**Molecular taxonomy**

It is possible to distinguish between even closely related organisms using a range of biochemical tests, but this approach does not necessarily give an accurate picture of the true taxonomic or evolutionary relationship between different organisms.

Conventional taxonomy therefore resulted in some quite different organisms being erroneously grouped together in the same genus or family. Molecular approaches have played an important role in resolving these issues.

One simple molecular characteristic that is used to classify bacteria is the base composition of the DNA, defined as the number of guanine and cytosine residues as a percentage of the total number of bases (%GC). It is not necessary tosequence the genome to determine this value (although of course that gives the most accurate and precise value). The %GC can be determined using physical techniques. The base composition of bacterial DNA varies widely from one species to another – over a range of 20–80 per cent – but closely related organisms tend to have similar DNA base compositions. There are good reasons for this.

Replication, transcription and translation are all, in different ways and to varying extents, sensitive to the base composition of the nucleic acids and so have evolved together. Earlier in this chapter it became evident that this can be used as evidence that a portion of the genome has been acquired more recently in the evolution of the organism, as its base composition is different from the rest of the genome.

If this concept is applied to bacterial taxonomy, it is sometimes found that organisms which are otherwise quite similar have quite different DNA base composition. For example the genera Staphylococcus and Micrococcus are morphologically similar Gram-positive cocci (although they can be distinguished biochemically). However Micrococcus has a high GC content (about 70 per cent GC), while Staphylococcus DNA has a low proportion of G‏C (30–39 per cent).

Ribosomal RNA sequencing is a much more powerful technique. The ribosomal RNA genes are very highly conserved, being remarkably similar in all bacteria, and yet there are small variations in the sequence from one species to another. These variations (most commonly in the 16S rRNA) not only distinguish between species but also indicate the degree of difference. In other words, by counting the number of bases that are different in two species a measure of the evolutionary distance that separates them can be calculated. If a number of such sequences is compared, a phylogenetic tree can be constructed which will show a possible route by which these species have diverged from a common ancestor.

A simple example (with a much shorter sequence than would be used in practice) is shown in Figure 9.12. The sequence of organism A is more similar to B than it is to C or D for example and this is reflected in the arrangement of the tree. A word of caution: construction of a phylogenetic tree is much more complicated than this simple description and many trees can be drawn from a single set of data. The computer produces the best fit, but it is only a model and does not necessarily reflect the true evolution of the organisms involved.

Cloning the ribosomal RNA genes to do this is not necessary. The variation in the 16S (or 23S) rRNA gene is not evenly spread across the gene. Some regions are particularly highly conserved, so a pair of PCR primers can be used which recognize conserved sequences on either side of a variable region and amplify the region which contains differences. This amplified product can then be sequenced . The degree of conservation is such that the same pair of primers can be used for any organism, without knowing anything about it. The sequence obtained can then be compared with sequences of rRNA from known organisms and thus the identity of the unknown bacterium and its relationship to known species can be determined, at least provisionally.



The power of PCR to amplify minute amounts of DNA means that the bacterium in question does not have to be cultured. This is significant as standard bacteriological techniques are designed to culture certain bacteria, especially medically important pathogens. The range of bacteria that can be isolated can be extended by using different media and different growth conditions. But however wide the range of conditions used, there will still be some bacteria – often a substantial majority – that are unable to grow. Applying PCR to such a sample, using primers directed at the 16S rRNA gene, will produce a very wide range of amplified products. Cloning this mixture of products, rather like constructing a gene library , enables each one to be isolated and sequenced so that the bacteria present in the sample can be identified (within the limitations of the known sequences in the database). For example, the bacterial flora of the human colon has been extensively investigated using cultural techniques and its constituent bacteria were thought to have been thoroughly characterized. However, the genotypic approach described above showed that 76 per cent of the 16S rRNA genes generated did not correspond to known organisms and were clearly derived from hitherto unknown and uncultured bacterial species.

**Diagnostic use of PCR**

Traditional methods for the detection and identification of bacteria rely on growing the organism in pure culture and identifying it by a combination of staining methods, biochemical reactions and other tests. This applies equally to detection of environmental organisms (in soil or water), bacteria in food (including milk and drinking water) or pathogens in samples from patients with an infectious disease. However these methods are slow, requiring at least 24 h or several weeks for slow-growing organisms such as Mycobacterium tuberculosis. In addition, there are some bacteria, such as Mycobacterium leprae (the causative agent of leprosy) that still cannot be grown in the laboratory.

In principle, gene probes could be used to provide quicker results by directly detecting the presence of specificDNAin the specimen. However, this only works if the bacteria present are plentiful. Gene probes are not sensitive enough to detect the small numbers of organisms that may be present, and significant, in such specimens.

The technique that is needed is the polymerase chain reaction (PCR) . This provides greatly enhanced sensitivity, being capable (in theory) of detecting a single organism. In order to apply this to the detection of a specific species, it is necessary to know the sequence of a gene that is characteristic of that species – that is, it is always present (and the sequence is conserved) in that species, but is absent or significantly different in other bacteria. A pair of PCR primers can then be designed which will anneal to this target sequence so that PCR will amplify a DNA fragment that can be easily detected. Other bacteria, lacking the specific binding sites for those primers, will not give an amplified product. In a research laboratory the amplified product (amplicon) would commonly be detected by gel electrophoresis, sometimes combined with Southern blotting and hybridization with specific gene probes to increase the sensitivity and specificity of the procedure. The commercial kits that are now available for detection of some bacterial pathogens (some using forms of gene amplification that are distinct from PCR) use other, quicker, ways of detecting product amplification.

Atechnique known as real-time PCR which produces results more rapidly than gel electrophoresis and has the additional advantage of quantifying the target present in the sample .

**Using the Polymerase Chain Reaction to Amplify DNA**

A major problem in working at the molecular level is that each gene is a tiny fraction of the total cellular DNA. Because each gene is rare, it must be isolated and amplified before it can be studied. Before mid-1980, the only procedure available for amplifying DNA was gene cloning—placing the gene in a bacterial cell and multiplying the bacteria. Cloning is labor intensive and requires at least several days to grow the bacteria. In 1983, **Kary Mullis** discovered a new technique for amplifying DNA in a test tube. The **polymerase chain reaction** allows DNA fragments to be amplified a billion fold within just a few hours. It can be used with extremely small amounts of original DNA, even a single molecule. The polymerase chain reaction has revolutionized molecular biology and is now one of the most widely used of all molecular techniques.

**The polymerase chain reaction in outline**

PCR is based on ability of DNA polymerase to synthesize complementary strand to the template strand. As DNA polymerase can add a nucleotide only onto a 3'-OH group, it needs an artificial DNA strand (called DNA primer) of about 18 to 25 nucleotides complementary to 3’ end of the DNA template. As shown below, each polynucleotide has a free 3’ –OH group and 5’ phosphate group. Moreover, a DNA strand has complimentary sequence, already paired by hydrogen bonding. Thus, primer can bind only when DNA strands are separated. This is generally done by heating. The primers anneal to the single-stranded DNA template at specific temperature (depends on primer sequence) and then DNA-Polymerase binds to this double stranded DNA produced. The again reaction mixture is heated to 72°C (extension); a temperature optimum for DNApolymerase functions. This starts synthesis of the new DNA strand. Than reaction mixture is cooled to lower temperature for short term storage, if required. This completes one cycle. After first cycle, one DNA molecule has become two. After multiple cycle of the PCR reaction, the specific sequence will be accumulated in billions of copies.

**The PCR reaction requires the following components**:

 **DNA template:** DNA template is DNA target sequence. As explained earlier, at the beginning of the reaction, high temperature is applied to separate both the DNA strands from each other so that primers can bind during annealing.

**DNA polymerase:** DNA polymerase sequentially adds nucleotides complimentary to template strand at 3’-OH of the bound primers and synthesizes new strands of DNA complementary to the target sequence. The most commonly used DNA polymerase is Taq DNA polymerase (from Thermus aquaticus, a thermophillic bacterium) because of high temperature stability. Pfu DNA polymerase (from Pyrococcus furiosus) is also used widely because of its higher fidelity (accuracy of adding complimentary nucleotide). Mg2+ ions in the buffer act as co-factor for DNA polymerase enzyme and hence are required for the reaction.

**Primers:** Primers are synthetic DNA strands of about 18 to 25 nucleotides complementary to 3’ end of the template strand. DNA polymerase starts synthesizing new DNA from the 3’ end of the primer



Two primers must be designed for PCR; the forward primer and the reverse primer. The forward primer is complimentary to the 3’ end of antisense strand (3’-5’) and the reverse primer is complimentary to the 3’ end of sense strand (5’-3’). If we consider the sense strand (5’-3’) of a gene, for designing primers, then forward primer is the beginning of the gene and the reverse primer is the reverse-compliment of the 3’ end of the gene.

 **Nucleotides (dNTPs or deoxynucleotide triphosphates):** All types of nucleotides are "building blocks" for new DNA strands and essential for reaction. It includes Adenine(A), Guanine(G), Cytosine(C), Thymine(T) or Uracil(U).

**Procedure**

 There are three major steps in a PCR, which are repeated for 30 or 40 cycles. This is done on an automated cycler, which can heat and cool the tubes with the reaction mixture in a very short time.



**1. Denaturation at 94°C :** During the heating step (denaturation), the reaction mixture is heated to 94°C for 1 min, which causes separation of DNA double stranded. Now, each strand acts as template for synthesis of complimentary strand.

**2. Annealing at 54°C :** This step consist of cooling of reaction mixture after denaturation step to 54°C, which causes hybridization (annealing) of primers to separated strand of DNA (template). The length and GC-content (guanine-cytosine content) of the primer should be sufficient for stable binding with template. Please recall our discussion about DNA structure during earlier lectures. Guanine pairs with cytosine with three hydrogen bonding adenine binds with thymine with two hydrogen bonds. Thus, higher GC content results in stronger binding. In case GC content is less, length may be increased to have stronger binding (more number of H bonding between primer and template).

 **3. Extension at 72°C :** The reaction mixture is heated to 72°C which is the ideal working temperature for the Taq polymerase. The polymerase adds nucleotide (dNTP's) complimentary to template on 3’ –OH of primers thereby extending the new strand.



**4. Final hold:** First three steps are repeated 35-40 times to produce millions of exact copies of the target DNA. Once several cycles are completed, during the hold step, 4–15 °C temperature is maintained for short-term storage of the amplified DNA sample.

**PCR-an exponential cycle:** As both strands are copied during PCR, there is an exponential increase of the number of copies of the gene as shown in the figure. Suppose there is only one copy of the desired gene before the PCR starts, after one cycle of PCR, there will be 2 copies, after two cycles of PCR, there will be 4 copies. After three cycles there will be 8 copies and so on.



**Various PCR methods**

**Nested PCR** - use to synthesize more reliable product - PCR using a outer set of primers and the product of this PCR is used for further PCR reaction using an inner set of primers.

**Inverse PCR** - for amplification of regions flanking a known sequence. DNA is digested, the desired fragment is circularise by ligation, then PCR using primer complementary to the known sequence extending outwards.

**AP-PCR (arbitrary primed)/RAPD** (random amplified polymorphic DNA) - methods for creating genomic fingerprints from species with little-known target sequences by amplifying using arbitrary oligonucleotides. It is normally done at low and then high stringency to determine the relatedness of species or for analysis of Restriction Fragment Length Polymorphisms (RFLP).

 **RT-PCR (reverse transcriptase)** - using RNA-directed DNA polymerase to synthesize cDNAs which is then used for PCR and is extremely sensitive for detecting the expression of a specific sequence in a tissue or cells. It may also be use to quantify mRNA transcripts. See also Quantiative RT-PCR, Competitive Quantitative RT-PCR, RT in situ PCR, Nested RT-PCR.

 **RACE (rapid amplificaton of cDNA ends)** - used where information about DNA/protein sequence is limited. Amplify 3' or 5' ends of cDNAs generating fragments of cDNA with only one specific primer each (+ one adaptor primer). Overlapping RACE products can then be combined to produce full cDNA. See also Gibco manual.

**DD-PCR (differential display)** - used to identify differentially expressed genes in different tissues. First step involves RT-PCR, then amplification using short, intentionally nonspecific primers. Get series of band in a high-resolution gel and compare to that from other tissues, any bands unique to single samples are considered to be differentially expressed.

 **Multiplex-PCR** - 2 or more unique targets of DNA sequences in the same specimen are amplified simultaneously. One can be use as control to verify the integrity of PCR. Can be used for mutational analysis and identification of pathogens.

