3.1.3 The detailed process of DNA replication

DNA replication ensures that exact copies of existing molecules are produced before a cell divides. The process is said to be semi-conservative and each strand of an existing DNA molecule acts as a template for the production of a new strand (see Nature of Science, Obtaining evidence: Meselson and Stahl's experiment and semi-conservative replication of DNA, earlier in this section).

In eukaryotes, replication is controlled through interactions between proteins, including cyclins and CDKs (Section 6.5) and takes up to 24 hours to complete. Each of the original DNA strands acts as a template to build up a new strand (Figure 3.1.7). The DNA double helix is unwound to expose the two strands for replication by the enzyme DNA helicase, at a region known as a replication fork. The action of helicase creates single-stranded regions, which are less stable than the double-stranded molecule. To stabilise these single strands, **single-stranded binding proteins** (SSBs) are needed. SSBs protect the single-stranded DNA and allow other enzymes involved in replication to function effectively upon it.

Replication must occur in the 5′→3′ direction (and also in transcription and translation, described in Section 3.2), because the enzymes involved only work in a 5′→3′ direction (adding new nucleotides to the 3′ end of the newly forming DNA molecule). As the two strands are antiparallel, replication has to proceed in opposite directions on the two strands. However, the replication fork where the double helix unwinds moves along in one direction only. This means that on one of the strands

replication can proceed in a continuous way, following the replication fork along, but on the other strand the process has to happen in short sections, each moving away from the replication fork (Figure 3.1.8). The strand undergoing continuous synthesis is called the **leading strand**. The other strand, in which the new DNA is built up in short sections, is known as the **lagging strand**.

KEY POINTS

lagging strand is the new strand that is synthesised in short fragments in the opposite direction to the movement of the replication fork.

leading strand is the new strand that is synthesised continuously and follows the replication fork.

single-stranded binding protein is the protein which binds to single-stranded regions of DNA to protect them from digestion and remove secondary structure.

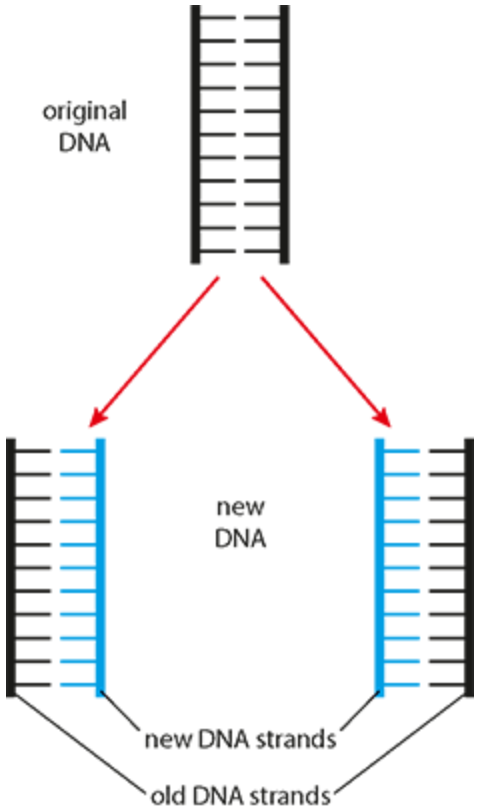

Figure 3.1.7: DNA replication is semi-conservative. As it is copied one original strand becomes paired with one new strand. One of the two strands in each new DNA molecule is conserved, hence 'semi-conservative'.

Copying the leading strand

Replication to produce the leading strand begins at a point on the molecule known as the 'origin of replication' site. First **RNA primase** adds a short length of RNA, attached by complementary base pairing, to the template DNA strand. This acts as a **primer**, allowing the enzyme **DNA polymerase III** to bind. DNA polymerase III adds free 'building units' called **deoxynucleoside triphosphates (dNTPs)** to the 3′ end of the primer and then to the forming strand of DNA. In this way the new molecule grows in a 5′→3′ direction, following the progress of helicase as it moves the replication fork along the DNA

double helix. The RNA primer is later removed by **DNA polymerase I**. In this way, a continuous new DNA strand is built up on the leading strand.

KEY POINTS

primer is short strand of nucleic acid that forms a starting point for DNA synthesis.

RNA primase is an enzyme that catalyses the synthesis of RNA primers as the starting point for DNA synthesis.

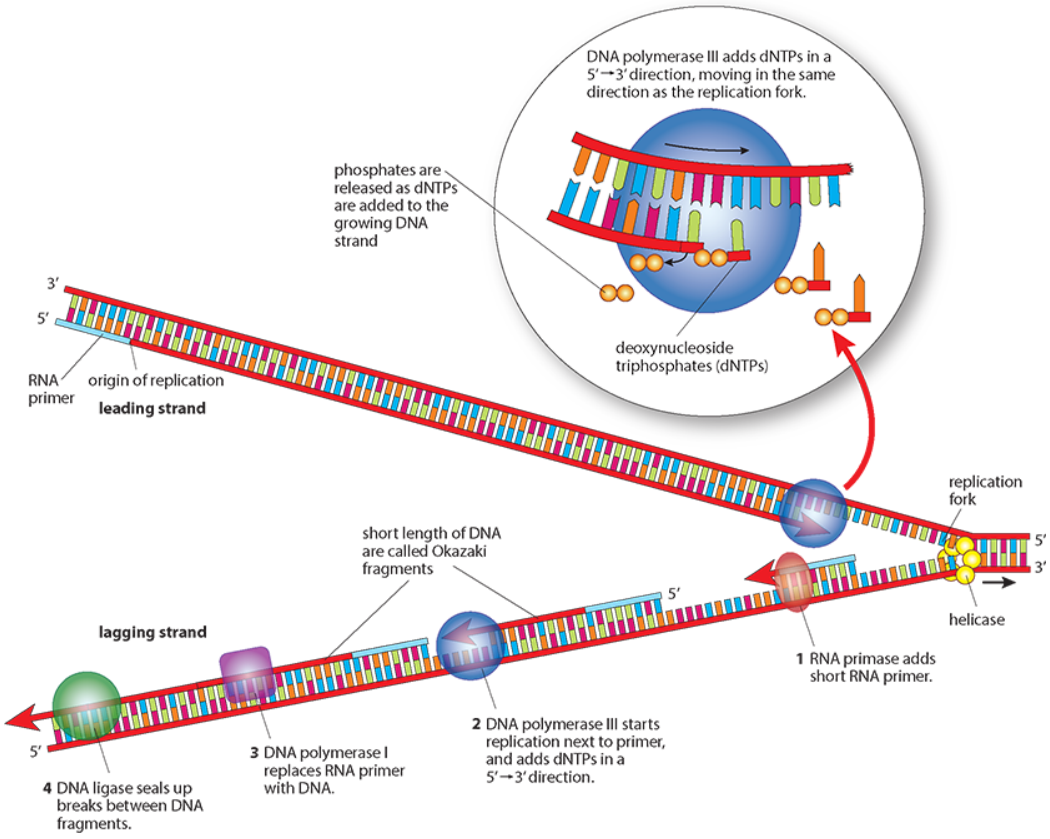

Figure 3.1.8: DNA replication showing the leading and lagging strands, and the direction of DNA synthesis.

KEY POINTS

DNA polymerase III extends the new DNA strand in a 5′ → 3′ direction from the RNA primer.

deoxynucleoside triphosphate (dNTP) is a building block for DNA: deoxyribose, three phosphate groups and one of the four bases.

DNA polymerase I removes the RNA nucleotides of the primers on the lagging strand and replaces them with DNA nucleotides.

EXTENSION

The dNTPs have two extra phosphate groups attached, and are said to be 'activated'. They pair up with their complementary bases on the exposed DNA strand and DNA polymerase III then link together the sugar and the innermost phosphate groups of adjacent nucleotides. The two extra phosphate groups are broken off and released.

Copying the lagging strand

Synthesis of the lagging strand is a little more complicated, as it has to occur in discontinuous sections, which are then joined together.

1 As for the leading strand, **DNA primase** first synthesises a short RNA primer, complementary to the exposed DNA. This happens close to the replication fork.

2 DNA polymerase III starts replication by attaching at the 3′ end of the RNA primer and adding dNTPs in a 5′→3′ direction. As it does so, it moves away from the replication fork on this strand.

3 DNA polymerase I now removes the RNA primer and replaces it with DNA using dNTPs. Short lengths of new DNA called **Okazaki fragments** are formed from each primer. The new fragment grows away from the replication fork until it reaches the next fragment.

4 Finally, **DNA ligase** seals up each break between the Okazaki fragments by making sugar–phosphate bonds so that a continuous strand of new DNA is created.

KEY POINTS

DNA primase a type of RNA polymerase catalyses the production of a short length of RNA called a primer which is base-paired to the parent DNA strand. The primer is removed when replication is complete and replaced by DNA.

DNA ligase joins adjacent Okazaki fragments by forming a covalent bond between adjacent nucleotides.

Okazaki fragments short fragments of a DNA strand formed on the lagging strand.

Proofreading new DNA

DNA polymerases 'proofread' their work as they build up new DNA strands. If the polymerase enzyme detects that an incorrect nucleotide has been added and does not pair up correctly the enzyme will remove and replace it.

EXAM TIP

There are several important enzymes to remember in the process of replication so it is helpful to keep a list of them and their jobs.

TEST YOUR UNDERSTANDING

4 Outline what is meant by antiparallel.

5 State the direction in which DNA replication occurs.

6 Outline the role of DNA primase.

7 Summarise the differences between forming the leading and lagging strands.

REFLECTION

Could you explain DNA profiling to someone who had never heard of it? Reflect on its importance to your understanding of ourselves.

Links

- How is the molecular structure of DNA linked to its function? (Chapter 1)
- Why must the genetic code carried by DNA be copied exactly? (Chapter 3.4)
- How is replication involved in cell division? (Chapter 6)

3.2 Protein synthesis

LEARNING OBJECTIVES

In this section you will:

- define transcription as the synthesis of RNA from a DNA template
- recognise that transcription is needed for the expression of genes
- learn that complementary base pairing between DNA and mRNA ensures that the polypeptides produced function properly
- define translation as the production of polypeptides from mRNA using tRNA
- recognise how complementary base pairing between codons and anticodons ensures accurate translation
- learn that ribosomes are the sites of translation; free ribosomes synthesise proteins for use within the cell, whereas bound ribosomes synthesis proteins for secretion or use in lysosomes

- understand the directional nature of transcription and translation
- recognise that transcription begins at a promoter region
- understand that in prokaryotes, translation occurs immediately after transcription but eukaryotes modify

- mRNA by removing introns to form mature mRNA composed of exons
- recognise that exons can be spliced in different ways to produce different proteins from a single gene
- understand how nucleosomes regulate transcription in eukaryotes
- understand that a large portion of the eukaryotic genome consists of non-coding sequences
- learn that non-coding DNA persists for many generations and has important functions
- recognise that polysomes allow many polypeptides to be made at the same time
- understand that translation does not always result in functional protein and that polypeptides are modified before they can function
- learn that amino acids are recycled in the cell by proteasomes.

GUIDING QUESTIONS

- How does complementary base pairing contribute to the resilience of the genetic code?
- How are enzymes involved in protein production?

3.2.1 Transcription

The main role of DNA is to direct the activities of the cell. It does this by controlling the proteins that the cell produces. Enzymes, hormones and many other important biochemical molecules are the proteins that control what the cell becomes, what it synthesises and how it functions. Protein synthesis can be divided into two sets of reactions: the first is transcription and the second is translation. In eukaryotes, transcription occurs in the nucleus and translation in the cytoplasm.

The sections of DNA that code for particular proteins are known as genes. Genes contain specific sequences of bases in sets of three, called triplets. Some triplets control where transcription begins and ends.

KEY POINTS

gene is a particular section of a DNA strand that codes for a specific polypeptide; a heritable factor that controls a specific characteristic.

transcription means copying a sequence of DNA bases to mRNA.

translation decoding mRNA at a ribosome to produce a polypeptide.

KEY POINTS

triplet a sequence of three bases that code for an amino acid.

messenger RNA (mRNA) is a single-stranded transcript of one strand of DNA, which carries a sequence of codons for

the production of protein.

Copying the DNA message to RNA

The first stage in producing a protein is the production of messenger RNA from a segment of DNA so that the genes that code for the required polypeptides can be moved to the cytoplasm. After this the message is translated and the necessary amino acids are used to build polypeptide chains.

The first stage in the synthesis of a protein is the production of an intermediate molecule that carries the coded message of DNA from the nucleus into the cytoplasm where the protein can be produced. This intermediate molecule is called messenger RNA (mRNA). RNA (or ribonucleic acid) has similarities and differences with DNA and these are shown in Table 2.5.1.

The building blocks for RNA are the RNA nucleotides that are found in the nucleus. Complementary base pairing of RNA to DNA occurs in exactly the same way as in the replication process but this time uracil (U) pairs with adenine since the base thymine (T) is not found in RNA. Transcription results in the copying of one section of the DNA molecule, not its entire length. Figure 3.2.1 describes the process.

1 DNA is unzipped by the enzyme RNA polymerase and the two strands uncoil and separate.

2 Free nucleotides move into place along one of the two strands.

3 The same enzyme, RNA polymerase, assembles the free nucleotides in the correct places using complementary base pairing. As the RNA nucleotides are linked together, a single strand of mRNA is formed. This molecule is much

shorter than the DNA molecule because it is a copy of just one section, a gene. The mRNA separates from the DNA and the DNA double helix is zipped up again by RNA polymerase.

Once an mRNA molecule has been transcribed, it moves via the pores in the nuclear envelope to the cytoplasm where the process of translation can take place. In prokaryotes, translation occurs immediately after transcription because there is no nuclear envelope.

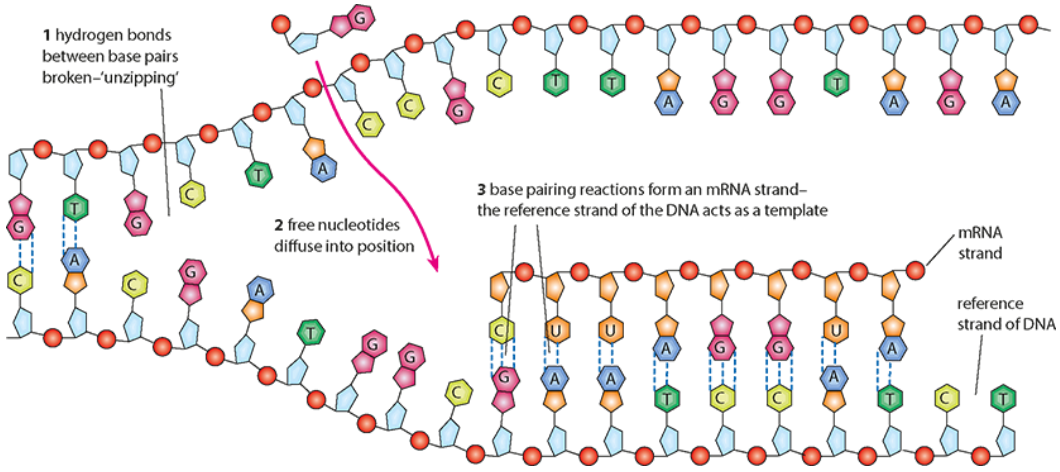

Figure 3.2.1: Transcription.

Initiating transcription

Before transcription begins, RNA polymerase must attach to a promoter region of DNA. This process is different in prokaryotes and eukaryotes. In prokaryotes the RNA polymerase binds directly to a promoter region, close to the region that will be transcribed. Transcription is controlled by another section of DNA called the regulator gene, which can produce a repressor molecule that binds to the operator region and prevents transcription (Figure 3.2.2).

Eukaryotes have several transcription factors that are needed to bind the RNA polymerase to the promoter region. Some of these bind to an enhancer region, away from the promoter. The transcription factors (labelled the transcription-initiation complex in Figure 3.2.3) bring the enhancer region close to the promoter and RNA polymerase can then bind and begin transcription (Figure 3.2.3).

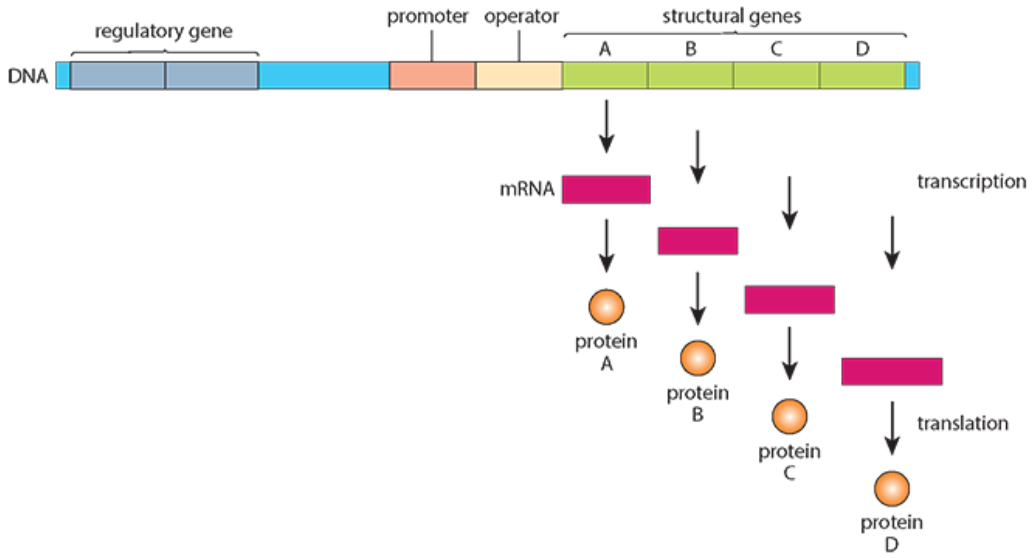

Figure 3.2.2: Controlling transcription in a prokaryote.

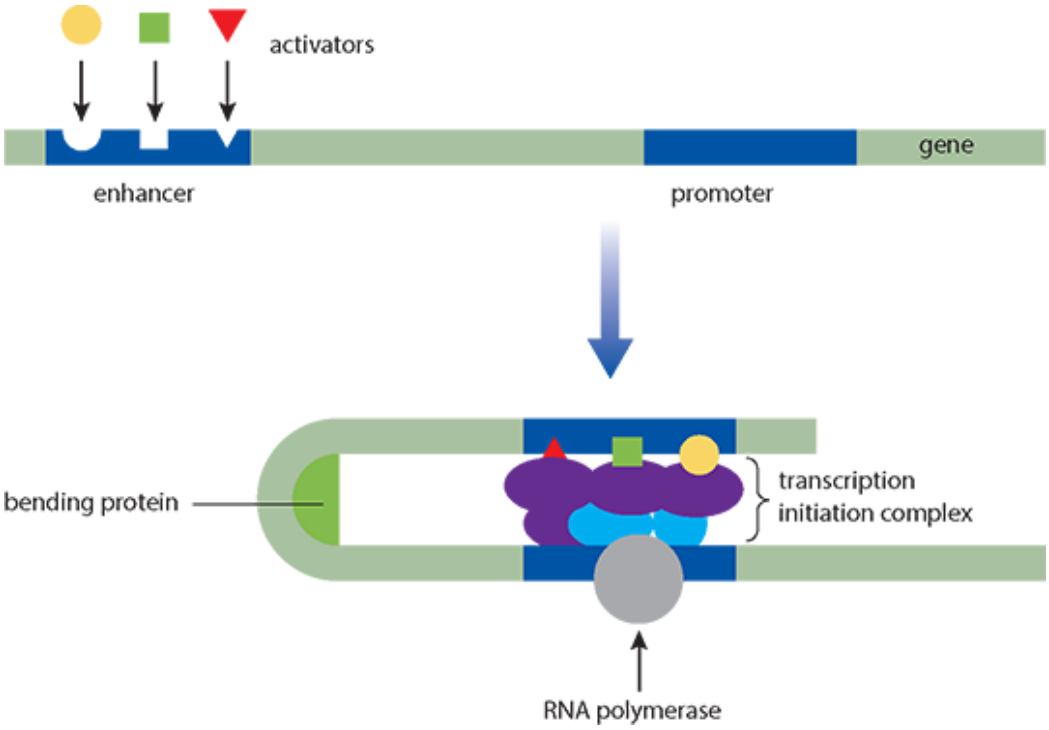

Figure 3.2.3: In eukaryotes transcription is controlled by proteins that bind to specific sequences of DNA. Bending proteins are discussed in the Higher Level Section 3.2.3.

Both prokaryotes and eukaryotes have regions of DNA that do not code for protein. In prokaryotes these include promoter regions that serve regulatory functions.

Regulation of transcription by nucleosomes

DNA in eukaryotes is incorporated into nucleosomes so that the genetic material can be stored in a compact form (Section 1.7.4). Nucleosomes are important because they can either inhibit or allow transcription by controlling whether the necessary molecules can bind to DNA.

KEY POINT

nucleosome a part of a eukaryotic chromosome important in regulating transcription, made up of DNA wrapped around histone molecules and held in place by another histone protein.

In order to transcribe genes, activators and enzymes involved in transcription must be able to gain access to DNA. In all eukaryotic species, the regions of DNA that contain the promoters and regulators, which are the binding sites for RNA polymerase and the starting point for transcription, have fewer nucleosomes than other areas, allowing greater access for binding proteins. Conversely, the regions that are transcribed have a higher density of nucleosomes. This suggests that nucleosomes have an important role in determining which genes are transcribed. This in turn can influence other factors such as cell variation and development.

DNA does not need to be completely released from a nucleosome to be transcribed and, although nucleosomes are very stable protein–DNA complexes, they are not static. They can undergo different structural rearrangements including so-called ‘nucleosome sliding’ and DNA site exposure: if a nucleosome is ‘unwrapped’ there is a significant period of time during which DNA is accessible. They can also be modified by methylation (Section 3.4.2). The new transcript of mRNA before modification is called pre-mRNA and it becomes known as mature mRNA after removal of introns’

3.2.2 Translation

Translation is the process by which the information carried by mRNA is decoded and used to build the sequence of amino acids that eventually forms a protein molecule. During translation, amino acids are joined together in the order dictated by the sequence of codons on the mRNA to form a polypeptide. This polypeptide eventually becomes the protein coded for by the original gene.

Complementary base pairing ensures that the sequence of bases along the mRNA molecule corresponds to the sequence on the original DNA molecule. Each sequence of three mRNA bases is called a **codon** and codes for one specific amino acid, so the order of these codons determines how amino acids will be assembled into polypeptide chains in the cytoplasm. The completed polypeptide chains will be folded to make functioning proteins. Translation is carried out in the cytoplasm by structures called ribosomes and molecules of another type of RNA known as transfer RNA or tRNA.

The mRNA codons that code for each amino acid are shown in Table 3.2.1. From the table you should be able to deduce which amino acid corresponds to any codon.

The genetic code is said to be a **degenerate code** because there are many codons that specify the same amino acid. It is also said to be **universal** because all living things use the same triplet code to specify the same amino acids.

Mutations are changes in sequence of bases that affect protein structure, you can read more about mutations in Section 3.3.

Transfer RNA

The process of translation requires a type of nucleic acid known as transfer RNA (tRNA). tRNA is made of a single strand of nucleotides that is folded and held in place by base pairing and hydrogen bonds (Figure 3.2.4). There are many different tRNA molecules but they all have a characteristic ‘clover leaf’ appearance with some small differences between them.

KEY POINTS

activating enzyme is an enzyme that catalyses the attachment of an amino acid to the appropriate tRNA.

anticodon is a triplet of bases in tRNA that pair with a complementary triplet (codon) in mRNA.

		Second base					

		U		C		A		G			
First base	U	UUU	phenylalanine	UCU	serine	UAU	tyrosine	UGU	cysteine	U	Third base
		UUC		UCC		UAC		UGC		C	
		UUA	leucine	UCA		UAA	‘stop’	UGA	‘stop’	A	
		UUG		UCG		UAG		UGG	tryptophan	G	
	C	CUU	leucine	CCU	proline	CAU	histidine	CGU	arginine	U	
		CUC		CCC		CAC		CGC		C	
		CUA		CCA		CAA	glutamine	CGA		A	
		CUG		CCG		CAG		CGG		G	
	A	AUU	isoleucine	ACU	threonine	AAU	asparagine	AGU	serine	U	
		AUC		ACC		AAC		AGC		C	
		AUA		ACA		AAA	lysine	AGA	arginine	A	
		AUG	methionine or ‘start’	ACG		AAG		AGG		G	
	G	GUU	valine	GCU	alanine	GAU	aspartic acid	GGU	glycine	U	
		GUC		GCC		GAC		GGC		C	
		GUA		GCA		GAA	glutamic acid	GGA		A	
		GUG		GCG		GAG		GGG		G	

Table 3.2.1: Amino acids and their associated mRNA codons.

At one position on the molecule is a triplet of bases called the anticodon, which pairs by complementary base pairing with a codon on the mRNA strand. At the 3′ end of the tRNA molecule is a base sequence CCA, which is the attachment site for an amino acid.

An amino acid is attached to the specific tRNA molecule that has its corresponding anticodon, by an activating enzyme. As there are 20 different amino acids, there are also 20 different activating enzymes in the cytoplasm.

KEY POINT

transfer RNA (tRNA) short lengths of RNA that carry specific amino acids to ribosomes during protein synthesis.

Ribosomes

Ribosomes are the site of protein synthesis. Some ribosomes occur free in the cytoplasm and synthesise proteins that will be used within the cell. Others are bound to the endoplasmic reticulum, forming rough endoplasmic reticulum (RER) (Section 5.2), and synthesise proteins that will be secreted from the cell or used within lysosomes.

Ribosomes are composed of two subunits, one large and one small. The subunits are built of protein and ribosomal RNA (rRNA). On the surface of the ribosome are three tRNA-binding sites (the entry site A, the P site and the exit site E), and one mRNA-binding site (Figure 3.2.5). Two tRNA molecules carrying amino acids can bind to a ribosome at one time. Polypeptide chains are built up in the groove between the two subunits.

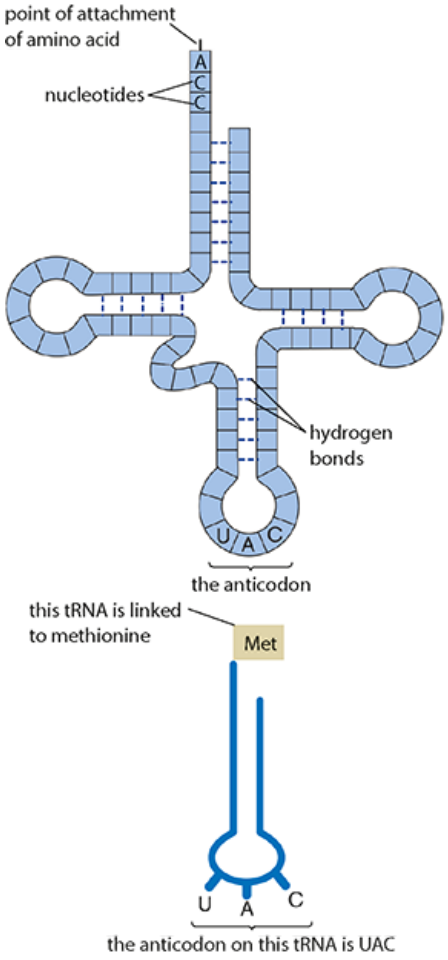

Figure 3.2.4: Transfer RNA (tRNA) has a 'clover leaf' shape.

Building a polypeptide

Ribosomes have binding sites for both the mRNA molecule and tRNA molecules. The ribosome binds to the mRNA and then draws in specific tRNA molecules with

anticodons that match the mRNA codons. Only two tRNA molecules bind to the ribosome at once. Each one carries with it the amino acid specified by its anticodon. The anticodon of the tRNA binds to the complementary codon of the mRNA molecule with hydrogen bonds.

When two tRNA molecules are in place on the ribosome, a peptide bond forms between the two amino acids they carry to form a dipeptide. A peptide bond links the amino group of one amino acid to the carboxyl group of the next.

Once a dipeptide has been formed, the first tRNA molecule detaches from both the amino acid and the ribosome. The ribosome moves along the mRNA one triplet to the next codon.

These processes, shown in Figure 3.2.5, are repeated over and over again until the complete polypeptide is formed. The final codon that is reached is a 'stop' codon, which does not code for an amino acid but tells the ribosome to detach from the mRNA. As it does so, the polypeptide floats freely in the cytoplasm or into the RER.

EXTENSION

A single chromosome contains DNA that codes for many proteins. Most genes are about 1000 nucleotides long, a few are longer and a very small number are less than 100 nucleotides. The size of a gene corresponds to the size of the polypeptide for which it codes.

TEST YOUR UNDERSTANDING

8 Define transcription.

9 Where does transcription take place in a prokaryotic cell and a eukaryotic cell?

10 Outline the structure of a ribosome.

11 Where in the cell are proteins for secretion or use in lysosomes produced?

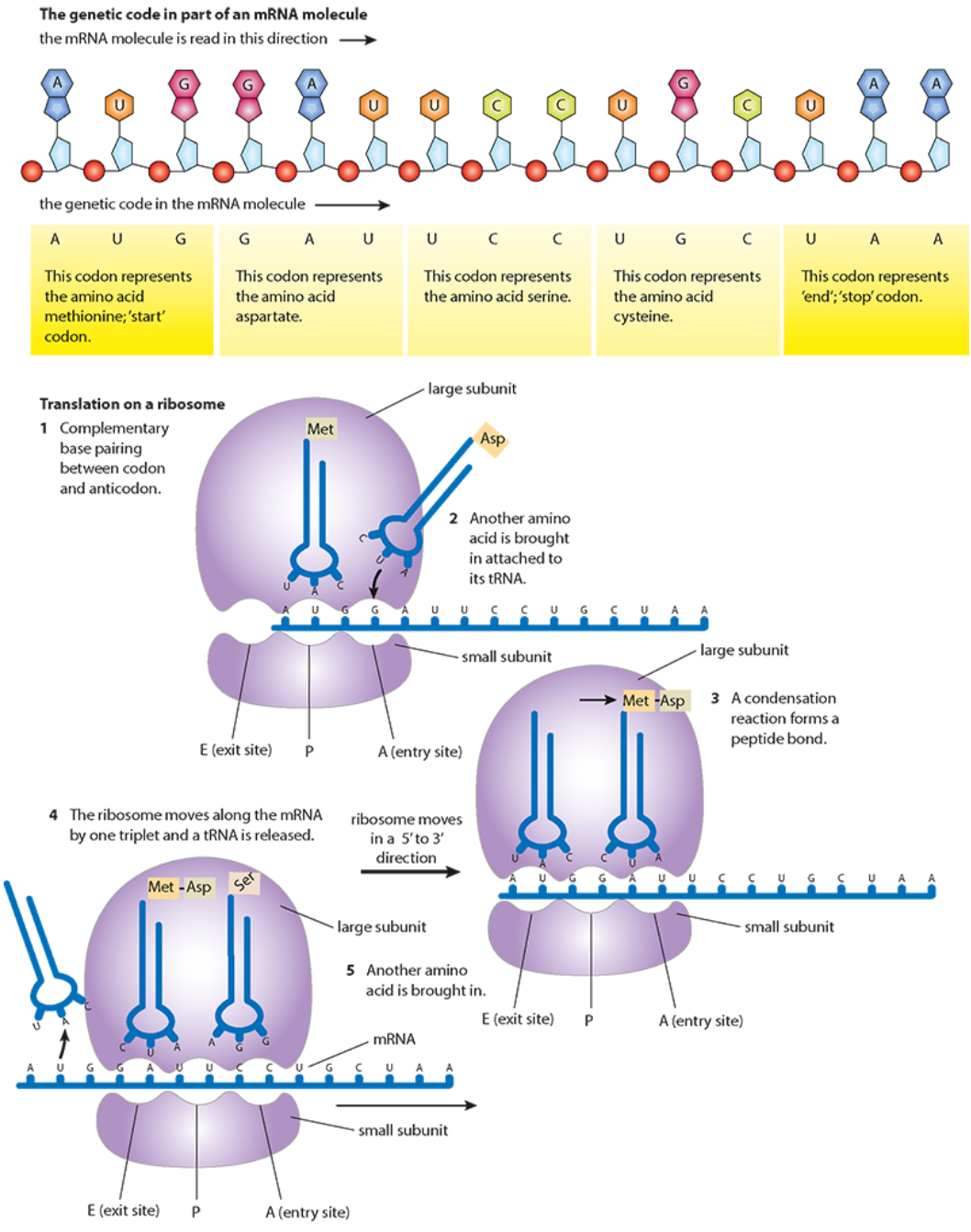

Figure 3.2.5: The stages of translation.

3.2.3 Non-coding regions of DNA

Repeated sequences

DNA molecules are very long but every strand has regions that do not code for proteins. For many years these regions were poorly understood but many of them have now been found to have important functions in the regulation of gene expression and other cell activities. Eukaryotic genomes contain mostly non-coding DNA; in fact, nearly 99% of the human genome is non-coding. About 7% of this DNA is thought to have regulatory functions but the exact proportion is not fully known. You may hear this type of DNA being referred to as 'junk DNA', but it is not a term that is now accepted by the scientific community because at least some of this DNA has a function.

When the highly repetitive sequences of non-coding DNA, called variable number tandem repeats (VNTRs), were first discovered they appeared to have no function but as genomes were mapped and compared it was found that several long repeated sequences in humans, mouse and rat DNA were common to all three species. These repeated non-coding sequences regulate and control the activity of genes and possibly embryo development. Studies of the genomes of many species have shown that non-coding DNA is conserved over hundreds of millions of years, suggesting that these regions have been conserved through evolution. It is likely that they give advantages in preserving certain vital genetic characteristics.

KEY POINT

tandem repeat is a repeated sequence of DNA base pairs where multiple repeats lie side by side on a chromosome.

Tandem repeats are generally associated with non-coding DNA. The number of times the DNA sequence is repeated is variable. Variable tandem repeats are used in DNA profiling because they are very similar in close relatives but very different in unrelated individuals.

Some regions of non-coding DNA act as ‘switches’ that determine when and where genes are expressed by controlling where and when transcription can begin. Others may be essential for chromosome structure and play a role in cell division. Introns are also transcribed but not translated into proteins.

Genes for transfer RNA

In humans, the genes that code for tRNA molecules are found on all chromosomes except 22 and Y. These genes do not code for protein but code either for cytoplasmic tRNA or for mitochondrial tRNA. The number of genes that code for tRNA is related to evolutionary history, so that organisms in the Archaea and Eubacteria domains have fewer than those in the domain Eukarya. This seems to be due to the duplication of the genes over time.

Promoter regions

Promoter regions are DNA sequences that define where transcription of a gene by RNA polymerase begins. They are usually found at the 5′ end of the area where transcription begins.

RNA polymerase requires the presence of a class of proteins known as ‘general transcription factors’ before transcription can begin. Interactions between the transcription factors, RNA polymerase and the promoter region allow the polymerase to

move along the gene so that transcription can occur. Many different transcription factors have been found and each one is able to recognise and bind to a specific nucleotide sequence in DNA. A specific combination of transcription factors is necessary to activate a particular gene.

Other DNA sequences, known as enhancer sequences, are also important and provide a place for regulatory proteins, called activators, to bind.

The role of binding proteins in gene expression is shown in Figure 3.2.3 and discussed in Section 3.4. The proteins bind to the enhancer, which may be some distance from the gene. 'Bending proteins' may then assist in bending the DNA so that the enhancer region is brought close to the promoter. Activators, transcription factors and other proteins attach, so that an 'initiation complex' is formed and transcription can begin. Some activator proteins affect the transcription of multiple genes.

Transcription factors are regulated by signals produced from other molecules such as hormones that are able to activate transcription factors and thus control transcription. Many other molecules in the environment of a cell or an organism can also have an impact on gene expression and protein production.

Telomeres

Telomeres are regions of repeated nucleotide sequences of non-coding DNA at each end of every eukaryotic chromosome. They protect chromosomes from damage and from fusing with adjacent chromosomes. They have been likened to the protection that a plastic tip on the end of a shoelace gives to protect a lace from fraying.

When chromosomes are replicated during cell division, the enzymes cannot copy the sequences at the end of the chromosomes. Telomeres act to protect important genes from being lost, by capping the end sequences. Cells contain enzymes called telomerases, which can replenish the repeated sequences after cell division in stem cells, but telomeres tend to shorten over time as cells replicate. Telomere shortening can block cell division and, by limiting the number of cell divisions, they protect cells from losing genetic information. The average cell will divide between 50 and 70 times before chromosomes are shortened too much and the cell dies.

EXTENSION

You may like to read about the work of Leonard Hayflick (b. 1928), a United States scientist who worked on cloning cells for vaccine production. He discovered that the number of times a cell can divide is not infinite, it is limited to a number called the Hayflick Limit.

Polysomes

Translation occurs at many places along an mRNA molecule at the same time. The electron micrograph in Figure 3.2.6 shows transcription and translation occurring simultaneously in a bacterium. A polysome is a group of ribosomes along one mRNA strand (Figure 3.2.7). Part of the bacterial chromosome can be seen as the fine pink line running horizontally along the top of the micrograph and two growing polypeptide chains are shown forming below it. DNA is being transcribed by RNA polymerase and the newly formed mRNA is being immediately translated by the ribosomes. In eukaryotes, the two processes

occur in the nucleus and cytoplasm, respectively, and so are separated not only in time but also in location.

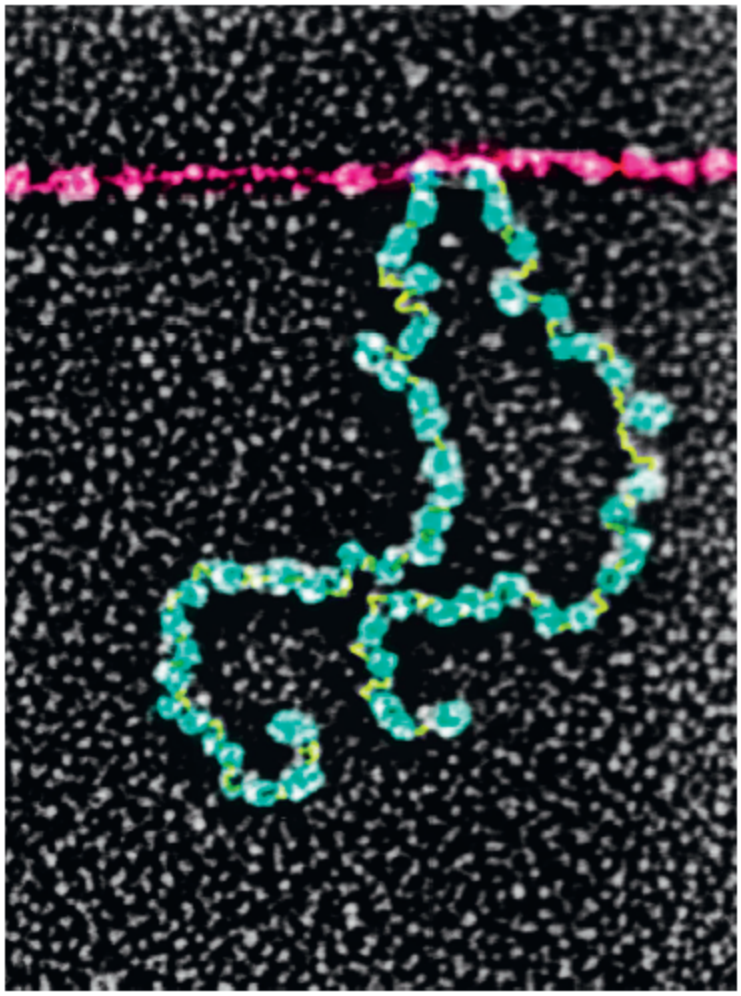

Figure 3.2.6: Electron micrograph of polysomes in a bacterium (× 150 000).

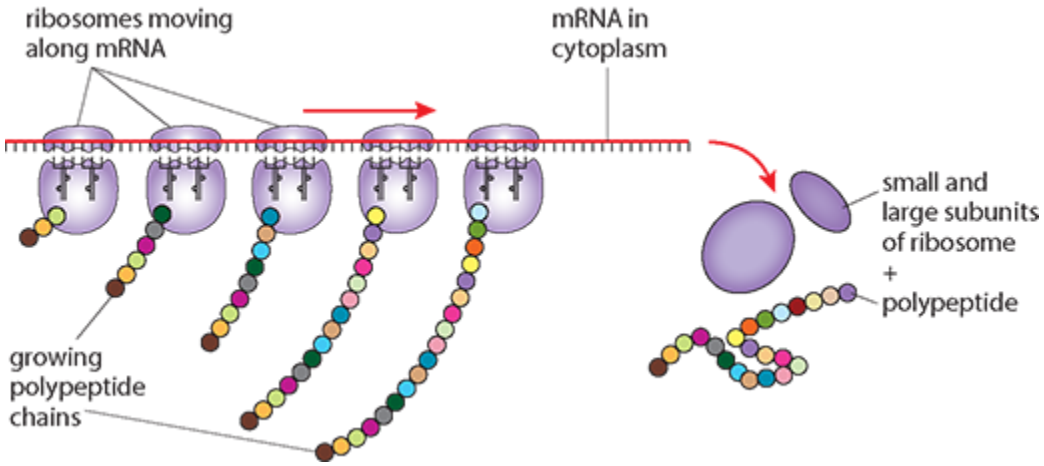

Figure 3.2.7: Diagram of a polysome in a eukaryotic cell. Polysomes appear like beads moving along a string of mRNA.

3.2.4 Post-transcriptional modification

1 Introns and exons

In eukaryotes many genes contain sequences of DNA that are transcribed but not translated. These sequences, known as introns, appear in mRNA but are removed before it is translated. Introns (or intragenic regions) are sequences of nucleotides *within* genes. After transcription of a gene, the introns are removed in a process known as post-transcriptional modification. Introns are removed in the nucleus before the mRNA moves to the cytoplasm for translation. Once introns have gone, the remaining sequences of bases, known as exons, are spliced together to form mature mRNA that will then be translated (Figure 3.2.8). Mature mRNA leaves the nucleus via the nuclear pores and moves to the cytoplasm.

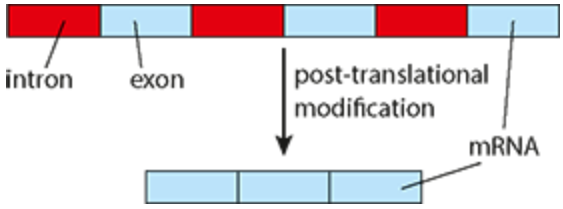

Figure 3.2.8: Introns and exons in mRNA.

Introns occur in many genes in all eukaryotic organisms and the number of introns per gene varies considerably between species. For example, introns are common in humans and mice, where genes almost always contain introns but are rare in other eukaryotes such as the yeast (*Saccharomyces cerevisiae*).

The human genome has been found to have about eight introns per gene but other organisms such as fungi may have fewer than 20 in their entire genome.

It also seems that there are more introns in species with smaller populations, and evolutionary and biological factors are thought to influence this.

Exons may be spliced together in different ways so that a number of different, but similar, protein sequences can be produced from a single gene. This alternative splicing means that a single gene can code for more than one protein. Different exons may be included or excluded from the final mRNA produced and as a result, the proteins produced from alternatively spliced mRNAs will contain different amino acid sequences and can have different biological functions. Mature mRNAs containing various combinations of exons from one original precursor mRNA increases the diversity of the proteins that can be produced. One example of this is in the production of immunoglobulins by B cells. A cell can splice together different exons and produce different immunoglobulins (antibodies) in response to different antigens that are present in the body (Chapter 10).

Splicing is controlled by molecules that respond to signals from both inside and outside the cell. Alternative splicing allows the human genome to synthesise many more proteins from its 20 000 genes than it could if there was no splicing.

2 5’ Caps and 3’ polytails

Two other modifications made to mRNA before translation are:

1 Adding a 5’ cap – a 5’ cap (a modified G nucleotide) is attached to the 5’ end of the mRNA

2 Adding a 3’ poly-A tail – poly-A tail is attached to the 3’ end of the mRNA and consists of a long string of A nucleotides

KEY POINTS

polysome an arrangement of many ribosomes along a molecule of mRNA so that multiple copies of the same polypeptide are produced at the same time.

alternative splicing including different exons in processed mRNA so that a cell can produce different proteins from the same gene.

exons sequences of bases in mRNA that are spliced together and translated after introns have been removed.

intron sequences of bases in mRNA that are removed after transcription.

Both these modifications help new mRNA strands leave the nucleus. They also help to protect the mRNA from damage. In the cytoplasm, the addition of the 5' cap and poly-A tail, help the ribosomes attach to the 5' end of the mRNA.

EXAM TIP

Remember: introns intervene in genes, but only exons are expressed.

3.2.5 Post-translational modification – producing functional proteins

Translation produces polypeptides which are the starting point for forming the working proteins that we need. But proteins are large, complex molecules, usually made up of hundreds of amino acid subunits linked in polypeptides. But it is the folding and linking of polypeptide chains into secondary, tertiary and, in some cases, quaternary structures that leads to the formation of functional proteins (Section 1.6). Folding takes place in the cytoplasm.

Another modification made to polypeptides and proteins after translation is the addition of a prosthetic group. Prosthetic groups are not polypeptides but they bind to different proteins or parts of them to enable them to function. An example is the respiratory pigment hemoglobin, which contains four polypeptide chains, each one containing a prosthetic heme group (Section 1.6).

Prosthetic groups

Many proteins contain **prosthetic groups** and those that do are called **conjugated proteins**. Prosthetic groups are non-protein groups that are able to bind to different proteins or parts of them. We can see two examples of prosthetic groups in the respiratory pigments myoglobin (Figure 1.6.4) and hemoglobin which both contain a prosthetic heme group. Hemoglobin consists of four polypeptide chains, each one containing a heme group. The heme group (Figure 1.6.1) consists of a central Fe (iron) atom and a porphyrin ring. The prosthetic heme group is vital to the structure of hemoglobin because the shape of the whole protein is changed as oxygen binds to it. The iron group not only allows

oxygen to bind but also holds the compact structure with four subunits in place and allows for progressively easier oxygenation as more oxygen molecules bind to the protein.

Protein modification and processing

Almost all proteins are chemically altered after they have been made. Modifications change the activity, life span and location of the protein. There are two main types of alteration that take place:

1. Chemical modification involves making additions to the side groups of amino acids or to the ends of the protein.
2. Processing involves removing peptide segments before the active protein is formed.

The most common chemical modifications are the addition of an acetyl group ($-CH_3CO$) to the amino group ($-NH_3$) at the end of a polypeptide. An estimated 80% of all proteins are modified in this way. Phosphorylation is another common modification and about 30% of proteins are modified by the addition of phosphate. The addition of phosphate regulates the activity of certain enzymes. Other modifications, such as the addition of a carbohydrate, can stabilise a protein and make it fully functional.

Processing involves the removal of segments of a protein. Active peptide hormones are produced in this way. One example is the polypeptide preproinsulin which is the inactive precursor to insulin. Preproinsulin is cut twice (Figure 3.2.9) so that the final active hormone insulin consists of two separate chains linked by disulfide bridges.

Preproinsulin is first converted to proinsulin by peptidase enzymes that remove a peptide from the $-NH_3$ end of the

molecule. Next, proinsulin is converted to active insulin by the removal of a second section of peptide at the C- end. Disulfide bridges link the remaining section of the molecule so that it becomes active insulin.

EXTENSION

The proteins present in human cells are far greater than the number that are coded for by our genes and this is due in part to modifications made to the proteins after they are formed. You can find out more about the proteins that are modified using the SWISS-PROT protein sequence database. Recent studies have identified more than 8000 proteins that are phosphorylated, more than 3000 that have acetyl groups added to them and around 5000 that have carbohydrates added to them.

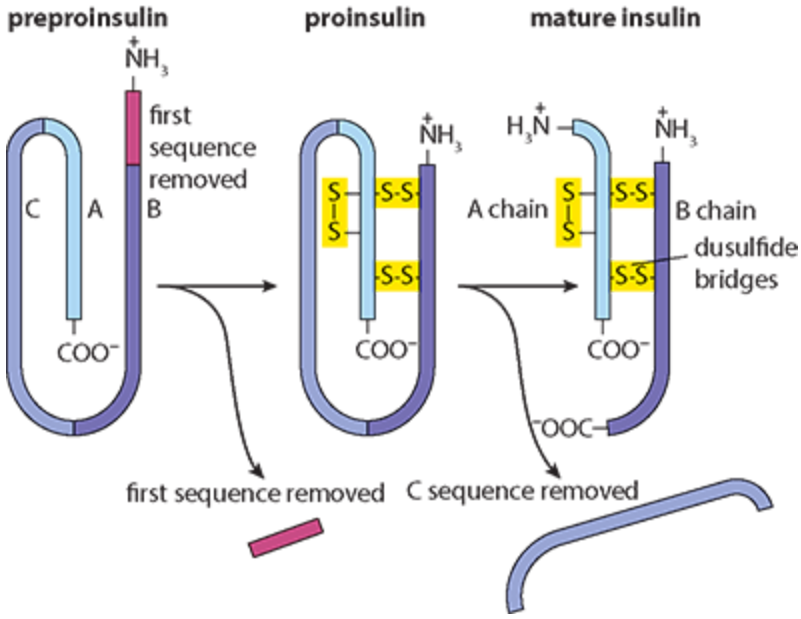

Figure 3.2.9: The formation of active insulin.

3.2.6 Protein transport molecules

Some proteins are able to operate as transport molecules and signalling systems inside cells. The bonds that hold these proteins together can, in some cases, be broken and remade to enable them to do so. For example, when molecular signals are received from outside a cell by receptors on its membrane surface they can be processed and transferred to the nucleus by the modification of transport proteins.

The enzyme protein kinase is important in this process because it can phosphorylate a transport protein. Phosphorylation changes the function of the protein by changing its activity, its location or the way it interacts with other proteins. A protein can be phosphorylated at the cell surface by protein kinase and dephosphorylated later in the cytoplasm. These changes in the protein's structure activate and then deactivate it while it acts as a signalling mechanism because phosphorylated proteins can be moved along many different pathways in the cell.

Recycling amino acids

Proteins synthesis requires a constant supply of amino acids. Some are provided from our diet but many come from recycled amino acids that have formed part of proteins in the body. Proteasomes are protein complexes which degrade unneeded or damaged proteins by proteolysis, a chemical reaction that breaks peptide bonds. Proteasomes are present in the cytoplasm and in the nuclei of all eukaryotic cells. The amino acids that are released can be used to build new polypeptides and proteins.

KEY POINT

protein kinase an enzyme that regulates the biological activity of proteins by phosphorylating specific amino acids using ATP.

NATURE OF SCIENCE

Looking for trends and discrepancies: do all organisms use only 20 common amino acids in their proteins?

Humans can make 10 of the 20 amino acids we need to build proteins but we do not have the enzymes needed for the biosynthesis of the others. Plants, on the other hand, must be able to make all the amino acids they require.

Researchers have also investigated the trends in amino acid compositions of proteins found in species of the important kingdoms of Archaea, Bacteria and Eukaryotes. International databases ProteomicsDB and SWISS-PROT (which contain information about the structure and composition of proteins) can compare amino acid frequencies for 195 known proteomes and all recorded sequences of proteins. They discovered that the amino acid compositions of proteins do differ substantially for different kingdoms.

In addition to the variations in amino acids in proteins, some microorganisms and plants are able to make so called ‘non-standard’ amino acids by modifying standard amino acids. Some species are also able to synthesise many uncommon amino acids. For example, some microbes synthesise lanthionine, which is a modified version of the amino acid alanine. Many other proteins are modified after they have been produced. This ‘post-translational modification’ involves the addition of extra side groups to the amino acids in a protein.

Considering all the evidence, it seems that, although we can observe many similar proteins in different species, we cannot always say that the same amino acids are used in their construction. The range of amino acids in proteins can vary considerably from species to species.

To consider:

1 What contribution have international databases made to our understanding of protein structure?

2 How can comparing proteomes and amino acids in different organisms help our understanding of evolutionary relationships?

TEST YOUR UNDERSTANDING

12 Explain the difference between introns and exons.

13 Name three types of non-coding DNA sequence.

14 Where are telomeres located and what is their role?

15 Why are polysomes important in cells?

REFLECTION

Reflect on the areas of this topic that you found particularly interesting. What was it about them that caught your attention?

Links

- How does the variety of proteins produced contribute to the functioning of a cell? (Chapter 6)
- How does the degenerate genetic code protect a cell against mutations? (Chapter 3.3)

3.3 Mutations

LEARNING OBJECTIVES

In this section you will:

- learn that mutations are structural changes to genes at the molecular level
- learn how new alleles form by mutation; changes may be neutral, harmful or beneficial
- understand that mutations in germ cells can be passed to offspring
- define a mutagen as a substance that can cause genetic change
- recognise that mutations can add, delete or substitute base or bases in genes
- recall examples of insertion and deletion mutations
- learn that the genetic code is degenerate and so it is resistant to some changes caused by mutations

- understand that DNA polymerases can make proofreading errors, and the errors remain permanently
- recognise that mutations do not always cause changes to a protein's function
- understand the technique of 'gene knockout' and its use in investigating gene function

- learn how CRISPR sequences are used in gene editing
- recognise the importance of highly conserved sequences in genes

GUIDING QUESTIONS

- Why are variation and mutation essential for evolution? (Chapter 11)
- Why must the cell cycle be regulated and controlled? (Chapter 6)

3.3.1 Chromosomes, genes and mutations

A DNA molecule comprises a pair of strands, each strand consisting of a linear sequence of nucleotides, with weak hydrogen bonds between the bases holding the two strands together. This linear sequence of bases contains the genetic code in the form of triplets of bases. A gene is a particular section of a DNA strand that, when transcribed and translated, forms a specific polypeptide (Figure 3.3.1). Some of the polypeptides will form structural proteins, while others become enzymes or pigments such as hemoglobin and it is the translation of the genes which gives each individual organism its own specific characteristics. (Transcription and translation are described in Section 3.2.) Each gene is found at a specific position on a chromosome and so, for example, it is possible to say that the gene for human insulin is always found on chromosome number 11.

Organisms that reproduce sexually almost always have pairs of chromosomes, with one of each pair coming from each parent. The members of the pair carry equivalent genes, so that – for example – in humans, both versions of chromosome number 11 carry the insulin gene. But there may be slight differences in the version of the gene on each chromosome. These slightly different forms of the gene are known as alleles. Alleles differ from one another by one or only a few bases and it is these differences in alleles that give rise to the variation we observe in living organisms.

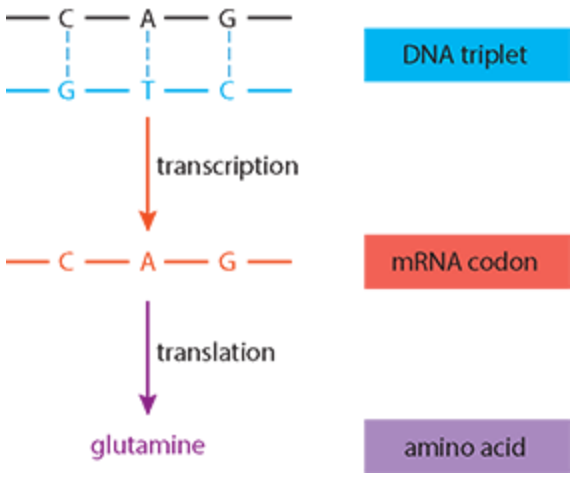

Figure 3.3.1: The base sequence in DNA is decoded via transcription and translation.

What are mutations?

The process of DNA replication is complex and mistakes sometimes occur – a nucleotide may be left out, an extra one may be added or the wrong one inserted. These mistakes are known as gene **mutations**. Mutations may occur spontaneously, as a result of errors in copying DNA, or they can be caused by factors in the environment known as **mutagens**, described in Section 3.3.2. A mutation that occurs in **germ cells** (**gametes**) that will go on to form a new offspring will be passed to every cell in those offspring, including their germ cells. As a result, the offspring may have a genetic condition that is not present in either of their parents. A mutation that occurs in **somatic cells** (body cells) will not be inherited. A mutation involving the change of a single nucleotide is called a **base substitution mutation**. When the DNA containing an incorrect nucleotide is transcribed and translated, errors may occur in the polypeptide that is produced. Errors may be beneficial, neutral or harmful.

Beneficial mutations change DNA and allow the synthesis of new proteins that may work slightly differently. One example of

a new, recently discovered, beneficial mutation has been found in a gene that codes for a receptor protein on the cell surface in the plasma membrane. Only a very few people carry this mutation but the change to their receptor protein gives them total immunity to infection by human immunodeficiency virus (HIV) because the virus cannot bind to their cells.

In 2009, United States researchers located another beneficial mutation in a gene (*SLC30A8*), which affects insulin. They found that subjects who were both overweight and elderly who carried the altered gene had considerable protection from developing type II diabetes. Evolutionary biologists also believe that our ability to discriminate three colours – red, green and blue – is due to a beneficial mutation that occurred in our primate ancestors' DNA millions of years ago.

However, beneficial mutations can be closely associated with harmful ones. Sickle-cell anemia, also called sickle-cell disease (SCD), is caused by a mutation that causes red blood cells to develop a crescent, or sickle shape; this abnormal shape can lead to a number of health problems but also has some advantages, which are described in the section on sickle-cell anemia.

How do mutations occur?

A mutation can involve the addition, deletion or substitution of a base in DNA or the inversion of a section of DNA so that it is turned backwards in the sequence.

Table 3.2.1 shows the amino acids that are specified by different mRNA codons. Most amino acids are coded for by more than one codon and so many substitution mutations have no effect on the final polypeptide that is produced. These are said to be neutral (or silent) mutations. For example, a mutation in the DNA triplet CCA into CCG would change the codon in the

mRNA from GGU to GGC but it would still result in the amino acid glycine being placed in a polypeptide. Other examples of neutral mutations are those that affect non-coding regions of the chromosome, or which result in changes to features such as blood type or eye colour in humans that do not adversely affect a person.

Some substitution mutations, however, do have serious effects. For example, one important human condition that results from a single base substitution is sickle-cell anemia.

Degeneracy in the genetic code

As Table 3.2.1 shows, there are 64 combinations of three-letter nucleotide sequences that can be made from the four nucleotides. Of these, 61 represent amino acids and three are stop signals. Although each codon is specific for only one amino acid or one stop signal, the genetic code is described as degenerate because a single amino acid may be coded for by more than one codon.

Degeneracy in the genetic code makes the code more resistant to changes. It means that single base substitutions can occur without disrupting protein synthesis or functioning of the organism.

EXAM TIP

You may be asked to explain inversions, deletions, substitutions or additions to a DNA sequence. When you do this, first remember to check whether you need to convert the DNA sequences to mRNA sequences before identifying the amino acids. Most tables used to decode the genetic code are shown as tables of mRNA codons.

Sickle-cell anemia: the result of a base substitution mutation

Sickle-cell anemia is a blood disorder in which red blood cells become sickle shaped and cannot carry oxygen properly (Figure 3.3.2). It occurs most frequently in people with African ancestry, about 1% suffer from the condition and between 10 and 40% are carriers of it. Sickle-cell anemia is due to a single base substitution mutation in one of the genes that make hemoglobin, the oxygen-carrying pigment in red blood cells.

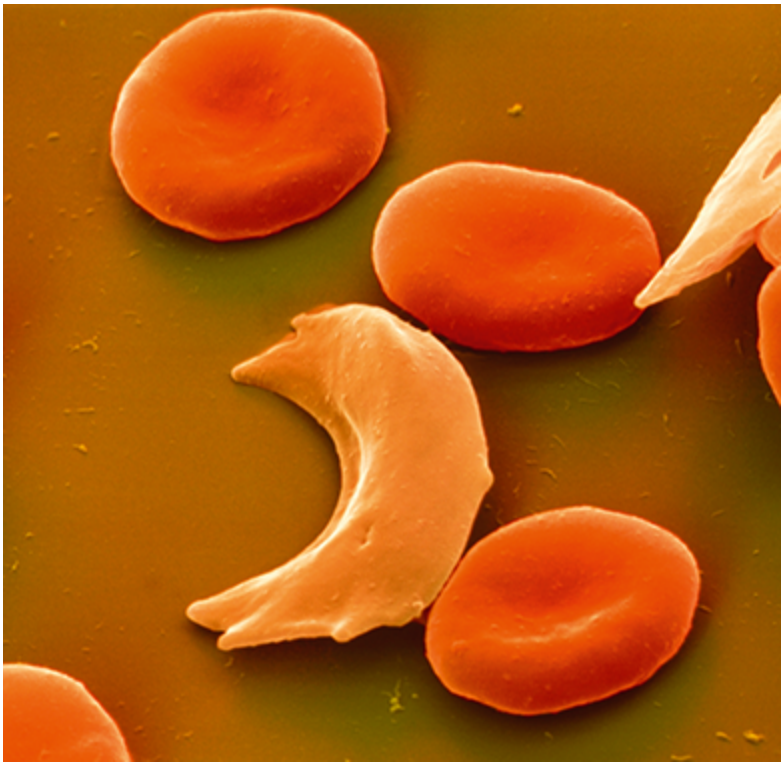

Figure 3.3.2: Coloured scanning electron micrograph showing a sickle-cell and normal red blood cells (× 7400).

Hemoglobin is made up of four subunits, as shown in Figure 3.3.3 – two a-chains and two b-chains. The b-chains are affected by the sickle-cell mutation. To form a normal b-chain, the particular triplet base pairing in the DNA is:

The C–T–C on the coding strand of the DNA (in blue here) is transcribed into the mRNA triplet G–A–G, which in turn is translated to give glutamic acid in the polypeptide chain of the b-chain.

If the sickle-cell mutation occurs, the adenine base A is substituted for thymine base T on the DNA coding strand, so the triplet base pairing becomes:

WORKED EXAMPLE 3.3.1

A mutation can involve the addition, deletion or substitution of a base in DNA. Consider the effect of these changes on this short length of DNA:

CTG GGG GGT GTG AAC

The sequence of amino acids produced by this sequence should be

Leu Gly Gly Val Asn

If the base highlighted in red above is *deleted* the consequence would be:

CTG GGG GGT TGA AC

This would result in the amino acids

Leu Gly Gly STOP

because TGA is a stop codon. The polypeptide produced will be shorter than it should have been.

Question

What type of mutation has occurred in these examples and what are the consequences?

a CTG GGG GGT AGT GAA C

Answer

In this case a base has been *added* to the sequence after the third codon. This will result in serine, coded for by AGT, being added as the fourth amino acid instead of valine. All the subsequent amino acids in the sequence will also be incorrect. This is known as a frameshift mutation because the 'reading' of the DNA in sets of three bases is changed completely. New amino acids will be inserted and produce a different translation of the code from the inserted bases onwards. Deletion of a base can also cause a frameshift mutation.

b CTG GGG GGT GTG CAA

Answer

Here the final triplet has been *inverted* (turned around, so it is backwards) and is CAA instead of AAC. In this case glutamine will be inserted into the amino acid chain instead of asparagine.

c CTG GGG GGG GTG AAC

Answer

In this example a *substitution* of a base has occurred. G has replaced T. In this case there will be no effect on the amino acid sequence that is assembled because GGT and GGG both code for the amino acid glycine. This is an example of a neutral (or silent) mutation.

This type of mutation is known as a **point mutation**. C–A–C on the coding strand of the DNA is now transcribed into the mRNA triplet G–U–G, which in turn is translated to give the amino acid valine. Valine replaces glutamic acid in the b-chain.

Valine has different properties from glutamic acid and so this single change in the amino acid sequence has very serious effects. The resulting hemoglobin molecule is a different shape and it is less soluble and, when in low oxygen concentrations, it deforms the red blood cells to give them a sickle shape. Sickle cells carry less oxygen, which results in anemia. They are also rapidly removed from the circulation, leading to a lack of red blood cells and other symptoms such as jaundice, kidney problems and enlargement of the spleen.

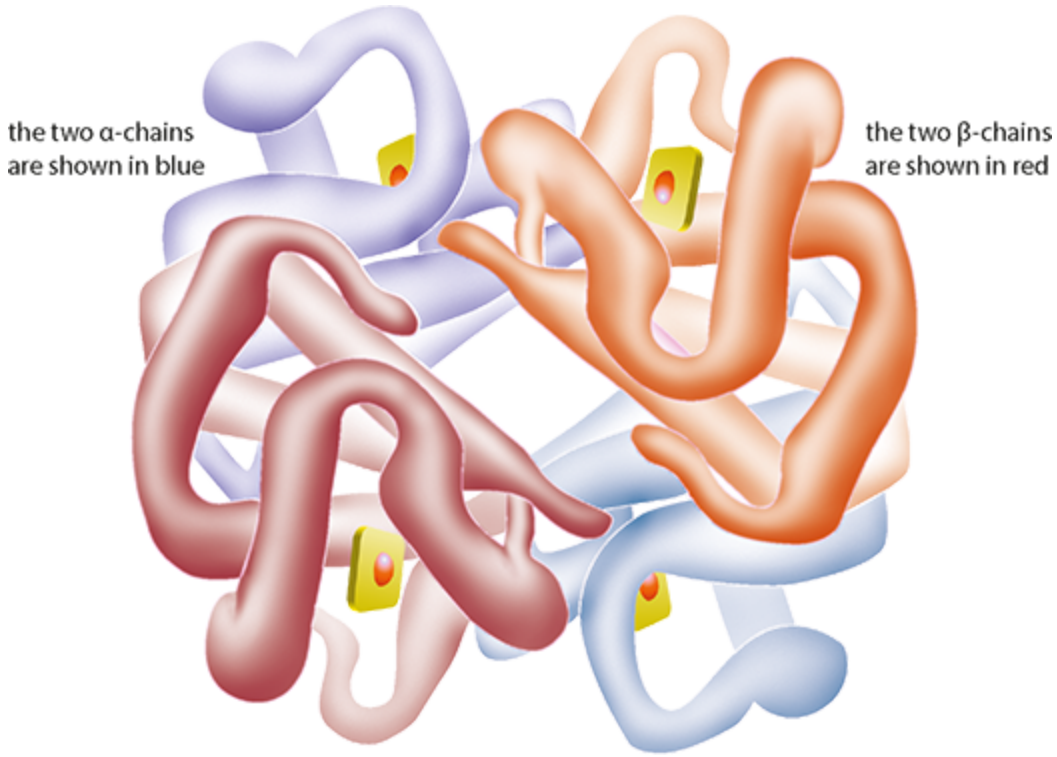

Figure 3.3.3: The structure of a hemoglobin molecule showing the 3D arrangement of the subunits that make it up.

People who have one sickle-cell allele and one normal allele are said to have sickle-cell trait and have some resistance to malaria. This benefit explains why the mutation persists in areas where malaria is endemic. But, in parts of the world where malaria is not a problem, the mutation no longer provides a survival advantage. Instead, it poses the threat of sickle-cell disease, which occurs in the children of carriers who inherit the sickle-cell gene from both their parents.

Insertions and Deletions of bases

Inserted and deleted bases can have serious consequences for protein production in cells. Sometimes changes in bases can cause serious illness but in other cases they lead to benefit. Two

examples are the mutations in the *HTT* gene and in the CCR5 gene.

Huntington's disease: result of insertions - trinucleotide repeats

The *HTT* gene provides instructions for making a protein called huntingtin which plays an important role in nerve cells in the brain. One region of the *HTT* gene contains a particular DNA segment known as a CAG trinucleotide repeat. This segment is made up of a series of three DNA bases (cytosine, adenine, and guanine) that appear multiple times in a row. Normally, the CAG segment is repeated 10 to 35 times within the gene. But if the number of repeats increases to more than 40 the result will be Huntington's disease. If more than 60 repeats are present a more severe form of the disease develops. The greater the number of repeats, the sooner the disease appears.

Huntington's disease stops parts of the brain working properly causing memory loss, difficulty with movement, mood swings and depression. It develops over a period of time and is usually fatal after about 20 years.

HIV resistance: result of a deletion

The human CCR5 gene located on chromosome 3 codes for a protein called CCR5. The protein is found on surface of lymphocytes and other cells of the immune system. The proteins form part of the receptors that are involved in cell signalling and the coordination of immune responses. The CCR5 receptors provide a point of entry for the HIV-1 virus to infect the cells. Some people have inherited a mutation called Delta 32 and have part of the CCR5 gene deleted. The gene changes alter the structure of the CCR5 receptors and make it difficult for the HIV virus to enter cells. Individuals who have this deletion mutation

live normal lives, but if they inherit copies of the mutation from their both parents, they are naturally immune to HIV.

NATURE OF SCIENCE

Tests for genetic diseases

There are many commercially available tests for potential health and disease risks. For example, the most effective and accurate method of testing for Huntington's disease is called a direct genetic test. DNA from a blood sample is taken and analysed for the number of CAG repeats it contains. The presence of 36 or more repeats is an indication that the person has, or will develop Huntington's disease. In a few cases, the test result is not clear and a definite answer is not possible. Any person who decides to take this test must be prepared to face not only the emotional effect it will have on themselves and their family, but also the affect on other aspects of their life. Life insurance and some job opportunities may become unavailable to them if the test is positive.

If any genetic test is taken it is important that the outcomes are interpreted correctly by an expert who can explain the consequences of the results. A negative result can eliminate the need for check-ups while a positive result can help direct a person to monitoring and treatment options. Some tests can help people decide about whether or not to have children and both genetic and non-genetic tests can provide information about a person's health in the future.

To consider:

1 Why do some people think that genetic test results can cause family discord, psychological distress and stigmatisation?

2 Why might some people decide not to have children as a result of a genetic test?

Gene knockout to investigate gene function

Gene knockout is a method that is used to damage or 'knock out' specific genes so that they no longer function and are not expressed. Gene knockout is used with model organisms such as mice and yeast which have specific genes knocked out so they can be used to study how those genes function and investigate what happens when the genes are lost. Researchers can draw inferences from the difference between the knockout organism and normal individuals with a similar genetic background. Knockout organisms are also used in the development of new drugs which target specific biological processes or genetic deficiencies. A library of the genomes of model organisms is available to researchers. The loss of gene activity often causes the phenotype of the model organisms such as mice to change so that living organisms can be used to study gene function. For example, 'Metheuselah' is a knockout model mouse which lives for far longer than an average animal and 'Frantic' is a model mouse which is used for studying anxiety disorders. The loss of the knocked out genes provides valuable information about what the gene normally does. Mice are useful model organisms because humans share many genes with mice.

CRISPR

CRISPR stands for Clustered Regularly Interspaced Short Palindromic Repeats. Repeated DNA sequences, called CRISPR, were first noticed in bacteria. They have 'spacer' DNA sequences in between the repeats that exactly match sequences found in viruses. These sequences which contain short repetitions of base sequences are involved in the defence

mechanisms of prokaryotic organisms such as bacteria and archaea to viruses that infect them. The sequences have come from DNA fragments of viruses that have previously infected the prokaryotes and the organisms use them to detect and destroy DNA from similar viruses that might infect them.

CRISPR is now used as a genetic engineering tool that uses repeated sequences of DNA to edit genes. Using CRISPR it is possible to find a specific section of DNA inside a cell so that a gene can either be modified or even turned on or off without altering the DNA sequences. The key to CRISPR is the many variations of 'Cas' proteins found in bacteria. These are the proteins produced by prokaryotes which help defend them against viruses. A protein called Cas9 is the most widely used by scientists. This protein can easily be programmed to find and bind to almost any desired target sequence, simply by giving it a piece of RNA to guide it.

When the CRISPR Cas9 protein is added to a cell along with a piece of guide RNA, the Cas9 protein links to the guide RNA and then moves along the strands of DNA until it finds and binds to a 20-DNA-letter long sequence that matches part of the guide RNA sequence.

One successful use of CRISPR has been in the treatment of human papillomavirus (HPV). This very common virus has more than 100 different strains; some of them affect the skin causing warts and are barely noticed, but others contribute to 99% of cervical cancers. Researchers have been able to turn off two genes in the virus and knock out the production of two viral oncoproteins. A constant supply of these viral proteins is needed to transform normal cells into cancer cells. Without the proteins, cells infected with the virus go into senescence which means that they stop dividing. Targeting the genes for the proteins can

potentially treat HPV-related cancers. CRISPR and Cas9 has also been used in the treatment of hepatitis B, in this case the ends of certain repeated sequences in the Hepatitis B viral genome are targeted. It has also been used experimentally to repair the mutations that cause cataracts in mice.

NATURE OF SCIENCE

Since the human genome was first sequenced, genetic research has expanded rapidly. The cost of sequencing the entire genome of one person has dropped from about US$1 billion to $1,000, and the speed of sequencing has become many times faster. Scientists around the world have a new and powerful way of understanding how genetic variation may affect not only organisms that humans make use of, but also human health and disease. CRISPR technology can be used to edit genes and has the potential to change genomes.

The CRISPR method was first used in 2012 and replaced costly methods of gene editing that had been used in some plants and animals previously. CRISPR has made gene editing cheap and easy. The technique is widely used in research and already has the potential to alter plants and animals on our farms. The technology also has the potential to treat and prevent many diseases.

It could even change the genomes of future generations, although many people think this is unethical. CRISPR is already being used to fingerprint cells and observing what happens inside them, and for directing evolution.

The most common use of CRISPR involves using Cas9 protein to cut the DNA at a target area. When the cut is repaired, mutations can be introduced that disable a gene. But

CRISPR can also be used to make precise changes such as replacing faulty genes. At present this is much more difficult. The knowledge gained from studying human gene knockouts gives scientists a tool to identify potential new targets for medical treatments and better understand safety concerns of treatments that are being developed. But the technique could have the potential to make permanent changes to a person's genome. Scientists around the world are subject to different rules in the use of genome technology. For this reason, there is an ongoing effort to make a system of regulation for all scientists working in this rapidly growing field of research.

Questions

1 Why is it important to have clear rules about what should and should not be attempted using technology such as CRISPR?

2 What are the potential benefits and dangers in gene editing and replacement?

Why are some gene sequences conserved in a species?

Conserved sequences are sequences of DNA (or protein) that are identical or very similar across a species or group of species. A highly conserved sequence is one that has remained relatively unchanged for long periods of evolutionary history. Examples of highly conserved sequences include those for the RNA found in ribosomes that are present in all domains of life, and the homeobox sequence which is a DNA sequence of around 180 base pairs that regulates the early stages of embryonic development. This sequence is found in many eukaryotes. Studies of sequence conservation now form part of

investigations in genomics, proteomics and evolutionary biology.

KEY POINT

conserved sequence a base sequence in a DNA molecule (or an amino acid sequence in a protein) that has remained relatively unchanged throughout evolution.

An explanation for the presence of these conserved sequences was put forward in 1965, by Emile Zuckerandl and Linus Pauling who proposed the hypothesis of a molecular clock. Amino acid sequences can be conserved to maintain the structure or function of a protein. Conserved proteins undergo fewer amino acid replacements or are more likely to substitute amino acids with similar biochemical properties. The molecular clock theory went on to suggest that steady rates of amino acid replacement could be used to estimate the time when two organisms diverged in evolution (Section 11.1.3). Many phylogenetic relationships worked out from studies of the fossil record seemed to support this theory but some other genes were found to evolve at different rates. This led to the development of theories of molecular evolution. In 1966 Margaret Dayhoff compared ferrodoxin (small proteins involved in a range of metabolic reactions) sequences in many organisms and proposed that natural selection will act to conserve protein sequences that are essential to life. This hypothesis explains why conserved sequences of DNA are found for many important proteins in many species.

3.3.2 Harmful mutations and mutagens

New cells are needed to replace cells that have died or to allow an organism to grow. The nucleus and cytoplasm of a cell divide in processes known as mitosis and cytokinesis, which are phases in a series of events known as the cell cycle (Section 6.5).

In normal circumstances the cell cycle is strictly controlled with cell division (mitosis) occurring to form new cells to replace damaged or dying cells.

In most cases, mitosis continues until a tissue has grown sufficiently or repairs have been made to damaged areas.

Most normal cells also undergo a programmed form of death known as **apoptosis** as tissues develop. Apoptosis can be caused when a cell experiences stress or if it receives signals that indicate it should die.

But sometimes mitosis does not proceed normally. Cell division may continue unchecked and produce an excess of cells, which clump together. This growth is called a **tumour**. Tumours can be either **benign**, which means they are restricted to that tissue or organ, or **malignant** (cancerous), in which some of the abnormal cells migrate to other tissues or organs and continue to grow further tumours there. If they are allowed to grow without treatment, tumours can cause obstructions in organs or tissues and interfere with their functions.

Mutagens are physical, chemical or biological agents that can cause mutations and modify DNA. Mutagens include ionising radiation – such as X-rays, gamma rays and ultraviolet light – and also chemical compounds, such as those found in tobacco smoke and aflatoxins produced by certain fungi. The DNA

changes caused by mutagens are not all harmful. However, because some of them cause cancer, some mutagens are said to be **carcinogens** (cancer causing). The development of a **primary tumour** can also be caused by mistakes in copying DNA, or a genetic predisposition as a result of inheritance. (You can learn more about the control of cell division and the development of cancer in Section 6.5)

Smoking and cancer

Smoking is a major cause of several types of cancer. There is strong evidence to show that it increases the risk of cancer of the bladder, cervix, kidney, larynx and stomach, and smokers are seven times more likely to die of these cancers than non-smokers. In the UK, approximately 70% of lung cancers in both males and females are related to smoking.

SCIENCE IN CONTEXT

Smoking and lung cancer

All tobacco products contain various amounts of carcinogenic substances. Tobacco smoke contains more than 70 chemicals, including many which are known to initiate or promote cancer. Recently the role of nicotine, the addictive drug in tobacco, has come under scrutiny as more people are turning to e-cigarettes and other non-tobacco sources of nicotine as substitutes for smoking. There is no clear evidence that nicotine is a direct carcinogen, but it seems to act as a promoter and may inhibit anti-tumour immune responses. In experiments, nicotine has been shown to induce breaks in DNA and enhance the growth of existing cancers.

The link between smoking and lung cancer was recognised in the 1940s and 1950s, with evidence from **epidemiology**,

animal experiments, examination of cells and chemical analysis. Cigarette manufacturers disputed the evidence, as part of a campaign to maintain cigarette sales. Their propaganda was successful in the short term and, as late as 1960, only one-third of all US doctors believed that the case against cigarettes had been established.

Today it is understood that the risk of contracting lung cancer increases with the number of cigarettes that a person smokes and the number of years that they continue to smoke. If a person gives up smoking, their risk of developing cancer decreases (Figure 3.3.4).

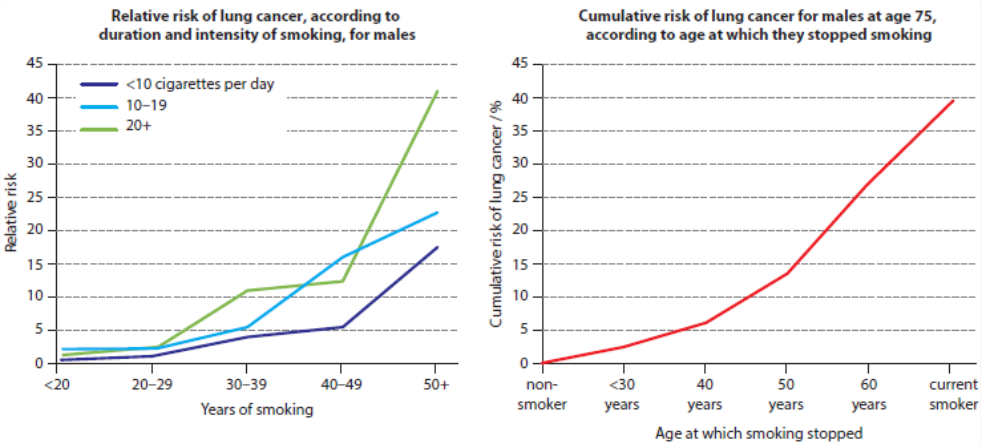

Figure 3.3.4: Graphs to show the relationship between smoking and lung cancer, and the cumulative risk of lung cancer among men in the UK at age 75 according to the age at which they stopped smoking (data from Cancer Research UK).

Lung cancer develops slowly and it takes years before the effects of the carcinogens become obvious. The number of males who suffered lung cancer in the UK was at its highest

levels in the early 1970s. This was as a result of a peak in smoking 20–30 years earlier. Cancer in females increased through the 1970s and 1980s because more females smoked in the 1950s and 1960s. Statisticians predicted that cancer in females would increase to reach the same levels as those in males over the next decade. New government education campaigns have persuaded people give up smoking and new laws have limited smoking in public places, and so the number of deaths have started to decrease. Figure 3.3.5 shows the incidence of lung cancer in the UK 1978–2017 and the incidence of smoking since 1948. As had been predicted, the rate for females increased by 15% but for males rates decreased by 11%.

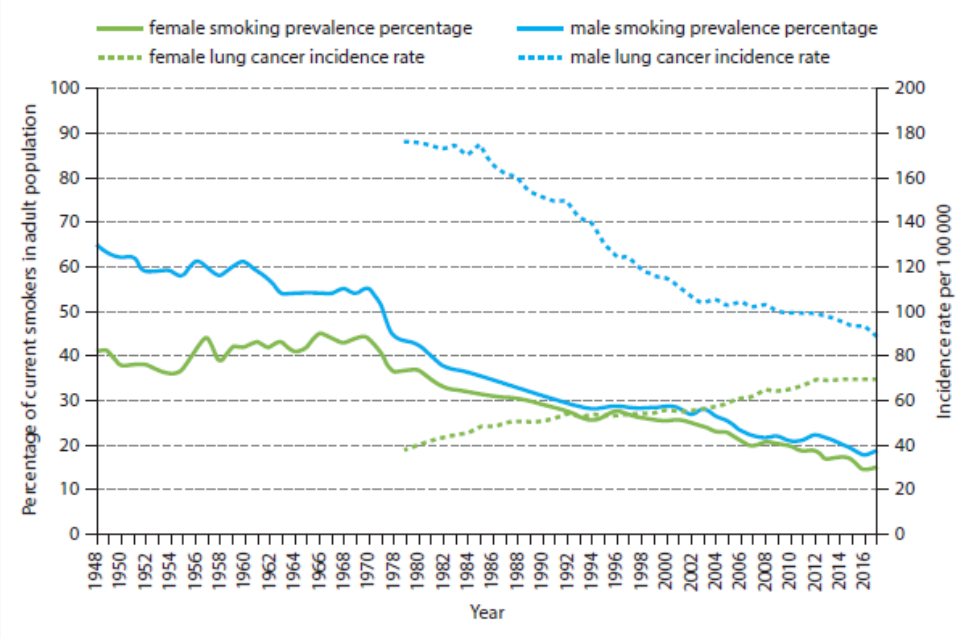

Figure 3.3.5: The incidence of smoking in UK males and females 1948–2017 and the incidence of lung cancer 1978–2017.

To consider:

- Lung cancer affects the economic performance of a country in health care costs and loss of production when people are unwell. Why do you think that governments and officials were reluctant to accept the link between smoking and health in the 1950s?
- How important were laws and restrictions in persuading people to modify their smoking behaviour?

3.4 Epigenetics

LEARNING OBJECTIVES

In this section you will:

- define gene expression as the mechanism in which genetic information affects the phenotype
- define epigenetics as the study of changes to gene activation in differentiated cells
- learn that gene expression is regulated by proteins that bind to base sequences in DNA and that degradation of mRNA can regulate translation
- understand that epigenetic changes modify the activation of certain genes but do not change their base sequences so that the phenotype will change but genotype does not
- learn that DNA methylation inhibits transcription
- understand that epigenetic changes are faster than changes caused by natural selection
- recognise that environment has an impact on gene expression and can trigger heritable changes in epigenetic factors
- learn that most epigenetic tags in gametes are removed from the embryo genome, but some remain and are

inherited leading to the appearance of phenotypic differences in organisms such as lion-tiger hybrids

- learn that external factors such as hormones and nutrients can affect the pattern of gene expression
- understand how environmental effects on DNA methylation can be studied using monozygotic twins.

GUIDING QUESTIONS

- How is DNA modified to influence gene expression?
- How is differentiation of cells brought about by epigenetics?

3.4.1 Epigenetics and gene expression

As our genes are transcribed and translated and proteins are built from the information they carry, an organism's appearance, enzymes and metabolism are all controlled and decided. **Gene expression** is the mechanism by which information carried in DNA has its effects on the phenotype of an organism. The stages of this process include transcription, translation and the function of the protein that is produced. Gene expression is regulated and controlled so that the genes that are expressed match the needs of the organism. Transcription can be regulated by proteins that binds to specific base sequences; these may be promoters, enhancers or transcription factors which either allow or prevent transcription of a gene. Translation can be regulated by the length of time that mRNA is present in a cell and this too is controlled in the cytoplasm. In addition, epigenetic changes can influence patterns of development and differentiation of cells without changing the genotype of the cell.

KEY POINTS

gene a particular section of a DNA strand that codes for a specific polypeptide; a heritable factor that controls a specific characteristic.

gene expression the mechanism by which genetic information affects the phenotype of an organism

genome the entire set of DNA instructions found in a cell

proteome the complete set of proteins expressed by an organism

transcriptome all the mRNA molecules expressed from the genes of an organism

Regulation of transcription by proteins that bind to DNA

Gene expression and binding proteins

Before transcription can begin and mRNA production start, RNA polymerase requires the presence of a class of proteins known as general transcription factors. **Transcription factors (TFs)** are regulatory proteins whose function is to activate transcription of DNA by binding to specific DNA sequences. TFs specifically bind to target sequences which are highly conserved. These sequence specific transcription factors are probably the most important mechanism of gene regulation in cells. In eukaryotes gene expression requires the co-ordinated interaction of several of these proteins. Interactions between the transcription factors, RNA polymerase and the **promoter region** of the DNA molecule allow the RNA polymerase to attach and move along a gene so that transcription can occur. There will be greater transcription of certain genes when specific transcription factors are present so that the proteins that a cell needs can be assembled. Many different transcription factors have been found and each one is able to recognise and bind to a specific nucleotide sequence in DNA.

KEY POINTS

transcription factors (TFs) proteins that bind to particular sites on DNA and activate transcription. Together with RNA polymerase and other proteins (activators), TFs form the

transcription apparatus and have a key role in regulating genes.

promoter a region of DNA to which proteins (RNApolymerase and TFs) bind to initiate transcription of a gene

The role of activators is shown in Figure 3.2.3. These proteins bind to a region of the DNA called the **enhancer**, which may be some distance from the gene. 'Bending proteins' may then assist in bending the DNA, so that the enhancer region is brought close to the promoter. Activators, transcription factors and other proteins attach, so that a 'transcription-initiation complex' is formed and transcription can begin.

Transcription factors are regulated by signals produced from other molecules. For example, hormones are able to activate transcription factors and thus control transcription of certain genes. Many other molecules in the environment of a cell or an organism can also have an impact on gene expression and protein production.

Regulation of translation by mRNA degradation

Once mRNA has reached the cytoplasm, translation can be regulated by mRNA persistence (length of time it is present). mRNA is degraded by nuclease enzymes and the degradation of mRNA and the efficiency with which it is translated are another essential stage in determining gene expression. Individual mRNA molecules can exist in an active state, a silent state or a state that is targeted for decay. In general, RNA is degraded at the end of its useful life, which is very short for introns and spacer fragments, but longer for other sections of mRNA. The time varies between a few minutes to a few days. RNA

molecules with defects in processing or assembly are rapidly identified and degraded by the nuclease enzymes. mRNA lifespan can be shortened if translation is incorrect affected or made more stable if translation elongation or termination are inhibited. mRNA lifespan can also be altered in response to developmental, environmental and metabolic signals.

RNA silencing

RNA silencing is one method of gene silencing that includes several pathways to control and regulate gene expression. Small non-coding strands of RNA (such as microRNAs and RNAi) may either block sections of transcribed mRNA by pairing with it, or degrade the mRNA in the cytoplasm. In both cases the mRNA is not translated and both methods therefore can prevent the translation of some genes. RNA silencing is also used to silence genes in research into the production of medicines to combat cancer and other diseases. This is because RNA silencing is used in the cells of most organisms to fight RNA viruses which are destroyed in the cytoplasm after transcription.

Epigenetics is a relatively new area of investigation in biology. It is the study of how the expression of DNA can be changed without changing the structure of DNA itself. The phenotype (characteristics) of an organism may be changed but its genotype (sequences of DNA) remain the same. Epigenetic changes can affect how cells read their genetic code and a few of the changes can be passed on to the next generation.

Environment and gene expression

The expression of genes can be influenced by the environment - not only the organism’s external environment, but also its internal environment, which is affected by chemicals such as hormones and various products of metabolism. Temperature,

light and chemicals are just some of the environmental factors that can cause some genes to be turned on or off and influence how an organism functions or develops.

KEY POINT

gene silencing interruption or suppression of the expression of a gene either at transcription or translation

3.4.2 Epigenetic changes

The activity of genes can be influenced by certain DNA modifications that do not change an organism's DNA sequence but which do affect which genes are active and which are not. These changes to gene activity will influence the phenotype of an individual but not its genotype. Chemical compounds attached to single genes can produce modifications known as epigenetic changes. When chemical compounds are attached to the genome, it is referred to as an epigenome. The additions or 'tags' are not part of the DNA sequence, but remain attached to DNA as cells divide and, in some cases, can be passed on to offspring. One group of chemical tags are methyl groups which attach to the DNA base cytosine when it is followed by a guanine, and it is their location that influences the expression of the associated gene. Tagging patterns vary from one cell to the next.

KEY POINTS

epigenetics the study of changes to gene activation in differentiated cells.

epigenome all the chemical compounds that have been added to a genome to regulate the expression of all the genes within the genome.

All cells in an organism contain the same DNA, but the many different cell types function differently, so that muscle cells, for example, produce different proteins from cells of the intestine. **Epigenetic changes** help to determine whether genes are turned on or off and determine which proteins are transcribed in each

cell. Cells are different because some of their genes are turned off, while others are turned on. **Epigenetic silencing** turns genes off so that only necessary proteins are produced and enable different cells to behave differently. Environmental influences, such as diet and exposure to pollutants, can also affect the **epigenome**. (See Science in Context, Nutrition and epigenetics). Here we discuss the three most important types of epigenetic change: DNA methylation, histone modification and RNA silencing.

SCIENCE IN CONTEXT

Nutrition and epigenetics

We know that an organism's development is influenced by genes being switched on or off at specific times and there has been much debate about how environmental factors can lead to such epigenetic modifications. The environment's effects can influence human health, and some of these effects can be inherited.

In the early 21st century, Swedish scientists investigated whether nutrition affected the death rate associated with cardiovascular disease and diabetes in a number of Swedish families. They were interested to learn whether the effects were passed from parents to their children and grandchildren.

Researchers examined records of harvests and food prices in Sweden from the 1890s onwards and the medical records of three generations of the families. Their studies revealed that if a father did not have sufficient food in his pre-pubescent years, his sons were less likely to suffer from cardiovascular disease.

For diabetes the picture was different: the children's death rate due to diabetes was unaffected if their father had had a plentiful supply of food at the same critical period. But if the children's grandfather had been well nourished at the same critical period of his life, the incidence of diabetes in his grandchildren was increased.

This suggests that diet can cause epigenetic changes to human genes that affect likelihood of disease. Furthermore, the changes have been passed on through males in a family in a similar way to the inheritance of coat colour in mice described in section 3.4.4.

To consider:

1 Why do you think that researchers used records from the 1890s in their investigations?

2 What other factors might be important in the development cardiovascular disease and diabetes?

DNA methylation

DNA methylation is the most common type of epigenetic modification. It involves attaching methyl groups, consisting of one carbon atom and three hydrogen atoms, to segments of DNA (Figure 3.4.1). When methyl groups are added to a particular gene, that gene is turned off or silenced, and no protein is produced from it.

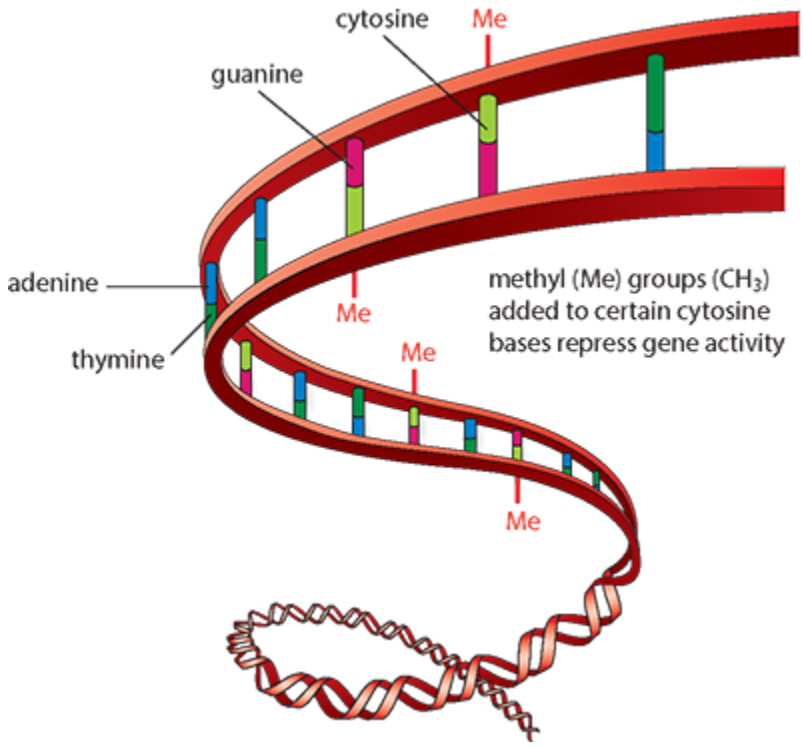

Figure 3.4.1: Methylation of DNA is one epigenetic factor affecting gene expression.

DNA methylation always happens in a region known as a CpG site where a cytosine nucleotide is located next to a guanine nucleotide that is linked by a phosphate. The enzyme DNA methyl transferase adds methylation markers to the base cytosine. Inserting methyl groups here changes the appearance and structure of DNA, and modifies the interactions between transcription factors that determine whether the gene will be expressed. Promoter regions of genes often lie within areas known as 'CpG islands' and if CpG is methylated, the gene will not be expressed.

Special binding proteins can recognise and 'read' these epigenetic markers. If binding protein is missing or a mutation occurs in it, genes which should not be expressed will transcribed.

KEY POINT

DNA methylation a process that adds a methyl group to the DNA base cytosine. It is an example of an epigenetic marker.

EXTENSION

Rett syndrome, a neurological disorder which affects brain development in girls, has been linked to the absence of binding proteins which recognise methylated markers in DNA.

Histone modifications

DNA in eukaryotes is packaged around histones and incorporated into nucleosomes so that the genetic material can be stored in a compact form (Figure 3.4.2) known as chromatin. In order to transcribe genes, enzymes involved in transcription must be able to gain access to DNA. In all eukaryotes, the regions of DNA that contain promoters and regulators have fewer nucleosomes than other areas, allowing greater access for binding proteins, while regions that are transcribed have a higher density of nucleosomes. Nucleosomes have an important role in determining which genes are transcribed and can influence cell variation and development.

KEY POINT

chromatin is an association of histone proteins and DNA which help to package DNA in a compact form in the cell nucleus.

DNA does not need to be completely released from a nucleosome to be transcribed. Nucleosomes are very stable protein–DNA complexes but they are not static. They can undergo structural rearrangements, including ‘nucleosome sliding’ and DNA site exposure. Nucleosomes are important because they can either inhibit or allow transcription by controlling whether binding proteins can access DNA.

Histones can be modified so that they influence the arrangement of chromatin about the chromosome and thus also DNA transcription. If chromatin is compact (condensed) it will prevent DNA transcription but if it is loose (active) DNA can be transcribed. Histones can either be methylated or acetylated by the addition of a methyl or acetyl group to the amino acid lysine in the histone. Acetylation produces active, less condensed chromatin so that proteins involved in transcription can bind to DNA and a gene can be transcribed. Histone methylation can indicate either active or inactive regions of chromatin (Figure 3.4.3).

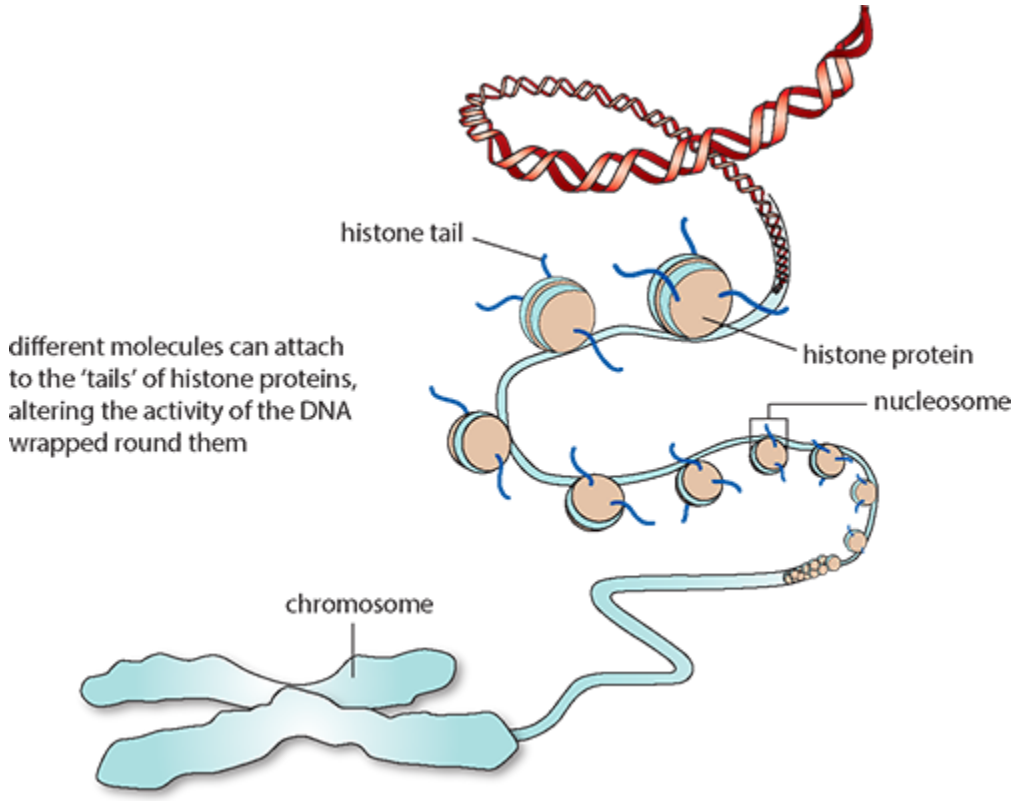

Figure 3.4.2: Histone modification is another epigenetic factor affecting gene expression.

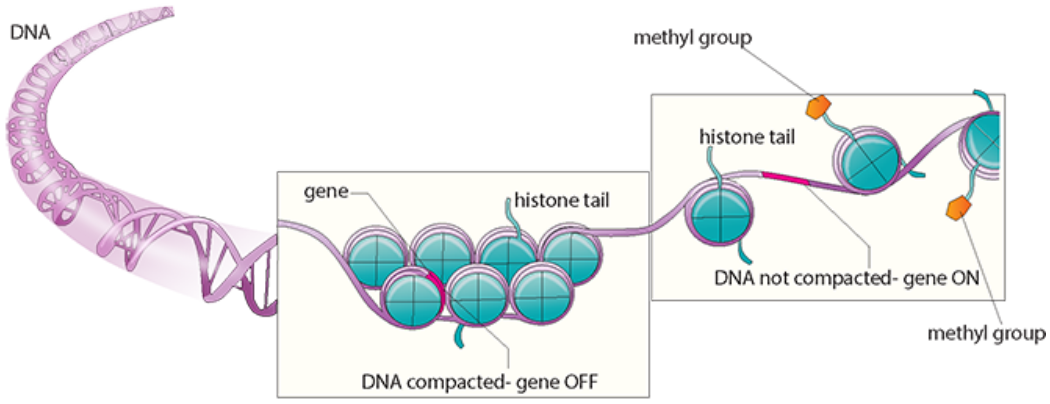

Figure 3.4.3: Effect of methylation on chromatin.

Histone methylation and deacetylation are also important in female mammals who have two X chromosomes. One of the two

chromosomes is inactivated so that females do not produce twice as many X-chromosome gene products as males.

3.4.3 Epigenetic markers and offspring

The genes in egg and sperm cells from the same species contain different epigenetic markers, which cause them to be differentially expressed in the zygote (the cell produced by the fusion two gametes) and developing embryo. These genes are known as imprinted genes. The expression of these genes depends upon which parent contributed them. Most epigenetic markers (tags) found in egg and sperm cells are removed from the epigenome of the embryo and so are not inherited. But in mammals a few do remain and are passed on. One example of this has been investigated in agouti mice. The *Avy* (agouti variable yellow) gene in mice influences the animal's coat colour. Research into the *Avy* gene established that coat colour was related to the degree of methylation of the gene. A high degree of methylation inactivates the gene so the mouse has a dark coat. Without methylation, the gene is active and the coat is yellow. An active gene is also linked to an increased likelihood of obesity and diabetes. Later research showed that if pregnant mice were fed with increased the levels of methylated molecules, such as folic acid and zinc, DNA methylation at the agouti locus was found in their offspring. Baby mice were born with darker coats and leaner bodies (Figure 3.4.4). Furthermore, the mother's diet affected not only characteristics of her own offspring, but also of her daughters' offspring in the next generation.

Scientists had previously believed that methylation markers were always removed from DNA as sperm and egg cells were produced but these experiments suggest that markers must remain, for at least some genes.

Another example of imprinted genes affecting phenotype can be seen in the cat family. Lions and tigers do not normally meet in nature, but in captivity they may mate and sometimes produce hybrid offspring. The offspring look different, depending on which animal is the mother. A male lion and a female tiger produce a liger – the largest of the big cats. A male tiger and a female lion produce a tigon, a cat that is about the same size as its parents. The difference in size and appearance between ligers and tigons is due in part to the parents' differently imprinted genes.

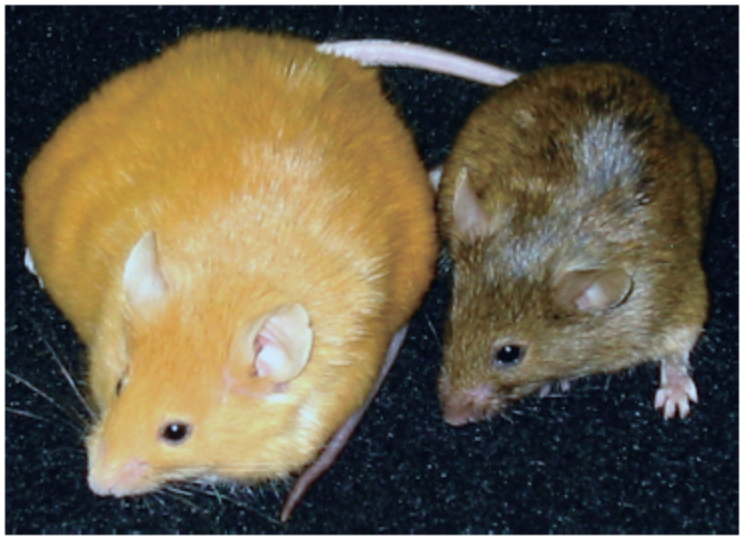

Figure 3.4.4: These mice are genetically identical and the same age. The mother of the left-hand mouse received a normal 'mouse diet' during pregnancy, while the mother of the mouse on the right was fed supplements including folic acid.

Genomic imprinting

Even though both parents contribute equally to the genetic content of their offspring, genomic imprinting sometimes leads to expression of certain genes from only one parent. A marker or imprint can affect particular genes on the maternal and paternal

chromosomes in such a way that only one copy of those genes is expressed in the offspring.

KEY POINT

genomic imprinting inheritance that is not controlled in a Mendelian way. Genes are silenced through DNA methylation but the pattern of gene expression is different depending on whether the gene comes from the father or mother.

Prader–Willi syndrome and Angelman syndrome are human genetic disorders caused by the same mutation on chromosome 15. Prader–Willi syndrome is a disorder that causes behavioural and cognitive problems, deficiencies in sexual development and obesity. It occurs when a mutation is inherited from the child's father and the gene from the mother is imprinted or silenced. Angelman syndrome occurs when the mutated gene from the mother is active. Although the same mutation is involved, Angelman syndrome causes developmental problems, sleep disorders and hyperactivity, but people with the condition have a normal life expectancy and laugh readily.

3.4.4 Rate of epigenetic change

Epigenetics shows us that gene expression can change in a more complex way than simple changes to the DNA sequence. If epigenetic changes occur in sperm or egg cells, the changes are inherited by offspring, as in the case of the *Avy* gene in mice.

Epigenetic changes occur more rapidly than changes due to natural selection (Section 11.2). The rate of changes, such as DNA methylation, is much higher than rates of mutations transmitted genetically and they are also easily reversed. This provides a way for variation within a species to increase quickly, especially if the environment is rapidly changing. This will be the case for both epigenetic effects within a single generation, as well as those which persist into the next generation. Epigenetic changes may also create new heritable variation that will enable organisms to adapt over a longer period of time. If a population is very small and has little genetic variation then epigenetic variation can help organisms adapt to new or changing environments. Many epigenetic effects are caused by the environment (refer to Section 3.4.5) and influence phenotypes in many ways. If the environment is changing slowly, the environment of a parent may serve to predict what their offspring may encounter.

Recent data also suggest that epigenetic patterns may change during the course of life, so that key genes in vital processes may be affected with age.

3.4.5 Pollution, methyl tags and twin studies

Imprinted genes are very sensitive to environmental signals because they have only a single active copy so any epigenetic changes will have a greater impact on gene expression.

Environmental signals can also affect the imprinting process itself. Imprinting happens during egg and sperm formation, when epigenetic tags are added to silence specific genes. Diet, hormones and toxins can all affect this process and the expression of genes in the next generation.

Chemicals in the environment are being linked to many processes that affect DNA methylation and histone modification. Investigations have identified a range of environmental chemicals that affect epigenetic markers and these include the metals cadmium, arsenic and nickel, and air pollutants including particulates and benzene.

Air pollution from car exhausts is an important cause of breathing problems and other respiratory disorders. High levels of the tiny particles (less than 2.5 μm in diameter) found in exhaust fumes not only irritate the lungs but can also enter the bloodstream and cause inflammation in the body. Inhaling these fine particles has been linked to DNA methylation in the T-helper cells of the immune system. These cells have a key role in our response to inflammation.

THEORY OF KNOWLEDGE

Nature or nurture?

Studies of identical and fraternal twins are used to separate the influences of genes and the environment on particular characteristics. If a characteristic is more common in identical twins than fraternal twins, it is likely that genetic factors are at least partly responsible. This is because identical twins have the same genes, whereas fraternal twins are likely to share only 50% of their genes. Twin studies allow scientists to study the influence of 'nature versus nurture', a phrase that was first used by British scientist Francis Galton. Galton came to realise how important studying twins could be and in 1875 he wrote 'The History of Twins' and tried to quantify the relative effects of nature versus nature on human intelligence. He believed strongly that intelligence is largely inherited and so, to improve humanity, the ablest and healthiest people should be encouraged to have more children. But his ideas were used by eugenicists who took them further and proposed that the human species could be improved by preventing the least able or those with 'undesirable' characteristics from having children at all.

In the early 21st century, Eric Turkheimer, a United States professor of psychology, looked again at the inheritance of IQ and studies involving twins. He noticed that most of the studies that reported that IQ is an inherited characteristic involved twins from affluent homes. When he looked at twins from low-income families, he found that the IQ of identical twins varied just as much as the IQ of fraternal twins. From his research he deduced that income can affect a child's natural intelligence. More recently, a much larger study showed that the relationship between income, genetics and IQ is not straightforward and there are many more variables and influencing factors.

Today twin studies are being used to investigate a range of factors including eating disorders, obesity and sexual orientation.

To consider:

1. To what extent should science be used to provide evidence for complex human characteristics such as intelligence?
2. How difficult is it to assess a person's intelligence?
3. How are such assessments influenced by the background and preconceptions of the assessor?

Monozygotic (identical) twins have the same genomes as they develop from a single fertilised ovum. But in many cases monozygotic twins are not identical in their phenotypes and in the diseases and conditions that affect them. When their genomes are investigated, results show that there are differences in methylation patterns in the twins' DNA. In studies, twins have shown substantial differences in the occurrence of schizophrenia, autism and diabetes, and these differences have been show to be related to methylation patterns of the twins' DNA. Other common diseases that may also be linked to epigenetic effects include heart disease and cancers, which are often influenced by environmental factors such as pollution.

Disease-discordant identical twins make good subjects for studying differences in disease linked to pollution and methylation patterns because their genes are matched and many non-genetic effects, such as their early environment, maternal influences and age, are also the same.

3.4.6 External factors affecting the pattern of gene expression

As well as factors such as TFs inside the nucleus, mRNA degradation and epigenetic factors, many external factors also influence the pattern of gene expression. These include hormones and chemicals which are present in the external environment of the cell and influence what happens inside it.

Hormones are factors produced from outside a cell that can affect the cell's gene expression. Steroid hormones such as estrogen influence the transcription of many genes because they interact with receptors inside cells. Estrogen binds to estrogen receptors in the plasma membrane and from there activates signalling pathways. Steroid hormones pass through the plasma membrane of a target cell and bind to intracellular receptors in the cytoplasm or in the nucleus. The cell signalling pathways induced by the steroid hormones regulate specific genes in the cell's DNA. The hormones and receptor complex act as transcription regulators by binding to the promoter region of the gene, this stimulates RNA polymerase binding and gene transcription and thus increasing or decreasing the synthesis of mRNA molecules of specific genes. This, in turn, determines the amount of corresponding protein that is synthesized. (You can read more about the details hormone activity in section 7.3.1)

In bacteria **biochemical factors** affect gene expression. One example is lactose which affects the expression of genes needed for lactose metabolism in *E coli* bacteria. The *lac* operon of *E. coli* contains genes involved in lactose metabolism. The bacteria are able to break down lactose, but if glucose is present, they prefer to use it as an energy source. Glucose metabolism

involves fewer steps and needs less energy to metabolise. But if lactose is the only sugar available, the *E. coli* will use it instead. The lac operon is only expressed if lactose is present, and glucose is absent.

Two regulators turn the operon on and off in response to lactose and glucose levels. The *lac* repressor acts as a lactose sensor. It usually blocks transcription of the operon but stops acting as a repressor when lactose is present. The other regulator is catabolite activator protein (CAP), which acts as a glucose sensor. CAP will bind to specific DNA sites in or near promoter regions and enhance the ability of RNA polymerase to bind and initiate transcription. It activates transcription of the operon, but only when glucose levels are low. (Fig 3.4.5)

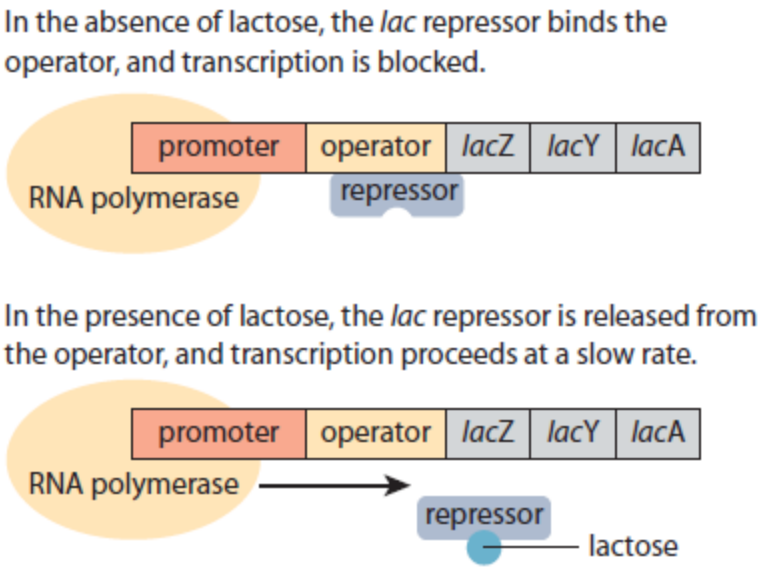

Figure 3.4.5: The *lac* operon.

TEST YOUR UNDERSTANDING

16 Define the term epigenetics.

17 Outline how DNA methylation affects activation of genes.

18 Suggest what effect pollution may have on DNA methylation.

19 Give an example of an epigenetic tag that is not removed from an embryo's epigenome.

20 Why are monozygotic twins useful in epigenetic studies?

REFLECTION

Epigenetics is a new area of scientific research. How much did you know about it before working on this section?

SELF-ASSESSMENT CHECKLIST

Think about the topics covered in this chapter. Which parts are you most confident with? Which topics require some extra practice?

I can...	Subsection	Needs more work	Nearly there	Confident to move on
explain that DNA replication is semi-conservative and produces two identical molecules	3.1.1			
state the two roles of the enzyme helicase	3.1.1			
outline the function of the PCR and list two examples of its use	3.1.2			
outline the technique of gel electrophoresis and its use	3.1.2			
	3.1.3			

explain the orientation of DNA strands and how DNA polymerases work in a 5′→3′ direction				
distinguish the leading and lagging strands	3.1.3			
describe the process of DNA replication in eukaryotes and the functions of primase, polymerases and ligase	3.1.3			
define and outline the process of transcription	3.2.1			
describe how nucleosomes regulate transcription	3.2.1			
explain the importance of introns and exons	3.2.1			
define translation and the	3.2.2			

importance of complementary base pairing				
explain the role of ribosomes in the formation of polypeptides	3.2.2			
recall that free ribosomes synthesise proteins for use within the cell	3.2.2			
summarise the types and importance of non-coding DNA	3.2.3			
outline the roles of promoter regions and telomeres	3.2.3			
describe a polysome and identify them in electron micrographs	3.2.3			
outline how functional proteins are produced after translation	3.2.3			

state that new alleles are formed by mutation and changes may be harmful, beneficial or neutral	3.3.1			
explain that mutations may add, delete or substitute a base, or invert a section of dna, and that substitution causes sickle-cell disease, addition of repeated sequences leads to Huntington's disease	3.3.1			
describe how the genetic code is degenerate and gives resilience to changes	3.3.1			
state that tumours are groups of cells that grow out of control and may be benign or malignant	3.3.2			

define mutagen and give some examples	3.3.2			
outline the importance of apoptosis	3.3.2			
define epigenetics	3.4.1			
explain how gene expression is regulated by binding proteins	3.4.1			
state that epigenetic changes affect phenotype but not genotype	3.4.2			
describe how epigenetic changes can be due to DNA methylation and modification of histones	3.4.2			
give an example of a heritable epigenetic change	3.4.3			
state that most epigenetic changes are not inherited but	3.4.4			

some can affect the epigenome of the offspring				
recall that epigenetics can cause variation more quickly than natural selection in a changing environment	3.4.4			
outline the importance of pollution in epigenetics	3.4.5			
describe the importance of monozygotic twins in the study of epigenetics.	3.4.5			

REFLECTION

Reflect upon the content of this chapter and identify those areas of strength and weakness in your understanding. How can you improve in those topics you have found difficult?

EXAM-STYLE QUESTIONS

You can find questions in the style of IB exams in the digital coursebook.

Chapter 4

Genetics

D1.3, D2.2, D3.2, D3.3

INTRODUCTION

Genetics is the study of genes and the way in which they are passed from one generation to the next. Every organism has a unique combination of genes making up their own genome. Genes determine all the characteristics of an organism and variations of genes, known as alleles, make each individual different from others members of the species.

4.1 Inheritance

LEARNING OBJECTIVES

In this section you will:

- define the genome is the whole genetic information of an organism
- understand that prokaryotes have one circular chromosome without histones
- learn that eukaryotes have a number of linear chromosomes associated with histones, and are normally enclosed within a nucleus
- recognise that diploid nuclei have pairs of homologous chromosomes, while haploid nuclei have one set of chromosomes
- define a gene as a length of DNA that occupies a specific locus on a chromosome and carries instructions for a specific characteristic
- define an allele as a form of a gene that differs from a corresponding allele by one or a few bases
- learn that homologous chromosomes carry the same sequence of genes but not necessarily the same alleles
- learn that one copy of each pair of homologous chromosomes is inherited from each gamete
- understand that alleles are inherited to form a genotype and interact to form a phenotype

- discover that different eukaryotes have different genome sizes, numbers of chromosomes and numbers of genes. The number of chromosomes is a characteristic of each species
- define an organism's karyotype as the number and type of chromosomes in the nucleus and understand that variations may be the result of whole chromosome mutation
- learn that a karyogram is a picture that shows an organisms' chromosomes
- recognise that most mammals have heteromorphic males (XY) and homomorphic (XX) females. In mammals X inactivation occurs so that one of the X chromosomes becomes and inactive Barr body

GUIDING QUESTIONS

- How is genetic information organised inside cells to carry genetic information for the whole organism?
- How can genes and chromosomes interact to affect the phenotype of an organism?
- What changes to genes and chromosomes cause interactions that lead to the appearance of a new phenotype or the loss of an existing phenotype?

4.1.1 The genome

Chimpanzees are set apart from all other organisms because their parents were chimpanzees and their offspring will also be chimpanzees. Every living organism inherits its own blueprint for life in the chromosomes and genes that are passed to it from its parents. The study of genetics attempts to explain this process of heredity and it also plays a very significant role in the modern world, from plant and animal breeding to human health and disease.

The genome of an organism is defined as the whole of its genetic information and every cell has a complete copy of the organism's genome. Genome analysis is an important field of modern biological research. Comparative genomics is an area that analyses genomes from different species. Genomes are compared to gain a better understanding of how species have evolved and to work out the functions of genes and also of the non-coding regions of the genome. Researchers look at many different features such as sequence similarity, gene location, the length and number of coding regions within genes and the amount of DNA that does not code for proteins. They use computer programs to line up genomes from different organisms and look for regions of similarity. There are many databases that store this information on DNA base sequences (Table 4.1.1). Most are freely available online to scientists and students, so that anyone with internet access can use them to compare DNA and protein sequences. You can find them by typing the database name into any search engine.

Database	Description
ENA	overview of all complete genomes deposited in

Genomes Server	the European Nucleotide Archive
Ensembl	a joint project between the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) and the Wellcome Sanger Institute in the UK to produce and maintain automatic annotation of eukaryotic genomes
Ensembl Genomes	provides access to genomes of non-vertebrate species
GenBank (US National Center for Biotechnology Information)	a collection of all publicly available DNA sequences, providing up-to-date and comprehensive DNA sequence information

Table 4.1.1: Some databases that hold information on genomes.

4.1.2 Chromosome structure

Chromosomes are made of DNA molecules and carry the genetic code for each organism but prokaryotic and eukaryotic chromosomes are different in their structure. Prokaryotes have a much simpler chromosome than eukaryotes. Prokaryotes contain a circle of DNA that is often concentrated in one area of the cell, whereas eukaryotes have linear DNA that is associated with histone proteins. Some of these proteins are structural and others regulate the activities of the DNA. Prokaryotes have additional genetic material in the form of small circular structures known as **plasmids**. Prokaryotes are much simpler organisms and so require fewer genes to maintain themselves. The differences between prokaryotic and eukaryotic genetic material are summarised in Table 4.1.2.

Prokaryotic DNA	Eukaryotic DNA
cell contains a circular chromosome, sometimes called a nucleoid	chromosomes are made of linear DNA molecules enclosed in a nucleus, bound by a double membrane
cell contains additional genetic material as small circular plasmids	no plasmids
DNA is ‘naked’ and is not associated with proteins	DNA is associated with histone proteins
cell contains just one circular chromosome	cell contains two or more chromosome types

Table 4.1.2: A comparison of prokaryotic and eukaryotic genetic material.

During the phase of the cell cycle known as interphase, eukaryotic chromosomes are in the form of long, very thin threads, which cannot be seen with a simple microscope. As the nucleus prepares to divide, these threads undergo repeated coiling and become much shorter and thicker (Figure 4.1.1). When stained, they are clearly visible even at low microscope magnifications.

Other aspects of chromosome structure are described in Sections 2.5 and 4.2.

Eukaryotic chromosomes

Eukaryotic species have two or more chromosome types. Their chromosomes form pairs, which are known as homologues. **Homologous** pairs are about the same length and carry the same sequence of genes at the same locations along their length. The form of the genes (alleles) on each of the pair is not necessarily the same because, in sexually reproducing organisms, one chromosome will have been inherited from each of the two parents. So a gene that determines flower colour in a plant would be at the same location on each chromosome in a homologous pair but the allele on the maternal chromosome might not be the same as that on the paternal chromosome.

Figure 4.1.1: Coloured scanning electron micrograph of human chromosomes. They have replicated prior to cell division and so consist of two identical copies (chromatids) linked at the centromere (× 7080). The chromatids split at the start of anaphase (Section 6.5) and become individual chromosomes.

Homologous chromosomes are found in the nuclei in the cells of **diploid** organisms. But if each chromosome exists alone with no partner, the cell is said to be **haploid**. Human somatic cells (or body cells) are diploid and contain 46 chromosomes in 23 homologous pairs; human gametes are haploid and contain only 23 chromosomes, one of each pair found in the body cells.

The number of chromosomes found in the cells of an organism is a characteristic feature of that organism. The diploid numbers of chromosomes in some well studied species are shown in Table 4.1.3.

Genome size

Genome size is the total number of nucleotide base pairs in one copy of a single genome. Measurements are often made in numbers of base pairs. It is interesting to note that an organism's complexity is not proportional to its genome size (Table 4.1.4): many organisms have more DNA than humans. Nor is variation in genome size proportional to the number of genes. This is due to the high proportion of non-coding DNA that is found in some organisms.

Organism	**Diploid number of chromosomes**
Canis familiaris (domestic dog)	78
Pan troglodytes (chimpanzee)	48
Homo sapiens (human)	46
Mus muscularis (house mouse)	40
Oryza sativa (rice)	24
Drosophila melanogaster (fruit fly)	8
Parascaris equorum (parasitic worm)	2

Table 4.1.3: The diploid numbers of different species.

Organism	**Genome size**	**Notes**

	in base pairs	
T2 phage (a virus which infects bacteria)	3569	first RNA genome sequenced
Escherichia coli (a bacterium)	4.6×10^6	
Drosophila melanogaster (fruit fly)	130×10^6	
Homo sapiens (human)	3200×10^6	
Protopterus aethiopicus (marbled lungfish)	$130\,000 \times 10^6$	largest known vertebrate genome
Paris japonica (Japanese native pale-petal)	$150\,000 \times 10^6$	largest known plant genome

Table 4.1.4: Genome sizes of some different organisms.