Chapter 12

Studying Genomes

Chapter contents

CHAPTER CONTENTS

At the start of the 21st century the emphasis in molecular biology shifted from the study of individual genes to the study of entire genomes. This change in emphasis was prompted by the development during the 1990s of methods for sequencing large genomes. Genome sequencing predates the 1990s—we saw in Chapter 10 how the first genome, that of the phage ϕX174, was completed in 1975—but it was not until 20 years later, in 1995, that the first genome of a free-living organism, the bacterium *Haemophilus influenzae*, was completely sequenced. The next five years were a watershed with the genome sequences of almost 50 other bacteria published, along with complete sequences for the much larger genomes of yeast, fruit fly, *Caenorhabditis elegans* (a nematode worm), *Arabidopsis thaliana* (a plant), and humans. Today, the sequencing of bacterial genomes has become routine, with over 900 completed, and almost 100 eukaryotic genomes have also been sequenced.

Genome sequencing has led to the development of a new area of DNA research, loosely called **post-genomics** or **functional genomics**. Post-genomics includes the use of computer systems in **genome annotation**, the process by which the genes, control sequences, and other interesting features are identified in a genome sequence, as well as computer-based and experimental techniques aimed at determining the functions of any unknown genes that are discovered. Post-genomics also encompasses techniques designed to identify which genes are expressed in a particular type of cell or tissue, and how this pattern of genome expression changes over time.

12.1 Genome annotation

Once a genome sequence has been completed, the next step is to locate all the genes and determine their functions. It is in this area that **bioinformatics**, sometimes referred to

Gene Cloning and DNA Analysis: An Introduction. 6th edition. By T.A. Brown. Published 2010 by Blackwell Publishing.

as molecular biology *in silico*, is proving of major value as an adjunct to conventional experiments.

Genome annotation is a far from trivial process, even with genomes that have been extensively studied by genetic analysis and gene cloning techniques prior to complete sequencing. For example, the sequence of the yeast *Saccharomyces cerevisiae*, one of the best studied of all organisms, revealed that this genome contains about 6000 genes. Of these, some 3600 could be assigned a function either on the basis of previous studies that had been carried out with yeast or because the yeast gene had a similar sequence to a gene that had been studied in another organism. This left 2400 genes whose functions were not known. Despite a massive amount of work since the yeast genome was completed in 1996, the functions of many of these **orphans** have still not been determined.

12.1.1 Identifying the genes in a genome sequence

Locating a gene in a genome sequence is easy if the amino acid sequence of the protein product is known, allowing the nucleotide sequence of the gene to be predicted, or if the corresponding cDNA has been previously sequenced. But for many genes there is no prior information that enables the correct DNA sequence to be recognized. How can these genes be located in a genome sequence?

Searching for open reading frames

The DNA sequence of a gene is an **open reading frame (ORF)**, a series of nucleotide triplets beginning with an initiation codon (usually but not always ATG) and ending in a termination codon (TAA, TAG, or TGA in most genomes). Searching a genome sequence for ORFs, by eye or more usually by computer, is therefore the first step in gene location. When carrying out the search it is important to remember that each DNA sequence has six reading frames, three in one direction and three in the reverse direction on the complementary strand (Figure 12.1).

The key to the success of **ORF scanning** is the frequency with which termination codons appear in the DNA sequence. If the DNA has a random sequence and a GC content of 50%, then each of the three termination codons will appear, on average, once every $4^3 = 64$ bp. This means that there should not be many ORFs longer than 30–40 codons in random DNA, and not all of these ORFs will start with ATG. Most genes are much longer than this: the average lengths are 317 codons for *Escherichia coli*, 483 codons for *S. cerevisiae*, and approximately 450 codons for humans. ORF scanning, in its simplest form, therefore takes a figure of 100 codons as the shortest length of a putative gene and records positive hits for all ORFs longer than this.

With bacterial genomes, simple ORF scanning is an effective way of locating most of the genes in a DNA sequence. Most bacterial genes are much longer than 100 codons

Figure 12.1
A double-stranded DNA molecule has six reading frames.

Figure 12.2

The typical result of a search for ORFs in a bacterial genome. The arrows indicate the directions in which the genes and spurious ORFs run.

in length and so are easily recognized (Figure 12.2). With bacteria the analysis is further simplified by the fact that most genes are closely spaced with very little intergenic DNA between them. If we assume that the real genes do not overlap, which is true for most bacterial genomes, then it is only in these short intergenic regions that there is a possibility of mistaking a spurious ORF for a real gene.

Simple ORF scans are less effective at locating genes in eukaryotic genomes

Although ORF scans work well for bacterial genomes, they are less effective for locating genes in eukaryotic genomes. This is partly because there is substantially more intergenic DNA in a eukaryotic genome, increasing the chances of finding spurious ORFs, but the main problem is the presence of introns (see Figure 11.1). If a gene contains one or more introns, then it does not appear as a continuous ORF in the genome sequence. Many exons are shorter than 100 codons, some fewer than 50 codons, and continuing the reading frame into an intron usually leads to a termination sequence that appears to close the ORF (Figure 12.3). In other words, many of the genes in a eukaryotic genome are not long ORFs, and simple ORF scanning cannot locate them.

Finding ways of locating genes by inspection of a eukaryotic sequence is a major challenge in bioinformatics. Three approaches are being followed:

- **Codon bias** can be taken into account. Not all codons are used equally frequently in the genes of a particular organism. For example, leucine is specified by six codons (TTA, TTG, CTT, CTC, CTA, and CTG), but in human genes leucine is most frequently coded by CTG and is only rarely specified by TTA or CTA. Similarly, of the four valine codons, human genes use GTG four times more frequently than GTA. The biological reason for codon bias is not understood, but all organisms have a bias, which is different in different species. Real exons display this bias, whereas chance series of triplets usually do not.

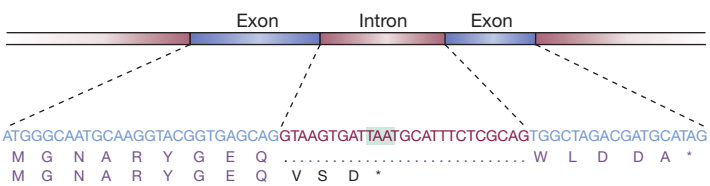


Figure 12.3

ORF scans are complicated by introns. The nucleotide sequence of a short gene containing a single intron is shown. The correct amino acid sequence of the protein translated from the gene is given immediately below the nucleotide sequence, using the single-letter amino acid abbreviations. In this sequence the intron has been left out, because it is removed from the transcript before the mRNA is translated into protein. In the lower line, the sequence has been translated without recognizing that an intron is present. As a result of this error, the amino acid sequence appears to terminate within the intron.

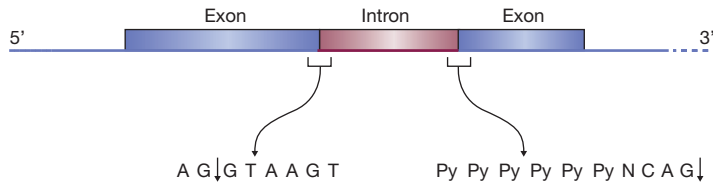


Figure 12.4

The consensus sequences for the upstream and downstream exon–intron boundaries of vertebrate introns. Py = pyrimidine nucleotide (C or T), N = any nucleotide. The arrows indicate the boundary positions.

- **Exon–intron boundaries** can be searched for as these have distinctive sequence features, although unfortunately the distinctiveness of these sequences is not so great as to make their location a trivial task. In vertebrates, the sequence of the upstream exon–intron boundary is usually described as 5′–AG↓GTAAGT–3′ and the downstream one as 5′–PyPyPyPyPyPyNCAG↓–3′, where "Py" means one of the pyrimidine nucleotides (T or C), "N" is any nucleotide, and the arrow shows the precise location of the boundary (Figure 12.4). These are **consensus sequences**, the averages of a large number of related but non-identical sequences, so the search has to include not just the sequences shown but also at least the most common variants. Despite these problems, this type of search can sometimes help delineate the locations of the exons within a region thought to contain a gene.
- **Upstream regulatory sequences** can be used to locate the regions where genes begin. This is because these regulatory sequences, like exon–intron boundaries, have distinctive sequence features that they possess in order to carry out their role as recognition signals for the DNA binding proteins involved in gene expression. As with exon–intron boundaries, the regulatory sequences are variable, and not all genes have the same collection of regulatory sequences. Using these to locate genes is therefore problematic.

These three extensions of simple ORF scanning, despite their limitations, are generally applicable to all higher eukaryotic genomes. Additional strategies are also possible with individual organisms based on the special features of their genomes. For example, vertebrate genomes contain **CpG islands** upstream of many genes, these being sequences of approximately 1 kb in which the GC content is greater than the average for the genome as a whole. Some 40–50% of human genes have an upstream CpG island. These sequences are distinctive and when one is located in vertebrate DNA, a strong assumption can be made that a gene begins in the region immediately downstream.

Gene location is aided by homology searching

Tentative identification of a gene is usually followed by a **homology search**. This is an analysis, carried out by computer, in which the sequence of the gene is compared with all the gene sequences present in the international DNA databases, not just known genes of the organism under study but also genes from all other species. The rationale is that two genes from different organisms that have similar functions have similar sequences, reflecting their common evolutionary histories (Figure 12.5).

To carry out a homology search the nucleotide sequence of the tentative gene is usually translated into an amino acid sequence, as this allows a more sensitive search. This is because there are 20 different amino acids but only 4 nucleotides, so there is less chance of 2 amino acid sequences appearing to be similar purely by chance

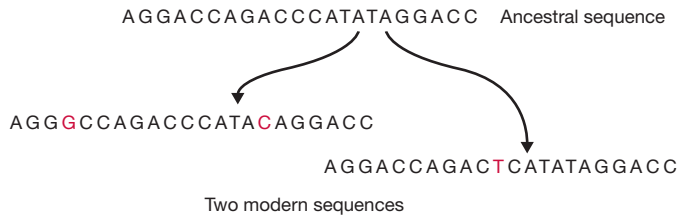


Figure 12.5

Homology between two sequences that share a common ancestor. The two sequences have acquired mutations during their evolutionary histories but their sequence similarities indicate that they are homologs.

Figure 12.6

Lack of homology between two sequences is often more apparent when comparisons are made at the amino acid level. Two nucleotide sequences are shown, with nucleotides that are identical in the two sequences given in red and non-identities given in blue. The two nucleotide sequences are 76% identical, as indicated by the asterisks. This might be taken as evidence that the sequences are homologous. However, when the sequences are translated into amino acids, the identity decreases to 28%, suggesting that the genes are not homologous, and that the similarity at the nucleotide level was fortuitous. Identical amino acids are shown in brown, and non-identities in green. The amino acid sequences have been written using the one-letter abbreviations.

(Figure 12.6). The analysis is carried out through the internet, by logging on to the web site of one of the DNA databases and using a search program such as **BLAST** (Basic Local Alignment Search Tool). If the test sequence is over 200 amino acids in length and has 30% or greater identity with a sequence in the database (i.e., at 30 out of 100 positions the same amino acid occurs in both sequences), then the two are almost certainly homologous and the ORF under study can be confirmed as a real gene. Further confirmation, if needed, can be obtained by using transcript analysis (p. 186) to show that the gene is transcribed into RNA.

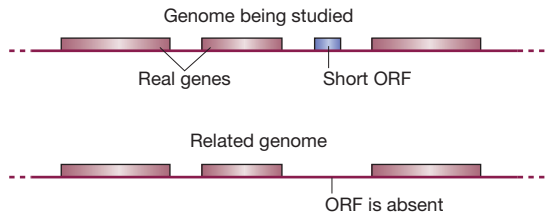
Comparing the sequences of related genomes

A more precise version of homology searching is possible when genome sequences are available for two or more related species. Related species have genomes that share similarities inherited from their common ancestor, overladen with species-specific differences that have arisen since the two species began to evolve independently. Because of natural selection, the sequence similarities between related genomes is greatest within the genes and least in the intergenic regions. Therefore, when related genomes are compared, homologous genes are easily identified because they have high sequence similarity, and any ORF that does not have a clear homolog in the related genome can discounted as almost certainly being a chance sequence and not a genuine gene (Figure 12.7).

This type of homology analysis—called **comparative genomics**—has proved very valuable for locating genes in the *S. cerevisiae* genome, as complete or partial sequences are now available not only for this yeast but also for several related species, such as *Saccharomyces paradoxus*, *Saccharomyces mikatae*, and *Saccharomyces bayanus*. Comparisons between these genomes has confirmed the authenticity of a number of *S. cerevisiae* ORFs, and also enabled almost 500 putative ORFs to be discounted on the grounds that they have no equivalents in the related genomes. The gene maps of

Figure 12.7

Using comparisons between the genomes of related species to test the authenticity of a short ORF. In this example, the questionable ORF is not present in the related genome and so is probably not a real gene.



these yeasts are very similar, and although each genome has undergone its own species-specific rearrangements, there are still substantial regions where the gene order in the *S. cerevisiae* genome is the same as in one or more of the related genomes. Conservation of gene order is called **synteny**, and it makes it very easy to identify homologous genes. More importantly, a spurious ORF, especially a short one, can be discarded with confidence, because its expected location in a related genome can be searched in detail to ensure that no equivalent is present.

12.1.2 Determining the function of an unknown gene

Homology searching serves two purposes. As well as testing the veracity of a tentative gene, identification it can also give an indication of the function of the gene, presuming that the function of the homologous gene is known. Almost 2000 of the genes in the yeast genome were assigned functions in this way. Frequently, however, the matches found by homology searching are to other genes whose functions have yet to be determined. These unassigned genes are called orphans and working out their function is one of the key objectives of post-genomics research.

In future years it will probably be possible to use bioinformatics to gain at least an insight into the function of an orphan gene. It is already possible to use the nucleotide sequence of a gene to predict the positions of α-helices and β-sheets in the encoded protein, albeit with limited accuracy, and the resulting structural information can sometimes be used to make inferences about the function of the protein. Proteins that attach to membranes can often be identified because they possess α-helical arrangements that span the membrane, and DNA binding motifs such as zinc fingers can be recognized. A greater scope and accuracy to this aspect of bioinformatics will be possible when more information is obtained about the relationship between the structure of a protein and its function. In the meantime, functional analysis of orphans depends largely on conventional experiments.

Assigning gene function by experimental analysis requires a reverse approach to genetics

Experimental identification of the function of an unknown gene is proving to be one of the biggest challenges in genomics research. The problem is that the objective—to plot a course from gene to function—is the reverse of the route normally taken by genetic analysis, in which the starting point is a phenotype and the objective is to identify the underlying gene or genes (Figure 12.8). In conventional genetic analysis, the genetic basis of a phenotype is usually studied by searching for mutant organisms in which the phenotype has become altered. The mutants might be obtained experimentally, for example by treating a population of organisms, such as a culture of bacteria, with ultraviolet radiation or a mutagenic chemical, or the mutants might be present in a natural population. The gene or genes that have been altered in the mutant organism are then studied by genetic crosses, which can locate the position of a gene in a genome and also

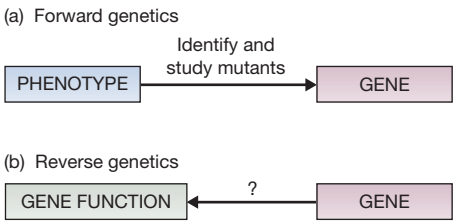


Figure 12.8
Forward and reverse genetics.

determine if the gene is the same as one that has already been characterized. The gene can then be studied further by molecular biology techniques such as cloning and sequencing.

The general principle of this conventional analysis—**forward genetics**—is that the genes responsible for a phenotype can be identified by determining which genes are inactivated in organisms that display a mutant version of the phenotype. If the starting point is the gene, rather than the phenotype, then the equivalent strategy—**reverse genetics**—would be to mutate the gene and identify the phenotypic change that results. This is the basis of most of the techniques used to assign functions to unknown genes.

Specific genes can be inactivated by homologous recombination

The easiest way to inactivate a specific gene is to disrupt it with an unrelated segment of DNA (Figure 12.9). This can be achieved by homologous recombination between the chromosomal copy of the gene and a second piece of DNA that shares some sequence identity with the target gene.

How is gene inactivation carried out in practice? We will consider two examples, the first with *S. cerevisiae*. Since completing the genome sequence in 1996, yeast molecular biologists have embarked on a coordinated, international effort to determine the functions of as many of the unknown genes as possible. One technique makes use of a **deletion cassette**, which carries a gene for antibiotic resistance (Figure 12.10). This gene is not a normal component of the yeast genome but it will work if transferred into a yeast chromosome, giving rise to a transformed yeast cell that is resistant to the antibiotic geneticin. Before using the deletion cassette, new segments of DNA are attached as tails to either end. These segments have sequences identical to parts of the yeast gene that is going to be inactivated. After the modified cassette is introduced into a yeast cell, homologous recombination occurs between the DNA tails and the chromosomal copy of the yeast gene, replacing the latter with the antibiotic resistance gene. The target gene therefore becomes inactivated. Cells which have undergone the replacement are selected by plating the culture onto agar medium containing geneticin, and their phenotypes examined to gain some insight into the function of the gene.

The second example of gene inactivation uses an analogous process, but with mice rather than yeast. The mouse is frequently used as a model organism for humans because

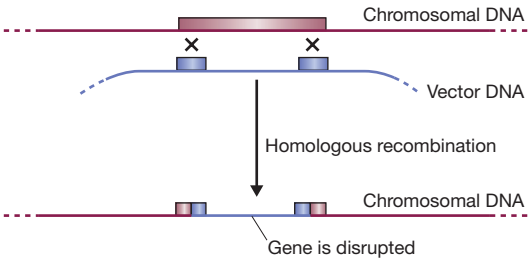
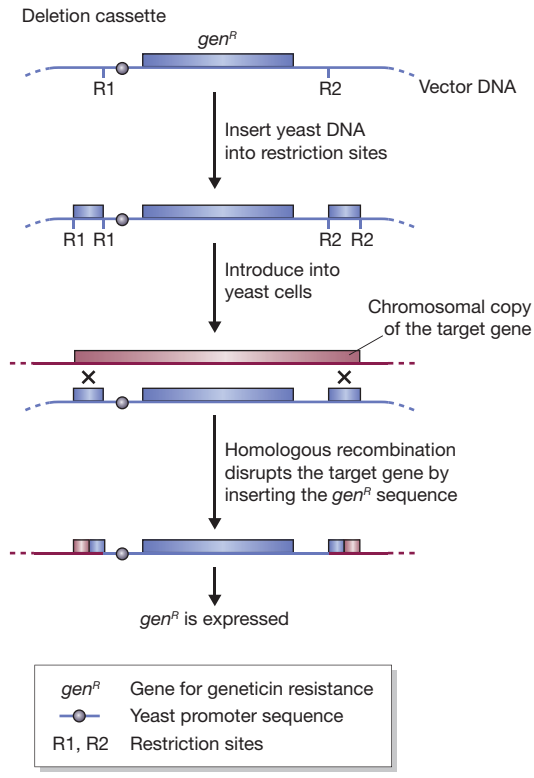


Figure 12.9
Gene disruption by homologous recombination.

Figure 12.10
The use of a yeast deletion cassette.



the mouse genome contains many of the same genes. Identifying the functions of unknown human genes is therefore being carried out largely by inactivating the equivalent genes in mice. The homologous recombination part of the procedure is identical to that described for yeast and once again results in a cell in which the target gene has been inactivated. The problem is that we do not want just one mutant cell, we want a whole mutant mouse, as only with the complete organism can we make a full assessment of the effect of the gene inactivation on the phenotype. To achieve this, the vector carrying the deletion cassette is initially microinjected into an embryonic stem cell (p. 124). The eventual result is a **knockout mouse**, whose phenotype will, with luck, provide the desired information on the function of the gene being studied.

12.2 Studies of the transcriptome and proteome

So far we have considered those aspects of post-genomics research that are concerned with studies of individual genes. The change in emphasis from genes to the genome has resulted in new types of analysis that are aimed at understanding the activity of the genome as a whole. This work has led to the invention of two new terms:

- The **transcriptome**, which is the messenger RNA (mRNA) content of a cell, and which reflects the overall pattern of gene expression in that cell;
- The **proteome**, which is the protein content of a cell and which reflects its biochemical capability.

12.2.1 Studying the transcriptome

Transcriptomes can have highly complex compositions, containing hundreds or thousands of different mRNAs, each making up a different fraction of the overall population. To characterize a transcriptome it is therefore necessary to identify the mRNAs that it contains and, ideally, to determine their relative abundances.

Studying a transcriptome by sequence analysis

The most direct way to characterize a transcriptome is to convert its mRNA into cDNA, and then to sequence every clone in the resulting cDNA library. Comparisons between the cDNA sequences and the genome sequence will reveal the identities of the genes whose mRNAs are present in the transcriptome. This approach is feasible but it is laborious, with many different cDNA sequences being needed before a near-complete picture of the composition of the transcriptome begins to emerge. Can any shortcuts be used to obtain the vital sequence information more quickly?

Serial analysis of gene expression (SAGE) provides one possible solution. Rather than studying complete cDNAs, SAGE yields short sequences, as little as 12 bp in length, each of which represents an mRNA present in the transcriptome. The basis of the technique is that these 12 bp sequences, despite their shortness, are sufficient to enable the gene that codes for the mRNA to be identified.

The first step in generating the 12 bp sequences is to immobilize the mRNA in a chromatography column by annealing the poly(A) tails present at the 3′ ends of these molecules to oligo(dT) strands that have been attached to cellulose beads (Figure 12.11). The mRNA is converted into double-stranded cDNA and then treated with a restriction enzyme that recognizes a 4 bp target site, such as *Alu*I, and so cuts frequently in each cDNA. The terminal restriction fragment of each cDNA remains attached to the cellulose beads, enabling all the other fragments to be eluted and discarded. A short linker is now attached to the free end of each cDNA, this linker containing a recognition sequence for *Bsm*FI. This is an unusual restriction enzyme in that rather than cutting within its recognition sequence, it cuts 10–14 nucleotides downstream. Treatment with *Bsm*FI therefore removes a fragment with an average length of 12 bp from the end of each cDNA. The fragments are collected, ligated head-to-tail to produce a catenane, and sequenced. The individual sequences can be identified within the catenane, because they are separated by *Bsm*FI sites.

Studying transcriptomes by microarray or chip analysis

SAGE enables individual mRNAs to be identified in a transcriptome, but provides only approximate information on the relative abundances of those mRNAs. The dominance of one particular type of mRNA in a transcriptome, such as the gliadin mRNAs in wheat seeds (p. 131), will be evident from the frequency of those mRNA sequences among the SAGE fragments, but more subtle variations in mRNA levels will not be apparent.

Techniques that enable more accurate comparisons of the amounts of individual mRNAs were first developed as part of the yeast post-genomics project. In essence, these techniques involve a sophisticated type of hybridization analysis. Every yeast gene—all 6000 of them—was obtained as an individual clone and samples spotted onto glass slides in arrays of 80×80 spots. This is called a **microarray**. To determine which genes are active in yeast cells grown under particular conditions, mRNA was extracted from the cells, converted to cDNA and the cDNA labeled and hybridized to the microarrays (Figure 12.12). Fluorescent labels were used and hybridization was detected by

Figure 12.11

Serial analysis of gene expression (SAGE).

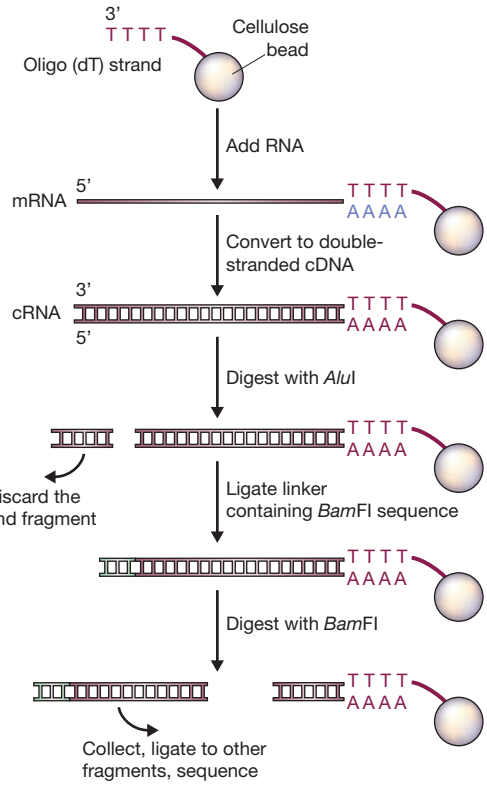
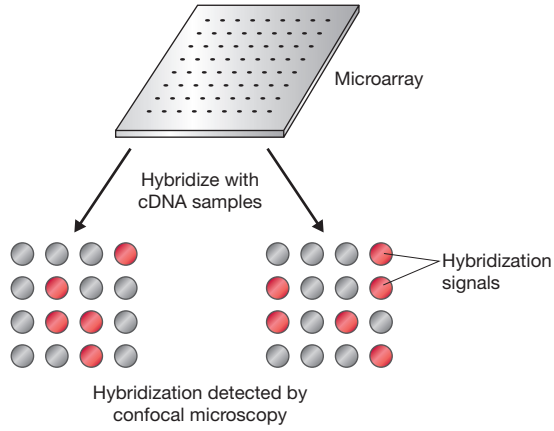


Figure 12.12

Microarray analysis. The microarray shown here has been hybridized to two different cDNA preparations, each labelled with a fluorescent marker. The clones which hybridize with the cDNAs are identified by confocal microscopy.



examining the microarrays by confocal microscopy. Those spots that gave a signal indicated genes that were active under the conditions being studied, and the intensities of the hybridization signals revealed the relative amounts of the mRNAs in the transcriptome. Changes in gene expression, when the yeast were transferred to different growth conditions (e.g., oxygen starvation), could be monitored by repeating the experiment with a second cDNA preparation.

Microarrays are now being used to monitor changes in the transcriptomes of many organisms. In some cases the strategy is the same as used with yeast, the microarray representing all the genes in the genome, but this is possible only for those organisms that