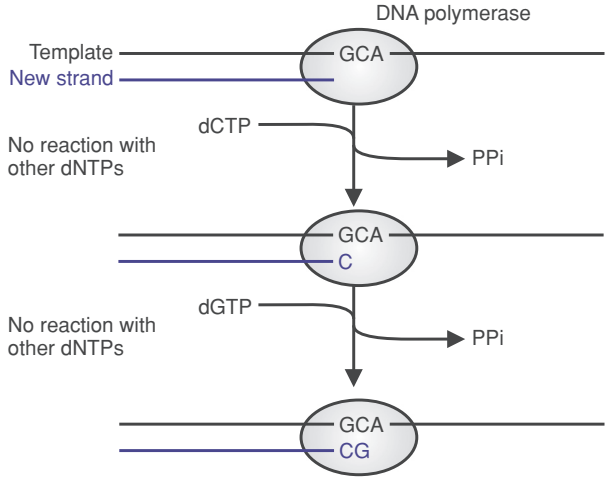
Other

Metzker, M.L. (2010) Sequencing technologies – the next generation. *Nat Rev Genet* 11, 31–46.

8.2.1 Pyrosequencing (454)

One alternative to Sanger sequencing that is used for genome sequencing is known as *pyrosequencing*. The principle of this, as illustrated in Figure 8.2, relies on the release of pyrophosphate when a nucleotide is incorporated into a growing DNA strand. The pyrophosphate released can be detected by a coupled enzymatic reaction in which ATP sulphurylase converts adenosine 5′-phosphosulphate (APS) to ATP, which enables a second enzyme, luciferase, to react with its substrate (luciferin), generating visible light, which is measured by a detector. Pyrophosphate will only be generated if the DNA synthesis reaction is supplied with the correct dNTP. So, in the example shown, there will be a reaction in the first step if dCTP is supplied, but not if dGTP, dATP or dTTP is added. At the second step, light is only emitted if dGTP is added – so the sequence of the new strand, in the 5′ to 3′ direction, is CG.

As described, this technique may be used to genotype single samples on a small scale. Its most important application, however, is the sequencing of a huge number of DNA fragments simultaneously (*parallel sequencing*), as shown in Figure 8.3. In this method, manufactured by 454 Life Sciences, a library is created by shearing DNA into small fragments (300–800 bases in length), after which an adaptor is ligated to each of the fragments. Each fragment is then attached to a small bead, and the beads are captured in an emulsion of PCR mix and oil. In this emulsion, each DNA bead (with its attached DNA fragment) is effectively contained in a microreactor, consisting of a



Pyrophosphate release at each step is detected by a coupled reaction involving ATP sulphurylase, with adenosine 5' phosphosulphate (APS) as substrate, and luciferase

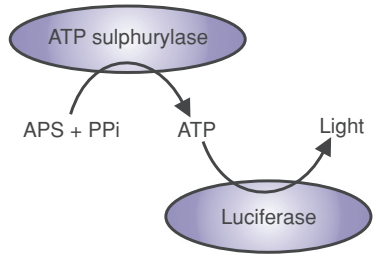


Figure 8.2 Pyrosequencing.

single droplet of water containing all the reagents necessary for a PCR reaction. This enables each DNA fragment to be amplified, within the droplet, without having to purify individual fragments, resulting in a collection of beads each of which now carries millions of copies of the original fragment.

This mixture is applied to a sequencing reaction slide that contains tiny reaction wells, less than 100 picolitres in volume, so that over 1 million wells are contained on a 60×60 mm slide. The size of the DNA beads ensures that each well will contain only one DNA bead, and therefore only one specific DNA fragment. The wells are also loaded with (smaller) enzyme beads carrying the pyrosequencing enzymes. The base of the slide is in optical contact with a fibre-optic bundle linked to a detector that enables the measurement of light emitted from each individual well. The reaction slide is then washed in turn with each dNTP, so that the detector senses which well reacts with each

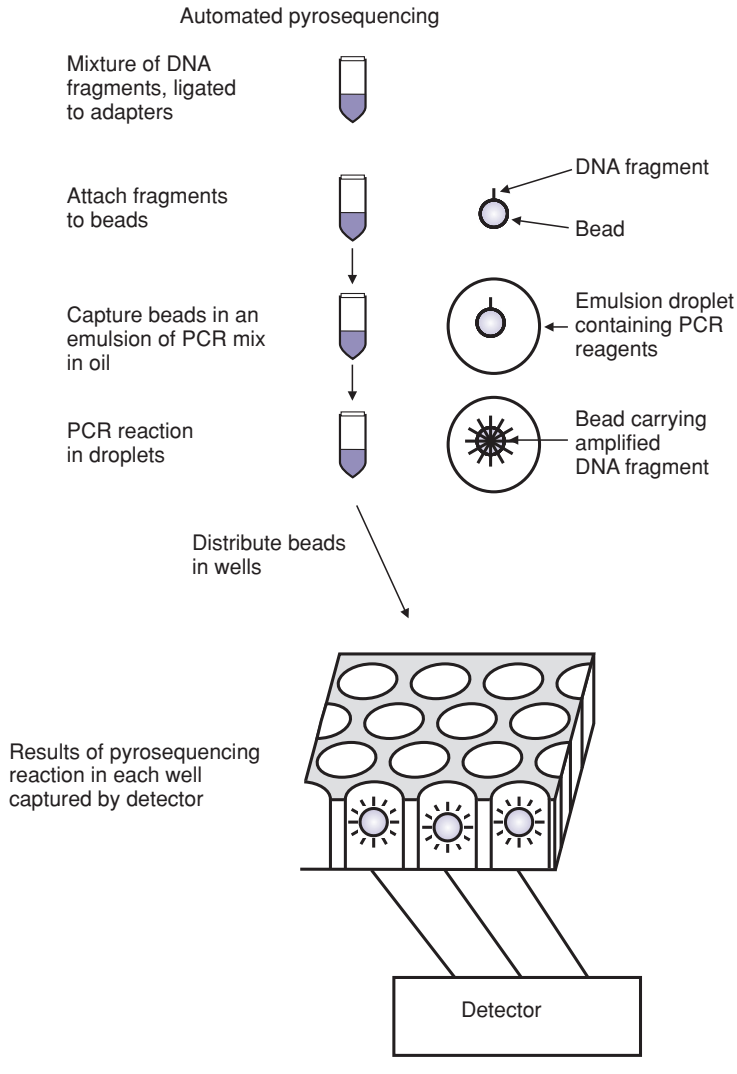


Figure 8.3 Automated pyrosequencing.

dNTP, and hence determines the sequence of the growing DNA strand. If a homopolymer is present (where one nucleotide is repeated), then the emitted light will be stronger, but this is only proportional to the length of the repeat for short homopolymers.

As with all next-generation sequencing, the real power lies in the ability to sequence very large numbers of fragments simultaneously – so that the sequence of 0.5 gigabases (500 million bases) can be obtained in one 4-hour run, or a few days from start to finish for the entire procedure. One strength of the 454 method is that a sequence of 400 bases can be read from each fragment,

which is not far short of that achievable with the Sanger method. This is sufficient for *de novo* sequencing of previously unmapped genomes, as opposed to *resequencing* in which one searches for variations between the genomes of individual members of a species for which a curated genome sequence is already available.

8.2.2 SOLiD sequencing (Applied Biosystems)

This sequencing method is based on ligation. Like the 454 method, the first step is to create a library of sheared DNA fragments. After ligation of adaptors at both ends, each fragment is immobilized onto a bead. As in the 454 method, PCR amplification is produced in an emulsion, where each bead is contained within a tiny single water droplet emulsified in oil. Following PCR, these beads are linked covalently to a glass slide. Universal sequencing primers are then annealed to the ends of the DNA fragments. The major difference with the SOLiD method is that ligation rather than polymerization is used to read the sequence. The principle of the method is illustrated in Figure 8.4. The slide is bathed in a mixture of short oligonucleotides and DNA ligase. Each oligonucleotide ends with a different pair of bases, so only that probe which ends with the correct dinucleotide will be able to be ligated to the primer. This means that there is a 1 : 16 chance that a given probe will be complementary to the two nucleotides following immediately after the universal primer in any given template. In Figure 8.4, the first two bases in the template, beyond the sequence to which the primer is annealed, are CA, so only the probe ending in GT will be ligated. The other probes will not be ligated, and can be washed off. Since each probe is labelled with a fluorescent dye that emits light of a specific wavelength, the machine will determine which of the probes has ligated, in each of the millions of wells.

The probe needs more than two bases to pair properly, so the characteristic two-base sequence is followed by three degenerate bases (i.e., positions where any one of the four bases is present). There is then a chemically labile bond linking to the last three bases and to the label, so that after the reaction is read, the label (and three bases) are removed, and a second ligation performed as before. This time, the new probe will ligate to the remains of the previous one, if correctly matched at the next two positions, thus giving the sequence of another two bases. This process is repeated for a number of cycles, so you would get a sequence of dinucleotides each separated by three unknown bases. By itself, this would not be much use. To fill in the gaps, the ligated DNA product is stripped off the template, and the process is repeated with a primer that is one base shorter, giving another set of two-base sequences, moved along the template by one base (Figure 8.4). Doing this repeatedly, each time moving the starting point one base to the left, means that, after five rounds of ligation and resetting the primer, you get back to the

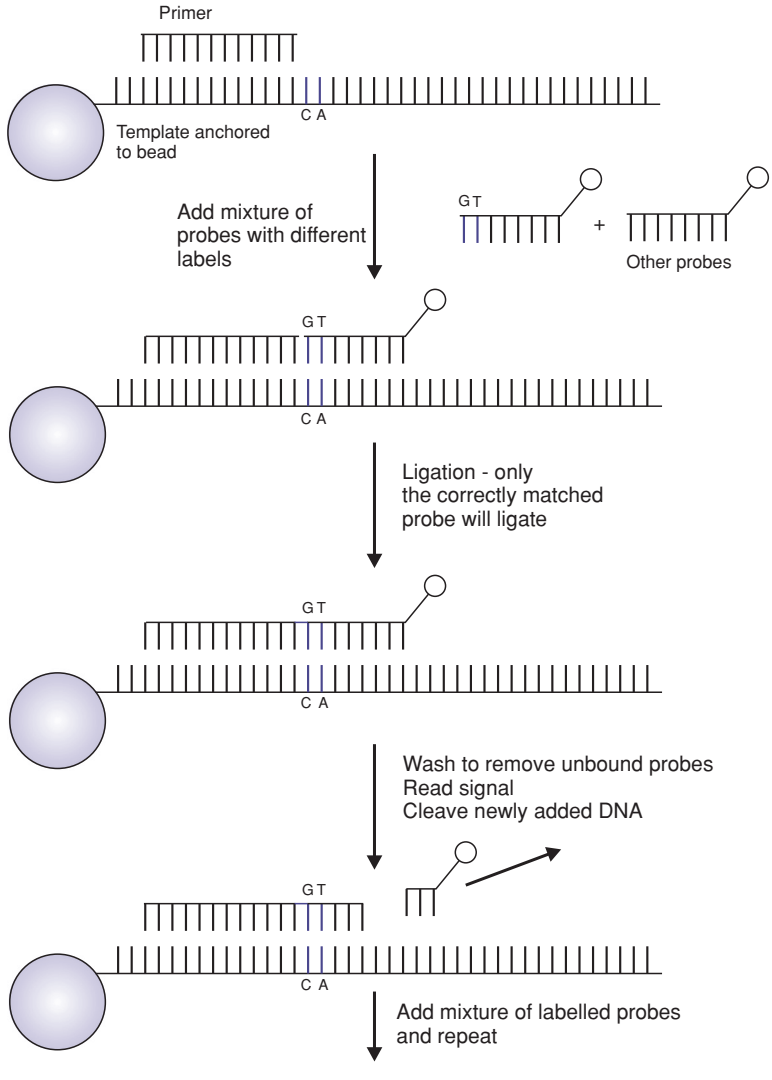


Figure 8.4 Principle of SOLiD sequencing. (*Continued*)

original set of two-base sequences. All of this information can then be put together by the computer to give a complete sequence.

Since this is all done automatically by the machine, it is very much easier to undertake experimentally than to describe (see Box 8.1 for more information). There is one further complication though. There are only four dyes available, and there are sixteen possible dinucleotides. However, careful design of the probes means that the computer can use the information from the overlapping sequences to determine which dinucleotide is present at a given position.

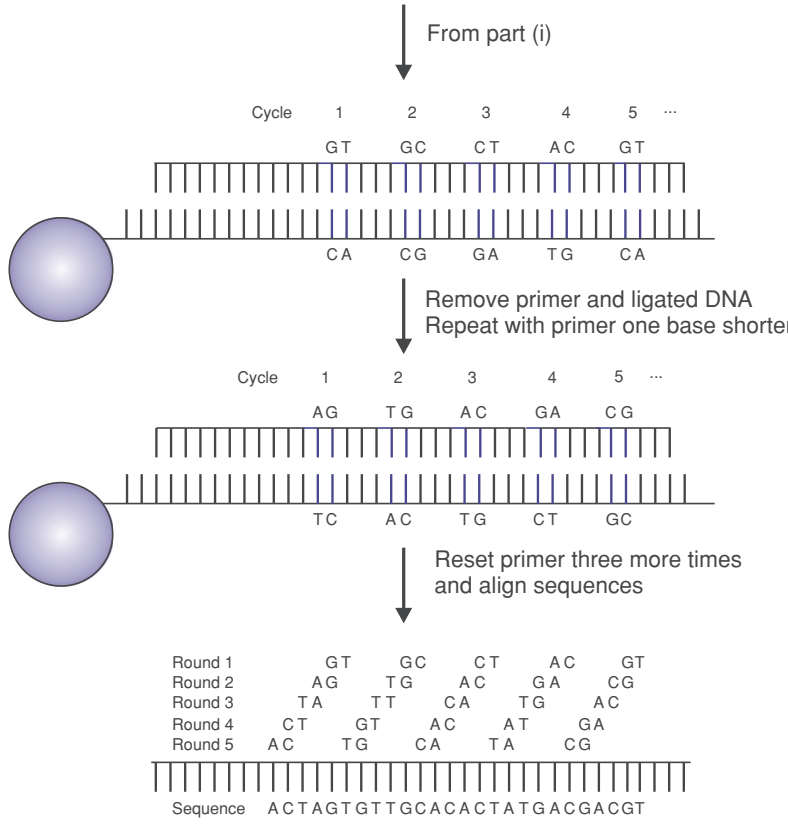


Figure 8.4 (*Continued*)

The high degree of accuracy (99.94%) is a major advantage of this method. Each sequencing round can obtain 60 Gb of data, which is computationally intensive. However, a disadvantage of SOLiD compared to the 454 method is that each read length is shorter (approximately 50 bases in length), meaning that significantly more computation is required to build longer sequence reads accurately. This makes this method more useful for *resequencing* than for *de novo sequencing*. Because 99% or so of the bases will be identical between any two individuals, the short reads obtained by this method can be arranged using the known full-length sequence as a reference (a process referred to as *scaffolding*), and any individual polymorphisms and mutations present in the individual identified.

8.2.3 Bridge amplification sequencing (Solexa/Ilumina)

In this technique (Figure 8.5), DNA is first randomly fragmented, and adaptors added to the ends of the fragments, in a manner similar to that in the

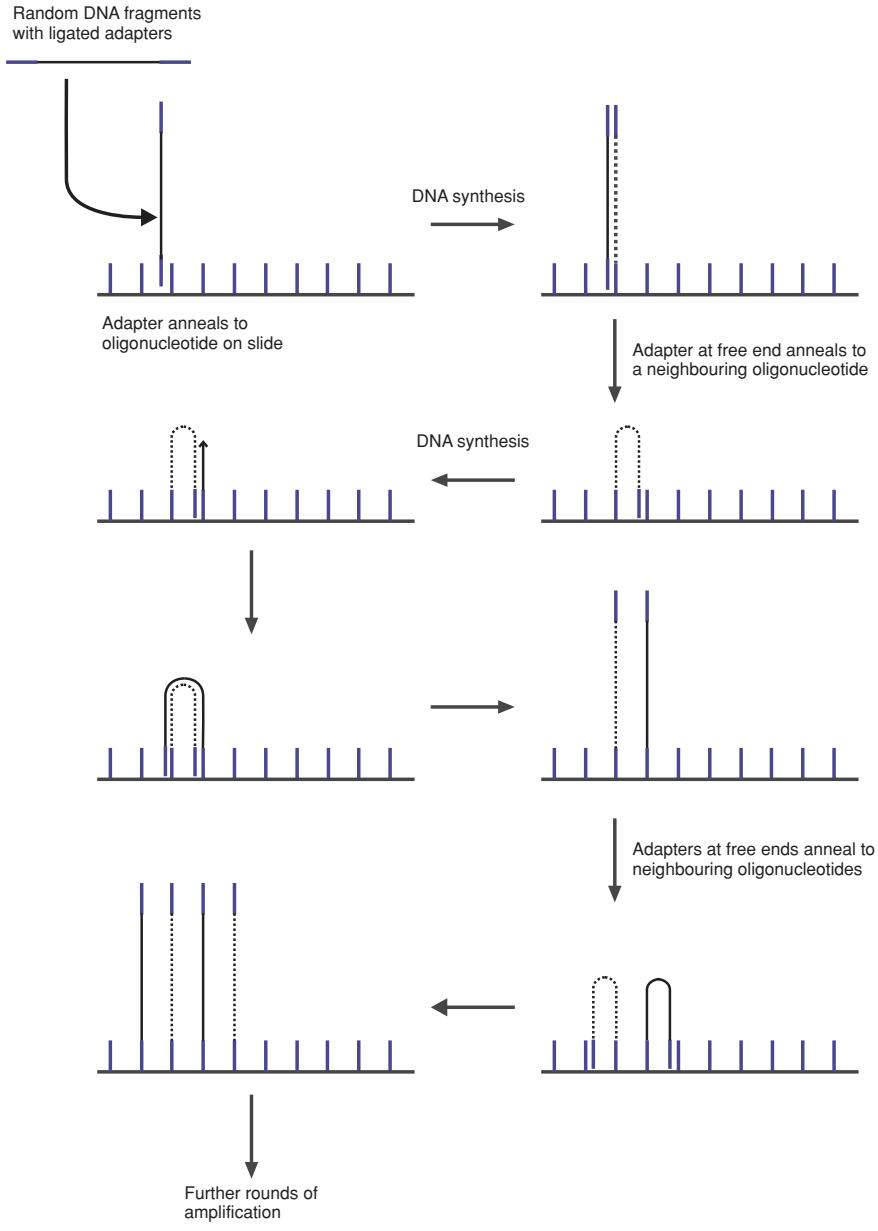


Figure 8.5 Principle of Solexa/Illumina sequencing.

previous technologies. Fifty million or more size-selected short fragments are then denatured, and attached individually to an ultra-high-density flow cell. The flow cell is covered with two types of single-stranded oligonucleotides, each one complementary to one strand of the adaptor fragments. These oligonucleotides serve both as an attachment site, through annealing to the

adaptors, and as primers for DNA synthesis. The first round produces a DNA strand, covalently attached to the substrate, which is complementary to the full length of the initial fragment, including the distal adaptor. These attached DNA fragments then undergo exponential amplification *in situ* using a procedure known as *bridge amplification*. The free end of each fragment bends over and binds to its complementary oligonucleotide, which serves as a primer for a round of DNA synthesis. This is repeated many times and results in each original fragment being amplified into a discrete cluster of a few thousand copies. At the end of the amplification, the reverse strands are cleaved and removed, so that each cluster consists of identical single strands, all in the same orientation.

Sequencing primers are then flushed through the flow cell, binding to the free end of each fragment. In each round of sequencing, a mixture of the four deoxynucleotides, each one bound to a fluorophore with a unique emission spectrum, is flushed through the cell. Wherever a nucleotide is complementary to the first free base in a cluster, it is added to the growing DNA strand and unincorporated nucleotides are flushed away. This results in fluorescence of a wavelength indicative of the incorporated base, and this occurs in parallel in each one of the 50 million clusters on the flow cell. The deoxynucleotides that are used are reversibly blocked in the 3′ position, thus ensuring that only one base can be added at a time, even where there are runs of the same base. Once the results are recorded, the block is removed, the dye is released, and the process is repeated to read the next base of the sequence. The computer then assembles the sequence of each cluster by analysing the series of images.

8.2.4 Other technologies

The description above is limited to the main methods currently available. Others have been introduced more recently or are expected to be released shortly. Some of these new technologies employ radically different approaches, including the omission of an amplification stage – i.e., they are capable of reading the sequence of a single DNA molecule – and thus have the potential to produce another dramatic increase in sequencing speed, as well as a cost reduction.

8.3 *De novo* sequence assembly

Regardless of methodology, the same basic steps apply for assembling a sequence from separate reads. As the sequence of individual fragments is obtained, they are assembled into contigs (see Chapter 5), which are then sequentially joined together until a stage is reached when a reasonably confident decision can be made as to where all the bits go. It can then be labelled as a *draft sequence*. Note that at this stage there are likely to be gaps in the

sequence, although informed predictions can be made as to the position and size of those gaps. There will also be some uncertainties and errors in the sequence, as no automated alignment program is 100% perfect. The next stage is the finishing process, which involves filling in the gaps by more targeted methods and correcting the more obvious errors and uncertainties. A *finished* sequence will contain no known gaps (or at least none that it is technically possible to overcome), and will be accurate to a defined level. The final stage is *annotation*, when coding sequences are identified, predictions made as to the nature of the products, and a wide variety of other features are identified. Aspects of the strategies available are discussed further in the next section, and annotation and further analyses are considered later in this chapter.

At the end of the process we have a finished (and annotated) sequence, and we would say that we have determined the complete genome sequence of the target organism. We need to consider what we mean when we say this. For the bacterial genomes that have been sequenced, this is literally true – we know (within the chosen limits of accuracy) which base is present at every single position in the genome. But for eukaryotic genomes, the definition of 'finished' is set lower, for example, to mean more than 95% of the euchromatic regions (which contain most of the genes) sequenced, to the set standard of accuracy, and with no gap more than 150 kb. This definition will vary from one project to another. The original (2001) human genome sequence was far short of this standard – it included about 90% of the euchromatic sequence, with some 250 000 gaps. An improved version was published in 2004, which had 99.7% of the euchromatic sequence, with about 300 gaps (mainly repetitive sequences), but lacked 200 Mb of heterochromatic sequence, including the centromeres. This is probably not a serious limitation. For example, many of the gaps are due to the occurrence of highly repetitive DNA, which can be difficult to sequence (and probably does not encode any genes). Most, if not all, of the significant coding sequences will be present in the sequence that has been determined – and therefore for practical purposes we 'know' the sequence of the human genome.

8.3.1 Repetitive elements and gaps

Two problems encountered in genome sequencing are worth considering: repetitive elements and gaps. This applies to a varied extent to all of the technologies described above.

Most genomes contain *repetitive elements*: identical sequences that occur more than once (often many times) in the genome. The amount of repetitive DNA varies substantially between organisms, ranging from a few percent to nearly 50% in the human genome. There are various classes of repeat sequences. *Dispersed* (or *interspersed*) *repeats* are mainly mobile elements such as insertion sequences and transposons, which occur at different sites

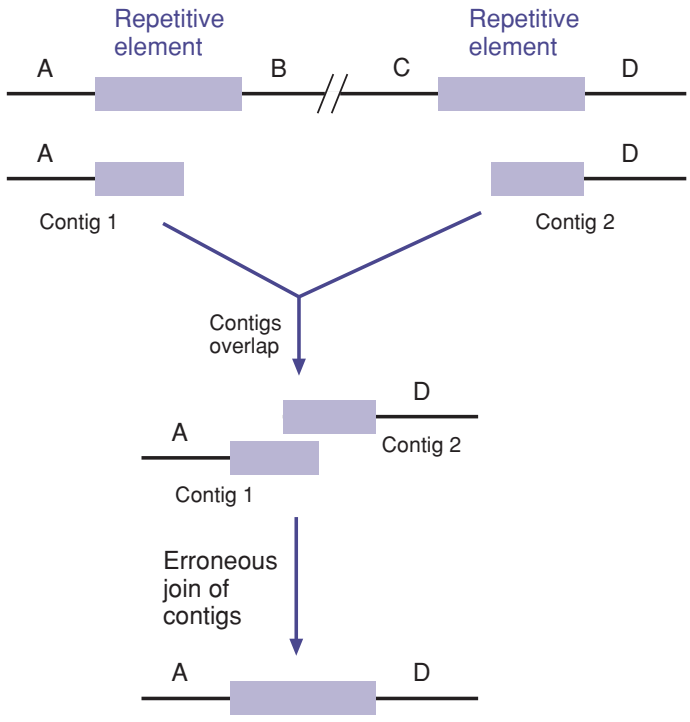


Figure 8.6 Repetitive elements cause a problem in sequencing.

distributed around the genome, while *tandem repeats* of shorter sequences (even as short as 1–3 nucleotides) may occur many times in succession at one locus. There may also be segmental duplications, where a block of perhaps several hundred kilobases of DNA has been copied to a different region of the genome, so called *copy-number variants* (see Chapter 9). Any repetitive element has the potential to cause problems in contig assembly. Two fragments of sequence that end with part of a repetitive element may be identified by the computer as overlapping, when in fact they are derived from quite different parts of the genome (see Figure 8.6). This is more of a problem with the total shotgun approach than when sequencing individual clones, because the separate clones are much less likely to carry more than one copy of an insertion sequence. Tandem repeats cause a rather different problem: the overlapping fragments are genuinely adjacent, but the number of copies of the tandem repeat may be miscounted. The length of individual sequence reads is an important consideration with any type of repetitive element. Longer sequence fragments are more likely to read right through the region containing the repeats, while shorter fragments may end within the repeated sequence, thus giving rise to false joins. All of the technologies described have some propensity to mis-sequence repetitive elements at the draft sequence stage.

In the discussion of smaller-scale sequencing projects (Chapter 5), we considered the problem of *gaps* remaining after contig assembly. These may be caused by difficulties in cloning certain fragments of DNA (with the Sanger method), or by problems in the sequencing reactions, especially with GC-rich regions that may cause termination of the polymerization reaction. Primer walking, as described in Chapter 5, is less likely to be useful in this context because many of the gaps will be too big. PCR can be used to amplify a fragment to bridge the gap, but only if the sequences either side can be aligned and if the gap is relatively short. If there are a number of gaps in the sequence, it is likely to be impossible to guess the order of the sequenced fragments, if you are sequencing an unknown genome. For resequencing (i.e., sequencing the genome of another individual from a species in which the full genome sequence is known), it is often possible to assemble these fragments in the correct order, using the existing sequence as a scaffold. Alternatively, if it is really important to close the gap, attempts can be made to clone the missing sequence, using a high-capacity vector such as a BAC (see Chapter 2), or a different sequencing method can be adopted. All these approaches are labour-intensive and not particularly conducive to the concept of high-throughput genome sequencing, but are not necessary for many purposes.

8.4 Analysis and annotation

Once the genome sequence has been finished, to meet the criteria of accuracy and completeness laid down in the project specifications, the final and important step is *annotation* – identifying the coding regions, the predicted products and other features of the sequence. Despite all the excitement that is generated by the publication of a genome sequence, knowing the exact sequence of As, Gs, Cs and Ts that make up a genome is only a means to an end, not an end in itself. The important factor is what you can do with that vast amount of information.

Genome sequence data are annotated in a similar way to that described in Chapter 5, but are often, for convenience, included in the databanks in sections that may correspond to actual clones (e.g., BACs). The fragments of the genome sequence may also be 'virtual clones' (i.e., arbitrarily divided sections of the genome). Whole chromosomes or whole genomes may also be included as a single sequence, and are useful for some forms of analysis – but downloading a whole chromosome sequence and analysing it may stretch the capability of your computer or your software. However, you can usually perform the analysis on line, using the capacity of the servers at the databank centres. See Box 5.1 for a selection of on-line resources.

8.4.1 Identification of ORFs

Identifying ORFs is relatively straightforward in bacteria, which typically lack introns. Initially, the identification of ORFs follows the methods outlined in Chapter 5, i.e., we find all start and stop codons in each of the six possible reading frames (three in each direction). But whereas with a specific clone we probably know the size of the protein, and may even have a partial amino acid sequence, in a genome sequencing project we want to identify all the possible ORFs, including ones that bear no relationship to any known protein. But we cannot simply label all these as potential protein-coding sequences. Unless we put a minimum size on it, the concept is meaningless – every reading frame between two stop codons is 'open' in the sense that it has no stop codons. On the other hand, by excluding ORFs below an arbitrary size, we would risk excluding genuine ORFs that do actually code for small polypeptides. But if we make the size limit too small, we run the opposite risk, of predicting expression of a small protein when no such protein exists. Therefore, a balance must be struck between the parameters we set and the expected false-positive (i.e., identification of ORFs that do not code for anything) or false-negative (i.e., missing small ORFs that actually code for something) rates that we are happy to accept.

To help identify real ORFs there are additional clues that we can use, with codon usage being a common parameter. As discussed in the previous chapter, the genetic code has many examples of different codons that are synonymous, i.e., they code for the same amino acid. Generally, these synonymous codons are not all used to the same extent in any one organism – there is a preferred codon usage, which in some cases can be very marked. Therefore, we can more accurately predict ORFs if we look for those that conform to the known codon bias for the organism in question. However, our knowledge of the codon usage of the target organism is often based on the sequence of a limited number of genes, so the predicted codon usage may not be accurate. Excluding ORFs that do not conform to this codon usage may, therefore, miss some ORFs, and we will reinforce our incorrect assignment of codon usage. Once again, such decisions on the use of these rules, and how rigorously they are applied, will be driven by the false-positive and -negative rates of ORF prediction that we are willing to accept. We may wish to err on the side of caution, and overpredict ORFs, and then use direct experiments to verify them.

These factors can be exemplified by the data shown in Figure 8.7. This is a portion (about 7 kb) of one of the cosmids used in the sequencing of the *Mycobacterium tuberculosis* genome. It is quite straightforward to locate all the stop codons on all six reading frames (three in each direction). In this case, the sequence (as a simple text file) has been read by a program known as Artemis (freely available from the Sanger Institute) to provide this display.

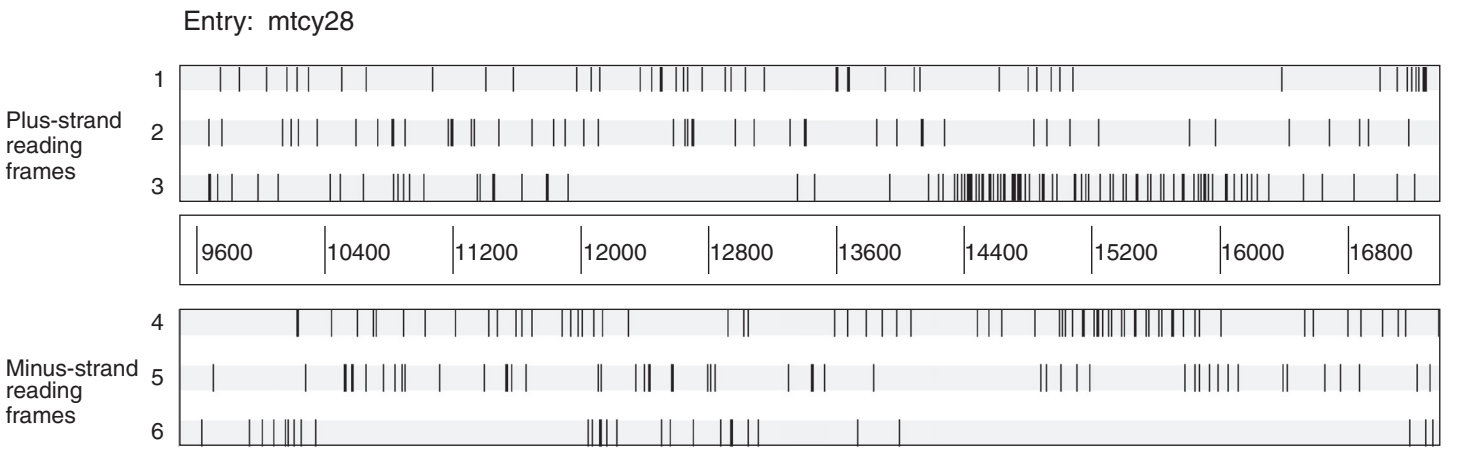


Figure 8.7 Open reading frames: computer mapping of stop codons. Edited display from an analysis of a DNA sequence using Artemis.

In the figure, the top three bands show the position of stop codons in the three reading frames that are read from left to right, while the bottom three bands show the complementary strand (read from right to left, 5′ to 3′). You will see that the distribution of stop codons is far from uniform. If we consider that only the largest open regions are likely to encode 'real' proteins, and therefore confine ourselves to regions of, say, more than 450 bases (coding for a sequence of 150 amino acids), there are three or more such sequences in each of the six reading frames. In some places, four of the six reading frames have no stop codons within such a distance, and it is highly unlikely that all of these are actually used to produce proteins. (In some organisms, especially bacteriophages, overlapping genes do exist, but this rarely happens in larger genomes.) So we then take start codons into account, and this reduces the number considerably, but still leaves more potential ORFs (including some overlapping ones) than are actually thought to exist.

In eukaryotes, *introns* are spliced out from the primary transcript as the *exons* are joined into an mRNA molecule (Figure 8.8). The genomic sequence will contain both introns and exons, and we need to be able to predict the intron-exon boundaries (splice sites) before we can attempt to identify and analyse the protein-coding sequence. In each eukaryotic species, intron/exon boundaries are surrounded by some more or less specific recurrent motifs. The most obvious and conserved of these is the GU-AG motif. Most introns begin with GU (encoded by GT in the sense strand of the DNA) and finish with AG. Obviously, this is not the whole truth. Both of these combinations of two bases will theoretically occur with a frequency of 1 : 16, so a longer, more variable, sequence is required to define a splice site. It is therefore not reliable simply to inspect a sequence and identify consensus splice sites. There are a variety of computer programs using statistical procedures that can be trained to recognize intron/exon boundaries in a specific organism, based on comparison of known cDNA and genomic sequences.

The computer can then effectively excise the predicted introns and thus assemble a *predicted* cDNA sequence. This is not totally reliable without some corroborating evidence. Fortunately, such evidence is often at hand. The most obvious thing to do is to search a databank for matching full-length cDNA sequences, from the same species or from a different one. Additional evidence can be obtained from screening databanks of *expressed sequence tags* (ESTs) (see Chapter 10).

Figure 8.9 shows the annotation for the sequence of a genomic clone corresponding to the cDNA sequence shown in Figure 5.7. It will be seen that the sequence contains seven exons, separated by six introns, making the total length of the gene (as measured by the primary transcript) over 15 kb, or about ten times the length of the cDNA.

A statistical technique that has been invaluable in the identification of protein-coding sequences, and in the prediction of intron-exon boundaries,

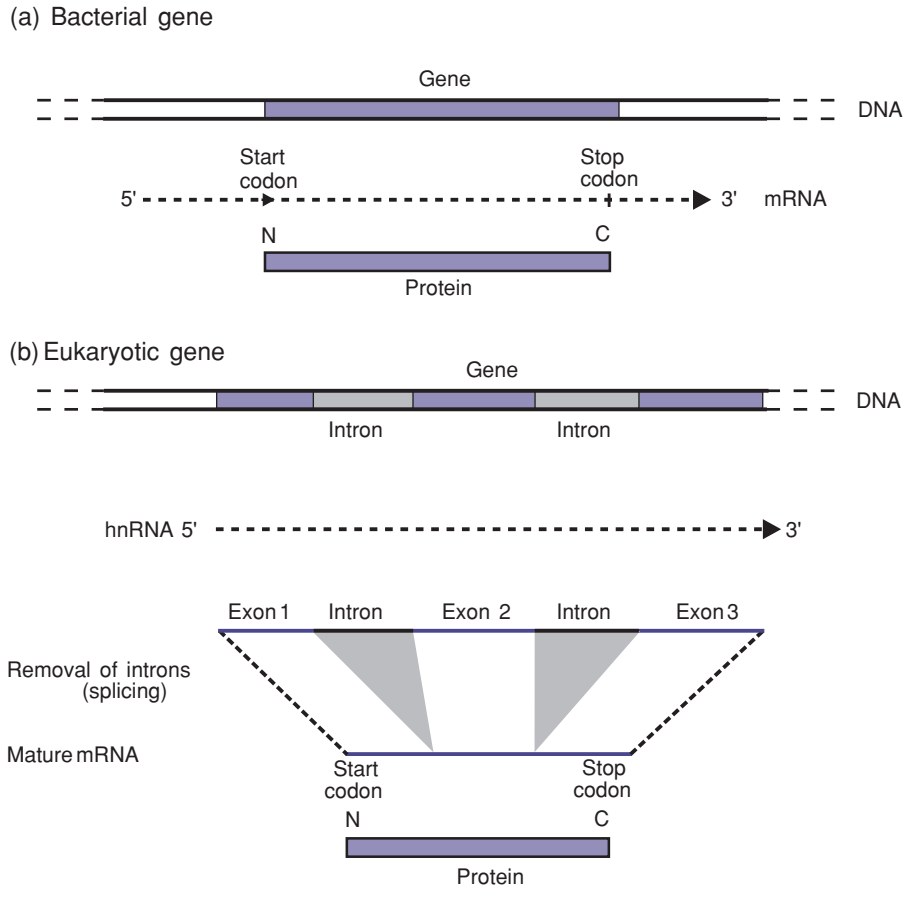


Figure 8.8 Introns and exons.

is known as a hidden Markov Model (HMM). A Markov model, or Markov chain, describes the probability of an event at one position (in this case, the occurrence of A, G, C or T), when this probability is influenced by events at one or more of the previous positions. For example, if you consider the words in this paragraph, the probability of finding the letter h is much higher if the previous letter is a t (there are 41 'th's, but only 33 'h's preceded by any other letter, or none at all). In a zero-order Markov model, the event at one position is independent of previous events, so you have a random DNA sequence, which may be true of non-coding DNA (although even non-coding regions are often not truly random). In a second-order model, the probability of a specific base at one position is influenced by the nature of the previous two bases, which is of course relevant to the situation with the triplet codons that occur within an ORF. Technically, an HMM considers a number of chains, only one of which consists of the observed states, the other

Accession number

```
ID   D00596; SV 1; linear; genomic DNA; STD; HUM; 18596 BP.
AC   D00596;
```

Definition

```
DT   17-JUL-1991 (Rel. 28, Created)
DT   14-NOV-2006 (Rel. 89, Last updated, Version 3)
DE   Homo sapiens gene for thymidylate synthase, exons 1, 2, 3, 4, 5, 6, 7,
DE   complete cds.
```

Link to the paper

Reference

```
RX   PUBMED; 2243092.
RA   Kaneda S., Nalbantoglu J., Takeishi K., Shimizu K., Gotoh O., Seno T., Ayusawa D.;
RT   "Structural and functional analysis of the human thymidylate synthase gene.";
RL   J. Biol. Chem. 265(33):20277-20284(1990).
```

Features table

Primary transcript

Coding sequence

Location of exons and introns

```
FH   Key             Location/Qualifiers
FT   source          1..18596
FT                   /organism="Homo sapiens"
FT                   /chromosome="18"
FT                   /map="18p11.32"
FT                   /mol_type="genomic DNA"
FT                   /clone="lambdaHTS-1 and lambdaHTS-3"
FT   prim_transcript 822..16246
FT                   /note="thymidylate synthase mRNA and introns"
FT   CDS             join(1001..1205,2895..2968,5396..5570,11843..11944,
FT                   13449..13624,14133..14204,15613..15750)
FT                   /translation="MPVAGSELPRRPLPPAAQ ..."
FT   exon            <1001..1205
FT                   /number=1
FT   intron          1206..2894
FT                   /number=1
FT   exon            2895..2968
FT                   /number=2
FT   intron          2969..5395
FT                   /number=2
FT   exon            5396..5570
FT                   /number=3
FT   intron          5571..11842
FT                   /number=3
FT   exon            11843..11944
FT                   /number=4
FT   intron          11945..13448
FT                   /number=4
FT   exon            13449..13624
FT                   /number=5
FT   intron          13625..14132
FT                   /number=5
FT   exon            14133..14204
FT                   /number=6
FT   intron          14205..15612
FT                   /number=6
FT   exon            15613..>15750
FT                   /number=7
```

DNA sequence

```
SQ   Sequence 18596 BP; 4521 A; 3991 C; 4479 G; 5605 T; 0 other;
     cctgtagtcc cagctacgcg agaggctgag gcagcagaat tacttgaacc caggaggcgg ...   60
```

Figure 8.9 Sequence annotation (EMBL) of a genomic clone. Some elements of the annotation have been omitted, and sequence data are truncated.

chains being 'hidden', but the full details of this approach are beyond the scope of this book. HMMs can be 'trained' to recognize the parameters associated with coding sequences, using DNA regions (from the organism being investigated) in which the genes have already been identified as 'templates' for further investigations. Note that we have previously encountered another use of HMMs, in the identification of protein domain families (see Chapter 5).

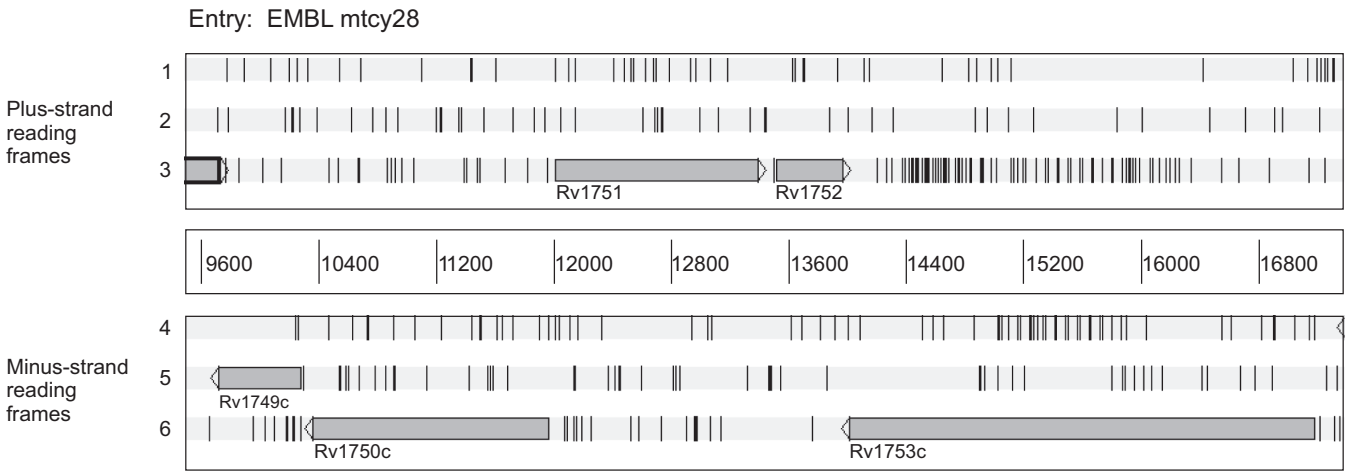
Another clue that we can use to aid in our identification of real ORFs is a comparison of all the identified ORFs with predicted proteins from other organisms. If the predicted polypeptide from our query sequence is similar to polypeptides that might be made by other organisms, then this supports the suggestion that our ORF is indeed a real coding sequence.

At this point it should be noted that if we do such a comparison using the DNA sequence itself, rather than the predicted protein sequence, we get information of a different sort. We will still find that the protein-coding regions are usually conserved (i.e., they are similar in other individuals from the same or closely related species), although the comparison will be rather noisier due to the existence of synonymous codons. However, we will also find that a substantial amount of non-coding sequence is also conserved, which would confuse our attempt to identify ORFs. Such conservation of non-coding sequences is, however, evidence that these regions actually do have a function, even if we don't always know what it is. If they were really 'junk' then we would expect them to vary much more as there would be no selective pressure to reduce the accumulation of variation between species.

As we saw in Chapter 5, databank entries contain a substantial amount of *annotation*, in addition to the sequence itself. This annotation shows the location of significant features of the sequence, and, in particular, the assignment of ORFs. Artemis is capable of reading the full databank entry and displaying these features. In Figure 8.10, we can see the output from Artemis when it is supplied with this extra information, for the same sequence as in Figure 8.7. This shows which of the potential ORFs in this region were considered (by those doing the annotation) to represent 'real' coding sequences, as well as (in some cases) the likely function of the product.

It is important to remember that the assignment of ORFs, and the possible function of the corresponding proteins, is only a prediction, and is subject to a degree of uncertainty in the absence of direct evidence as to the existence and properties of the encoded protein. In eukaryotes, alternative splicing is very common, meaning that further diversity is created by exons being joined together in different combinations. The necessary assumptions about the likely minimum size of a protein means that statistics on the number of predicted proteins have to be treated with care as they may miss an unknown number of small proteins. Hence, laboratory experiments are vital to confirm these predictions, and to extend our knowledge of the function of any potentially expressed protein.

It must also be remembered that the genome also encodes important RNA transcripts that do not encode proteins. The best known of these are ribosomal and transfer RNAs (rRNA and tRNA). A more recent discovery is microRNA (miRNA) in eukaryotes, which regulates expression of other genes (see Section 8.10 and Chapter 11). It is a salutary observation that the discovery of miRNAs was delayed by several years because of incorrect



CDS	9674	10231	Rv1749c	Possible membrane protein
CDS	10315	11913	Rv1750c	Possible CoA ligase
CDS	11967	13349	Rv1751	Possible hydroxylase
misc_feature	13224	13250		Dihydrofolate reductase signature
RBS	13464	13468		Possible ribosome binding site upstream of Rv1752
CDS	13476	13925	Rv1752	Unknown
CDS	13960	17121	Rv1753C	PPE family of Gly-, Asn-rich proteins

Figure 8.10 Open reading frames: display of coding sequences. Edited display from analysis of a DNA sequence and databank annotations, using Artemis.

assumptions about the length and sequence properties needed for a transcript to have a biological function.

8.4.2 Identification of the function of genes and their products

The methods described above will lead to a provisional assignment of ORFs, or coding sequences, in our genome sequence. The next question is the nature and function of those proteins. When we were dealing with the sequence of a specific clone (see Chapter 5), we started with a strong hypothesis as to what the protein actually does. However, if we are dealing with a whole genome sequence, we would be starting with no preconceived idea as to what each putative ORF actually codes for. In this case, we have to resort to comparing each predicted sequence against all the known protein sequences held in the databanks. The methods for doing this were discussed in Chapter 5. As we saw then, it is possible to carry out such comparisons with DNA sequences as well as with predicted protein sequences, but proteins usually give better results. There are a number of possible outcomes to such a search. If we are lucky, our protein may have a high level of similarity to one or more well-characterized proteins from another source. In this case, we are on reasonably safe ground in attaching that label to our protein as well. For genome sequences, the number of genes/proteins involved would make this a tedious task to do manually. In general, the initial annotation of genome sequences (including the assignment of ORFs) is done automatically by a computer.

There is a snag. Many proteins in the databanks, especially those predicted from gene sequences, have a functional label attached to them not because of an established function, but because of their similarity to another protein with an established function. Unless we look carefully at the evidence for the identity of these proteins, there is a risk of building up a chain of such similarities, becoming less and less reliable. Furthermore, we cannot be sure in all cases that enzymes with a similar structure actually carry out the same biochemical reaction. The enzyme beta-lactamase, responsible for ampicillin resistance in many bacteria, is similar in many respects to a serine protease, but it is not a proteolytic enzyme. It is also important to be clear about the concept of orthologues and paralogues. An orthologue is used to identify the same gene, producing a protein with the same function, in different organisms, and this is what we are trying to identify here. However, paralogues refer to two genes in a single species that evolved from a single antecedent gene; these paralogues are likely to encode proteins with highly similar functions. For example, the *CAR* and *PXR* genes in humans are paralogues, having evolved from a duplication of a single gene (such as the chicken *CXR*). Therefore, when making predictions it is important to ensure that we correctly match orthologues, for

example, *PXR* in rat and *PXR* in humans, and not mistakenly cross-match paralogues (i.e., *CAR* in rat with *PXR* in humans). To achieve this with a high degree of accuracy it may well be necessary to undertake formal phylogenetic analysis, as described in Chapter 9.

We may be less lucky. Our predicted gene product may not have a high degree of similarity to any known protein, meaning that we cannot use this comparative approach to gain insight into its function. But it may show some features that are characteristic of certain classes of protein, which would enable us to provisionally label it as, for example, 'probable membrane protein', or 'possible oxidoreductase'. Such features could be identified by the methods described in Chapter 5, using PROSITE or Pfam, for example. Or, it may show no discernible resemblance to any sequence in the database, in which case it has to be labelled as 'unknown function'. This applies to a substantial proportion of the predicted proteins in a typical genome sequence.

However good these predictions look, they are still only predictions, which need direct experimental evidence to prove them. This includes not only verification of the biochemical functions of the gene product, but also testing of its role in the physiology of the whole organism – including its potential role in causing a specific disease. The approaches outlined in Chapter 5, for testing a specific cloned gene, are not so easy to apply to a whole genome, but are currently being developed to allow such ambitious investigations to be undertaken. Later in this chapter we will consider some ways of surveying the whole genome for specific gene functions.

In the course of these sequence comparisons, we may also come across *pseudogenes*. These are DNA sequences which have significant sequence similarity to 'real' genes – i.e., to DNA sequences that are known or believed to code for proteins – but the pseudogenes contain changes in the sequence that make it unlikely to be functional. Pseudogenes occur through the process of gene duplication, whereby a single gene (or DNA region) is duplicated during replication; this 'copy' may then evolve to form a paralogue of the original gene, as we saw earlier in the case of *PXR* and *CAR*. However, it is possible for this copy to degenerate, such that it does not produce a functional protein, and it is then known as a pseudogene. The simplest type of change is one that puts a stop codon (or several stop codons) within what should be the coding sequence. Some pseudogenes may be transcribed into mRNA (*transcribed pseudogenes*), but these mRNA molecules cannot be translated into functional proteins.

8.4.3 Other features of nucleic acid sequences

The analysis of our DNA sequence does not end with the identification of protein-binding sites. One of the simplest, and yet very informative, analyses is the base composition of the DNA, i.e., the ratio between G+C and A+T

bases. Overall, this ratio is characteristic of a particular species, and tends to be similar between different species within a genus, while varying more widely between less closely related organisms (with ratios sometimes as low as 30% G+C or as high as 70%). Base composition can therefore be used as an aid to establishing the taxonomic relationship between different species.

Within bacterial genomes, the base composition tends to be reasonably uniform from one region to another, but there are notable exceptions. Figure 8.11 shows an example from the genome of *E. coli*, where the region indicated has a G+C content of 40.9%, compared to the average overall composition of 50.8% G+C. Such a region is known as an *island*, and is usually (as it is here) marked by a sudden change in base composition at each end. Many of the islands that have been studied in bacteria concern groups of genes that are connected with bacterial pathogenicity, and are hence referred to as *pathogenicity islands*, but the phenomenon is not restricted to virulence determinants. What is the significance of these islands?

Processes such as DNA replication and transcription, and particularly the regulation of these processes, are to some extent sensitive to the base composition of the DNA, as this will affect the amount of energy required to break the hydrogen bonds in order to access the single strands. Therefore over an extended evolutionary period, the enzymes involved in these processes, and the composition of the DNA, have evolved together to produce a well-balanced system. Furthermore, the codon usage of the genes is also related to the base composition – so as the codon usage and the specificity of the available tRNAs coevolves, this will also be reflected in the base composition of the DNA. The inference from this is that these islands, with a different base composition, are relatively recent arrivals. They represent DNA that has been acquired by the bacterium by horizontal gene transfer from a different species. For example, the genome of *Vibrio cholerae* contains an island in which the genes for the cholera toxin are found. It has been established that this island is in fact phage DNA that has been integrated into the bacterial chromosome. Many other islands have been shown to be integrated bacteriophages, either by direct evidence or by comparison of the sequence with that of known bacteriophages. In the example shown in Figure 8.11, the genes indicated are predicted to be associated with the production of fimbriae (pili), and may have been introduced by a bacteriophage.

Mammalian DNA also contains ‘islands’ with a different base composition from that of the remainder of the genome, but, as well as those derived from the integration of viruses, mammalian islands include regions with a different significance. The dinucleotide CG (usually written as CpG to emphasize that we are referring to consecutive bases on one strand rather than a base-pair) occurs much less commonly than would be expected from a random distribution of bases. Yet in some regions, the frequency of this doublet is very much higher, forming regions (CpG-rich islands) up to 2 kb in length, with a

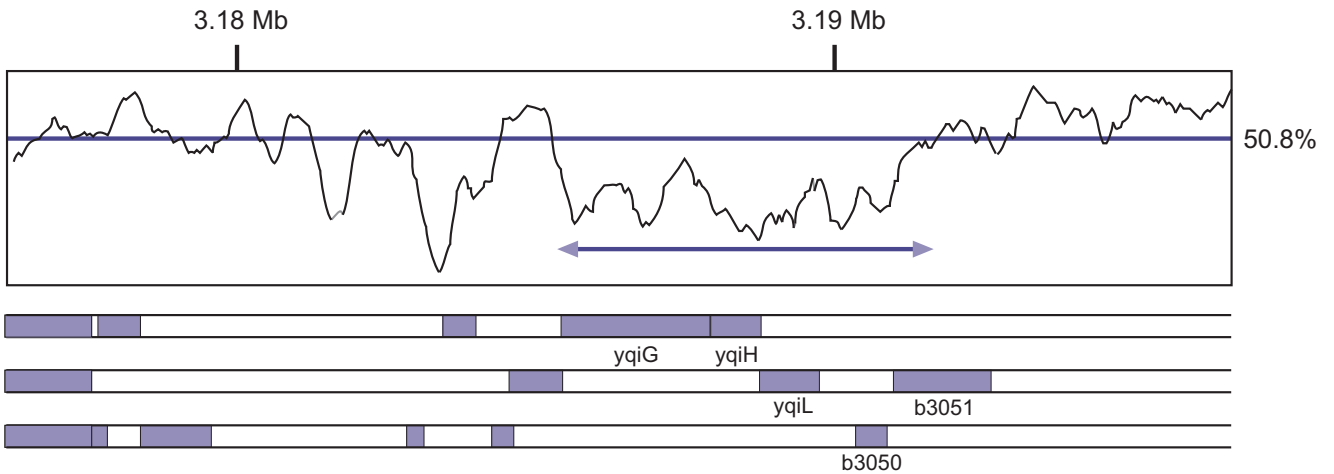


Figure 8.11 Variation in genomic G+C content. The figure shows edited Artemis output displaying part of the genome sequence of *E. coli*. The upper part shows the base composition (G+C%) and the lower part shows the genes identified in each reading frame.

much higher G+C content than that of the whole genome. There are many thousands of such regions in the genome. CpG islands are formed as a consequence of an epigenetic modification to DNA (see Chapter 10) that is used to regulate transcription, namely the methylation of cytosine within a CpG dinucleotide. In areas where transcription occurs within the genome, CpG is likely to be unmethylated, while transcriptionally silent regions are methylated. This methylation greatly increases the chance of CpG being mutated to TpG, meaning that the CpG is lost. Hence, CpGs tend to become associated with transcriptionally active regions of the genome.

Amongst the many other features of DNA that are amenable to computer analysis, we can single out the occurrence of inverted repeat sequences. It is important to be clear about the meaning of an 'inverted repeat'. Because DNA strands have a direction to them, and the two strands are in opposite directions, an inverted repeat of, say, CAT is not TAC but ATG (Figure 8.12). A pair of inverted repeats, in close succession, can anneal together, so that a single strand containing such a sequence will give rise to a *hairpin* structure, or if they are separated by a few bases, a *stem-loop* structure (Figure 8.13). A well-known example of such a structure is the tRNA molecule. Ribosomal RNA also exists in a highly folded conformation, and mRNA transcripts will also form folded structures. Although such structures are most common in single-stranded nucleic acids (mainly RNA), they can also give rise to localized destabilization of double-stranded DNA, most notably when the double helix is unwound during replication. One of the significant aspects of such structures is the role that they play in transcriptional termination; the formation of a stable stem-loop structure in the mRNA favours the dissociation of the nascent mRNA strand from the template DNA, allowing the two

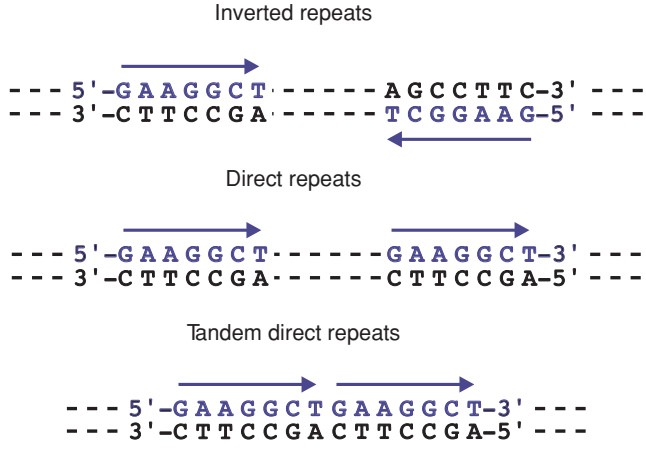


Figure 8.12 Inverted and direct repeats.

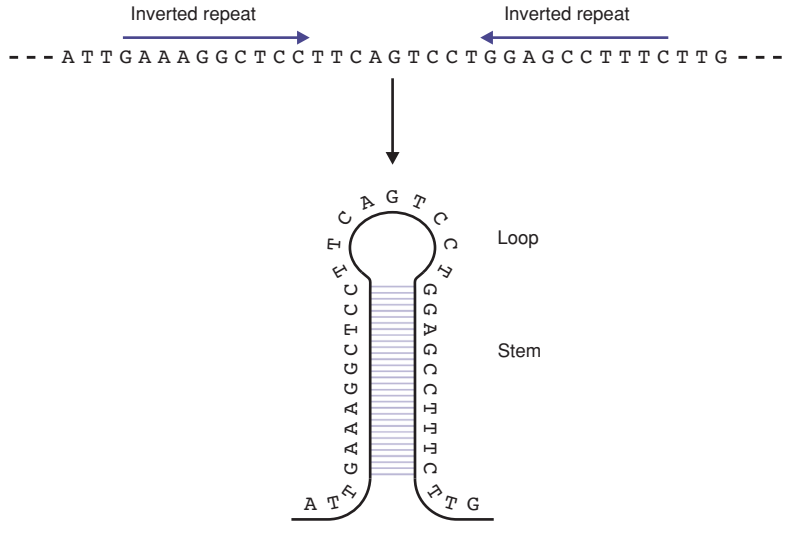
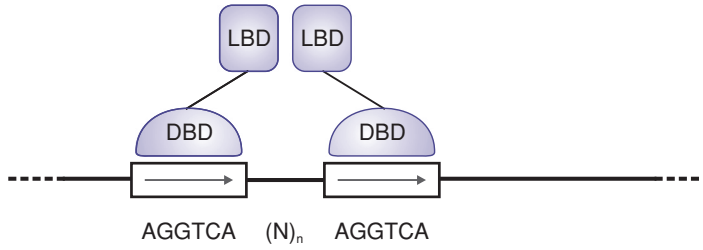


Figure 8.13 Formation of a stem-loop structure.

DNA strands to reanneal. This causes the RNA polymerase to pause, and ultimately stop transcription.

In addition to their role in transcription termination, inverted (or sometimes direct) repeats in DNA are often associated with the binding of regulatory proteins. Many such proteins are dimeric, so one subunit can bind to one copy of the repeat, and the other subunit to the second copy. The requirement for two binding sites gives a higher degree of specificity to the interaction, as well as establishing a strong interaction. Figure 8.14 shows examples from the nuclear receptor family of proteins.

Repeated sequences are also a frequent cause of variation (see Chapter 9). Inverted repeats are typically found at the ends of mobile genetic elements such as transposons and insertion sequences. In addition, recombination between inverted repeat sequences will lead to inversion of the region between them. In contrast, recombination between two direct repeats will cause deletion of the region between them. These sources of variation are not only an interesting natural phenomenon; they can be a nuisance in gene cloning as the occurrence of repeated sequences within your insert can cause instability of the construct. Tandem direct repeats (i.e., direct repeats with no intervening region) also cause variation in another way – by *replication slippage*. When the replication apparatus encounters a tandem direct repeat, it may (very occasionally) jump forwards or backwards, causing a loss or gain of additional copies of the repeated sequence. One example of this is found in Huntington's disease, where a CAG repeat in a gene called huntingtin has a tendency to slip and extend itself. Above a certain length, the encoded polyglutamate stretch in the resulting protein will cause the disease.



Nuclear Receptor	Binding site	Notes*
Peroxisome Proliferator-activated Receptor (PPAR)	→ $(N)_1$ →	DR1
Retinoic Acid Receptor (RAR)	→ $(N)_{2/5}$ →	DR2 or DR5
Vitamin D Receptor (VDR)	→ $(N)_3$ →	DR3
Thyroid Hormone Receptor (TR)	→ $(N)_4$ →	DR4
Glucocorticoid Receptor (GR)	→ $(N)_3$ ←	IR3
Oestrogen Receptor (ER)	→ $(N)_3$ ←	IR3
Pregnane X-Receptor (PTR)	← $(N)_6$ →	ER6
	or	
	→ $(N)_{3/4}$ →	DR3/4

*DR = direct repeat; IR = inverted repeat; ER = everted repeat (like an inverted repeat but facing in the opposite direction

Nuclear receptors tend to bind to DNA as dimers, and have a DNA-binding domain (DBD) and ligand-binding domain (LBD)

Figure 8.14 DNA-binding of regulatory proteins: nuclear receptor family.

Because the number of copies of a direct repeat at a specific site may therefore vary from one individual to another, it forms the basis of one method of molecular typing (see Chapter 9).

8.5 Comparing genomes

8.5.1 BLAST

In principle, the comparison of two genomes is not fundamentally different from the searches described in Chapter 5, using BLAST. However, in this case, the query sequence is not an individual gene or DNA fragment, but a whole genome. The sheer size of the query sequence presents some problems both with the analysis and the presentation of the results. Modified versions of BLAST are therefore used for this purpose. The problems of course become more severe if you want to compare the whole of one genome with the complete database, which contains many billions of sequences.

8.5.2 Synteny

Approaches such as BLAST will identify regions of two genomes that match one another, within the degree of matching required. But it does not tell us whether the genes are arranged in the same order. Generally speaking, the sequence of individual genes is more highly conserved than is their arrangement in the genome – which is interesting since it indicates that genes get shuffled around in the course of evolution. The matching of the gene order between organisms is known as *synteny*. A useful tool for displaying the degree of synteny, and the presence of insertions and deletions (*indels*), between two organisms is the Artemis Comparison Tool (ACT), from the Sanger Institute (see Box 5.1). An example of the comparison of part of two genomes (*Shigella sonnei* and *E. coli*) is shown in Figure 8.15. The blue blocks (red in the original) show regions that match, in the same orientation, while the grey blocks indicate regions that are in the opposite orientation. While there is a very large amount of agreement between the two genomes, there is one large block of DNA (and many smaller ones) that is inverted, and a number of smaller regions that are found in different places in the two genomes. In part (b) of Figure 8.15, we have zoomed in on a small part of the genome, which shows a small inverted region. Checking the annotations, we find that this contains the insertion sequence IS2. The mobility of insertion sequences can make a substantial contribution to the plasticity of bacterial genomes.

Another useful tool for genome sequence comparison is gMAP (see Box 5.1); an example of the use of this tool is shown in Figure 8.16. In the first step, a comparison of the entire genome sequence, based on pairwise BLAST comparisons, is shown for three strains of the bacterial pathogen *Streptococcus pyogenes*. As in the previous example, it can be seen that the genome sequences are made up of blocks that are broadly similar, but that these blocks vary in order and orientation between the strains. Clicking on region 3 produces a comparison of this part of the genome (Figure 8.16b), which confirms the overall similarity of this region (despite the different position and orientation in the genome), but also identifies a small region that is absent in one of the strains. This is labelled 3 in part (b) – the regions identified are renumbered each time, so this should not be confused with region 3 in part (a). Clicking on this region produces the full comparison shown in part (c), with individual genes identified by additional coloured bars. Each of these has a link to the sequence annotation so the genes involved can be identified. Three of them are labelled in the figure – identified as a phage tail protein, a phage repressor and a phage integrase, on the basis of similarity to known bacteriophage genes. This suggests strongly that this region consists of an integrated bacteriophage, which is absent from strain MGAS8232.

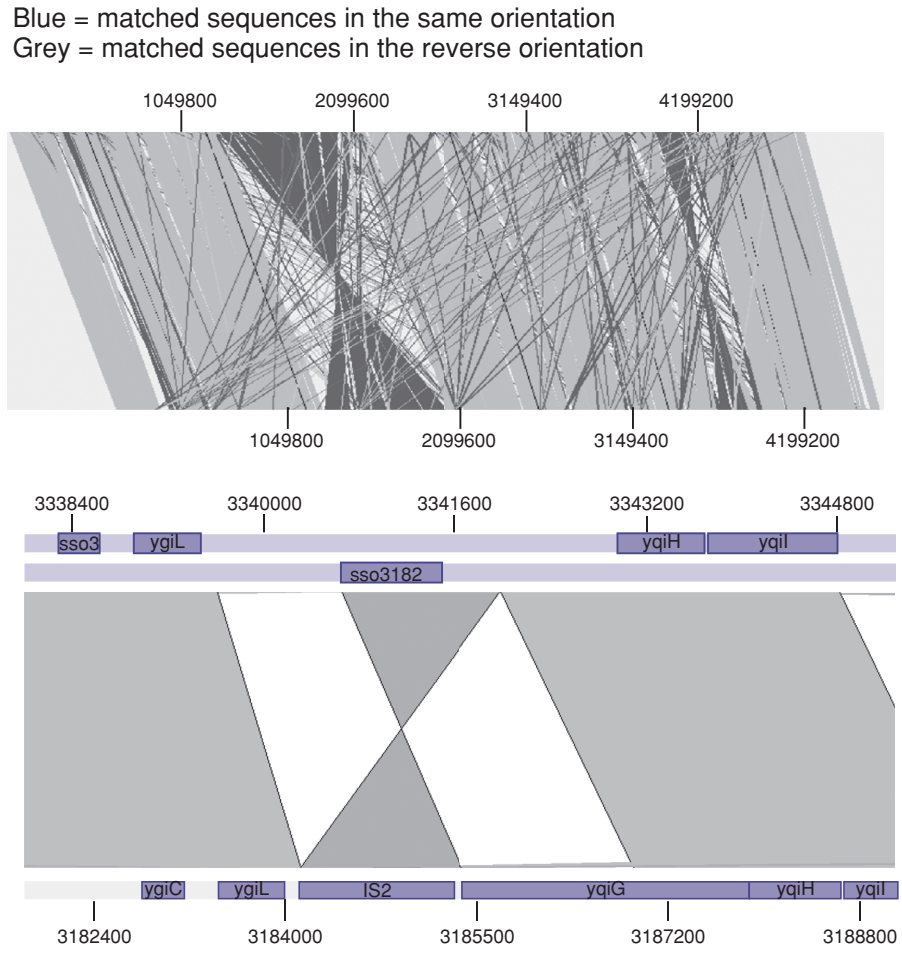


Figure 8.15 Genome sequence comparisons using ACT. (a) Comparison of the genome sequences of Shigella sonnei (top) and E. coli (bottom). (b) Detailed comparison of a shorter region of the same sequences.

8.6 Genome browsers

So far in this chapter we have described some of the ways in which we can make sense of genome sequence data. Fortunately, we have access to some immensely useful tools for viewing longer or shorter chromosome segments and their annotations using web-based *genome browsers.* These are basically viewing tools that access all the information from the databases from the latest version of each finished genome sequence. Because the data they access are the same in each case, you have the choice of using whichever genome browser you feel most comfortable with. The original genome browser is

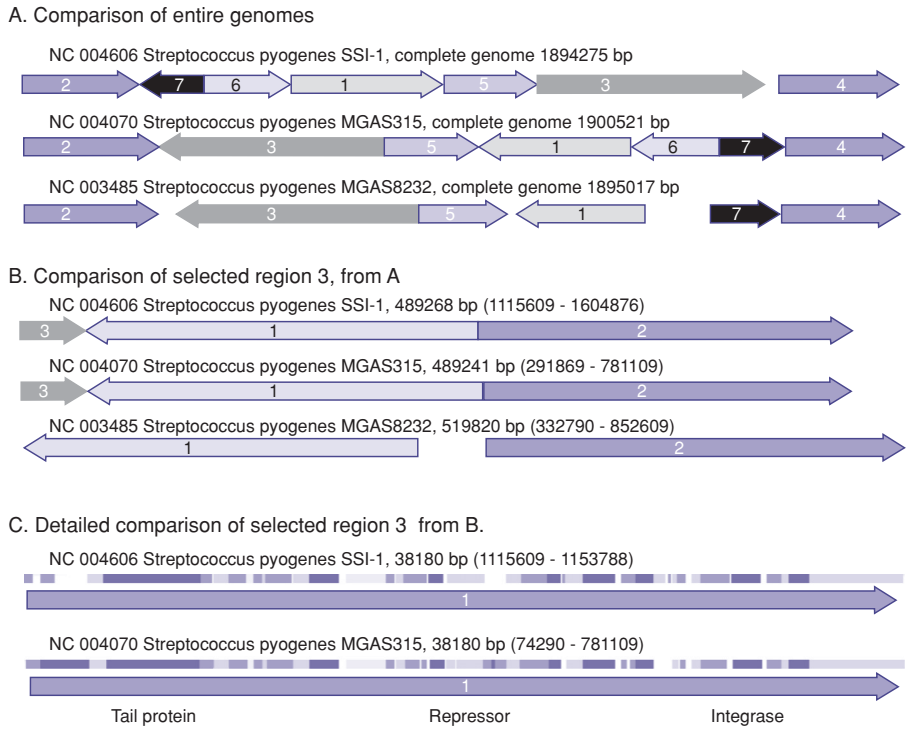


Figure 8.16 Genome structure comparison using gMAP.

hosted by the University of California at Santa Cruz (genome.ucsc.edu), while the most versatile and commonly used one is called Ensembl (www.ensembl.org). A third, simpler browser, the NCBI (National Center for Biotechnology Information) Map Viewer (www.ncbi.nlm.nih.gov/Genomes) is directly connected to GenBank. All genome browsers display your choice amongst the currently available sequenced animal genomes, with many other eukaryotes thrown in for good measure.

Basically, all of these browsers allow you to zoom back and forth between the chromosome view and the sequence view, with a number of different degrees of resolution in between. Through the browser's default annotation tracks, and through a number of optional ones, the sequence view is decorated with annotation information retrieved from the databases, including the position of genes, exons and coding regions, but also with optional features such as known polymorphisms, alternative splice variants, regions that are conserved in selected related organisms, etc. Apart from browsing, you are able to search for a specific gene (or a specific chromosomal region) and see it displayed. Also you can access links that produce graphic displays of predicted transcripts and proteins, and domains in these. The browsers also

allow you to define and export a smaller or larger part of a chromosome, and to compare it with other species.

8.7 Relating genes and functions: genetic and physical maps

We have already, in Chapter 5, discussed how we can use the primary DNA sequence of a gene to make predictions about the structure and function of its protein products. However, this approach has limitations. A substantial proportion of the genes identified in any genome sequence are not related to any gene with a known function, although there may be apparent ORFs coding for similar hypothetical proteins in other genome sequences. And always there is the cautionary note that bioinformatics can make predictions about function, but does not provide definitive answers. In Chapter 5, we also considered some of the ways in which we can address the relationship between a cloned gene and its function. We now want to look at some of the ways that are applicable on a genome-wide scale.

Another way of looking at the problem is to contrast the data obtained by classical and molecular genetics. Classical genetics works primarily with phenotypes, and produces a *genetic map*, which shows the position on the chromosome of genes associated with specific phenotypes. However, the genetic map does not tell us the structure of those genes, nor does it directly tell us about the biochemical function of those genes.

In contrast, molecular techniques are primarily concerned with the structure of genes and their sequence. This could start with cloning and sequencing a fragment of DNA; you could then use hybridization techniques to find the position of that sequence on the chromosome, producing a *physical map* of the chromosome. Ultimately, as we have seen, you can determine the complete sequence of the genome, which is the definitive physical map. It tells you exactly the DNA sequence at any position on the genome, but, taken in isolation, tells you nothing about the nature or function of any genes located within that region of DNA, or their products, and still less about the phenotype associated with them. Although we can make inferences on the basis of sequence comparisons, to advance our understanding further we have to be able to relate the genetic and physical maps.

To some extent we can tackle this problem from either end. For example, we could start with the classical approach, i.e., isolating specific variants and mapping the genes concerned. We would then fuse the techniques described in the previous chapters to isolate and clone the DNA region that is different in the mutant and wild-type organisms – in this way we are able to link the genetic map to the structure of a specific gene. However, to undertake this for all genes, and their multiple variants, using solely the techniques described

earlier in this book for individual genes would be rather time consuming. Instead there are a number of other methods that can be used to extend our observations on single genes for genome-wide studies, and we will concentrate on these during the latter part of this chapter.

The opposite approach to starting with a known variant, and then identifying the DNA sequence responsible for this effect, is to start with the DNA sequence instead. If we know the sequence of a piece of DNA, or the entire genome, we can infer the likely nature of the enzyme or other product coded for by each gene, by comparison with the sequence of known genes from other organisms. So we can work backwards from the physical map towards the genetic one. However, this approach has limitations, as discussed in Section 5.6.

There is a final limitation to this approach that is more fundamental in nature. We may have correctly identified the biochemical reaction carried out by the enzyme for which our gene is responsible. But this does not necessarily tell us what role that gene plays in the characteristics of the cell. For bacteria, and other unicellular organisms, it may be relatively straightforward to understand the role of enzymes that are components of a simple metabolic pathway, such as synthesis of an amino acid – but, even at this simplest of levels, such understanding is not always completely straightforward. The organism may have more than one gene coding for enzymes that carry out the same reaction, meaning that we would have to ask under what conditions each of those genes is used, or if the product for one gene is able to *compensate* for the absence of the other gene's product. With more subtle processes, it may be very difficult to ascertain the role of specific proteins – and if we consider complex processes such as the regulation of cell division it is likely to be impossible to determine the role of individual proteins just by examining their structure. If we then move on to consider a multicellular organism such as an animal, there is an even bigger jump from knowing the biochemical function of the protein to understanding its role in the whole animal. A multicellular organism may also use the gene product for more than one function. For example, in many vertebrates the gene encoding the enzyme alpha-enolase is also expressed at high levels in the lens of the eye, where it has no catalytic function but contributes to the transparent and refractive properties of the lens.

We can now look at some of the techniques that are available for constructing more direct links between genetic and physical maps, i.e., for establishing (or confirming) more directly the actual function in the cell of specific genes.

8.7.1 Linkage analysis

Linkage analysis is a classic technique for establishing how close two points are on the chromosome. So, if our genetic mapping data tell us that the gene

we are interested in is closely linked to another marker that has been characterized, we can narrow down the search for the gene of interest to a much smaller region of the chromosome. However, this requires the mapping of a very large number of genes if we are to be sure that there will be a mapped gene very close to our unknown gene. The distances separating known linked genetic markers in mammals commonly run to thousands of kilobases (1% recombination corresponds to about 1 Mb of DNA). But the second marker does not have to be a functional gene. It can be a polymorphic marker such as the microsatellites and single nucleotide polymorphisms (SNPs) described in Chapter 9. The process of recombination, where sister chromosomes exchange DNA strands, is central to this approach. It is logical that the closer together two points are on a chromosome, the less likely it is that a recombination event will occur between them; this is known as *linkage disequilibrium* (*LD*). Hence, if the unknown gene is often co-inherited with such a polymorphism (i.e., the two regions of DNA are in LD), linkage analysis can identify its position to a comparatively short region, which can then be cloned and characterized to identify the nature of the mutation that is responsible for the observed variation. This technique, known as *positional cloning*, has been used for the identification of important human genes such as *BRCA1,* variations in which predispose to breast cancer. The identification of genes associated with human diseases is covered further in Chapter 9.

8.7.2 Ordered libraries and chromosome walking

An *ordered library* is a special type of gene library that consists of a set of overlapping clones so that the position of each clone is known with respect to the clones on either side on the genome. This provides, in essence, a form of physical map of the genome. Any gene that has been cloned can easily be located to one of these clones by hybridization. (Of course if the genome sequence is known this is not necessary.) We can then use that as a starting point for locating other genes that are known to be linked to the first marker.

Construction of ordered libraries, especially of large genomes, is a laborious undertaking. A more generally applicable version of the technique is known as *chromosome walking* (Figure 8.17). Again, this requires as a starting point a marker that is known to be linked to the gene in question. This marker (A in the figure) is used to identify a clone from a gene library, by hybridization. That clone is then used to screen the gene library in order to identify overlapping clones; one (or more) of these clones is then in turn used as a probe to identify other clones that overlap with it. These steps are repeated until the required sequence (labelled B) is reached.

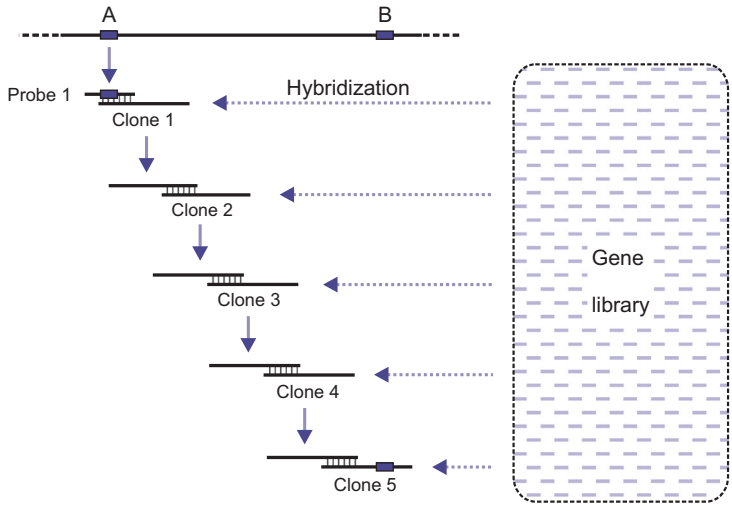


Figure 8.17 Chromosome walking.

8.8 Transposon mutagenesis and other screening techniques

8.8.1 Transposition in bacteria

Transposons are natural DNA sequences that have the ability to move from one position within DNA to another, hence their name. Part of the DNA of a transposon codes for an enzyme (transposase) that is capable of carrying out a special form of recombination, involving inverted repeat sequences at each end of the transposon, which results in insertion of the transposon at a new position, either on the same DNA molecule or on a different one. Transposons can thus move from one site to another on the chromosome, or they can move from a plasmid to the chromosome, or from one plasmid to another. Some transposons can insert themselves more or less at random while others have varying degrees of specificity. One of the most commonly used transposons in molecular biology is Tn*5* (or derivatives thereof), which is not very specific in its insertion site requirements, and hence can insert at a large number of positions; the reason why this non-specific integration is an advantage will become clear shortly.

A further feature of transposons that is relevant for undertaking experiments is that they generally carry antibiotic resistance genes. Indeed they play, together with plasmids, a major role in the spread of antibiotic resistance genes amongst pathogenic bacteria. But transposition does not only move genes between different sites. Insertion of a transposon within a coding sequence will usually inactivate that gene, thus producing a mutation. The site

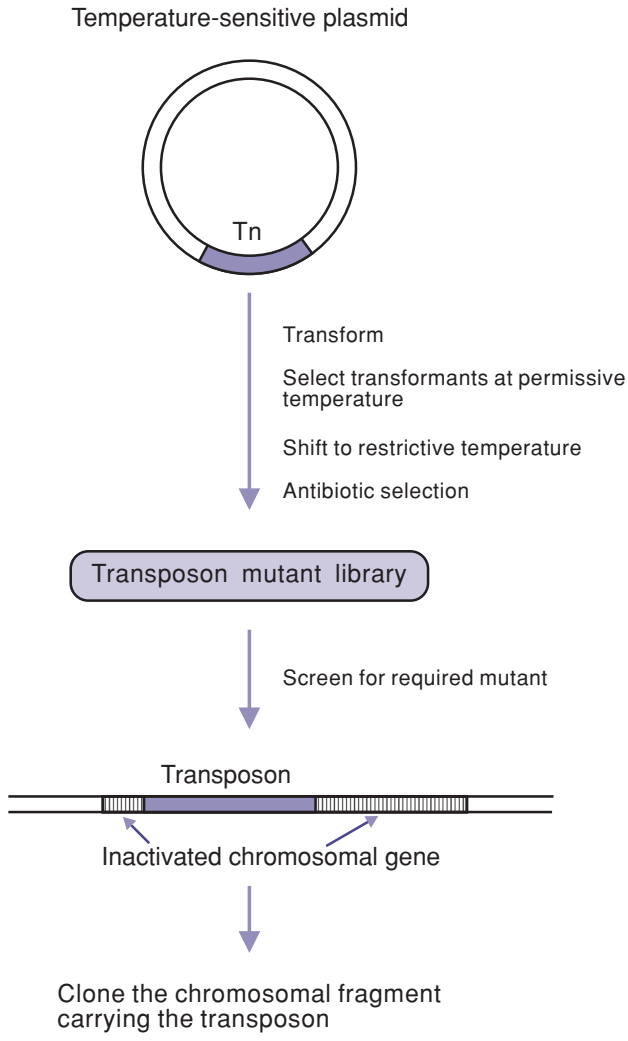


Figure 8.18 Transposon mutagenesis.

of that mutation is now marked by the presence of the resistance gene, which makes it relatively easy to clone, and thus to identify, the affected portion of the DNA.

The procedure in practice (illustrated in Figure 8.18) is to use a plasmid, carrying the transposon, which is unable to replicate in the host species being investigated; this is known as a *suicide* plasmid. Even better is to use a plasmid that is temperature-sensitive for replication, so you can establish the plasmid at a low temperature (e.g., 30°C, the *permissive* temperature) and subsequently prevent its replication by shifting the incubation temperature to say 42°C (the *restrictive* temperature). Inside the bacterial cells, the plasmid

is unable to replicate at the restrictive temperature. Thus, if we plate the transformed bacteria on a medium containing the relevant antibiotic, only those cells in which the transposon has hopped onto the chromosome will be able to survive and grow to form colonies. If this happens, the transposon will be replicated as part of the bacterial chromosome.

Of course we do not know where the transposon will have jumped to, and due to the non-specific nature of integration site for transposons such as Tn*5* there will be a large number of possibilities. To make the most of all of these possibilities, we can create a collection of cells, each with a random transposon integration, a *transposon mutagenesis library*. We can now use this library to identify all the genes that are related to a specific phenotype, by screening the library for mutants that are altered in this characteristic. We can even, using more powerful methods described later in this chapter, screen the library for inserts in essential genes (including genes necessary for virulence).

Once we have identified the mutant clones of interest, the next step is straightforward. We can extract genomic DNA from those cells, digest it with a restriction enzyme, and ligate these fragments with a suitable vector. In effect, we create a genomic library. But we do not need the complete library. We are only interested in those fragments that carry the transposon. We can identify these quite easily because they will contain the antibiotic resistance gene that is part of the transposon. So we just need to plate the library onto agar containing the relevant antibiotic, and only those clones that carry the transposon will be able to grow.

These clones will contain not only the transposon but also a portion of the DNA either side of the insertion site. Determining the sequence of this flanking DNA will therefore enable us to identify the gene into which the transposon has inserted, and we thus have a direct link between the sequence and the phenotype, i.e., we know (subject to certain limitations that are discussed below) that inactivation of that gene gives rise to that phenotype, and hence we can infer the function of that gene in the normal life of the cell.

As is so often the case, PCR provides us with an alternative to cloning for identifying the insertion site of the transposon. However, we cannot do a traditional PCR, because that would require knowledge of the flanking sequence for designing the primers, and that is exactly what we do not know. One strategy that we can adopt in such circumstance is known as *inverse PCR* (Figure 8.19). If we cut the DNA with a restriction enzyme (using one that does not cut the transposon itself), then, instead of ligating these fragments with a vector, we can carry out a ligation in the absence of a vector and under conditions that promote self-ligation (*intra*molecular rather than *inter*molecular ligation). Reference to Chapter 4 will show that this requires ligation at *low* DNA concentrations, whereas usually we do ligation at high concentrations of DNA to promote ligation of DNA fragments with the vector DNA. The consequence of self-ligation is that, amongst a lot of other fragments, we have

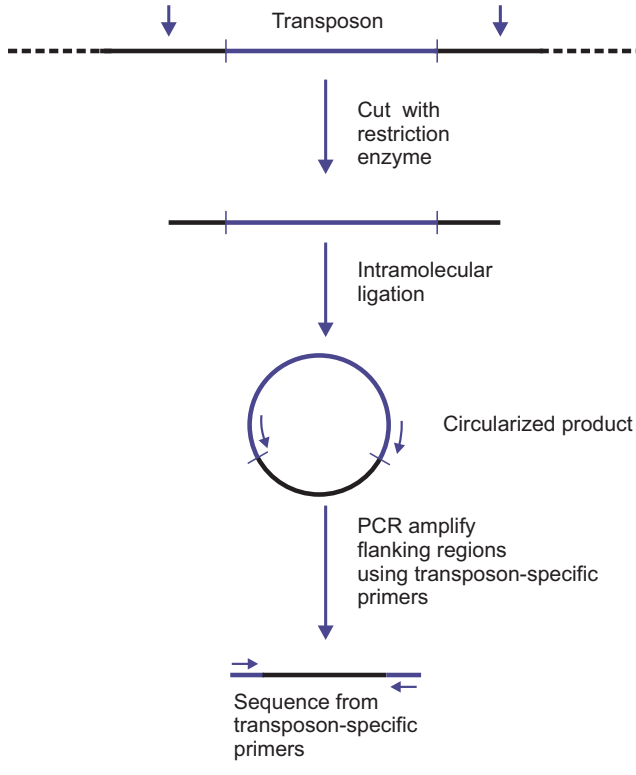


Figure 8.19 Locating an integrated transposon by inverse PCR.

circular molecules containing the transposon and the flanking sequences. Although this will be only one amongst thousands of products, the flanking sequences can be amplified by PCR, using primers derived from the known sequence of the transposon.

8.8.2 Transposition in *Drosophila*

The discussion of transposons and transposition has so far focused on bacteria. However, transposable elements of one sort or another are common in all types of organisms. The family of transposable elements known as *P* elements, which occur in the fruit fly *Drosophila melanogaster*, are especially important – both in providing vectors for the integration of foreign genes into the *Drosophila* genome and in providing a system for transposon mutagenesis of *Drosophila*. *Drosophila* is one of the most highly studied multicellular organisms, with many of the fundamentals of eukaryotic genetic regulation having been undertaken in these humble fruit flies. The ability to

use transposons in a similar way to that we saw with bacteria is an extremely useful tool in these investigations.

Transposition of *P* elements, as with bacterial transposons, requires the action of a transposase, acting on inverted repeat sequences at the ends of the element. In a P strain, which carries multiple copies of the P element dispersed throughout the genome, the transposase is repressed and so no further transposition occurs. However, if sperm from a P strain fertilizes an egg from a strain that does not contain a *P* element, the temporary absence of the repressor causes extensive transposition, resulting in a high rate of mutation.

The *P* element is also able to transpose into the genome from an injected piece of DNA. Therefore, if we insert a piece of foreign DNA into a *P* element contained on a plasmid vector, and then inject that construct into a fruit fly embryo, the *P* element will transpose into the genome, carrying our inserted DNA fragment with it. However, it is not easy to insert DNA into a *P* element without disrupting the transposase gene. Figure 8.20 shows how we can get round this problem. The transposase can act *in trans*, i.e., it can be expressed from a different piece of DNA. In the example shown, the foreign DNA fragment has replaced most of the *P* element genes, leaving the inverted repeat ends intact. The transposase is expressed from a second copy of the *P* element, and it will recognize the inverted repeats flanking the foreign DNA, resulting in transposition of the insert into the chromosome (with the IR ends). At the same time, we do not want the element with the intact transposase to be inserted as well, as it would cause additional mutations. So we remove the inverted repeat ends from the *P* element that has the transposase, rendering it non-mobile; this is referred to as a '*wings-clipped*' element.

The main applications of this approach lie in identifying or confirming the relationship between specific genes and identified phenotypes. As a simple example, Rosy$^-$ flies have brown eyes rather than red ones. Insertion of a

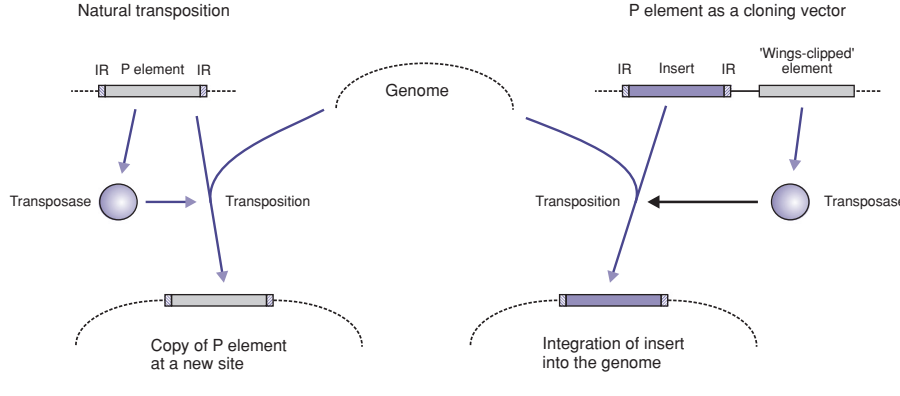


Figure 8.20 Transposition of *P* elements in *Drosophila*.

DNA fragment coding for the enzyme xanthine dehydrogenase will restore the wild-type eye colour, thus confirming the function of the *rosy* gene; this is an example of *complementation* (see Chapter 5). Fruit flies have been used extensively as a model research system for a multitude of simple and advanced functions, especially the differentiation and development of multicellular organisms. The ability to link phenotypes with specific DNA sequences in this way has been an important component of these advances.

The applications of *P* elements do not end there. Insertion of a *P* element into the chromosome can cause a mutation, and since the affected gene is tagged with the transposon, it is readily identified, as in bacterial transposon mutagenesis described above. Another application involves a *P* element containing a reporter gene (such as the beta-galactosidase gene, *lacZ*) with a weak promoter. Random insertion of this element into the genome will occasionally result in integration adjacent to an enhancer element, resulting in activation of expression of the reporter gene. This technique, known as *enhancer trapping*, enables the identification of enhancers and their specific activity in certain cell types. Related ways of using reporters to identify regulatory sequences, such as bacterial promoter-probe vectors, were described in Chapter 6.

These applications to fruit flies represent an example of *transgenics*, in that they include the manipulation not just of individual cells but of the whole organism. Further examples of transgenics, as applied to higher animals and plants, are discussed in Chapter 11.

8.8.3 Transposition in other organisms

Having seen the importance of transposon mutagenesis for the study of bacterial and *Drosophila* genetics, it is reasonable to ask if this approach can be used in a similar way with other organisms; the answer is yes. In yeast (*S. cerevisiae*), the endogenous *Ty* element provides a way of obtaining libraries of mutants. This is a retrotransposon, i.e., it transposes via an RNA intermediate, produced by a reverse transcriptase encoded by the element. However, insertion is not random, as it tends to insert preferentially within genes that are transcribed by RNA polymerase III (mainly tRNA genes), rather than protein-coding genes (which are transcribed by RNA polymerase II), which is a limitation. An alternative strategy is to use bacterial transposons such as those based on Tn*3*. These do not transpose in yeast, but transposition can be carried out in *E. coli* containing cloned yeast DNA. The mutagenized DNA is then reintroduced into yeast cells where it combines with the genome, replacing the native sequence.

Retroviruses and genetically engineered retrotransposons have also been used for gene tagging in mice and other vertebrates. Currently, such research

is undertaken primarily within cultured cell lines, although techniques are becoming available for application to whole animals.

In plants, T-DNA from *Agrobacterium tumefaciens*, which is also used as a vector for introducing foreign genes (see Chapter 11), can be used for insertional mutagenesis. Alternatively, endogenous transposons such as the maize transposon *Activator* (*Ac*) have been used to generate libraries of mutants in various plants including the model genus *Arabidopsis*.

8.8.4 Signature-tagged mutagenesis

A form of transposon mutagenesis can also be used to identify genes that are necessary for the virulence of pathogenic bacteria. Insertion of the transposon into such a gene will attenuate the organism – i.e., it will destroy (or reduce) its virulence. This will be manifested by a reduced ability to grow or survive following administration to an experimental animal, or in some cases by a reduction in its ability to survive attack by macrophages in culture.

It is not possible to select directly such a mutation, and testing the thousands of mutants in a transposon library is impractical. However, we can modify the transposon by incorporating a highly variable sequence tag so that each copy of the transposon is uniquely identifiable (Figure 8.21). We then produce a transposon mutant library, with these tagged transposons, and infect mice with a pool of transposon mutants. Those clones in which the transposon has inserted into a gene that is essential for virulence will be unable to replicate in the mice, and will therefore be absent when we recover the bacteria from the infected mice. We then use PCR to amplify all the tags that are present in the recovered bacteria, and label the collection of PCR products for use as a probe. Identification of the tags that are absent in this mixture is carried out by probing a membrane that contains a gridded array of DNA from the clones from the original transposon mutant library. Absence of hybridization means that the clone concerned was not present in the material recovered from the mice – and hence identifies this as a mutant in which an essential virulence gene has been inactivated by the transposon. The gene can then be recovered and identified as described above. This technique, known as *signature-tagged mutagenesis*, has proved to be an extremely powerful tool for the identification of virulence genes, or for the identification of any gene that is essential for the growth of the bacteria under defined conditions.

Another method for identifying virulence genes in bacteria (although not actually involving transposons) is known as *in vivo* expression technology, or IVET. This uses a promoter-probe approach (see Chapter 6). In the example shown in Figure 8.22, the host strain was a *purA* mutant of *Salmonella typhimurium (*i.e., it had a defect in purine biosynthesis, which makes it unable

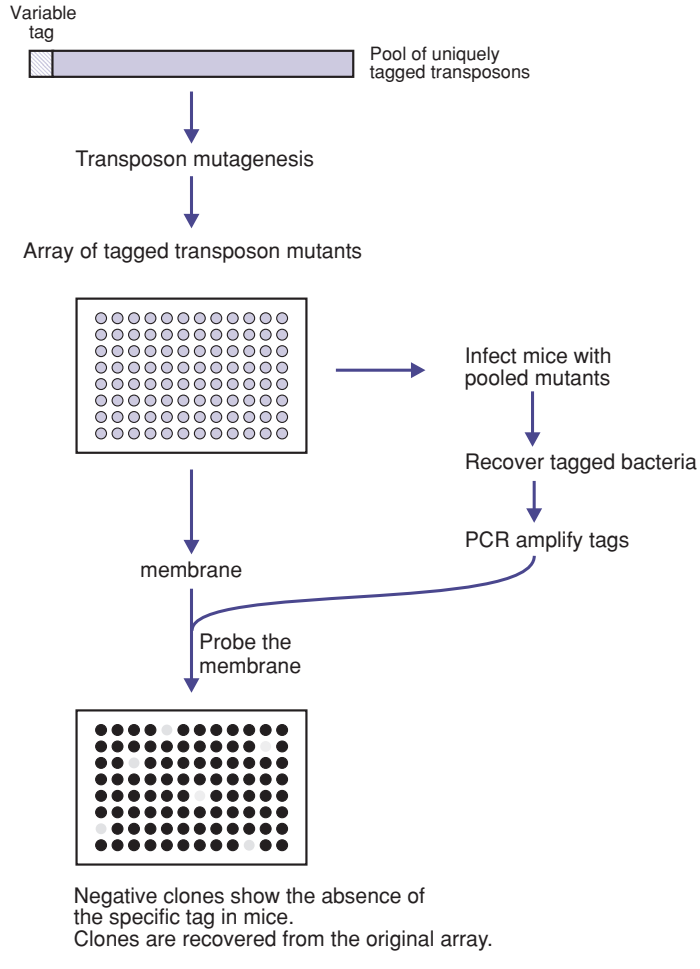


Figure 8.21 Signature-tagged mutagenesis.

to grow in experimental animals), and the vector contained a promoterless *purA* gene. Random fragments of *S. typhimurium* DNA were inserted in the vector to produce a gene library, and the constructs were integrated into the chromosome by homologous recombination. When animals were challenged with the mixture of clones, only those clones with a promoter that was active in this situation were able to express the *purA* gene – thus providing direct selection for promoters that were active during infection. The inference from this is that the genes naturally driven by those promoters are important for the infection process.

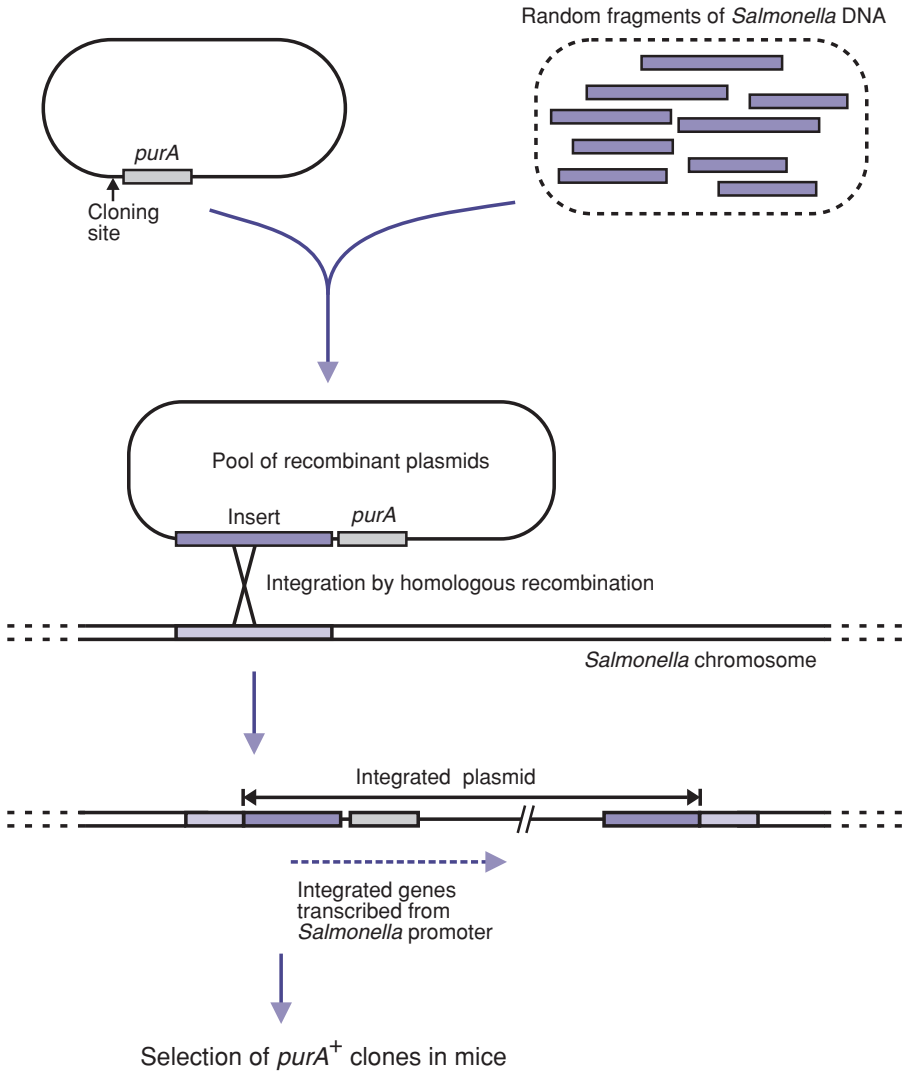


Figure 8.22 *In vivo* expression technology (IVET).

8.9 Gene knockouts, gene knockdowns and gene silencing

In Chapter 5, we described the use of gene knockouts, or allelic replacement, for studying the function of a specific gene. Although it might seem that this procedure is not suitable for genome-wide analysis, as it has to be done on a gene-by-gene basis, it is still possible to attempt it – for example, a large-scale collaborative project succeeded in knocking out 95% of the 6200 predicted

protein-coding gene in the yeast *S. cerevisiae*. A project is underway to knock out every predicted mouse gene (the NIH Knockout Mouse Project, or KOMP).

However, the discovery of RNA interference (RNAi) has made an alternative, and much more convenient, approach available. RNAi, or *gene silencing* (see Chapter 11), was discovered by finding that injecting double-stranded RNA into the nematode *Caenorhabditis elegans* resulted in switching off the corresponding gene. Andrew Fire and Craig Mello were awarded the Nobel Prize in 2006 for this discovery. The mechanism responsible (Figure 8.23) is

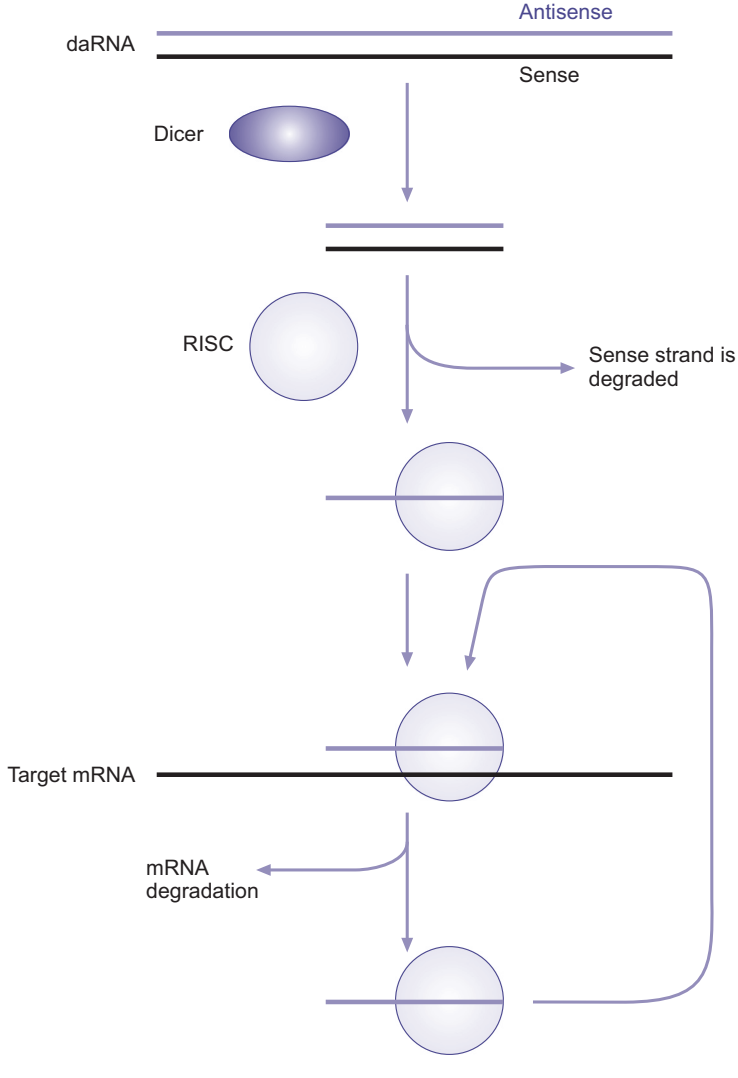


Figure 8.23 RNA interference.

that the dsRNA is cleaved by the enzyme Dicer into short fragments (short interfering RNA, or siRNA), which are 19 bases long with an extra two unpaired bases at each end. These are recognized by a ribonucleoprotein complex called RISC (RNA-induced silencing complex), which specifically degrades the sense strand of the siRNA. The antisense strand then binds specifically to the complementary region of the target RNA, and guides the protein complex to it, leading to specific degradation of that message. This is a very widespread phenomenon, in organisms ranging from protozoa to mammals, indicating that it is an evolutionarily ancient mechanism for genome defence, against viruses, and may also act in controlling mobile genetic elements such as transposons.

The relevance of RNAi technologies for this chapter is that instead of the laborious process of knocking out genes one by one to determine their function it is possible, with *C. elegans*, to inject the worms with dsRNA, or simply to put them in a solution containing dsRNA. Alternatively, you can feed them with bacterial clones expressing dsRNA targeted to specific genes. This is simple enough to do one gene at a time, or you can use mixtures (libraries) of dsRNA to screen for genes associated with complex properties of the organism. Human cells in culture can also be screened by transfection with libraries of siRNAs; such libraries are commercially available. Use of an expression vector designed to express short hairpin RNAs (shRNAs), which are also cleaved by Dicer to siRNAs, enables long-term studies of the effects of silencing specific genes. Since this procedure may leave a small amount of residual gene activity, as opposed to the complete removal of activity seen with gene knockouts, it is referred to as a *gene knockdown*.

RNAi also has a role in transgenesis and, potentially, in gene therapy, and we will return to those subjects in Chapter 11.

8.10 Metagenomics

Conventional genome sequencing, as described earlier in this chapter, assumes that your starting material is derived from a single species. But it doesn't have to be. Since the sequence is assembled from reads of random fragments, these can be derived from many different organisms, so you will end up with genome sequences, complete or partial, from all the organisms present in your original mixture. This is the basis of the powerful technique known as *metagenomics*.

The great advantage of this is that in many situations – ranging from seawater to the human gut – there are very many organisms present that we know nothing about. The vast majority of organisms present in such environments have never been isolated, or grown in the laboratory. And yet, by simply taking a sample and extracting and sequencing all the DNA in it, we

can obtain fundamental knowledge of all the organisms the sample contains, whether viruses, bacteria or eukaryotes. By comparing the sequence data to those of known organisms, we can establish their taxonomic relationships. And by analysing their gene content, we can often establish many of their properties, such as their nutritional requirements, and if they are capable of photosynthesis.

8.11 Conclusion

In this chapter we have moved from a consideration of genome sequencing and analysis of gene structure to linking the genome sequence with genome-wide studies of gene function. Further evidence of gene function comes from genome-wide studies of transcription and translation, producing lists of all transcribed genes (*transcriptome*) and all translated products (*proteome*). These are covered in Chapter 10. But before that, in Chapter 9, we want to look further at how we can study the variation between strains and species at the genome sequence level.

9
Analysis of Genetic Variation

The basic structure and organisation of genes, in eukaryotes and prokaryotes, was considered in Chapter 1. It is a central dogma of molecular biology that the inherited characteristics of an organism are a reflection of the structure and organisation of its genes. However, it is important to realize that this includes an extremely complex set of interactions between different genes, and their products, as well as environmental factors. For example, the gene(s) (i.e., genotype) directly responsible for an observed characteristic (i.e., phenotype) may be absolutely identical in sequence between two individuals, but the observed phenotype may be different because of variation in other genes that affect their expression, or because of alteration in other cellular components that affect the activity of the proteins encoded by those genes. Environmental influences in particular have a major role in determining the observed characteristics of the organism. Excessive reductionism – ascribing every change to one single gene – is an ever-present pitfall in the analysis of variation, as many traits are much more complex than that. In addition, we also need to remember that epigenetic factors (i.e., heritable changes in gene expression caused by mechanisms other than changes in the underlying DNA sequence, such as CpG methylation) play a major role in determining gene expression, and we will discuss these in more detail in Chapter 10.

We can use the study of genetic variation to examine differences between members of the same species, ranging from the study of bacterial characteristics (such as antibiotic resistance) to investigation of human genetic diseases, or to differentiate between individuals (e.g., in forensic analysis). Or we can compare the genetic composition of members of different species – even over wide taxonomic ranges – which can throw invaluable light on the processes

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

of evolution as well as helping to define the taxonomic relationship between species. Some methods of comparing genomes were considered in Chapter 8; we now want to look further at some methods that can be applied to the analysis of variation within a species.

The words *polymorphism* and *mutation* are often used in an interchangeable fashion as descriptors of genetic variants; however, it is important to recognize that they do have precise definitions. Formally, a polymorphism is the stable, multi-generational existence of multiple alleles at a gene locus (i.e., different versions of the gene). By comparison, a mutation is an individual event leading to an alteration in base sequence in one individual. In practice, the word polymorphism is used to describe a variant that occurs quite frequently (e.g., >1%) in a population, whereas mutation is used to describe rarer variants. Thus, by this definition, there is no fundamental difference between the two; every polymorphism, no matter how frequent it is today, will have started off as a mutation in one single individual before it became fixed in the population. However, the word mutation is also often used to describe alleles that cause disease. Most of these are rare enough to fit below the 1% cut-off, but some, such as the mutation in beta-globin that causes sickle-cell anaemia, are not. As we will see later, where we set the boundary is largely a matter of convenience. It is easier to screen for the more common polymorphisms (>1%), but the rarer ones (0.5%, or even less frequent) may be more informative.

When studying genetic variation, it is important to keep in mind the differences that *diploidy* makes in species such as our own. Clearly, having two copies of each gene allows more leeway for evolutionary drift in one allele, as long as the other one functions normally. This is why carriers of a recessive disease gene generally are not negatively affected by it, Indeed, sometimes (as in the case of the sickle-cell anaemia mutation in beta-globin) it actually confers an evolutionary advantage – in this case by conferring some resistance to malaria. When analysing genetic variation in diploid organisms, it is obviously paramount to keep in mind that the organism has two copies of each chromosomal gene (with some exceptions such as the sex chromosome genes in male mammals), and that the individual can be either homozygous or heterozygous for any given genetic variation.

9.1 Single nucleotide polymorphisms

The simplest form of genetic variation consists of a change in the sequence of bases at a single point. Such differences are called *single nucleotide polymorphisms* (SNPs). It is worth reflecting briefly on the consequences of such a simple change. The first possibility is that the polymorphism is within a region of DNA that does not undertake any important function; indeed in eukaryotes that is the most common occurrence (this is both because most

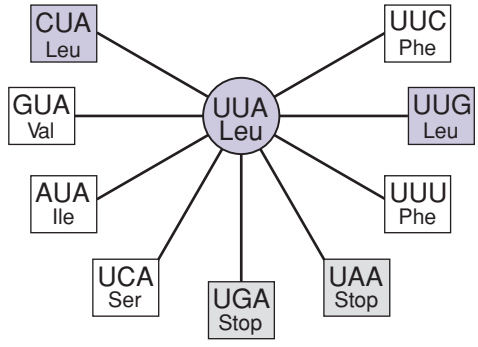


Figure 9.1 Codons arising by single base substitutions from UUA.

eukaryotic DNA is not a part of any gene, and because a non-coding sequence is where a polymorphism is least likely to have a negative effect). Secondly, if the change is within a gene, the chances are (again in eukaryotes) that it is in an intron and (unless it affects a splice site) is unlikely to affect the resulting gene products. Finally, even if it is within an exon, there may be no change at all in the sequence of the protein encoded by that gene. A single base change may alter for example a leucine codon such as UUA into CUA, which also codes for leucine (Figure 9.1). This is referred to as a *synonymous* substitution. It would be expected to have little or no effect on the organism (and hence is also referred to as a 'silent mutation') – although the change in codon usage and/or secondary structure could have some effect on the translation of the mRNA. Other changes may also be silent, although not synonymous. A change from UUA to GUA would replace the leucine with valine at that position in the protein, which may have little consequence for protein structure and function, since leucine and valine are similar amino acids. Other changes, such as UUA to UCA (serine), are more likely to have an effect, although such a change (and even more radical changes) can be tolerated at some positions. These non-synonymous (or missense) substitutions therefore may or may not give rise to changes in phenotype.

Most commonly, these changes will either have no significant effect or will simply impair or destroy the function of the product of the allele (loss of function). Rarely, they will lead to a protein that can function in a detectably altered way – for example, an enzyme that may be able to use a different substrate (gain of function). An accumulation of such mutations may eventually give rise to an enzyme that carries out quite a different reaction. This is of course the basis of the gradual process of evolution.

A further possibility is exemplified by the change from UUA to UGA in Figure 9.1. This is a *stop codon* (in the standard genetic code), i.e., it causes termination of protein synthesis (because of the absence of a tRNA that could recognize it). Such a change will therefore cause a truncated protein,

which will usually have lost activity but may also cause a pathological gain of function.

If instead of changing a single base we envisage a situation where a base is removed, then we have a *frameshift mutation*. Once the ribosomes pass that point in the mRNA, they are reading the message in the wrong frame – and the translation product will be totally different from that point onwards. Usually they will encounter a stop codon quite quickly, as the unused reading frames are generally well supplied with stop codons.

A point mutation may also occur within an exon but outside of the region that encodes the protein; these are the 5′ or 3′ untranslated regions (UTRs). One might assume that such mutations would not cause any discernible differences in the phenotype, but they may have an effect, because the UTRs have a function too, although a less easily predicted one than the coding region. In a similar manner to promoters, UTRs contain short regions of sequence that impact upon gene expression, although this time at the level of translation. These sequences control factors such as the stability of the RNA or its ability to dock with ribosomes and initiate translation, which is accomplished by the action of RNA-binding proteins. An alteration to the mRNA in this region may therefore affect its recognition by such a protein, which may have a significant effect on the translatability of the mRNA.

Whatever the consequences of a SNP may be, it is important to remember that the association of a specific SNP with a certain genetic trait does not prove that it is responsible for that trait (i.e., causative), directly or indirectly. It may simply be in a region that is close to a gene that is involved (i.e., it is linked to that gene) and therefore tends to occur more often in individuals that show the trait in question (i.e., it is associated with the trait, but not causative). It may therefore be a useful marker, but further work is needed to establish the nature of the gene that is causative for the trait under study.

We can analyse the occurrence of sequence polymorphisms in different strains, or different individuals, either by direct sequencing of PCR products, or by using allele-specific probes (including the use of microarrays). In addition, whole genome sequencing is now becoming a viable alternative to investigations of targeted regions of the genome.

9.1.1 Direct sequencing

The principle here is to PCR-amplify the desired DNA fragment, and to sequence it. This method does not require prior knowledge of the nature or exact location of any polymorphisms, so long as they are located between the two chosen PCR primers.

In a diploid organism your amplicon may be heterozygous for any polymorphism. This will mean that, in the case of an SNP, you will have two signals of more or less equal strength in the same position. In the case of a frameshift mutation, the whole sequence will become unreadable from the site of the

mutation, and the two PCR product variants would have to be cloned first in order to be sequenced. If the polymorphism affects the binding site for one of the PCR primers, the mutant allele will fail to amplify. However, the wild-type allele will amplify perfectly, and give you the incorrect impression that you are dealing with a homozygous wildtype.

9.1.2 SNP arrays

Microarrays consist of a glass slide containing a large number of tiny spots of a series of DNA fragments to which you hybridize labelled DNA fragments from the sample you are investigating. For SNP detection, your slide would contain a very large number of synthetic oligonucleotides corresponding to each version of the DNA sequence at each of the possible SNPs of interest. A robot is used to deposit each spot in a closely defined position on the slide. Once the DNA fragments are made, the machine can produce many identical copies of the same microarray. Hybridization to a labelled DNA sample can then detect which of the spots contain a DNA fragment that is present in the sample, thus enabling you to determine which sequence is present at each position (Figure 9.2). In this way, an enormous number of potential SNPs can be characterized for each sample.

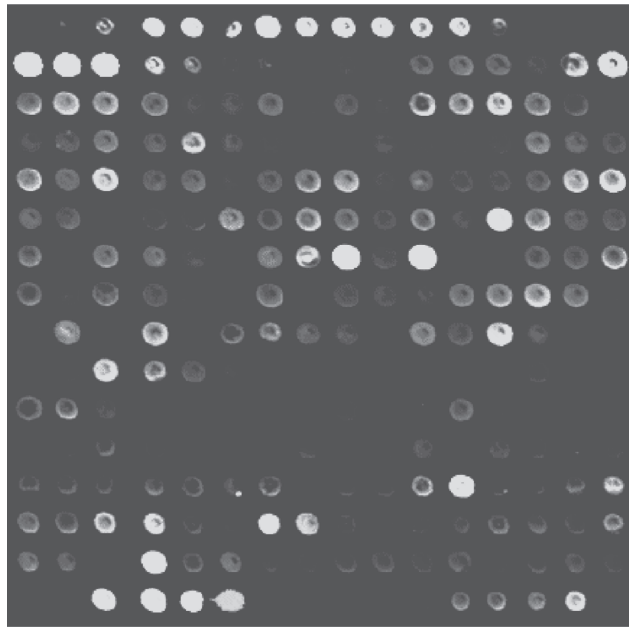


Figure 9.2 Use of a microarray. The illustration shows a representation of a part of a microarray, hybridized with a fluorescent-labelled probe. In practice, two probes with different labels would normally be used and the results displayed in red, green or yellow, depending on which probe hybridized to the greater extent.

Arrays containing over 2 million known human SNPs are now commercially available, and arrays with 5 million SNPs are likely to be on the market soon. This enables rapid screening of large populations of individuals for the presence of this enormous range of SNPs, which forms the basis of genome-wide association studies (GWAS). By comparing the presence of individual SNPs in people with or without suspected genetic diseases, or their occurrence/absence in other traits such as height, the statistical association of such SNPs with the trait in question can be determined. Note that this does not actually identify the gene in question, but it can locate the region of the chromosome where there may be a relevant gene, linked to the identified SNP. Later in this chapter, we will look further at the use of GWAS for locating genes associated with specific traits, as well as the use of microarrays for other aspects of whole genome comparison.

9.2 Larger scale variations

In addition to SNPs and other *point mutations*, there is a wide range of larger scale variations in the structure of the genome. These include *deletions*, *duplications*, *insertions*, *transpositions*, *inversions* and other rearrangements. Of particular importance in the context of this chapter is the activity of mobile genetic elements, including insertion sequences and transposons. Genome sequencing projects have disclosed large numbers of such elements in most species. In bacteria, in particular, they play an important role in the generation of variation by inactivation of the genes into which they are inserted; they also provide a convenient tool for the differentiation of distinct strains and species. Strains may also differ in their genome structure through the insertion of viruses. Examples of phage insertion and insertion sequences were given in Chapter 8.

In eukaryotes, duplication of entire or partial genomes has long been recognized as an important method of evolution. In some species, such as salmon, this process has occurred so recently that four copies of what is essentially the same chromosome can be recognized; such species are termed *tetraploid*. However, the ultimate fate of a tetraploid genome is that each chromosome pair will begin to amass additional changes, and thus revert to an overtly diploid state.

Until quite recently, subchromosomal *copy number variations* (CNVs) were assumed to be quite rare in genomes such as human, with increased variant number associated with severe abnormalities. However, as more and more individual genomes began to be sequenced, it was discovered that CNVs are in fact very common, and that essentially all human genomes are distinguished by CNVs that are either inherited or have occurred *de novo* during meiosis. Most of these are seemingly harmless (which is probably a major reason why they hadn't created a strong enough phenotype to be

noticed previously), but others are recognized as risk factors for diseases such as cancer and mental illness.

The best way of identifying CNVs is by genome sequencing. However, even in the modern era of next-generation high-throughput sequencing this is too cumbersome for large-scale surveys. A more common approach is to identify CNVs using microarrays, as described below. The basic premise behind this technique is the variation in the strength of the hybridization signal that will occur for different copy numbers, although this is less reliable than traditional microarrays, and routine microarrays do not identify precisely the position of the variation. This can be overcome by the use of short, closely spaced oligonucleotides. In one such study, using a microarray with 42 million spots at an average spacing of 56 base pairs, 8000 CNVs were identified, overlapping 13% of human genes, some of which were associated with conditions ranging from Crohn's disease to obesity. As before, it is important to remember that these CNVs are merely associated with the disease condition at this time, with further work needing to be undertaken to examine if any causal relationship exists

9.2.1 Microarrays and indels

In earlier sections, we referred to the use of microarrays for detecting SNPs and CNVs. They also provide a convenient way of finding other differences between closely related organisms (such as two strains of the same species), especially insertions and deletions. For this, you could either use a microarray spotted with larger clones or PCR products, or a high-density array of short synthetic oligonucleotides that are *tiled* in order to represent the whole genome. Whereas in traditional microarrays there may only be one or two probes for each genomic region, often just representing the predicted ORFs, a tiling array has many more probes, regularly spaced over the whole genome. Ideally, each probe overlaps slightly with its neighbours, producing a complete and unbiased coverage of the complete genomic sequence. Some tiling arrays fall short of this ideal, and contain short (but equal) gaps between probes. Hybridization of this tiled array to a labelled DNA sample can then detect which of the spots contain a DNA fragment that is present in the sample (see Figure 9.2). Further details of various types of arrays are considered in Chapter 10.

We can illustrate the principle of using such arrays to distinguish closely related genomes by a comparison of two bacterial strains, where the microarray is prepared using probes derived from the sequence of a standard reference strain. These are often obtained by PCR amplification of each identified open reading frame. Genomic DNA fragments from the standard strain are labelled with a red fluorescent dye, and fragments from the test strain are

labelled with a second, green, dye. The two DNA samples are then mixed and hybridized to the microarray. A microarray reader is used to compare the fluorescence arising from the two dyes on each spot. If both hybridize equally, there will be a signal from both dyes and the reader will display a yellow spot. If a gene is absent in the test strain, only the standard probe will hybridize, and you will see a red spot. If the converse were to happen – only the test DNA hybridizing – you would get a green spot, but in this example this would not be expected to happen, as the standard probe should hybridize to all the spots. This is therefore a very effective way of detecting regions of DNA that are present in only one of the two strains. As it is not easy to determine whether this represents an insertion in one strain or a deletion in the other, the term *indel*, meaning insertion or deletion, is often used.

In Chapter 10 we will look at another important use of microarrays, for genome-wide analysis of gene expression (*transcriptomics*).

9.3 Other methods for studying variation

Ultimately, the best way of detecting all the possible types of genetic differences that may occur between two or more individuals is the comparison of their genome sequences. The development of next-generation sequencing methodologies (as described in Chapter 8) has reduced the time and cost to a point where this is no longer an extravagant or unrealistic ambition – and both factors are still coming down rapidly. Whole-genome sequencing is playing a major role in our understanding of the evolution of the genome as a whole, and indeed is now a routine matter for organisms such as bacteria with relatively small genomes. However, there is still an important role for methods that do not rely on such knowledge.

9.3.1 Genomic Southern blot analysis: restriction fragment length polymorphisms (RFLPs)

One of the classic mainstays of the analysis of genetic variation is the Southern blot, which enables us to detect which band on an agarose gel is able to hybridize to a specific probe. The technique was described in Chapter 3 as a way of verifying the nature of the insert in a recombinant clone. In that case, there were only a few fragments on the gel. If we digest a total DNA preparation from an organism, we will usually get so many fragments that they will appear as a smear on the gel rather than as discrete bands. (We will consider later on, in Section 9.3.3, what happens if we use an enzyme that cuts so rarely that we get only a few very large fragments.) Southern blots enable us not only to identify specific gene fragments in such a smear, but also to compare the DNA samples according to the pattern of hybridizing bands.

If we use a probe that hybridizes to a DNA sequence that is present as a single copy in the genome, then we will detect a single fragment on the Southern blot. (This assumes that the enzyme used to digest the genomic DNA does not cut within the region that the probe hybridizes to; if it does, we will get more fragments.) The size of that band (or bands) will be determined by the position of the restriction sites flanking the detected sequence. If the structure and location of that gene is the same in each organism tested, then the band will be in the same position. However, if there is variation in the distance separating the two restriction sites, there will be a difference in the size of the fragment, and we will see a difference in the position of the detected band. This is referred to as a *restriction fragment length polymorphism* or *RFLP*.

There are a variety of changes that can give rise to such a polymorphism (Figure 9.3). The most obvious one – a point mutation causing loss (or gain)

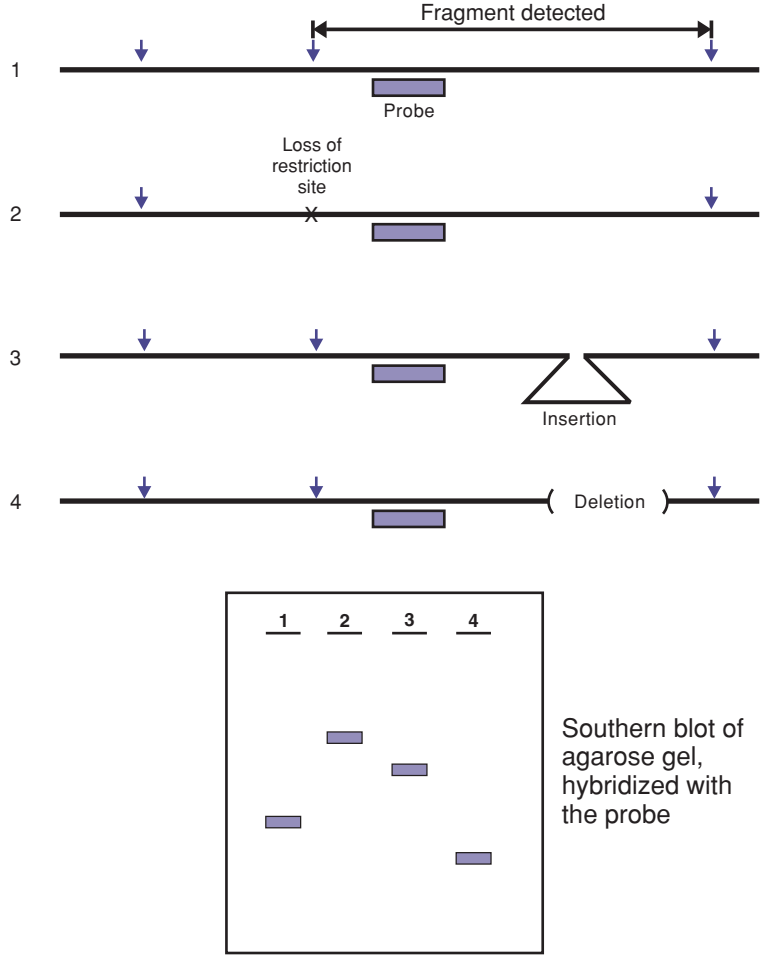


Figure 9.3 Restriction fragment length polymorphism (RFLP).

of a specific restriction site – will usually happen at a very low frequency. Insertions and deletions will make the fragment larger or smaller, respectively. Of course, if the insert itself contains a cleavage site for the enzyme used (or if the deletion removes a restriction site) this will also affect the sizes of the observed fragments.

For RFLPs to be useful for routine comparison of individuals or strains, we need to look at events that happen more frequently. The most useful polymorphisms arise through the duplication or transposition of repetitive sequences. Duplication is often associated with the occurrence of tandem repeated sequences, i.e., a short sequence of bases occurs twice (or more) in succession. When this region is replicated, mistakes can occur: the replication machinery may slip back from the second copy to the first, giving rise to an extra copy of the repeat; or it may (less commonly) slip in the other direction, causing a reduction in the number of copies. The repeat sequence can be quite short – for example, just a pair of bases, or even a run of the same base – in which case the change in the length of the restriction fragment will be small. PCR-based techniques are then more useful in detecting these changes (see Section 9.3.2 below).

If we use a probe that hybridizes to a mobile DNA element such as a transposon or insertion sequence, then when a copy of that element moves to a different site on the chromosome we will pick up a new fragment on the Southern blot (Figure 9.4). This is a widely used technique for the differentiation of bacterial strains (*typing* or *fingerprinting*) for epidemiological purposes. For example, the standard method for typing strains of *Mycobacterium*

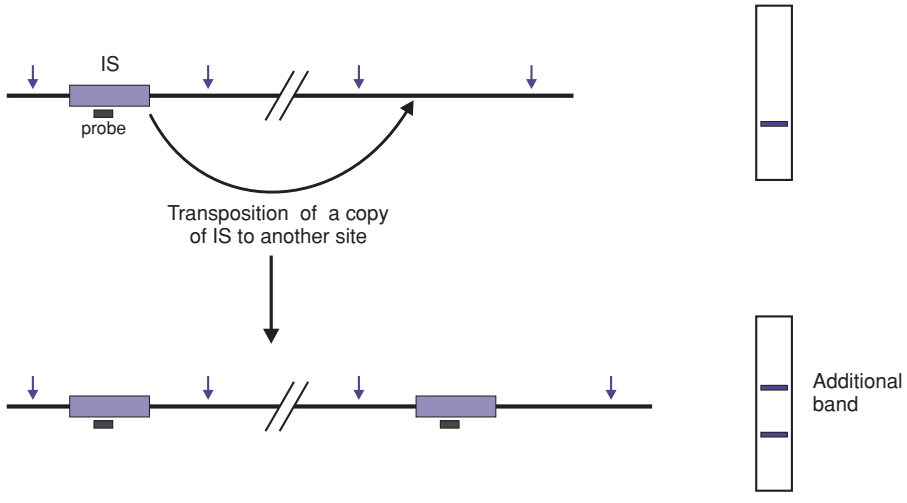


Figure 9.4 Restriction fragment length polymorphism arising from transposition of an insertion sequence.

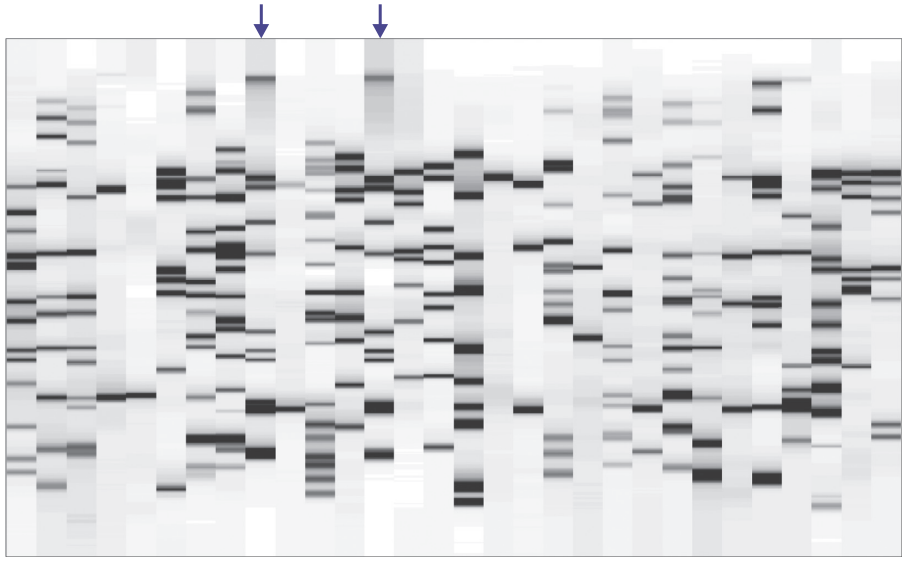


Figure 9.5 Fingerprinting of *Mycobacterium tuberculosis* using IS*6110*. Each track shows a different isolate of *M. tuberculosis*. Total DNA preparations were cut with *Pvu*II and electrophoresed through an agarose gel. Following Southern blotting, the membrane was hybridized with a labelled probe derived from the insertion sequence IS*6110*. The arrowed tracks show identical patterns, which may represent two strains from the same outbreak of tuberculosis.

tuberculosis, the bacterium that causes tuberculosis (TB), is Southern blotting of digested DNA using a probe that hybridizes to an insertion sequence known as IS*6110*. If two patients have caught TB from the same source, the IS*6110* patterns will be identical (see the arrowed tracks in Figure 9.5). If they are different, we can conclude that they are not part of the same outbreak. It should be noted that in this case we are mainly looking at changes arising from the location of the sequence detected by the probe, rather than alterations in a specific region of the genome, although the latter can contribute to the overall polymorphism.

9.3.2 VNTR and microsatellites

There are practical limitations to the conventional RFLP method described above: it requires a relatively large amount of DNA for the signal to be detectable, and you need to blot and probe it in order to obtain a readout, which is time consuming. For many practical applications, including forensic examination and many types of clinical diagnosis, this approach is either not convenient or not feasible. Generally, we have to use PCR to amplify the tiny amount of DNA that is available for investigation, and we then need a different approach to identify variations rapidly and robustly.

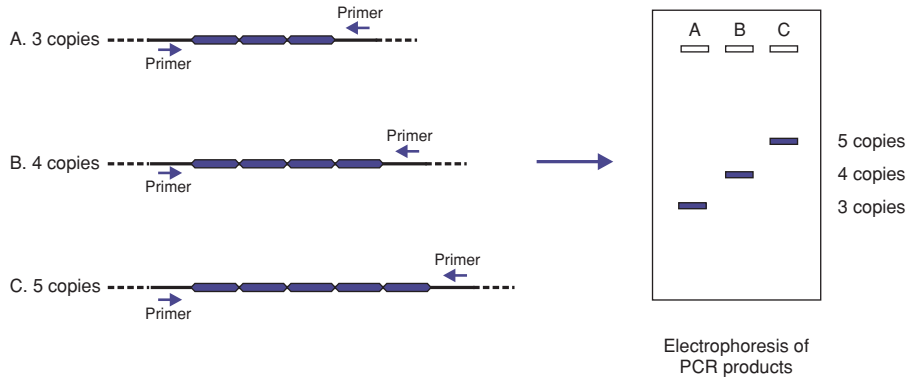


Figure 9.6 Variable number tandem repeats (VNTRs).

One approach exploits the tandem repeats referred to previously. But instead of using a probe and a genomic Southern blot, we can amplify just the region containing the tandem repeat, using primers that hybridize either side of the tandem repeats. The product is then separated in an acrylamide gel, which has greater resolution than agarose gels and can distinguish sequences differing by only a single base pair. We will see that the size of the product varies in a stepwise fashion according to the number of copies of the repeated sequence (Figure 9.6). This gives the technique its name – VNTR (variable number tandem repeats). Each strain can then be given a number indicating the number of copies of that tandem repeat. By identifying several loci that contain VNTRs, and determining the number of copies at each position, we arrive at a composite designation such as 3433243. This is now widely used in bacterial genotyping instead of RFLPs.

A very similar concept, which has proved extremely useful in human genotyping, involves the use of shorter tandem repeats known as *microsatellites*, or *short tandem repeats* (STRs). These are curious short stretches of sequence of 200 bases or less. They look almost as if the genome got bored with randomly distributed bases and instead decided to go for a straight stretch of one or two (less commonly three or four) nucleotide bases repeated a hundred times or so (such as TTTTTTTTT. . . or GTGTGTGTGTGTG. . .). Nobody knows what these repeats do, but they are extremely variable in length (no doubt because of 'slipping' during replication) as well as randomly distributed throughout the genome, and it is this that makes them so useful. By performing PCR with a battery of primers flanking microsatellite markers, a genetic fingerprint can be obtained that makes it possible to draw conclusions about an individual's identity and genetic relationship with others with great confidence. Such STRs are now commonly used to provide proof of identity for paternity or forensic observations, due to the highly variable nature of

these STRs. Studying only 12 such loci produces a power of exclusion of over 99.99%, meaning that the chances of incorrect DNA matches are incredibly small.

9.3.3 Pulsed-field gel electrophoresis

Most of the methods we have looked at so far are based on the examination of variation in limited regions of the genome. For example, even with a method such as RFLP fingerprinting of bacteria using an insertion sequence, where the insertion sites giving rise to the polymorphism may be more or less randomly distributed over the whole chromosome, the actual regions involved in the comparison of two strains represent less than 1% of the whole chromosome. Although this might be sufficient to allow us to distinguish the two strains from each other at the genetic level, it does not really provide a full comparison of their genomes. Short of complete sequencing of all the genomes to be compared, how can we look at the global structure of the genome?

In principle, you might expect that since a restriction endonuclease will cut genomic DNA at defined positions, you would get a characteristic pattern of restriction fragments from a given individual. Any deletions, insertions or transpositions (and some base substitutions) would give rise to a change in that pattern. And in principle you would be correct. However, if we consider an enzyme that requires a six-base recognition sequence (a 'six-base cutter') then, assuming an equal proportion and random distribution of all four bases, we would expect this enzyme to cut on average once every 4096 bases. If you are working with a bacterial genome of, say, 4 million (4×10^6) bases, you would expect about 1000 fragments. Standard agarose gel electrophoresis does not have enough resolving power to separate out such a large number of fragments. This problem would become even more acute for organisms with larger genomes, such as eukaryotes, where you would produce many hundreds of thousands of fragments. The obvious solution to this problem is, therefore, to produce fewer fragments to resolve in our electrophoresis. There are restriction enzymes that will cut DNA less frequently because they have a longer recognition sequence and/or a recognition site that occurs less frequently than expected on a random distribution (see Chapter 2). Such an enzyme will produce fewer fragments – but the problem now is that the fragments are too big to be separated by conventional gel electrophoresis. These large fragments *will* actually move through the gel, but in a size-independent manner. This is most easily visualized as considering the DNA fragments as very long and very thin structures; once they eventually get lined up end-on, parallel to the direction of the electric field, they can then slide through the pores in the gel. As long as the field direction is constant, these long fragments will stay orientated and move through the gel. As the movement is size

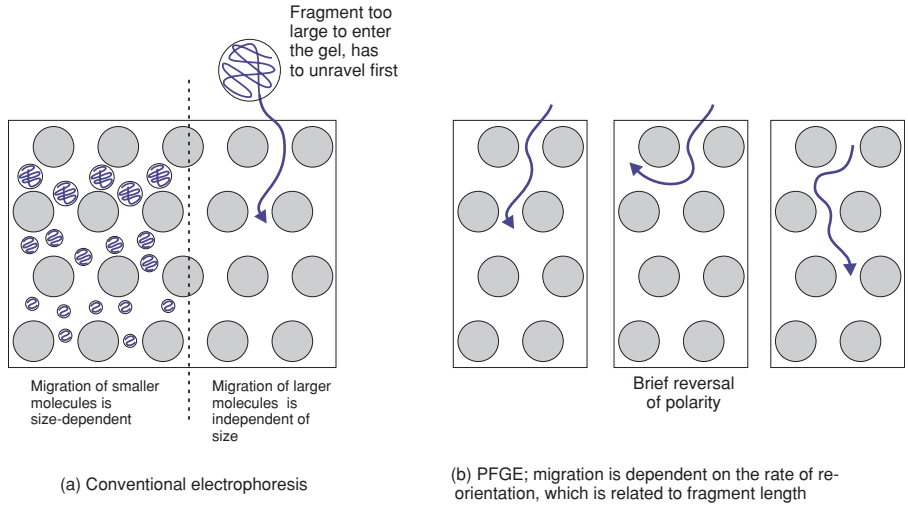


Figure 9.7 Pulsed-field gel electrophoresis (PFGE).

independent, the limiting factor for the rate of movement is actually the time taken to orientate the fragment into a long, thin structure.

The technique of *pulsed-field gel electrophoresis* (PFGE) exploits this phenomenon. The simplest version to describe is one where the polarity of the electric field is briefly reversed periodically throughout the run (Figure 9.7). Each time the field is reversed, the molecule has to be reorientated in order to move through the pores in the opposite direction, and the longer the molecule is, the slower this process will be. Consequently, the net migration of the fragments through the gel will be size-dependent. (In reality, this approach leads to unacceptable distortions of the lanes, and the apparatus actually used for PFGE employs more complex systems for changing field direction in order to achieve straight and coherent lane patterns).

Successful application of PFGE requires the isolation of genomic DNA intact or at least as extremely large fragments. This is virtually impossible by conventional means, as the DNA tends to fragment when the cells are lysed. Instead, intact cells are embedded in agarose, and both lysis and restriction digestion are carried out in a plug of agarose excised from the initial block. The agarose plug is then inserted into the gel and the DNA fragments migrate from the plug into the gel proper during electrophoresis.

Pulsed-field gel electrophoresis has been employed for bacterial typing, but the technical demands of the process make it less popular than the other methods described. It should be noted, however, that it does provide a method (although not the only one) for the preparative separation of chromosomes from eukaryotic cells.

9.4 Human genetic variation: relating phenotype to genotype

One of the most important applications of gene technology lies in the ability to detect genetic differences that result in specific traits, or in an increased predisposition to the development of these traits. These traits include factors such as obesity or height, as well as a wide range of diseases. We have already discussed, in different chapters of this book, various methods employed for identifying mutations that cause distinguishing traits or diseases. Here, we will take the opportunity to look back at the general strategy and integrate the different parts we have touched upon earlier.

The power of these techniques relies on a combination of approaches. Classical genetics (often termed *forward* genetics) starts with a phenotype (whether a complex disease state or a simple trait such as eye colour) and works towards an identification of the gene(s) responsible. In contrast, the molecular approach offers the possibility of starting with differences in a defined DNA sequence, and investigating the phenotypic effects of such a change: This is often called *reverse* genetics.

9.4.1 Linkage analysis

The classical approach towards identifying a gene associated with a specific trait is based on the concept of linkage analysis, as introduced in Chapter 8, which enables the position of genetic markers to be mapped to specific regions of one of the chromosomes. Before the advent of molecular methods, the construction of a genetic map in model organisms (especially *Drosophila* and mice) was carried out by painstaking hard work, involving large numbers of controlled crosses between strains with different genetic markers. The extent to which two different genetic markers are inherited together (co-segregate), or in other words the degree of linkage between the markers, provides an estimate of their relative position. Genes that are close to one another will tend to be inherited together, and display *linkage disequilibrium*. In this way, a genetic map (or linkage map) can be gradually built up, showing the relative position of the available genetic markers. The position of specific disease-associated genes can be located by the degree of linkage to other markers. Of course linkage analysis in humans is much more difficult, as controlled crossing is not an option; in this case we must rely on the use of extended family trees where possible, or even the use of large numbers of unrelated individuals who all share the same trait.

This can be supplemented with another form of mapping, known as *cytogenetic mapping*. Staining of eukaryotic chromosomes with a dye such as Giemsa stain produces a characteristic banding pattern, as some parts of

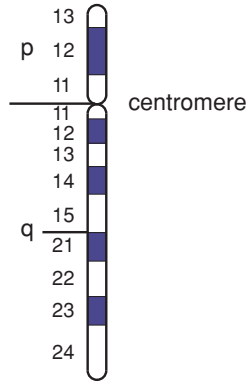


Figure 9.8 Chromosome banding patterns. A simplified representation of the banding pattern of human chromosome 12.

the chromosome take up the stain much more heavily than others (see Figure 9.8). A major change such as a translocation will create a change in the banding pattern, which can be detected microscopically. However, this is an imprecise method of mapping, with a resolution of several million base pairs.

One limitation of conventional genetic linkage analysis is the need for markers that have an identifiable phenotype. Since there are a limited number of such markers, even in well-studied organisms, the linkage map will be relatively crude. However, we can use molecular markers, such as RFLPs and microsatellites (see earlier in this chapter), to locate a disease-associated gene more precisely, by mapping its linkage to a specific polymorphism. This approach can be applied to human genetics by examining very large pedigrees from which DNA samples are available, rather than by making controlled crosses. Nowadays, following the determination of the human genome sequence, *single-nucleotide polymorphisms* (SNPs) are the most commonly used molecular markers.

These forms of gene mapping simply show you the chromosomal location of a gene that is associated with a trait. They do not identify the gene, its product or its function, nor (importantly) do they prove a causal relationship between any gene, or its products, and the trait under study. To gain such information, and to begin to examine if the gene has a causal relationship to the trait, the next step is to clone the gene – and such gene mapping approaches provide a route for such cloning to be undertaken. It is important to realize that we may not have any information about the function of this gene, beyond knowing that there is a gene in which mutations result in a certain disease, so none of the approaches outlined in Chapter 3 for identifying genes in a library are available to us. But if we have mapped its position well enough, so

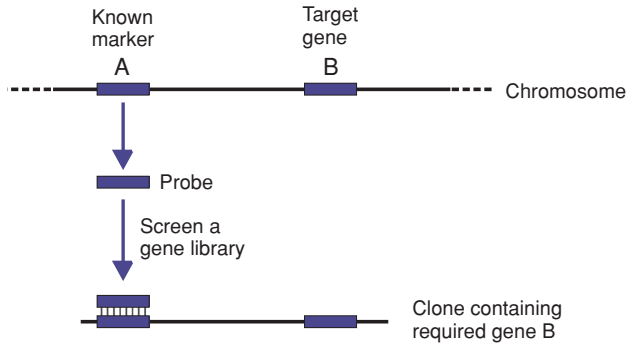


Figure 9.9 Positional cloning.

that we know it is close to a defined marker (whether this is another gene, a polymorphism or a microsatellite), then we can use a probe for that locus to screen a large insert gene library (using, e.g., YAC vectors) to obtain a clone that carries the defined marker and a substantial amount of adjacent DNA, which we hope contains the disease gene. Since this form of cloning relies solely on the position of the gene, it is known as *positional cloning* (Figure 9.9).

The distance that can be covered by this approach can be extended by *chromosome walking* (see Chapter 8). In this technique, the first clone, carrying the known marker, is used to probe the gene library again, so as to identify clones carrying adjacent sequences; these are in turn used as probes to obtain the next region of the chromosome, and so on. When your 'walk' reaches the region where you expect the gene to be, analysis of the sequence by the methods described in Chapter 5 will enable the identification of genes and their putative products.

An alternative approach is to base the search on gene products (mRNA or proteins) rather than on the DNA itself. This rests on the assumption that expression of the relevant gene will be altered in the diseased state, or that the structure of the product will be different. Both of these assumptions can be used – either separately or in combination – to devise protocols for the identification of candidate genes. The most important of these protocols are described in Chapter 10.

The final identification of the relevant gene, and examination of its causal relationship with the trait, will usually rely on a combination of techniques. The putative genes identified by positional cloning and chromosome walking can be tested for their association with the disease state by specifically testing the expression of those genes in the diseased tissue, or by analysing the structure of the protein. In experimental animals, final confirmation will come

from a combination of gene replacement and complementation experiments (see Chapter 11).

Sometimes one or more candidates are obvious without any further experiments being necessary. For example, no reverse genetics was needed to conclude that mutations in the red and green cone photopigment genes were candidates for the cause of red/green colour blindness. Nor was any forward genetic analysis needed – the linkage of these conditions to the X chromosome is well established from observations of their pattern of inheritance. Thus, the mapping of these two *a priori* candidate genes to the X chromosome constituted strong supporting evidence for the initial hypothesis, although it is of course important to remember that experimental testing is always required as proof that the candidate genes are directly responsible, however obvious it may seem. In this way, forward and reverse genetics interact – the chromosomal location and the expression pattern are both important parts of the jigsaw puzzle.

9.4.2 Genome-wide association studies (GWAS)

However, the scenario where obvious candidate genes present themselves is actually rather rare. Fortunately, modern methodologies make it possible to find the genetic background for an altered phenotype through a procedure that entirely eliminates the need to hypothesize about a candidate gene. *Association studies*, made feasible by the development of DNA chip microarrays, enable investigators to compare wild-type individuals or those with a disease or another trait with respect to a series of markers evenly interspersed throughout the genome. For each marker, a LOD score (logarithm of odds) is calculated, which describes the tendency of the marker to be inherited together with the trait. As two chromosomal locations are likely to be inherited together if they are physically close to each other on the chromosome, then a high LOD score suggests that a gene associated with your disease is located close to the marker being investigated. In common practice, a LOD score greater than 3, representing a 1000:1 chance that there is linkage between the marker and the disease trait, is used as a threshold. If the markers used are spaced closely enough, then it will be possible to narrow down on a defined region of the chromosome containing only a few genes, or even one gene, which can then be further analysed to determine its potential role in producing the phenotype under investigation.

A central piece of evidence, however, is quite naturally the sequencing of the allele from affected individuals and the confirmation that it is different from the wildtype. This is not, however, sufficient on its own, as this merely shows an association and not the all-important causal link; to prove this link we must undertake several further steps. The mutation must first be

analysed, using methods that we have described in Chapter 5. Is it located where it could be predicted to have a biological effect? This would often involve the alteration of the encoded amino acid sequence in a subtle or more drastic way. However, even if the coding region is completely unaltered from the wildtype, mutations in the promoter region that affect the expression of the gene, or mutations in splice sites or insertions that otherwise affect mRNA processing, could still cause an altered phenotype. But even if you did show the alteration to have a biological impact, the scientific community would normally demand even more evidence before your discovery could be regarded as causal for the trait. Some polymorphisms, such as amino acid substitutions, occur completely normally in the population without causing any abnormality. One way to show that the mutation causes disease is to express the mutated protein and compare its biological properties to its wild-type counterpart. Another one, preferably used in combination, is of course to show that the mutation that you have discovered co-segregates with the trait to a significant degree in a genetic pedigree.

In practice, things are often not that straightforward. Such studies have a good chance of success if there are only a few genes that determine a specific trait. But many traits involve a large number of genes, each of which has a very small effect on its own. In addition, genome-wide association studies (GWAS) only consider variations (SNPs) that occur commonly (>1%), with rarer variations not being detected. One such example is height. There is a strong inherited component to height, with the height of the parents being 80–90% predictive of the height of their offspring. Genome-wide association studies have identified 200 loci that are associated with height, but these only explain about 16% of the inherited component. More sophisticated statistical analyses, looking at the combination of these loci, have resulted in claims to explain up to half of the heritability of height. The remaining components are presumed to be either less common SNPs, or epigenetic factors.

One of the limitations of GWAS using SNP arrays arises from the focus on the more common SNPs. Because of their prevalence, we can deduce that they must have evolved a long time ago. If they are linked to a trait that is in any way deleterious, then selection would have weeded them out. Consistent with this concept is the observation that GWAS tends to work better for late-onset disease, such as Alzheimer's, than for conditions such as autism that occur at a younger age. There is therefore a case for including rarer SNPS. But not only does that mean using much larger arrays, but also, because the SNPs occur less often, you need to increase the size of the population studied to get enough examples for your statistical analysis. One of the largest studies involved testing 100 000 individuals for the presence of 2 million SNPs – and identified 95 loci associated with variation in cholesterol and triglyceride levels.

9.4.3 Database resources

Of course the availability of the human genome sequence has had an enormous impact on the ease with which genes associated with disease can be identified. There are a number of databases that can be used in this context (see Box 5.1 for a list of on-line resources with web addresses). In particular, the dbSNP database contains information about short insertion and deletion polymorphisms as well as single nucleotide substitutions; HapMap describes the common patterns of human DNA sequence variation; and OMIM (Online Mendelian Inheritance in Man) is a catalogue of human genes and genetic disorders.

9.4.4 Genetic diagnosis

The identification of polymorphisms that have a causal relationship to human genetic diseases, either directly causing the disease or leading to a significant increase in the predisposition towards developing the disease, not only leads to significant insights into the nature of those diseases but also provides a mechanism for molecular diagnosis. This is of particular benefit in prenatal diagnosis, which involves obtaining samples containing fetal cells by amniocentesis or chorionic villus sampling – the latter technique having the advantage of being able to be carried out at an earlier stage of pregnancy, which is important if abortion is considered as an option. More recently, tests have been developed that are based on detecting fetal DNA in a sample of blood from the mother, which is much simpler and safer.

The increasing use of *in vitro* fertilization makes other options available, since it is possible to remove a single cell (blastomere) at the eight-cell stage, by micromanipulation; this cell can then be tested (preimplantation genetic diagnosis). The remaining seven cells continue to develop normally and can subsequently be implanted in the uterus, if the genetic tests are satisfactory. The gene(s) to be tested can be amplified by PCR and specific polymorphisms detected by the use of a pair of labelled oligonucleotides, one of which will hybridize to the normal sequence and the other being specific for the mutant (see earlier in this chapter). Alternatively, real-time PCR procedures (see Chapter 4) can be devised that will differentiate the two versions of the sequence directly, producing much quicker results.

A further advantage of molecular techniques for detecting mutations is the ability to detect the presence of the defective gene in heterozygous carriers. If the mutation is recessive, the heterozygotes will show no symptoms of the condition, but a proportion of the offspring of two heterozygous parents will have the disease. In the past, genetic counselling of prospective parents from families with a history of a genetic disease has relied on statistical probabilities to assess the likelihood that both parents carry the affected gene. With

molecular techniques, it is possible to be certain, one way or the other, about whether a fetus is carrying a genetic variant that is bound to cause severe disease. The situation becomes much more complicated and controversial with respect to the multiple genetic variants that are known to be associated with an increased predisposition to develop a disease. In these more complicated cases, our lack of complete understanding of how all these predisposing factors influence the total risk of developing a disease means that any advice given will still be one of statistical probabilities, rather than a simple yes or no; however, it is still far better to have some information to help reach a decision than none at all.

Similar considerations apply to genes that only show their effects later in life, and to those genes that only cause a predisposition to a specific disease. For example, approximately 5–10% of breast cancers are familial (i.e., they have a genetic component). In many cases this has been traced to a specific gene, *BRCA1,* with individuals carrying a specific variant of the gene having a higher risk of developing breast cancer, although the presence of the affected gene does *not* mean that it is certain that they will develop the disease. Under these circumstances it is prudent to increase the frequency of medical examinations for these individuals, as this increases the chance of identifying any breast tumour as early as possible, leading to an improved prognosis for the affected individual.

There are substantial ethical issues raised by the availability of these techniques. Even if one accepts abortion as an option under some circumstances, its use as a method to prevent the birth of a 'defective' child raises the uncomfortable spectre of eugenics. Furthermore, the ability to test people for the presence of genes that may affect their subsequent health or lifespan has serious implications for their ability to obtain life or medical insurance, subject to legal restrictions that may limit the powers of insurers in this respect.

9.5 Molecular phylogeny

Traditionally, systematics – the classification of organisms into species, genera and higher groupings – relied on phenotypic properties, such as physiology and anatomy for higher organisms, and biochemical characteristics for bacteria. Once the concept of evolution had been established, scientists began relating identified similarities in traits, or *homologies*, to degrees of evolutionary relatedness. Homology denotes any character that has been derived from a common ancestor. A classic example of homology is the vertebrate forelimb – the human arm, the front leg of a horse, the wing of a bird or a bat, or the flippers of a seal or a turtle. By contrast, the wings of a bird and the wings of a butterfly perform the same function in the animal, but have evolved entirely independently from a distant wingless ancestor by the process of *convergent evolution* – they are not homologous but *analogous*.

Once the role of DNA as the genetic material was established, in the middle of the last century, it became apparent that molecular methods, based on DNA similarity, could be used to supplement the phenotypic approach. The advent of the ability to sequence proteins and genes caused a revolution in systematics. And since these species, genera and other groups have arisen by the accumulation of gradual changes as a result of evolution, a study of the current similarities and differences between different genes in different organisms can be used to infer the order and time of divergence of both species and genes. This is molecular phylogeny.

There are two different types of homology between different genes which are descended from one ancestral gene in one ancestral species. *Orthology* denotes two genes in two different species where they have the same function and, typically, the same name. *Orthologous* genes are descended from a gene with the same function in a species that is ancestral to both current species – e.g., the sex-determining gene (*Sry*) in humans and mice. The other alternative is *paralogy*. Here, we are dealing with two genes in the *same* species that are derived from one gene in one ancestral species that was duplicated, and where the daughter genes started to take on different functions. An example of *paralogues* are the green-sensitive and red-sensitive cone opsin genes in humans (*OPN1MW* and *OPN2LW*), which arose through the duplication of a single ancestral gene in the common ancestor of all Old World primates. In molecular phylogeny, we will often simultaneously compare orthologues and paralogues.

For evolutionary analysis, establishing the order in which species diverged from one another, we do need to be confident that the rate of accumulation of variations is similar for each species over time (i.e., there is a reasonably constant *molecular clock*). This is not necessarily true. We can illustrate this by reference to Figure 9.10. Here we can see that A and B are more similar

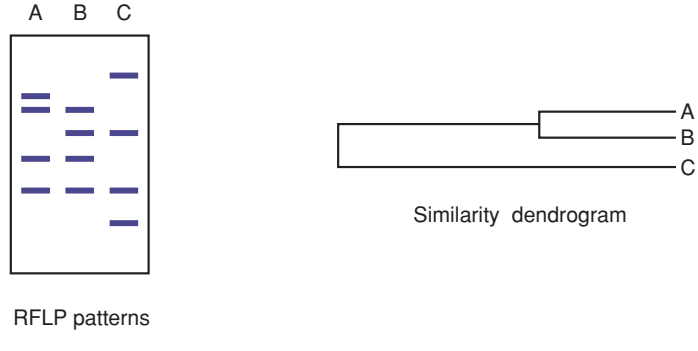


Figure 9.10 Analysis of similarity of restriction fragment length polymorphism (RFLP) patterns. Based on the RFLP patterns, A and B are more similar to one another than either is to C. The dendrogram is a way of displaying these relationships visually. However, it does not necessarily represent the true evolutionary history of these organisms.

to one another than either is to C, and a computer program will generate a dendrogram to display that relationship visually. However, unless we know that the rate of variation is constant, we should not jump to the conclusion that C diverged from a common ancestor of A and B, and that the difference between A and B happened more recently. If for some reason variation has occurred more rapidly in C (and its ancestors) than in A and B, then C might have diverged more recently, but has undergone more rapid variation since. To make matters more complex, we are now aware that different gene families may evolve at different rates, meaning that an average rate of variation is not always applicable. For example, genes that encode proteins involved in the body's defence from external chemicals tend to evolve at a faster rate than average, and this reflects the need for rapid change within this system to meet the challenges of new environments, while genes that encode proteins vital for life tend to evolve more slowly. Thus, to construct accurate phylogenetic trees we must take into account not only the average rate of variation, but also those cases where the rate is significantly faster or slower, and we will discuss this in more detail in the next section.

Comparison of DNA or protein sequences provides the most reliable method of establishing phylogenetic relationships. In Chapter 5, we looked at a simple example, based on comparison of the sequence for a single protein (thymidylate synthetase) from a variety of organisms. There are a number of potential problems with such an approach. If the protein is an essential one, it may be highly conserved, so you would get less variation than expected. This would not matter much if the effect was the same in all organisms, but it might not be. Some organisms could be more tolerant of variation, for example, because of other differences in their physiology. More seriously, you could find that proteins with the same function have evolved from different ancestral proteins – *convergent evolution*. You may therefore find that you get different phylogenetic trees depending on your choice of protein. A common approach to counter these problems is to use a number of protein sequences for the comparison. You do not need to compare them individually – you can join the sequences together to produce a single artificial protein sequence. However, you are still left with the objection that by examining protein sequences you are ignoring all the non-coding DNA, and you are also ignoring any synonymous changes in the DNA sequence (i.e., changes that replace one codon with another one coding for the same amino acid).

The way to circumvent this is to base your analysis on the DNA sequence itself. This could be selected fragments of the genome, coding or non-coding, or the complete genome sequences, provided that you can identify the regions that align with one another. One advantage is that this can be applied to incomplete genome sequence data, without having to identify the coding sequences.

It is worth noting that the choice between protein and DNA sequence comparisons depends on a number of factors and judgments. A deterministic view would regard evolution as occurring solely through natural selection, and therefore only changes in the protein sequences are relevant. The opposing view is that a substantial component of the variation between species occurs by random genetic drift, with neutral changes becoming fixed because of limitations in the population size. These variations will only be detected at the DNA level. In practice, evolution is a mixture of both effects. Often, DNA sequence comparisons are preferred when dealing with fairly shallow phylogenies – that is, where the sequence difference is quite limited, and the protein sequence is therefore less informative.

9.5.1 Methods for constructing trees

There are many methods available for inferring a phylogenetic tree from sequence data, none of them perfect, and all are quite complex, so a full treatment is beyond the scope of this book. But it might be helpful to introduce some of the basic concepts.

As an example, we can start with four short sequences (Figure 9.11). The first step is to count the number of base differences between each pair of sequences, and construct a difference matrix, as shown in the figure. We immediately come up against two problems. Firstly, the sequences to be compared may not be the same length, so we should really divide each count by the length of the sequence to get a true measure of dissimilarity. In this case, the sequences are the same length, so it is clearer to use the actual number of differences. The second problem is less obvious. As the sequences diverge, an A may change to a G, for example; as they diverge further, there may be a second mutation that changes the G back to an A. A simple count of the number of differences will therefore underestimate the true evolutionary difference. A statistical correction can be applied to account for this, but in this case we have left it as it is, to keep it simple.

Essentially, the methods employed are either *algorithmic* or *tree-searching*. The former approach utilizes specific calculations, as the name implies, tends to produce a result rapidly, and produces only a single, unambiguous output. In comparison, tree-searching methods construct multiple possible phylogenetic trees, and then use chosen criteria to select between them.

Clustering methods, such as UPGMA (Unweighted Pair Group Method with Arithmetic means), have often been used to develop trees from a similarity matrix. To understand a similarity matrix, we must first introduce the concept of Operational Taxonomic Units (OTUs). This simply means the sequences, or groups of related sequences, that we are comparing at each point. In the example in Figure 9.11 we start with four OTUs, and the aim is to combine these sequentially until all the branches in the tree have been defined.

Four DNA sequences for comparison

A:	C	A	T	A	G	A	C	C	T	G	A	C	G	C	C	A	G	C	T	C
B:	C	A	T	A	G	A	C	C	C	G	C	C	A	T	G	A	G	C	T	C
C:	C	G	T	A	G	A	C	T	G	G	G	C	G	C	C	A	G	C	T	C
D:	C	C	T	A	G	A	C	G	T	C	G	C	G	G	C	A	G	T	C	C

Distance matrix, showing the number of base differences in each pairwise combination

	A	B	C
B	5		
C	4	7	
D	7	10	7

A and C show the fewest base differences, so they are combined into a single Operational Taxonomic Unit (OTU) and the matrix is recalculated using the average number of differences (e.g., A to B = 5, C to B = 7; mean = 6).

	AC	B
B	6	
D	7	10

The fewest changes are shown between B and the combined OTU AC, so B and AC are combined into one OTU, and the matrix is recalculated:

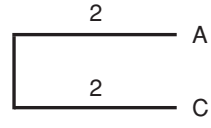
	ACB
D	8.5

These data can now be used to construct a tree – see Figure 9.12

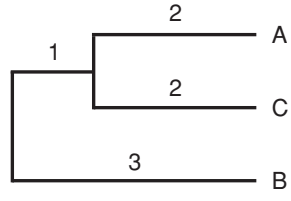
Figure 9.11 Phylogenetic trees based on DNA sequences; sample data and Unweighted Pair Group Method with Arithmetic means (UPGMA) calculation.

To achieve this, we inspect the matrix and identify the pair of sequences (OTUs) that are most similar, which in this case are A and C, with only four differences. (Note that we will come back to this later on, as it is an example of how this method can give rise to an incorrect tree.) We combine A and C into a single OTU (AC), and calculate the average distance from this combined OTU to each of the others. This gives us a new matrix, now with only three OTUs. Again, we combine the two most similar OTUs (which would be B and AC) into one OTU (ACB), and calculate a further matrix, now with only two OTUs. The final step is to calculate the average distance from ACB to D, and we have a tree (Figure 9.12), which appears to show the relationships between these OTUs and the genetic distance separating them. The depth of the divergence of each branch is shown as half the distance separating the OTUs – for example, A and C differ by four bases, so each branch is given a length of two. This is a limitation, as it means that the individual path lengths are not accurately represented. So while the true distance from A to B is five

Using the data in Figure 9.11, A and C are first combined into a single OTU. The four sequence differences are distributed equally between the two branches of the initial tree



Following recalculation of the matrix, AC and B are combined into one OTU. Again, the mean number of differences (6) is divided equally between the two branches. Add the numbers together as you trace the path from A or C to B



Finally, ACB is combined with D into a single OTU

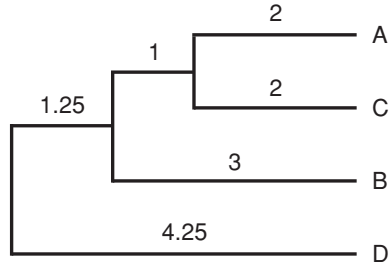


Figure 9.12 Derivation of a phylogenetic tree using Unweighted Pair Group Method with Arithmetic means (UPGMA).

changes, and that from B to C is seven, both lengths are shown on the tree as $(2+1+3)=6$ (the average of 5 and 7).

Furthermore, this method only gives viable results if the rate of change is the same for all branches, i.e., there is a uniform linear molecular clock. As this condition is frequently not met, clustering methods can easily give erroneous trees. They were popular initially because of their relative simplicity and being less demanding on computer power than other methods. The increased computer power now widely available makes this much less of a problem. There are ways of overcoming the limitations of clustering methods, but the availability of superior methods means there is little reason to persist with clustering for phylogenetic inference.

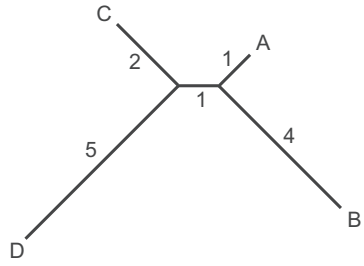


Figure 9.13 An unrooted phylogenetic tree produced by the Neighbour Joining (NJ) method, from the data in Figure 9.11.

One of the better methods, which allows for variation in the clock speed, is the Neighbour Joining (NJ) method. Neighbours are two OTUs that are connected by a single node in the tree – so if you look at Figure 9.13, A and B are neighbours, as are D and C, but A and D are not neighbours. Of course we do not yet know the tree, so there are a number of possible trees, and potential neighbours. The first step is to select a pair of OTUs on the basis that separating that pair from the others will give the smallest sum of all the branch lengths in the hypothetical tree. Those two OTUs are then combined to form a single OTU (or node) and a new distance matrix is computed. This process is repeated until all the branch points have been defined.

With the data in Figure 9.11, the NJ method gives the tree shown in Figure 9.13. This is an *unrooted* tree. It does not imply any direction to evolution, nor what the ancestral organism looked like. The usual way of obtaining a *rooted* tree is to include an *outgroup* in the comparison – i.e., an organism that we know (or can be reasonably confident) diverged from all the others before they diverged from one another. This may be from independent evidence, such as fossil records, or we may infer it from the sequence data itself. In this case, we may be able to take D as an outgroup, since its sequence is so different from each of the others. Using D as an outgroup, we obtain the rooted tree shown in Figure 9.14. Note that this method does not tell us exactly where the root is, except that it is somewhere along the branch joining D to the node where C diverges from A and B.

We can now consider why the UPGMA and NJ trees are different. The reason lies in the length of the branch leading to B, presumably as a consequence of a higher rate of variation in this lineage. As a result, A appears to be more similar to C than to B, and therefore in the clustering (UPGMA) method, A is initially clustered with C. In contrast, the NJ method considers all the path lengths, which show, for example, that B is much closer to A than it is to either C or D. The NJ method therefore produces a much more reliable tree. A final trick used in the construction of phylogenetic trees is the concept of *maximum parsimony*: this is the principle that the simplest

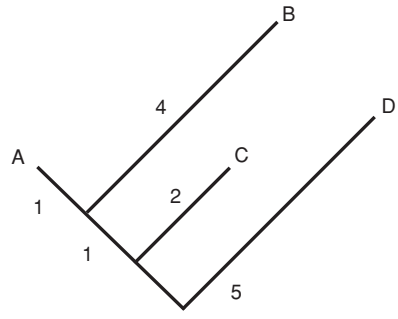


Figure 9.14 A rooted phylogenetic tree, produced with the same data as in Figure 9.13, using D as an outgroup to root the tree. Note that although the root must be along the branch between D and (ABC), its precise position cannot be determined in this analysis.

phylogenetic tree (i.e., the one with the fewest divergence events) that can explain the relationships between our sequences is most likely to reflect the biological scenario. This is referred to as the *maximum parsimonious reconciliation,* or MPR.

A real phylogenetic study would use longer sequences, and would usually include a larger number of sequences. Ideally, we would compare entire genome sequences, and this is indeed now being done, at least for bacteria. In such a case, there are a very large number of possible trees. How confident can we be that the tree we have constructed is the right one? The short answer is that we cannot be certain. We are attempting to infer evolutionary history from an examination of the current state of things, and that is inherently unreliable. But we can obtain a measure of how reliably the tree depicts current relationships. A common procedure for doing this is known as *bootstrap* analysis. Essentially this means taking random samples of the data and repeating the construction of the tree each time. If the tree is reliable, we will get the same tree every time. The extent to which this is true gives us the measure we need of the robustness of the tree. If we knew not only that the clock speed was uniform and constant, but also how fast it ran – i.e., how many base changes occurred in a period of time – then we would be able to estimate how long ago these species diverged.

The MPR phylogenetic tree we have produced provides us with a good overview of the likely relationships between all of the sequences under study. It does not, however, necessarily tell us about when these divergence events (i.e., the nodes of the tree) occurred, or indeed if they represent the formation of paralogues (i.e., a duplication event) or orthologues (i.e., a speciation event). To achieve this, more complex statistical methods, such as probabilistic orthology analysis (POA), are required. In POA, the MPR tree is compared to a tree of estimated species divergence times. Through the simulation of many different variation rates for our sequences, these trees are combined

to produce an output tree that estimates not only where the divergence in sequences occurred, but also when. As we know when the species diverged, it is therefore possible to determine if any particular divergence is most likely to be due to a gene duplication event within one species, or as a result of a speciation event. Not only that, we also get an approximation of the rate of variation for our particular sequences, which we can compare to the average, and hence see if our sequences are under an evolutionary pressure to change rapidly or are relatively stable.

10
Post-Genomic Analysis

In a previous chapter we considered ways in which we can study the expression of individual genes. Now, having dealt with the sequencing of complete genomes, we can move on to look at how we can analyse gene expression across the genome as a whole. These methods, collectively described as post-genomic analysis, include *transcriptomics* (the study of the complete collection of mRNA transcripts in a cell, tissue or organism at a particular time) and *proteomics* (analysis of the complete set of translated proteins)

10.1 Analysing transcription: transcriptomes

In Chapter 6 we described methods such as Northern blotting and qRT-PCR for studying the expression of specific genes. There are a variety of reasons why we might want to study the expression of individual genes, such as the investigation of disease-associated genes once a limited number of strong candidates have been identified, as discussed in Chapter 9. However, in many cases we don't have strong candidate genes, and may want to identify and compare the complete spectra of genes that are expressed under different circumstances. This could include genes that are expressed in one organism but not in another; genes that are only expressed in specific tissues; genes that are differentially expressed under varying environmental conditions; and genes whose expression is altered during disease conditions. Over the years, a number of techniques have been developed for identifying the mRNA species that are expressed in different samples, and (more importantly) for allowing a quantitative comparison of these levels of expression. Such methods are able to answer some very important questions, or at least give us a strong clue. For example, they can tell us which genes are associated with a

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

degenerative disease, by comparing diseased and healthy tissue. They can tell us which genes of a pathogenic bacterium are expressed during infection of a suitable host, as opposed to those that are expressed in laboratory culture. And comparisons of gene expression can indicate which genes are involved in various stages during development of an organism. This is far from an exhaustive list. However, as we saw with association studies in the previous chapter, it is important to remember that identification of a change in expression for a gene from, for example, normal to tumour tissue tells us that the change occurs, but not whether it causes the disease process or is a consequence of it. We need to undertake more targeted experiments to answer such questions, such as the use of transgenic animals (see Chapter 11).

Modern methods rely mainly on the use of microarrays and transcriptome sequencing (see below), but we will first describe some of the older procedures. These still have some potential applications, especially as they can be performed in the absence of genome sequence information.

10.1.1 Differential screening

One approach is to identify cloned genes that are differentially expressed by using *differential screening* of a gene library (Figure 10.1). This is similar to the method for screening a gene library, as described in Chapter 3, except that here we are comparing the effects of screening with two different,

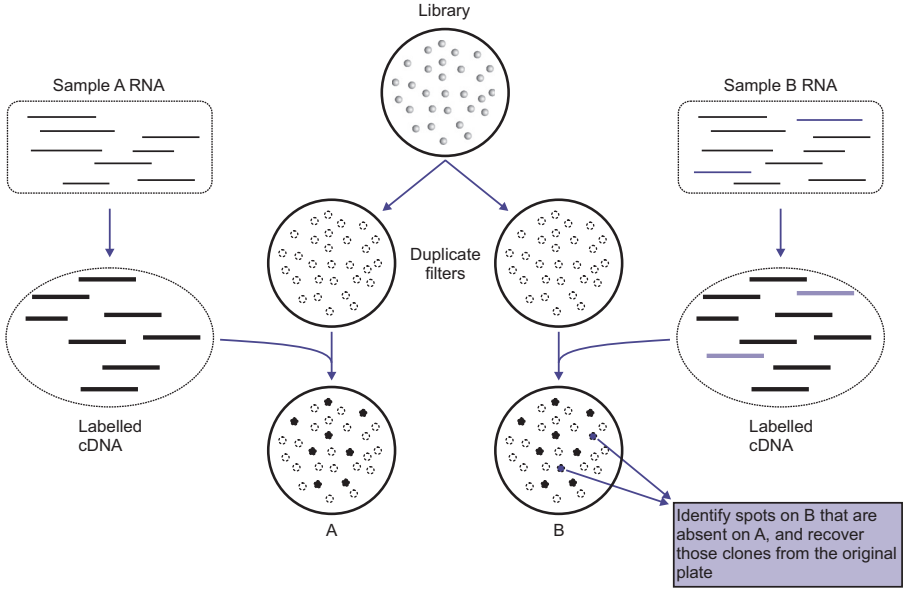


Figure 10.1 Differential screening of a gene library.

mRNA-derived, probes. Starting with the library in the form of colonies or phage plaques on an agar plate, the procedure involves producing two identical filters containing DNA imprints of the gene library. To obtain the probes for screening the library, we extract total mRNA from two samples (e.g., two strains or two tissues), and subject each of these complex mRNA mixtures to reverse-transcription in the presence of labelled nucleotides. One of the filters is hybridized with one of these pools of labelled cDNA, while the second pool is used to probe the second filter. Clones that show a positive signal with one probe but not the other are candidate clones of differentially expressed genes.

This description assumes that the specific transcript is completely absent from one sample (sample A in the figure). But genes are not necessarily turned off absolutely; there may still be a small amount of the specific mRNA present. This might result in a different intensity of some spots with the two probes, and we could use that for our differential screening. But the situation is rather better than that. If there is only a very small amount of a specific mRNA species (a very low abundance), then the amount of the corresponding cDNA in the probe will be too low to produce a visible signal. Furthermore, the synthesis of cDNA is likely to be more efficient for the high-abundance mRNA, leading to even further depression of the signal from the low-abundance species. So any low-abundance mRNA in sample A may give, not just a weak signal, but no signal at all. This actually causes a problem in some cases, in that mRNA species that are less abundant, but are still differentially expressed, may not be detectable in either sample. RT-PCR can be used to generate more representative cDNA probes, but the more representative the probes become, the more difficult it is to pick up differential expression where the ratio between the two sources is not high.

A related procedure, known as *subtractive hybridization*, involves removing from the test probe any sequences that are also present in the control. The result is detection of only those transcripts that are present in the test sample. A further variation, known as differential display, is a PCR-based method of comparing two cDNA populations. One primer is designed to bind to the poly(A) tail, while the other is a random sequence. By running PCR at a very low annealing temperature, a subpopulation of the cDNA molecules in the sample is amplified. This subpopulation will be small enough to be resolved in a polyacrylamide gel. If you perform identical PCR reactions with two cDNA populations, you will be able to identify cDNA species that are expressed at different levels in the two samples, or even completely absent in one of them. The presence of these cDNA species amongst the amplified mixture is a random event governed by the degree of sequence similarity with the primers; the method itself does not favour the amplification of differentially expressed messages. Thus, a considerable number of primer combinations will have to be tried, and even so, the analysis will be far from exhaustive.

10.1.2 Other methods: transposons and reporters

Transposons are mobile genetic elements that are able to move (transpose) from one site to another in the genome, or between plasmids and chromosome. In Chapter 8 we discussed the use of transposons to produce mutations, due to inactivation of a gene into which the transposon has inserted. We can extend the concept to the study of gene expression, by inserting a *reporter* gene (see Chapter 6) into the transposon.

This is especially valuable in bacteria for identification of the range of genes that are expressed under a specific set of conditions. For this purpose, we would first make a *transposon library* (see Chapter 8), containing a large number of clones with transposon inserts at different, random sites. We would then expose the transposon library to the environmental conditions in question, such as anaerobic growth, and identify those clones that show expression of the reporter gene. Cloning, or PCR amplification, of the transposon and its flanking DNA then enables the insertion sites to be identified.

10.2 Array-based methods

The development of methods such as subtractive hybridization allowed us to pan for differentially expressed gene products, rather than merely examining the expression of a specific gene. With the availability of genome sequence data, it became possible to use array technology to address the global question of identifying the complete spectrum of genes that are expressed under certain conditions and not otherwise. This is a much more powerful technique, and much simpler, than those described above.

The basis of the use of *arrays* was described in Chapter 9. However, in that case the arrays were probed with a mixture of genomic DNA fragments, to identify genes that were present in one sample and not another. Now we want to identify which genes have different levels of mRNA transcripts between samples. The only change is that we will hybridize the array with labelled cDNA, rather than genomic DNA. Those spots in the array that correspond to an expressed gene will yield a positive signal. Using machines to produce many identical copies of the array, you can test for the presence of a wide range of specific cDNA molecules derived from cells from different environments. Various types of arrays, and how they are used, are described below.

Conceptually, this is not really different from differential screening. Subtractive cDNA libraries in plasmid vectors have often been painstakingly picked clone by clone and inoculated in a grid formation on an agar plate, and duplicate colony lifts used for their analysis. The advent of robotic picking and high-density spotting technology, and of electronic image analysis, not only made it technically simpler to produce such arrays, but also largely

eliminated the need to limit the array to a subpopulation of particularly suspect transcripts. Furthermore, knowledge of the complete genome sequence of an organism makes it possible to produce arrays of small DNA fragments covering the whole genome, or selected parts of it. If we want an array specifically for testing expression (as opposed to genomic comparisons), we may want to restrict the array to protein-coding regions, either by using cDNA or by identifying ORFs. There are a variety of possible approaches for many organisms, with the commonest being expressed sequence tag (EST) arrays, PCR product arrays and synthetic oligonucleotide arrays.

10.2.1 Expressed sequence tag (EST) arrays

If you want a cDNA array, the simplest method would be to plate an unamplified cDNA library, say from the pancreas, and pick the clones one by one. This has distinct disadvantages, however. Firstly, the use of your array will be restricted to samples from one tissue. Your array will have limited use with samples from brain tissue, for example, because the brain and the pancreas will each have their specific repertoires of genes that are expressed in the respective tissue but not in both. Since you will want to use your array to test a variety of tissues, you want to make it as widely applicable as possible.

Secondly, when you made your library, you were wise to base it on poly(A)-enriched mRNA (if you were working with a eukaryote), thus avoiding a library where 95% of the material consists of endlessly repeated rRNA and tRNA sequences. However, even so you will have quite a lot of repetition. This is because the frequency of any given gene in a cDNA library (unlike a genomic library) is biased according to the levels of expression of that particular gene. Thus, you will find major components such as structural genes (e.g., tubulin), housekeeping genes (metabolic enzymes) and major tissue-specific products (e.g., insulin) greatly over-represented in your library.

The use of *expressed sequence tags* (ESTs) simultaneously helped to solve the problems of tissue-specific libraries and the problem of over-representation. ESTs are cDNA clones obtained by a laboratory equipped with a robot that, using the method that we outlined above, has picked every clone from a library. This has been done with libraries from many different tissues, creating a sort of transcriptome fingerprint for each one of them. They have all been subjected to one sequencing run, yielding a few hundred (unconfirmed) bases from one end, numbered, and archived jointly in large public facilities. These ESTs are so named because they are derived from mRNA (i.e., they must have been *expressed*), they are known nucleotide *sequences*, and they are not the complete sequence for a transcript, only a *tag*. A database is available through GenBank containing all of these ESTs, so that you can search it for similarity to a fully characterized gene you are interested in, and obtain the clones for a nominal charge.

Systematic computer comparison has made it possible to identify which cDNA clones are unique. It is thus possible to produce cDNA arrays with non-redundant clones. By spotting these onto a membrane (macroarray) or a glass slide (microarray), resources have been created that enable simultaneous study of the majority of the genes in the human, mouse and rat genomes, independently of the sequencing of these genomes. A virtually unlimited number of copies of each array can be produced at a marginal cost. Screening a pair of identical arrays with cDNA populations from two different tissues will give you, in one experiment, a complete survey of all the significant differences in gene expression.

10.2.2 PCR product arrays

One objective of the EST programme was to try to obtain at least one clone representing each gene, and indeed from every product from every gene. Although, in the case of the human genome, this objective was not quite achieved, ESTs have been immensely helpful in the task of identifying those genes that were represented, and their open reading frames. Some genes, however, lack EST correlates, often because the mRNA transcripts are either very rare or extremely labile. In addition, even if we have an EST from a gene, the method of spotting bacterial clones carrying plasmids is far from ideal, meaning that further ESTs may be missed. Finally, some eukaryotic probes will have high enough sequence identity to the bacterial sequences to bind non-specifically to the array, reducing their usefulness. Because of these confounding factors, EST-based arrays do not present 100% coverage, meaning that other approaches must be used.

In bacteria, by contrast, ESTs are not available, meaning that we cannot even begin to make an EST-based array. However, we now know the entire genome sequence for very many bacteria, making possible even more efficient approaches for array production (Figure 10.2). If we want our array to focus on protein-coding regions, we can use a computer to identify all the open reading frames above a certain size; these are the predicted genes. Alternatively, if we want to make sure we pick up all transcripts, including RNA that does not code for proteins, we can simply select regions spaced at an arbitrary distance apart along the genome sequence. For each selected region, we use the computer to design a pair of primers that will amplify a suitably sized part of that region. Automated DNA synthesizers are used to make each of these primers, and robotic methods are used to carry out the thousands of PCR reactions needed to make the DNA fragments, which are then robotically spotted onto microarrays.

Figure 10.2 shows an example of how such an array could be used. In this example, we want to compare the bacterial genes that are expressed under aerobic and anaerobic conditions. From bacteria grown under the different

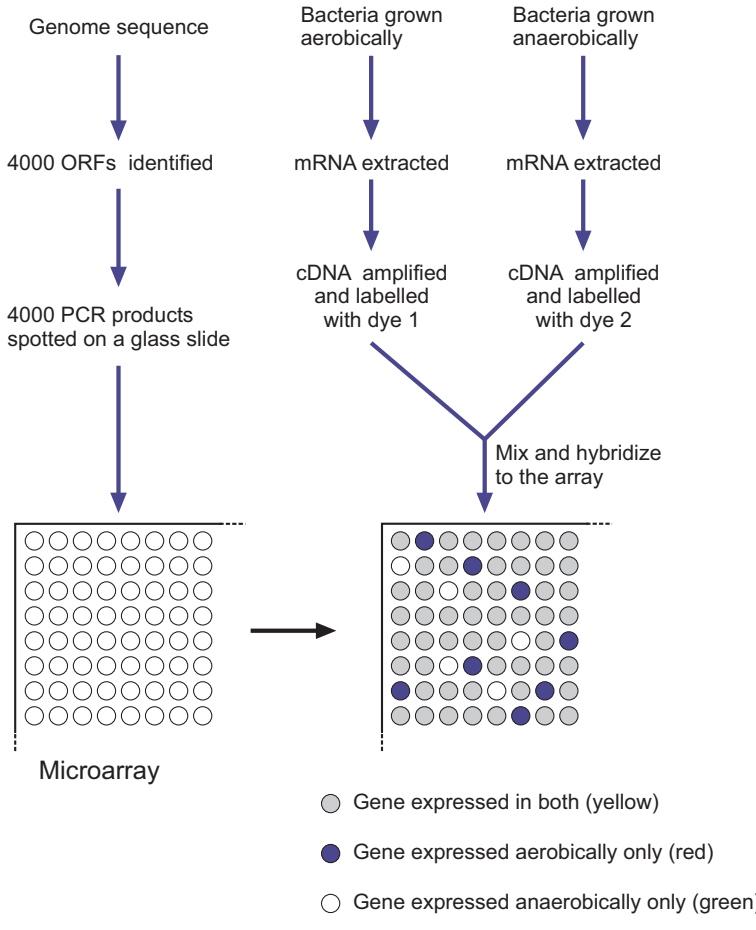


Figure 10.2 Transcriptomic comparison using a microarray.

conditions, mRNA extracts are made, and reverse transcribed to produce a mixture of cDNA molecules. These are amplified by PCR and labelled with different fluorescent dyes. The two probes are then mixed and hybridized to the microarray. A microarray reader is used to measure the different fluorescence of the two dyes, for each spot on the microarray. The results are presented visually in different colours. If the probe from the aerobically grown bacteria, labelled with dye 1, hybridizes preferentially (i.e., that gene is expressed at a higher level in aerobic culture), you will see a red spot. Conversely, if the gene is expressed more when grown anaerobically, the spot will be green, because there is more hybridization with the probe labelled with dye 2. If both probes hybridize equally (i.e., there is no difference in expression of that gene) you will see a yellow spot. For a reliable

quantitative analysis, more sophisticated techniques are required (see Section 10.2.4 below).

Mammalian genomes are much larger, of course, but the availability of complete genome sequences makes it possible to produce a complete array of PCR-amplified products in the same way. However, the larger number of genes involved makes it desirable to use even higher density arrays (see below).

10.2.3 Synthetic oligonucleotide arrays

Microarrays, as described above, can equally well be produced with a set of oligonucleotides, synthesized in the conventional manner and then attached to the glass slide. Such oligonucleotide microarrays are used in the same way as the PCR product arrays.

However, oligonucleotide arrays can also be made, at a much higher density, by synthesizing the oligonucleotides *in situ*, using photolithographic techniques analogous to those used in the manufacture of computer chips. For this reason, such arrays are also known as *gene chips*. (The term is sometimes used for microarrays in general, but it should be restricted to those produced by this method; the term GeneChip is a trademark of the company that produces them, Affymetrix. Hence these arrays are also known as Affymetrix arrays.)

The procedure, as illustrated in simplified form in Figure 10.3, involves the sequential addition of nucleotides to each position in the array. The use of photolabile blocking groups prevents the addition of nucleotides to positions where they are not wanted. The blocking group is destroyed by the action of a laser to activate selected positions.

Initially, after attachment of linker molecules with a protective blocking group, to the substrate, a mask is positioned over the array, so that the laser will only be allowed to reach the positions where the first nucleotide is to be attached. After the selected positions have been activated, the first nucleotide, in this case G (with a blocking group), is attached. A second mask is then positioned, to allow activation of positions where the next nucleotide is an A, followed by laser exposure, attachment of A residues (with a blocking group) – and so on. Such an array can contain millions of oligonucleotides on a single chip, and they are therefore of great benefit for screening organisms with larger genomes – either for expression studies, or for screening for nucleotide polymorphisms (see Chapter 9).

The principle of the use of a gene chip is similar to that described for microarrays in general, involving hybridization with fluorescently labelled probes.

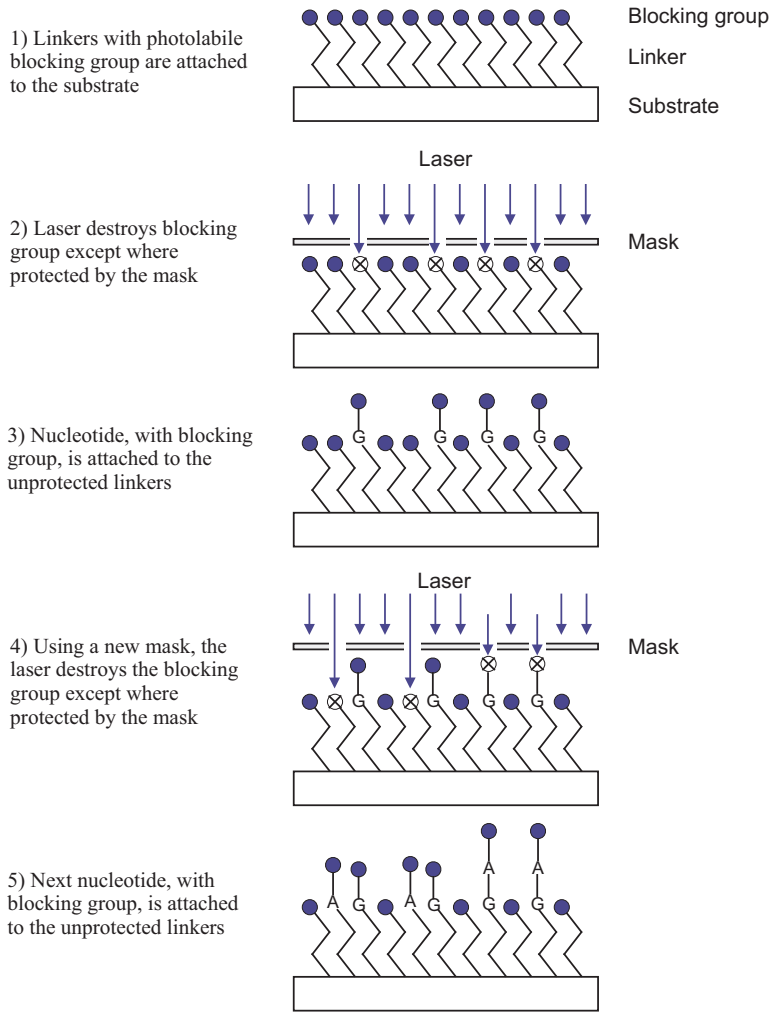


Figure 10.3 Synthesis of oligonucleotides in an array.

10.2.4 Important factors in array hybridization

It should be noted that array hybridization is in effect the reverse of conventional hybridization. The conventional techniques involve attaching the target DNA or RNA to a filter (e.g., by Southern or Northern blots, or using dot blots), and then hybridizing that filter with a specific labelled probe. With macroarrays or microarrays, the specific fragments are attached, unlabelled, to the solid phase, and they are hybridized with the labelled target molecule.

The production of arrays requires a high initial investment. Not only does it require expensive equipment, but also the synthesis of the large number

of PCR primers, and the carrying out of large numbers of PCR reactions, is costly. This is especially true for Affymetrix gene chips, which are only practicable where a substantial level of use of a specific chip is envisaged. However, once you (or someone else) has got to that stage, large numbers of arrays can be made, and the subsequent hybridization and reading of the results is much less expensive.

Although the visual display of red, green and yellow spots (often called a *heatmap*) is attractive, the real comparison demands a quantitative approach. The microarray reader actually compares the relative levels of the two fluorescent signals and records those data. At a simple level, you might expect that for genes that are expressed to the same extent under both conditions you would get a ratio of 1, and that values greater than or less than 1 would indicate differential expression. But what represents a change in expression level? At the simplest level, a cut-off value is commonly employed, taking, for example, values greater than 2, or less than 0.5, as indicating a difference in expression. Whereas this approach is attractive for its simplicity, it does have the disadvantage that you are making an arbitrary decision for the cut-off level; for example, why is a two-fold change important? Indeed, for many key regulators in biology, smaller changes may have biologically significant effects. A more reliable approach is to use a microarray that contains a number of spots representing each gene (randomly arranged on the array), and/or do the whole experiment several times, and then use statistical analyses to estimate whether any differences are statistically significant. Note, however, that a simple statistical test, as you would usually use, would be inadequate here due to the large number of comparisons to be made. For example, if you choose to regard $p < 0.01$ as indicating significance, this means that the event is expected to occur by chance 1 in 100 times. If you are looking at 4000 genes, you would expect to get 40 of them showing such a difference, even if the two samples are identical, meaning that these are *false positives*; on the flip-side an equal number of genes would not come up as significantly different even if they were (*false negatives*). To get away from these potential problems, more complex analyses, such as SAM (statistical analysis of microarrays) can be used, which allows you to see what both the potential false positive and negative rates are for any particular analysis. Whereas this approach does not remove the false positives and false negatives, it does at least allow you to decide what level of uncertainty you are willing to accept in your experiment

Furthermore, there are a host of assumptions behind the microarray procedure. Most importantly, it assumes that every step in the process (mRNA extraction, cDNA production, labelling, hybridization) is equally efficient for each gene in both samples. This is unlikely to be true, and carrying out repeat experiments will only eliminate random bias, not systematic errors. For example, if a specific gene does not amplify well in one sample (perhaps because of competition with another gene), then no amount of repetition will

solve the problem. You also need to remember that, as with most of these techniques, what you are measuring is the *amount* of each mRNA, not the level of transcription of that gene. The amount of each mRNA will be influenced by its stability as well as its rate of production.

Despite all of these potential confounders, microarray experiments such as these have generated an enormous amount of extremely valuable data. Although they are subject to unreliability and artefacts, they do provide a rapid screening method that can identify a large number of candidate genes. Indeed, a lot of work is being undertaken to help minimize these artefacts, with international collaborations examining the best technology platforms and analysis methods to ensure robust results whenever an experiment is undertaken. However, it is important not to regard microarrays as definitive, but rather as the first step in a scientific journey. The data need to be verified by more robustly quantifiable techniques, such as quantitative RT-PCR (see Chapter 6), applied to each of the candidate genes individually. In addition, even once these candidate genes have been shown to have their expression altered, it is important to undertake further experiments to show the biological significance of these changes.

It is worth reiterating that these methods only provide information, at best, about the *levels* of the relevant mRNA species. This is a function of the rate of transcription of that gene and of the stability of the message. So a poorly transcribed but highly stable mRNA may be present at a higher level than that of a more strongly expressed gene where the mRNA is very unstable. The stability of the mRNA is an especially important parameter if you are considering conditions under which specific genes are turned on or off. It is also important to remember the next step of gene expression – translation of the message into protein. The level of mRNA is not necessarily a foolproof predictor of the amount of protein that is being made. In a later section we will look at *proteomics,* which is the study of the spectrum of proteins that are produced by the organism.

10.3 Transcriptome sequencing

Although the use of microarrays in surveys of transcriptional activity represented a great step forward, it is already becoming superseded, as a consequence of the advent of rapid sequencing methods (see Chapter 8). These methods now make it possible to sequence all the transcripts present in a given cell or tissue in a relatively short time period. This is referred to, interchangeably, as transcriptome sequencing, RNA-seq, or whole transcriptome shotgun sequencing (WTSS).

This technique involves using RT-PCR with random primers to produce a complete cDNA library representing all the RNA present in the sample. Each of these thousands of cDNAs is then sequenced, using the massively

parallel sequencing methods described in Chapter 8, and a computer will automatically check each sequence against the known genome sequence to identify the transcribed regions. This alignment is carried out in much the same way as we saw for simple sequence alignments in Chapter 5, except that this time we have the additional complication of intronic sequences to deal with. The cDNA that we have sequenced will contain only exonic information, whereas the genomic sequence we are aligning to will contain both exons and introns, which could obviously cause some problems with achieving a good alignment. To get around this potential problem, a 'proxy genome' is created, which contains only exonic information, and the computer uses special alignment algorithms that are optimised for aligning the short reads produced from RNA-seq to these proxy genomes.

RNA-seq has an additional advantage over traditional microarray technologies for looking at transcriptomes. As the transcriptome is shotgun sequenced (i.e., randomly), this means that novel splice junctions and transcriptional units can be identified, which cannot be done with array data as you only detect what is on the array to start with. Where array-based technologies have reigned supreme for a number of years is their ability to produce *quantitative* data on gene expression; they tell you not only which genes are being transcribed, but to what extent. While RNA-seq was originally a purely qualitative technique, recent developments mean that it is now quantitative in the same manner as arrays. In the case of RNA-seq this is achieved by the simple assumption that if there are more transcripts present for an individual gene product, then this will produce more cDNA molecules and, hence, more sequence reads. Experimental work has proven this assumption to be correct, with RNA-seq quantitation being linear over five orders of magnitude.

10.4 Translational analysis: proteomics

As with the analysis of transcripts, the study of the levels of proteins in different samples can be divided into those methods that are applicable to specific proteins, such as Western blots (Chapter 6), and other methods that are intended to demonstrate the full spectrum of proteins that are present; the latter constitute the discipline of *proteomics*. It should be emphasized that while the genome of an organism is a fixed entity (give or take a bit of variation), the profile of expressed mRNA transcripts will be different according to factors such as growth conditions and cellular differentiation, as indeed will be the profile of the encoded protein products. An important reason for studying at the level of the proteome is that the activity of many proteins is regulated post-translationally, through the addition of post-translational modifications (PTMs). These modifications include, for example, the proteolytic cleavage of insulin during its activation, or the phosphorylation of proteins to switch their activity on or off; indeed, the study of those proteins that

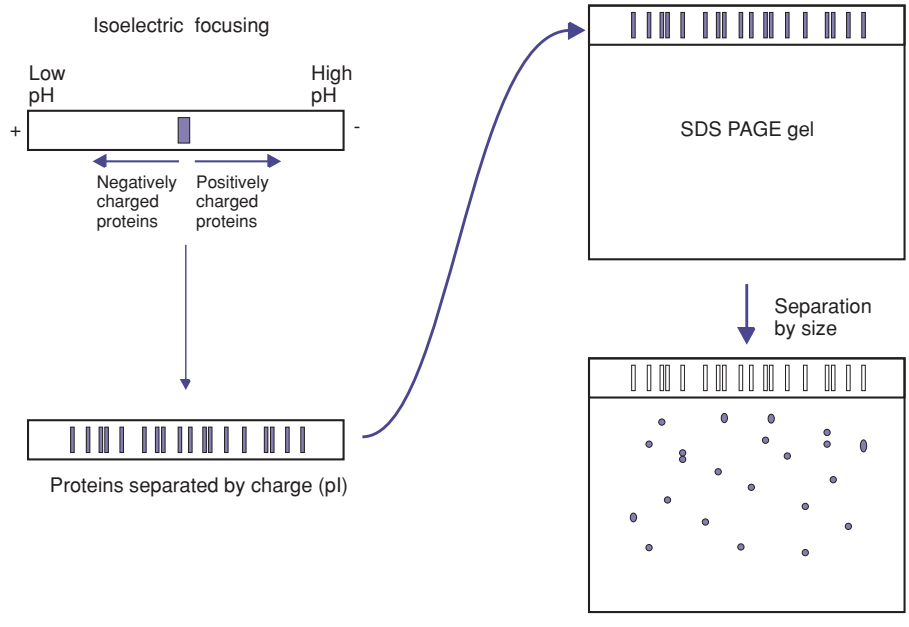


Figure 10.4 Two-dimensional (2D) protein gel electrophoresis.

are (de)phosphorylated in response to a stimulus (the *phosphoproteome*) is a specific subdiscipline within proteomics.

10.4.1 Two-dimensional electrophoresis

The oldest established method of attempting to identify a large number of proteins in a sample is *two-dimensional (2D) gel electrophoresis*, which involves isoelectric focusing (IEF) in one direction and SDS-PAGE in the other (Figure 10.4). Isoelectric focusing consists of applying an electric field across a stable pH gradient. All proteins have a characteristic *isoelectric point* (or pI), which is the pH value at which they have no net negative or positive charge; it is determined by their amino acid composition and the effect of pH on the ionization of basic and acidic amino acids. At pH values above the pI, the protein will have a net negative charge; conversely, below the pI the net charge will be positive. Negatively charged proteins will move through the pH gradient towards the positively charged electrode (anode) until they reach a pH equal to their pI, at which point they are no longer charged. Similarly the positively charged proteins will move towards the cathode until they reach their pI. The end result is that each protein will focus at a characteristic point on the pH gradient, dependent upon their exact amino acid sequence. The

second dimension (SDS-PAGE) separates the proteins according to their size, as discussed in Chapter 6.

After 2D gel electrophoresis, you will get a pattern of spots that can be compared with patterns from other samples to enable the identification of proteins that are differentially expressed between these samples. Web-based resources are available that can help you with the interpretation; see the Proteome Database System for Microbial Research hosted by the Max Planck Institute for Infection Biology (www.mpiib-berlin.mpg.de/2D-PAGE/). However, it can be extremely difficult to obtain reproducible results from 2D gel electrophoresis. It is necessary to control the conditions used very carefully to enable a reliable comparison to be made. Despite this increase in technical complexity, 2D gel electrophoresis has a marked advantage over one-dimensional SDS PAGE with respect to the number of proteins that can be resolved. Whereas, only approximately 100 protein bands can be resolved in a single dimension, the addition of the second dimension (pI focusing), allows several thousand protein spots to be resolved. Note, however, that this is still short of the actual number of proteins that may be present in a cell, which probably runs into the tens of thousands for a mammalian cell.

A second advantage of the use of pI focusing in 2D gel electrophoresis is that, by its very nature, it can separate very closely related proteins, providing that their overall charge is different. Hence, changes such as post-translational modifications – phosphorylation, glycosylation or ubiquitination – may mean that a single protein can occur at different positions in a 2D gel. As previously noted, post-translational modifications, such as phosphorylation, are often used to control the activity of proteins, and this means that the apparent disappearance of a spot, and the appearance of another one in a different place, may indicate a difference in activity of a specific protein, rather than the appearance of a totally new protein. We can therefore use this approach to get an idea of not only the proteins expressed in a cell at any one time, but also how they change, in terms of both *absolute expression* and *biological activity*, in response to a stimulus.

10.4.2 Mass spectrometry

Initially, a 2D gel merely shows us a pattern of spots, some of which may be more interesting than others as their intensity varies according to the nature of the sample. However, all you currently know is that a particular spot varies in intensity, but that does not tell you what protein the spot actually represents. It is possible, using microsequencing techniques, to obtain protein sequences from individual spots. However, this would be a rather cumbersome way of identifying all the spots on a 2D protein gel. Fortunately, if the genome sequence is available, we can use quicker methods. One commonly

used procedure is known as Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry, or MALDI-TOF for short. The first step in this procedure is to recover the protein from the gel by cutting out the individual spot and then soaking it in an elution buffer. Next, the eluted protein is digested with a proteolytic enzyme such as trypsin. In an analogous fashion to restriction enzymes for DNA, proteolytic enzymes cut proteins at specific sequences (e.g., trypsin cuts after lysine and arginine residues), meaning that for a given protein we will produce a characteristic set of fragments. This peptide mixture is added to a matrix, which is then subjected to a laser pulse to vaporize and ionize the peptides. The machine then accelerates the ionized peptides towards a detector, which records the time of arrival of peptide ions. The time taken to reach the detector is a function of the mass-to-charge ratio of each ion. The result is a peptide mass fingerprint with a very precise measurement of the relative mass of each peptide. If we know the genome sequence of the organism in question, it is a simple matter for a computer to predict all the tryptic peptides that would be obtained from each protein. We can now compare the experimentally determined peptide mass fingerprint with the peptides predicted from each protein, and thus determine which of those proteins is present in the spot on the 2D gel.

An alternative, even more powerful procedure, is known as Electrospray Ionization Mass Spectrometry (ESI-MS). In this case, the sample is introduced into the machine in liquid form to produce charged droplets; as the solvent evaporates, ions are produced, which are analysed by the mass spectrometer. The main advantage of this is that it can be readily automated to analyse peaks eluted from high-performance liquid chromatography (HPLC), and so bypass the need for 2D gel electrophoresis, and the subsequent picking and isolation of spots

The previous two approaches rely on us knowing the genome sequence of the organism under study, so that we can predict the proteolytic digest pattern for the encoded proteins. As we saw in Chapter 8, the increase in high-throughput sequencing technologies means that more and more genome sequences are available. However, complete sequences are by no means available for every organism. How then do we undertake proteomic analysis in the absence of a genome sequence to predict the proteolytic digest pattern? Mass spectrometry can be used even if we do not know the genome sequence. Under suitable conditions, peptides can be randomly fragmented by a procedure known as Collision-Induced Dissociation, resulting in a series of ions differing, sequentially, by the mass of single amino acids. This requires two mass analysers, operating in tandem within the same instrument (known as MS/MS). The first separates the original peptides and selects individual peptides, which are then subjected to collision-induced fragmentation before the second analyser determines the sizes of the fragments produced. From these data, the sequence of the peptide can be determined. This sequence

can then be compared to databanks of known protein sequences, resulting in an identification of the likely nature of the protein in question.

10.5 Post-translational analysis: protein interactions

Most of the methods for post-translational analysis fall outside the scope of this book. However, procedures for studying which proteins interact with one another are worth considering as they provide a useful supplement to proteomic analysis in the genome-wide study of protein function.

10.5.1 Two-hybrid screening

One important clue as to the function of a protein is its ability to interact with other proteins in the biological environment. Two-hybrid screening is a commonly used method to identify such interactions; since this method is normally used with yeast cells as a host organism, it is therefore often called *yeast two-hybrid screening* (although versions are available for bacterial and mammalian cells as well). Although this procedure is commonly used for studying individual proteins, it can also be used for genome-wide screening.

The basis of this procedure is similar to yeast one-hybrid screening, as described in Chapter 6, in that it depends on the modular nature of transcriptional activators such as GAL4. The activator domain (AD), the part that stimulates transcription, does not have to be covalently attached to the DNA-binding domain (BD). If the activator domain can interact with a second protein, and if that protein is bound to DNA adjacent to a reporter gene, transcriptional activation will occur.

The basis of the technique is illustrated in Figure 10.5. As implied by the name, two recombinant plasmids have to be constructed. The first is designed to express one of the proteins that we wish to study (the *bait*) as a fusion protein with a specific DNA-binding domain, while the second plasmid expresses another protein (the *prey*) as a fusion with an activator domain. The host is an engineered yeast strain that contains, inserted into the genome, a reporter gene with a specific operator or upstream activator sequence to which the BD domain will bind. Expression of the reporter gene will only occur if the two fusion proteins interact.

The system can be set up to test the interaction between the products of two specific cloned genes, and as such is well suited to test the interaction of proteins from two candidate genes. This approach can be expanded to meet the needs of the post-genomic era by using a library of DNA fragments in the prey vector, and comparing them against a single bait vector. Such screening allows you to identify multiple proteins that are able to interact with the

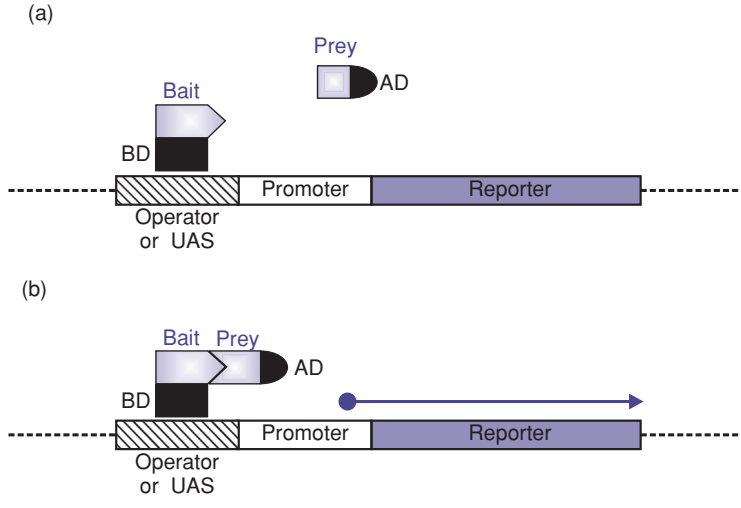


Figure 10.5 Yeast two-hybrid system. (a) Non-interacting proteins: no expression of reporter. (b) Interacting proteins: expression of reporter.

product of your bait cDNA. Finally, an array-based approach can be undertaken, allowing the examination for binding between multiple bait and prey vectors.

One word of caution should be noted for two-hybrid approaches. The interaction of a bait and prey protein in an artificial yeast system does not necessarily mean that these interactions occur *in vivo*; nor indeed does the absence of interaction mean that no interaction is possible. Some proteins are naturally 'sticky' and are likely to interact with many proteins in these assays, even though they will have no such interaction *in vivo*. Conversely, we may not be able to reproduce the exact conditions in the assay that allow proteins to interact *in vivo*, and they will hence come back as negative in the two-hybrid assay. Hence, it is important that we use further techniques to confirm these interactions and examine their biological consequences.

10.5.2 Phage display libraries

This procedure employs a filamentous *E. coli* phage called M13. As M13 replicates, it produces single-stranded copies of its DNA, which are extruded through the cell envelope, producing large numbers of phage particles without lysing the host cell. During passage of the DNA through the cell membrane, it becomes coated with phage-encoded proteins, which thus form the coat of the mature phage. If we insert foreign DNA fragments into a cloning site within the M13 gene coding for the coat protein, the result will be a hybrid protein that (we hope) will be incorporated into the phage particle in such a

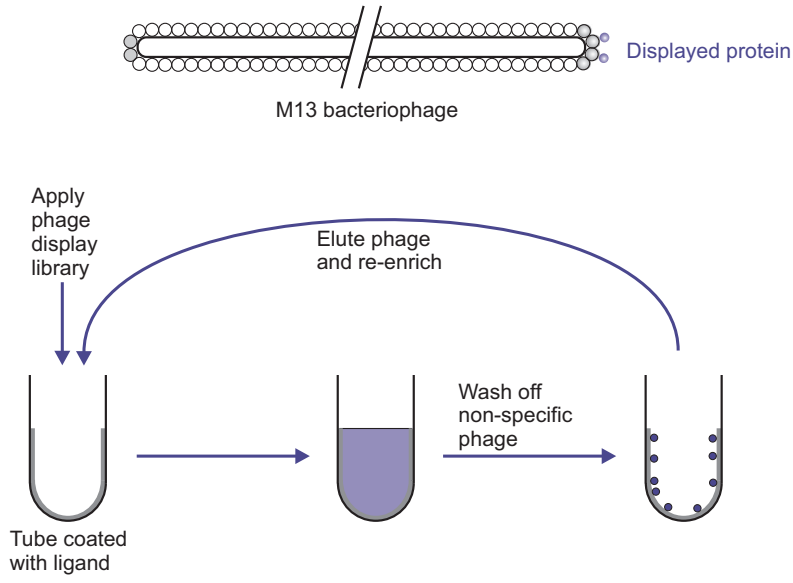


Figure 10.6 Phage display.

way as to *display* the foreign protein or peptide on the surface of the bacteriophage particles that result from infection of host bacteria (Figure 10.6). A phage display library can be created either by cloning DNA fragments that code for actual proteins or parts thereof, or by inserting synthetic oligonucleotides designed to give rise to a representative collection of random peptides. The phage particles displaying proteins or peptides with the required properties can be recovered by adsorption to a tube or a well in a microtitre tray coated with the target protein, so as to recover all the phage particles containing different proteins that will bind to the target. This approach is not confined to the identification of protein–protein interactions. Other ligands such as an antibody or a hormone receptor can be used. Following binding, non-specific phage are then washed off and the retained phage, enriched for the specific recombinant, are eluted; further rounds of enrichment can be carried out. The power of this technique lies in the ability to test very large numbers of phage particles – M13 preparations can contain 10^{12} phage particles per millilitre – and to carry out repeated rounds of enrichment for the phage displaying proteins or peptides with the properties that you want.

Phage display libraries therefore provide a versatile technique, not only for genome-wide surveys of protein interactions, but also for screening for specific proteins or peptides that can interact with smaller molecules of pharmacological interest. Identification of a peptide sequence that binds to the ligand can be of direct use either in itself (such as in the search for a

therapeutic agent), or can serve to provide clues about the partial sequence of a longer ligand-binding protein.

10.6 Epigenetics

The standard view of genetics is that all those characteristics of a cell that can be passed on to progeny cells (i.e., excluding variation due to environmental influences) are determined by the sequence of bases in the DNA. We now know that this is far from the truth, with other modifications to DNA and chromatin representing another level of heritable information; taken together these changes are known as epigenetics. At the level of DNA sequence itself, in addition to the base sequence, chemical modification of the bases in the DNA (especially methylation of the cytosines in CpG dinucleotides) can form the fifth (5-methylcytosine) and sixth (5- hydroxymethylcytosine) DNA bases, and this information is passed on from one cell to its progeny, although the fidelity is orders of magnitude less than that of DNA replication. Evidence of the importance of this mechanism in humans comes, for example, from studies that have shown that babies born to mothers on the verge of starvation have a life-long risk from a variety of conditions ranging from obesity to schizophrenia, and that this is associated with an altered methylation pattern in specific genes. Indeed, evidence exists to suggest that these changes can, like the sequence of DNA bases, be inherited through multiple generations (*transgenerational inheritance*), suggesting that exposure to certain environmental factors may affect not only you, but several generations of your progeny as well.

Due to the undoubted importance of epigenetics, if we want to get a complete picture of the genetic make-up of a cell, we need more than just the base sequence. We need to ascertain the methylation pattern as well. Currently, the commonest method to achieve this involves bisulphite treatment of the DNA, although newer technologies still under development claim to be able to detect methylation concurrently with sequencing. Bisulphite sequencing involves treating the DNA with bisulphite, which converts cytosine residues to uracil, but does not affect 5-methylcytosine (and 5-hydroxymethylcytosine) residues. Sequencing of the resulting DNA can then be used to determine which cytosine bases were methylated and which were not. A more recent innovation is the use of methylation-specific PCR, which allows the analysis of bisulphite-treated DNA without the need for sequencing. In this approach, primers are designed to be 'methylation-specific' by locating them over the CpG to be tested and designing two primer sets. The first primer pair contains a guanine to base-pair with the unconverted 5-methlycytosine (and 5-hydroxymethylcytosine), while the second pair contains a thymine to base-pair with the converted uracil base. Hence, by seeing

which primer pair produces an amplification product, we can see if the base is methylated or not. This approach is particularly useful for examining so-called CpG islands, which are collections of CpG dinucleotides that often occur upstream of genes, as primers can be designed to examine several CpG dinucleotides at once.

However, this is still not the end of the epigenetics story. Methylation of cytosine nucleotides is the best understood epigenetic mechanism, but it is not the only one. Prominent amongst other mechanisms is the post-translational modification of histones, especially by methylation and acetylation of specific lysine residues on histone tails. These alterations change the association of histones into nucleosomes, and their subsequent interaction with DNA to form higher-order chromatin structures. Chromatin conformation not only affects expression of certain genes, as closed chromatin is transcriptionally repressive (and vice versa), but is also transmitted from one cell to its progeny at mitosis, in a similar way to cytosine methylation.

10.7 Integrative studies: systems biology

Although the techniques described in this chapter are designed to look at the complete spectrum of gene expression in an organism, under the chosen conditions, they still do not really reflect the contribution that each gene, or its product, makes to the overall characteristics of that organism. The overall metabolic processes of a cell are defined not only by which enzymes are present, and the extent to which they are produced, but by a complex network of interactions, including physical interactions between the enzymes, interactions with regulatory proteins, and activation or inhibition by small molecules (substrates, products and messengers such as cyclic AMP). The extent to which we can predict the flow of material through any metabolic pathway by studying the individual enzymes in isolation is severely limited.

As the simplest example, we can consider the situation in bacteria. These can be grown under carefully controlled conditions in a chemostat, under which conditions it is possible to measure accurately all the input substrates and output products, and use the information from the genome sequence to predict the possible metabolic pathways. These data can be used to construct a model of the flow of material through each step of the pathways. By analogy with the genome, transcriptome and proteome, this sort of analysis is known by the rather ungainly term *metabolomics*.

10.7.1 Metabolomic analysis

The metabolome is the complement of small molecules (typically <1 kDa) within a cell, such as metabolic intermediates, hormones and other signalling molecules. Metabolomics is therefore the analytical description of these

molecules, and is important as it tells us about the biological output from the proteomic and transcriptomic levels of organisation. How then do we capture this level of information? We can use mass spectrometry, as we did for proteomic analysis, but this is complicated by the complex nature of the chemical soup that we are trying to analyse. Prior to MS analysis, metabolites need to be separated from their biological matrix, and this is usually achieved through either HPLC or gas chromatography (GC). Such HPLC-MS or GC-MS approaches tend to be very sensitive, but are prone to misleading results caused by the analysis of such complex mixtures. Rather than MS-based techniques, an alternative technology to perform metabolomic analysis is the use of nuclear magnetic resonance (NMR). Whilst NMR is not as sensitive as MS-based approaches, the analytes detected tend to be more reproducible, and NMR has the added advantage of being non-destructive to the sample, meaning that it can be used for further experiments.

For metabolomic studies, ^{1}H-NMR is commonly used, which produces spectra from all hydrogen-containing chemicals in the sample. The initial stage of analysis is to extract the chemicals, using either buffer-containing heavy water (D_2O), or deuterated organic solvents such as acetonitrile. Samples are then loaded into the NMR spectrometer, which first applies a magnetic field across the sample, aligning the spin of all nuclei in the sample. Next, irradiation of the sample with electromagnetic radiation is carried out; a unique absorption frequency is seen for each hydrogen nucleus, dependent upon the atoms that surround it. This will produce a spectrum showing peaks for each proton within the mixture. Comparison of two samples can identify differences in the spectra, and further analysis can then be used to identify the chemicals that have caused these differences, in a similar way to that we saw for proteomic analysis.

10.7.2 Pathway analysis and systems biology

The technologies described above allow us to gain a view of the different levels of organisation within biological samples, from the genome through transcriptome and proteome, to the metabolome. However, at present we have only discussed how these approaches can be used to identify differences between two samples, without considering how this all fits together to impact on the biological functioning. Such analysis is the remit of *systems biology*, which aims to understand the complex interaction networks that form a biological organism. Rather than studying a single gene, transcript, protein or chemical, systems biology understands the place of each of these species in the total biological environment, and how changing any of these species may impact upon biological functioning.

A complete examination of systems biology is beyond the scope of this book, and indeed would form a book on its own. However, it is worth

noting the basic premise of this approach, and its potential outcomes. Many computer programs now exist that contain biological information gathered from the published scientific literature. This information is linked together into pathways, such that the computer knows which proteins or chemicals have been shown to interact together *in vivo*. Overlaying '-omic level' analyses onto these databases allows you to see not only which individual species changes, but how this may impact upon individual pathways, or indeed whole networks. More complex analysis is also available, allowing these pathways to be simulated in a computer; this is the first step toward the generation of *in silico* cells, tissues and, eventually, whole organisms.

11 Modifying Organisms: Transgenics

Much of this book deals with the genetic manipulation of individual cells in culture, either by introducing and expressing genes from other sources, or in other ways specifically altering the genetic composition of those cells. In this chapter, we will deal with the stable introduction of a gene into another *organism*, which is referred to as *transgenesis*, the result being a *transgenic* organism. For a bacterial or yeast geneticist, there is no distinction between manipulating individual cells and manipulating an organism. In effect, all the previous discussion of genetic modification of bacteria, and unicellular eukaryotes such as *Saccharomyces*, could be labelled as transgenics. And yet for somebody who works with truly multicellular organisms, such as plants or animals, the distinction between manipulating a cell line in culture and genetic modification of a whole plant or animal is very real and important. A gene can be stably transfected into a cell line quite cheaply, and this can be used to provide fast and reasonably useful answers to many questions – but is inevitably limited in scope. In particular, it cannot provide definitive answers to questions about the differentiation of cells in the normal animal, nor about the interaction between different cells. The creation of a genetically manipulated whole organism is much more difficult, but often much more informative.

11.1 Transgenesis and cloning

It is important to distinguish the production of transgenic plants and animals from cloning of the plants or animals themselves. The former involves the introduction of foreign DNA, in the form of cloned genes. In contrast, the

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

latter means obtaining progeny that are genetically identical to the original plant or animal, which includes methods ranging from the widely publicized procedure that led to the birth of Dolly the sheep to techniques for asexual propagation of plants (such as taking cuttings) that have been in use for hundreds of years or more.

When you introduce a *transgene* into an organism you insert into the genome a foreign gene, be it from the same species or from a different one. Owing to the virtual universality of the coding properties of DNA, huge evolutionary distances can be bridged in transgenesis – the example of a fish gene for cold hardiness inserted into tomato plants achieved notoriety. In order to create a transgenic multicellular organism, the transgene must be inserted into germline DNA, so that it can be propagated in subsequent generations. Other applications where genes are inserted into *somatic cells* (cells that are not involved in reproduction of the organism), for example, for gene therapy, are considered later in this chapter.

11.1.1 Common species used for transgenesis

In principle, the technology is available to insert transgenes into any species. In some species it is also technically possible to do what is conceptually the opposite – to *knock out* a gene (which may involve disrupting the gene rather than deleting it entirely). The use of gene knockouts in bacteria was considered in Chapter 5; as we will see later on, obtaining gene knockouts in whole animals and plants follows a similar procedure. Before we consider the practicalities of transgenesis, it is useful to examine which species are commonly used for transgenesis experiments, and why. These species are chosen, in general, due to their ease of manipulation and the ability for discoveries in these 'simple' organisms to be extrapolated to the 'more complex' scenario of humans.

Amongst the simplest organisms are the slime moulds, such as *Dictyostelium discoideum*. When food is plentiful, these behave as unicellular organisms, but they switch to an essentially multicellular organisation when food becomes limiting. Under these circumstances they aggregate to form a tiny 'slug' (1–2 mm long) that crawls around and eventually develops into a structure containing a fruiting body and a stalk. Many of the signals that control the differentiation of the initial cells into stalk or spore cells, and the genes responsible, have been identified by the transgenic techniques described below, and have increased our understanding of development and differentiation in higher species, including humans.

A second important model species is the nematode *Caenorhabditis elegans (C. elegans)*. This is small (about 1 mm long), easily maintained in the laboratory, and has the attraction for developmental biologists that it contains a precise number of cells (959), and the entire ancestry of every cell is known. RNA interference (RNAi; see later in this chapter) has proved to be an

especially valuable technique with *C. elegans*, since it can be achieved by feeding the nematodes with bacterial clones expressing dsRNA targeted to specific genes. Genome-wide RNAi screens have successfully identified the function of many of the genes of the *C. elegans* genome. Furthermore, the organism is transparent, which greatly facilitates the use of reporter genes (such as that coding for green fluorescent protein, GFP) to study *in situ* gene expression and localization of the product.

Third in the list of model species is the fruit fly *Drosophila melanogaster*. Like those of *C. elegans*, *D. melanogaster* individuals are small (2–4 mm long), easy to maintain, well characterized and have been used extensively in the field of developmental biology. These fruit flies are particularly tractable for genetic manipulation, with many genotype-phenotype relationships being well understood. The obvious advantage of fruit flies over nematodes is one of evolutionary distance: Humans and nematodes diverged nearly 1000 million years ago, whereas the human–fly divergence was only approximately 600 million years ago. This means that it is often easier to extrapolate findings from fruit flies to humans, making them an ideal model species for studying human biology.

Whilst the fruit fly *D. melanogaster* has many biological systems in common with humans, there are still major differences, typified by the use of a vertebral spine in humans and an exoskeleton in flies. It is hence desirable to use model species that are more closely related to humans, thus making any extrapolations of findings more robust. Amongst more complex organisms, the zebrafish *Danio rerio* is popular due to its ease of use in the laboratory. This is a small (4 cm) freshwater fish, native to south Asia. The embryos develop rapidly (within a few days) and are transparent – both features being useful for their study. Given all of these advantages, plus the shorter divergence time between humans and fish (approximately 450 million years), then it is perhaps not surprising that zebrafish are often termed the 'vertebrate *Drosophila*'. The roles of many of the genes involved in embryonic development of zebrafish have now been identified, and show close correlation to the processes seen in humans. In addition to developmental studies, zebrafish are an attractive model for studying many other aspects of human biology. These include microbial infections, as zebrafish possess innate and adaptive immune systems with many of the features of mammalian immunity. They are also useful in studies of drug action (since many of their core biological functions, such as the cardiovascular and ocular systems, are similar to humans), as well as for circadian biology (there are many common features between the human and zebrafish circadian machinery). A further example is the development of diseases such as cancer, where transgenic zebrafish have been used to study how tumour cells invade and proliferate within the body.

However, in terms of ease of use and reliability of extrapolation of findings to humans, transgenic mice are often the organism of choice, although

they do come with increased time, cost and ethical burdens. More controversially than these research-oriented applications, a transgenic organism may be created not to answer a question, but for a specific, usually commercial, purpose. Commonest amongst these are the production of transgenic plants, but animals may also be altered both for commercial gain or to create living bioreactors for the production of pharmaceutical products, so called 'pharming'. Examples of all these will be provided in later sections.

11.1.2 Control of transgene expression

The simplest method of creating a transgenic organism is through the process of *non-targeted integration*. Such an approach means that the position for successful integration into the genome of the target organism is more or less completely random. The desired outcome is obviously that the gene lodges itself into a region where it does not disrupt another gene. Because most of the DNA in plants and animal does not actually form part of a gene, the odds are stacked against any such problems occurring. Nonetheless, occasionally they do, either by the transgene lodging itself directly into a gene or into a region that is important for gene regulation. In some instances, this is lethal. In others, there is an obvious risk that the disruption of a different gene creates a phenotype that has nothing to do with the transgene itself. The random integration site of the transgene will also affect its level of expression due to two important factors. First, there exist large-scale regional differences on the chromosomes with regard to their ability to support expression of the transgene, predominantly due to altered chromatin status. Second, the number of copies of the transgene that integrate is not fixed, and may range from a single copy to several hundred. Such variation in localization and insert number may result in *position effects*, causing significant variability in the level of expression of the transgene from one organism to another, meaning that it is often necessary to screen many progeny before the required expression level is found.

Even if a transgene does integrate into a region of the genome that is conducive to expression, and the integration does not disrupt any other genes, there is a final problem to consider. In Chapter 7 we discussed the expression of cloned genes in mammalian cells in culture. In that context, the emphasis was on promoters that were constitutively active, or could be turned on and off by altering the culture conditions. Neither approach is really suitable for a transgenic animal, where we would normally want to target gene expression to specific tissues. To ensure that we optimise expression of the transgene as much as possible, it is necessary to include other regions of the DNA beyond the gene coding region in the targeting plasmid. These regions are used to control and modify expression of the transgene once it has been successfully integrated into the genome of the target organism.

For some purposes, the ideal objective would be to introduce the native transcription unit into a genotype that lacks it. This would require a cloned sequence covering the whole gene and its associated regulators, including activator sequences and promoter elements as well as the introns and exons. The conceptually simplest, and often preferable, method of constructing such a transgene exploits the large coding capacity of yeast artificial chromosome (YAC) vectors, as described in Chapter 2.

One alternative is to mix and match promoter and coding sequences. For example, you could combine a genomic clone of the promoter region of the native gene with a cDNA clone of the same gene. The absence of introns may have some effect, but introns are not *necessary* for mammalian gene expression – some mammalian genes do not have them in the first place. What may make more of a difference, however, is the promoter. You would certainly have pinpointed it using reporter gene constructs before spending a minor fortune on producing a transgenic organism. However, the conditions in the cell line where this was done do not necessarily apply to the conditions in the native cells. Moreover, there may be important elements – positive or negative – that are located some considerable distance away from the transcription start site.

Apart from the obvious combination of a cDNA with its own promoter, there are many inventive and informative combinations that can be – and have been – made, either to obtain tissue-specific expression of the transgene or to identify the tissues or cell types in which that gene would naturally be expressed. Many of these examples are related to topics that have been discussed in earlier chapters in relation to the manipulation of unicellular microorganisms or of cell lines in culture. Here are a few examples:

1. If the promoter region of the gene you want has not been characterized, a cDNA encoding your protein may be combined with a different promoter that has the same cell or tissue specificity.

2. In order to investigate the exact cell/tissue specificity of a promoter, or its activity under different conditions, it may be combined with a reporter gene such as *lacZ,* luciferase or green fluorescent protein (which are easy to detect *in vivo*). Similar applications of reporter genes for individual cells in culture were described in Chapter 6.

3. A promoter may be combined with a gene, such as the gene for the A subunit of diphtheria toxin (DTA), which will specifically damage any cell type in which it is expressed. (The inserted gene does not include the part of the toxin, the B subunit, that is needed for cell entry, so the toxin subunit will only affect the cells in which it is produced.)

4. A promoter may be combined with a gene that adds a specific functionality to a particular cell type. A common variant of this is the rescue of a recessive mutated phenotype by introducing the wild-type gene, which will provide conclusive proof of the identity of the gene. This can obviously be accomplished with a YAC clone as well. See also the discussion of complementation in Chapter 5.

5. The inclusion of boundary elements (insulators) to minimize position effects. As the name suggests, insulators prevent regulatory elements from one region of the genome/gene from affecting expression of other regions/genes. By placing insulators either side of the transgene construct, it is possible to prevent enhancers that regulate genes near to the integration site from affecting expression of the transgene.

6. In order to ensure expression of the transgene in a specific tissue, it is possible to use a promoter that is only active in that tissue. For example, albumin is only produced in the liver, and hence any transgene under the control of the albumin promoter will also only be expressed in the liver.

7. As with tissue-specific expression, it is also possible to control temporal expression using promoters that are only active during specific times in the development of an organism. For example, animals express different subunits of β-globin during development, producing haemoglobins with different affinities for oxygen (Figure 11.1). By using the promoter from a specific β-globin linked to your transgene it is therefore possible to control at which point in an animal's development the transgene is expressed.

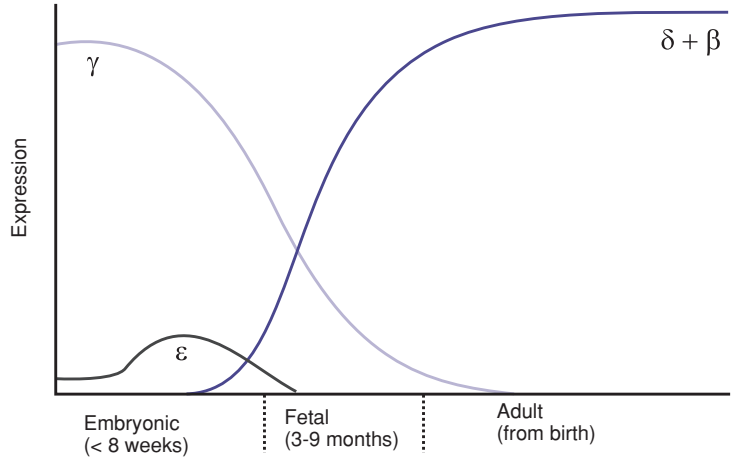


Figure 11.1 Differential expression of beta-globins during development.

11.2 Animal transgenesis

11.2.1 Basic methods

In order to ensure expression of the transgene in every cell of the animal, the transgenic construct must first be introduced into the germline. This means that it will be passed on to the offspring of the animal, with every cell of that offspring having the transgene integrated into its genome. There are three major ways that this can be achieved, each with its own advantages and disadvantages. These methods are direct injection, retroviral vectors and embryonic stem (ES) cell culture, and each will be discussed in detail in the following sections. As we will see, transgenesis may result in a number of genetically different offspring, each of which is referred to by a specific term. First, transgenesis may be used to either add an extra piece of DNA, usually the coding region for a gene, into an organism, creating a *knock-in*, or to disrupt the DNA sequence of a genome, usually destroying a gene (*knock-out*). As the majority of organisms are diploid, having two copies of each chromosome and, hence, of each gene, it is therefore possible to create a knockout animal where the sequence on only one chromosome has been disrupted (*heterozygous knockout*), or an animal where the sequences on both sister chromosomes have been disrupted (*homozygous knockout*). Later in this chapter, we will also encounter *knock-downs*. Finally, whereas the ultimate aim of transgenesis is to create an organism in which every cell contains the altered sequence, several of the procedures outlined below first produce a *chimera* (or mosaic). In these animals some cells will contain the transgenic region while others will not, and further rounds of breeding will be required to create a fully transgenic animal.

11.2.2 Direct injection

The most obvious, and most widely applicable, method for transgenesis is to physically inject the transgenic construct into the nucleus of a fertilized egg. This egg is then cultured *in vitro* for several cell generations before being implanted in a pseudopregnant female (Figure 11.2). The creation of a pseudopregnant female is achieved by mating a female animal with a vasectomized male, and is necessary to ensure that the hormonal status of the female is conducive to the successful survival and growth of the reimplanted egg. The resulting offspring will be completely transgenic, with every cell containing the transgene, and will most likely be heterozygous, with the transgene integrated into one chromosome only.

Whereas direct injection is conceptually simple, it is by no means a trivial affair. Not all eggs survive the injection without damage, with many dying following overnight incubation. From those that survive, not all of them will

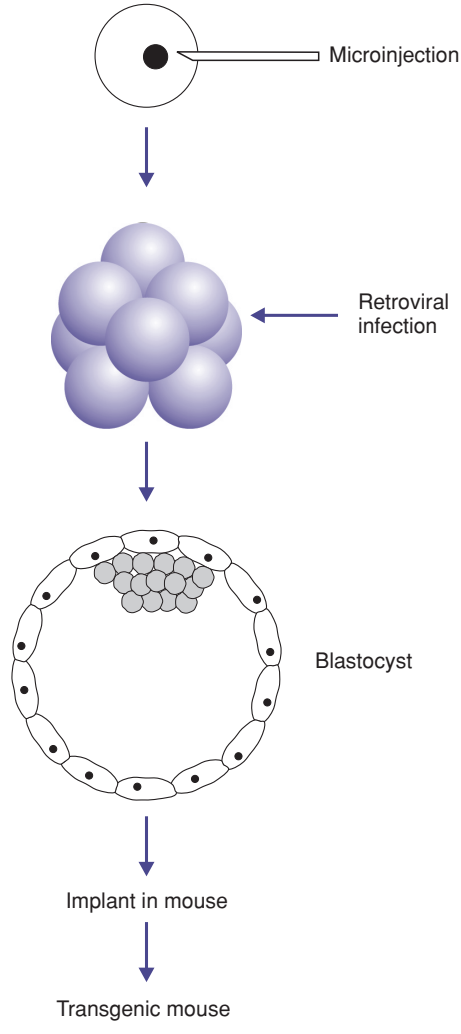


Figure 11.2 Two methods for producing transgenic mice.

develop, and not all of those will contain the transgene. Only a proportion of the resulting embryos will be transgenic, so it is necessary to screen the embryos (using gene probes and/or PCR) for the presence of the foreign gene. Taken together, this can mean that producing a transgenic animal using direct injection can represent a significant time and financial commitment, and the large number of animals required for this procedure can be ethically challenging. Despite these apparent obstacles, direct injection is still used widely today due to two major advantages that this approach has. First, the result for each embryo is all or nothing, with chimeras not usually occurring. Second,

the procedure is applicable to a wide variety of different organisms, and is especially good for organisms with larger eggs, such as fish.

11.2.3 Retroviral vectors

One alternative method to direct injection for the creation of transgenic animals is to use retroviral vectors, as described in Chapter 7. Technically, this is much simpler, with the absence of an injection stage vastly reducing the number of eggs lost due to non-survival. To create a transgenic animal using retroviral vectors, a fertilized egg is allowed to develop to the eight-cell stage, making it significantly more robust to manipulation, before it is infected with the (defective) recombinant virus (Figure 11.2). Following incubation to allow the infection to occur, these eggs are then treated in a similar fashion to those created by direct injection, with reimplantation into pseudopregnant females.

This method has distinct limitations, however. Firstly, retroviral vectors can only accommodate approximately 8 kb of DNA, which may not be sufficient for the whole coding region and the necessary promoter elements, as discussed earlier. Secondly, the modified retrovirus does not contain the genes necessary for replication as a virus. It needs assistance from a *helper virus* to be infective. This helper virus could lodge itself into the genome as well and, like a prophage, become infective again at some stage in the future. If the helper virus is also present in the embryo, it can result in the unwanted spread of the recombinant construct to other animals. The method described in Chapter 7 uses *helper cells* as an intermediate stage, producing infective virus particles that do not need a helper virus, and such an approach can be used here to prevent this particular problem. However, there is still the possibility that a subsequent retrovirus infection of the transgenic animals might result in mobilization of the recombinant DNA. A further limitation is that, although infection of eight-cell embryos is technically simpler than microinjection, the fact that cell division has occurred before infection takes place means that there is a possibility that not all the cells in the embryo will be infected. Mosaicism is therefore a potential problem, where the derived transgenic animals do not contain the transgene in every cell. In such cases, it is then necessary to carry out several rounds of breeding between the chimeric mice and wild-type mice, known as a *back-cross*; selection of the progeny will ultimately result in stable homozygous mutant mice.

In summary, whereas retroviral-mediated transgene integration has the advantages that it is both technically easier and tends to use fewer animals than direct injection there are a number of distinct disadvantages. The animals produced are generally chimeras, the region of DNA that can be integrated is relatively small, the integration site of the transgene is non-targeted (potentially resulting in position effects), and reinfection with the retrovirus

is possible. Due to these numerous disadvantages, retroviruses are limited in their application for transgenic animals, but do play a significant role in the development of potential techniques for gene therapy (see later in this chapter).

11.2.4 Embryonic stem cell technology

Some types of cells taken from an adult animal can be grown in culture, for a short time. In some cases, these can give rise to established *cell lines*, which can in principle be maintained indefinitely through serial subculture. Manipulation of animal cells in culture (as considered in previous chapters) is considerably easier than introducing genes into a fertilized egg. But the differentiation of such cells is not usually reversible; they cannot be used to produce a whole animal. This is in marked contrast to the situation in many plants, where cells from various parts of the plant, although differentiated, are still capable of producing whole plants. This is why it is so easy to propagate many plants asexually, for example, by taking cuttings, or even by starting from *in vitro* cell cultures. It is also why there was so much excitement over Dolly the sheep, which was a demonstration that it *is* actually possible, through some intricate techniques, to reverse the differentiation of some cells from an adult animal.

The important characteristic of stem cells is that they are able to differentiate into multiple different cell types. In general, such cells can be described as multipotent, pluripotent or totipotent. *Multipotent* cells have the potential to differentiate into many, but not all cell lineages, and hence would not be applicable for transgenesis, where we wish to grow an entire organism. The converse is *totipotent* cells, which may differentiate into all cell lineages, both embryonic and extraembryonic, such as the placenta. However, for the production of transgenic animals we use *pluripotent* cells, which are capable of forming all the cell lineages within an embryo, but cannot form extraembryonic lineages.

Cells at an early stage in the development of an embryo still have the potential to develop into all the cell lineages in an animal, being pluripotent; these are known as *embryonic stem cells* (ES cells). Such cells can be collected at the blastocyst stage of development – the blastocyst is a large sphere of cells containing an inner cell mass, or embryoblast, that forms a few days after fertilization of the egg (Figure 11.2). ES cells are collected from the inner cell mass, and can be grown in culture for a limited period of time provided they are kept in special medium to prevent them differentiating. Indeed, a distinct advantage of ES cells is that ES cell lines can be established, or bought commercially, meaning that it is not necessary to harvest cells from embryos on a regular basis. The availability of cells in

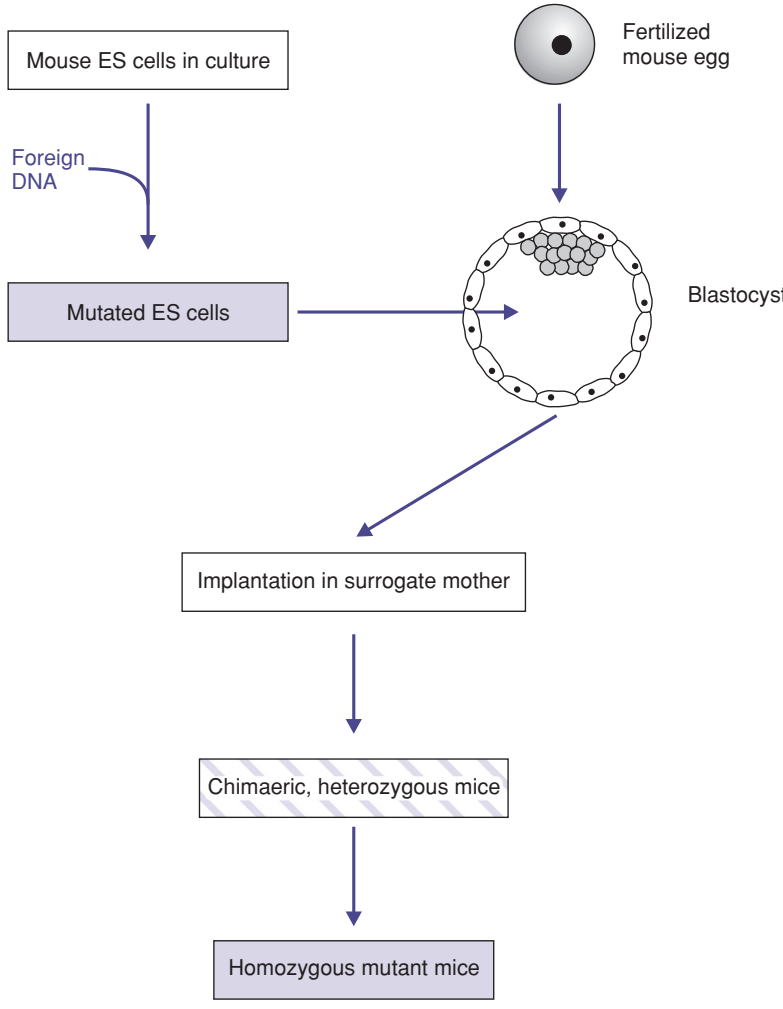


Figure 11.3 Embryonic stem cell technology.

culture, whether they are fresh ES cells or a cell line, means that many of the techniques described in earlier chapters can be employed – in particular, the investigator can include selectable markers in the construct. These convey resistance to toxic compounds that are then added to the growth medium. This is similar to antibiotic selection for plasmids in bacteria: only cells that have successfully integrated the transgene will survive. These selected cells are then injected into blastocysts and implanted into a surrogate mother (Figure 11.3). The progeny will be chimeric, since the blastocyst will still contain some of the non-manipulated cells, and heterozygous, since normally only one of the chromosomes will be altered. As we saw for retroviral-produced

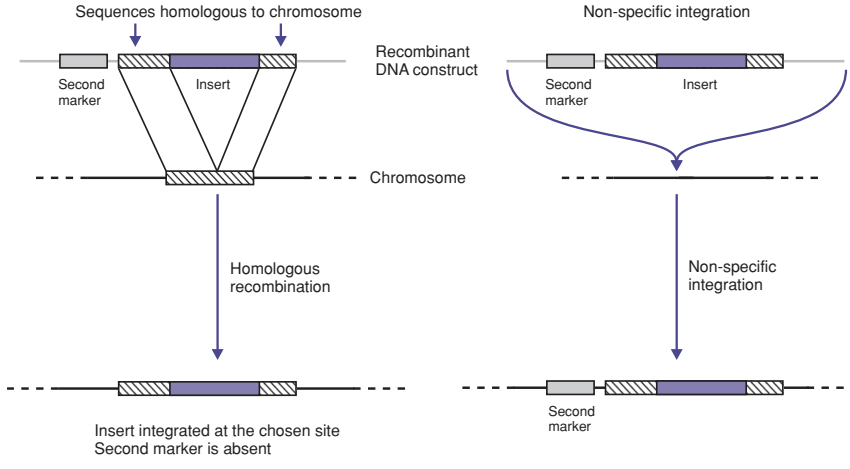


Figure 11.4 Gene insertion at a specific site.

transgenic animals, it is therefore necessary to undertake several rounds of back-crossing to produce progeny that contain a stable homozygous insertion in every cell.

A further advantage of ES technology is that it allows the introduction of the construct in an exactly specified site in the host genome, so called *targeted-integration*. This is done by including two sequences that are identical to two particular adjacent regions in a host chromosome. Recombination can occur at these sites, and will lead to the specific insertion of the transgene exactly in the desired place, by *homologous recombination* (Figure 11.4). The inserted gene would normally be accompanied by a selectable marker gene, so that we can select the cells in which the insert has become integrated into the chromosome. However, it has to be remembered that in animal cells random integration of transfecting DNA occurs quite frequently, so it may be necessary to screen a number of clones in order to find one that has the insert at the right place. This can be made easier by including a second marker gene on the construct, outside the region flanked by the homologous sequences. Random integration of the transfecting DNA would lead to integration of the whole construct, while integration by homologous recombination will only insert the region between the two recombination sites (see Figure 11.4). The cells we are looking for will therefore lack the second marker from the construct, and can thus be easily identified.

In summary, ES cell-mediated transgenesis has a number of advantages over both direct injection and retroviral-mediated transgenesis: large regions of DNA can be successfully integrated into the target organism; this integration can be targeted to specific regions of the genome, allowing gene

knockouts; and the growth of ES cells in culture allows rapid and simple screening of transgenic cells *before* reimplantation, which is much more cost-effective and ethically acceptable. However, it should be remembered that, like retroviral-mediated transgenesis, the resultant offspring will be chimeric, and thus require further breeding to achieve a stable, completely transgenic animal. In addition, ES cells are not available for many species, and thus the application of this technology is really limited to the creation of transgenic mice.

11.2.5 Gene knockouts

Embryonic stem cell technology with homologous recombination also makes it possible not only to *add* a gene, but also to *remove* (or *modify*) an existing one. In the above description of targeted integration, our choice of homologous sequences would be designed so that integration will take place in a part of the genome that allows the gene to be freely expressed without disrupting any other genes. However, disrupting a gene may be exactly what we want to do, creating a *knockout animal*, and we can achieve it in a very specific way. A transgenic construct is produced that contains resistance markers, and a part of the target gene with a specific manipulation. This could be the disruption of the promoter, the introduction of a stop codon early on in the coding region, or the deletion of part of the gene. In the version shown in Figure 11.5, the target gene is disrupted by the incorporation of a *neo* gene. This gene gets its name because it conveys resistance to neomycin in bacteria, but in animal cells another aminoglycoside antibiotic, known as G418 or Geneticin, is used. It simultaneously fulfils the roles of disrupting the target gene, creating the knockout, and providing a selectable marker for insertion of the transgene into the genome. This procedure is basically the same as that described in Chapter 5 for allelic replacement in bacteria. A second selectable marker, such as the thymidine kinase gene, is often also incorporated into the targeting vector, outside of the regions of homologous recombination, to allow selection of ES cells where integration has occurred in a targeted manner only.

Knockout mice have been absolutely crucial in elucidating the exact function of many genes, by investigating the anatomical, physiological, biochemical and behavioural characteristics of the knockout strain. By studying what happens when a gene is removed, scientists are able to draw important conclusions about its function. The method is not without its pitfalls, however. The absence of the gene may be lethal to the embryo. Alternatively, overexpression of other genes may compensate for the absence of the knocked-out gene, thus obscuring the phenotype.

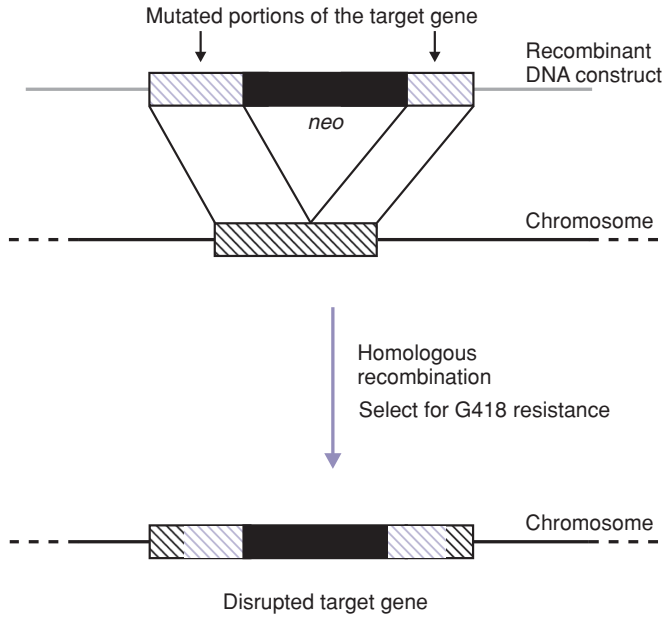


Figure 11.5 Gene knockouts. The *neo* gene codes for an aminoglycoside phosphotransferase that gives resistance to neomycin and to G418 (Geneticin).

11.2.6 Gene knock-down technology: RNA interference

The genomes of probably all multicellular organisms encode a large number (as many as 1000) different *microRNAs* (miRNAs). In Chapter 8 (see Figure 8.23) we introduced the concept of RNA interference (RNAi) or gene silencing, which is an artificial equivalent of the role that miRNAs play in regulating gene expression. In summary, double-stranded RNA is introduced into cells, where it is processed by the enzyme Dicer and loaded into the RNA-induced silencing complex (RISC) as a single-stranded, antisense RNA molecule (short interfering RNA, or siRNA). Once loaded, RISC will then pair with mRNA molecules that show complementarity to the loaded siRNA, and reduce protein production through one of two methods. If the complementarity between the siRNA and mRNA is perfect then Argonaute2-mediated RNA degradation occurs, whereas if the complementarity is only partial then inhibition of translation occurs. This causes a loss of function of that gene (*gene knock-down*), and is a much more convenient way of achieving that end than knocking out the gene.

RNA interference is widely used with cells in culture, but the challenge was to transfer this to whole organisms. Initial attempts revolved around

the addition of siRNAs to adult animals, using techniques such as hydrodynamic perfusion, or injection into the target organ. However, these approaches generally only last a single generation, and indeed may not even be stable for the entire life of the animal. The combination of traditional transgenic approaches with RNAi has, more recently, allowed germline transmission of siRNA constructs in mice and rats. There is thus considerable potential for RNAi technologies to be used not only to understand basic biological functioning of genes, but also for gene therapy in humans (see below).

A major advantage of this approach is that it is very versatile, and will basically work with any eukaryotic cell, ranging from unicellular organisms (such as yeast) to multicellular ones (animals or plants). In addition, unlike classical knockout technology, you can observe the knock-down in progress. If the knock-down is lethal to the cells, you have the opportunity to see what happens to the cells before they die, which can be very informative. A final advantage is that you are less likely to see compensatory overexpression of other genes when using RNAi to disrupt gene expression compared to traditional gene knockout approaches. There are, however, limitations to this approach. Non-specific adverse events have been reported in mice transfected with too high a level of siRNAs. Also, it is uncommon to see 100% knock-down of gene expression using RNAi, meaning that some residual biological activity may remain and must be accounted for when interpreting the data. Furthermore, RNAi is not as specific as gene knockout technology, and it is possible that a single siRNA might alter the expression of many different mRNAs, which is obviously undesirable when attempting to examine the role of only one gene product.

11.2.7 Gene knock-in technology

Most diseases are not caused by the complete disruption of gene function, but by mutations that alter it only in part. These effects are more likely to be subtle and less likely to be lethal. Such mutations can be introduced by *knock-in* technology, where one or more exons are replaced with altered ones. Indeed, producing animal models of disease is not the only important application of this technology, or even necessarily the most important one. The concept is important in any situation where we do not want to lose a gene completely, so knockout technology is not appropriate, but rather where we want to introduce specific changes into that gene. One important example of the application of knock-in technology is the production of mouse strains in which critical portions of significant genes have been replaced with, or engineered to be more similar to, their human equivalents. The potential value of such

strains, for example, in the testing of the toxicity and/or action of new drugs, is obvious.

The procedure for producing knock-in mice is very similar to that described above for gene knockouts, or to the procedure in Chapter 5 for allelic replacement in bacteria. The desirable mutation(s) is introduced into the exon(s) in question, using site-directed mutagenesis or other procedures described in Chapter 7, and a selectable marker such as the *neo* gene (see above) is inserted into an intervening intron. Homologous recombination will then result in the replacement of the chromosomal gene with the modified sequence.

From this description of the procedures that can be used, it should be clear that transgenic technology is extremely valuable for research into the role and regulation of specific gene products in the development and characteristics of a variety of animals. We now want to turn to a few examples of practical applications of these techniques.

11.3 Applications of transgenic animals

The use of transgenic technology for commercially important animals such as cattle and sheep is highly controversial. This should be put in perspective. Conventional properties of farm animals, such as their rate of growth or milk production, have been manipulated by selected breeding for thousands of years, resulting in the specialized breeds we see today that bear little resemblance to their original ancestors. The novel aspect of transgenic technology is the ability to introduce additional genes, from other sources. This means that the animals can be used as a source of a variety of useful products.

An example of the potential for genetic engineering to enhance commercially important animal products involves attaching the foreign genes to a promoter that will be active in mammary tissue, such as the casein or β-lactoglobulin promoters. Consequently, the protein concerned will be produced in large quantities in milk. This opens the field not only for enriching the milk itself with various proteins, but also for the use of cattle or sheep as bioreactors for producing large amounts of proteins with various uses. Another example is using the lysozyme promoter in a hen to drive expression of a gene for a protein that you wish to express. The transgenic animal will then lay 'bioreactor' eggs that are full of the protein you wish to express. The production of pharmaceutically useful proteins by farm animals has been termed 'pharming'. This can potentially offer substantial advantages over the use of recombinant bacterial cultures. In the first place, it avoids the need to build and run expensive industrial-scale fermenters. In addition, using animal hosts means that the product is likely to have the appropriate post-translational modifications, which is often not the case for products obtained from bacteria.

Recombinant plants also offer considerable potential in this way, as described subsequently.

11.4 Disease prevention and treatment

11.4.1 Live vaccine production: modification of bacteria and viruses

Genetic modification has a relatively long history of application to microorganisms (bacteria and viruses), especially the use of bacteria for the production of novel products. This was dealt with in Chapter 7. Here, in keeping with our Gene to Genomes theme, we want to focus on modifications that result in an *organism* with novel properties. The main example, in respect of microorganisms, is in the use of genetic modification to produce novel live vaccines. The impact of genetic modification on production of other types of vaccines was covered in Chapter 7.

Live attenuated vaccines One of the limitations of non-living vaccines, including both conventional killed vaccines and subunit vaccines, is that the protective effect on initial immunization can be relatively poor and short-lived. Boosters are commonly used, but even so the protection is usually inferior to that conferred by a live vaccine.

The conventional route for development of a live attenuated vaccine involves repeated laboratory subculture, especially under conditions unfavourable to the pathogen. For example, a virus may be repeatedly cultured using a cell line that is a poor host for the wild virus, or a bacterium may be repeatedly grown using a medium in which the pathogen grows very weakly. The principle is that as the pathogen adapts to these unusual conditions it is likely simultaneously to lose the ability to infect human or animal hosts. Although this empirical procedure has been successful in producing several widely used vaccines, it is rather hit or miss. Many mutations are likely to accumulate, apart from those that give rise to the desired attenuation, so there is a good deal of uncertainty in the nature of the resulting strain. Although the sequence can be determined, it is still difficult to know which of these mutations is associated with loss of virulence.

It can be important to understand the nature of the loss of virulence. If it is due to just a single mutation, and especially if that is a simple point mutation, then it is quite possible for that mutation to revert. In other words, a further mutation at that site may restore the original sequence of the gene, and the strain is no longer attenuated.

It is also necessary to recognize the possibility that the strain will not only have lost virulence, but also may have lost the ability to induce protection.

The structure of the key antigens may have changed, for example. Furthermore, it is possible for a strain to become too attenuated to be useful as a vaccine. The most effective protection is achieved by a microorganism that is able to undergo at least a limited degree of replication in the tissues before being eliminated by the host defences. If the strain is over-attenuated, the body will eliminate it before it has a chance to evoke a full immune response.

With the classical empirical procedure, it is difficult to control these factors. Genetic technology provides rational ways of achieving these goals in a controlled manner, by site-directed mutagenesis (Chapter 7) or by gene replacement (Chapter 8). The relatively small size of viral genomes makes them especially attractive targets for these techniques.

This requires some knowledge of the genes that are necessary for virulence. If pathogenicity of a bacterium is due to the production of a protein toxin, then you would expect the inactivation of the toxin gene to produce an attenuated derivative. Usually it is not so easy. With most bacteria, pathogenicity is multifactorial. Much effort is being devoted to attempts to identify genes that are needed for virulence, especially those that are selectively expressed during infection (using, e.g., the techniques described in Chapter 8) – which would identify targets that could be knocked out in order to attenuate the bacterium.

The most successful approach so far has been to target genes that are involved in certain central biochemical pathways rather than the more obvious virulence genes. This takes advantage of knowledge that the bacterium at the site of infection will not have a ready supply of certain key substrates, such as some amino acids. It will have to make them for itself. Mutants in which key genes in that pathway have been disrupted will therefore be unable to multiply inside the body. Of course we can still grow these strains in the laboratory, simply by adding the required amino acid to the medium. One such pathway is that for aromatic amino acid biosynthesis. *Salmonella* mutants in which the *aroA* gene has been disrupted have reduced virulence.

Live recombinant vaccines One of the novel advances in vaccine technology that has been made possible by genetic manipulation is the addition of genes from another pathogen into an existing live vaccine strain. Much of this work has been carried out using the vaccinia virus (the smallpox vaccine) as the carrier. Vaccinia is a rather large and inconvenient virus to use for genetic manipulation (with a genome size of 187 kb), and the insertion of genes requires the use of *in vivo* recombination. This is shown schematically in Figure 11.6. The required gene is inserted into an *E. coli* vector, adjacent to a promoter derived from vaccinia. This plasmid also carries a defective thymidine kinase (TK) gene from vaccinia, interrupted by the promoter and the cloning site. If an animal cell is infected with vaccinia virus, and at the same time transformed with the recombinant plasmid, the homology between the TK genes

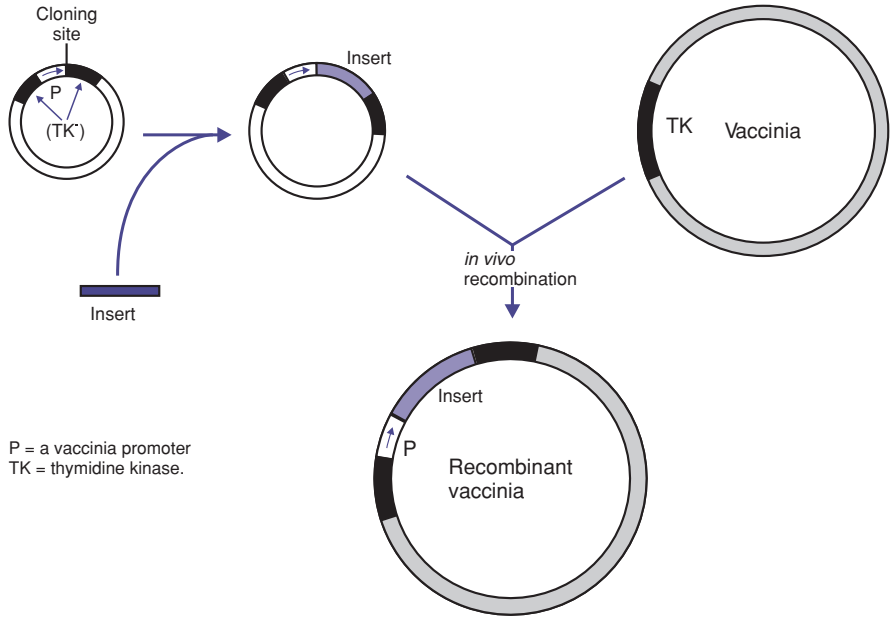


Figure 11.6 Construction of recombinant vaccinia virus. The plasmid carries an interrupted vaccinia TK gene. Recombination occurs between the TK sequences on the plasmid and vaccinia DNA, after co-infecting animal cells. The resulting recombinant is TK^-, and expresses the insert gene from the vaccinia promoter.

on the plasmid and the virus will allow specific homologous recombination. This results in the incorporation of the insert from the recombinant plasmid, together with the adjacent promoter, into the viral DNA. The recombinant virus now lacks a functional TK gene, which provides a way of selecting the recombinants.

Antigens from a wide variety of viruses and other pathogens have been expressed in this way as recombinant vaccinia viruses. In one example, a recombinant vaccinia virus expressing a gene from the rabies virus was formulated into a bait for oral immunization against rabies of animals in the wild in parts of Europe. The applications of vaccinia recombinants are not limited to single antigens. It is possible to insert several genes, from different sources, into vaccinia, producing candidate vaccines that are capable of immunizing simultaneously against several different diseases.

However, vaccinia is not an ideal candidate vector for human vaccines. Despite its widespread use in the smallpox eradication campaign, it is not a very safe vaccine: for example, it is not suitable for people with skin conditions such as eczema, nor for people who are immunocompromised. However, there are other candidate carriers, notably attenuated strains of *Salmonella* and also the tuberculosis BCG (bacille Calmette–Guérin) vaccine. These are

very safe vaccines, and the manipulations involved are relatively straightforward. At a basic level, you simply clone the genes you want into a shuttle vector, including appropriate expression signals, and transform the chosen host. Ultimately there are further considerations. Antibiotic resistance genes are often used as a selectable marker for such manipulations, but for practical use as a vaccine, the inclusion of resistance genes is unlikely to be acceptable, and alternative strategies have to be adopted. Furthermore, plasmids are generally not completely stable, and insertion into the chromosome is preferred. One way of achieving this is to use a suicide plasmid with a bacteriophage integration system. The expression of the phage integrase then recombines the plasmid with the attachment site on the chromosome. There is a wide range of such potential vaccines at different stages of development.

11.4.2 Gene therapy

Knowing which mutation, in which gene, causes a disease is not a prerequisite for treating that disease. Many genetic diseases have been successfully treated for decades or even centuries thanks to insights into the pathophysiology of the conditions, or indeed because of the discovery of treatments by trial and error or even sheer coincidence. Nonetheless, discovering the exact genetic nature of a condition is a very important step. Firstly, as described in Chapter 9, it makes it possible to devise precise diagnostic methods to confirm who has, or is predisposed to developing, a specific condition. Secondly, knowing which gene is defective in the disease means you know which protein may be targeted for the development of new drugs to treat the disease. It also allows you to express the resulting protein (if any) and characterize it in the laboratory (see Chapter 7). Furthermore, it puts you in the position of being able to produce a mouse engineered to carry the same mutation, so that you have an animal model that closely approximates the human condition you wish to study (see above). All this will potentially be of great help in the development of improved treatments for the condition. However, there will probably always be conditions where treatment is not enough to alleviate the symptoms. This is the rationale for developing treatments involving gene transfer (*gene therapy*).

Gene therapy treatments can potentially take place by specifically targeting the diseased organs in an affected individual (*somatic gene therapy*), or by correcting the gene defect in gametes or fertilized eggs (*germline gene therapy*). There are strong reasons why most gene therapy research focuses on the former. Firstly, the genetic manipulation of embryos is inherently a very contentious issue, and the controversy is exacerbated by the fact that germline therapy would lead to a heritable alteration in the genetic material. Even the prenatal diagnosis and specific abortion of embryos carrying an affected gene has a far greater acceptance in society as a whole.

Secondly, there is an obvious need to treat people who are alive and suffering now, and indeed there will always be people born suffering from genetic disease. However, somatic gene therapy would only lead to a cure for that patient, and would not remove the underlying genetic cause – so there would still be the risk of the individual concerned passing the defective gene on to his/her offspring.

11.4.3 Viral vectors for gene therapy

The principle of delivering genes to the cells of a human being is not different from that of delivering genes to any other cells. Gene expression can be either *transient* or *permanent*. Delivery can be either directly into the affected tissues – *in vivo* – or into cells transiently removed from the patient's body– *ex vivo*. For *ex vivo* delivery, the methodology is virtually the same as for transfecting mammalian cell lines in culture – genes can be delivered using liposomes, calcium phosphate precipitation or electroporation. However, the methodology that is attracting the greatest interest by far, for the development of both *in vivo* and *ex vivo* methods, is the use of viral vectors

Amongst these are the *adenoviruses*. These are common causes of respiratory tract infections, especially pharyngitis (sore throat), but it is possible to produce defective viruses that do not cause disease, and to replace the deleted genomic regions with a cloned gene. The recombinant DNA will persist for some time within the infected cell, but not indefinitely. This limitation has both advantages and disadvantages. The advantage is that any negative effects are likely to be reversible, while the disadvantage is that reinfection is needed to sustain the effect. Nonetheless, trials with adenovirus vectors have been conducted, and one case where they were particularly promising was in the therapy of cystic fibrosis. Cystic fibrosis is a recessive disorder caused by a mutation in a chloride transporter gene, and manifests itself most strongly in respiratory problems, which both reduce the quality of life of individuals and limit their average lifespan to only 35 years on average. However, the use of adenoviruses as carriers received a major setback in 1999 when a patient in a trial of gene therapy for ornithine transcarboxylase deficiency died from multiple organ failure believed to have been caused by a severe immune response to the adenovirus. Due to these potential safety issues, the cystic fibrosis gene therapy consortium has discounted the use of an adenovirus-mediated approach, instead opting for non-viral delivery methods. Such approaches have had some success during pilot studies in humans, and a larger, clinical trial is currently underway that may lead to the first gene therapy treatment to reach general practice.

Another DNA virus that can be used for gene therapy is herpes simplex virus I (HSV-I). Unlike adenovirus infections, HSV-I infections persist throughout the lifetime of the patient, although they remain latent through

most of it. The preference of this virus for a particular cell type, i.e., nerve cells, makes it especially interesting for correcting mutations affecting these cells.

Retroviruses are the natural choice of vector for stable expression. The biology of these viruses, and their use as cloning vectors, was described in Chapter 7. The main feature that is important here is that, after infection, the RNA genome of the virus is copied into DNA by reverse transcriptase and integrated efficiently into the host genome. By replacing the genes needed to form new viral particles, two birds are killed with one stone – space is created for the introduction of the gene that is to be transferred, and the recombinant virus is prevented from reinfecting other cells.

Integration into the host genome is an effective way of ensuring reliable and lasting expression, which is also transferred to the progeny of the infected cell, and can therefore be used as a vector for the production of germline transgenics. Disadvantages include the limited space available for the gene that is to be inserted, and the fact that retroviruses are not very efficient in cells that do not divide, such as nerve cells. In addition, we have to remember that this class of viruses includes some notable pathogens, not only HIV but also other viruses that cause some forms of cancer in humans and animals. Although the vectors described are defective, their widespread use for gene therapy would require extensive checks to ensure that untoward effects do not occur by interaction with other naturally occurring retroviruses. However, the major problem is that random retrovirus insertion can lead to activation of neighbouring genes, which can include genes with oncogenic potential (i.e., they can give rise to the development of cancer). In one trial, two children who had been successfully treated for severe combined immunodeficiency disease, by gene therapy using a retroviral vector, subsequently developed leukaemia. Lentiviruses are claimed to be safer, and there have been reports of successful treatment of patients using these vectors.

Despite the problems encountered with both adenovirus and retrovirus vectors, gene therapy has enormous potential for the treatment of life-threatening genetic diseases. The potential does not stop there. In particular, there is considerable potential for novel cancer treatments by suppressing the activity of oncogenes, for example, by using *RNAi technology* (see Chapter 8 and earlier in this chapter) or by targeting toxic genes to cancerous tissue. The serious nature of these diseases provides a powerful argument for employing an approach that has such potential. But there are concerns, apart from the safety aspects. Partly these rest on a feeling that altering the DNA of a human being in this way is somehow unnatural, and that by developing this technology we are in some way in danger of letting the genie out of the bottle and paving the way for altering human genes that do not cause disease. Could the techniques developed for the praiseworthy objective of eliminating the suffering caused by diseases such as cystic fibrosis be exploited for other

ends? These are serious questions that demand a wider debate, and scientists have a role in trying to ensure that a broader audience is sufficiently educated about the issues to achieve this on a rational level.

11.5 Transgenic plants and their applications

11.5.1 Introducing foreign genes

Producing transgenic plants is much simpler than producing transgenic animals, as all plant cells, unlike animal cells, are *totipotent* and hence capable of regenerating an entire plant. In many cases, plant cells can be grown in culture, and manipulated, and whole plants regenerated from these individual cells. There are a variety of ways of introducing DNA into plant cells, including protoplast transformation/electroporation, microinjection, biolistics and transfer from *Agrobacterium tumefaciens*.

For protoplast transformation, the plant cell wall is digested with suitable enzymes to obtain protoplasts. These will take up DNA quite readily, and the efficiency can be improved by the use of electroporation (see Chapter 2), or by fusion with DNA-containing liposomes. Microinjection involves the direct injection of DNA into the nucleus of the plant cell, in much the same way as was described above for animal cells.

Biolistics involves coating tiny beads of gold with DNA and bombarding the plant cells at high velocity. This technique has also been used with other types of cell (see also the discussion of DNA vaccines in Chapter 7). For plants it has the advantage of being also applicable to embryonic plants, and so avoids the difficulty of regenerating some types of plants from individual cells.

The most distinct way of integrating a transgene into a plant cell involves the bacterium *Agrobacterium tumefaciens*. In nature this bacterium causes a type of plant tumour known as a *crown gall*. The pathogenic ability of these bacteria is associated with the presence of a particular type of plasmid, known as a *tumour-inducing* (Ti) plasmid. The tumour is produced by the transfer of plasmid DNA (or more specifically a 23 kb fragment of the plasmid, known as T-DNA) from the bacterium to the plant cells, where it becomes integrated into the plant chromosomal DNA. We can thus exploit this ability to integrate into plant cell DNA by hooking foreign genes up to the plasmid, so that the genes concerned are also integrated into the plant cell DNA along with the T-DNA.

Unfortunately, this is actually more difficult than it sounds. The natural Ti plasmids are very large (over 200 kb), and so it is not possible to clone genes directly into them. One strategy for overcoming this is illustrated in Figure 11.7. In this procedure, the initial construct in *E. coli* is made in an *intermediate vector*, a small plasmid containing a part of the T-DNA fragment

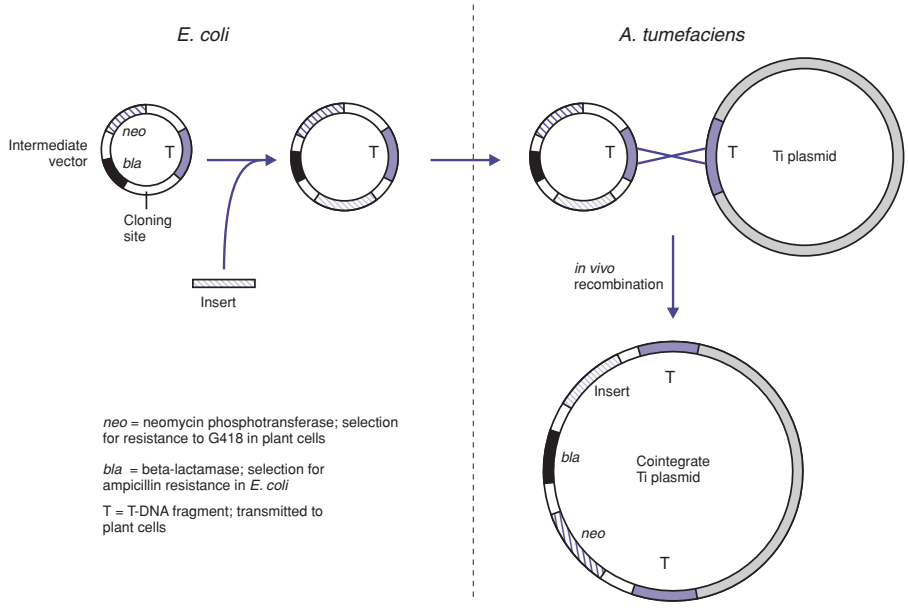


Figure 11.7 Cloning in plant cells using a Ti plasmid.

from a Ti plasmid. The recombinant intermediate vector can be transferred to *A. tumefaciens* by conjugation. Within the *A. tumefaciens* cell, a single recombination event between the homologous T-DNA sequences on the intermediate vector and the resident Ti plasmid will result in the incorporation of the entire intermediate vector plasmid into the T-DNA region of the Ti plasmid. The intermediate plasmid is unable to replicate in *A. tumefaciens*, and hence selection for antibiotic resistance will enable recovery of the bacterial cells carrying the cointegrate plasmid. Infection of a plant will now result in the transmission of the genes from the intermediate vector (including the cloned insert) into the plant cells along with the T-DNA.

Rather than infecting an intact plant, *A. tumefaciens* can be used to infect plant cells in culture and transformants can be isolated on a selective medium. All cells in a plant regenerated from a transformed cell will contain the cloned gene, and so the gene will be inherited in the subsequent progeny of that plant. However, plant cells infected with a wild-type Ti plasmid will not regenerate properly, because of the oncogenic effects of the T-DNA. But only short sequences at the ends of the T-DNA region are necessary for transmission to the plant cells. The oncogenic genes between these sequences can be removed, resulting in *disarmed* Ti plasmids, which will insert into plant DNA, but will not cause any of the adverse effects associated with the full Ti plasmid. Any DNA that is inserted between these two sequences, in place of the normal oncogenic genes, will be transmitted to the plant cells, and, in the

absence of the oncogenes, the plant cells can be regenerated into mature plants carrying the foreign DNA.

Just as in animal transgenesis, a suitable combination of promoter and gene is chosen to express the gene in the desired location and quantity. One promoter that is commonly used for this purpose is derived from the cauliflower mosaic virus. Clearly, producing genetically modified (GM) plants is much simpler than GM animals, and unsurprisingly the commercial exploitation of this fact has gone much further, and is already a reality.

11.5.2 Gene subtraction

In addition to adding genes to plants, it is possible to inactivate specific plant genes, by the use of RNAi, as we saw with animals previously. However, while the use of RNAi to knock down protein production in animals is well established, this is not true in plants; indeed, the whole field of miRNA control of gene expression is less well understood in plants compared to animals. In addition, there is one important difference between how siRNAs are handled in plant and animal cells: Unlike animal cells, plants possess an RNA-dependent RNA polymerase, which can copy the antisense strand of an siRNA, making more dsRNA. This amplifies the effect, and the dsRNA can spread throughout the plant by cell-to-cell transfer. The natural significance of this is that it plays a role in the development of plant resistance to viral infection, but it also has implications for the use of RNAi-based approaches to reduce protein production in plants.

Rather than use RNAi in plants, it is more common to use *antisense RNA*. This involves cloning all or part of the coding region of a relevant gene in an expression vector, but in the 'wrong' orientation. Stimulation of the promoter leads to the production of an RNA molecule that is complementary to that which codes for the required protein. Binding of the antisense RNA to the genuine mRNA probably interferes with translation, or leads to degradation by ribonuclease, so resulting in reduced production of the corresponding protein (Figure 11.8), although effects on transcription or processing of the mRNA are also possible. There are obvious similarities between this description of antisense RNA and the model presented earlier for RNA interference (see Chapter 8, and earlier in this chapter) in that both act via dsRNA to cause a reduction or loss of gene expression. However, this model for antisense RNA is stoichiometric – in other words, one molecule of antisense RNA will inhibit one molecule of mRNA – whereas the RNAi model is catalytic, so a small amount of the siRNA can result in complete removal of all the relevant mRNA. However, the distinction is not totally clear-cut.

One example of an application of this is in delaying the ripening, and spoilage, of tomatoes. In the later stages of ripening, an enzyme (polygalacturonase) is produced. This enzyme breaks down polygalacturonic acid in the

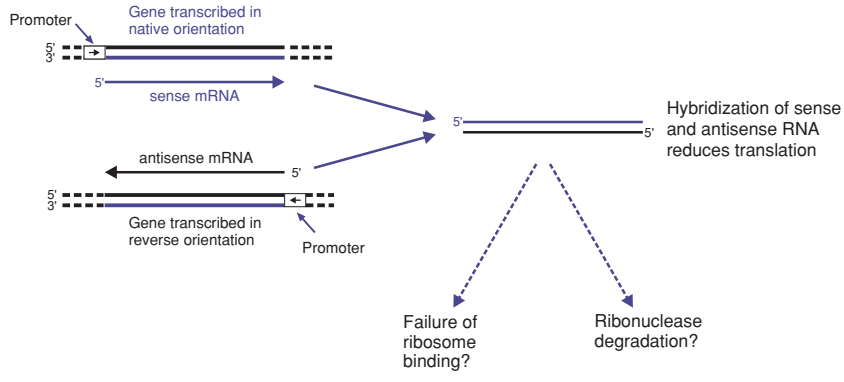


Figure 11.8 Gene subtraction in plants, using antisense RNA.

cell walls of the tomato and thus softens the tomato. Softening is desirable, but if it goes too far, you get a spoilt, squishy tomato. Introduction of part of the polygalacturonase gene, under the control of a cauliflower mosaic virus vector but in the 'wrong' orientation, produces an antisense RNA that binds to the normal polygalacturonase mRNA and greatly reduces the enzyme levels in the fruit. Enough enzyme is made to achieve a gradual softening, but the fruit can be stored for a much longer period before they spoil.

11.5.3 Applications

An extremely wide variety of modifications have been made, and the list is growing very rapidly. One of the applications that has received a lot of publicity is the production of plants that are resistant to insect attack. This most commonly involves the insertion of a gene from the bacterium *Bacillus thuringiensis* coding for an insecticidal toxin that is highly poisonous to certain groups of insects. Expression of this gene by a plant makes it resistant to insect attack, and thus substantially reduces the need for spraying the crop with insecticides. This is widely used in cotton and maize (corn) in particular, and has resulted in dramatic increases in yield.

The second widely publicized application is that of resistance to herbicides (principally glyphosate). This is often confused with insecticide resistance, but the cases are quite different. Expression of a herbicide-resistance gene in a crop plant enables the farmer to spray the crop with a broad-spectrum herbicide to eliminate weeds.

Other examples include genes that convey tolerance to environmental factors such as cold or salt, or influence texture, taste or colour (the latter important both in agricultural and horticultural crops). Drought tolerance is of especial importance. US farmers lose up to 15% of their maize crop because of water stress. A drought-tolerant maize strain that includes a gene derived

from the bacterium *Bacillus subtilis* is likely to be on the market soon. We have mentioned previously the potential of using plants for the production of pharmaceutically important proteins. Another important application consists of modifications that improve the nutritional content of the product. Many people in poorer countries rely extensively on a single food crop, such as rice, which often does not provide an adequate level of all the essential nutrients and vitamins in their diet. For example, vitamin A deficiency affects up to 250 million children worldwide, with some 500 000 per year becoming blind. The development of so-called 'Golden Rice', which accumulates provitamin A due to the incorporation of genes from various sources, could make a major contribution to alleviating vitamin A deficiency. This has been available in the laboratory since 1999, but its introduction into the field has been delayed by a variety of problems, both technological and regulatory.

In recent years, perhaps the most exciting development in the field of transgenic plants is the development of purple tomatoes by the John Innes Centre in Norwich, UK. Through genetic manipulation of red tomatoes, scientists were able to activate a pathway in the plants, causing overproduction of anthocyanins, which led to the purple coloration. What makes this finding exciting is the fact that anthocyanins have been shown to offer protection against certain cancers, cardiovascular disease and age-related degenerative diseases. Cancer-susceptible TP53−/− mice fed a diet enriched in the purple tomatoes had their lives extended by over 40 days, equivalent to several years in human terms. Thus, this development presents the real prospect of a genetically modified food that could significantly increase our protection against a range of life-threatening diseases.

11.6 Transgenics: a coda

In this chapter, as in the rest of the book, we have introduced the techniques and some examples of their applications. Although this only scratches the surface of a rapidly expanding subject, we hope it will have given you some understanding of the underlying concepts involved, of the advances in knowledge that have already been attained, and of their future possibilities. Some of these possibilities may be undesirable, but in many ways these techniques will play a major role in improvements in health, to use just one example. From your study of this book, you should be better placed to understand, and hopefully contribute to, the debate over the use and control of genetic modification.

Glossary

adaptor: a short, double-stranded oligonucleotide used to add sticky ends to a blunt-ended fragment, or to change one sticky end to a different one (see *linker*).

adenovirus(es): medium-sized DNA viruses, widespread in animals and birds. Derivatives are used as vectors in animal cells, and for gene therapy.

affinity chromatography: use of a specific *ligand*, attached to an insoluble matrix, to bind the required protein and enable its purification.

Agrobacterium tumefaciens: a bacterium that produces tumour-like growths (crown galls) on certain plants; used for transfer of DNA into plant cells.

alkaline phosphatase: enzyme that removes the terminal 5′ phosphate from DNA, and thus can be used to prevent ligation.

allele: one version of the DNA sequence at a specific *locus*.

allelic replacement: see *gene replacement*.

allosteric effect: alteration of the conformation of a protein, through the binding of a *ligand*, resulting in a change of the activity of a different site on the protein.

amplicon: the amplified product of a PCR reaction.

amplification: (i) an increase in plasmid copy number under changed growth conditions; (ii) multiplication of the number of bacteria or phages in a library (without increasing the number of independent clones); (iii) production of a large number of copies of a specific DNA fragment by PCR.

analogues: genes or proteins with similar functions but different sequences, arising by convergent evolution from different ancestral genes (see also *orthologues, paralogues*).

annealing: formation of double-stranded nucleic acid from two single-stranded nucleic acid molecules; see *hybridization*.

anticodon: the region of tRNA that pairs with the codon.

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

antisense RNA: RNA that is complementary to a specific mRNA, and which can interfere with translation.

autoradiography: detecting radioactively labelled material using X-ray film or a phosphoimager.

auxotroph: a mutant that requires the addition of one or more special supplements to its growth medium (cf. *prototroph*)

avidin: a protein with high affinity for biotin; used for detecting biotinylated probes.

back translation (reverse translation): predicting all possible nucleic acid sequences coding for a specific amino acid sequence.

bacterial artificial chromosome (BAC): a vector based on the *F plasmid*, used to clone very large DNA fragments.

bacteriophage: a virus that infects bacteria (often shortened to 'phage').

baculovirus: an insect virus used as a vector for eukaryotic gene expression.

bioinformatics: computer-based analysis of biomolecular data, especially large-scale datasets derived from genome sequencing.

biolistics: the use of high-velocity microprojectiles for introducing DNA into cells.

biotin: a small molecule that can be attached to dUTP and incorporated into a nucleic acid as a non-radioactive label. Detected with avidin.

BLAST: Basic Local Alignment Sequencing Tool; a computer program for searching databases for sequences similar to a query sequence.

blunt end: the end of a double-stranded DNA molecule in which both strands terminate at the same position, without any single-strand extension (as opposed to a *sticky* end); also known as a *flush* end.

bootstrap: a technique in molecular phylogeny for assessing the robustness of a phylogenetic tree.

box: a short sequence of bases in DNA conforming (more or less) to a consensus for that particular type of box; usually has a regulatory function.

broad host-range plasmid: a plasmid that is capable of replicating in a wide range of bacteria.

Caenorhabditis elegans: a nematode species widely used for studies of genetics and cell differentiation.

calf intestinal phosphatase (CIP): a commonly used type of alkaline phosphatase.

capsid: a viral component that surrounds and contains the nucleic acid.

cassette: a DNA sequence containing several genes and/or regulatory components that can be inserted as a unit into a vector, or that moves naturally as a unit. The term is used with similar meanings in other contexts.

cDNA: complementary (or copy) DNA; DNA synthesized using mRNA as a template in a reaction carried out by reverse transcriptase.

cDNA library: a collection of cDNA clones that together represent all the mRNA present in a sample at a particular time (cf. *genomic library*).

chaperone: a protein that affects the folding of other proteins or the assembly of complex structures.

chemiluminescence: light given off when a substrate interacts with a specific enzyme.

chimera (or chimaera): an animal or plant containing a mixture of cells with different genotypes.

chromatin: the higher-order structure resulting from the interaction of genomic DNA and protein complexes, such as nucleosomes. See also *euchromatin* and *heterochromatin*.

chromatin immunoprecipitation (ChIP): a technique for studying DNA–protein interactions within chromatin.

chromosome: a DNA-containing structure in the nucleus of a eukaryote. Also used to refer to the main genetic component of a prokaryote, especially when distinguishing it from a *plasmid*.

chromosome walking: a technique for identification of DNA regions adjacent to a known marker using the sequential hybridization of clones.

***cis*-acting**: a control region that influences genes on a more or less adjacent region of the same DNA molecule only, and has no effect on other DNA molecules (cf. *trans-acting*).

clone: one of an identical set of organisms, or cells, descended asexually from a single ancestor.

cloning: (i) for cell cultures, obtaining a homogeneous population of cells by repeated single colony isolation; (ii) for multicellular organisms, production of one or more whole organisms from a somatic cell. Also used to refer to obtaining copies of recombinant DNA carried by such a cell ('gene cloning'), and, by further extension, for the reactions used to make such recombinants.

Clustal: a computer program for multiple alignment of protein or DNA sequences.

CMV: cytomegalovirus; used as a source of expression signals in mammalian expression vectors. Also used (but not in this book) as abbreviation for cauliflower mosaic virus.

codon: a group of three bases in mRNA that codes for a single amino acid.

codon bias: difference in the frequency of occurrence of different *synonymous codons*.

codon usage: a measure of the relative use of different *synonymous codons*.

cohesive end: see *sticky end*.

colony/plaque lift: transfer of colonies or plaques from a plate onto a membrane for hybridization.

competent: a bacterial cell that is able to take up added DNA.

complementary strand: a nucleic acid strand that will pair with a given single strand of DNA (or RNA).

complementation: restoration of the wild-type phenotype by the introduction of a second DNA molecule, without recombination.

complexity: the number of independent clones in a library.

conjugation: transfer of genetic material from one bacterial cell to another by means of cell-to-cell contact.

consensus sequence: a sequence derived from a number of related, non-identical sequences, representing the nucleotides that are most commonly present at each position.

constitutive: describing gene expression in which the gene product is always formed irrespective of the presence of inducers or repressors (see *induction*, *repression*).

contig: a composite DNA sequence built up from a number of smaller overlapping sequences during a sequencing project.

convergent evolution: the evolution of organisms, structures or proteins with similar functions but descended from different sources.

copy number variations (CNVs): differences between individuals in the number of copies of a repetitive sequence.

***cos* site**: the sequence of bases of bacteriophage lambda that is cut asymmetrically during packaging, generating an unpaired sequence of 12 bases at each end of the phage DNA.

cosmid: a plasmid that contains the *cos* site of bacteriophage lambda. After introducing a large insert, the recombinant cosmid forms a substrate for *in vitro* packaging.

CpG island: a chromosomal region rich in CG dinucleotides.

ddNTP: any of the dideoxynucleotides ddATP, ddCTP, ddGTP or ddTTP.

denaturation: (a) reversible separation of the two strands of DNA by disruption of the hydrogen bonds (usually by heat or high pH); (b) disruption of the secondary and tertiary structure of proteins.

***de novo* sequencing**: sequencing the genome of a species for the first time (see *resequencing*).

deoxyribonuclease (DNase): an enzyme that degrades DNA.

dideoxynucleotide: a nucleotide lacking an OH at both the 2′ and 3′ positions, which therefore terminates DNA synthesis. Used in DNA sequencing by the *Sanger* method.

direct repeat: two identical or very similar DNA sequences, reading in the same direction (see also *inverted repeat*).

DNA ligase: an enzyme for joining DNA strands by formation of phosphodiester bonds.

dNTP: any of the deoxynucleotides dATP, dCTP, dGTP or dTTP.

domain: a region of a protein that folds into a semi-autonomous structure.

domain shuffling: a natural process in which various domains of a protein appear to have originated, evolutionarily, from different sources (see *mosaic proteins*).

dot (or slot) blot: a hybridization technique in which the samples containing the target nucleic acid are applied directly to the membrane in a regular pattern, usually defined by a manifold.

double cross-over: the integration of a gene into the chromosome by two recombinational events, either side of the gene (cf. *single cross-over*).

dsDNA/dsRNA: double-stranded DNA or RNA.

electrophoresis: separation of macromolecules (DNA, RNA or protein) by application of an electric field, usually across an agarose or acrylamide gel.

electroporation: the technique of inducing cells to take up DNA by subjecting them to brief electrical pulses.

embryonic stem cell: a *totipotent* cell derived from the embryo of an animal. Used in transgenics.

endonuclease: an enzyme that cuts a DNA molecule at internal sites (cf. *exonuclease*).

enhancer: a *cis*-acting sequence that increases the utilization of a promoter.

epigenetic: describing changes in a chromosome that do not alter the DNA sequence, but result in a stably inherited phenotype.

episome: a plasmid that can integrate into the chromosome. In eukaryotic systems, often used to emphasize the extrachromosomal state.

epitope: a portion of a protein that is recognized by a specific antibody.

EST: expressed sequence tag; a partial cDNA molecule, picked at random from a library and sequenced.

euchromatin: a complex of DNA and protein that is loosely wound and, generally, permissive for transcription.

eukaryote: a cell that has a discrete nucleus, bounded by a membrane; e.g., fungal, protozoan, plant and animal cells (cf. *prokaryote*).

exon: a coding sequence, part of a single gene and flanked by *introns*.

exonuclease: an enzyme that removes nucleotides from the ends of a DNA molecule (cf. *endonuclease*).

expression vector: a cloning vector designed for expression of the cloned insert using regulatory sequences present on the vector.

fingerprint: polymorphic pattern of bands on a gel for differentiating individual genomes.

FISH: fluorescent *in situ* hybridization; a technique that uses a probe labelled with a fluorescent dye to locate specific genes on a chromosome, or within a cell or tissue.

fluorogenic: a dye that emits visible light when stimulated by ultraviolet radiation.

flush end: see *blunt end*.

footprinting: identification of sequences that bind a specific protein, by visualization of the protection of positions that are protected from attack by DNase I.

F plasmid: a natural, low-copy plasmid of *E. coli*. See *bacterial artificial chromosome*.

frameshift: an insertion or deletion of bases other than in multiples of three. This changes the reading frame of protein synthesis beyond that point.

gel retardation: the reduction in the electrophoretic mobility of a DNA fragment due to protein binding (also known as gel shift or band shift).

gene chip: a high-density array, produced by the synthesis of oligonucleotides on a substrate.

gene knock-down: the reduction of gene expression by antisense RNA.

gene knock-in: see *gene replacement*.

gene knockout: the inactivation of a gene by homologous recombination or deletion (see also *gene replacement*).

gene replacement (gene knock-in): the replacement of a chromosomal gene by recombination with a homologous sequence, inactivated or otherwise modified *in vitro*.

gene subtraction: the use of antisense RNA for partial or controllable reduction in the activity of a specific gene.

gene therapy: the use of specific DNA to treat a disease, by correcting the genetic lesion responsible or by alleviating its effects.

genetic map: a map of the genetic structure of a genome, showing the relative position of known genes (cf. *physical map*).

genome: the entire genetic material of an organism. Sometimes (e.g., 'human genome') used in a more restrictive sense, to refer to the nuclear material only.

genome plasticity: larger-scale variations in the organisation of the genome, by rearrangements, inversions, etc., and by incorporation of DNA from other sources (see *horizontal gene transfer*).

genomic library: a collection of recombinant clones that together represent the entire genome of an organism. See also *cDNA library*.

genomics: the analysis of data derived from the complete DNA sequence of an organism.

genotype: the genetic make-up of an organism (cf. *phenotype*).

germline: cells involved in sexual reproduction, in a multicellular organism (cf. *somatic cells*).

GFP: green fluorescent protein; a naturally occurring protein that emits green light (or other colours in modified versions) when irradiated with UV; commonly used as a reporter.

hairpin: a region of DNA or RNA that contains a short inverted repeat, which can form a base-paired structure resembling a hairpin (similar to a *stem-loop structure*).

heterochromatin: a complex of DNA and protein that is tightly wound and, generally, non-permissive for transcription.

heteroduplex: a double-stranded nucleic acid molecule formed by base- pairing between two similar but not identical strands. Since the two strands are not identical, some regions will remain single-stranded.

heterologous probe: a probe that is similar but not identical to the target.

heterozygote: a diploid organism that carries two different versions of a specific gene (cf. *homozygote*).

HMM (Hidden Markov Model): a statistical technique with many applications in bioinformatics such as predicting open reading frames in genome sequences.

homologous recombination: a natural process involving the pairing of similar DNA molecules followed by breakage, crossing-over and rejoining of DNA strands.

homology: similarity in the sequence of two genes, from different organisms, that share a common evolutionary origin. Often used, more loosely (as in *homologous recombination*), to describe DNA molecules with a sequence that is sufficiently similar for complementary strands to hybridize, without evidence of common evolutionary origin or function.

homopolymer tailing: see *tailing*.

homozygote: a diploid organism in which the two copies of a specific gene are the same (cf. *heterozygote*).

horizontal gene transfer: the transfer of genetic material from one organism to another.

hybridization: the formation of double-stranded nucleic acid molecules by the production of hydrogen bonds between wholly or partially complementary sequences.

hybridoma: a cell line producing a *monoclonal antibody*.

hydrophilic: 'water-loving'; describing substances, or parts of a structure, that are polar and interact with water, and therefore tend to be exposed to an aqueous environment.

hydrophobic: 'water-hating'; describing substances, or parts of a structure, that are non-polar and do not interact with water, and therefore tend to remove themselves from an aqueous environment.

identity: nucleotides or amino acids that are exactly the same in two or more DNA or protein sequences, respectively (see also *similarity*).

inclusion body: an insoluble intracellular protein aggregate formed during high-level expression of a recombinant protein.

indel: insertion/deletion; a term used in a comparison of two genome sequences to indicate that there has been either an insertion in one or a deletion in the other.

induction: (i) increasing the synthesis of a gene product through a specific environmental change; (ii) applying a treatment to a lysogenic bacterium that results in the bacteriophage entering the lytic cycle.

insertional inactivation: destruction of the function of a gene by insertion of a foreign DNA fragment, either by transposition or by gene cloning.

insertion sequence: a DNA sequence that is able to insert itself, or a copy of itself, into another DNA molecule; it carries no information other than that required for transposition (see also *transposon*).

insertion vector: a lambda cloning vector into which DNA can be inserted at a single site (cf. *replacement vector*).

***in situ* hybridization**: a form of hybridization in which the immobilized target is part of a whole chromosome, or in some cases a whole cell (see also *FISH*).

intermolecular: describing an interaction between two different molecules.

intramolecular: describing an interaction between different parts of the same molecule.

intron: an intervening sequence in a eukaryotic gene; removed by splicing. See also *exon*.

inverse PCR: a version of PCR designed to amplify the sequences flanking the target rather than the target DNA itself.

inverted repeat: two identical or very similar DNA sequences, reading in opposite directions (see also *direct repeat*).

***in vitro* mutagenesis**: see *site-directed mutagenesis*.

***in vitro* packaging**: the assembly of mature bacteriophage particles *in vitro* by mixing suitable DNA with cell extracts that contain bacteriophage heads and tails and the enzymes needed for packaging.

IPTG: iso-propylthiogalactoside; an inducer of beta-galactosidase that is not hydrolysed by the enzyme.

island: a region of the genome with different base composition from that of the overall genome (see also *CpG island, pathogenicity island*).

isoelectric focusing: a technique for the separation of proteins by electrophoresis in a stable pH gradient, so that each protein will move to its *isoelectric point*.

isoelectric point: the pH at which a specific protein has no net overall charge.

isogenic: describing strains that are identical in their genetic composition; normally used to mean identical in all genes except the one being studied or used.

isoschizomers: restriction endonucleases that recognize the same nucleotide sequence (but do not necessarily cut in the same fashion).

IVET: *in vivo* expression technology; a procedure for identifying bacterial genes that are expressed during infection rather than during growth in the laboratory.

kilobase: a nucleic acid region that is 1000 bases long (abbreviated to kb).

Kozak sequence: a consensus sequence, in eukaryotes, adjacent to the translation start site.

labelling: adding a detectable signal to a nucleic acid probe or to a protein.

lambda: a temperate bacteriophage of *E. coli*, used as a cloning vector.

leader: a nucleotide sequence at the 5′ end of mRNA, before the start point for translation of the first structural gene. Often involved in regulating gene expression.

ligand: a (usually small) molecule that binds non-covalently to specific site(s) on a protein.

ligation: joining two DNA molecules using *DNA ligase*.

linkage: the degree to which two genes are inherited together.

linkage analysis: mapping the relative position of two genes (or other markers) by determining the extent to which they are co-inherited.

linkage disequilibrium: describes the situation when two genetic markers tend to be inherited together rather than randomly distributed in the progeny.

linker: a short double-stranded oligonucleotide with blunt ends and an internal restriction site, used to add sticky ends to a blunt-ended fragment (see *adaptor*).

locus: a broad term meaning a genetic site of any nature, including regulatory and structural genes.

lysogeny: a (more or less) stable relationship between a bacteriophage (*prophage*) and a host bacterium (lysogen).

lytic cycle: the multiplication of a bacteriophage within a host cell, leading to lysis of the cell and infection of other sensitive bacteria.

macroarray: a large set of DNA spots immobilized on a membrane. Used for comparative and differential studies of genomes and transcriptomes (cf. *microarray*).

MALDI-TOF: matrix-assisted laser desorption ionization time-of-flight; a mass spectrometry technique used in proteomics for identification of peptides.

mammalian two-hybrid (M2H): an interaction assay undertaken in mammalian cells *in vitro* to study protein–protein interactions (cf. *yeast 2H*).

mapping: the determination of the position of genes (*genetic map*), or of physical features such as restriction endonuclease sites (*physical map*).

massively parallel sequencing: various techniques by which the sequences of millions of relatively short fragments are determined simultaneously (see *next generation sequencing*).

melting: the separation of double-stranded DNA into single strands (see *denaturation*).

melting temperature (T_m): the temperature at which the two strands of a nucleic acid molecule separate (denature).
messenger RNA (mRNA): an RNA molecule used by ribosomes for translation into a protein.
metabolome: the cellular composition of metabolic intermediates and other small molecules.
metagenomics: the study of genome sequences of a mixture of organisms.
microarray: a large set of DNA spots immobilized on a glass slide. Used for comparative and differential studies of genomes and transcriptomes (cf. *macroarray*).
microinjection: the direct injection of DNA into the nucleus of a cell.
micro RNA (miRNA): small non-translated RNA molecules, involved in regulation of gene expression.
microsatellite: tandem repeats of a short sequence of nucleotides. Variation in the number of repeats causes *polymorphism*.
mobilization: transfer by conjugation of a non-conjugative plasmid in the presence of a conjugative plasmid.
modification: the alteration of the structure of DNA (usually by methylation of specific residues) so that it is no longer a substrate for the corresponding *restriction endonuclease*.
molecular beacons: a technique for real-time PCR in which a probe shows fluorescence only when annealed to a PCR product.
molecular clock: the rate of variation of a gene or an organism.
monoclonal antibody: a homogeneous population of identical antibody molecules produced by an immortalized lymphocyte cell line.
mosaic: a transgenic animal in which the cloned gene is present in only a proportion of the cells or tissues (cf. *chimera*).
mosaic protein: a protein composed of domains from different sources (see *domain shuffling*).
motif: a conserved sequence within a family of proteins indicating a specific function.
multiple cloning site: a short region of a vector containing a number of unique restriction sites into which DNA can be inserted (see *polylinker*).
multiplex PCR: a PCR reaction carried out with several pairs of primers, amplifying different sequences.
multipotent: describing cells that can differentiate into many but not all cell types (cf *pluripotent*, *totipotent*).
mutagenesis: the exposure of an organism to chemical or physical agents so as to induce alterations in the genetic material (see also *site-directed mutagenesis*).
mutant: a cell (or virus) with a change in its genetic material (cf. *mutation*).
mutation: an alteration in the genetic material (cf. *mutant*).

Neighbour Joining (NJ): a technique for constructing phylogenetic trees, e.g., from sequence data.

nested PCR: a technique for increasing the sensitivity and/or specificity of PCR, by using a second set of primers internal to the first pair.

next-generation sequencing (NGS): a term used to describe various modern sequencing techniques (see also *massively parallel sequencing*).

nick: a break in one strand of a double-stranded DNA molecule.

nonsense mutation: a base substitution creating a *stop codon* within the coding sequence, causing premature termination of translation.

Northern blot: a membrane with RNA molecules transferred from an electrophoresis gel for hybridization (cf. *Southern blot*).

nucleoside: a purine or pyrimidine base linked to the 1′ position of ribose or deoxyribose (the latter should strictly be a deoxynucleoside).

nucleotide: a nucleoside with a phosphate group at the 5′ position.

oligonucleotide: a short nucleic acid sequence (usually synthetic).

open reading frame (ORF): a nucleic acid sequence with a reading frame that contains no stop codons; it therefore defines a potentially translated polypeptide.

operational taxonomic unit (OTU): in the construction of phylogenetic trees, a group of organisms, or sequences, that are treated as a single organism or sequence.

operator: a region of DNA to which a repressor protein binds to switch off expression of the associated gene. Usually found adjacent to, or overlapping with, the promoter.

operon: a group of contiguous genes (in bacteria) that are transcribed into a single mRNA, and hence are subject to coordinated induction/repression.

ordered library: a collection of clones containing overlapping fragments, in which the genomic order of the fragments has been determined.

origin of replication: a position on a DNA molecule at which replication starts. Most commonly used to mean the part of a plasmid that is necessary for replication.

orthologues: genes or proteins with the same function and similar sequences, having arisen from a common evolutionary ancestral gene (see also *analogues, paralogues*).

outgroup: in phylogenetic analysis, a sequence (or organism) that is distantly related to those being studied. Enables the production of a *rooted tree*.

P1: a bacteriophage that infects *E. coli*. Used as the basis for some cloning vectors.

packaging: the process of incorporating DNA into a bacteriophage particle (see also *in vitro packaging*).

packaging limits: the range of DNA sizes that can be packaged into a specific bacteriophage particle.

palindrome: a sequence that reads the same in both directions (on the complementary strands).

parallel sequencing: sequencing techniques in which millions of sequences are read simultaneously (see also *next-generation sequencing*).

paralogues: genes or proteins with similar sequences but different (although possibly related) functions, arising from a common ancestral gene.

partial digest: cutting DNA with a restriction endonuclease using conditions under which only a fraction of available sites are cleaved.

partitioning: the distribution of copies of a plasmid between daughter cells at cell division.

pathogenicity island: a DNA region (in bacteria) carrying virulence determinants; often with a different base composition from the remainder of the chromosome.

P element: a transposable element in *Drosophila*, used in the construction of cloning vectors, and in transposon mutagenesis.

phage: see *bacteriophage*.

phage display: a technique for expressing cloned proteins on the surface of a bacteriophage.

pharming: the production of recombinant proteins from genetically modified farm animals.

phenotype: the observable characteristics of an organism (cf. *genotype*).

phylogenetic tree: a graphical display of the relatedness between organisms or sequences, used to infer possible evolutionary relationships.

phylogeny: the study of the evolutionary relationships amongst groups of organisms or sequences.

physical map: a map of the physical structure of a genome, e.g., showing restriction sites, positions of specific clones, or ultimately the complete sequence (cf. *genetic map*).

pilus: in bacteria, a filamentous appendage on the surface, often involved in *conjugation*.

plaque: a region of clearing, or reduced growth, in a bacterial lawn, as a result of phage infection.

plasmid: an extrachromosomal genetic element, capable of autonomous replication.

pluripotent: describing cells that can differentiate into all the cell lineages within an embryo, but cannot form extraembryonic lineages (cf. *multipotent*, *totipotent*).

plus and minus strands: mRNA is defined as the plus (sense) strand, and the complementary sequence as the minus (antisense) strand. DNA sequences maintain the same convention, so it is the minus strand of DNA that is transcribed to yield the mRNA (plus strand).

point mutation: an alteration (or deletion/insertion) of a single base in the DNA.

polar mutation: a mutation in one gene that affects the expression of others (e.g., genes downstream in an operon). The phenotypic effect may not be directly caused by the original mutation.

polyadenylation: a natural process (mainly in eukaryotes) that produces a long string of adenyl residues at the 3′ end of the mRNA. (Should strictly be 'polyadenylylation' but 'polyadenylation' is commonly used.)

polycistronic mRNA: messenger RNA coding for several proteins (see *operon*).

polylinker: a synthetic oligonucleotide containing several restriction sites (see *multiple cloning site*).

polymerase chain reaction (PCR): enzymatic amplification of a specific DNA fragment, using repeated cycles of denaturation, primer annealing and chain extension. (Note that there are other methods for amplification of specific DNA sequences, not covered in this book.)

polymorphism: the stable, multigenerational existence of multiple alleles at a gene locus. Generally describes variations that occur in a population at a frequency higher than an arbitrarily defined frequency.

positional cloning: using the mapped position of a gene to obtain a clone carrying it.

post-genomics: studies that go beyond the genome sequence, including genome-wide studies of the products of transcription (*transcriptomics*) and translation (*proteomics*).

post-translational modification: modification of the structure of a polypeptide after synthesis, e.g., by phosphorylation, glycosylation or proteolytic cleavage.

primary structure: the base sequence of a nucleic acid, or the amino acid sequence of a protein.

primer: a specific oligonucleotide, complementary to a defined region of the template strand, from which new DNA synthesis will occur.

primer walking: a technique in DNA sequencing whereby information from one sequence run is used to design another primer to extend the sequence determined.

probe: a nucleic acid molecule that will hybridize to a specific target sequence.

prokaryote: a cell that does not have a discrete nucleus bounded by a membrane (cf. *eukaryote*). Includes bacteria and the evolutionarily distinct Archaea. For simplicity, we have ignored the Archaea, and have (largely) used the term 'bacteria' instead.

promoter: a region of DNA to which RNA polymerase binds in order to initiate transcription.

promoter probe vector: a vector carrying a promoterless reporter gene, so that inserts carrying a promoter can be detected.

proofreading: the ability of DNA polymerase to check, and correct, the accuracy of the newly made sequence.

prophage: the repressed form of bacteriophage DNA in a lysogen; it may be integrated into the chromosome or exist as a plasmid.

protein engineering: altering a gene so as to produce defined changes in the properties of the encoded protein.

proteome: the complete content of different proteins in a cell (cf. *genome, transcriptome*).

proteomics: the global study of protein expression in an organism.

protoplast: the cell unit formed by complete removal of the cell wall, using osmotically stabilized conditions.

prototroph: a nutritionally wild-type organism that does not need any additional growth supplement (cf. *auxotroph*).

pseudogene: a gene, usually recognizable as a copy of another gene, that does not produce a protein product because of the presence of numerous changes (often in-frame stop codons), or because it is truncated.

pulsed-field gel electrophoresis (PFGE): a technique whereby large DNA molecules are separated by application of an intermittently varying electric field.

purine: one of the two types of bases in nucleic acids (adenine, guanine). See also *pyrimidine*.

pyrimidine: one of the two types of bases in nucleic acids (cytosine and thymine in DNA; cytosine and uracil in RNA). See also *purine*.

random primers: synthetic oligomeric nucleotides (usually hexamers) designed to act as primers for DNA synthesis at multiple sites.

reading frame: a nucleic acid sequence is translated in groups of three bases (*codons*). There are three possible ways of reading the sequence (in one direction), depending on where you start. These are the three reading frames.

real-time PCR: a PCR technique, using fluorescent dyes, that makes it possible to monitor the progress of the amplification as it occurs.

recombinant: (i) an organism containing genes from different sources, either by natural horizontal gene transfer, or as a result of gene cloning; (ii) a DNA molecule formed by joining two different pieces of DNA.

recombination: (i) the production of new strains by mating two genetically distinct parents; (ii) the generation of new DNA molecules by breaking and rejoining the original molecules.

relaxation: the conversion of supercoiled circular plasmid DNA to an open circular form, usually by nicking one strand.

replacement vector: a lambda cloning vector in which a piece of DNA (the *stuffer fragment*) can be removed and replaced by the cloned fragment (cf. *insertion vector*).

replica plating: the transfer of colonies from one plate to others, in the same position, for differential screening.

replication: the synthesis of a copy of a DNA molecule.

replication slippage: errors introduced during replication of repeated DNA sequences, which increase or reduce the number of copies of the repeated sequence

replicon: a DNA molecule that can be replicated in a specific host cell; also used to refer to the replication control region of a plasmid (cf. *origin of replication*).

reporter gene: a gene that codes for a readily detected protein for study of the regulation of gene expression.

repression: (i) a reduction in the transcription of a gene, usually due to the action of a repressor protein; (ii) the establishment of lysogeny with temperate bacteriophages.

repressor: a protein that binds to a specific DNA site to switch off transcription of the associated gene.

resequencing: sequencing all or part of the genome of individual organisms for which a reference genome sequence is already available (see *de novo sequencing*).

response element: a region of DNA to which a transcription factor can bind and influence the rate of transcription of a gene.

restriction: reduction or prevention of phage infection through the production of *restriction endonucleases* that degrade foreign DNA. See also *modification*.

restriction endonuclease: an enzyme that recognizes specific DNA sequences and cuts the DNA, usually at the recognition site.

restriction fragment length polymorphism (RFLP): variation between individuals or strains in the size of specific restriction fragments; used for strain typing, and for locating particular genes.

restriction mapping: determination of the position of restriction endonuclease recognition sites on a DNA molecule.

retrotransposon: a transposon that transposes by means of an RNA intermediate (cf. *retrovirus*).

retrovirus: a virus with an RNA genome that is copied, by reverse transcriptase, into DNA after infection.

reverse genetics: making specific changes to the DNA and then examining the phenotype; contrasts with classical genetics in which you select mutants by their phenotype and then study the nature of the mutation

reverse hybridization: hybridization in which the specific nucleic acid fragments are fixed to a substrate and hybridized with a labelled mixture of nucleic acids such as a cell extract.

reverse transcription: the production of cDNA from an RNA template, by reverse transcriptase.

reverse translation: see *back translation*.

ribonuclease: an enzyme that digests RNA.

ribosome binding site: the region on an mRNA molecule to which ribosomes initially attach, in bacteria.

RNA interference: reducing expression of a specific gene through the action of dsRNA (see *siRNA*).

RNA polymerase: an enzyme that synthesizes RNA, generally using a DNA template.

rooted tree: a phylogenetic tree in which the position of the putative common ancestor can be inferred (see *outgroup*, *unrooted tree*).

RT-PCR: reverse transcription PCR; a technique for producing an amplified DNA product from an mRNA template.

Sanger sequencing: a DNA sequencing method using *dideoxynucleotides*.

scaffold: in genome sequencing, a known sequence of a closely related organism (usually another individual of the same species) that enables pieces of sequence to be arranged in the right order. Hence also used to describe a series of contigs that are in the right order but not necessarily connected in one continuous stretch of sequence.

SDS-PAGE: sodium dodecyl sulphate polyacrylamide gel electrophoresis; a form of gel electrophoresis in which proteins are separated, according to molecular weight, in the presence of sodium dodecyl sulphate.

secondary structure: the spatial arrangement of amino acids in a protein, or of bases in nucleic acid.

selectable marker: a gene that causes a phenotype (usually antibiotic resistance) that can be readily selected.

Shine–Dalgarno sequence: see *ribosome binding site*.

short tandem repeats (STRs): see *microsatellites*.

shotgun cloning: the insertion of random fragments of DNA into a vector.

shotgun sequencing: a genome sequencing strategy involving the sequencing of large numbers of random fragments; the individual sequences are subsequently assembled by a computer.

shuttle vector: a cloning vector that can replicate in two different species, one of which is usually *E. coli*. It facilitates cloning genes in *E. coli* initially and subsequently transferring them to an alternative host without needing to re-clone them.

sigma factor: a polypeptide that associates with (bacterial) RNA polymerase core enzyme to determine promoter specificity.

signal peptide: an amino acid sequence at the amino terminus of a secreted protein; involved in conducting the protein through the membrane, or targeting it to specific cellular locations.

signal transduction: the process whereby extracellular conditions alter the conformation of a transmembrane protein, which in turn alters the regulation of metabolic pathways within the cell.

silent mutation: a change in the DNA sequence that has no effect on the phenotype of the cell.

similarity: amino acids in two or more protein sequences that occur at the same point and confer a similar biological function – e.g., lysine and arginine, which are both basic amino acids (see also *identity*).

single cross-over: recombination at a single position, e.g., between a plasmid and chromosome, leading to integration of the whole plasmid (cf. *double cross-over*).

single nucleotide polymorphism (SNP): a single base difference in the DNA sequence of some individuals in a population.

siRNA: short interfering RNA; small dsRNA molecules, responsible for *RNA interference* (see also *micro RNA*).

site-directed mutagenesis: a technique for specifically altering (*in vitro*) the sequence of DNA at a defined point.

somatic cell: any cell (in a multicellular organism) other than those involved in sexual reproduction (cf. *germline*).

Southern blot: a membrane with DNA fragments transferred from an electrophoresis gel, preparatory to hybridization (cf. *Northern blot*).

splicing: removal of introns from RNA and joining together of the exons.

start codon: the position at which protein synthesis starts, usually the codon AUG.

stem cell: a cell that is capable of differentiating into various types of cells. See *embryonic stem cell*.

stem-loop structure: a nucleic acid structure formed from two complementary sequences that can fold so that these sequences are paired (stem), with the region between them forming a loop of unpaired bases (cf. *hairpin*).

sticky end: the end of a DNA molecule where one strand protrudes beyond the other; also known as a c*ohesive end.* See *blunt end*.

stop codon: a codon that has no corresponding tRNA, and that signals the end of a region to be translated.

stringency: conditions affecting the *hybridization* of single-stranded DNA molecules. Higher stringency (higher temperature and/or lower salt concentration) demands more accurate pairing between the two molecules.

structural genes: genes coding for enzymes (or sometimes other products), as distinguished from regulatory genes.

stuffer fragment: a piece of DNA that is removed from a vector such as a replacement lambda vector, and replaced by the cloned DNA fragment.

suicide plasmid: a vector that is unable to replicate in a specific host. Maintenance of the selected marker requires integration into the chromosome.

supercoiling: coiling of a double-stranded DNA helix around itself.

superinfection immunity: resistance of a lysogen to infection by the same (or related) bacteriophage.

suppression: the occurrence of a second mutation that negates the effect of the first without actually reversing it.

SV40: simian virus 40; a small virus isolated from monkeys. It is used as a vector, and also as a source of expression signals in mammalian expression vectors.

SYBR Green: a dye that fluoresces when bound to dsDNA. Used for *real-time PCR*.

synonymous codons: different codons that code for the same amino acid.

synteny: conservation of the overall organisation of genes in the genomes of related organisms. Also used to refer to the comparison of overall genome structure.

systems biology: the integrated study of all the interacting networks within (or between) cells.

T7: a lytic bacteriophage of *E. coli*. The requirement of T7 RNA polymerase for a highly specific promoter is used in several contexts.

TA cloning: a method for cloning PCR products, exploiting the tendency of *Taq* polymerase to add a non-specific adenyl residue to the 3′ end of the new DNA strand.

tagging: constructing a recombinant so that the protein formed has additional amino acids at one end, facilitating purification by affinity chromatography, targeting the protein to specific destinations, or recognition by specific antibodies.

tailing: adding a number of nucleotides to the 3′ end of DNA, using *terminal transferase*.

tandem repeat: the occurrence of the same sequence two or more times, directly following one another.

***Taq* polymerase**: a thermostable DNA polymerase, commonly used for PCR.

TaqMan: a technique for real-time PCR in which fluorescence develops as a consequence of destruction of the labelled probe during DNA synthesis.

T DNA: the part of the *Ti plasmid* that is transferred into plant cells.

telomere: the end region of a eukaryotic chromosome containing sequences that are replicated by a special process, counteracting the tendency of linear molecules to be shortened during replication.

temperate: describing a bacteriophage that is able to enter *lysogeny*.

template: a single strand of nucleic acid used for directing the synthesis of a complementary strand.

terminal transferase: an enzyme that adds nucleotides, non-specifically, to the 3′ ends of DNA (see *tailing*).

terminator: a site at which transcription stops.

tertiary structure: the folding of secondary structure components of a protein.

Ti plasmid: tumour-inducing plasmid; a plasmid from *Agrobacterium tumefaciens*, used as a vector in genetic manipulation of plants (see *T-DNA*).

topoisomerase: an enzyme that alters the supercoiling of DNA by breaking and rejoining DNA strands.

totipotent: describing a cell that is capable of giving rise to all types of cell within a whole animal or plant (cf. *multipotent*, *pluripotent*).

***trans*-acting**: describing a gene that influences non-adjacent regulatory DNA sequences, through the production of a diffusible protein (cf. *cis-acting*).

transcription: the synthesis of RNA according to a DNA template.

transcriptional fusion: a recombinant construct in which a promoterless insert is transcribed from a promoter on the vector (cf. *translational fusion*).

transcriptome: the complete mRNA content of a cell (cf. *genome, proteome*).

transduction: the bacteriophage-mediated transfer of genes from one bacterium to another.

transfection: the introduction of viral nucleic acids into a cell.

transformation: (i) the introduction of extraneous DNA into a cell; (ii) the conversion of an animal cell into an immortalized, tumour-like cell.

transgenic: describing an animal or plant possessing a cloned gene in all its cells, so that the introduced gene is inherited by the progeny of that animal or plant.

translation: the synthesis of proteins/polypeptides by ribosomes acting on a mRNA template.

translational fusion: a recombinant construct in which an insert lacking a translation start site is joined (in frame) to a fragment carrying translational signals (cf. *transcriptional fusion*).

transposon: a DNA element carrying recognizable genes (e.g., antibiotic resistance) that is capable of inserting itself into the chromosome or a plasmid, independently of the normal host cell recombination machinery (cf. *insertion sequence*).

transposon mutagenesis: disruption of genes by insertion of a transposon.

two-dimensional gel electrophoresis: the separation of a complex mixture of proteins by a combination of *isoelectric focusing* and *SDS-PAGE*.

Ty element: a retrotransposon, used for transposon mutagenesis in yeast.

universal primers: sequencing primers derived from the sequence of the vector (pUC series); any insert can be sequenced with the same primers

unrooted tree: a phylogenetic tree in which the position of the putative common ancestor cannot be inferred (cf. *rooted tree*).

UPGMA: Unweighted Pair Group Method with Arithmetic means; a technique for constructing phylogenetic trees, e.g., from sequence data.

upstream activator sequence: a sequence, occurring upstream from the promoter, that is required for efficient promoter activity.

vaccinia: smallpox vaccine virus, used as a vector for recombinant vaccine construction.

variable number tandem repeats (VNTR): repeated sequences, where the number of copies of the repeat varies between individuals or strains. Used in bacterial typing. See also *microsatellites*.

vector: a replicon (plasmid or phage) into which extraneous DNA fragments can be inserted, forming a recombinant molecule that can be replicated in the host cell.

Western blot: a membrane with proteins transferred from an electrophoresis gel, usually for detection by means of antibodies.

wildtype: the normal, naturally occurring form of an organism (cf. *mutant*).

X-gal: 5-bromo-4-chloro-3-indolyl-β-D-galactoside; a chromogenic substrate for beta-galactosidase.

yeast artificial chromosome (YAC): a vector that mimics the structure of a yeast chromosome; used to clone very large DNA fragments.

yeast centromere plasmid (YCp): a yeast vector containing a centromere; replicates at low copy number.

yeast episomal plasmid (YEp): an autonomously replicating vector based on a yeast plasmid (the '2 μm circle').

yeast integrating plasmid (YIp): a yeast vector that relies on integration into the yeast chromosome.

yeast one/two hybrid (Y1H/Y2H): an interaction assay undertaken in yeast to study DNA–protein (Y1H) or protein–protein (Y2H) interactions (cf. mammalian two-hybrid).

Bibliography

General books

Brown, T.A. (2010) *Gene Cloning and DNA Analysis – an Introduction*, 6th edn. John Wiley and Sons, Inc.

Dale, J.W. and Park, S.F. (2010) *Molecular Genetics of Bacteria*, 5th edn. John Wiley and Sons, Inc.

Gibson, G. (2008) *A Primer of Genome Science*, 3rd edn. Sinauer Associates.

Lewin, B., Krebs, J.E., Goldstein, E.S. and Kilpatrick, S.T. (2009) *Lewin's Essential Genes*, 2nd edn. Jones and Bartlett.

Latchman, D.S. (2010) *Gene Control*. Garland Science.

Primrose, S.B. and Twyman, R. (2011) *Principles of Gene Manipulation and Genomics*, 8th edn. Blackwell Publishing.

Norman, R.I. and Lodwick, D. (2007) *The Flesh and Bones of Medical Cell Biology*. Elsevier.

Laboratory manuals

Sambrook, J. and Russell, D.W. (2006) *Condensed Protocols from Molecular Cloning*. Cold Spring Harbor Laboratory Press.

King, N. (ed.) (2010) *RT-PCR Protocols*, 2nd edn. Humana Press.

Also, many commercial suppliers of molecular biological materials produce very informative catalogues and methods sheets, usually available on-line.

Special topics

Lemey, P., Salemi, M. and Vandamme, A-M. (2009) *The Phylogenetic Handbook – A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, 2nd edn. Cambridge University Press.

Lesk, A.M. (2008) *Introduction to Bioinformatics*, 3rd edn. Oxford University Press.

Strachan, T. and Read A.P. (2010) *Human Molecular Genetics*, 4th edn. Taylor & Francis.

Sudbery, P. (2009) *Human Molecular Genetics*, 3rd edn. Pearson Education.

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.

Websites

Note that this is only a small selection of the available sites. Many other sites not mentioned are also valuable resources. Note also that website addresses may change, although the ones selected are likely to be reasonably stable, at least in their home pages. Within each site, the structure is likely to be more fluid, so only the home pages are listed. You will need to explore the site from that point to find the facilities that you want.

See Boxes 5.1 and 8.1 for further and more specific sites.

Address	Organisation	Facilities include
http://www.ebi.ac.uk/	European Bioinformatics Institute (EBI)	Databases: EMBL, UniProt, TrEMBL Tools: BLAST, FASTA, MPsrch, Clustal
http://www.ncbi.nlm.nih.gov/	National Center for Biotechnology Information (NCBI)	GenBank database plus search and comparison tools
http://www.ddbj.nig.ac.jp/	DNA Data Bank of Japan	DDBJ database plus search and comparison tools
http://www.isb-sib.ch/ http://www.expasy.org/	Swiss Institute of Bioinformatics (SIB)	UniProt database; PROSITE; SWISS-2DPAGE ExPASy (Expert Protein Analysis System) proteomics server
http://www.sanger.ac.uk/	Wellcome Trust Sanger Institute (WTSI)	Genome sequence data
http://www.ensembl.org http://genome.ucsc.edu	EMBL, EBI, WTSI University of California Santa Cruz	Genome browsers
http://www.HapMap.org	International HapMap Project	Genotype database
http://bioinformatics.weizmann.ac.il/cards/	GeneCards database, Weizmann Institute	Human gene information
http://www.ebi.ac.uk/arrayexpress/	EBI	Functional genomics database

Address	Organisation	Facilities include
http://www.hmdb.ca/	Genome Canada and Genome Alberta	Human Metabolome database
http://www.reactome.org/ReactomeGWT/	EMBL, NIH, CSH, NY School of Medicine, Ontario Institute of Cancer Research	Biological pathway database
http://jjj.biochem.sun.ac.za/	Multicentre collaboration	Systems modelling database
http://www.jcvi.org/	J. Craig Venter Institute	Genomics
www.genomesonline.org		Sequenced genomes

Index

From Genes to Genomes: Concepts and Applications of DNA Technology, Third Edition.
Jeremy W. Dale, Malcolm von Schantz and Nick Plant.