Mendel’s studies of garden peas in the 18th century provided the first evidence of inheritance and so began the science of genetics.1 The one gene–one protein theory had been established by 1941 and in 1952 Hershey proved that these genes were made of deoxyribonucleic acid (DNA). This was followed quickly by Watson and Crick’s description of the double helix structure of DNA and the codon that expresses its language.2 In 1986, Mullis published details of the polymerase chain reaction to multiply lengths of DNA3 and the language and tools of today’s science of molecular biology were born.

**DNA and RNA structure and function**

The nucleus of human cells contain 22 pairs of chromosomes in addition to the sex chromosomes XX or XY, consisting of as many as 100 000 genes. Each chromosome consists of a tightly wound strand of DNA, regulating production of specific proteins, one gene for each protein. Genes and their regulatory components represent less than 10% of the total chromosomal material, the rest being repetitive and variable DNA sequences, thought to provide flexibility and protection. DNA and ribonucleic acid (RNA) are linear polymers of pentose sugars (deoxyribose for DNA and ribose for RNA). A purine or a pyrimidine base is attached to the 1’ carbon end of each pentose molecule. Two purines (adenine (A) and guanine (G)) and two pyrimidines (thymine (T) and cytosine (C)) are found in DNA. In RNA, uracil (U) is present instead of thymine, but otherwise the bases are the same as in DNA. The combination of a pentose and a base is called a nucleoside. With the addition of a phosphate group to the 5’ hydroxyl, the compound is called a nucleotide and these are the units of DNA and RNA.

RNA ranges in length from tens of bases to thousands while the number of units in DNA can be millions. When nucleotides polymerize to form nucleic acids, the hydroxyl group attached to the 3’ carbon of a sugar of one nucleotide forms an ester bond to the phosphate of another nucleotide (Fig. 1). Thus a single nucleic acid strand is a phosphate–pentose polymer with purine and pyrimidine bases as side groups. This gives each strand a definite polarity. This directionality has given rise to the convention that nucleotide sequences are written and read in the 5’ to 3’ direction (from left to right); for example the sequence AUG is assumed to be (5’)AUG(3’). The orientation of the nucleic acid strand is an extremely important property of the molecule.

Double helix DNA consists of two strands that wind around each other; the base pairs are stacked in between
the strands (Fig. 2). The geometry of the sugar phosphate backbone favours a right-handed spiral conformation and it is this form that is most commonly seen in natural DNA. The orientation of the two strands is complementary and anti-parallel. They are complementary because adjacent nucleotides always pair in a specific way (adenine to thymine and guanine to cytosine) so that the base sequence of one strand defines that of the other. They are anti-parallel because the orientation of their 5' to 3' directions are opposite. The strands are held together by hydrogen bonds and hydrophobic interactions. This relatively weak bonding means that the two strands can separate at specific locations so that each strand can serve as a template for the formation of another complementary strand. The process of breaking and rejoining is called hybridization. The double strands of DNA can also be separated and rejoined (annealed) by heating and cooling. DNA replication begins with unwinding and separation of the two strands at a specific location by the action of a polymerase enzyme. The two DNA templates are 'read' from the 3'-5' end with the new DNA (or RNA) strand being synthesized from the 5'-3' end in short sections, joined together by a ligase enzyme (Fig. 3).

Proteins are coded by the arrangements of bases, in regions of DNA called exons; intervening non-coding sequences called introns separate the exons. Regulatory regions respond to hormonal or other regulatory signals by turning on or off production of protein from each particular gene. Synthesis of pre-mRNA complementary in base pair sequence to the DNA by the action of RNA polymerase is called transcription. The introns are then removed such that messenger RNA (mRNA) consists only of coding regions. Other modifications include addition of a poly-adenosine (poly A) tail, and a 7-methylguanylate cap, which aid stability and cellular transport of the mature mRNA. mRNA is then transferred from the nucleus to the ribosomes in the cytoplasm, the organelles which are the site of protein synthesis.

RNA has many functions and structures. Most RNA is single-stranded with possible conformational structures which in many ways resemble proteins. The bulk of RNA in cells is bound to proteins in large complexes of very specific structure. Single strand RNA is the genetic material of many viruses. It is thought that in early evolution, RNA served both as a genetic store in addition to having enzyme functions, with DNA appearing much later.

Protein synthesis

The relationship in cells between the synthesis of DNA, RNA and protein is circular. DNA directs the synthesis of RNA and RNA then directs the synthesis of protein; special proteins catalyse the synthesis of both RNA and DNA. Proteins are the active working components of the cellular machinery. Whereas DNA stores the information for protein synthesis and RNA carries out the instructions encoded in DNA, most biological activities are carried out by proteins; their synthesis is the heart of cellular function.

Proteins are made by ribosomal RNA (rRNA) with the
help of transfer RNA (tRNA) which decodes mRNA and directs the activity of rRNA (translation). DNA and mRNA are considered in units of three base pairs called codons, with each codon potentially directing the insertion of one amino acid into the protein being synthesized. Only 61 of the 64 possible codons specify individual amino acids, and because there are 61 codons for 20 amino acids, many amino acids have more than one codon. The different codons for a given amino acid are said to be synonymous. The genetic code itself is termed degenerate, which means that it contains redundancies (Table 1). The start (initiator) codon AUG specifies the amino acid methionine and all protein chains begin with this amino acid. The three codons UAA, UGA and UAG do not specify amino acids but constitute stop (terminator) signals at the ends of protein chains. All organisms use the same codons, so that an organism as simple as bacteria can translate human DNA to produce a hormone product such as insulin. If either the DNA or RNA sequence is known, then the protein sequence can be derived.

Table 1 The degeneracy of the genetic code. Total number of codons for amino acids, 61; number of stop codons, 3; total number of codons, 64

<table>
<thead>
<tr>
<th>No. of synonymous codons</th>
<th>Amino acid</th>
<th>Total number of codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Leu, Ser, Arg</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>Gly, Pro, Ala, Val, Thr</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>Ile</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Phe, Tyr, Cys, His, Gin, Glu</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>Asn, Asp, Lys, Met, Trp</td>
<td>2</td>
</tr>
</tbody>
</table>

DNA or RNA translation (transcription) results in functionally mature proteins. The three codons UAA, UGA and UAG do not specify amino acids but constitute stop (terminator) signals at the ends of protein chains. All organisms use the same codons, so that an organism as simple as bacteria can translate human DNA to produce a hormone product such as insulin. If either the DNA or RNA sequence is known, then the protein sequence can be derived.

Transfer RNA (tRNA) delivers the appropriate amino acids to the rRNA–mRNA complex for assembly into polypeptides. Finally, post-translational modification such as glycosylation (addition of sugar molecules) or polymerization then results in functionally mature proteins. The triplets in an mRNA molecule do not select amino acids directly. Instead, the protein synthesizing system uses tRNA as an adaptor molecule to translate the information in each mRNA triplet so that the appropriate amino acid is added to the chain. Each tRNA molecule must be recognized by one of 20 enzymes (amino acyl tRNA synthetases) and each of these enzymes adds one of the 20 amino acids to tRNA.

The mRNA with its encoded information and the individual tRNAs loaded with their correct amino acids are brought together by their mutual binding to the ribosome. For the polypeptide chain to begin to grow, a second amino acyl-tRNA must be positioned correctly on the ribosome. This is the function of one of the elongation factors. Two ribosomal sites are occupied by tRNA molecules: the A site accommodates the incoming amino acyl-tRNA that is to contribute a new amino acid to the growing chain; the P site contains the peptidyl–tRNA complex (the tRNA still linked to all the amino acids added to the chain so far). A peptide bond is formed between the carboxyl group of one amino acid and the amino group of the other. In this process the amino acid occupying the A site then moves along to now occupy the P site displacing the amino acid previously occupying the P site. Hydrolysis of GTP provides the necessary energy for this process.

Techniques in molecular biology

Separation and identification of proteins (Table 2)

Nucleic acids, peptides or proteins in a mixture can be separated according to size by electrophoresis. This technique depends on the fact that dissolved molecules in an electric field move at a speed determined by their charge:mass ratio. For example, two molecules with the same mass can be differentiated if one has a higher electrical charge as it will move faster towards an electrode. A small amount of sample is placed on a strip of filter paper or some other porous material which is then soaked with a conducting solution. When an electric field is applied at the ends of the strip, small molecules dissolve in the conducting solution and move along the strip at a rate corresponding to their charge.

Nucleic acids generally have a negative charge because their phosphate groups are ionized and they thus migrate towards a positive electrode. However, nucleic acids often have the same charge:mass ratio whatever their length because each charged residue contributes approximately the same mass and charge. Many proteins also have similar charge:mass ratios. Electrophoresis is now commonly performed in gels in which the pores in the gel limit the rate at which molecules can move through the medium so that some separation also occurs in relation to molecular length. Denaturing electrophoresis is when the gel contains a substance such as sodium dodecyl sulphate (SDS, a detergent which denatures the separated proteins).

Western blotting is a technique used to identify specific proteins. Proteins are separated by electrophoresis on a polyacrylamide gel, blotted by capillary transfer onto a
nylon membrane and then identified using radio- or enzyme-labelled antibodies to the proteins of interest. An example of the end product of western blotting is shown in Figure 4.

Recombinant DNA techniques

Restriction enzymes
Much of molecular biology is directed at the identification, multiplication (cloning) and analysis of genes. As genes are made up of the same four nucleotides, they cannot be readily identified, and therefore the gene in question must be physically removed from the chromosomal DNA and amplified. To remove a piece of DNA, restriction enzymes are required. These enzymes are naturally generated in bacteria where they serve a protective function, cutting up foreign DNA at specific sites. Each enzyme, of which hundreds have been identified, cleaves the DNA into distinct patterns depending on the DNA sequence. The pattern of DNA fragments is characteristic for each chromosome, as the enzyme always cuts the DNA at its specific site of action and nowhere else. Southern blotting (see below) can be used for restriction enzyme mapping. Variation within the population of the size of restriction fragments can be used clinically in a variety of ways, including identification of specific markers for clinical conditions. Knowledge of restriction enzyme maps in clinical conditions has enabled the use of restriction fragment length polymorphism in screening for genetic disorders such as the thalassaemias, muscular dystrophy or sickle cell anaemia. Polymorphism is defined as the presence of more than one structural gene at a single genetic locus within a population. Every individual has a distinct pattern of DNA fragments (DNA fingerprinting) now used widely in the field of forensic medicine.

Pharmacogenetics

The field of pharmacogenetics or pharmacogenomics or the study of the genetic basis of the response to xenobiotics (drugs, environmental or toxicological agents) is expanding rapidly. For example, polymorphic variation in the genes for the cytochromes P-450 enzymes can affect drug metabolism leading to excessive or decreased therapeutic effects, or toxicity as a result of altered metabolism. Alcohol is metabolized to acetaldehyde by cytochrome P-450E1 and genetic variations in its activity can influence susceptibility to alcoholic liver disease. In addition, pharmacogenetics may also provide the biological basis of the ethnic differences in the metabolism of nicotine in smokers, and the racial variation in cancer incidence. Pharmaceutical companies are also becoming interested in the concept of customized drugs for subgroups of patients identified by polymorphic screening, such as cancer therapy, where genetic variation in enzymes leads to different metabolism of chemotherapeutic agents and different responses in different patients.

Recombinant DNA

To produce specific quantities of DNA for analysis or other purposes, DNA is cut with restriction enzymes and the fragments separated by electrophoresis. Using the enzyme DNA ligase, cut pieces of DNA can be made to rejoin because of base pairing. As rejoining occurs at complementary base pairs, the pieces of DNA are referred to as sticky ends of the DNA. The DNA fragments with sticky ends can be amplified by inserting them into a segment of DNA capable of independent growth, called a vector. Bacterial plasmids (small circular segments of non-chromosomal self-replicating DNA) and bacteriophages (viruses which infect bacteria) are commonly used vectors. The plasmids are cut with restriction enzymes and then incubated with DNA fragments cut with the same enzymes. The sticky ends attach to the opened plasmid DNA and the result is DNA containing both plasmid and foreign (human) DNA (i.e. recombinant DNA). When re-introduced back into bacteria (transfection) the DNA replicates itself, forming clones of a specific piece of human DNA.

Although plasmids are useful to amplify a given piece of DNA containing a discrete gene, it is difficult to screen large numbers of bacteria to see which one contains the desired gene. Screening large numbers of different cloned DNA segments is best done with the plasmid, phage λ. Pieces of foreign (human) DNA are recombined with DNA from phage λ using DNA ligase. An agar plate with a thin
layer of bacteria is then infected with the virus. The phage infects the bacteria, killing them and leaving phage particles containing identical DNA fragments, which can then be hybridized with a labelled DNA probe.

Probes are DNA or RNA sequences complementary to, and so will hybridize to, the gene of interest. The nucleotide configuration ensures specificity despite the presence of non-complementary sequences. Probes are labelled with either radioactive or chemiluminescent markers to enable detection. Probes are made using primers (short sequences of nucleotides corresponding to each end of the DNA or RNA of interest and which are used to extend the nucleotide chain in both directions to produce longer sequences for use as probes).

Cloning of a gene allows its function and regulation to be studied, and its product (one gene–one protein) to be produced in large quantities. For example, E. coli transfected with the DNA encoding somatostatin produces the protein for clinical use.

Analysis of DNA and RNA

Southern and northern hybridization or ‘blotting’

Southern blotting for detecting DNA is named after its inventor E. M. Southern from Oxford University, while northern blotting is a tongue in cheek name given to the same technique but for the detection of RNA. In Southern blotting, restriction fragments of DNA are electrophoresed on a denaturing gel and the denatured DNA bands are blotted onto nitrocellulose filters or charged nylon membranes. This is achieved by capillary action under the pressure of a stack of absorbent paper towels and a heavy book. The blots are then probed with labelled probes specific to the area of interest and detected by either exposure of the blot to a photographic plate or, more recently, by phosphor imaging.

In contrast, northern blotting detects specific mRNA sequences. Total cytoplasmic RNA can be isolated using phenol–chloroform–guanidinium thiocyanate extraction in the presence of an RNAse inhibitor. (RNAses are enzymes which rapidly degrade RNA.) mRNA can be purified using oligo-dT cellulose columns which capture the poly-A tail present on all eukaryotic RNA. Relative quantification of mRNA by northern blotting can be by densitometry or by phosphor imaging, provided the same amount of RNA is used as a starting point (Fig. 5). In this way, an increase in production of mRNA to, for example, the cytokine tumour necrosis factor-α, in response to an infection can be detected. Monitoring of the mRNA to a protein rather than the protein itself may suggest at which level regulation may be occurring. Dot blotting is when RNA is placed directly onto the membrane thus missing out the electrophoresis and transfer steps. This method is not as specific as Southern or northern blotting because the species of interest is identified by hybridization alone, and not by molecular weight and charge.

In situ hybridization

In situ hybridization allows the detection of specific mRNA sequences within tissue sections. Routinely unfixed or formalin-fixed and paraffin wax-embedded tissue are treated with protein-K, which partially digests tissue and exposes (unmasks) target sequences. Probes labelled with chemiluminescent or enzyme markers are used to detect the mRNA of interest using light or fluorescence microscopy. Specific genes of interest can also be detected within cells on intact chromosomes, called fluorescent in situ hybridization or FISH. An example of the image obtained by this technique is shown in Figure 6.

Polymerase chain reaction

The polymerase chain reaction (PCR) is a powerful enzymatic technique for the amplification of minute specific DNA sequences in vitro from pure or impure preparations from fixed or unfixed tissue or cells. The process is performed in a test tube containing small quantities of the DNA to be replicated and primers complementary to each end of the sequence to be replicated (generally 50–2000) base pairs in length. A heat-stable DNA polymerase (commonly ‘Taq polymerase’ from Thermus aquaticus, a bacterium which lives in hot springs) and a large excess of bases are also included. The tube is then heated to separate the DNA strands, then cooled so that the primers anneal to the ends of the specific DNA strands by base pairing. The polymerase enzyme now synthesizes a single strand of DNA beginning
Fig 6 Fluorescence in situ hybridization on human cancer cell lines using a probe targeted to the centromere of chromosome 18. (Courtesy of S Marsh and H McLeod, Department of Medicine and Therapeutics, University of Aberdeen, Aberdeen, UK.)

Fig 7 Diagrammatic representation of a single cycle in the polymerase chain reaction. Taq is a heat stable DNA polymerase enzyme from Thermus aquaticus.

at each end of the primer attachment point. The excess bases included ensure that the process is complete. The tube is then re-heated and double-stranded DNA forms. The process of heating and cooling is repeated many times to produce an exponential increase in identical DNA strands. The process is shown in detail in Figure 7.

PCR can also be used to amplify mRNA. To do this, a DNA strand complementary to the mRNA of interest (called
complementary or copy DNA, a single DNA strand without introns) must be first synthesized using an RNA-dependent polymerase called reverse transcriptase. The technique is called RT-PCR. The rest of the process is the same as for DNA PCR. In situ PCR involves amplification of RNA or DNA directly within tissue sections or cells. Visualization of an amplified signal is achieved using chemiluminescent or enzyme labels as for in situ hybridization. This technique allow the detection of very small amounts of the RNA or DNA of interest.

Transgenic animals
Animals in which foreign genes are inserted are called transgenic and often express traits coded by the genes such that they provide an experimental model of, for example, a clinical condition. The genes can be inserted biochemically, by microinjection or using retroviruses. Retroviruses are RNA viruses which convert host viral RNA to DNA and incorporate this DNA into host chromosomes. Retroviruses containing foreign genes of interest have been used as vectors for inserting genes into both somatic and germ cells. Coding sequences (exons) and adjacent regulatory sequences that direct transcription of a protein of interest along with regulatory sequences from another gene with high basal transcriptional activity can be inserted. This creates a hybrid gene which expresses higher levels of its protein than that normally found in nature. Such models can be used to study the effect of mediators of metabolic processes. For example, hybrid genes encoding growth hormone releasing factor were inserted into embryos from mice, to investigate the effect of chronic exposure on growing organisms. Genes can also be inserted into tumour cell lines, producing tumours which secrete the protein of interest. The tumour can then be transplanted into animals to simulate chronic exposure to the protein. When tumours secreting TNFα were transplanted into mice, the chronic effects of the protein were seen, including its ability to cause cachexia and death.

Knockout animals, commonly mice, because of the ease and speed of breeding, can be used to determine the phenotypic effects of missing genes, and also the overlapping functions or compensation between genes. The genomes of mice and humans are closely related and the detailed study of genetically altered mice can provide many insights into human biology and disease.

References
2 Watson JD, Crick FHC. A structure for deoxyribose nucleic acid. Nature 1953; 171: 737–8

Further reading