# 2 Nucleic acids and proteins: a review

KEY KNOWLEDGE

This chapter is designed to enable students to:

- enhance their knowledge and understanding of nucleic acids
- enhance their understanding of the structure of proteins and polymerisation
- develop an understanding of the structure of DNA, RNA and triplet codes
- explain the differences between structural genes and regulatory genes
- understand the lac operon.

(a)

FIGURE 2.1 In celebration of the 50-year anniversary of the discovery of the structure of DNA **(a)** Australian stamps depicting, from top to bottom, a chromosome, a DNA molecule and the base sequence of a DNA molecule **(b)** An English two-pound sterling coin showing the double helix structure of DNA.

The year 2003 marked a significant celebration. It was the 50-year anniversary of the discovery of the structure of DNA by James Watson and Francis Crick. The discovery was a giant step forward in the field of science and opened the way to extensive research that has given us the molecular biology knowledge, understanding and applications that we have today. Many countries around the world celebrated the anniversary through the issue of special stamps. Australia was one of those countries. The country in which Watson and Crick did their work, England, also issued a coin.
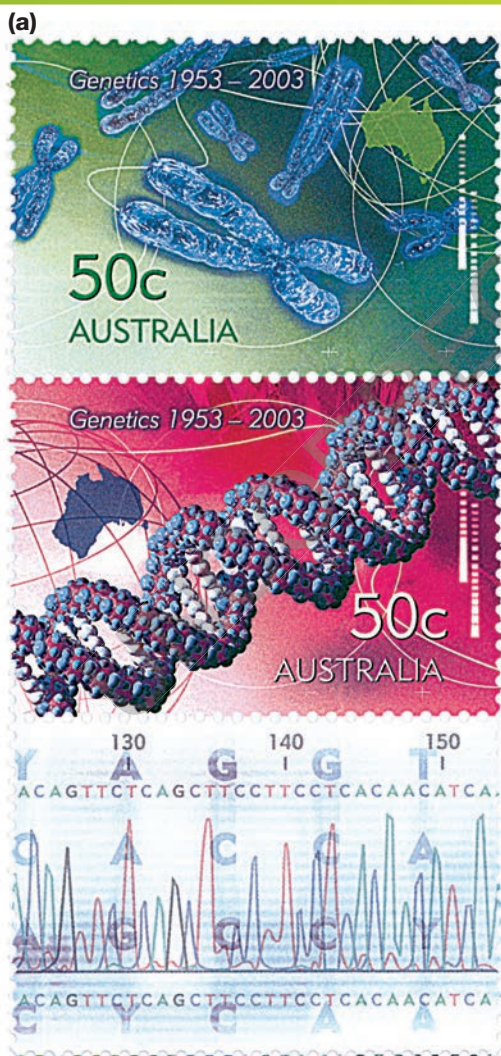
(b)

eBookplus

Interactivity
Organic compounds
int-3026

Organic molecules

Many organic molecules are made of large numbers of smaller sub-units that are linked together by specific covalent bonds. Nucleic acids, proteins and polysaccharides are examples of this type of organic molecule. **Compounds formed in this way are called polymers. The sub-units are called monomers.** The joining of monomers involves the release of a water molecule. Reactions of this kind are termed condensation reactions. We classify polymers on the basis of the kind of sub-unit they contain (refer to figure 2.2). These polymers are important to the functions of living organisms and, in the case of polysaccharides and proteins, also form part of their structures.

In addition, other organic compounds are involved in the structure and function of living organisms, with one important group being lipids. Lipids, which include fats, oils and phospholipids, are not regarded as polymers because they are not linked together in a chain of many repeating identical sub-units.
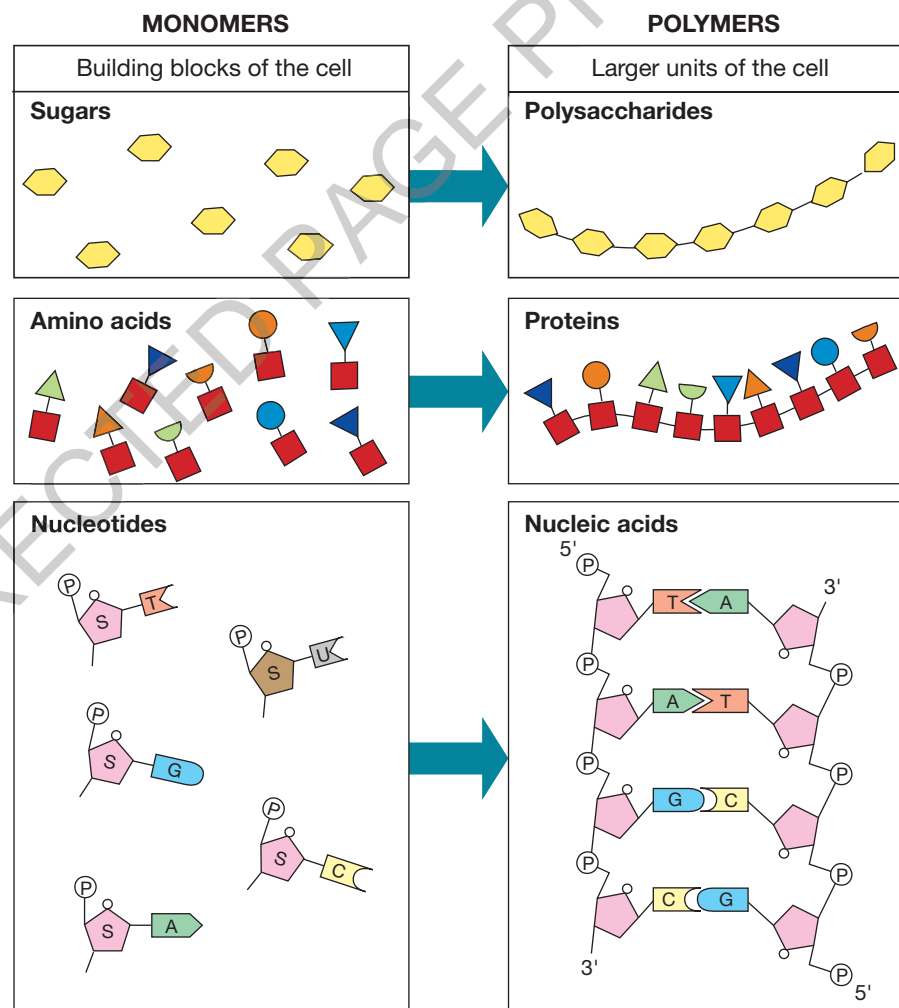

FIGURE 2.2 Three main groups of polymeric organic molecules present in cells. Note the type of monomer that makes up each type of polymer. Examples of polysaccharides include starch, glycogen and chitin; examples of protein groups include enzymes, immunoglobulins and contractile proteins; examples of nucleic acids include DNA and RNA.

The kinds of organic molecules that we will consider are **proteins** and **nucleic acids**. For each of these, we will examine:

- the basic unit of structure
- how the units combine to form complex molecules
- where each kind of molecule is found in a cell
- the functions of the molecules.

Nucleic acids

There are two kinds of nucleic acid:

- **deoxyribonucleic acid (DNA)**, which is located in chromosomes in the nucleus of eukaryotic cells (see figure 2.3). It is the genetic material that contains hereditary information and is transmitted from generation to generation.
- **ribonucleic acid (RNA)**, which is formed against a template strand of DNA.

DNA — an influential molecule

The genetic material **deoxyribonucleic acid is a polymer of nucleotides**. Each nucleotide unit, or monomer, has:

- a sugar (deoxyribose) part
- a phosphate part
- an N-containing base.

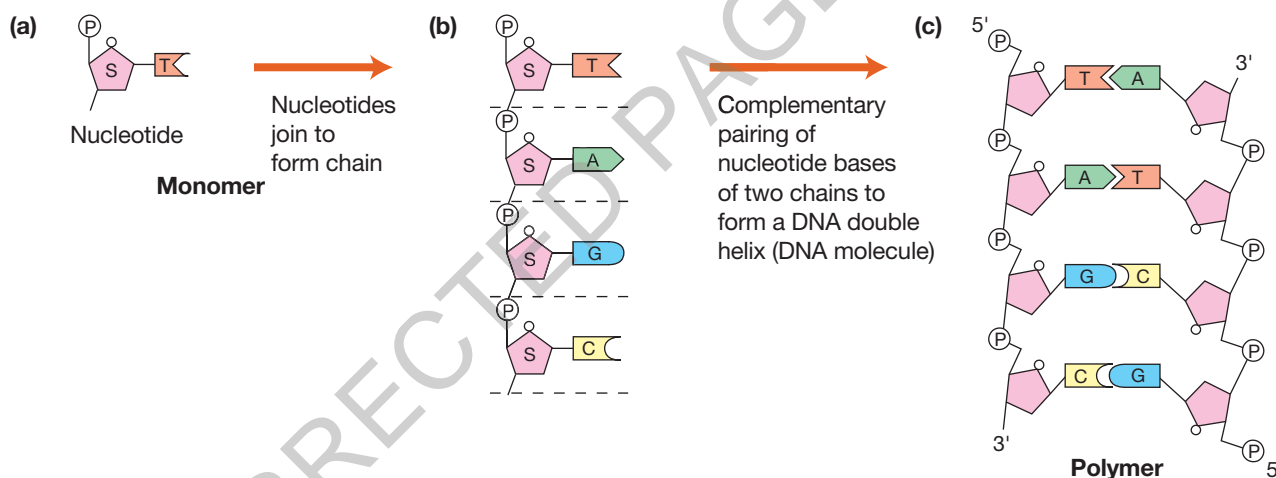**FIGURE 2.3** Deoxyribonucleic acid is made from nucleotide sub-units. Each DNA molecule is made of two complementary chains of nucleotides.
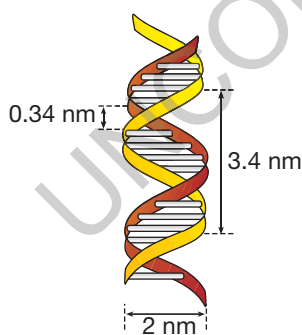
FIGURE 2.4 The double helix structure and dimensions of DNA. The two chains are held together by hydrogen bonds between complementary bases.

The sugar and phosphate parts are the same in each nucleotide.

There are four different kinds of nucleotides because **four different kinds of N-containing bases are involved:** adenine (A), thymine (T), cytosine (C) and guanine (G). Examine figure 2.3. The nucleotide sub-units (a) are assembled together to form a chain (b) in which the sugar of one nucleotide is linked to the phosphate of the next nucleotide in the chain. Each DNA molecule contains two chains (c) that form a double helix, with the bases in one strand pairing with the bases in the other strand. **The base pairs between the two strands, namely A with T, and C with G, are said to be complementary pairs.** The two chains form a double-helical molecule of DNA (see figure 2.4).

The DNA double helix combines with certain proteins, **in particular histones, as it condenses to form a chromosome** (see figure 2.5a). **As the DNA winds around clusters of histone proteins, it forms structures called nucleosomes** (see figure 2.5b).

ODD FACT

The total length of the DNA double helix molecule in an 'average' human chromosome is about five centimetres. By coiling and supercoiling, this long DNA molecule becomes compressed into a microscopic chromosome.

(a)

(b) nucleosome core particle

FIGURE 2.5 **(a)** Diagram showing the coiling and supercoiling of one molecule of a DNA double helix to form a eukaryotic chromosome. Note that a key process involves the coiling of DNA around histone proteins to form structures called nucleosomes that then become supercoiled. **(b)** A model of a nucleosome showing the DNA double helix (grey) coiling around a cluster of histone proteins (shown in colours).
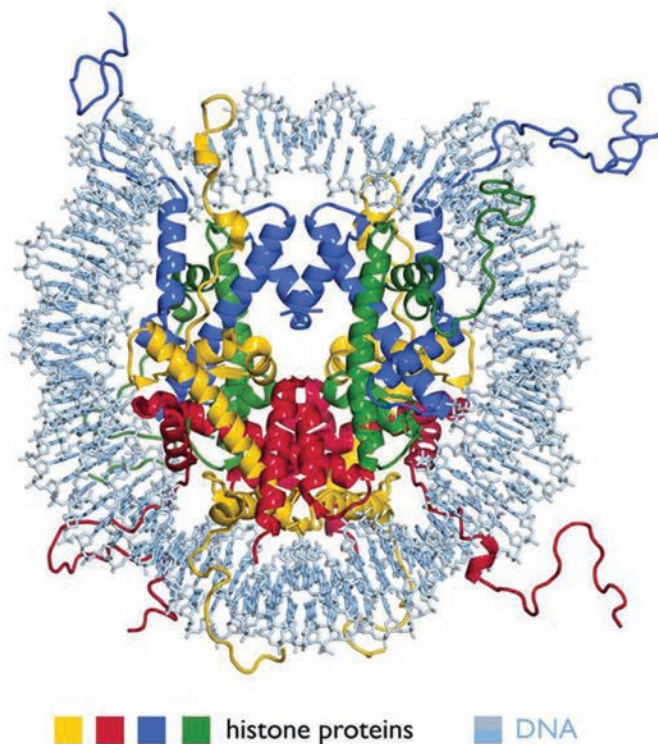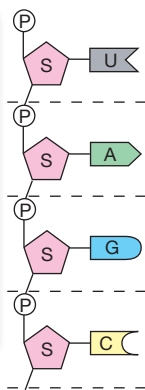
FIGURE 2.6 The four nucleotide sub-units, uracil, adenine, guanine and cytosine, from which the three forms of RNA are constructed

How does DNA control functions within cells? All metabolic reactions in cells are controlled by enzymes that, almost without exception, are proteins built of one or more polypeptide chains — chains of amino acids. Hence, **the DNA in the nucleus of a eukaryotic cell controls all the metabolic processes in a cell, through the polypeptide chains for which the DNA dictates production.**

Ribonucleic acid

Ribonucleic acid (RNA) is also a polymer of nucleotides (see figure 2.6). It differs from DNA in that it is **an unpaired chain of nucleotides** that contain the sugar ribose. RNA nucleotides include four different bases, three of which — adenine, guanine and cytosine — are identical to those in DNA. **The fourth nucleotide is uracil,** which is capable of pairing with A.

The three different forms of RNA are all produced in the nucleus alongside DNA as a template:

- **messenger RNA** (mRNA), which carries the genetic message to the ribosomes where the message is translated into a particular protein (see figure 2.7)
- **ribosomal RNA** (rRNA), which, together with particular proteins, makes the ribosomes found in cytosol
- **transfer RNA** (tRNA), molecules that carry amino acids to ribosomes where they are used to construct proteins.

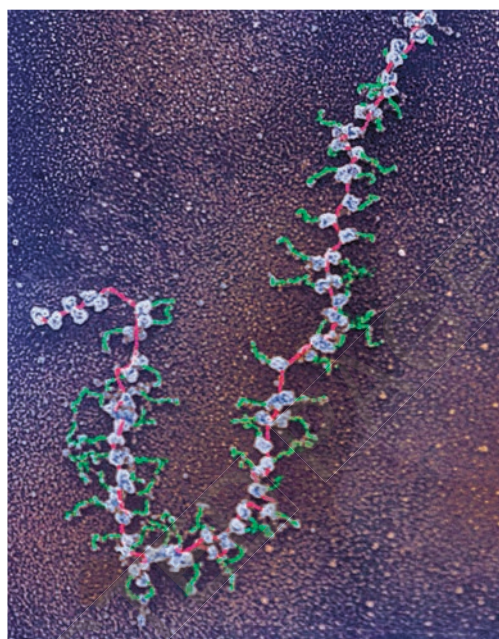The strand of nucleotides in each of the forms of RNA is folded in a different way.

FIGURE 2.7 A transmission electron micrograph (TEM) of a fragment of an mRNA translation unit from the salivary gland cell of a midge (*Chironomus* sp.). Ribosomes (blue) attach to the mRNA strand (pink) and read its code. A tRNA molecule carrying a corresponding amino acid binds to the ribosome. As the ribosome moves onto the next bases along the mRNA, a protein (green) grows from the ribosome.

KEY IDEAS

- Proteins, polysaccharides and nucleic acids are polymeric organic molecules built out of a very large number of repeating sub-units.
- The nucleic acids, double-helical DNA and single-stranded RNA are built out of a very large number of repeating sub-units called nucleotides.
- Each nucleotide consists of a sugar, a phosphate and an N-containing base, with the sugar in DNA being deoxyribose and that in RNA being ribose.
- Each DNA molecule consists of two chains of nucleotides that are complementary to each other and held together by hydrogen bonds.
- Each RNA molecule consists of a single strand of nucleotides.

QUICK CHECK

1 What do the initials DNA and RNA stand for?
2 What are the four kinds of nucleotides (a) in DNA and (b) in RNA?
3 Where in a cell would you find DNA and what is its function?
4 There are three different kinds of RNA molecules. What are they and where is each found in the cell?
5 In what ways do the kinds of RNA differ from one another?
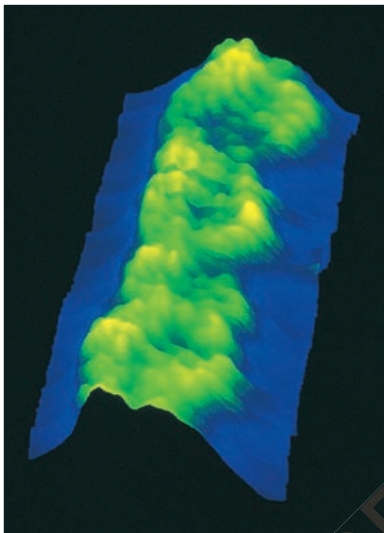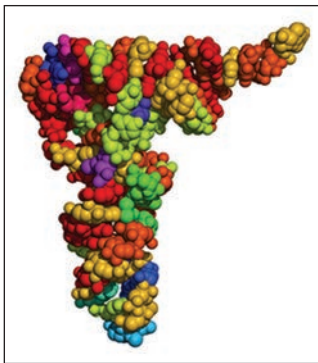
FIGURE 2.8 Part of a DNA double helix revealed through scanning tunnelling microscopy
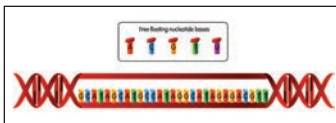
Gene structure

The template strand and its partner

A gene consists of a particular part of a double-helical molecule of DNA. **Only one of the two chains contains the information present in a particular gene, and this is called the** template strand. The complementary chain is sometimes called the nontemplate strand, also called the mRNA-like strand or the coding strand.

Representations of DNA

DNA cannot be seen with a light microscope. However, a technique known as scanning tunnelling microscopy (STM) allows DNA molecules to be visualised (see figure 2.8).

Part of a single chain of DNA could be shown as follows:

… -nucleotide-nucleotide-nucleotide-nucleotide-nucleotide- …

OR it could be shown as:

$$… \text{-P-sugar-P-sugar-P-sugar-P-sugar-P-sugar-} …$$
$$| \quad\quad | \quad\quad | \quad\quad | \quad\quad |$$
$$\text{base} \quad \text{base} \quad \text{base} \quad \text{base} \quad \text{base}$$

OR the specific bases in the nucleotides in one chain could be shown as:

$$5'…A\,T\,T\,A\,G\,C\,T\,T\,G\,A\,G\,G\,C\,G…3'$$

Which representation is correct? All are correct. Which representation is the most informative? The third is the most informative because it gives the information about the order of nucleotides, the only variable part of the genetic material. What information does the second representation provide?

DNA is not always represented in diagrams as a double helix. Figure 2.9 shows some of the many ways of representing DNA. The representation used will depend on the purpose of the diagram. Each provides different information about DNA. For example, (a) gives information about the coding regions (exons) and the noncoding regions (introns) within a gene, while (c) gives the base sequence.
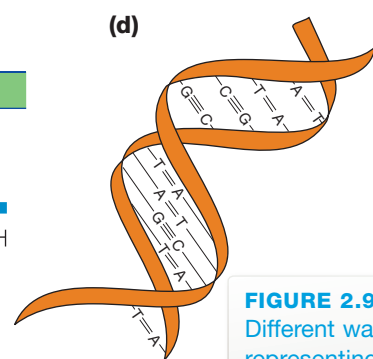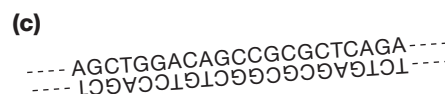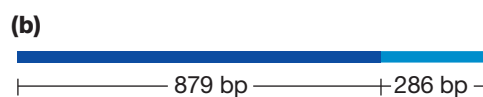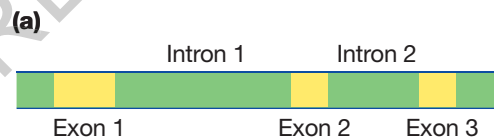
FIGURE 2.9 Different ways of representing DNA

Gene sequencing

What is this?

A T G G T G C A C C T G A C T C C T G A G G A G A A

This is part of the **nucleotide sequence** of the template strand of the human **HBB** gene, which encodes the information for one of the protein chains found in haemoglobin. What is the sequence of the complementary strand? **When the order of the nucleotides in a gene is identified, the gene is said to be**
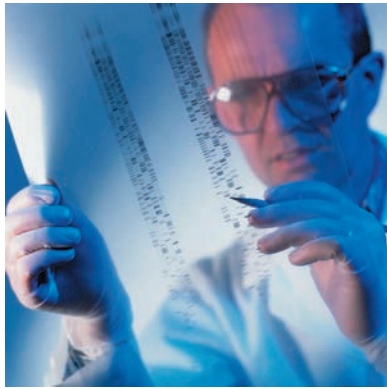
FIGURE 2.10 A scientist examining the sequence of bases in DNA

'sequenced'. The order in many genes from animals, plants and bacteria has been identified, and data continue to be added.

Gene sequencing involves the process of identifying the order of nucleotides along a gene. Figure 2.10 shows a scientist examining some sets of bands arranged in columns. Each band represents one nucleotide and the order of the bands down the column corresponds to the **gene sequence**. New techniques of sequencing are described in the box below.

Are gene sequencers used only for human genes? No! The genetic material of all organisms is DNA and the structure of that DNA is identical, regardless of whether it comes from wheat, jellyfish, ducks, *Bacillus* bacteria, insects or people. **In all organisms, genes are built of the same 'alphabet' of four letters**, namely, the nucleotides A, T, C and G of DNA. In contrast, non-cellular agents, such as hepatitis C and Ebola viruses, include many types that have RNA as their genetic material rather than DNA. Refer to chapter 6, page xxx for further detail.

So the genetic instruction kit to 'make a human being' or 'make an oak tree' or 'make a white shark' consists of thousands of instructions, each consisting of DNA with different base sequences.
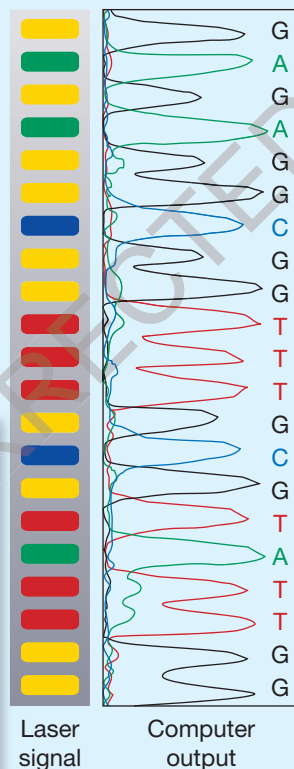
DNA SEQUENCERS

(a)

(b)

Laser signal   Computer output

FIGURE 2.11 (a) Applied Biosystems 3500 genetic analyser by Thermo Fisher Scientific **(b)** Output from a DNA sequencer showing the laser signals and the output from the computer, which identified the base sequence in part of a DNA fragment. What is signalled by a red band?

which binds to a specific base (A, T, C or G) in DNA. The DNA chain is sequenced using a stepwise procedure that makes a complementary copy using the unknown DNA as the template, with each copy being one nucleotide longer than the previous one as shown below:

Unknown DNA:	**CTCTCCGCCAAACGCATAACC**
1st copy	G*
2nd copy	GA*
3rd copy	GAG*
4th copy	GAGA*
etc.	
21st copy	GAGAGGCGGTTTGCGTATTGG*

In each case, the nucleotide at the end of each copy becomes attached to the specific fluorescent dye (shown as *). The copies move in turn, shortest first, past a scanning laser that activates the dye so that it emits a fluorescent signal, which is captured by a detector. This detector transfers the signal to a microcomputer, which determines the entire base sequence. The output from a DNA sequencer shows the base sequences as a series of coloured signals (see figure 2.11b) with a yellow peak denoting G, a red peak denoting T, a green peak denoting A and a blue peak denoting C.

The latest developments in sequencing are the next-generation DNA sequencers (NGS). These sequencers enable **millions of DNA fragments from tiny samples to be strung together** *at the same time*, **in contrast to the one-at-a-time method of the Sanger technique.**

The process of gene sequencing has now been automated and is done using instruments known as **DNA sequencers** (see figure 2.11a). One automated system, known as the Sanger method, involves the use of four different coloured fluorescent dyes, each of

ODD FACT

In October 2001, the cost of DNA sequencing was almost US$5000 per 1000 bases (1 megabase). By 2015, this cost had dropped to about 0.33 cents per raw megabase. (Data from National Human Genome Research Institute.)

Comparing DNA from different organisms

Databases hold information about the base sequences in the DNA of genes of many organisms. Table 2.1 shows part of various genes from different organisms — a duck, a *Bacillus* bacterium, a corn plant and a human being. Could you pick the human gene? It's not possible. The genes look similar because the genetic language of all living things is written in a common language based on an 'alphabet' of four letters (A, T, C and G) that correspond to the nucleotides (and bases) present in DNA.

TABLE 2.1 Part of the sequences of different genes from various organisms. Numbers are placed above the sequences for ease of locating a particular nucleotide.

	1			10			20			30			40	
Organism P	ATG	GCT	ACC	AAG	ATA	TTA	GCC	CTC	CTT	GCG	CTC	CTT	TCC	CTT	TTA
	1			10			20			30			40	
Organism Q	ATG	AAG	TGT	AAT	GAA	TGT	AAC	AGG	GTT	CAA	TTA	AAA	GAG	GGA	AGC
	1			10			20			30			40	
Organism R	ATG	ACG	CTG	ACT	CAA	GCT	GAG	AAG	GCT	GCC	GTG	ATC	ACC	ATC	TGG
	1			10			20			30			40	
Organism S	ATG	AGG	CTC	TTG	TGG	TTG	CTT	TTC	ACC	ATT	GGG	TTC	TGC	TGG	GCT

P: corn plant
Q: *Bacillus* bacterium
R: duck
S: human being

How do genes differ?

The total human DNA contains a 'make a human being' instruction kit; the total geranium DNA contains a 'make a flowering plant, geranium type' instruction kit. The human genetic instruction kit consists of tens of thousands of separate instructions, such as 'make the hair protein, keratin' and 'make growth hormone'. **Different genetic instructions within and between species are due to different nucleotide sequences in the genes.**

ODD FACT

Global databases with publicly available DNA, RNA and protein sequences include the National Center for Biotechnology Information (NCBI) database and the European Molecular Biology Laboratory (EMBL) database.

KEY IDEAS

- The length of a double-helical DNA molecule can be expressed as the number of base pairs (bp) it contains, and one chain of this DNA as the numbers of nucleotides.
- Each human chromosome contains one long molecule of double-stranded DNA with millions of base pairs.
- A typical gene consists of tens of thousands of nucleotides.
- The estimated total number of human genes is 21 000.
- Of the two DNA chains in a gene, the one containing the genetic information is known as the template strand of DNA, while its complementary chain is called the nontemplate strand.
- Genetic instructions are coded in an 'alphabet' of four letters only: (the nucleotides) A, T, C and G.
- Identification of the order of nucleotides along a length of DNA is called DNA sequencing.
- Different genes vary in the nucleotide sequences along their DNA.

6 A piece of DNA comprises 20 000 bp. Is this more likely to be a whole chromosome or a whole gene? Explain.
7 If genes were isolated from a cat, a cyanobacterium and a cauliflower, what similarity would be seen?
8 If the two human genes for making blood-clotting factor and salivary amylase were compared:
 a in what way would they be similar
 b in what way would they be different?

Nature of the genetic code

Codes and more codes

The genetic instructions for all organisms are written in a code that uses an 'alphabet' of four letters only, namely A, T, C and G. What is a code?

Figure 2.12 shows some codes. Codes contain or encode information. To translate or decode a coded message, the information held in the code elements must be known. When **decoded**, this information may be verbal (words) or numerical (numbers) or auditory (musical sounds) or may specify an object.

If the coded information exists in a permanent form, the information can be decoded at any time by returning to read the code. For example, you can play a CD many times; each time the CD player decodes (translates) the information that is permanently **encoded** as microscopic pits on the disk surface.
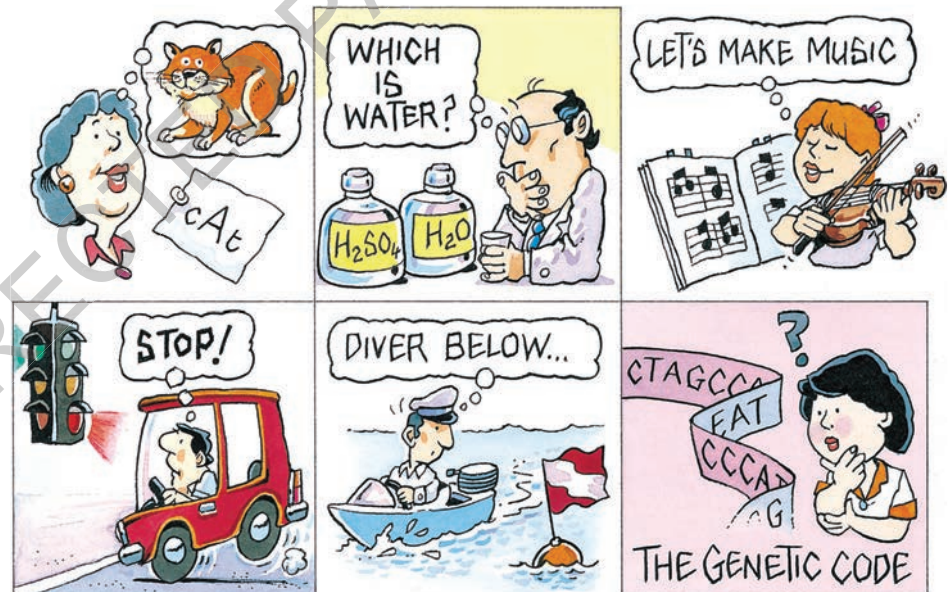
FIGURE 2.12 Various ways of representing information codes. Can you think of others?

The genetic code

The DNA of genes is an information-carrying molecule. Genetic information is encoded using just four elements: A, T, C and G. Before DNA was identified as the genetic material, many biologists thought that DNA was too simple a molecule to contain complex genetic instructions. How can a large amount of information be encoded by a code that has a small number of elements?

A code consisting of a few elements can encode a large amount of information. The morse code has two elements only, a dot (.) and a dash (–). By using various combinations of these elements, morse code can convey very complex information, such as all the words in this chapter.

Information encoded in the genetic code

The **genetic code** in the DNA of protein-encoding genes typically contains information for joining amino acids to form polypeptides. We can show this as:

Coded information nucleotide sequences in DNA template strand	Decoded information order of amino acids in polypeptides

A polypeptide on its own may be a functional protein or, in other cases, a functional protein may be built of several different polypeptide chains.

Proteins have many functions, and the various types include:

- structural proteins, which occur in connective tissues and in cell membranes
- contractile protein of muscle and myofilaments
- enzymes that regulate chemical processes
- proteins of the immune system, such as the antibodies
- oxygen-carrying proteins, such as haemoglobin
- hormonal proteins, such as insulin and growth hormone.

By encoding the sets of instructions on how to make the various types of proteins, genes control the structure and the biochemical and physiological functioning of an organism. The estimated 21 000 genes of a human organism contain all the instructions on 'how to make a human organism' that, if printed as the base sequences, would fill 1000 volumes of an encyclopedia.

Organisation of the genetic code

Consider two observations:

1. Genes typically contain coded information for assembling amino acids to form polypeptides.
2. Polypeptides are made of combinations of 20 different amino acid sub-units.

From these observations, it may be inferred that the genetic code must have at least 20 different instructions or pieces of information.

Examine table 2.2.

TABLE 2.2

Number of nucleotides in one instruction	Total number of different instructions possible
1 (e.g. T)	4
2 (e.g. AA, AT, GA)	16
3 (e.g. TTA, GCC, AAA)	64
4 (e.g. GGGA, TGCA, AATG)	256

In fact, one genetic instruction consists of a group of three bases, such as AAT, GCT and so on. Because of this, **the genetic code is referred to as a triplet code**. This form of code is sufficient to account for the pieces of information that must be encoded.

Consider a piece of template-strand DNA with the base sequence T A C A A A C A A G C T C C T A C T . . . This DNA has six coded instructions (shown underlined) that are decoded or translated as follows:

1. TAC = start building a protein, commencing with the amino acid *met*
2. AAA = now add the amino acid *phe*
3. CAA = now add the amino acid *val*
4. GCT = now add the amino acid *arg*
5. CCT = now add the amino acid *gly*
6. ACT = now stop.

Cracking the genetic code

The genetic code was originally unknown and had to be broken. In 1961, the first piece of the genetic code was broken when the three-base triplet 'AAA' in DNA was decoded as 'Add the amino acid *phe* into a protein being constructed' (see figure 2.13). We will see later in this chapter that the translation of each DNA triplet occurs through an intermediate molecule, messenger RNA (mRNA). The AAA triplet in DNA is transcribed into a complementary UUU codon in mRNA that is then translated into the amino acid phenylalanine (*phe*). By 1966, all 64 pieces of the genetic code had been deciphered.

Refer to table 2.5 on page xxx to see the full translation of the genetic code in terms of the 64 mRNA codons. Note that some codons contain the instructions 'START adding amino acids' and 'STOP adding amino acids'. Refer to the Appendix (page xxx) to see the structures of the amino acids and their three-letter and single-letter abbreviations.

The main features of the genetic code are:
- Pieces of information in the genetic code consist of triplets or three-base sequences (e.g. TCA in DNA, or UGT in mRNA).
- The code is non-overlapping. So, a fragment of DNA consisting of 12 bases contains four pieces of information or instructions.
- The code is essentially the same in bacteria, in plants and in animals — it is said to be universal (but see the Odd Fact at left).
- The code is said to be redundant because, in many cases, more than one triplet of bases codes for one particular amino acid.
- The information encoded in DNA is the set of instructions to assemble amino acid sub-units into polypeptides.
- The information in the DNA template strand also includes a START instruction (TAC) and three STOP instructions (ATT, ATC or ACT).

FIGURE 2.13 In 1961, Nirenberg and Matthaei 'cracked' the first element of the genetic code.

KEY IDEAS

- DNA contains information encoded in the base sequence of its template strand.
- Genes contain coded instructions for joining specific amino acids into polypeptides.
- The genetic code in DNA is a non-overlapping triplet code consisting of groups of three bases.
- One piece of genetic code typically contains the information to add one amino acid to a protein.

QUICK CHECK

9 Give an example of a code.
10 In what form is information held in a DNA molecule?
11 Which of the following statements is most accurate? Explain your choice.
   a DNA is converted to the amino acid sub-units of proteins.
   b DNA contains the coded information for joining amino acids to form polypeptides.
   c DNA turns into polypeptides.
12 How many instructions (for adding amino acids) are present in the base sequence TTAGGG?
13 Which code means 'START joining amino acids to form a polypeptide'?
14 What are the meanings of the following codes in DNA: CAA and ACT?

Proteins

Although water is the main compound in living cells, **more than half of the remainder, about 18 per cent is protein**. There are thousands of different proteins in each cell and many of these control all metabolic processes within cells. **All proteins contain nitrogen (N), hydrogen (H), carbon (C) and oxygen (O). Some also contain sulfur (S)**. They are large complex compounds as indicated by the following examples:

- zian from Indian corn $C_{736}H_{1161}N_{184}O_{208}S_3$
- gliadin from wheat $C_{685}H_{1068}N_{196}O_{211}S_5$
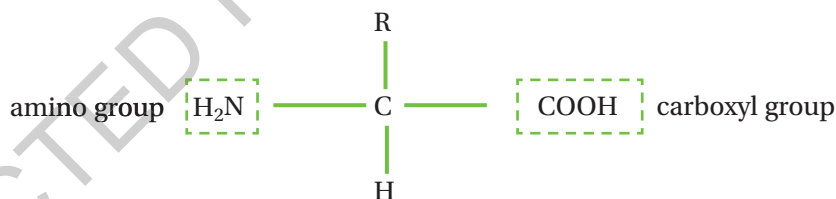- casein from milk $C_{708}H_{1130}N_{180}O_{224}S_4P_4$.

Some plants contain granules of solid protein that they store as reserve food. Animals, as we will see in the next section, have most of their energy stores in fat.

Amino acids: the building blocks of proteins

Proteins are large molecules built of sub-units called amino acids.

There are 20 naturally occurring amino acids (refer to the Appendix). Humans are unable to make all 20 amino acids and must rely on their food for the nine they are unable to make (see the note at left). Not all plants can make all 20 amino acids so a vegetarian diet should be well planned to ensure a balanced intake of appropriate amino acids. Generally, animal proteins are a better source of amino acids for humans because animal protein is more like that of humans. Different proteins contain different numbers and proportions of each of the amino acids.

The general formula of an amino acid is:

amino group $\quad$ H$_2$N $\quad$ —— $\quad$ C $\quad$ —— $\quad$ COOH $\quad$ carboxyl group

with R above C and H below C.

Each amino acid has one part of its molecule different from other amino acids. **The R group in the general formula is the part that varies.** Refer to the Appendix and note that the R group of the amino acid cys contains sulfur (S). When several cys amino acids are present, disulfide bonds (-S-S-) can form (see figure 2.19).

Two amino acids join together as a dipeptide when a peptide bond forms between the amino group of one amino acid and the carboxyl group of another amino acid and a water molecule is released (see below and figure 2.14). **When a number of amino acids join in this way, a polypeptide is formed.** Each type of protein has its own particular sequence of amino acids. Polypeptide chains become folded in different ways depending on their amino acid sequences.

These nine amino acids cannot be made by human cells but they are essential and must be obtained in food eaten:

threonine	isoleucine
methionine	histidine
lysine	phenylalanine
valine	tryptophan.
leucine	

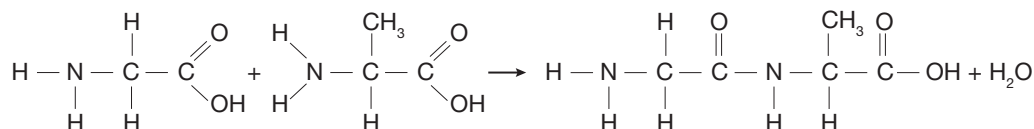glycine + alanine $\longrightarrow$ dipeptide + water

FIGURE 2.14 Proteins are assembled from amino acids that are joined by peptide bonds. Each peptide bond forms by the linkage of an amino group from one amino acid and a carboxyl group of another amino acid. A number of amino acids joined by peptide bonds form a polypeptide chain. Note that the different R groups are represented by different colours. The remainder of each amino acid molecule is identical.
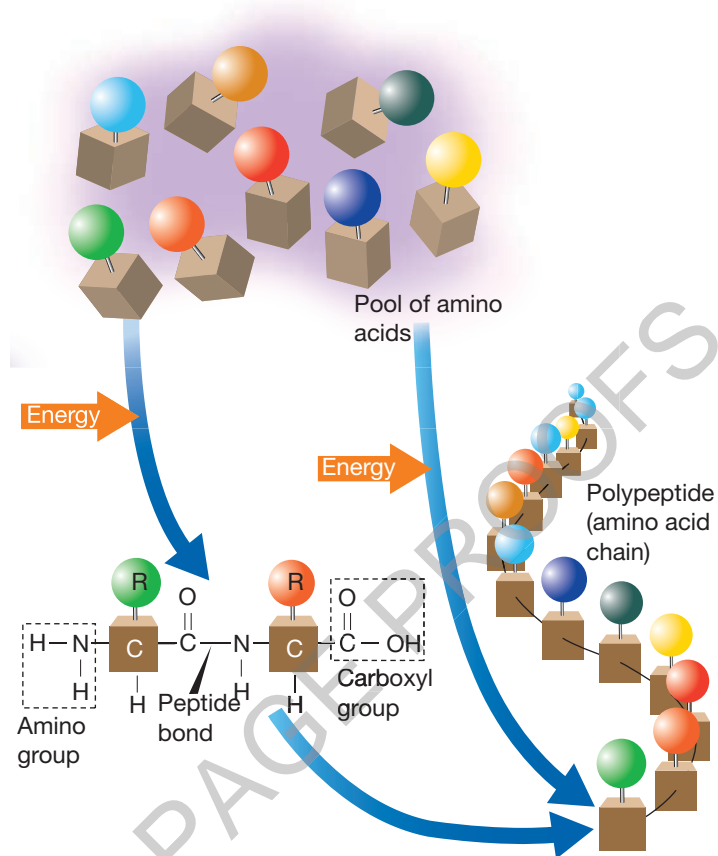
The structure and shape of proteins

Proteins are very large molecules; some contain thousands of amino acids. These large molecules fold and organise into different shapes. Protein structure is described at four different levels of organisation (see figure 2.15).

Primary structure

The primary structure of a protein is the specific linear sequence of amino acids in the protein. Different proteins have different primary structures and hence have different functions. The sequence of amino acids in a protein is determined by the genetic material in the nucleus (this is discussed further in the next section).

Secondary structure

Amino acid chains fold in three different ways (figure 2.15b). Hydrogen bonds form between segments of the folded chain that have come close together and help stabilise the three-dimensional shape. The following are some examples of secondary structure:

1. The major proteins of wool are keratins that have a spiral secondary structure, known as an **alpha helix**. If the fibre is stretched and the hydrogen bonds are broken, the fibre becomes extended. If the fibre is then 'let go', the hydrogen bonds reform and the fibre returns to its original length.
2. The major protein of silk is fibroin, which is fully extended and lacks the coiling found in the structure of wool. The silk molecules form a **beta-pleated sheet** (see figure 2.15b). The polypeptide chains of silk are already extended and cannot be extended further.
3. The secondary structure of portions of a protein is called **random coiling** if the portions do not conform to the shape of an alpha helix or a beta-pleated sheet.
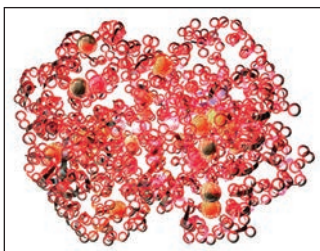
study on

Unit 3
AOS 1
Topic 3
Concept 3

Do more
Protein structure

(a) Primary structure   Amino acid sequence
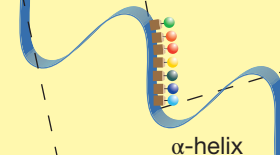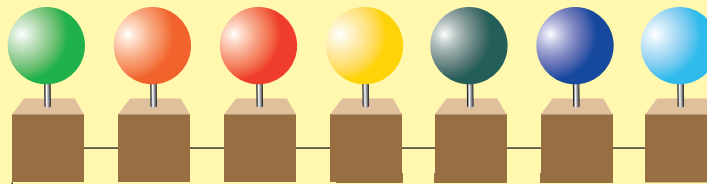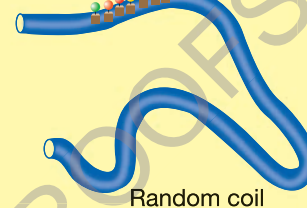
(b) Secondary structure (3 configurations)

α-helix

Random coil

β-pleated sheet

ODD FACT

The secondary structure of myoglobin, the oxygen-binding protein of muscle, consists mainly (75%) of a coiled alpha helix structure.
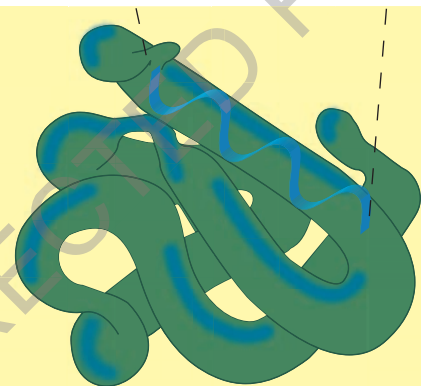
Myoglobin

(c) Tertiary structure
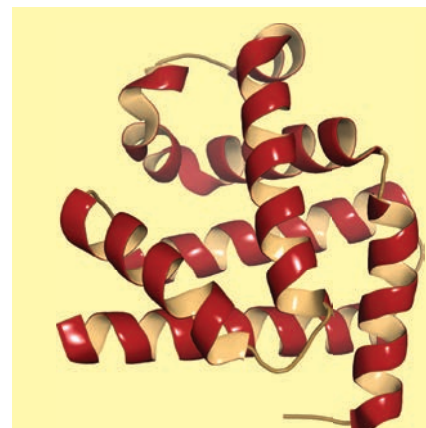
(d) Ribbon model of tertiary structure
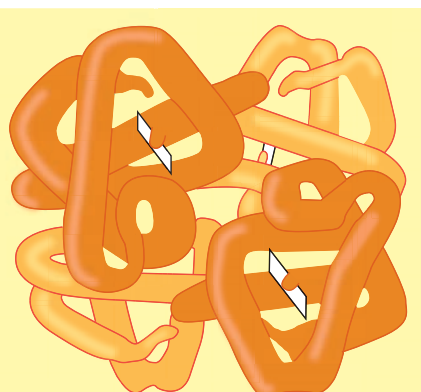
Haemoglobin

(e) Quaternary structure

FIGURE 2.15 The four levels of protein structure: **(a)** primary — order of amino acids in the molecule; **(b)** secondary — folding of some portion of the amino acid chain (note the three different modes); **(c)** tertiary — describes the shape of the entire polypeptide chain; **(d)** tertiary ribbon model — note that some of the secondary structure of this protein can also be seen, namely the regular coils, each showing an alpha coil or an alpha helix; **(e)** quaternary — some proteins comprise a number of polypeptide chains.

The forces that maintain the tertiary structure of proteins are:
1. hydrogen bonds
2. ionic attractions between charged R groups
3. interactions between hydrophobic R groups in the protein interior
4. covalent disulfide cross links.

Tertiary structure

The tertiary structure refers to the total irregular folding held together by ionic or hydrogen bonds forming a complex shape such as that of myoglobin. The bonds form between side chains of amino acids to form a complex internal structure.

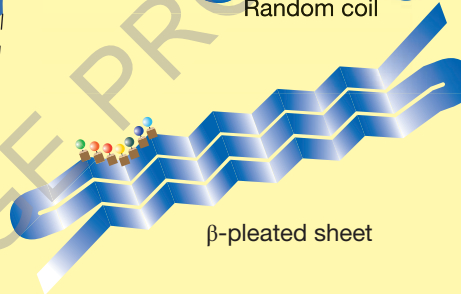The 3D shape that constitutes the tertiary structure of a protein is critical for its function. For example, if the shape of an enzyme is changed, particularly at its active site, the protein can no longer function as an enzyme.

The regular folding of alpha coils or beta-pleated sheets that is present in the secondary structure of some proteins may be seen when you examine the overall tertiary structure of the protein. Figure 2.15d shows a ribbon model of the tertiary structure of the muscle protein myoglobin. Note the regular coils that form part of this protein — each of these is an alpha helix or alpha coil that is part of its secondary structure. Other regions of the myoglobin molecule do not have a regular secondary structure — these regions are shown as thin lines with random shapes.

Quaternary structure

Quaternary structure describes a structure in which two or more polypeptide chains interact to form a protein. The resulting structure can be, for example, **globular** as in haemoglobin (figure 2.15e and figure 2.16) or **fibrous** as in collagen, the most common protein in skin, bone and cartilage.
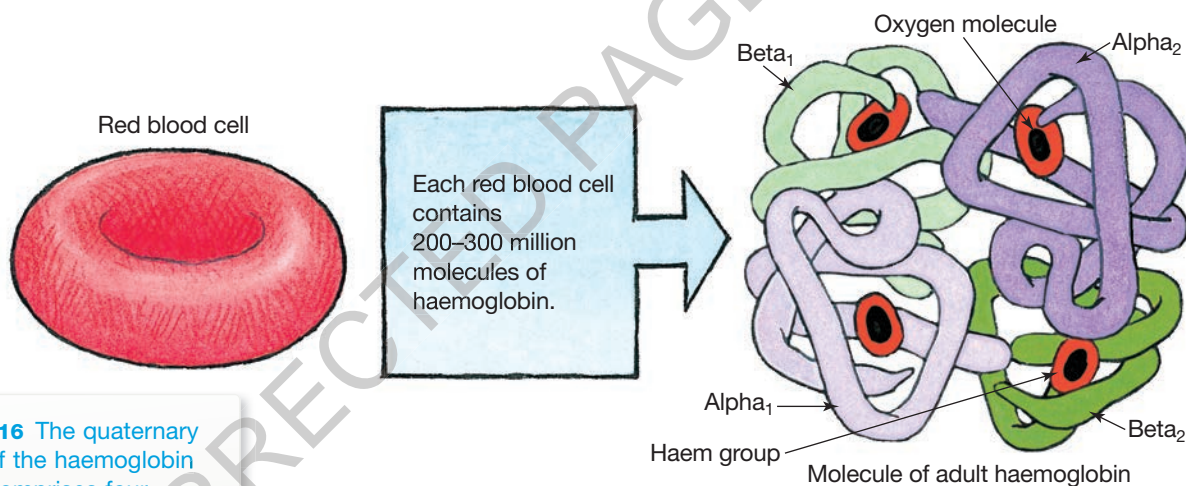
FIGURE 2.16 The quaternary structure of the haemoglobin molecule comprises four chains: two alpha chains and two beta chains. Note how many of these molecules are present in each red blood cell.

Table 2.3 lists a number of different types of protein and their functions. Some examples of proteins are also shown in figure 2.17.

TABLE 2.3 Examples of proteins and their functions

Type of protein	Function	Example
structural	fibrous support tissue in skin, bone, tendons, cartilage, blood vessels, heart valves, and cornea of the eye	collagen, keratin
enzyme	catalyse reactions	ATP synthase
contractile	muscle movement	myosin, actin
immunoglobulin	defence against disease	antibodies
hormone	regulate body activity	insulin
receptor	respond to stimuli	insulin receptors
transport	carry other molecules	haemoglobin
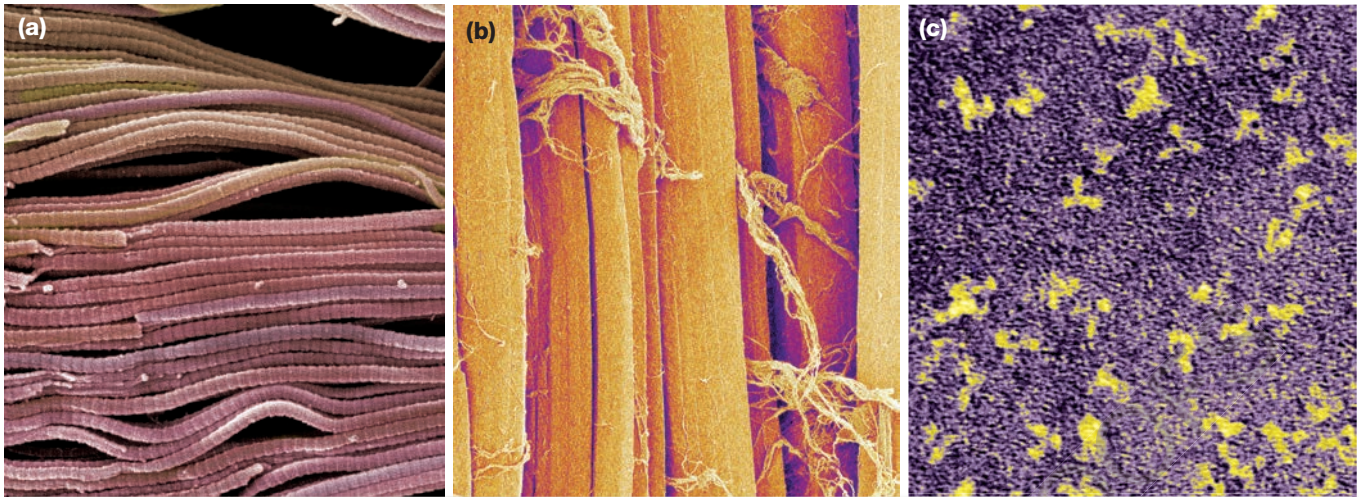
FIGURE 2.17 **(a)** A scanning electron micrograph (SEM) of collagen bundles from connective tissue that wraps around and supports nerve fibres (notice the characteristic banding of collagen fibres). **(b)** A scanning electron micrograph (SEM) of skeletal muscle fibre showing the thick filaments that are made up of myosin. **(c)** A transmission electron micrograph (TEM) showing Y-shaped structures (yellow), which are molecules of the immunoglobulin G antibody, an important part of the immune system.

FORENSICS — DETECTING HAEMOGLOBIN

The popularity of television shows about forensics has brought some knowledge about the detection of blood into many homes. Turn off the light, spray a chemical at the crime scene and examine the scene for signs of luminescence. The luminescence produced is because the chemical used reacts with any haemoglobin present (figure 2.18).

Although the chemical Luminol has been the star in many television shows, a new star has been developed. BLUESTAR® Forensic is an improved product that gives a brighter luminescent result visible in normal light. It is also nontoxic and causes no change in the DNA so that minute quantities that are present can also be used for DNA-typing.

Although the presence of blood is indicated, further testing is required to confirm that the substance luminescing is in fact blood, because other compounds also react with the chemical. The important factor is that the position of minute blood remains are indicated even though no blood was visible to the naked eye. Without this, testing for blood would not be possible.

FIGURE 2.18 Images from a crime scene investigation. Although no blood was visible to the naked eye, use of BLUESTAR® Forensic detected the presence of minute levels of haemoglobin. The luminescence indicates possible blood. Additional tests can prove it to be so.

Conjugated proteins

Many of the proteins we have described above are *simple* proteins — the final molecule contains amino acids only. **With some proteins, the chains of amino acids conjugate with other groups.** This is particularly the case for those proteins in the nucleus. They are mostly **nucleoproteins** — they comprise a molecule containing both protein and nucleic acid, as, for example, nucleosomes that are complexes of DNA coiled around a group of eight histone proteins (refer back to figure 2.5b).
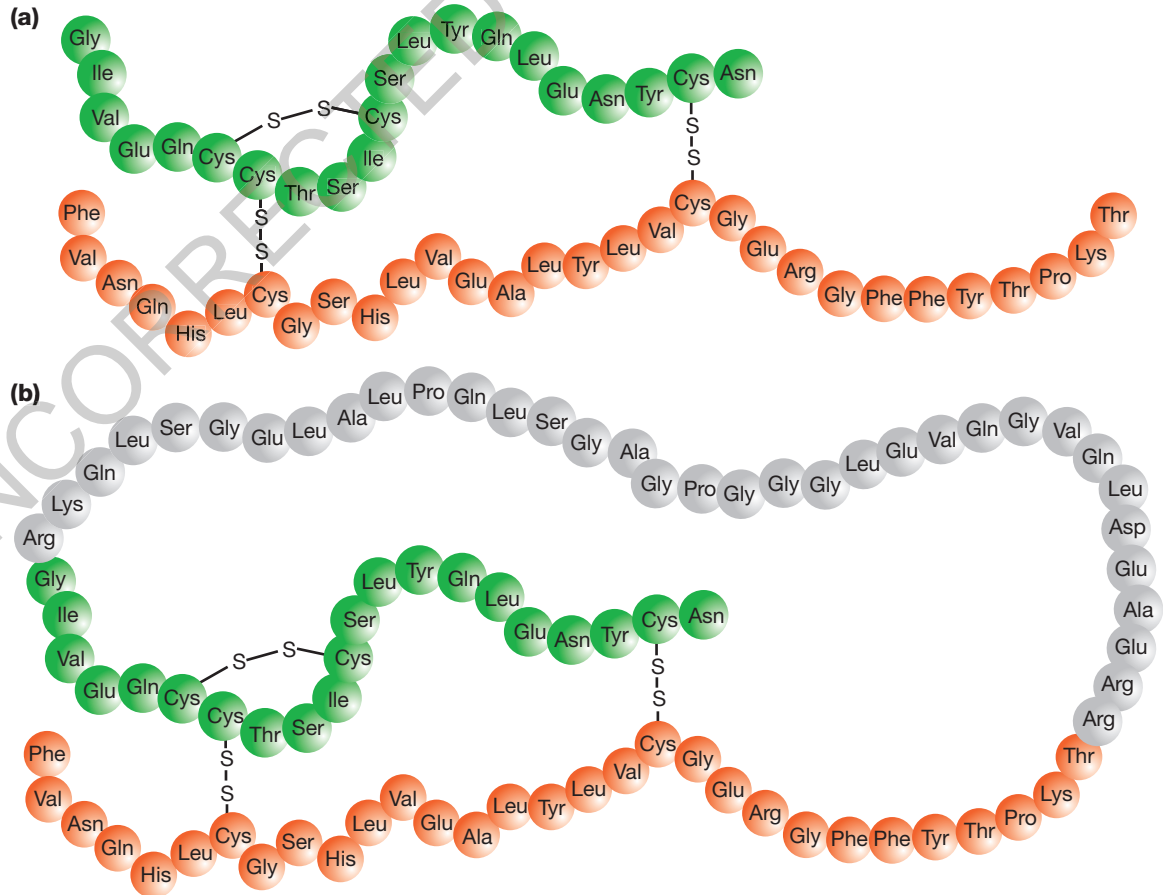
Another conjugated protein is haemoglobin. Each tertiary structure component produced associates with a haem group. The quaternary molecular structure comprises four chains: two alpha chains and two beta chains. The amino acid sequence in a protein is important. If the order of amino acids in either chain is altered, a defective chain results. An individual inherits, from each parent, the DNA that encodes the beta chain. If a defect in this DNA is inherited from both parents, an individual is unable to produce any normal haemoglobin and has the genetic disorder beta thalassaemia.

Inactive to active molecule

Insulin, produced by beta cells in the pancreas, is a protein hormone. It controls the level of glucose in the blood by facilitating the uptake of glucose from the blood into tissue cells. When this occurs the level of blood glucose declines. An insulin molecule comprises two chains of amino acids held together by disulfide bonds (figure 2.19a). This is the active state of the hormone.

When initially produced, insulin is inactive as a hormone. It is produced as a single chain of amino acids with folds that are held together by three disulfide bonds (figure 2.19b). A section of this molecule is removed by an activating enzyme leaving the active hormone as two chains of amino acids held together by the three disulfide bonds.

FIGURE 2.19 (a) A molecule of insulin comprises two chains of amino acids held together by three disulfide bonds. This is the active state of the molecule. **(b)** Inactive insulin is produced as a single chain of amino acids folded on itself and held together by the three disulfide bonds.

Hence, although a molecule may be made from a number of molecules linked together by sulfide or other bonds, they may be derived from the same initial inactive protein. Many enzyme proteins are produced in an inactive form and become active only if the appropriate enzyme is present to convert them for active service. Can you think of a situation that you considered in your previous studies that involved an enzyme being converted from the inactive to the active state?

What is the proteome?

In living organisms, proteins are involved in one way or another in virtually every chemical reaction. They may be the enzymes involved, they may be the reactants or the products, or they may be all three. **The complete array of proteins produced by a single cell or organism in a particular environment is called the proteome** of the cell or organism. The study of the proteome is called **proteomics**.

Scientists are moving away from investigating single proteins because no protein acts in isolation from other proteins. They are now exploring the total pattern of proteins produced by a cell and analysing these patterns to compare them with patterns from different kinds of cells. What are the differences? What are the similarities? What is the proteome profile of diseased tissue or even the fluids surrounding the tissue? In what ways do they differ from the healthy state?

There is also an emphasis on structural proteomics (see figure 2.20). **Knowing that a protein exists is different from knowing how it operates.** Knowing how some operate and knowing about their structures may enable testable predictions about the roles of other proteins on the basis of their structures.
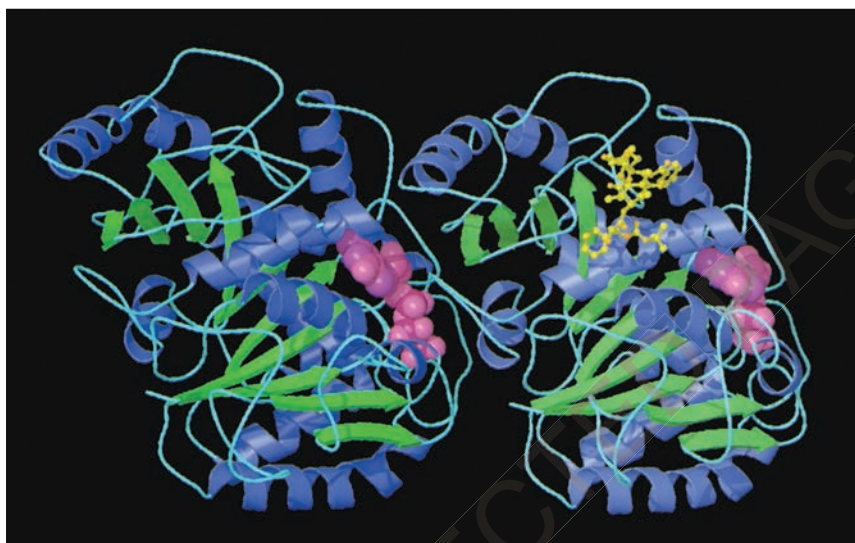
FIGURE 2.20 Knowledge of three-dimensional structures of well-known proteins may give insights into their functions. This three-dimensional model shows the overall tertiary structure of two tubulin molecules. Tubulin, a globular protein, is a component of microtubules that are found, for example, in the cilia of the human lung. The secondary structure of tubulin is denoted by the blue regions showing the alpha helices, the green regions showing the beta-pleated sheets, and the thin lines denoting regions of random coiling. (The pink denotes the nucleotide GTP, which binds to the protein.)

KEY IDEAS

- All proteins contain N, C, H and O, and some contain S.
- The structure of all proteins can be identified at the primary, secondary and tertiary levels, while proteins that consist of more than one polypeptide chain have an additional quaternary level of structure.
- Proteins can also be classified on the basis of their different functions.
- The proteome of an organism is the complete array of proteins produced by that organism.

QUICK CHECK

15 What is the basic formula of an amino acid molecule?
16 How is a peptide bond formed?
17 List the four basic structures of protein molecules and draw an example of each.
18 Give an example of (a) a structural protein, (b) a contractile protein and (c) a conjugated protein.
19 Why is proteomics considered important?

A closer look at a gene

The part of a gene that contains the coded information for making a protein is called the coding region of a gene. The regions on either side of the coding region of a gene are called **flanking regions**. The flanking region before the start of the coding region is called the upstream region. The flanking region after the end of the coding region is called the downstream region.

The box on page XXX provides further information about these regions.

Introns: just an interruption or two

An unexpected discovery about the genes in eukaryotes was made in 1977. Until then, the coding region of a gene was thought to be continuous (see figure 2.21a).

Instead, the coding region is 'interrupted' by other segments of DNA. **Each segment of the coding region of a gene is called an exon**. The exons are separated by lengths of DNA that do not contain instructions relating to the protein chain. These **noncoding segments are called introns** (see figure 2.21b).
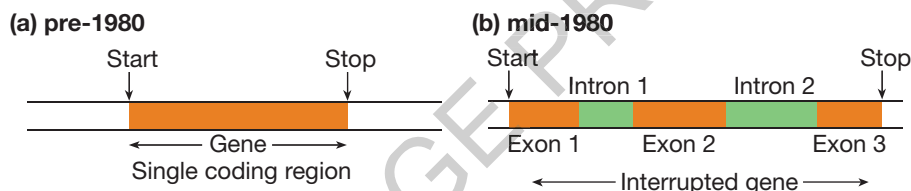
To avoid confusing the two terms *exon* and *intron*, think about INterruption and INtron and that will help you remember the functional difference. Introns are transcribed but not translated.

(a) pre-1980

Start — Stop

Gene
Single coding region

(b) mid-1980

Start — Intron 1 — Intron 2 — Stop

Exon 1 — Exon 2 — Exon 3
Interrupted gene

FIGURE 2.21 (a) Pre-1980 view of a gene, with a continuous coding region. **(b)** Actual situation with coding region (orange) broken into segments (exons) by noncoding segments (green).

CODING AND FLANKING REGIONS

If it were possible to travel in a miniaturised vehicle along a gene, what would we see?

The coding region

The coding region of a gene is the segment of DNA double helix that includes the DNA template strand, which encodes the information that will later be translated into the amino acid sequence of a polypeptide. This region of a DNA template strand begins with a start signal (TAC) and, some distance away, there is a stop signal (ATT or ATC or ACT).

Upstream from the coding region

The region of DNA on the template strand upstream from the coding region contains some particular base sequences. One part of the upstream region is rich in As and Ts and is often called the **TATA box**, because the sequence TATA AA (or similar) occurs there. Another sequence commonly found further upstream is called the CAT or CAAT box.

Role of upstream sequences

Consider the following observations:

- Upstream sequences are invariably found in all organisms. It is reasonable to suggest that these upstream sequences serve an important function since they have been maintained during evolution.
- If upstream sequences are altered by mutation, the activity of the coding region of the gene may be reduced or even become inactive. The absence of the correct upstream signal is a cause of some inherited human diseases. One form of thalassaemia is due to a missing TATA group in the upstream region of the DNA of both copies of the specific gene in the people concerned.
- The upstream region includes segments of DNA to which hormones can attach. The fact that some hormones can bind to DNA provides one clue as to how hormones can influence the action of genes.

These observations support the conclusion that specific DNA sequences upstream of the coding region of a gene initiate transcription, the process by which the encoded information in the DNA coding region is transcribed into mRNA. **Promoters** also act as sites where proteins called transcription factors can bind and regulate the expression of genes.

(continued)

Downstream from the coding region

The DNA following the end of the coding region is referred to as the 'downstream' region (see figure 2.22). About 20 bases downstream, the sequence AATAAA is usually found. If this sequence is altered, the gene action is altered. The downstream region includes an 'end transcription signal', which terminates the process of transcription of mRNA from the DNA template.

Critical sites where mutations can affect normal functioning of a gene include:
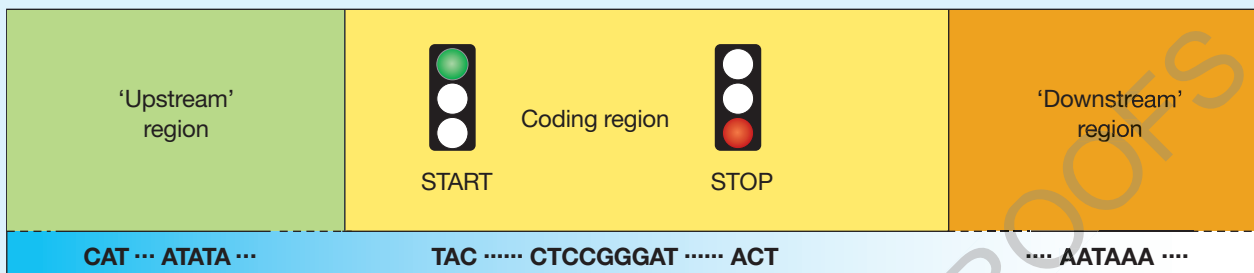- the upstream promoter region
- the START and STOP signals.

FIGURE 2.22 Regions of the template strand of a typical gene.

So, eukaryote genes are not like nursery rhymes in a book, where the reader starts at the beginning and reads through to the end. The information in genes is broken up into segments and the sections in between are filled with other printed material that is unrelated to the rhyme.

Here is an interrupted rhyme:

'Hey, diddle diddle the cat and the fid HERE IS AN INTERRUPTION dle, the cow jumped over the AND HERE IS ANOTHER INTERRUPTION moon. The little dog laughed to see such fun and the dish HERE'S ANOTHER ran away with the spoon.'

If this interrupted rhyme were thought of as a gene, how many exons and how many introns would it contain? The underlined portions are like exons, and there are four of them. The interruptions are like introns, and there are three of them; they are removed from the mRNA before translation.

The number of exons and introns in genes varies. The DNA making up the **HBB** gene, which controls the production of one chain of the haemoglobin molecules, consists of three exons and two introns. The **F8C** gene, which controls the production of factor VIII, which assists in blood clotting, consists of 26 exons and 25 introns.

KEY IDEAS

- Each gene in eukaryote organisms contains a coding region, and also includes flanking regions upstream and downstream of the coding region.
- The coding region of a gene typically consists of several exons separated or interrupted by introns.

QUICK CHECK

20 Using words or diagrams, distinguish between the members of each of the following pairs:
   a intron and exon
   b coding region and flanking region.
21 True or false? All genes contain the same number of exons.
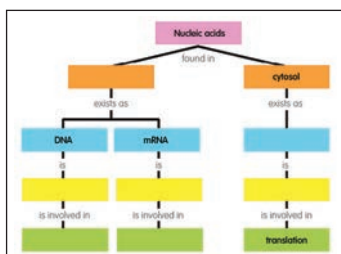
FIGURE 2.23 Because of pairing of complementary bases in DNA and RNA, one DNA chain can act as a template to build an RNA chain with a predictable nucleotide sequence. What does an A in DNA put in place in RNA?

Protein synthesis

The genetic instructions for producing proteins are found within the DNA of the chromosomes. A gene is a segment of DNA that codes for a protein. In eukaryote organisms how do genetic instructions get from the nucleus to the ribosomes?

When a gene becomes active, it first makes a mobile copy of the coded instruction that it contains. This occurs by a process known as **transcription**. This mobile copy of a genetic instruction can leave the nucleus and move to the cytoplasm where the instruction is decoded. This occurs by a process known as **translation**. So, in the case of protein-encoding genes, **gene action** involves two processes: transcription, which occurs in the nucleus, and translation, which occurs on ribosomes in the cytoplasm.

Transcription: copying the original

The nucleus of a eukaryote cell is like a safe that contains the genetic masterplan in the form of DNA. The genetic masterplan containing the entire set of instructions for an organism is like the complete plan for the construction of a complex structure, such as a jumbo jet. One gene or instruction for a particular protein is like the plan for making one component of the jet, such as a wing flap.

The workers at the site where the wing flaps are made do not work directly from the complete masterplan; instead, they have copies of the relevant section of the plan. Likewise, **before a genetic instruction in DNA is decoded, that instruction is copied (transcribed) from the genetic masterplan, which remains in the nucleus.** This copy is encoded in a different nucleic acid called ribonucleic acid (RNA). Because the role of this particular RNA is to carry a copy of a genetic instruction from the nucleus to the cytoplasm, it is known as **messenger RNA (mRNA)**.

Pairing or hybridisation can occur between the bases in one DNA strand and complementary bases in an RNA strand as follows:

A in DNA	pairs with	U in RNA
T	pairs with	A
C	pairs with	G
G	pairs with	C.

This pairing means that a DNA chain can act as a template to guide the construction of RNA with a complementary base sequence (see figure 2.23). This means that the genetic information in DNA can be accurately copied into RNA during the process of transcription.

Consider a DNA template with the base sequence:

DNA template . . . **ATGCCTGAAT** . . .

This DNA can act as a template to guide the formation of an RNA molecule with the complementary base sequence as follows:

mRNA transcript (copy) . . . **UACGGACUUA** . . .

The base sequence of the mRNA primary transcript is not identical to that of the template DNA strand; instead, the mRNA has a complementary sequence. However, the mRNA base sequence matches that of the complementary nontemplate DNA strand, except that U replaces A. For this reason, the nontemplate DNA strand is also called the coding DNA strand.

| Nontemplate DNA:	A	T	C	C	C	G	G	T	A	A	C	A
Template DNA:	T	A	G	G	G	C	C	A	T	T	G	T
mRNA transcript:	A	U	C	C	C	G	G	U	A	A	C	A

From DNA to mRNA: step by step

A particular gene in the nucleus is switched on at a specific stage of development. In the case of the **HBB** gene, this gene becomes active during late fetal development in certain bone marrow cells. **Transcription takes place in a series of steps** (see figure 2.24).

1. An enzyme, known as **RNA polymerase**, attaches to a specific promotor sequence of DNA in the upstream region of the template strand. The double-stranded DNA of the gene unwinds and exposes the bases of the template strand.
2. The base sequence of the DNA template guides the building of a complementary copy of the mRNA sequence. The RNA polymerase enzyme moves along the DNA template in a 3′ to 5′ direction and, as it moves, complementary nucleotides are brought into place and, one by one, are joined to form an RNA chain.
3. After the RNA polymerase moves past the coding region and into the downstream region of the gene, transcription stops and the mRNA molecule is released from the template.

   The result of this process is a single-stranded molecule, called pre-messenger RNA. The base sequence in the pre-mRNA molecule is complementary to the base sequence of the DNA of the template strand.

FIGURE 2.24 Transcription occurs in the cell nucleus. The enzyme RNA polymerase moves along the DNA template building an mRNA molecule at the rate of about 30 bases per second.
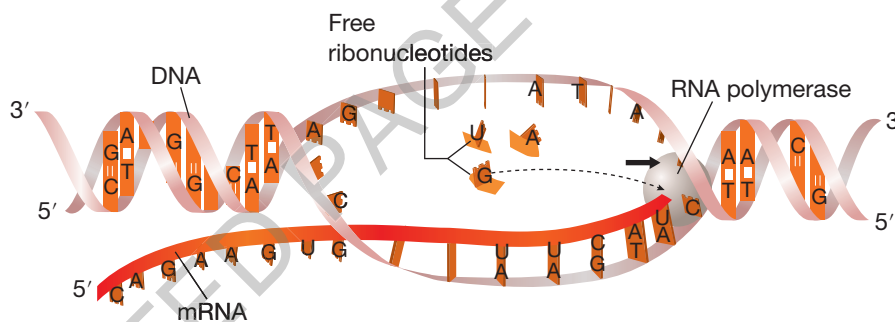
ODD FACT

Splicing of pre-mRNA is carried out by a complex known as the spliceosome. This complex consists of protein and snRNA.

Pre-mRNA is modified after transcription

The primary product of translation is pre-mRNA, also known as the primary transcript. The sequence of bases in the pre-mRNA is complementary to all the DNA bases of a gene, both introns and exons (refer back to chapter xx, pages 000–0). The primary mRNA transcript then undergoes a process termed **post-transcription modification** (see figure 2.25) before it leaves the nucleus.

FIGURE 2.25 Various processes in the post-transcription modification of pre-mRNA (primary transcript) produce mRNA. Post-transcription modification occurs in the nucleus. The **HBB** gene comprises three exons and two introns and, in the post-transcription modification of the primary transcript, the two introns are excised and the three exons are spliced to form the mRNA.
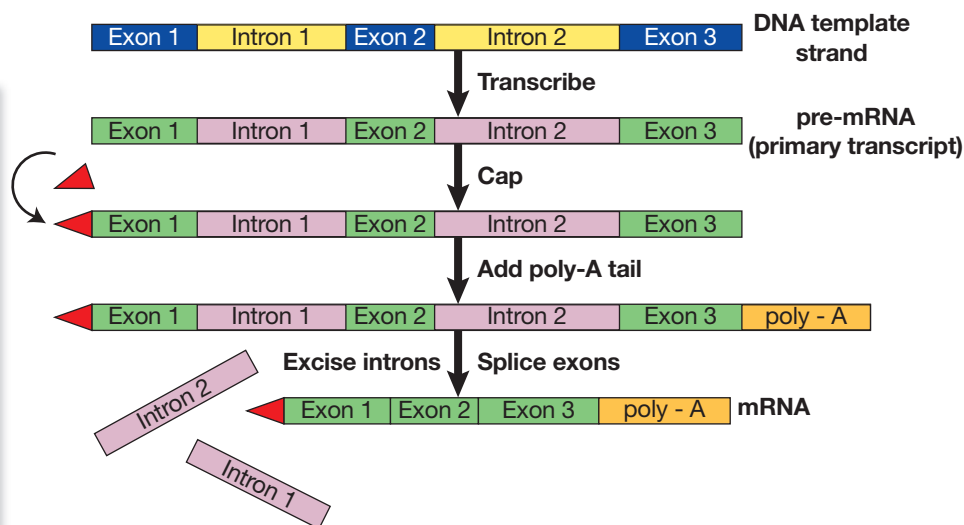
studyon

Post-transcription modification includes the following processes:

- *Capping:* The 5′ end of the pre-mRNA is capped with an altered guanine (G) base (methyl guanosine). The cap protects the pre-mRNA from enzyme attack and contributes to its stability.
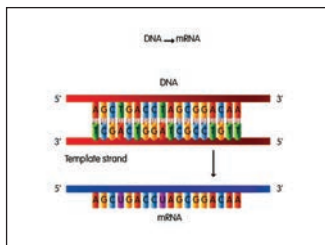- *Adding a tail:* The primary transcript is clipped at a specific point downstream of the coding region and a poly-adenine (A) tail, with up to 250 As, is then added at the 3′ end. The tail contributes to the stability of the mRNA.
- *Splicing:* The regions in the pre-mRNA that correspond to the introns are cut out and the remaining exons are spliced together. This cutting and splicing is done by **spliceosomes**, which recognise specific base sequences at the ends of the introns: GU at the 5′ end, and AG at the 3′ end.

Post-transcription occurs within the nucleus. The final mRNA product now moves across the nuclear membrane into the cytosol, carrying with it a copy of the information originally encoded in the DNA of the gene.

In the next section, we will examine how the genetic information that is copied into mRNA is decoded (translated) into a particular protein chain.

KEY IDEAS

- Coded genetic instructions are located in the DNA of the nucleus of eukaryote organisms.
- DNA and RNA are both nucleic acids, but differ in several ways.
- During transcription, the information in the template strand of the DNA of a gene is copied into a pre-RNA molecule.
- The final mRNA molecule results when post-transcription modification is complete, including intron excision and exon splicing.

QUICK CHECK

22 Where does transcription occur? What is the end product of this process?
23 A template strand of DNA includes the base sequence:
. . . T A T C G G C A T . . .
Write the base sequence of the complementary nontemplate strand.
24 A strand of mRNA includes the base sequence:
. . . A U G U A U C C G . . .
Write the base sequence of the DNA coding strand.
25 Where does transcription occur in a cell?
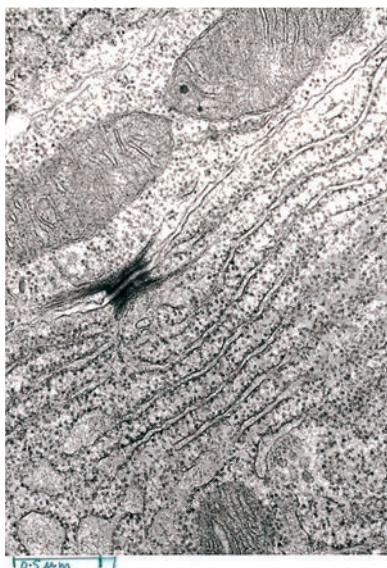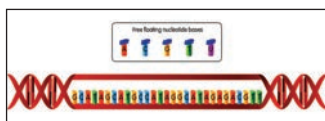26 List two differences between pre-mRNA and mRNA.

FIGURE 2.26 Ribosomes are composed mainly of a type of ribonucleic acid known as ribosomal RNA (rRNA). In this electron photomicrograph, ribosomes appear as dark dots located on the endoplasmic reticulum.

Translation: decoding genetic instructions

The decoding of the genetic instructions occurs through the process of translation, which takes place in the cytoplasm. By the end of this process, the genetic instructions carried in mRNA have been decoded and translated into a protein chain built of amino acids. The coded instruction in the mRNA is not changed in this process, just as the plan of a jumbo jet part is unaltered after the part is made.

Translation involves the combined action of several agents (see table 2.4).

The mRNA moves from the nucleus to the cytoplasm where it attaches to submicroscopic organelles known as **ribosomes** (see figure 2.26).

TABLE 2.4 Players and places in translation (Many enzymes are also involved.)

Agents	Analogy
DNA in the nucleus	masterplan with complete set of instructions
mRNA	working copy of one instruction
ribosomes	construction site
tRNA	carriers of raw material
amino acids	raw material
protein chain	end product

The construction of a protein according to the coded instructions in mRNA involves the assembly of amino acid sub-units. The various **amino acids** are present in solution in the cytosol. How are the correct amino acids selected from this solution?

Each amino acid is brought to the mRNA on the ribosomes by a carrier molecule called **transfer RNA (tRNA)**. Each tRNA molecule consists of a single strand of 76 nucleotides coiled and paired with themselves. At one end of each tRNA molecule are three bases that make up an **anti-codon**. At the other end of a tRNA molecule is a region that attaches to one specific amino acid (see figure 2.27). An enzyme, amino acyl tRNA synthetase, catalyses the linking of each amino acid to its specific tRNA carrier.

In this text, the term *protein chain* is used to refer to a single chain built of amino acid sub-units. Polypeptide is a term for a short chain consisting of amino acids.
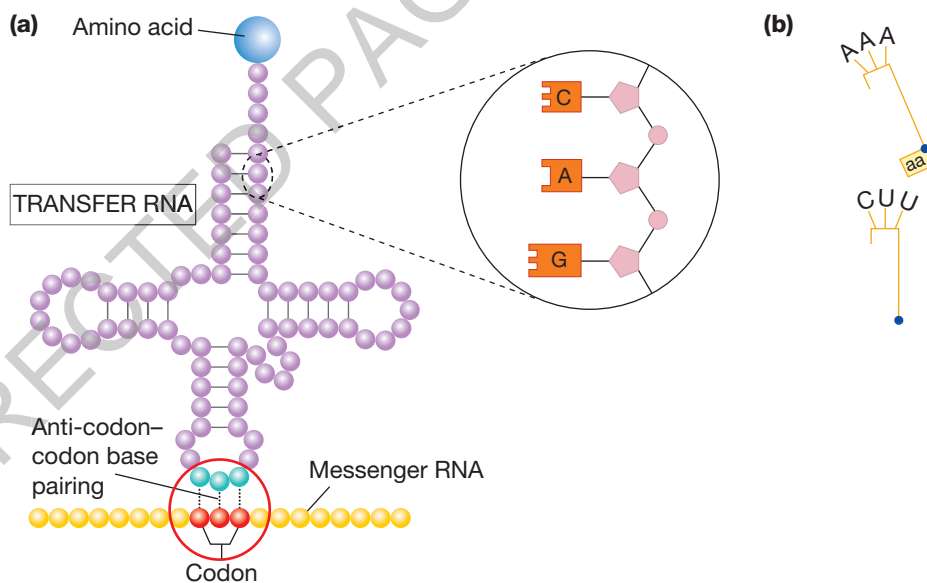

FIGURE 2.27 (a) Transfer RNA (tRNA) molecule. At one end is an amino acid attachment site and, at the opposite end, an anti-codon region. How many bases make up an anti-codon? **(b)** A simpler representation as used in this textbook.

From mRNA to protein: step by step

The information in mRNA is present in coded form as sets of three bases or triplets. These triplets, such as AGG and UCU, are called **codons**. **Most codons contain the information to add one specific amino acid to a protein chain.** In addition, one codon (AUG) is a START TRANSLATION instruction, and three different codons (UAA, UAG and UGA) are STOP TRANSLATION instructions (see table 2.5).

The instructions in an mRNA molecule are decoded three bases (or one codon) at a time. Translation begins at the 'start adding amino acids' signal (AUG codon) (see figure 2.28a). This codon both starts the process of building a protein chain and puts the amino acid *met* into place as the first amino acid in the chain. Then, the next three bases are translated by adding the next amino acid to the growing protein, and so on (see figure 2.28b).

The colour coding in table 2.5 groups amino acids according to their side chains; for example, those with a positive charge are shown in blue.

TABLE 2.5 Genetic code shown as the 64 mRNA codons and information they specify. (See Appendix for the full names of the amino acids.) The codon AUG is a start signal and it also codes for the amino acid *met*. Note the three stop signals.

mRNA codon	Amino acid	mRNA codon	Amino acid	mRNA codon	Amino acid	mRNA codon	Amino acid
UUU	*phe*	UCU		UAU	*tyr*	UGU	*cys*
UUC		UCC	*ser*	UAC		UGC	
UUA	*leu*	UCA		UAA	STOP	UGA	STOP
UUG		UCG		UAG		UGG	*trp*
CUU		CCU		CAU	*his*	CGU	
CUC	*leu*	CCC	*pro*	CAC		CGC	*arg*
CUA		CCA		CAA	*gln*	CGA	
CUG		CCG		CAG		CGG	
AUU		ACU		AAU	*asn*	AGU	*ser*
AUC	*ile*	ACC	*thr*	AAC		AGC	
AUA		ACA		AAA	*lys*	AGA	*arg*
AUG	START /met	ACG		AAG		AGG	
GUU		GCU		GAU	*asp*	GGU	
GUC	*val*	GCC	*ala*	GAC		GGC	*gly*
GUA		GCA		GAA	*glu*	GGA	
GUG		GCG		GAG		GGG	

(a)

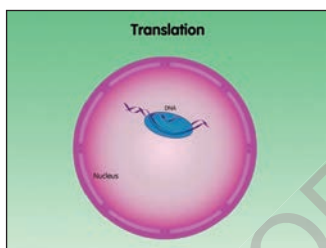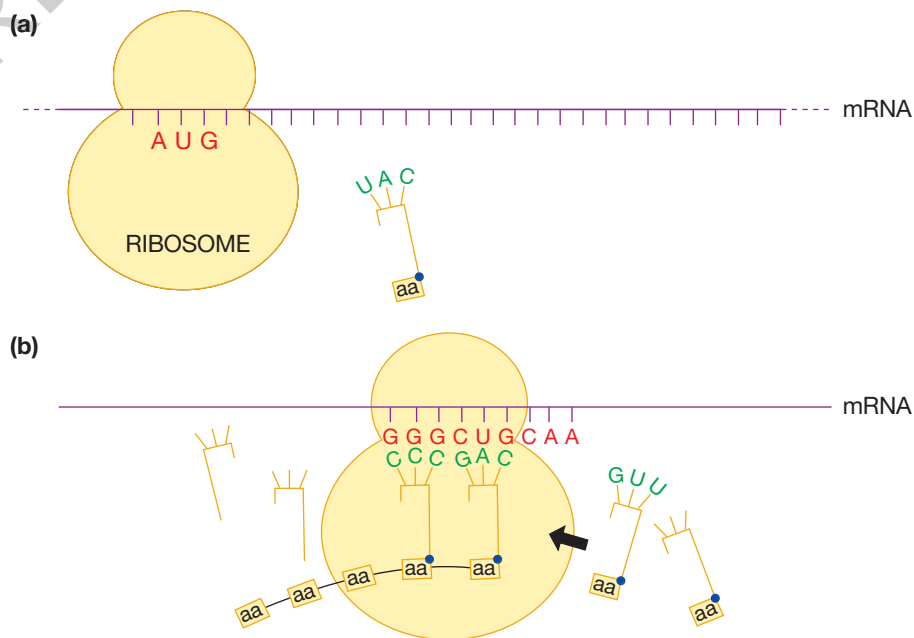
(b)

FIGURE 2.28 **(a)** The mRNA molecule attaches to a ribosome. In turn, as the ribosome moves along the mRNA molecule, each codon pairs with the tRNA with the complementary anti-codon. **(b)** The amino acids carried by each tRNA molecule are joined to form a chain. The final product is a protein consisting of a chain of amino acid sub-units, which, if short, is termed a polypeptide.

As each codon is translated, the tRNA molecule with the complementary anti-codon pairs momentarily with the mRNA. The pairing between bases in codons and the complementary bases in an anti-codon is as follows:

A pairs with U
U pairs with A
C pairs with G
G pairs with C.

So, when the mRNA codon UUU is reached, the tRNA carrier molecule that has the anti-codon AAA comes into place with its specific cargo of the amino acid *phe*. The amino acid carried by that tRNA is brought into the correct position to be joined into the growing protein chain. Amino acids continue to be added until a STOP signal is reached, which stops the addition of amino acids to the protein chain.

What is the mRNA codon for STOP? Will this have a corresponding anti-codon on a tRNA carrying an amino acid?

Putting it all together: transcription and translation

The production of a particular protein, such as beta chains of haemoglobin, starts in the cell nucleus. It is here that pre-mRNA molecules are produced by transcription from a template DNA chain. The mRNA leaves the nucleus and moves to the cytosol where it becomes attached to ribosomes. It is here that protein chains are formed by translation of mRNA.

Figure 2.29 shows **a summary of the processes of transcription and of translation.** For simplicity, this diagram omits one important step. Can you identify the missing step? By referring to the genetic code shown in table 2.5 (see page xxx), identify the amino acid that is about to be added to the growing protein chain in this diagram.
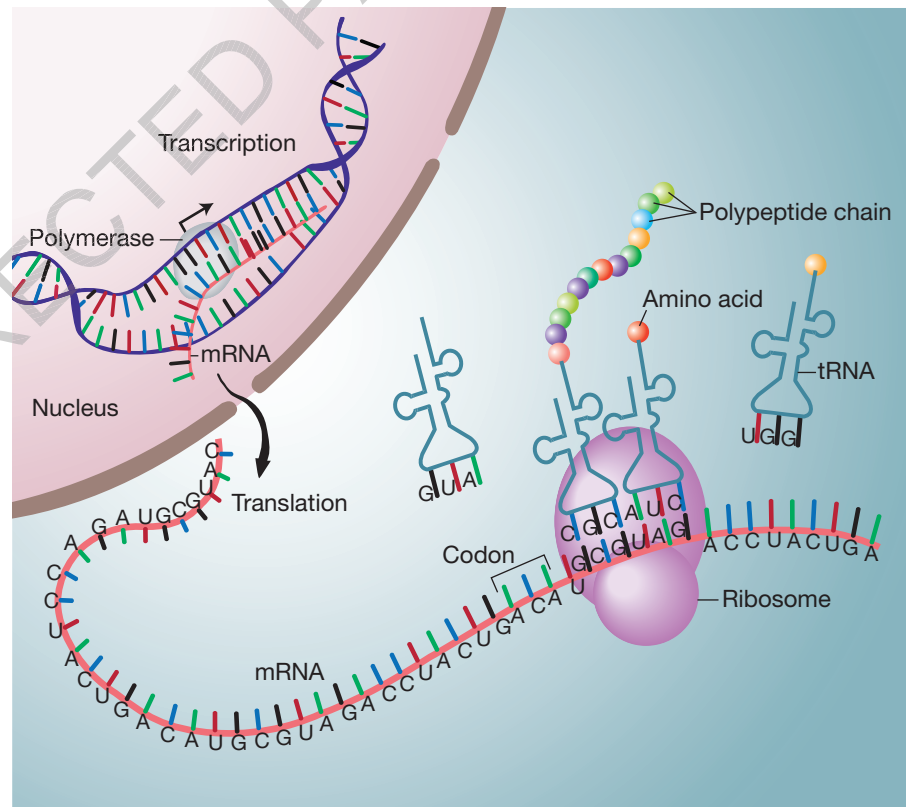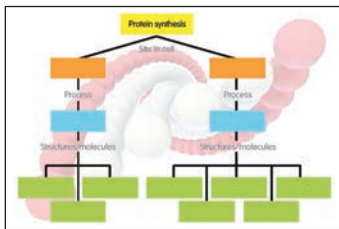

FIGURE 2.29 Representation of the processes of transcription and translation in a eukaryote cell (*Source:* National Human Genome Research Institute). Which of these processes involves the action of three different kinds of RNA?

Alternative splicing of pre-mRNA

The human genome contains only about 21 000 genes and this range is typical of other mammals. In the past, it was accepted that one gene had a single function as, for example, producing one particular protein — this is the 'one gene, one polypeptide' concept. The question remains: How can a relatively small number of genes produce the complexity of structure and function of a living mammal?

Research is now revealing that **one gene can be regulated in different ways so that it can produce more than one protein**. This means that:

- one gene could produce one protein at one stage of development but a different protein at another stage of development
- one gene could produce a particular protein in one tissue but a different protein in another tissue.

It is estimated that more than 60 per cent of human genes are regulated to operate in this way. The complexity of mammals is not in the number of genes that they have, but in the complex processes by which their genes are regulated.

How might one gene produce different protein products at different developmental stages or in different tissues? One way is through **alternative splicing** of the pre-mRNA molecules from a single gene.

Alternative splicing can occur during post-transcription modification of pre-mRNA and is the major mechanism enabling the same gene to produce different proteins in different tissues. Usually, alternative splicing involves **exon juggling**, where different exons are combined to form several kinds of mRNA, each with a different base sequence (see figure 2.30).
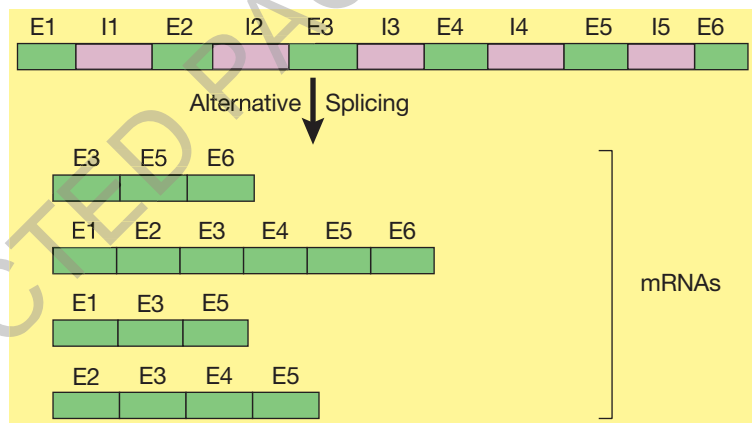
FIGURE 2.30 Exon juggling of pre-mRNA (E = exon and I = intron). Juggling of exons generates mRNAs that differ in their base sequences and so code for different proteins. Are any mRNAs possible other than those shown?

These different mRNAs are then translated into proteins that differ in their amino acid sequences. For example, the human **TPM** gene has 11 exons. Juggling of these exons produces different mRNAs in different tissues with the result that each tissue has a different form of the protein (see table 2.6).

TABLE 2.6 Result of exon juggling of the pre-mRNA from one human **TPM** gene in different tissues. The resulting proteins differ in their amino acid sequences. Why?

Tissue	Exons in mRNA	Excised exons
skeletal muscle	exons 1, 3–11	exon 2
smooth muscle	exons 1, 2, 4–9, 11	exons 3, 10
liver	exons 1, 4–6, 8, 9, 11	exons 2, 3, 7, 10
brain	exons 1, 4–9	exons 2, 3, 10, 11

Alternative splicing means that the number of proteins from the genetic instructions (genes) in a genome is far greater than the number of genes. Identifying how one gene can be regulated in different tissues, or at different times, to produce different products is an exciting area of ongoing research.

Genes have various functions

Structural and regulator genes

Genes vary in the functions that they carry out in the cells of an organism. **Some genes produce proteins that become part of the structure and the functioning** of the organism. These genes are termed **structural genes**.

Some genes produce proteins that control the action of other genes. These genes are termed **regulator genes** and their actions determine whether other genes are active ('on') or not ('off') and, if active, the rate at which their products are made.
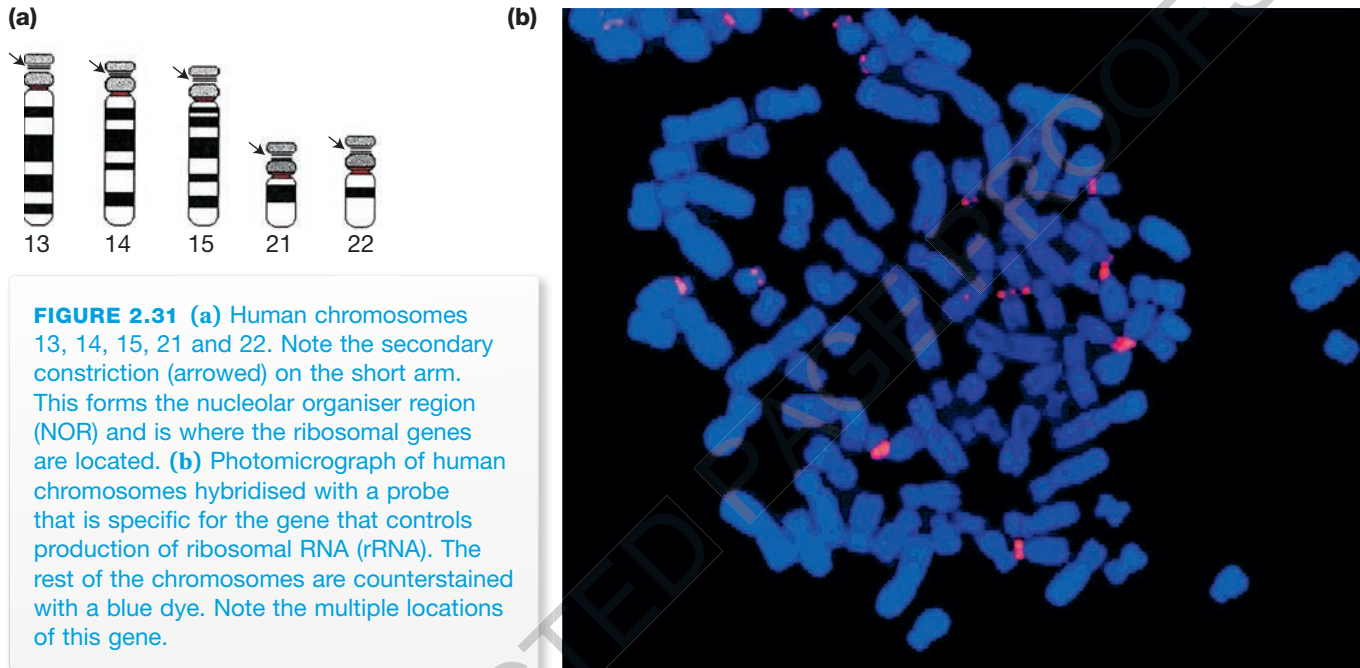
(a)

(b)

FIGURE 2.31 **(a)** Human chromosomes 13, 14, 15, 21 and 22. Note the secondary constriction (arrowed) on the short arm. This forms the nucleolar organiser region (NOR) and is where the ribosomal genes are located. **(b)** Photomicrograph of human chromosomes hybridised with a probe that is specific for the gene that controls production of ribosomal RNA (rRNA). The rest of the chromosomes are counterstained with a blue dye. Note the multiple locations of this gene.
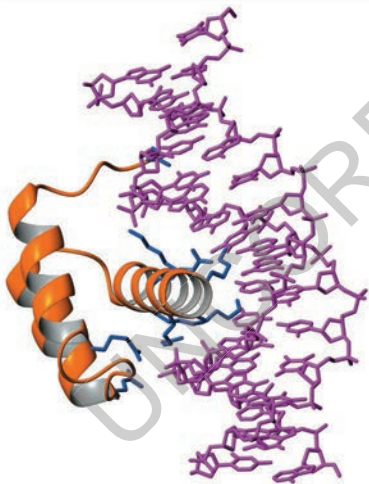
FIGURE 2.32 Homeotic genes produce a polypeptide with a region that carries + + charges that can bind to DNA. Here, the DNA-binding region of the polypeptide is shown bound to DNA (purple).

Regulator genes switch other genes 'on' or 'off' by producing proteins that act in one of two different ways:
1. Some proteins, known as **DNA-binding proteins**, bind to regions of nuclear DNA near genes and directly switch these genes on or off. These proteins carry a net positive charge that enables them to bind to DNA, which has a net negative charge. By binding to the DNA near their target genes, these proteins can switch these genes on and off (see figure 2.32).
2. Some proteins bind to receptors on the membrane of cells in their target tissue and trigger a series of intercellular reactions that switch genes on or off; these are **signalling proteins**.

Genes that control embryonic development in insects and in vertebrates are master genes, known as **homeotic genes**, and are examples of regulator genes. These genes act by producing DNA-binding proteins. Homeotic genes control the action of hundreds of other genes that are needed to build the various parts of an animal body in their correct locations. In insects, for example, when a particular homeotic gene malfunctions, the result is an insect that has, instead of antennae, legs on its head (see figure 2.33, page xxx). If another homeotic gene does not operate normally, the affected insect may have an antenna where it would normally have a wing.

Homeotic genes that organise the body plan of mammals are known as **HOX** genes. Mutations in these genes reveal how they control the pattern of

the body plan during embryonic development. For example, in mice, mutation of the HOXC8 gene results in an extra pair of ribs. In humans, mutation of the HOXD13 gene results in limb abnormalities, with an extra digit between the third and fourth digits and with all three digits fused. If the genes at one end of the HOXD complex are removed by a chromosomal deletion, severe limb and genital deformities result.

(a) Halteres
Antennae

(b) Wings replace halteres
Legs replace antennae
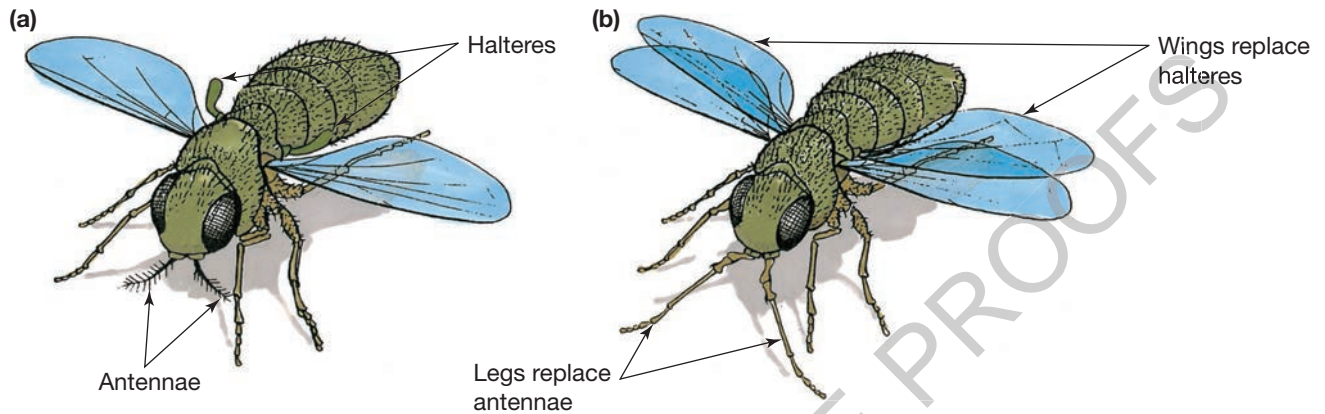
FIGURE 2.33 Mutations in 'master' or homeotic genes result in the appearance of body parts in unexpected locations. **(a)** Normal fly **(b)** Fly has mutations in homeotic genes that result in the appearance of a second pair of wings instead of halteres and legs instead of antennae.

Homeotic genes produce DNA-binding proteins. By binding to the DNA near other target genes, these proteins can switch these genes on and off.

KEY IDEAS

- All active genes produce RNAs of some kind.
- Most genes transcribe mRNAs that are then translated into proteins.
- Some genes produce other kinds of RNA, such as tRNA and rRNA, and these are the end products.
- Genes can be classified as structural genes or as regulator genes.
- Regulator genes control the activity of other genes, switching them on or off, either directly through the action of DNA-binding proteins, or indirectly through the action of signalling proteins.

QUICK CHECK

27 List one difference between a structural gene and a regulator gene.
28 What is a homeotic gene?

Time and place for everything

When are genes active?

Genes vary in the time of their action. Some genes are active in making mRNA and proteins only during a short period of the life span of a person, while other genes are active throughout a person's life.

The **HBZ** gene on the number-11 chromosome controls production of one kind of chain found in a type of haemoglobin that occurs only in mammalian embryos. After the first few weeks of embryonic development, this gene is switched off and remains silent during fetal development and after birth.

A baby boy with the genotype *d* (Y) for the **DMD** gene on the X chromosome is healthy at birth and for several years after birth. However, the signs of Duchenne muscular dystropy gradually appear during childhood, so that, by about 10 years old, he is confined to a wheelchair (refer back to figure x.xx).

Some genes are not expressed in the phenotype until a person is well into adulthood. A person with the genotype *Hh* has the allele for Huntington's disease on the number-4 chromosome. The *H* allele remains inactive often until middle age when the person shows the first signs of this devastating disease.

The **AD1** gene, on the number-21 chromosome, is involved in a familial form of Alzheimer disease. The expression of the specific allele (*A*) typically occurs after the onset of middle age when the person shows the first signs of familial Alzheimer disease.

Many genes remain active throughout the life of a person. These genes include the genes responsible for controlling the production of enzymes that are essential for cellular respiration, such as the **SDH** gene located on the number-1 chromosome. This gene controls production of a sub-unit of one of the enzymes (succinic dehydrogenase) essential for cellular respiration.

Where are genes active?

Some genes are active only in the cells of specific tissues:
- The **HBB** gene on the number-11 chromosome, which controls the production of the beta chains of haemoglobin, is active only in those cells of the bone marrow destined to become red blood cells.
- The **DMD** gene on the X chromosome, which controls production of the protein dystrophin, is active only in skeletal muscle tissue.
- The **GH** gene on the number-17 chromosome, which controls production of growth hormone, is active only in the pituitary gland at the base of the brain.

In contrast, genes involved in controlling production of enzymes involved in cellular respiration are active in all living cells.

The term *gene action* refers to the processes of gene transcription and gene translation. The end products of gene action typically are proteins, such as beta protein chains of the **HBB** gene or amylase enzyme of the **AMY** gene. However, the final phenotype of an organism is more complex than just proteins produced by the genes.

Identifying active genes

Of the approximately 21 000 genes in a human cell, only some genes are expressed or 'switched on' at a given time. A 'switched on' gene is one that is transcribing the mRNA, while a 'switched off' gene is one that is not producing mRNA. But which genes are active? The pattern of gene activity or gene expression differs between:
- cells of different tissues in the same organism
- cells of the same tissue but at different stages of development
- normal and abnormal cells, such as cancerous cells
- cells under different environmental conditions, such as cells exposed to chemical pollutants or drugs.

In the past, it was possible to examine whether only one or a few genes were active. Through a new technology involving **microarrays** (also known as **DNA arrays** and **gene chips**), **it is now possible to study large numbers of**

ODD FACT

Commercially available microarrays include one with 32 996 spots covering the mouse genome and one with 32 878 spots covering the human genome. In both cases, each spot comprises a segment of DNA, 60 bases long, from a different gene.

genes simultaneously or even the entire genome. This means that it is now possible to:

- identify which genes are active and which genes are switched off
- compare gene expression in different cell types
- compare active genes in the same cells under different conditions.

A microarray consists of a glass slide (or other solid surface) on which thousands of tiny spots of different DNA molecules are present. Each spot is in a precise location and consists of a short segment of single-stranded DNA from one particular gene (see figures 2.34 and 2.35). The spots are put in place using a robotic instrument to ensure their precise locations. Typically, the DNA spots in a microarray are organised into segments, with each often having up to 20 columns and 20 rows.

Microarray (chip) Segment of a chip Spot containing copies of a single DNA molecule Part of one DNA strand

AGGACGT — DNA bases
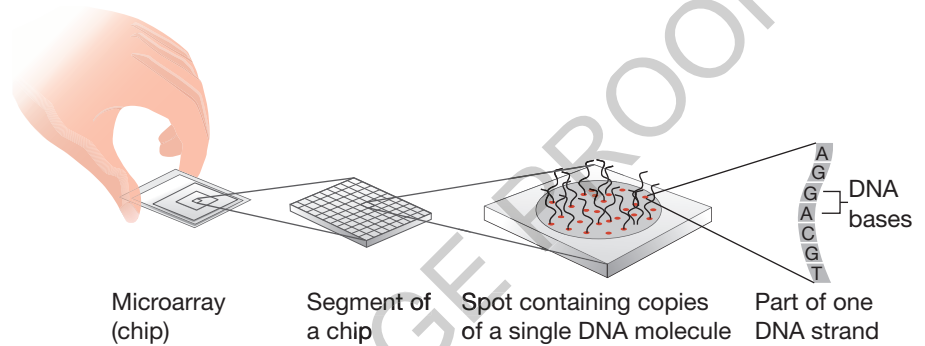
FIGURE 2.34 Diagram showing a microarray (at left), one segment from this microarray (middle) and one spot from a segment (at right). Note that the DNA in each spot contains single-stranded DNA from one specific gene. The DNA spots are typically 150 to 200 micrometres (μm) in diameter and the centres of the spots might be 300 μm (0.3 mm) apart. So, one segment with 20 rows and 20 columns would fit in a square with a side of 6 millimetres.
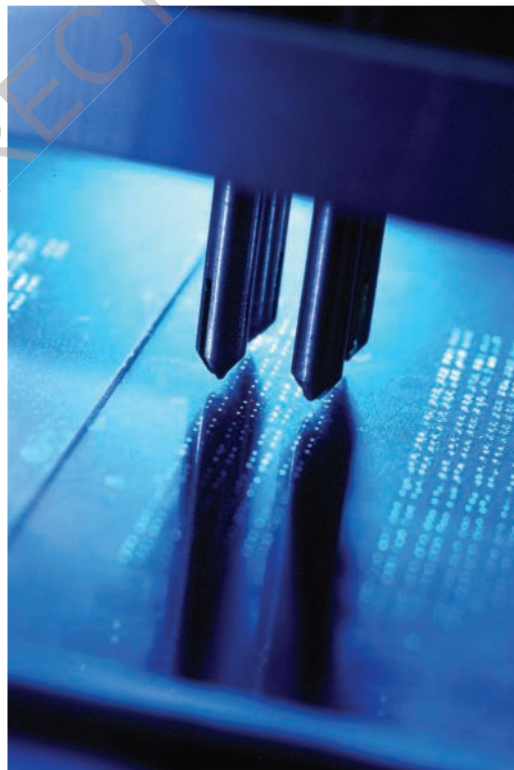
FIGURE 2.35 DNA microarrays enable the simultaneous identification of the active genes in a cell. This figure shows the printing tool that applies the spots to a microarray.

Each of the thousands of spots in the microarray acts as a probe to detect the activity of one different gene. How? Active genes produce mRNAs that can bind at the spots where the complementary single-stranded DNA is located. (If a gene is not active, no mRNA is being transcribed so no binding occurs.)

Applications of microarray technology

The ability to recognise which genes are active can assist our understanding of the role of genes in various diseases. For example, Swedish researchers used microarrays to identify the active genes in breast cancer cells from 159 patients who were followed up after surgery to remove the cancer. They found that patients whose cancer returned or who died during the study period shared 64 active genes in common, but this set of genes was not active in other patients where the cancer did not reoccur. This study showed how microarrays might be used to predict how people with breast cancer will respond to treatments and these findings may lead to new treatments for cancer. The following box outlines how microarrays work.
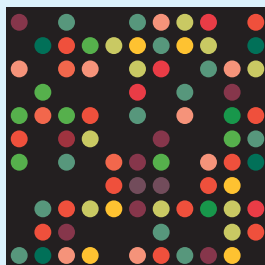
HOW DO MICROARRAYS WORK?

Imagine that some research scientists want to identify how gene activity changes when mouse cells are exposed to a particular drug. Remember that mRNA production signals an active gene. To do this, the scientists need:

1. a 'mouse' microarray with the single-stranded DNA of the spots coming from genes of the mouse genome
2. two cell samples, one from cells that have been exposed to the drug treatment and the second from normal cells that have not been exposed to the drug treatment. (Why two cell samples?)

The next step is to separately extract all the mRNA from each cell population. This mRNA is used as a template to make single-stranded DNA known as **copy DNA (cDNA)**. The unknown cDNA molecules from the two cell samples are labelled with a coloured fluorescent marker so that they can later be detected, with a different colour being used for the control cells (green label) and the drug-treated cells (red label).

Next, the mixtures of unknown cDNA molecules from both samples are exposed to the microarray.

When a cDNA sequence is complementary to one of the spots on the microarray, the cDNA with its fluorescent label binds to this probe. Possible results for one gene are shown in table 2.7.

TABLE 2.7 Results expected when cDNA from untreated (control) cells with a green fluorescent label and drug-treated cells with a red label are probed using a microarray

Control cells	Drug-treated cell	Result on microarray
Gene ON	Gene OFF	GREEN spot
Gene ON	Gene ON	YELLOW spot
Gene OFF	Gene ON	RED spot
Gene OFF	Gene OFF	no colour

Finally, the microarray is scanned using a confocal laser microscope and the locations of the fluorescent spots are identified and displayed on a computer screen (see figure 2.36).

🔴 Gene that strongly increased activity in drug-treated cells

🟢 Gene that strongly decreased activity in drug-treated cells

🟡 Gene that was equally active in treated and untreated cells

⬛ Gene that was inactive in both groups

FIGURE 2.36 Because of the large amount of data involved, computer software is used to analyse the microarray results. Here we see one segment of a microarray and the information that it provides. Does this result provide evidence that the drug treatment changed the activity of genes in the treated cells compared with the control cells?

Gene regulation in bacteria

In 1961, the French scientists Jacques Monod (1910–1976) and François Jacob (1920–2013) made a significant discovery — they were the first to document a gene regulation system. Rather than producing the same enzymes all the time, Jacob and Monod found that the bacteria produced enzymes only when needed — for example, they discovered that the bacteria *Escherischia coli (E. coli)* produced the enzymes needed to metabolise the sugar lactose only when lactose was present and when glucose was not available. In other words, the bacterial genes that encode the enzymes to metabolise lactose are expressed only in the presence of lactose. Glucose is the favoured metabolite of these bacteria, and, when glucose is available, the genes governing lactose metabolism are silent.

Clearly, there was a mechanism in place to control the activity of the genes encoding enzymes involved with lactose. What Jacob and Monod discovered was how the activity of structural genes that encode particular enzymes are regulated by other genes known as regulator genes. They introduced **the term operon to describe a group of linked structural genes with a common promoter and operator, that is transcribed as a single unit**. The expression of operons is controlled by regulator genes that produce repressor proteins.

This was a significant discovery because it revealed, for the first time, the existence of a new class of genes, **regulator genes**, that control the activities of structural genes, switching them on and off as conditions require. For their discovery, Jacob and Monod shared the Nobel Prize in Physiology or Medicine in 1956.

Since the discovery of the **lac operon** in bacteria, other operons have been identified in bacteria, such as the *trp* operon.

Let's look at some details of the lac operon in E. coli.

The lac operon

Figure 2.37 shows the structure of the lac operon. Also shown is the regulator gene that controls the operon.
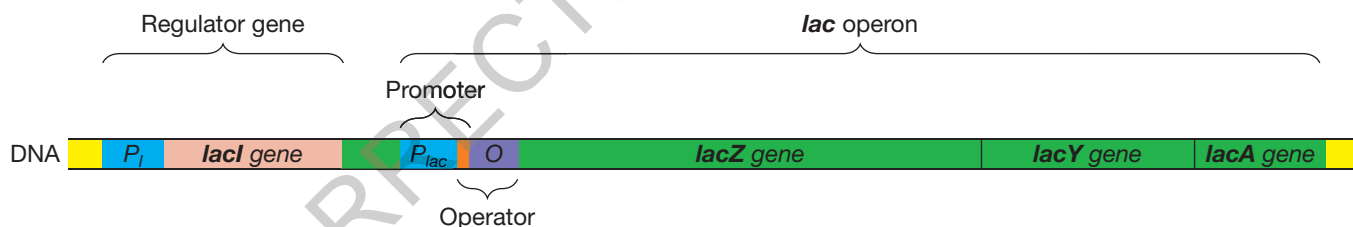
FIGURE 2.37

Check out this figure and note that the lac operon consists of:
- Three structural genes concerned with the metabolism of lactose as follows:
  - **lac Z** encodes the enzyme galactosidase, which breaks down lactose into glucose and galactose
  - **lac Y** encodes the enzyme permease, which enables lactose to enter cells
  - **lac A** encodes the enzyme acetyltransferase, which also has a role.
- A **promoter** (P_{lac}): The promoter is a short DNA segment where RNA polymerase can attach and start transcription of the three downstream lac genes. The three lac genes are transcribed as a single entity with one long mRNA transcript being produced.
- An **operator** (**O**): An operator is a short DNA segment that provides a binding site for a repressor.

Regulator genes may be located adjacent to an operon or be at a distance from the operon or may even be within it. There is no fixed location, unlike the requirement that the DNA sequences of the promoter and operator must be just upstream from the structural genes of the operon.

This operon is controlled by a regulator gene as follows:

• The **lac I** gene (with its upstream promoter): The lac I gene is a regulator gene that encodes a **repressor**. The repressor is a protein that, when active, down-regulates gene expression by preventing gene transcription.

What happens when lactose is absent? When lactose is absent, the repressor protein is active and binds to the operator, physically blocking RNA polymerase from attaching to the promoter. Transcription cannot start and the lac operon is repressed (see figure 2.38a).

What happens when lactose is present? When lactose is present, it binds to the repressor protein changing its shape and inactivating it so that it cannot bind to the operator. This means that RNA polymerase can attach to the promoter and start transcription of the structural genes (see figure 2.38b). Then the long mRNA transcript is translated to produce the enzymes needed to metabolise lactose. So, in the presence of lactose, the enzymes needed to transport it into cells and break it down into galactose and glucose are synthesised.

The operon system enables the bacteria to adapt rapidly to their environment and to use their resources economically by synthesising enzymes only when they are needed: a just-in-time strategy, rather than just-in-case.
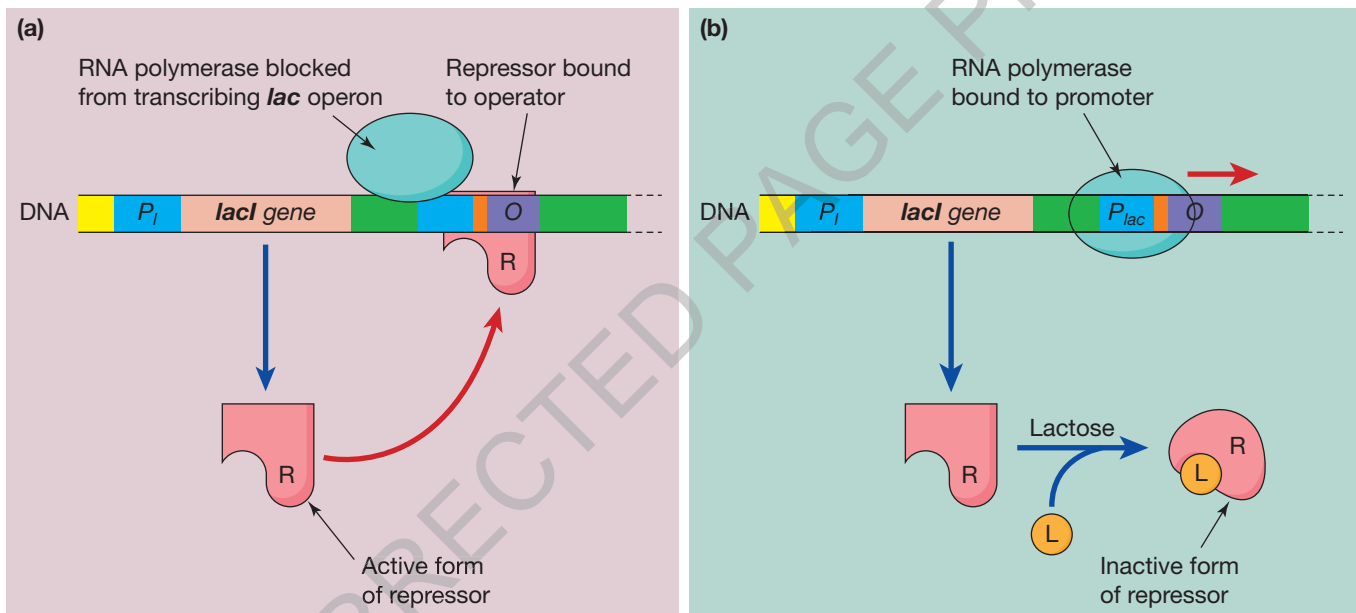
FIGURE 2.38 Details showing a regulator gene and part only of the lac operon it controls. **(a)** When an active repressor (R) is bound to the operator (O) of the operon, the RNA polymerase is blocked from attaching to the promoter (P_{lac}) and transcription of the structural genes of the operon cannot occur. **(b)** The repressor protein is inactivated by binding with lactose and cannot attach to the operator. RNA polymerase can now attach to the common promoter of the operon (P_{lac}) and transcription can occur so that the genes of the lac operon can be expressed.

BIOCHALLENGE

1 Examine each of the molecules in the diagram. For each one, answer the following:
   a Decide whether it is a monomer or a polymer and, if a polymer, name the monomers of which it is made.
   b Classify the molecule as comprehensively as possible with regard to its organic molecule category.
   c Name where, in either a plant or an animal cell, you could expect to find such a molecule.
   d When the sub-units shown in (b) join to form a larger unit,
      i What kind of bond is formed?
      ii Is this a condensation reaction or a hydrolysis reaction?

(a)

(b)

(c)

(d)
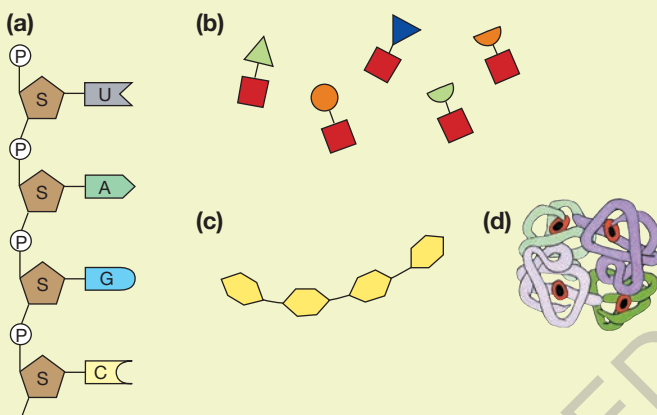
Thousands of years ago, woolly mammoths (*Mammathus primigenius*), now extinct, lived in frigid regions including present-day Siberia (see figure 2.39). The woolly mammoth evolved from a line that originated in Africa and diverged to give rise also to the African elephant (*Loxodonta africana*) and the Asian elephant (*Elephas maximus*). The woolly mammoth shows features that equip it for life in very cold climates.

2 The DNA sequences of the coding strand of the **HBA** and the **HBB** genes of the woolly mammoth were compared with those of the African and the Asian elephants. Unique mutations in the exons of the **HBB** gene of the woolly mammoth were found:

Exon	Base number*	Mutation
1	37	A → G
2	259	G → T
2	304	G → C

(*Base number is given relative to the ATG start codon, which begins at base number 1.)

Refer to the genetic code, shown as mRNA codons on page XXX, to answer the following:
   a In the woolly mammoth, the A → G mutation at base number 37 in the DNA coding strand produces an amino acid substitution in the beta-globin chain of haemoglobin, with *thr* being replaced by *ala*. Identify what bases number 38 and number 39 could be in the DNA coding strand.
   b In the woolly mammoth, the G → T mutation at base number 259 produces a new sequence (TCC instead of GCC) in the DNA coding strand of the **HBB** gene. What effect will this have on the amino acid sequence in the beta-globin chain of haemoglobin?
   c A mutation present in the exon 3 of the Asian elephant is a G → A substitution at the third position of the 119th codon. This mutation, however, is silent as it does not change the amino acid in the beta-globin chain. Give another example of a silent mutation in the coding strand of the DNA of a gene.
   d A codon consists of a three-base sequence, either in the coding strand of DNA or in the mRNA strand. A single mutation in a codon may affect the first, second or third base in the codon. In terms of changing the amino acid sequence, which position of a codon has the most impact?

FIGURE 2.39 Woolly mammoth, an extinct mammal, closely related to Asian and African elephants.

Chapter review

Key words

alternative splicing
amino acids
anti-codon
coding region
codons
conjugate
copy DNA (cDNA)
decoded
deoxyribonucleic acid (DNA)
DNA arrays
DNA-binding proteins
DNA sequencers
encoded

exon
exon juggling
flanking regions
gene action
gene chips
gene sequence
gene sequencing
genetic code
homeotic genes
introns
lac operon
messenger RNA (mRNA)
microarrays
monomers

nucleic acids
nucleoproteins
nucleotides
nucleotide sequence
operon
polymers
polypeptide
post-transcription modification
promoters
proteins
proteome
proteomics
regulator genes

ribonucleic acid (RNA)
ribosomes
RNA polymerase
signalling proteins
spliceosomes
structural genes
TATA box
template strand
transcription
transfer RNA (tRNA)
translation
triplet code

1 Applying your knowledge and understanding ➔
a When two monomers such as amino acids join together, a molecule of water is produced. Use two amino acid molecules to explain where this water comes from.
b How many water molecules would be required to completely hydrolyse a carbohydrate polymer that contained 100 monomers?

2 Applying understanding and drawing conclusions ➔ Before the introduction of genetically engineered insulin for use by people with diabetes, the protein hormone was extracted from beef or pig pancreas. Explain how you would expect the sequence of amino acids in the beef and pig insulin to compare with that of humans.

3 Analysing data and drawing conclusions ➔ A particular small polypeptide contains nine amino acids. The polypeptide has been fragmented in various experiments by breaking particular peptide bonds. The fragments obtained were:

ser – cys – his – pro – arg – cys
pro – arg – cys
X – gly – met – cys
his – pro – arg – cys
X – gly – met – cys – ser – cys

X is known to be the first amino acid in the polypeptide. What is the primary structure of the polypeptide?

4 Analysing data and drawing conclusions ➔ Assume that the nitrogen base sequence in one of the strands in a DNA molecule is:
A-T-C-G-A-C-A-T-G-G-A-A-T-A-C-C-T-C
a What is the base sequence in the complementary strand?

b How many amino acids does this piece of DNA code for?

5 Making connections ➔ Use at least eight of the key words from this chapter to form a map for the concept 'gene action'. In drawing your map, add any other concepts that you wish.

6 Demonstrating understanding ➔ The following is part of the nucleotide sequence in the template strand of part of a gene:
. . . T A T G G G C A T G T A A T G G G C . . .
a Identify the base sequence in each of the following:
  i the complementary DNA strand
  ii the mRNA that would be transcribed from this template.
b How many codons are present in this mRNA?
c List the anti-codons that correspond with each codon.

7 Applying knowledge ➔ Refer to the genetic code (see table 2.5 on page XXX) and answer the following questions:
a Which codons in mRNA control the addition of the amino acid *gly*?
b How many codons contain the information to add the amino acid *lys* to a protein? For each codon, write the complementary anti-codon.
c When the mRNA codon UUU is translated, which amino acid is added to a protein chain?

8 Demonstrating understanding ➔ A protein includes the following amino acids in part of its structure:
. . . - *val - thr - lys - pro* - . . .
a How many codons are needed for the instruction to put these amino acids into place?

b Write this instruction in genetic code, as it would appear in mRNA. Would you predict that your code would be identical to that written by all your fellow students? Explain.

9 **Analysing information and drawing conclusions** ➜ A mutation in one gene (G1) affects just one kind of protein produced by a cell. A mutation in another gene (G2) affects a large number of different kinds of proteins produced by that cell. One of these genes carries the coded instructions to make one kind of tRNA. The other carries coded information to make one kind of salivary enzyme.

Which is more likely to be the gene for the tRNA: **G1** or **G2**? Explain.

10 **Demonstrating understanding** ➜ A segment of mRNA has the base sequence:

... C A U A A G A A U C U U G C ...

a Write the base sequence of the DNA template strand.

b Write the amino acids that would be translated from this mRNA segment.

c Assume that a base substitution occurs in the original DNA so that the third base (U) of the mRNA is replaced by a G:

... C A G* A A G A A U C U U G C ...

Write the amino acid sequence that would result from this change.

d Assume that a base addition occurs in the original DNA so that a G is added between the third and fourth bases:

... C A U G*A A G A A U C U U G C ...

Write the amino acid sequence that would result from this change.

e On the basis of this information, what kind of change in DNA has a more extensive effect on the protein resulting from gene translation — a base substitution or a base addition? Explain.

11 **Applying knowledge** ➜

a *Where is it?*

Where would you find the following:

   i a codon

   ii gene transcription in action

   iii an anti-codon

   iv gene translation in action

   v nucleolar organiser region?

b *What is it?*

Suggest possible identities for the following:

   i a STOP codon

   ii a START codon

   iii an amino acid that has six codons

   iv a self-replicating molecule that carries information in coded form

   v the cell organelle to which an mRNA molecule attaches for translation.

12 Undertake some online research to find the gene that causes cystic fibrosis (the search term OMIM can be useful).

Answer the following questions about the gene that causes cystic fibrosis:

a How many exons are there in the DNA of this gene?

b The pre-mRNA is longer than the mRNA. Explain.

c The coding sequence within the mRNA is shorter than the mRNA. Explain.

d What relationship exists between the number of bases in the coding sequence of the mRNA and the number of amino acids in the protein product?

13 *E. coli* bacteria have a requirement for amino acids, including tryptophan (*trp*). These bacteria can take up *trp* from their environment, but if it is not available, *E. coli* can synthesise this amino acid. The five genes involved in the synthesis of tryptophan are part of a system called the ***trp operon***. (This is similar to the lac operon)

a As well as the five structural genes needed to synthesise *trp*, what other DNA segments form part of the *trp* operon?

b Draw a rough line diagram showing the essential components of the *trp* operon.

Consider a situation in which tryptophan is present in the environment in which the *E. coli* bacteria are growing.

c Under these conditions, do the *E. coli* need to synthesise *trp*?

d Under these conditions, would you expect that the *trp* operon would be repressed or be activated?

If tryptophan is *not* available from the environment in which the bacteria are growing, they will manufacture it themselves.

e Under these conditions, would you expect that the *trp* operon would be repressed or be active?

When the *trp* operon is repressed, the structural genes that encode the various enzymes needed are silent and the *trp* operon is said to be repressed.

f Identify a possible means by which the *trp* operon might be repressed.

When the *trp* operon is activated, the structural genes that encode the various enzymes needed for its synthesis and transport from cells are transcribed and translated and the *trp* operon is said to be activated.

g Identify a possible means by which the *trp* operon might be activated.