I1 THE DNA REVOLUTION

Key Notes

Genomics

Genomics refers to studies of an organism's genome. The analysis of gene function (functional genomics) relies upon a range of techniques including reverse genetics and RNA interference (RNA_i).

Transcriptomics and proteomics

Transcriptomics is the study of the transcriptome (the expressed RNA) whereas proteomics is the study of the proteome (the proteins that are synthesized).

Metabolomics

Metabolomics is the study of the small molecule components of a cell, i.e. the metabolome.

Transgenic organisms

Transgenic organisms are those that have been modified by the insertion of a cloned gene(s).

Related topics

Protein sequencing and peptide synthesis (B8)
Restriction enzymes (I2)
Nucleic acid hybridization (I3)
DNA cloning (I4)
DNA sequencing (I5)
Polymerase chain reaction (I6)

Genomics

We are living in an unprecedented era of biological discovery and the application of biological knowledge. Automated DNA sequencing (Topic I5) delivered, in 2001, over 2.6 billion base pairs of DNA sequence in the human genome (the **Human Genome Project**) and the genomes of many other organisms have either already been or are being sequenced too. This vast, and ever-increasing, wealth of DNA sequence data is a major asset for genomics, the study of an organism's genome. One of the key goals is to understand the functions of the large numbers of new genes predicted by genome sequencing, a field known as **functional genomics**. One approach to revealing the function of a gene is called **reverse genetics** whereby a mutation can be created in a cloned gene and the modified gene can be introduced into a host cell to monitor the effects of the mutation. The approach is called *reverse* genetics since starting with a gene and then creating a mutant is the reverse of traditional genetics whereby mutants were used to identify genes. Another useful approach is **RNA interference (RNAi)** whereby a double-stranded RNA (dsRNA) corresponding to the sense and antisense strands of the gene under investigation is introduced into cells. The dsRNA is degraded *in vivo* but the resulting fragments base pair with mRNA from the target gene and cause that to be degraded also, in effect terminating expression of the gene. Monitoring the cellular effects of this loss of function provides evidence of the role of the gene *in vivo*.

Transcriptomics and proteomics

By analogy with the term 'genome', the **transcriptome** is all of the RNA sequence transcribed from a cell's genome and the **proteome** is all of the expressed proteins of that cell. Whereas all of the cells of an organism such as a human contain essen-

tially the same genome, the transcriptome and proteome of different cell types varies depending on the genes that are expressed. For example, the transcriptome and proteome of a liver cell are different from that of a neurone. **Trancriptomics** refers to studies of the transcriptome and includes, for example, the use of DNA microarrays (Topic I3) to determine expression profiles. **Proteomics** (see Topic B8) is the study of the proteome and currently relies heavily on using two-dimensional gel electrophoresis (see Topic B8) to separate the proteins expressed by a cell or tissue, followed by mass spectrometry to produce peptide mass fingerprints of each protein (see Topic B8). The expressed proteins can then be identified by comparing these fingerprints with databases of fingerprints, including those predicted from DNA sequence data. Given the large scale of such techniques, many of these procedures are necessarily automated.

Metabolomics

Together, genomics, transcriptomics and proteomics are powerful approaches to increasing our understanding not only of how cells function normally but also what the key changes are in disease and so will find increasing use in the diagnosis of specific diseases and in the identification and assessment of new chemotherapeutic agents. **Metabolomics** is perhaps the newest of these fields of study and relates to the study of the small molecule components of the cell, that is, the **metabolome**. By analyzing the metabolites of a cell, one can generate a **metabolic profile** that indicates the cell's metabolic activity; information which should prove useful for a range of applications, including clinical diagnostics and drug discovery. The techniques of metabolomics center on methods to separate the various classes of small molecules, such as gas liquid chromatography, high performance liquid chromatography and capillary electrophoresis, followed by identification using, for example, mass spectrometry. So far, techniques for large-scale metabolomic analyses along the lines of those employed for genomics, proteomics and transcriptomics have yet to be developed.

Transgenic organisms

As the functions of individual genes become known, the power of this new biology can be used to modify organisms in predictable and desirable ways. Organisms that have been modified by the insertion of a cloned gene are called **transgenic organisms**; **transgenic plants** and **transgenic animals** are both possible. Such approaches are not only of academic importance but are increasingly finding commercial applications. For example, modified plants with improved pest or virus resistance have obvious attractions for agriculture. However, there are important ethical considerations that must be applied to the use of this new technology both for plants and, even more controversially, for animals (including humans).

This wealth of knowledge about the genome and its expression, and the application of that knowledge, relies upon a wide range of recombinant DNA tools and techniques. Many of the present-day experimental approaches are extremely sophisticated and are beyond the scope of this introductory text. However, the following Topics provide an understanding of some of the core methodology: the ability to cut DNA at specific sites using restriction endonucleases (**restriction enzymes**; see Topic I2), procedures that allow the detection of specific DNA (and RNA) sequences with great accuracy (**nucleic acid hybridization**; see Topic I3), methods for preparing specific DNA sequences in large amounts in pure form (**DNA cloning**; see Topic I4) and rapid **DNA sequencing** (see Topic I5). More recently, the development of the **polymerase chain reaction (PCR)** (see Topic I6) has revolutionized the field of molecular biology. A more extensive description of recombinant DNA technology is provided in the companion book *Instant Notes in Molecular Biology*.

I2 Restriction enzymes

Key Notes

Restriction enzyme digestion

Restriction enzymes recognize specific recognition sequences and cut the DNA to leave cohesive ends or blunt ends. The ends of restricted DNA molecules can be joined together by ligation to create new recombinant DNA molecules.

Nomenclature

Restriction enzymes have a three-letter name based on the genus and species name of the bacterium from which they were isolated, together with a roman numeral designed to indicate the identity of the enzyme in cases when the bacterium contains several different restriction enzymes.

Gel electrophoresis

DNA fragments in a restriction digest can be separated by size by electrophoresis in polyacrylamide or agarose gel. Polyacrylamide gel is used to separate smaller DNA molecules whilst agarose gel has larger pore sizes and so can separate larger DNA fragments.

Restriction maps

A map showing the position of cut sites for a variety of restriction enzymes is called the restriction map for that DNA molecule. Restriction maps allow comparison between DNA molecules without the need to determine the nucleotide sequence and are also much used in recombinant DNA experiments.

Restriction fragment length polymorphisms

A restriction fragment length polymorphism (RFLP) is a common difference between the DNA of individuals in a population (i.e. a polymorphism) that affects the sizes of fragments produced by a specific restriction enzyme. If the RFLP lies near a gene, changes in which can cause a human genetic disease, it can be used as a marker for that gene. In the past, RFLPs have proved valuable both for screening patients for the gene defect and also in studies directed at cloning the gene. However, RFLPs are becoming less commonly used in such work as the genes themselves are identified. The polymerase chain reaction (PCR) is increasingly the method of choice for screening.

Related topics

DNA structure (F1)
Nucleic acid hybridization (I3)
DNA cloning (I4)

Restriction enzyme digestion

Restriction enzymes recognize specific nucleotide sequences (**recognition sequences**) in double-stranded DNA, that are usually four, five or six nucleotides long, and then cut both strands of the DNA at specific locations. There are basically three ways in which the DNA can be cut; a staggered cut to leave a **5′ overhang** (i.e. a short single-stranded region of DNA is left that has a 5′ end and overhangs the end of the double-stranded DNA), a staggered cut to leave a **3′**

overhang, or a cut in the same place on both strands to leave a **blunt end** (*Fig. 1*). For enzymes that cut in the staggered manner, the single-stranded tails are called **cohesive ends** because they allow any two DNA fragments produced by the same restriction enzyme to form complementary base pairs (*Fig. 1*). The cut ends can

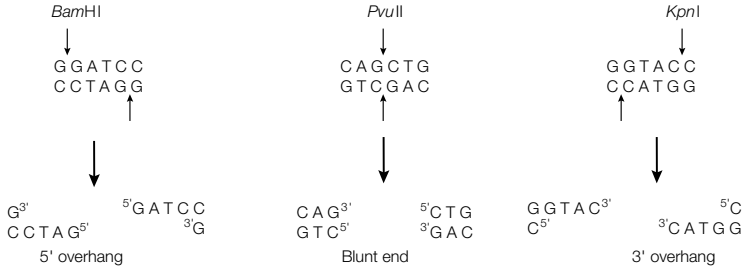

Fig. 1. The three types of cleavage by commonly used restriction enzymes.

then be joined together (**ligated**) by an enzyme called **DNA ligase**. The new DNA molecule that has been made by joining the DNA fragments is called a **recombinant DNA molecule** (*Fig. 2*). Blunt-ended DNA molecules can also be joined together by DNA ligase but the reaction is less favorable.

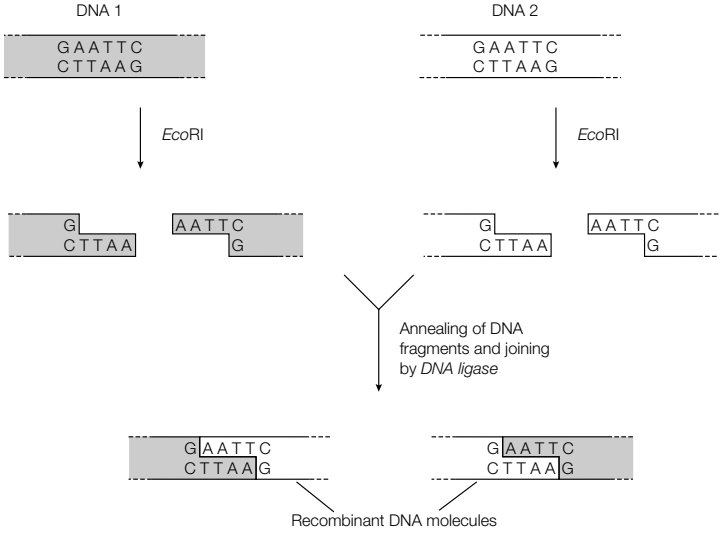

Fig. 2. Using a restriction enzyme to create recombinant DNA.

Nomenclature

Restriction enzymes are isolated from bacteria, where they play a role in protecting the host cell against virus infection. Over 100 restriction enzymes have now been isolated and have been named according to the bacterial species from which they were isolated. The first three letters of the enzyme name are the first letter of the genus name and the first two letters of the species name. Since each bacterium may contain several different restriction enzymes, a Roman numeral is also used to identify each enzyme. *Eco*RI, for example, was the first enzyme isolated from *Escherichia coli*.

Gel electrophoresis

When a DNA molecule is cut by a restriction enzyme, the DNA fragments (called **restriction fragments**) from that **restriction digest** can be separated by **gel electrophoresis** (*Fig. 3*). Electrophoresis on a polyacrylamide gel will separate small DNA fragments of less than about 500 bp in size, but agarose gels (which have larger pores) are needed to separate larger DNA fragments. The DNA digest separates into a series of bands representing the restriction fragments. Since small fragments travel further in the gel than larger fragments, the size of each fragment can be determined by measuring its migration distance relative to standard DNA fragments of known size. The DNA can be located after gel electrophoresis by staining with ethidium bromide that binds to the DNA and fluoresces a bright orange. Alternatively, if the DNA is labeled with a radioisotope such as ^{32}P, the bands can be detected after electrophoresis by laying the gel against an X-ray film whereby the radioactivity causes silver grains to be formed in the film emulsion, giving black images corresponding to the radioactive bands **(autoradiography)**.

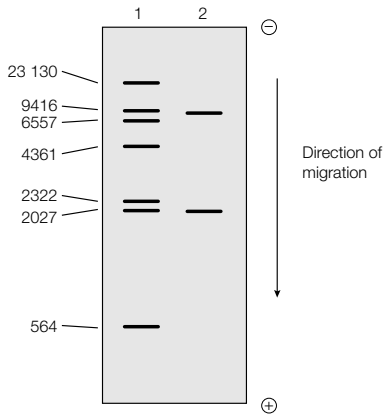

Fig. 3. Agarose gel electrophoresis of DNA fragments. DNA fragments of known size were electrophoresed in lane 1 (the sizes in bp are given on the left). A restriction digest of the sample DNA was electrophoresed in lane 2. By comparison with the migration positions of fragments in lane 1, it can be seen that the two sample DNA fragments have sizes of approximately 9000 bp and 2000 bp. The sizes could be determined more accurately by plotting the data from lane 1 as a standard curve of log DNA size vs. migration distance and then using this to estimate the size of the sample DNA fragments from their measured migration distances.

Restriction maps

Any double-stranded DNA will be cut by a variety of restriction enzymes that have different recognition sequences. By separating the restriction fragments and measuring their sizes by gel electrophoresis, it is possible to deduce where on the DNA molecule each restriction enzyme cuts. A **restriction map** of the DNA molecule can be drawn showing the location of these cut sites (**restriction sites**) (*Fig. 4*). It is then easy to compare two DNA molecules (for example, to examine the evolutionary relationship between two species) by looking at their restriction maps without the need to determine the nucleotide sequence of each DNA. Restriction maps are also important experimentally during recombinant DNA work, both to plan where individual DNA molecules should best be cut and to monitor the progress of the experiment.

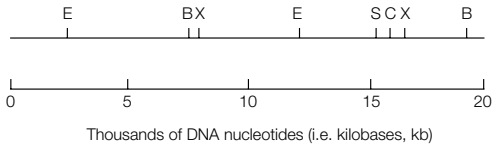

Fig. 4. A typical restriction map of a DNA molecule. The cleavage sites of different restriction enzymes, indicated by letters, are shown. For example, E denotes sites where EcoRI *cuts.*

Restriction fragment length polymorphisms

Analysis of human genomic DNA has revealed that there are many differences in DNA sequence between individuals that have no obvious effect, often because the changes lie in introns or between genes. Some of these changes are very common in individuals in a population and are called **polymorphisms**. Some polymorphisms affect the size of fragments generated by a particular restriction enzyme, for example by changing a nucleotide in the recognition sequence and so eliminating a cut site. Instead of two restriction fragments being generated from this region, a single large restriction fragment is now formed (*Fig. 5*). Alternatively the polymorphism may result from the insertion or deletion of

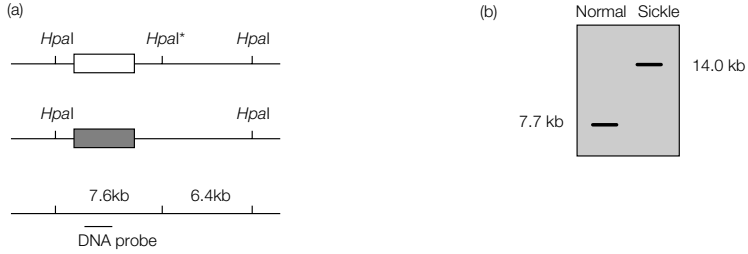

Fig. 5. Analysis of human genetic disease using RFLPs. The analysis concerns two individuals from a family, one of whom has normal β-globin and one of whom has an abnormal β-globin gene that leads to sickle-cell anemia. (a) The sickle β-globin is associated with a nucleotide change that results in the loss of the HpaI site marked with an asterisk. The presence or absence of this HpaI site is detected by hybridization and Southern blotting (see Topic I3) using a DNA probe for the 7.6 kb fragment; (b) normal DNA with three HpaI sites yields a 7.6 kb fragment detected by the DNA probe but sickle DNA yields a 14.0 kb fragment due to loss of the HpaI* site.*

sequences between two cut sites, so increasing or decreasing the size of that restriction fragment produced. A polymorphism that affects restriction fragment sizes is called a **restriction fragment length polymorphism (RFLP)**. Provided that a DNA probe (see Topic I3) exists for a sequence of DNA within the affected region, so that this sequence can be detected by hybridization, RFLPs can be detected by Southern blotting (see Topic I3).

The value of RFLPs has been in the ability to use these as markers for particular human genetic diseases. Consider a polymorphism that happens to be very near to the site of a change in a key gene that results in a human genetic disease (*Fig. 5*). Because these two changes, the polymorphism and the genetic defect, lie close together on the same chromosome, they will tend to be co-inherited. Identifying such a **closely linked** RFLP has two major advantages. First, experiments can be directed to cloning DNA near the RFLP in the hope of identifying the gene itself which can then be sequenced and studied. Secondly, even in the absence of the gene, the RFLP acts as a screening marker for the disease; individuals who have the RFLP have a high probability of having the associated gene defect. Of course other RFLPs that are located a very long way from the gene, or even on a different chromosome, will essentially be **unlinked** (i.e. because of the high probability of cross-over events during meiosis to produce germ-line cells, the gene and RFLP will have only a 50:50 chance of being co-inherited). Thus a large amount of very painstaking work has to be carried out to identify a useful RFLP for a particular human genetic disease. Large numbers of individuals in family groups, some of whom suffer from the disease, need to be screened for a range of likely RFLPs to attempt to locate an RFLP that is routinely co-inherited with the gene defect.

As the genes themselves are identified and sequenced, so the need for RFLP markers declines since specific DNA probes (see Topic I3) for the most common types of gene defect can be employed. In addition, the use of the polymerase chain reaction (PCR; see Topic I6) in screening for human genetic disease is increasingly the method of choice rather than RFLP analysis since it is much faster to perform and requires far less clinical material for analysis.

I3 Nucleic acid hybridization

Key Notes

The hybridization reaction

Double-stranded DNA denatures into single strands as the temperature rises but renatures into a double-stranded structure as the temperature falls. Any two single-stranded nucleic acid molecules can form double-stranded structures (hybridize) provided that they have sufficient complementary nucleotide sequence to make the resulting hybrid stable under the reaction conditions.

Monitoring specific nucleic acid sequences

The concentration of a specific nucleic acid sequence in a sample can be measured by hybridization with a suitable labeled DNA probe. After hybridization, nuclease is used to destroy unhybridized probe and the probe remaining is a measure of the concentration of the target sequence. The hybridization conditions can be altered to ensure that only identical sequences (high stringency conditions) or identical plus related sequences (low stringency conditions) will hybridize with the probe and hence be detected.

Southern blotting

Southern blotting involves electrophoresis of DNA molecules in an agarose gel and then blotting the separated DNA bands on to a nitrocellulose filter. The filter is then incubated with a labeled DNA probe to detect those separated DNA bands that contain sequences complementary to the probe.

Northern blotting

Northern blotting is analogous to Southern blotting except that the sample nucleic acid that is separated by gel electrophoresis is RNA rather than DNA.

***In situ* hybridization**

For *in situ* hybridization, a tissue sample is incubated with a labeled nucleic acid probe, excess probe is washed away and the location of hybridized probe is examined. The technique enables the spatial localization of gene expression to be determined as well as the location of individual genes on chromosomes.

DNA microarrays

A DNA microarray is a large number of DNA fragments or oligonucleotides arrayed in known positions on a glass slide. After hybridization with fluorescently-labeled target cDNA, examination using automated scanning laser microscopy indicates which DNA sequences are expressed.

Related topics

DNA structure (F1)
Restriction enzymes (I2)
DNA cloning (I4)

The hybridization reaction

As double-stranded DNA is heated, a temperature is reached at which the two reaction strands separate. This process is called **denaturation**. The temperature at which half of the DNA molecules have denatured is called the **melting temperature** or T_m for that DNA. If the temperature is now lowered and falls below the T_m, the two complementary strands will form hydrogen bonds with each other once more to reform a double-stranded molecule. This process is called **renaturation** (or **reannealing**). In fact, double-stranded structures can form between any two single-stranded nucleic acid molecules (DNA–DNA, DNA–RNA, RNA–RNA) provided that they have sufficient complementary nucleotide sequence to make the double-stranded molecule stable under the conditions used. The general name given to this process is **hybridization** and the double-stranded nucleic acid product is called a **hybrid**.

Monitoring specific nucleic acid sequences

The rate of formation of double-stranded hybrids depends on the concentration of the two single-stranded species. This can be used to measure the concentration of either specific DNA or RNA sequences in a complex mixture. The first task is to prepare a single-stranded **DNA probe** (i.e. a DNA fragment that is complementary to the nucleic acid being assayed). This can be one strand of a DNA restriction fragment, cloned DNA or a synthetic oligonucleotide. It must be labeled in order to be able to detect the formation of hybrids between it and the target nucleic acid. Whereas most labeling used to be via the incorporation of a radioisotope, nonradioactive chemical labels are now often used instead. For example, a DNA probe can be labeled with digoxygenin, a steroid, by using digoxygenin-labeled dUTP during DNA synthesis. Hybrids containing the digoxygenin-labeled DNA probe can then be detected using antidigoxygenin antibody linked to a fluorescent dye. Irrespective of the method of probe labeling, the DNA probe is incubated with the nucleic acid sample (the **'target'** DNA or RNA) and then nuclease is added to degrade any unhybridized single-stranded probe. The amount of labeled probe remaining indicates the concentration of the target nucleic acid in the sample.

The hybridization conditions (e.g. temperature, salt concentration) can be varied so as to govern the type of hybrids formed. The conditions may be arranged so that only perfectly matched hybrids are stable and hence assayed (conditions known as **high stringency**). Alternatively, the conditions may be such that even poorly complementary hybrids are stable and will be detected (**low stringency**). Thus by varying the reaction conditions it is possible to detect and quantify only those target sequences that are identical to the DNA probe or, alternatively, to detect and quantify related sequences. Hybridization of nucleic acid probes with genomic DNA, for example, can be used to measure the copy number of particular DNA sequences in the genome. Hybridization of a DNA probe with cellular RNA as the target will indicate the concentration of the corresponding RNA transcript and hence give information about the level of gene expression. Variants of the methodology even allow determination of the transcriptional Start and Stop sites and the number and location of intron sequences in protein-coding genes.

Southern blotting

Gel electrophoresis is widely used to separate and size DNA molecules during recombinant DNA experiments. After gel electrophoresis, there is often a need to detect one or more DNA fragments containing a specific nucleotide sequence. This is easily carried out by Southern blotting. After electrophoresis of the restriction fragments through an agarose gel, the gel is soaked in alkali to denature the

DNA to single strands and the pH is then neutralized. The gel is placed in contact with a nitrocellulose or nylon membrane filter sheet arranged so that buffer flows through the gel and carries the DNA fragments to the membrane (*Fig. 1*). The membrane binds the single-stranded DNA and so the band pattern in the gel is now transferred to it. The membrane filter is peeled from the gel, baked at high temperature to fix the DNA to it, and then incubated with a radiolabeled DNA probe. After hybridization, the probe will have bound only to DNA fragments with complementary sequences. These can be visualized by washing away excess probe and then placing the filter against an X-ray film for autoradiography. The images produced on the autoradiogram indicate those bands that contain the probe sequence (*Fig. 1*).

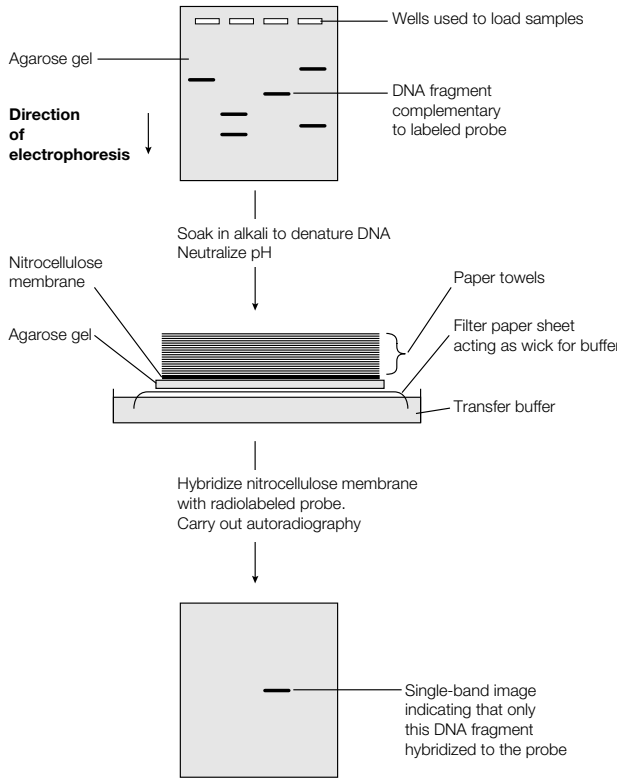

Fig. 1. Southern blotting. The procedure shown is the original method of Southern using capillary action to blot the DNA bands from the gel to the nitrocellulose membrane. Electrolytic transfer is now often used instead.

Northern blotting

Northern blotting follows much the same procedure as Southern blotting except that the sample analyzed by gel electrophoresis and then bound to the filter is RNA not DNA. Therefore the technique detects RNA molecules that are complementary to the DNA probe. If cellular RNA is electrophoresed, for example, a DNA probe for a specific mRNA could be used to detect whether that mRNA

was present in the sample. The migration distance of the RNA in the gel would also allow estimation of its size. Note that Southern blotting (for DNA) obtained its name after its inventor (E. Southern); the name Northern blotting (for RNA) was devised later and is a geographical pun!

In situ hybridization

It is also possible to incubate radioactive or fluorescent nucleic acid probes with sections of tissues or even chromosomes, wash away excess probe and then detect where the probe has hybridized. This technique (***in situ* hybridization**) has proved to be very powerful in determining which cells in a complex tissue such as the mammalian brain express a particular gene and for locating specific genes on individual chromosomes.

DNA microarrays

The techniques described above are limited in that only a relatively small number of samples can be analyzed at any one time. In contrast, **DNA microarrays** (DNA chips, gene chips) can analyze the expression of tens of thousands of genes simultaneously. A DNA microarray is a large number of DNA sequences that are spotted onto a glass slide in a pre-determined grid pattern; given the large numbers of DNAs involved, this is done robotically. In other cases, instead of spotting DNA fragments, the DNA microarray may be produced by synthesizing thousands of oligonucleotides on the glass slide *in situ*; as many as 40 000 oligonucleotides per square centimeter. The microarray, containing DNA fragments or oligonucleotides, can then be used to explore the expression of each of these DNA sequences in the tissue or sample of interest by hybridization. In the terminology of hybridization (see above), the DNAs or oligonucleotides on the microarray are the 'probe' and the RNA in the tissue or sample is the 'target'.

To understand how DNA microarrays are used, consider the following typical application (*Fig.* 2). Imagine that the goal is to determine which genes are regulated by a newly-discovered hormone. Cells are exposed to the hormone (test sample) or left untreated (control sample), RNA is isolated from each sample and is used to synthesize cDNA using reverse transcriptase. When synthesizing cDNA from the test sample RNA, one of the nucleotide precursors is labeled with, for example, a red fluorescent dye, so that the resulting cDNA is also tagged with this label. The control sample cDNA is similarly labeled but with, for example, a green fluorescent dye. The two cDNAs are mixed together and allowed to hybridize to the DNA microarray. Any cDNA that does not hybridize is washed away and the DNA microarray is examined using an automated scanning-laser microscope. Laser excitation of the microarray with light of the appropriate wavelength to excite the relevant fluorophore and measurement of the intensity of the resulting fluorescence for each DNA or oligonucleotides spot allows the extent of hybridization with test (red) and control (green) cDNA to be determined. Since the exact location of every DNA or oligonucleotide in the microarray is known, these data immediately indicate which genes are activated by the hormone (red spots), which genes are expressed only in the absence of the hormone (green spots) and which genes are unaffected and are expressed both in the absence and presence of the hormone (yellow spots = red + green).

DNA microarrays are now widely used to examine changes in gene expression in both plants and animals. For example, in humans, they can be used to determine how particular diseases affect the pattern of gene expression (the **expression profile**) in various tissues, or the identity (from the expression profile) of the infecting organism. Thus, in clinical medicine alone, DNA microarrays have huge potential for diagnosis.

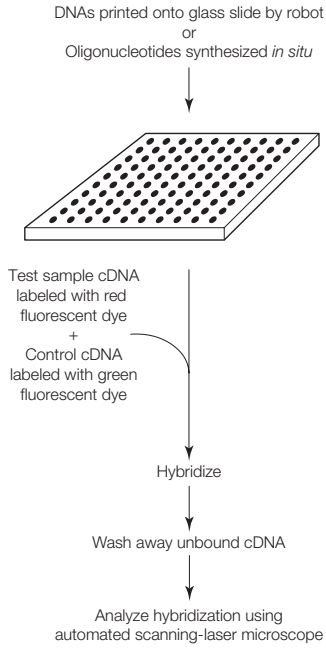

Fig. 2 A typical use of a DNA microarray to determine the expression of thousands of genes simultaneously (see the text for details).

I4 DNA CLONING

Key Notes

The principle of DNA cloning

Most foreign DNA fragments cannot self-replicate in a cell and must therefore be joined (ligated) to a vector (virus or plasmid DNA) that can replicate autonomously. Each vector typically will join with a single fragment of foreign DNA. If a complex mixture of DNA fragments is used, a population of recombinant DNA molecules is produced. This is then introduced into the host cells, each of which will typically contain only a single type of recombinant DNA. Identification of the cells that contain the DNA fragment of interest allows the purification of large amounts of that single recombinant DNA and hence the foreign DNA fragment.

The basics of DNA cloning

To clone into a plasmid vector, both the plasmid and the foreign DNA are cut with the same restriction enzyme and mixed together. The cohesive ends of each DNA reanneal and are ligated together. The resulting recombinant DNA molecules are introduced into bacterial host cells. If the vector contains an antibiotic resistance gene(s) and the host cells are sensitive to these antibiotic(s), plating on nutrient agar containing the relevant antibiotic will allow only those cells that have been transfected and contain plasmid DNA to grow.

DNA libraries

Genomic DNA libraries are made from the genomic DNA of an organism. A complete genomic DNA library contains all of the nuclear DNA sequences of that organism. A cDNA library is made using complementary DNA (cDNA) synthesized from mRNA by reverse transcriptase. It contains only those sequences that are expressed as mRNA in the tissue or organism of origin.

Screening DNA libraries

Genomic libraries and cDNA libraries can be screened by hybridization using a labeled DNA probe complementary to part of the desired gene. The probe may be an isolated DNA fragment (e.g. restriction fragment) or a synthetic oligonucleotide designed to encode part of the gene as deduced from a knowledge of the amino acid sequence of part of the encoded protein. In addition, expression cDNA libraries may be screened using a labeled antibody to the protein encoded by the desired gene or by using any other ligand that binds to that protein.

Related topics

DNA structure (F1)
Restriction enzymes (I2)
Nucleic acid hybridization (I3)

The principle of DNA cloning

Consider an experimental goal which is to make large amounts of a particular DNA fragment in pure form from a mixture of DNA fragments. Although the DNA fragments can be introduced into bacterial cells, most or all will lack the ability for self-replication and will quickly be lost. However, two types of DNA

molecule are known which can replicate autonomously in bacterial cells; bacteriophages and plasmids. Plasmids are small circular double-stranded DNA molecules that exist free inside bacterial cells, often carry particular genes that confer drug resistance, and are self-replicating. If a recombinant DNA molecule is made by joining a foreign DNA fragment to plasmid (or bacteriophage) DNA, then the foreign DNA is replicated when the plasmid or phage DNA is replicated. In this role, the plasmid or phage DNA is known as a **vector**. Now, a population of recombinant DNA molecules can be made, each recombinant molecule containing one of the foreign DNA fragments in the original mixture. This can then be introduced into a population of bacteria such that each bacterial cell contains, in general, a different type of recombinant DNA molecule. If we can identify the bacterial cell that contains the recombinant DNA bearing the foreign DNA fragment we want, the cell can be multiplied in culture and large amounts of the recombinant DNA isolated. The foreign DNA can then be recovered from this in pure form; it is then said to have been **cloned**. The vector that was used to achieve this cloning is called a **cloning vector**. Vectors are not limited to bacterial cells; animal and plant viruses can also act as vectors.

The basics of DNA cloning

There are a wide variety of different procedures for cloning DNA into either plasmid or viral vectors but the basic scheme of events is often the same. To clone into a plasmid vector, the circular plasmid DNA is cut with a restriction enzyme (see Topic I2) that has only a single recognition site in the plasmid. This creates a linear plasmid molecule with cohesive ends (*Fig. 1*). The simplest cloning strategy is now to cut the foreign donor DNA with the same restriction enzyme. Alternatively, different restriction enzymes can be used, provided that they create the same cohesive ends (see Topic I2). The donor DNA and linear plasmid DNA are now mixed. The cohesive ends of the foreign DNA anneal with the ends of the plasmid DNA and are joined covalently by DNA ligase. The resulting **recombinant plasmid DNA** is introduced into bacterial host cells that have been treated to become permeable to DNA. This uptake of DNA by the bacterial cells is called **transfection**; the bacterial cells are said to have been **transfected** by the recombinant plasmid. The bacterial cells are now allowed to grow and divide, during which time the recombinant plasmids will replicate many times within the cells. One useful procedure is to select as cloning vector a plasmid that carries one or more antibiotic resistance genes plus a host that is sensitive to those antibiotics (*Fig. 1*). Then, after transfection, the cells are grown in the presence of the antibiotic(s). Only cells containing plasmid DNA will be resistant to the antibiotic(s) and can grow. If the cells are spread on an agar plate, each cell will multiply to form a bacterial colony where all the cells of that colony contain the same recombinant plasmid DNA bearing the same foreign DNA fragment. Thus all that is now needed is to identify the particular bacterial colony that contains the foreign DNA sequence of interest.

DNA libraries

A DNA library is a collection of cloned DNA fragments in a cloning vector that can be searched for a DNA of interest. If the goal is to isolate particular gene sequences, two types of library are useful:

- **Genomic DNA libraries**. A genomic DNA library is made from the genomic DNA of an organism. For example, a mouse genomic library could be made by digesting mouse nuclear DNA with a restriction nuclease to produce a large number of different DNA fragments but all with identical cohesive ends. The

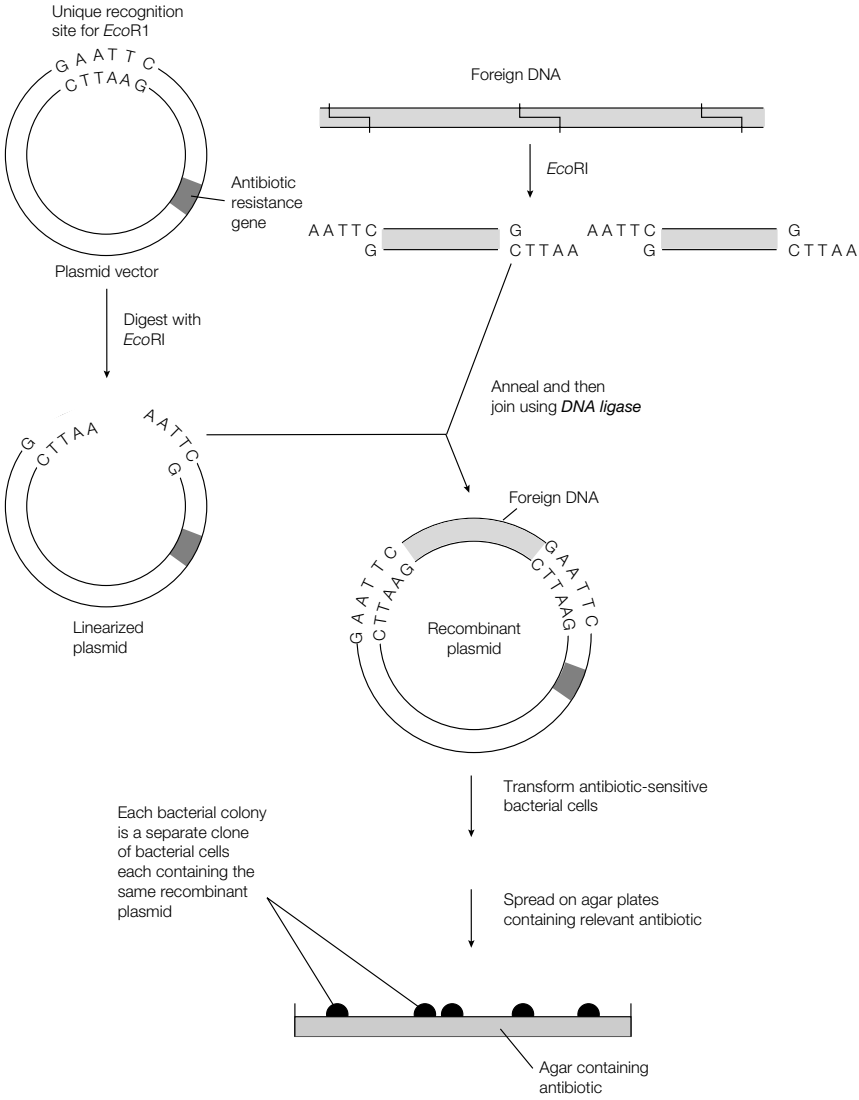

Fig. 1. A simple method of DNA cloning using a plasmid vector.

DNA fragments would then be ligated into linearized plasmid vector molecules or into a suitable virus vector. This library would contain all of the nuclear DNA sequences of the mouse and could be searched for any particular mouse gene of interest. Each clone in the library is called a **genomic DNA clone**. Not every genomic DNA clone would contain a complete gene since in many cases the restriction enzyme will have cut at least once within the gene. Thus some clones will contain only a part of a gene.

- **cDNA libraries**. A cDNA library is made by using the reverse transcriptase of a retrovirus to synthesize **complementary DNA (cDNA)** copies of the total mRNA from a cell (or perhaps a subfraction of it). The single-stranded cDNA is converted into double-stranded DNA and inserted into the vector. Each clone in the library is called a **cDNA clone**. Unlike a complete genomic library that contains all of the nuclear DNA sequences of an organism, a cDNA library contains only sequences that are expressed as mRNA. Different tissues of an animal, that express some genes in common but also many different genes, will thus yield different cDNA libraries.

Screening DNA libraries

Genomic libraries are screened by hybridization (see Topic I3) with a DNA probe that is complementary to part of the nucleotide sequence of the desired gene. The probe may be a DNA restriction fragment or perhaps part of a cDNA clone. Another approach is possible if some of the protein sequence for the desired gene is known. Using the genetic code, one can then deduce the DNA sequence of this part of the gene and synthesize an oligonucleotide with this sequence to act as the DNA probe.

When using a plasmid vector, a simple procedure for screening would be to take agar plates bearing bacterial colonies that make up the genomic library and overlay each plate with a nitrocellulose membrane (*Fig. 2*). This is peeled off and is a **replica** of the plate in that some of the bacterial cells in each colony will have adhered to it and in the same pattern as the colonies on the plate. This filter is often called a '**colony lift**'. It is treated with alkali to lyse the bacterial cells and denature the DNA and then hybridized with a radiolabeled DNA probe. After washing away unreacted probe, autoradiography of the filter shows which

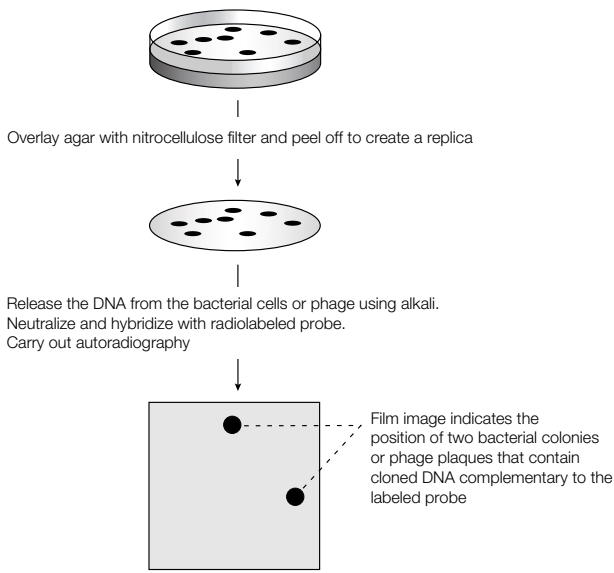

Fig. 2. Screening a gene library by hybridization.

colonies have hybridized with the probe and thus contain the desired sequences. These are then recovered from the agar plate.

When a bacteriophage is used as the cloning vector, the gene library is screened as an array of plaques in a bacterial lawn. A hybridization screening method is used similar to that described for plasmid screening (see Fig. 2); in this case the replica filter is called a '**plaque lift**'.

For cDNA libraries, screening can similarly be carried out by hybridization. In addition, it is possible to make the cDNA library using a vector that will actually transcribe the inserted cDNA and then translate the resulting mRNA to form protein corresponding to the cloned gene. A library made with such an **expression vector** is an **expression cDNA library**. It can be screened using a labeled antibody that recognizes the specific protein and hence identifies those bacteria which contain the desired gene and are synthesizing the protein. Not just antibody but any ligand that binds to the target protein can be used as a probe. For example, labeled hormone may be used to identify clones synthesizing hormone receptor proteins.

I5 DNA SEQUENCING

Key Notes

Two methods for DNA sequencing

DNA can be sequenced by the chemical method or the chain termination procedure. The latter is now the standard method; the (single-stranded) DNA to be sequenced serves as the template for the synthesis of a complementary strand when supplied with a specific primer and *E. coli* DNA polymerase I.

Chain termination method

Four incubation mixtures are set up, each containing the DNA template, a specific DNA primer, *E. coli* DNA polymerase I and all four deoxyribonucleoside triphosphates (dNTPs). In addition, each mixture contains a different dideoxynucleoside triphosphate analog, ddATP, ddCTP, ddGTP or ddTTP. Incorporation of a dideoxy analog prevents further elongation and so produces a chain termination extension product. The products are electrophoresed on a polyacrylamide gel and the DNA sequence is read from the band pattern produced.

Automated DNA sequencing

Automated DNA sequencing uses the chain termination method but with an oligonucleotide primer labeled with a fluorescent dye. Each of the four reactions receives a primer labeled with a different dye. After incubation, the reaction mixtures are pooled and electrophoresed on one lane of a polyacrylamide gel. The order in which the different fluorescently labeled termination products elute from the gel gives the DNA sequence. More advanced systems use multiple capillary sets in which sample preparation, loading and data analysis are automated for maximum throughput.

Related topics

DNA structure (F1)
DNA replication in bacteria (F3)
Nucleic acid hybridization (I3)

Two methods for DNA sequencing

Two main methods have been devised to sequence DNA; the **chemical method** (also called the *Maxam–Gilbert method* after its inventors) and the **chain termination method** (also known as the *Sanger dideoxy method* after its inventor). The chain termination method is now the method usually used because of its speed and simplicity. In this procedure, the DNA to be sequenced is prepared as a single-stranded molecule so that it can act as a template for DNA synthesis (see Topic F3) in the sequencing reaction. ***E. coli* DNA polymerase I** is used to copy this DNA template. However, this enzyme needs a primer to start synthesis (see Topic F3). The primer used can be either a DNA restriction fragment complementary to the single-stranded template or it can be a short sequence of complementary DNA that has been synthesized chemically (a **synthetic oligonucleotide**).

Chain termination method

An incubation mixture is set up containing the single-stranded DNA template, the primer, DNA polymerase I and all four deoxyribonucleoside triphosphates (dATP, dGTP, dCTP, dTTP), one of which is radioactively labeled, plus a single 2′3′ dideoxyribonucleoside triphosphate analog, say ddGTP. In this incubation,

the DNA polymerase begins copying template molecules by extending the bound primer. As the new DNA strand is synthesized, every time that dGTP should be incorporated there is a chance that ddGTP will be incorporated instead. If this happens, no further chain elongation can occur because dideoxy analogs lack the 3′-OH group needed to make the next 3′5′ phosphodiester bond. Thus this particular chain stops at this point. In this first incubation mixture, a large population of templates is being copied and each new strand will stop randomly at positions where a G must be added to the newly synthesized strand. Thus, for every G in the complementary sequence there will be some new DNA strands that have terminated at that point (*Fig. 1*).

In fact, four incubation mixtures are set up, each containing the same components except that each contains a different dideoxy analog; one of ddATP, ddCTP, ddTTP or ddGTP (*Fig. 1*). This produces four sets of chain-terminated fragments corresponding to the positions of A, C, T and G in the sequence. After the incubation, all four reaction mixtures are electrophoresed in parallel lanes of a polyacrylamide gel and then subjected to autoradiography. The DNA sequence is determined simply by reading the band pattern on the autoradiogram from the bottom of the gel towards the top (i.e. reading the DNA sequence as it is synthesized from the primer; *Fig. 1*). The sequence read off the

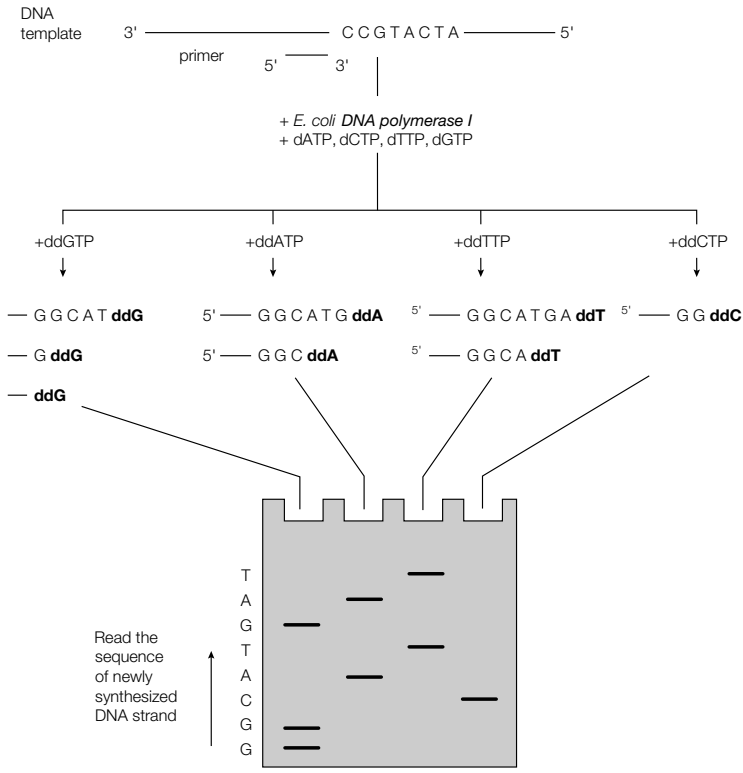

Fig. 1. DNA sequencing by the chain termination (Sanger) method.

gel is the sequence of the synthesized DNA strand and hence is the complementary sequence to the original DNA template strand.

Automated DNA sequencing

Automated DNA sequencing is now commonplace, based on the chain termination method but using a fluorescent dye attached to an oligonucleotide primer instead of using radioactive labeling. A different fluorescent dye is attached to the primer for each of the four sequencing reactions but, after incubation, all four mixtures are combined and electrophoresed on one gel lane. Laser detection systems then distinguish the identity of each termination product as it elutes from the gel. The sequence in which the different fluorescent products elutes from the gel gives the DNA sequence.

Modern automated DNA sequencing systems designed to generate over half a million bases of sequence per day use sequencing gels contained in multiple capillary tubes (rather than in a slab gel format). Preparation and loading of the samples onto the capillary gels is carried out by robots and data analysis is also automated.

I6 Polymerase chain reaction

Key Notes

Principles of PCR

The polymerase chain reaction (PCR) allows an extremely large number of copies to be synthesized of any given DNA sequence provided that two oligonucleotide primers are available that hybridize to the flanking sequences on the complementary DNA strands. The reaction requires the target DNA, the two primers, all four deoxyribonucleoside triphosphates and a thermostable DNA polymerase such as *Taq* DNA polymerase. A PCR cycle consists of three steps; denaturation, primer annealing and elongation. This cycle is repeated for a set number of times depending on the degree of amplification required.

Applications of PCR

PCR has made a huge impact in molecular biology, with many applications in areas such as cloning, sequencing, the creation of specific mutations, medical diagnosis and forensic medicine.

Related topics

DNA structure (F1)
DNA replication in bacteria (F3)
Restriction enzymes (I2)
Nucleic acid hybridization (I3)
DNA cloning (I4)
DNA sequencing (I5)

Principles of PCR

PCR (polymerase chain reaction) is an extremely simple yet immensely powerful technique. It allows enormous amplification of any specific sequence of DNA provided that short sequences either side of it are known. The technique is shown in *Fig. 1*. A PCR reaction contains the target double-stranded DNA, two primers that hybridize to flanking sequences on opposing strands of the target, all four deoxyribonucleoside triphosphates and a DNA polymerase. Because, as we shall see, the reaction periodically becomes heated to high temperature, PCR depends upon using a heat-stable DNA polymerase. Many such heat-stable enzymes from thermophilic bacteria (bacteria that live in high temperature surroundings) are now available commercially. The first one used was *Taq* polymerase from the thermophilic bacterium *Thermus aquaticus*.

PCR consists of three steps:

- **Denaturation**. The reaction mixture is heated to 95°C for a short time period (about 15–30 sec) to denature the target DNA into single strands that can act as templates for DNA synthesis.
- **Primer annealing**. The mixture is rapidly cooled to a defined temperature which allows the two primers to bind to the sequences on each of the two strands flanking the target DNA. This **annealing temperature** is calculated carefully to ensure that the primers bind only to the desired DNA sequences. One primer binds to each strand (*Fig. 1*). The two parental strands do not re-anneal with each other because the primers are in large excess over parental DNA.

- **Elongation**. The temperature of the mixture is raised to 72°C (usually) and kept at this temperature for a pre-set period of time to allow DNA polymerase to elongate each primer by copying the single-stranded templates. At the end of this incubation, both single-stranded template strands have been made partially double stranded. The new strand of each double-stranded DNA extends for a variable distance downstream.

The three steps of the PCR cycle are repeated. Thus in the second cycle, the four strands denature, bind primers and are extended. No other reactants need to be added. The three steps are repeated once more for a third cycle (*Fig. 1*) and so on for a set number of additional cycles. By the third cycle, some of the PCR products (indicated by asterisks in *Fig. 1*) represent DNA sequence only between the two primer sites and the sequence does not extend beyond these sites. As more and more reaction cycles are carried out, this type of double-stranded DNA molecule becomes the majority species present. After 20 cycles,

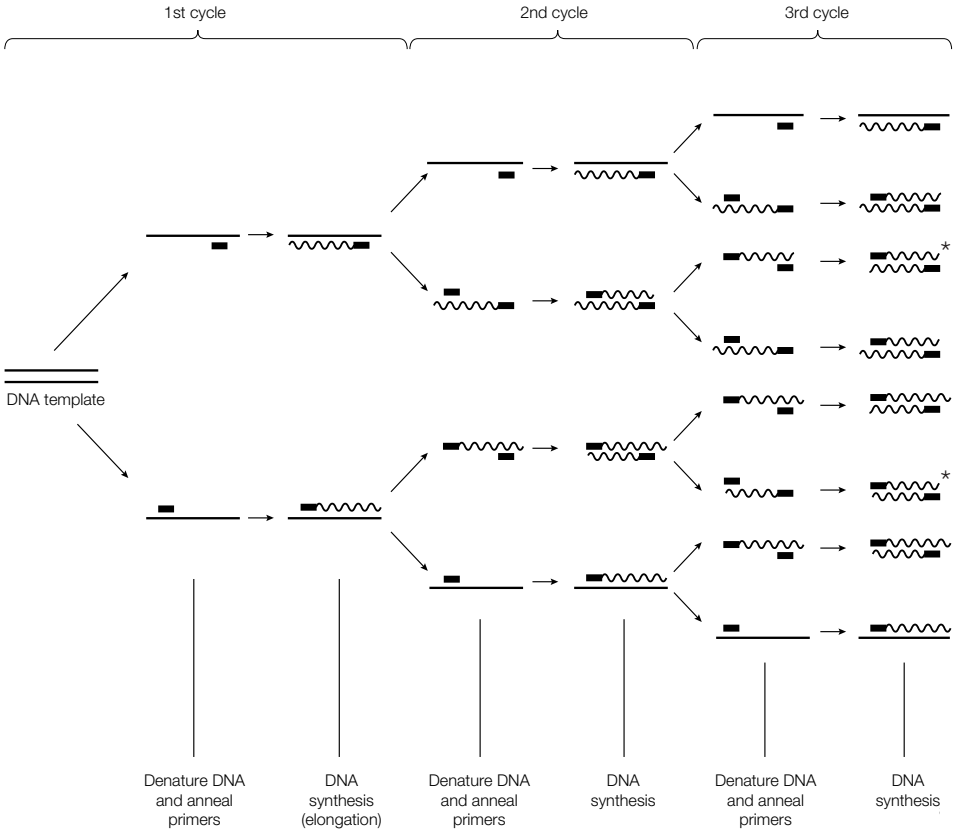

Fig. 1. The polymerase chain reaction (PCR). Asterisks indicate the first PCR products that arise (in the third cycle) which consist of DNA sequence only between the two primer sites.

the original DNA has been amplified a million-fold and this rises to a billion-fold (1000) million after 30 cycles. At this point the vast majority of the products are identical in that the DNA amplified is only that between the two primer sites. Automated **thermocyclers** are now routinely used to cycle the reaction without manual interference so that a billion-fold amplification of the DNA sequence between the two primer sites (30 cycles) can take less than one hour!

Applications of PCR

PCR already has very widespread applications, and new uses are being devised on a regular basis. Some (and certainly not all) of the applications are as follows:

- PCR can amplify a single DNA molecule from a complex mixture, largely avoiding the need to use DNA cloning to prepare that molecule. Variants of the technique can similarly amplify a specific single RNA molecule from a complex mixture.
- DNA sequencing has been greatly simplified using PCR, and this application is now common.
- By using suitable primers, it is possible to use PCR to create point mutations, deletions and insertions of target DNA which greatly facilitates the analysis of gene expression and function.
- PCR is exquisitely sensitive and can amplify vanishingly small amounts of DNA. Thus, using appropriate primers, very small amounts of specified bacteria and viruses can be detected in tissues, making PCR invaluable for medical diagnosis.
- PCR is now invaluable for characterizing medically important DNA samples. For example, in screening for human genetic diseases, it is rapidly replacing the use of RFLPs (see Topic I2). The PCR screen is based on the analysis of **microsatellites**. These are di-, tri- and tetranucleotide repeats in the DNA of the type $(CA)_n$ or $(CCA)_n$ where n is a number from 10 to more than 30. The microsatellites can be used as markers in the same way that RFLPs were used in the past (see Topic I2). Thus two primers are chosen that bind to the DNA flanking the microsatellite. PCR is then carried out and the different sizes of microsatellite give different sizes of amplified DNA fragments that can then be used as screening markers. The method is very fast, reliable and uses very small amounts of clinical material.
- Because of its extreme sensitivity, PCR is now fundamentally important to forensic medicine. It is even possible to use PCR to amplify the DNA from a single human hair or a microscopic drop of blood left at the scene of a crime to allow detailed characterization.