DNA is a polymer containing chains of nucleotide monomers. Each nucleotide contains a sugar, a base and a phosphate group. The sugar is 2'-deoxyribose which has five carbons named 1' (prime) 2' etc. There are four types of base: adenine and guanine have two carbon-nitrogen rings and are purines; thymine and cytosine have a single ring and are pyrimidines. The bases are attached to the 1' carbon of the deoxyribose. A sugar plus a base is termed a nucleoside. A nucleotide has one, two or three phosphate groups attached to the 5' carbon of the sugar. Nucleotides occur as individual molecules or polymerized as DNA or RNA.

Nucleotide triphosphates of the four bases are joined to form DNA polynucleotide chains. Two phosphates are lost during polymerization and the nucleotides are joined by the remaining phosphate. A phosphodiester bond forms between the 5' phosphate of one nucleotide and the 3' hydroxyl of the next nucleotide. The polynucleotide has a free 5' phosphate at one end (5' end) and a free 3' OH (3' end) at the other end. The sequence of bases encodes the genetic information. It can be read 5'→3' or 3'→5'. Polynucleotides are extremely long. It is possible to have 4ⁿ different sequences.

DNA molecules are composed of two polynucleotide strands wrapped around each other to form a double helix. The sugar–phosphate part of the molecule forms a backbone. The bases face inwards and are stacked on top of each other. The two polynucleotide chains run in opposite directions. The double helix is right-handed and executes a turn every 10 bases. The helix has a major groove which mediates interactions with proteins. Variant DNA structures have been identified including Z DNA which has a left-handed helix.

Hydrogen bonds between bases on the two DNA strands stabilize the double helix. The available space between the strands restricts the bases that can interact such that a purine always interacts with a pyrimidine. Thus, A interacts only with T and G only with C. This is called complementary base pairing. The restriction on base pairing means that the sequence of bases on the two strands are related to each other, such that the sequence of one determines and predicts the sequence of the other. This allows genetic information to be preserved during replication of the DNA and expression of the genes. Disruption of the hydrogen bonds between the bases by heat or chemicals or by the action of enzymes causes the strands of the double helix to separate.

In RNA thymine is replaced by uracil and 2-deoxyribose by ribose. RNA normally exists as a single polynucleotide strand however, short stretches of base pairing may occur between complementary sequences.
Related topics

<table>
<thead>
<tr>
<th>Gene transcription (A4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA replication (A9)</td>
</tr>
<tr>
<td>DNA mutation (B5)</td>
</tr>
</tbody>
</table>

### Nucleotides

The ability of DNA to carry the genetic information required by a cell to reproduce itself is closely related to the structure of DNA molecules. DNA is a polymer and consists of a long chain of monomers called **nucleotides**. The DNA molecule is said to be a polynucleotide. Each nucleotide has three parts: a sugar, a nitrogen containing ring-structure called a base, and a phosphate group. The sugar present in DNA is a five carbon pentose called 2'-deoxyribose in which the -OH group on carbon 2 of ribose is replaced by hydrogen (Fig. 1). The carbon atoms in the sugar are numbered 1-5. The numbers are given a dash (') referred to as prime to distinguish them from the numbers of the atoms in the base. The numbering is important because it indicates where other components of the nucleotide are attached to the sugar.

Nucleotides contain one of four bases: adenine, guanine, thymine or cytosine (Fig. 2). These are complex molecules containing carbon and nitrogen ring structures. Adenine and guanine contain two carbon–nitrogen rings and are known as purines. Cytosine and thymine contain a single ring and are called pyrimidines. The bases are attached to the sugar by a bond between the 1' carbon of the sugar and a nitrogen at position 9 of the purines or position 1 of the pyrimidines. A sugar plus a base is called a **nucleoside** (Fig. 3a).

Nucleotides contain phosphate groups (PO₄) attached to the 5' carbon of the sugar (Fig. 3b). A nucleoside is called a nucleotide when a phosphate group is attached, the attachment can consist of one, two or three phosphate groups joined together. The phosphate groups are called α, β and γ, with α directly attached to the sugar. Nucleotides may exist in cells as individual molecules (nucleotide triphosphates play an important role in cells as the carriers of energy used to power enzymatic reactions) or polymerized as nucleic acids (DNA or RNA).

### DNA polynucleotides

Nucleotide triphosphates are joined together to give polynucleotides. There are four used to synthesize DNA polynucleotides, 2'-deoxyadenosine 5'-triphosphate (dATP or A), 2'-deoxythymidine 5'-triphosphate (dTTP or T), 2'-deoxycytosine 5'-triphosphate (dCTP or C) and 2'-deoxyguanosine 5'-triphosphate (dGTP or G). The β and γ phosphates are lost during polymerization and the nucleotide units are joined together by the remaining phosphate. The 5' phosphate of one nucleotide forms a bond with the 3' carbon of the next nucleotide eliminating the -OH group on the 3' carbon during the reaction. The bond is called a 3'-5' **phosphodiester bond** (C–O–P) (Fig. 4). The polynucleotide chain has a free
Fig. 3. Structure of (a) nucleosides, (b) nucleotides.

5' triphosphate at one end known as the 5' end and a free 3' hydroxyl group at the other end called the 3' end. This distinction gives the DNA polynucleotide polarity so that a DNA molecule can be described as running 5'→3' or 3'→5'.

It is the sequence of the bases in the DNA polynucleotide that encodes the genetic information. This sequence is always written in the 5'→3' direction (polymerase enzymes copy DNA molecules in this direction). Polynucleotides can be extremely long with no apparent limit to the number of nucleotides and no restrictions on the sequence of the nucleotides. The maximum number of possible base sequences for a polynucleotide is $4^n$, where $n$ is the number of nucleotides. This is an enormous number. For example, a polynucleotide containing just six bases could be arranged as $4^6 = 4096$ different sequences.

The double helix DNA molecules have a very distinct and characteristic three-dimensional structure known as the double helix (Fig. 5). The structure of DNA was discovered
in 1953 by Watson and Crick working in Cambridge using X-ray diffraction pictures taken by Franklin and Wilkins. DNA exists as two polynucleotide chains wrapped around each other to form the double helix. The sugar-phosphate part of the molecule forms a spine or backbone which is on the outside of the helix. The bases, which are flat molecules, face inwards towards the center of the helix and are stacked on top of each other like a pile of plates.

X-ray diffraction pictures of the double helix show repeated patterns of bands that reflect the regularity of the structure of the DNA. The double helix executes a turn every 10 base pairs. The pitch of the helix is 34Å so the spacing between bases is 3.4Å. The diameter of the helix is 20Å. The double helix is said to be 3 antiparallel. One of the strands runs in the 5'→3' direction and the other 3'→5' direction. Only antiparallel polynucleotides form a stable helix. The double helix is not absolutely regular and when viewed from the outside a major groove and a minor groove can be seen. These are important for interaction with proteins, for replication of the DNA and for expression of the genetic information. The double helix is right-handed. This means that if the double helix were a spiral staircase and you were climbing up, the sugar-phosphate backbone would be on your right.

A number of variant forms of DNA occur when crystals of the molecule are formed under different conditions. The form present in cells is called the B form. Another form called the A form has a slightly more compact structure. Other forms that exist are C, D, E and Z, which is striking because it exists as a left-handed helix. Regions in chromosomes containing nonstandard structures such as Z-DNA have recently been identified.

The bases of the two polynucleotide chains interact with each other. The space between the polynucleotides is such that a two-ring purine interacts with a single-ring pyrimidine. Thus, thymine always interacts with adenine and guanine with cytosine. Hydrogen bonds form between the bases and help to stabilize the interaction. Two bonds form between A and T and three between G and C. Thus, G-C bonds are stronger than A-T bonds. The way in which the bases form pairs between the two DNA strands is known as complementary base pairing.
Complementary base pairing and is of fundamental importance (Fig. 6). Combinations other than G-C and A-T do not work because they are too large or too small to fit inside the helix or they do not align correctly to allow hydrogen bond formation. Because G must always bond to C and A to T the sequences of the two strands are related to each other and are said to be complementary with the sequence of one strand predicting and determining the sequence of the other. This means that one strand can be used to replicate the other. This is a vital mechanism for retaining genetic information and passing it on to other cells following cell division. Complementary base pairing is also essential for the expression of genetic information and is central to the way DNA sequences are transcribed into mRNA and translated into protein.

The double helix is stabilized by hydrogen bonds between the base pairs. These can be disrupted by heat and some chemicals. This results in separation of the double helix into two strands and the molecule is said to be denatured. In cells enzymes can separate the strands of the double helix for the purposes of copying the DNA and for expression of the genetic information.

RNA structure

The structure of RNA is similar to that of DNA but a number of important differences exist. In RNA ribose replaces 2'-deoxyribose and the base thymine is replaced by another base, uracil, which can also base pair with adenine (Fig. 7). In addition, RNA molecules normally exist as a single polynucleotide strand and do not form a double helix. However, it is possible for base pairing to occur between complementary parts of the same RNA strand resulting in short double-stranded regions.

Fig. 7. Structures of ribose and uracil.
A gene is a unit of information and corresponds to a discrete segment of DNA that encodes the amino acid sequence of a polypeptide. Human cells contain 50-100,000 genes arranged on 23 chromosomes. The genes are dispersed and are separated by noncoding intergenic DNA. Information is encoded on the template strand which directs the synthesis of an RNA molecule. Both DNA strands can act as the template strand. DNA molecules have an enormous capacity to store genetic information.

Some genes are arranged as clusters known as operons and multigene families. Operons occur in bacteria and contain coregulated genes with a related function. Multigene families occur in higher organisms and contain genes that are identical or similar that are not regulated coordinately. Simple multigene families contain identical genes whose product is required in large amounts. Complex multigene families contain genes that are very similar and encode proteins with a related function.

The biological information encoded in genes is made available by gene expression. In this process, an RNA copy of a gene is synthesized which then directs the synthesis of a protein. The central dogma states that information is always transferred from DNA to RNA to protein. The functioning of cells is dependent on the coordinated activity of many proteins. Gene expression ensures that proteins are synthesized in the correct place at the correct time.

Gene expression is highly regulated. Not all of the genes present in a cell are active and different types of cell express different genes. The expression of a gene is regulated by a segment of DNA upstream of the coding sequence called the promoter. This binds RNA polymerase and associated transcription factor proteins and initiates synthesis of an RNA molecule.

The coding sequence of a gene is split into a series of segments called exons which are separated by noncoding sequences called introns which usually account for most of the gene sequence. The number and sizes of the introns vary between genes. Introns are removed from RNA transcripts by a process called splicing prior to protein synthesis. Introns are not usually present in bacteria.

Copies of some genes exist which contain sequence errors acquired during evolution that prevent them from producing proteins. These are called pseudogenes and they represent evolutionary relics of original genes. Examples include the globin pseudogenes.

Related topics
Regulation of gene expression in prokaryotes (A10)
Regulation of gene expression in eukaryotes (A11)
The human genome (B4)
Structure of genes

The biological information needed by an organism to reproduce itself is contained in its DNA. The information is encoded in the base sequence of the DNA and is organized as a large number of genes, each of which contains the instructions for the synthesis of a polypeptide. In physical terms, a gene is a discrete segment of DNA with a base sequence that encodes the amino acid sequence of a polypeptide. Genes vary greatly in size from less than 100 base pairs to several million base pairs. In higher organisms the genes are present on a series of extremely long DNA molecules called chromosomes. In humans there are an estimated 50-100,000 genes arranged on 23 chromosomes. The genes are very dispersed and are separated from each other by sequences that do not appear to contain useful information; this is called intergenic DNA. The intergenic DNA is very long, such that in humans gene sequences account for less than about 30% of the total DNA. Only one of the two strands of the DNA double helix carries the biological information: this is called the template strand and it is used to produce an RNA molecule of complementary sequence which directs the synthesis of a polypeptide. The other strand is called the nontemplate strand. Both strands of the double helix have the potential to act as the template strand: individual genes may be encoded on different strands. Other terms are used to describe the strands of the double helix as alternatives to template and nontemplate. These include sense/antisense and coding/noncoding: the terms antisense and noncoding are equivalent to the template strand.

The capacity of DNA molecules to store information is enormous. For a DNA molecule n bases long, the number of different combinations of the four bases is 4^n. Even for very short DNA molecules the number of different sequences possible is very large. In practice, there are limitations to the sequences that can contain useful information. However the capacity to encode information remains vast.

Gene families

Most genes are spread out randomly along the chromosomes, however some are organized into groups or clusters. Two types of cluster occur: these are operons and multigene families.

Operons are gene clusters found in bacteria. They contain genes that are regulated in a coordinated way and encode proteins with closely related functions. An example is the lac operon in E. coli which contains three genes encoding enzymes required by the bacterium to break down lactose. When lactose is available as an energy source, the enzymes encoded by the lac operon are required together. The clustering of the genes within the operon allows them to be switched on or off at the same time allowing the organism to use its resources efficiently (Fig. 1).

In higher organisms, operons are absent and clustered genes exist as multigene families. Unlike operons, the genes in a multigene family are identical or are very similar and are not regulated coordinately. The clustering of genes in multigene families probably reflects a requirement for multiple copies of that

<table>
<thead>
<tr>
<th>DNA</th>
<th>lac Z</th>
<th>lac Y</th>
<th>lac A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regulatory DNA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. The lac operon. Three genes (lac Z, Y and A) are arranged and regulated together.
gene which was fulfilled by duplication during evolution. Some multigene families exist as separate clusters on different chromosomes; this probably arose by rearrangements of the DNA during evolution which resulted in the breaking up of clusters. Multigene families may be simple or complex. In simple multigene families the genes are identical. An example is the gene for the 5S ribosomal RNA. In humans, there are about 2000 clustered copies of this gene reflecting the high demand of cells for the gene product (Fig. 2a). Complex multigene families contain genes that are very similar but not identical. An example is the globin gene family that encodes a series of polypeptides (α β γ ε ζ globins) that differ from each other by just a few amino acids. Globin polypeptides form complexes with each other and with a cofactor molecule called heme to give the adult and embryonic forms of the oxygen carrying blood protein, hemoglobin (Fig. 2b).

Gene expression

The biological information in a DNA molecule is contained in its base sequence. Gene expression is the process by which this information is made available to the cell. The use of the information is described by the central dogma, originally proposed by Crick, which states that information is transferred from DNA to RNA to protein (Fig. 3). During gene expression, DNA molecules copy their information by directing the synthesis of an RNA molecule of complementary sequence. This process is known as transcription. The RNA then directs the synthesis of a polypeptide whose amino acid sequence is determined by the base sequence of the RNA. This process is known as translation. The amino acid sequence of the protein determines its three-dimensional structure which in turn dictates its function. The central dogma states that the transfer of information can only occur in one direction – from DNA to RNA to protein – and cannot occur in reverse. An exception to this rule is found in retroviruses which have an enzyme called reverse transcriptase which can copy RNA into DNA. The functioning of cells, and in turn of living organisms, is dependent on the
coordinated activity of many different proteins. The biological information contained within the genes acts as a set of instructions for synthesizing proteins at the correct time and in the correct place.

**Gene promoters**

The expression of the biological information present in genes is highly regulated. Not all the genes present in a cell's DNA are expressed and different genes are active in different cell types. The overall complement of genes that are active determines the characteristics of a cell and its function within the organism. Thus, for example, many of the genes that are active in muscle cells are different from those that are active in blood cells. Expression of genes is regulated by a segment of DNA sequence present upstream of the coding sequence known as the promoter. Conserved DNA sequences in the promoter are recognized and bound by the RNA polymerase and other associated proteins called transcription factors that bring about the synthesis of an RNA transcript of the gene. The expression of a gene in a cell is determined by the promoter sequence and its ability to bind RNA polymerase and transcription factors.

**Introns and exons**

One of the more surprising features of genes is that in higher organisms the coding information is usually split into a series of segments of DNA sequence called exons. These are separated by sequences that do not contain useful information called introns (Fig. 4). The number of introns varies greatly, from zero to more than 50 in some genes. The length of the exons and introns also varies but the introns are usually much longer and account for the majority of the sequence of the gene. Before the biological information in a gene can be used to synthesize a protein, the introns must be removed from RNA molecules by a process called splicing which leaves the exons and the coding information continuous. Introns are a feature of higher organisms only and are not usually found in bacteria.

![Fig. 4. Structure of a gene.](image)

**Pseudogenes**

Some genes exist which resemble other genes but examination of their base sequence shows errors that make it impossible for them to contain useful biological information. These are called pseudogenes and they represent genes that have acquired errors or mutations in their DNA sequence during evolution causing their biological information to be scrambled so that they are no longer able to direct the synthesis of a protein. As such, pseudogenes are evolutionary relics. During evolution, the initial base changes causing loss of biological information are followed by more rapid changes so that the sequence of the pseudogene eventually deviates substantially from the original gene. Examples include several globin pseudogenes that are present in the globin gene clusters.
Section A - Molecular genetics

A3 THE GENETIC CODE

Key Notes

Gene expression
Genetic information is encoded in the base sequence of DNA molecules as a series of genes. Gene expression is the term used to describe how cells decode the information to synthesize proteins required for cellular function. The expression of a gene involves the synthesis of a complementary RNA molecule whose sequence specifies the amino acid sequence of a protein. The DNA sequence of the gene is collinear with the amino acid sequence of the polypeptide.

Genetic code
Amino acids are encoded by 64 base triplets called codons which encode the 20 amino acids. Most amino acids have more than one codon. This is known as the degeneracy of the genetic code and it helps to minimize the effect of mutations. Codons that specify the same amino acid are called synonyms and differ at their third base, known as the 'wobble' position. AUG is the initiation codon and encodes methionine. There are three stop codons: UAG, UGA and UAA.

Reading frames
Three possible sets of codons can be read from any sequence depending on which base is chosen as the start of a codon. Each set of codons is known as a reading frame. The initiation codon determines the reading frame of the protein coding sequence. Other reading frames tend to contain stop codons and are not used for protein synthesis. An open reading frame is a sequence of codons bounded by start and stop codons.

Universality of the code
The genetic code applies universally with all organisms using the same codons for each amino acid. However, some exceptions to the standard codon usage occur in mitochondrial genomes and in some unicellular organisms.

Related topics
Transfer RNA (A5) Translation (A8)

Gene expression
The information required by an organism to reproduce itself is carried by its DNA, encoded in the base sequence and organized as a series of genes. Gene expression is the term used to describe the process by which cells decode and make use of this information to synthesize the proteins that are responsible for cellular function. During gene expression, information is copied from DNA to RNA by the synthesis of an RNA molecule whose base sequence is complementary to that of the DNA template. The RNA then directs the synthesis of a protein whose amino acid sequence is specified by the base sequence of the RNA. For every gene the DNA sequence is collinear with the amino acid sequence of the polypeptide it encodes such that the 5'→3' base sequence of the coding strand specifies the amino acid sequence of the encoded polypeptide from the amino to the carboxy terminus.
Genetic code

The genetic code describes how base sequences are converted into amino acid sequences during protein synthesis. The DNA sequence of a gene is divided into a series of units of three bases. Each set of three bases is called a codon and specifies a particular amino acid. The four bases in DNA and RNA can combine as a total of $4^3 = 64$ codons which specify the 20 amino acids found in proteins (Table 1). Because the number of codons is greater, all of the amino acids, with the exceptions of methionine and tryptophan, are encoded by more than one codon. This feature is referred to as the degeneracy or the redundancy of the genetic code. Codons which specify the same amino acid are called synonyms and tend to be similar. For example, ACU, ACC, ACA and ACG all specify the amino acid threonine. Variations between synonyms tend to occur at the third position of the codon, which is known as the wobble position. The degeneracy of the genetic code minimizes the effects of mutations so that alterations to the base sequence are less likely to change the amino acid encoded and possible deleterious effects on protein function are avoided. Of the 64 possible codons, 61 encode amino acids. The remaining three, UAG, UGA, and UAA, do not encode amino acids but instead act as signals for protein synthesis to stop and as such are known as termination codons or stop codons. The codon for methionine, AUG, is the signal for protein synthesis to start and is known as the initiation codon. Thus all polypeptides start with methionine although this is sometimes removed later.

Table 1. The genetic code

<table>
<thead>
<tr>
<th>First position (5' end)</th>
<th>Second position</th>
<th>Third position (3' end)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>Phe UUU</td>
<td>Ser UCU</td>
<td>Tyr UAU</td>
</tr>
<tr>
<td>Phe UUC</td>
<td>Ser UCC</td>
<td>Tyr UAC</td>
</tr>
<tr>
<td>Leu UUA</td>
<td>Ser UCA</td>
<td>Stop UAA</td>
</tr>
<tr>
<td>Leu UUG</td>
<td>Ser UCG</td>
<td>Stop UAG</td>
</tr>
<tr>
<td>C</td>
<td>Leu CUU</td>
<td>Pro CCU</td>
</tr>
<tr>
<td>Leu CUC</td>
<td>Pro CCC</td>
<td>HisCAC</td>
</tr>
<tr>
<td>Leu CUA</td>
<td>Pro CCA</td>
<td>Gln CAA</td>
</tr>
<tr>
<td>Leu CUG</td>
<td>Pro CCG</td>
<td>Gln CAG</td>
</tr>
<tr>
<td>A</td>
<td>Ile AUU</td>
<td>Thr ACU</td>
</tr>
<tr>
<td>Ile AUC</td>
<td>Thr ACC</td>
<td>Asn AAC</td>
</tr>
<tr>
<td>Ile UAA</td>
<td>Thr ACA</td>
<td>Lys AAA</td>
</tr>
<tr>
<td>Met AUG</td>
<td>Thr ACG</td>
<td>Lys AAG</td>
</tr>
<tr>
<td>G</td>
<td>Val GUU</td>
<td>Ala GCU</td>
</tr>
<tr>
<td>Val GUC</td>
<td>Ala GCC</td>
<td>Asp GAC</td>
</tr>
<tr>
<td>Val GUA</td>
<td>Ala GCA</td>
<td>Glu GAA</td>
</tr>
<tr>
<td>Val GUG</td>
<td>Ala GCG</td>
<td>Glu GAG</td>
</tr>
</tbody>
</table>

Reading frames

In addition to identifying the start of protein synthesis, the initiation codon determines the reading frame of the RNA sequence. Depending on which base is chosen as the start of a codon, three possible sets of codons may be read from any base sequence. In practice, during protein synthesis, normally only one reading frame contains useful information; the other two reading frames usually contain several stop codons which prevent them from being used to
Reading frame 1, 5'– AUG Acc AAG AGA UCC G–3'  
Met Thr Lys Arg Ser

Reading frame 2, 5'– AGA OL A AGA GAU OCG G–3'  
Stop Leu Arg Asp Pro

Reading frame 3, 5'– AU GAC Acc UAA GAG ULA OCG G–3'  
Asp Stop Glu Ile Arg

Fig. 1. Every DNA sequence can be read as three separate reading frames depending on which base is chosen as the start of the codon.

direct protein synthesis (Fig. 1). A set of codons that runs continuously and is bounded at the start by an initiation codon and at the end by a termination codon is known as an open reading frame (ORF). This characteristic is used to identify protein coding DNA sequences in genome sequencing projects.

Universality of the code

Initially the genetic code was believed to apply universally, that is all organisms would recognize individual codons as the same amino acids. However, it has now been shown that some variation in the code exists, although this is rare. For example, mitochondria have a small DNA genome containing about 20 genes in which deviations from the genetic code occur. Changes are mostly associated with start and stop codons. For example, UGA, which is normally a termination codon, codes for tryptophan whereas AGA and AGG which normally encode arginine are termination codons, and AUA, normally isoleucine, specifies methionine. It is thought that these changes tend to be viable because the mitochondrion is a closed system. A few examples of nonstandard codon usage have now been found outside mitochondrial genomes in unicellular organisms. For example UAA and UAG which are normally stop codons, encode glutamic acid in some protozoa.
A4 Gene transcription

Key Notes

Transcription
This is the first stage of gene expression and it involves the synthesis of RNA from a DNA template by RNA polymerase. The RNA is synthesized from the template DNA strand and has the same sequence as the non-template (sense/coding) strand. RNA synthesis involves the polymerization of ribonucleotide triphosphates and occurs 5'→3' in the opposite direction to the template strand which runs 3'→5'.

Transcription in prokaryotes
Transcription occurs in three phases (initiation, elongation and termination). In E. coli a single RNA polymerase composed of five subunits (α2, β, β', σ) exists and is known as the holoenzyme. The σ subunit may dissociate leaving behind the core enzyme.

Initiation
Transcription is initiated at the gene promoter. In E. coli the RNA polymerase recognizes the −10 box and −35 box sequence elements, and the σ subunit binds to the −35 box. Initially, a closed promoter complex forms. The double helix then dissociates at the −10 box to form an open promoter complex. The σ subunit dissociates and RNA synthesis begins.

Elongation
During elongation RNA polymerase adds ribonucleotides to the 3' end of the RNA molecule in an order specified by the template DNA sequence. The enzyme moves along the DNA disrupting the hydrogen bonds between the bases (melting) and unwinding the double helix. To avoid strain on the helix only a short region (12–17 bases) is unwound at any one time. RNA synthesized is initially base-paired to the template strand but is subsequently released allowing the double helix to reform.

Termination
Pallindromic sequences can form stem–loop structures in RNA and act as signals for transcription termination. The RNA polymerase is thought to pause after a stem–loop and weak A–U base pairs that follow it break releasing the transcript. Alternatively, ρ protein disrupts base pairing and releases the transcript. On termination the RNA polymerase is released from the template and may reassociate with a σ factor.

Transcription in eukaryotes
In eukaryotes transcription initiation is complex and termination does not involve stem–loop structures. Three RNA polymerases (I, II and III) transcribe different sets of genes.

RNA polymerase II
This enzyme transcribes genes that encode proteins. Promoter sequences usually contain a TATA box about 25 bp upstream of the transcription start site which binds the enzyme. Transcription factors (TFII A, B etc.) bind DNA around the TATA box in a specified order and form a platform to which RNA polymerase II binds. Genes that lack a TATA box may have an initiator element; however, some genes have neither element. Other promoter elements
such as the CAT box act as binding sites for other transcription factors that influence the rate of transcription initiation. Distant elements called enhancers and silencers also influence the transcription rate. The signals that mediate transcription termination are uncertain.

This enzyme transcribes 18S, 28S and 5.8S ribosomal (r) RNAs. The promoter contains two elements essential for transcription: a core element that overlaps the transcription start site and an upstream control sequence at around position –100. An 18 bp termination signal is present approximately 600 bp after the end of the gene.

This enzyme transcribes short genes encoding transfer (t) RNAs and the 5.8S rRNA. The promoter sequences occur within the coding sequence and are called internal control regions (ICRs). The tRNA gene has two important sequence elements, the A box and the B box; transcription of the 5S rRNA gene requires a sequence called the C box. A second sequence called the A box is important. Transcription terminates at a signal sequence containing a run of A residues.

**Related topics**

- Transfer RNA (A5)
- Ribosomal RNA (A6)
- Messenger RNA (A7)
- Regulation of gene expression in prokaryotes (A10)
- Regulation of gene expression in eukaryotes (A11)

**Transcription**

In this process an RNA copy of the DNA sequence of a gene is produced as the first stage of gene expression. The RNA is synthesized by enzymes called RNA polymerases using DNA as a template. The two strands of the double helix are called the template and the non-template strands. RNA is produced using the template strand and the RNA molecule synthesized is a copy of the non-template strand (Fig. 1). Gene sequences usually refer to the non-template strand. Other names used to describe the non-template strand are the sense (+) strand or the coding strand. The RNA molecule synthesized is called a transcript and may subsequently undergo translation to produce a protein or may be used as ribosomal or transfer RNA.

During transcription RNA is synthesized by the polymerization of ribonucleotide triphosphate subunits (ATP, UTP, GTP, CTP). The 3'-OH of one ribonucleotide reacts with the 5' phosphate of another to form a phosphodiester bond. The order in which the ribonucleotides are added to the growing RNA chain is determined by the order of the bases in the template DNA. New ribonucleotides are added to the growing chain at the free 3' end. The transcript is synthesized in the 5'→3' direction but because the chains must be antiparallel for base pairing the template strand runs in the opposite, 3'→5', direction.

**Transcription in prokaryotes**

In prokaryotic organisms transcription occurs in three phases known as initiation, elongation and termination. RNA is synthesized by a single RNA polymerase enzyme which contains multiple polypeptide subunits. In *E. coli* the RNA polymerase has five subunits: two α, one β, one β' and one σ subunit (α2ββ'σ).
This form is called the **holoenzyme**. The σ subunit may dissociate from the other subunits to leave a form known as the **core enzyme**. These two forms of the RNA polymerase have different roles in transcription.

**Initiation**

Transcription cannot start randomly but must begin specifically at the start of a gene. Signals for the initiation of transcription occur in the **promoter** sequence which lies directly upstream of the transcribed sequence of the gene. The promoter contains specific DNA sequences that act as points of attachment for the RNA polymerase. In *E. coli*, two sequence elements recognized by the RNA polymerase known as the **−10 sequence** and the **−35 sequence** are present. The exact sequences can vary between promoters but all conform to an overall pattern known as the **consensus sequence**. The σ subunit of the RNA polymerase is responsible for recognizing and binding the promoter, probably at the −35 box. In the absence of the σ subunit the enzyme can still bind to DNA but binding is more random. When the enzyme binds to the promoter it initially forms a **closed promoter complex** in which the promoter DNA remains as a double helix. The enzyme covers about 60 base pairs of the promoter including the −10 and −35 boxes. To allow transcription to begin, the double helix partially dissociates at the −10 box, which is rich in weak A–T bonds, to give an **open promoter complex**. The σ subunit then dissociates from the open promoter complex leaving the core enzyme. At the same time the first two ribonucleotides bind to the DNA, the first phosphodiester bond is formed and transcription is initiated (Fig. 2).

**Elongation**

During elongation the RNA polymerase moves along the DNA molecule melting and unwinding the double helix as it progresses. The enzyme adds ribonucleotides to the 3' end of the growing RNA molecule with the order of addition determined by the order of the bases on the template strand. In most cases, a **leader sequence** of variable length is transcribed before the coding sequence of the gene is reached. Similarly, at the end of the coding sequence a noncoding
**transcription sequence** is transcribed before transcription ends. During transcription, only a small portion of the double helix is unwound at any one time. The unwound area contains the newly synthesized RNA base-paired with the template DNA strand and extends over 12-17 bases. The unwound area needs...
to remain small because unwinding in one region necessitates overwinding in adjacent regions and this imposes strain on the DNA molecule. To overcome this problem, the RNA is released from the template DNA as it is synthesized allowing the DNA double helix to reform (Fig. 3).

Termination

The termination of transcription occurs nonrandomly and takes place at specific points after the end of the coding sequence. In *E. coli*, termination occurs at sequences known as *palindromes*. These are symmetrical about their middle such that the first half of the sequence is followed by its exact complement in the second half. In single-stranded RNA molecules this feature allows the first half of the sequence to base pair with the second half to form what is known as a *stem–loop structure* (Fig. 4). These appear to act as signals for termination. In some cases the stem–loop sequence is followed by a run of 5–10 As in the DNA which form weak A–U base pairs with the newly synthesized RNA. It is thought that the RNA polymerase pauses just after the stem–loop and that the weak A–U base pairs break causing the transcript to detach from the template. In other cases the run of As is absent and a different mechanism occurs based on binding of a protein called *Rho* (ρ) which disrupts base-pairing between the template and the transcript when the polymerase pauses after the stem–loop. The termination of transcription involves the release of the transcript and the core enzyme which may then reassociate with the σ subunit and go on to another round of transcription.

Transcription in eukaryotes

Transcription occurs in eukaryotes in a similar way to prokaryotes. However, initiation is more complex, termination does not involve stem–loop structures and transcription is carried out by three enzymes (RNA polymerases I, II, III) each of which transcribes a specific set of genes and functions in a slightly different way.

RNA polymerase II

This enzyme transcribes genes that encode proteins. Binding of RNA polymerase II to its promoter involves several different DNA sequence elements and a number of proteins called *transcription factors*. The promoter usually (but not always) contains a DNA sequence element called the *TATA box* which acts as the attachment site for the RNA polymerase II. This has the consensus sequence
5' TATA(A/T)A(T/A) 3' and is present about 25 bp upstream from the transcription start site. Its function is to locate the RNA polymerase at the start of the gene in the correct position to begin transcription. Attachment of the RNA polymerase at the TATA box is achieved with the help of a series of transcription factors specific to RNA polymerase II, referred to as TFIIA, TFIIB etc. These bind to the DNA around the TATA box and form a platform to which the RNA polymerase II is bound. The transcription factors bind in a specific order. TFIID binds first followed by TFIIA and TFIIB. The RNA polymerase II then binds followed by TFIIF, E, H and J to produce a functional complex capable of initiating transcription (Fig. 5).

Genes that lack a TATA box may contain an alternative initiator element around the transcription start site. Other gene promoters lack either a TATA box or an initiator element. These genes are usually transcribed at low levels and transcription initiation can occur at several different points. These genes often contain a GC-rich sequence 100-200 bp upstream of the transcription start site.

A number of other promoter elements with characteristic consensus sequences influence transcription. These sequences act as binding sites for other transcription factors that regulate transcription by stimulating or repressing
initiation. Examples include the CCAAT (pronounced CAT) box found upstream of the TATA box in some genes. Many genes also contain sequence called enhancers that greatly stimulate transcription. These are located outside the promoter often far away from the gene they influence and work independently of their orientation. Similar sequences known as silencers also occur which inhibit transcription.
Termination of transcription by RNA polymerase II occurs at a point after the end of the protein coding sequence by a mechanism which is uncertain. Termination signals are difficult to identify because the 3' end of the message is removed shortly after it is transcribed. It may be that definite termination signals do not exist but that the dissociation of a transcription factor, at some point, destabilizes the transcription complex causing the RNA polymerase to fall off the template at a later time.

Initiation of transcription by RNA polymerase I and III is similar to RNA polymerase II and involves specific promoter sequences that act as binding sites for the RNA polymerase and associated transcription factors (TIF-I for RNA polymerase I and TFIIIA-C for RNA polymerase III).

**RNA polymerase I** This enzyme transcribes genes encoding three of the four ribosomal RNAs (18S, 28S and 5.8S). The promoter recognized by RNA polymerase I has two important sequence elements required for efficient transcription: a core element overlapping the transcription start site and an upstream control sequence located approximately 100 bp upstream of the transcription start site. These sequences bind RNA polymerase I and its associated transcription factors. The core sequence is essential for transcription to occur and the upstream control sequence is involved in stimulating the rate of transcription initiation. A termination signal consisting of an 18 bp consensus sequence is present about 600 bp after the end of the gene.

**RNA polymerase III** This enzyme transcribes a set of short genes that encode transfer RNAs and the 5S ribosomal RNA. The promoter sequences recognized by RNA polymerase III are unusual in that they occur after the transcription start site within the transcribed sequence of the gene. These are known as internal control regions (ICRs) and they occur within about 100 bases of the transcription start site. The 5S rRNA gene contains a sequence known as the C box which acts as a binding site for transcription factors and RNA polymerase III. A second sequence upstream known as the A box is also important. For the transfer RNA genes the ICR is present as two highly conserved sequences known as the A box and the B box which together act as a binding site for transcription factors and the RNA polymerase III. Termination of transcription by RNA polymerase III occurs at a DNA sequence recognized by the enzyme which contains a run of A residues and which occurs soon after the end of the gene.
Section A – Molecular genetics

A5 Transfer RNA

Key Notes

Transfer RNAs (tRNAs) are small molecules that bring amino acids together for protein synthesis in an order specified by a messenger RNA (mRNA) sequence. Cells contain a number of different tRNAs each of which binds a specific amino acid. Each tRNA also binds a specific codon in the mRNA allowing it to place its amino acid in the correct position.

Base-pairing of tRNA molecules produces a cloverleaf structure composed of stem-loops called arms. These include: the acceptor arm which is the point of attachment of the amino acid; the anticodon arm which recognizes codons in the mRNA sequence; the DHU arm which contains dihydrouracil; the optional arm; and the TψC arm which contains pseudouracil. Transfer RNAs have conserved nucleotides that correspond to base paired regions. The tertiary structure predicts similar base pairing to the 2D doverleaf representation and shows the acceptor and anticodon arms at opposite ends of the molecule.

Transfer RNA genes occur as multiple copies each of which encodes several tRNAs. They are transcribed by RNA polymerase III as pre-tRNAs which are processed by ribonucleases to release mature tRNAs. RNaseP contains ribozyme activity. The sequence CCA is added to eukaryotic tRNAs after transcription.

The nucleotides in tRNA molecules are modified following transcription. Modifications include: methylation, base rearrangements, double bond saturation, deamination, sulfation and addition of larger groups. The function of the modifications is uncertain.

Related topics

Gene transcription (A4)  Translation (A8)
Ribosomal RNA (A6)

Role in translation

Cells contain three types of RNA – transfer, ribosomal and messenger – which are produced by transcription from DNA. Transfer and ribosomal RNAs form part of the machinery of protein synthesis and messenger RNAs act as the template for the synthesis of proteins during translation.

Transfer RNAs (tRNAs) are small molecules that act as adapters during protein synthesis; they link the nucleotide sequence of the messenger RNA (mRNA) to the amino acid sequence of the polypeptide. Cells contain a number of tRNAs each of which can bind a specific amino acid. Each tRNA recognizes a codon in the mRNA allowing it to place its amino acid in the correct position in the growing polypeptide chain as determined by the sequence of the mRNA.
Transfer RNA molecules contain between 74 and 95 nucleotides. Base-pairing takes place between complementary parts of the nucleotide sequence resulting in a structure known as the cloverleaf which is characteristic of tRNA molecules (Fig. 1). The cloverleaf is composed of a series of stem-loop structures known as arms. These include:

- The acceptor arm which is formed by base-pairing between nucleotides at the 5' and 3' ends of the tRNA. The sequence CCA, which occurs at the 3' terminus, is not base-paired and is the point of attachment for amino acids.
- The D or DHU arm is a stem-loop structure containing dihydrouracil, an unusual pyrimidine nucleotide.
- The anticodon arm is responsible for recognizing and binding codons in the mRNA.
- The extra, optional or variable arm occurs only in some tRNAs. It may be small containing only 2–3 nucleotides (class I tRNAs) or larger containing 13–21 nucleotides with up to five base pairs in a stem (class II tRNAs).
- The TψC arm contains the sequence TψC, where ψ is a modified nucleotide called pseudouracil.

Comparison of different tRNAs shows that parts of the nucleotide sequence are conserved. At certain positions the nucleotide present is invariant and is the same in all tRNAs. At other positions, the nucleotide is always a purine or a pyrimidine and is said to be semi-invariant and at other points the sequence is not conserved and the nucleotide present varies.

The cloverleaf structure is a two-dimensional description of the tRNA molecule. A more accurate representation is obtained from its three-dimensional (3D) structure which has been determined by X-ray diffraction (Fig. 2). The base-paired nucleotides described in the cloverleaf structure are still present but some nucleotides which appear far apart in the cloverleaf are base-paired in the 3D structure. Many of the nucleotides that are base-paired are invariant.
or semi-invariant. In the 3D structure the acceptor arm and the anticodon loop are at opposite ends of the molecule consistent with their role in translation.

**Synthesis and processing**

Transfer RNAs are synthesized by transcription of tRNA genes by the RNA polymerase III enzyme (see Topic A4). The genes exist as multiple copies, especially in eukaryotic cells, reflecting the large requirement of cells for tRNA. The tRNAs are produced as precursor RNA molecules called **pre-tRNAs** which are processed to give mature tRNAs (Fig. 3). Several tRNA genes may be transcribed together as a single pre-tRNA which is then processed by ribonucleases that cleave at the 5’ and 3’ ends of each tRNA sequence. In prokaryotes, processing is carried out in an ordered series of steps by the ribonucleases, RNaseD and P. RNaseP, which is found in prokaryotes and eukaryotes, is unusual in that it has an RNA component with catalytic activity known as a **ribozyme**. Eukaryotic tRNAs differ from their prokaryotic counterparts in that many of them are transcribed containing a short intron which is removed during processing. The sequence CCA is present at the 3’ terminus of all tRNAs and is the point of attachment for amino acids. In eukaryotes, CCA is not present in the DNA of the tRNA gene but is added later by a tRNA nucleotidyl transferase. In prokaryotes, the CCA is present in the coding sequence but is sometimes removed by RNaseD and then replaced by a prokaryotic nucleotidyl transferase.

**Modification of nucleotides**

Transfer RNAs contain unusual nucleotides produced after transcription by chemical modification. The most common modifications are:

- **Methylation** of the ribose sugar of the nucleotide. For example, guanosine is methylated to 7-methylguanosine.
- **Base rearrangements.** These involve interchanging of the positions of atoms in a purine or pyrimidine ring. An example is the conversion of uridine to pseudouridine.
- **Double-bond saturation.** An example is the conversion of uridine to dihydro-uridine.
- **Deamination.** This involves removal of amino groups from bases. For example, guanosine is deaminated to produce inosine.
- **Sulfur substitution.** For example, the oxygen atom of uridine is replaced by sulfur to give 4-thiouridine.
Section A - Molecular genetics

![Diagram of transcription and processing of tRNA molecules]

**Fig. 3.** Transcription and processing of tRNA molecules.

- **Addition of larger groups.** An example is the conversion of guanosine to queuosine.

Over 50 types of modification have been described, each carried out by a different tRNA modifying enzyme. The reasons for most of the modifications are unknown but in some cases roles have been assigned for modified nucleotides within the anticodon loop.
**Ribosomes**

Ribosomes are macromolecular structures composed of ribosomal RNA (rRNA) bound to protein. They occur in the cell cytoplasm where they bind to messenger RNA and translate it to produce proteins. Large numbers of ribosomes are required to fulfill the cell's requirement for protein. A typical bacterium contains 20,000 ribosomes which account for 80% of its RNA and 10% of its protein. Because ribosomes are very large, estimates of their molecular weight are difficult to obtain. The size of a ribosome is measured by its S value (Svedberg units) which is related to the rate at which it passes through a dense solution such as sucrose when centrifuged at high speed. The S value is determined by the size, shape and the macromolecular structure of the ribosome.

Each ribosome is composed of two parts called the large and small subunits (Fig. 1). In prokaryotes such as *E. coli*, the ribosome is 70S and is made up of 50 and 30S subunits (S values are not additive). The 50S subunit contains two rRNAs (23S and 5S) complexed with 31 polypeptides. The 30S subunit contains a single rRNA (16S) and 21 polypeptides. In eukaryotes, the ribosome is 80S.
Fig. 1. Composition of typical prokaryotic and eukaryotic ribosomes.

and is composed of 60S and 40S subunits. The 60S subunit contains three rRNAs (28S, 5.8S and 5S) and about 49 polypeptides; the 40S subunit contains one rRNA (18S) and about 33 proteins.

In ribosomes the rRNA molecules adopt a characteristic three-dimensional structure which is stabilized by complementary base-pairing both within and between RNA molecules. The RNA molecules are believed to form a framework to which the proteins, which provide most of the functional activity of the ribosome, are attached. Some of the RNA molecules also have enzymatic activity. These are known as ribozymes and they may contribute to the functioning of the ribosome.

Transcription and processing of rRNA genes in prokaryotes

Cells contain large numbers of ribosomes which must be replicated when the cell divides. As a result, cells have a huge requirement for rRNA which is produced by transcription of RNA genes by RNA polymerase. To ensure that correct numbers of each of the different rRNAs are produced, they are transcribed together from a single gene present in the genome as multiple copies. In prokaryotes such as E. coli, seven rRNA genes occur scattered throughout the genome. Each gene contains one copy each of the 16S, 23S and 5S rRNA sequences arranged consecutively. In addition, between one and four transfer RNA (tRNA) sequences are present in each gene. The gene is transcribed to produce a single RNA molecule called pre-rRNA (30S) which is processed to produce individual rRNAs and tRNAs (Fig. 2). Processing involves a series of defined steps. Following transcription, the RNA molecule folds and complementary parts of the sequence base-pair to give a series of stem-loop structures. The ribosomal proteins then bind to the folded RNA. At this stage some of the bases in the RNA are modified by the addition of methyl groups. Finally, the RNA is cleaved at specific points by the ribonuclease, RNAse III, to release the 5S, 23S and 16S rRNAs. Further trimming at the 5' and 3' ends by other ribonucleases called M5, M16 and M23 then yields mature rRNAs.
Fig. 2. Transcription and processing of rRNA genes in E. coli.

(a) Organization of eukaryotic rRNA genes.
(b) Transcription and processing of eukaryotic rRNA genes.
In eukaryotes, the sequences of the 28S, 18S and 5.8S rRNAs are present in a single gene which exists as multiple copies separated from each other by short nontranscribed regions (Fig. 3a). In humans, there are about 200 genes arranged as a series of five clusters of about 40 genes on separate chromosomes. The genes are transcribed by RNA polymerase I in the cell nucleus in a region known as the nucleolus. In humans, a single pre-rRNA (45S) is synthesized which is processed to give individual 28S, 18S and 5.8S rRNAs (Fig. 3b). The 5S rRNA is transcribed separately by the RNA polymerase III enzyme from unlinked genes as a short 121 base transcript which does not undergo processing. The eukaryotic pre-rRNA is processed in a similar way to its prokaryotic counterpart. Following transcription, the pre-rRNA folds and ribosomal proteins bind to it. Some of the bases are then modified by methylation of the ribose sugar. This reaction is catalyzed by molecules composed of RNA and protein called small nuclear ribonucleoproteins (snRNPs, pronounced snurps). Mature 28S, 18S and 5.8S rRNAs are then produced by a series of steps in which the pre-rRNA is cleaved by ribonucleases. Initial cleavage of the 45S pre-rRNA occurs in regions known as the external transcribed spacers (ETSs). This is followed by cleavage in regions known as the internal transcribed spacers (ITSs) to produce 20S and 32S precursor rRNAs. Further cleavage produces mature 28S, 18S and 5.8S rRNAs. In the final processing step the 5.8S rRNA base pairs with the 28S rRNA.
Messenger RNA (mRNA) acts as a template for protein synthesis. It is produced in the nucleus by transcription of protein coding genes by RNA polymerase II. The mRNA is initially transcribed as a precursor called pre-mRNA which contains noncoding intron sequences that are subsequently removed by splicing. The 5' end of the mRNA is modified by capping and the 3' end is modified by polyadenylation. RNA transcribed by RNA polymerase II exists in the nucleus as a population of molecules known as heterogeneous nuclear RNA (hnRNA).

This process involves the removal of introns from pre-mRNAs. The sequences GT and AG occur at the ends of introns and are part of larger 5' and 3' splicing signal sequences. Another signal sequence called the branchpoint sequence occurs in the intron. Splicing involves cleavage of the 5' end of the intron and its attachment to the branchpoint sequence to form a tailed loop. The intron is then released by being cleaved at its 3' end and the exons are brought together and joined. Splicing is catalyzed by small nuclear ribonucleoproteins (snRNPs): U1 binds to the 5' splice site and U2 to the branchpoint sequence; U5 and U4/6 then form a complex with U1 and U2 called the spliceosome which holds the mRNA in the correct orientation for splicing and provides the enzymatic activities required for excision of the intron and joining of the exons.

Eukaryotic mRNAs are modified at the 5' end by the addition of the modified nucleotide, 7-methylguanosine, in an unusual 5'→5' triphosphate linkage to the first nucleotide of the mRNA. This modification is known as capping and it protects the mRNA from degradation by 5' exonucleases.

Most eukaryotic mRNAs are modified at their 3' ends by the addition of a poly A tail (polyadenylation). The pre-mRNA is cleaved about 20 bases downstream of the polyadenylation signal sequence, 5' AAUAAA 3', and poly A polymerase adds a run of adenine residues. Polyadenylation is thought to protect the 3' end from degradation by exonucleases.

mRNA is relatively unstable compared with ribosomal and transfer RNA. This allows cells to regulate protein levels by altering the rate of gene transcription. Prokaryotic mRNAs have much shorter half-lives than eukaryotic mRNAs.

Variations in splicing patterns produce mRNAs with different sequences from a single pre-mRNA allowing production of variant proteins. Splicing patterns can vary to include or exclude one or more exons. Use of alternative polyadenylation signals also produces variant mRNAs. mRNA sequences can be altered by RNA editing which involves changing sequences by the insertion, deletion or substitution of individual bases.
In eukaryotes, messenger RNA (mRNA) produced by transcription of protein coding genes by the RNA polymerase II enzyme acts as a template for protein synthesis during translation. The coding information in eukaryotic genes is discontinuous and is arranged as a series of exons separated by noncoding introns. mRNA is synthesized as a precursor known as pre-mRNA by transcription of the exon and intron sequences. Before acting as a template for protein synthesis, the pre-mRNA undergoes a series of processing events to produce mature mRNA. Noncoding intron sequences are removed by a process called splicing which makes the coding sequences continuous and ensures that the mRNA is an accurate template for protein synthesis. In addition, the 5' end of the RNA is altered by the addition of a modified nucleotide in a process known as capping and the 3' end is modified by the addition of a tail of up to 250 adenines in a process called polyadenylation. RNA transcribed by RNA polymerase II exists in the nucleus as a population of molecules of different lengths (reflecting variations in gene size) and at different stages of processing and is known collectively as heterogeneous nuclear RNA (hnRNA).

In prokaryotes, mRNA is not processed and translation of the message begins even before transcription is complete. Prokaryotic genes do not normally contain introns and so splicing is unnecessary.

This process takes place in the nucleus and involves the removal of noncoding intron sequences from pre-mRNAs to produce mature mRNAs in which the coding sequences, corresponding to the exons, are continuous. The mature spliced mRNA is then exported to the cytoplasm where it acts as a template for protein synthesis.

Splicing depends on the presence of signal sequences in the pre-mRNA. In almost all genes the first two nucleotides at the 5' end of an intron are GT and the last two at the 3' end are AG. These are part of larger signal sequences present at the 5' and 3' ends of the introns. The complete 5' signal sequence is 5' AGGTAAGT 3' and the 3' sequence is 5' YYYYYYNCA 3' (Y = pyrimidine; N = any nucleotide). In addition, in vertebrates the sequence, 5' CURAY 3' (R = purine), which is called the branchpoint sequence, is present in the intron 10–40 bases upstream of the 3' signal sequence. A more specific sequence, 5' UACUAAC 3', occurs in introns of yeast. Splicing occurs in two steps (Fig. 1). In the first step the 2' hydroxyl group of the adenine of the branch-point sequence attacks the phosphodiester bond 5' to the G of the GT (5' splice site). The bond is broken releasing the 5' end of the intron and attaching it to the branchpoint sequence. The intron now forms a tailed loop structure called a lariat. In the second step the 3' end of the intron is cleaved after the G of the AG (3' splice site), the intron is released and the two exon sequences are joined together.

Splicing is catalyzed by a group of molecules called small nuclear ribonucleoproteins (snRNPs, pronounced snurps). These are composed of small RNA molecules rich in uracil called U RNAs or small nuclear RNAs (snRNAs) that exist complexed with proteins. Many different snRNPs exist but the most
Fig. 1. Splicing of pre-mRNA in eukaryotes.

abundant are U1, U2, U4, U5, and U6 which catalyze the splicing reaction. Each snRNP contains a single U RNA molecule except U4 and U6 which exist base-paired to each other in the same snRNP. The RNA components of the snRNPs interact by base-pairing with the splicing signal sequences. The U1 snRNP binds to the 5' splice site and the U2 snRNP binds to the branchpoint sequence. The remaining snRNPs, U5 and U4/U6, then form a complex with U1 and U2 causing the intron to loop out and the exons to be brought together. The combination of the pre-mRNA and the snRNPs is called the spliceosome and this is responsible for folding the pre-mRNA into the correct conformation for splicing (Fig. 2). The spliceosome also catalyzes the cutting and joining reactions that excise the intron and ligate the exons. Once splicing is completed the spliceosome dissociates. The functioning of the spliceosome is not fully understood and the components responsible for all of its enzymatic activities have not been identified.

Although almost all introns are spliced by a spliceosome, there are some examples of intron splicing which occur by different mechanisms. Introns in ribosomal RNA genes in some unicellular organisms can adopt a three-dimensional shape by base-pairing; this then acts as an RNA cutting enzyme, known as a ribozyme that contributes to its own splicing. Introns also occur in transfer RNA genes which are removed by the action of ribonucleases in a similar way to the processing of transfer RNA molecules.
**Section A - Molecular genetics**

**Fig. 2. Spliceosome formation.**

**Capping**

Eukaryotic mRNAs are altered at their 5’ end by a modification known as capping which involves addition of the modified nucleotide, 7-methylguanosine. The cap is added by the enzyme guanylyltransferase which joins GTP by an unusual 5’→5’ triphosphate linkage to the first nucleotide of the mRNA. Methyl transferase enzymes then add a –CH₃ group to the 7-nitrogen of the guanine ring and, usually, to the 2’ hydroxyl group on the ribose sugar of the next two nucleotides. Capping protects the mRNA from being degraded from the 5’ end by exonucleases in the cytoplasm and is also a signal allowing the ribosome to recognize the start of a mRNA molecule.

**Polyadenylation**

Most eukaryotic pre-mRNAs are modified at their 3’ ends by the addition of a sequence of up to 250 adenines known as a poly A tail. This modification is called polyadenylation and requires the presence of signal sequences in the pre-mRNA. These consist of the polyadenylation signal sequence, 5’ AAUAAA 3’, which occurs near the 3’ end of the pre-mRNA. The sequence YA (Y = pyrimidine) occurs in the next 11–20 bases and a GU rich sequence is often present further downstream. A number of specific proteins recognize and bind these signal sequences forming a complex which cleaves the mRNA about 20 nucleotides downstream of the 5’ AAUAAA 3’ sequence. The enzyme poly(A) polymerase then adds adenines to the 3’ end of the molecule. The purpose of the poly A tail is uncertain but it may serve to protect the mRNA from
degradation of the coding sequence at the 3' end by exonucleases. However, some mRNAs, notably those encoding histone proteins, have no poly A tail.

Unlike ribosomal and transfer RNAs which are stable within cells, mRNA is relatively short-lived. This is because cells regulate protein levels in the cytoplasm primarily by changing the rate of gene transcription. Because mRNAs are short-lived, changes in the rate of transcription of genes are reflected by changes in the amount of mRNA available for protein synthesis. In bacterial cells the half-life for a mRNA is just a few min. In eukaryotic cells a typical half-life might be as much as 6 h, although some mRNAs, such as those encoding the globin polypeptides that make up hemoglobin, are very long-lasting.

Variations can occur in the way pre-mRNAs are processed which generate different mRNAs and hence different proteins from a single gene sequence. This occurs by alternative splicing of the pre-mRNA in which cells vary the splice sites they use such that particular exons may be removed or retained during splicing. In addition, alternative polyadenylation signals present in the pre-mRNA can lead to the production of mRNAs with different sequences at the 3' end (Fig. 3). For example, the use of an upstream version of alternative polyadenylation sites may exclude exons downstream of it producing an mRNA encoding a truncated protein.

The same pre-mRNA may be alternatively processed within a single cell type, between different cell types and in the same cell type at different stages of development. The proteins produced following alternative processing of a pre-mRNA are related to each other but may have some different functions or characteristics. For example, alternative processing of immunoglobulin pre-mRNAs leads to the synthesis of proteins that may or may not contain

Fig. 3. Alternative splicing of pre-mRNA. (a) Exon skipping; (b) use of alternative polyadenylation sites.
hydrophobic amino acid sequences that allow them bind to cell membranes; this leads to the production of alternative membrane-bound and secreted forms of immunoglobulin proteins.

Pre-mRNAs may also undergo alternative processing by RNA editing. In this process the sequence of the pre-mRNA is altered by the insertion, deletion or substitution of bases. RNA editing was first identified in association with some parasitic protozoa in which the transcripts of many of the mitochondrial genes were found to be extensively modified by the insertion of uracil residues. Examples of RNA editing have also been identified in vertebrates, although the modifications found were much less extensive. In humans, the pre-mRNA of the apolipoprotein B gene is edited in intestinal cells by a C to U substitution that produces a stop codon; this leads to the synthesis of a shortened form of the protein. In liver cells, where editing does not occur, a full-length version of the protein is produced.
A8 TRANSLATION

Key Notes

Role of transfer RNA

Transfer RNAs (tRNAs) deliver amino acids to the ribosome for protein synthesis in an order specified by the messenger RNA (mRNA) sequence. Each amino acid is bound by one or more tRNAs (isoacceptors) which recognize codons for that amino acid. Amino acids are covalently attached (by aminoacylation) to the end of the acceptor arm of the tRNA by aminoacyl tRNA synthetase enzymes.

Codon recognition

Complementary base-pairing between a codon in the mRNA and the anticodon of a tRNA ensures that the amino acids are placed in the correct order during protein synthesis. The genetic code is degenerate; most amino acids are encoded by more than one codon. Each tRNA can recognize more than one codon specifying its amino acid because the 5' base in the anticodon can bind alternative 3' codon bases. This is known as wobble.

Translation

Translation is similar in prokaryotes and eukaryotes and occurs in three stages (initiation, elongation, and termination). Each stage involves a set of accessory proteins. Energy is provided by hydrolysis of adenosine triphosphate (ATP) and guanosine triphosphate (GTP).

Initiation

Translation begins with binding of the small ribosomal subunit to the mRNA at the Shine-Dalgarno sequence in prokaryotes and the 5' cap in eukaryotes. The subunit migrates downstream to the AUG initiation codon and the initiator tRNA methionine (tRNA Met) binds to form the initiation complex. In prokaryotes, the methionine of the tRNA Met is formylated. IF1, 2, and 3 are bacterial initiation factors. IF1 and IF3 prevent binding of the large ribosomal subunit before initiation is complete. IF2 brings the tRNA Met to the initiation complex. In eukaryotes at least nine initiation factors are involved. eIF1 and eIF2 have similar roles to IF1 and IF2. Several factors remove mRNA secondary structure.

Elongation

Following initiation the large ribosomal subunit binds the initiation complex forming the A and P sites. The P site is occupied by tRNA Met. A second charged tRNA enters the A site and peptidyltransferase forms a peptide bond between the two amino acids. tRNA deacylase breaks the bond between methionine and its tRNA leaving a dipeptide bound to the second tRNA. In prokaryotes an elongation factor, EF-Tu, is associated with entry of the tRNA to the A site. GTP is hydrolyzed and EF-Tu is released attached to guanosine diphosphate (GDP). EF-Ts regenerates EF-Tu. In eukaryotes eEF1 has a similar role to EF-Tu. Following peptide bond formation the ribosome translocates to the next codon, the dipeptide bound to the second tRNA moves to the P site expelling the initiator tRNA. A third charged tRNA enters the A site and the elongation cycle repeats. In prokaryotes translocation is mediated by EF-G and requires hydrolysis of GTP. In eukaryotes eEF2 has a similar function.
Translation ends when a termination codon enters the A site. Release factors enter the A site and cause release of the polypeptide. In *E. coli* RF 1, 2 and 3 cause termination. In eukaryotes a single protein, eRF, is involved. Following termination the ribosome dissociates and the mRNA is released.

Following translation, polypeptides may be modified by the addition of chemical groups to amino acid side chains and the N and C termini or by proteolytic cleavage. Modifications may be required for full functional activity.

**Related topics**
The genetic code (A3)  
Transfer RNA (A5)  
Ribosomal RNA (A6)

**Role of transfer RNA**
Translation is the process by which cells synthesize proteins. During translation information encoded in mRNA molecules is used to specify the amino acid sequence of a protein. Transfer RNA molecules play a key role in this process by delivering amino acids to the ribosome in an order specified by the mRNA sequence; this ensures that the amino acids are joined together in the correct order (Fig. 1). Cells usually contain between 31 and 40 individual species of tRNA, each of which binds specifically to one of the 20 amino acids. Consequently, there may be more than one tRNA for each amino acid. Transfer RNAs that bind the same amino acid are called *isoacceptors*. Before translation begins, amino acids become covalently linked to their tRNAs which then recognize codons in the mRNA specifying that amino acid. The attachment of an amino acid to its tRNA is called *aminoacylation* or charging. The amino acid is covalently attached to the end of the acceptor arm of the tRNA which always

---

Fig. 1. Role of tRNAs in translation.
ends with the base sequence 5’ CCA 3’. A bond forms between the carboxyl group of the amino acid and the 3’-hydroxyl of the terminal adenine of the acceptor arm. Charging is catalyzed by enzymes called **aminoacyl tRNA synthetases** in a reaction requiring the hydrolysis of ATP. A separate enzyme exists for each amino acid and each enzyme can charge all the isoacceptor tRNAs for that amino acid. The aminoacyl tRNA synthetase recognizes both its appropriate amino acid and the corresponding tRNA. The amino acid is recognized primarily by its side chain. It is not clear exactly how the tRNA is recognized but variant nucleotides that are specific for individual tRNAs may be involved.

When the correct amino acid has been attached to the tRNA it recognizes the codon for that amino acid in the mRNA allowing it to place the amino acid in the correct position, as specified by the sequence of the mRNA. This ensures that the amino acid sequence encoded by the mRNA is translated faithfully. Codon recognition takes place via the anticodon loop of the tRNA and specifically by three nucleotides in the loop known as the **anticodon** which binds to the codon by complementary base-pairing. The four bases present in DNA can combine as 64 codons. Three codons act as signals for translation to stop and the remaining 61 encode the 20 amino acids present in proteins. Consequently, most amino acids are represented by more than one codon, a feature referred to as the **degeneracy of the genetic code**. To deal with this, individual tRNAs can recognize more than one of the codons for their amino acid. They can achieve this because the anticodon is capable of binding to alternative bases present at the third position of the codon, a feature known as **third base degeneracy or wobble**. Binding between the codon and the anticodon can tolerate variation at the third base because the anticodon loop is not linear and when the anticodon binds to the codon it does not form a perfect RNA double helix. This permits the formation of a few nonstandard base pairs. For example, G can base-pair with U as well as with C at the wobble position, and inosine, a deaminated form of guanine sometimes present in the wobble position of the anticodon loop, can base-pair with C but also with A and U. Wobble allows a single tRNA to decode more than one member of a codon family decreasing the number of tRNAs required by the cell. The rules of the genetic code however are not violated and polypeptides are synthesized strictly in accordance with the nucleotide sequence of the mRNA. An important consequence of wobble is that it serves to minimize the effects of mutations.

**Translation**

Translation occurs by similar mechanisms in prokaryotes and eukaryotes and is conveniently described as occurring in three stages: **initiation**, **elongation** and **termination**. Initiation involves binding of a ribosome to mRNA. Elongation involves repeated addition of amino acids and termination involves release of the new polypeptide chain. Sets of **accessory proteins** assist the ribosome in each of the three stages. Translation requires the use of energy by the cell which is provided by the hydrolysis of guanosine triphosphate (GTP) and adenosine triphosphate (ATP). GTP is used for ribosome movement and in binding of accessory factors. ATP is used to charge tRNAs and in removing secondary structure from mRNA. Up to 90% of ATP produced in a bacterium is used for translation.

**Initiation**

When they are not actively involved in translation ribosomes exist as separate large and small subunits. The first step in translation involves the binding of
the small ribosomal subunit to the mRNA (Fig. 2). Translation usually begins at the sequence AUG (bacteria sometimes use GUG or UUG) which encodes methionine and is known as the translation initiation codon. The small subunit binds to the mRNA at a specific point upstream of the AUG. In prokaryotes, this is the Shine–Dalgarno sequence (5' AGGAGGU 3') found near the start of the mRNA. Once bound, the small subunit migrates in a 3' direction.
along the mRNA until it finds the AUG, usually about 10 nucleotides downstream. In eukaryotes, the small ribosomal subunit recognizes the cap structure at the 5' end of the mRNA. It then moves downstream until it encounters the first AUG, although sometimes other AUGs are recognized. A tRNA charged with methionine binds to the AUG located by the small ribosomal subunit. In bacteria, the methionine is modified by the addition of a formyl group (-CHO) to one of the hydrogens of the amino group (tRNA\textsuperscript{met}). This blocks the amino group, thus preventing it from forming a peptide bond and ensuring that polymerization of the polypeptide can only occur in the amino to carboxy direction. The combination of the mRNA, the small ribosomal subunit and the tRNA\textsuperscript{met} is called the initiation complex. Two tRNAs recognize AUG and carry methionine. One is used for initiation (tRNA\textsuperscript{met}) and the other recognizes internal AUGs. Only the initiator tRNA is capable of binding to the initiation complex.

A number of accessory proteins called initiation factors are required for initiation. Bacteria have three, known as IF1, IF2, and IF3. Initiation begins with binding of IF1 and IF3 to the small ribosomal subunit. This helps to prevent binding of a large subunit before the mRNA has bound. Next, IF2 complexed with GTP binds the small subunit. Its purpose is to assist binding of the initiator tRNA. The small subunit then binds the mRNA and locates the AUG initiation codon. The initiator tRNA charged with methionine binds to the complex and IF3 is released. This marks the end of initiation. Elongation begins with binding of a large ribosomal subunit to the initiation complex to form a complete ribosome. This is accompanied by the release of IF1 and IF2 and hydrolysis of GTP (Fig. 2).

Initiation in eukaryotes is similar to that in prokaryotes; however, the initiation codon is almost always AUG and the methionine is not formylated. Many more initiation factors (at least nine) are involved, some have roles analogous to those of their bacterial counterparts. Two eukaryotic factors, eIF2 and eIF3 have roles similar to bacterial IF2 and IF3. Several of the eukaryotic initiation factors are involved in removing secondary structure from the mRNA before it is translated. Energy for this is provided by hydrolysis of ATP.

### Elongation

As soon as the initiation complex has formed, a large ribosomal subunit binds to it. The complete ribosome contains two binding sites for tRNA molecules (Fig. 3). The first site is the P or peptidyl site and is occupied by the tRNA\textsuperscript{met} base-paired to the AUG. The second site is the A or aminoacyl site and is positioned over the second codon. Elongation begins when a tRNA enters the A site and base pairs with the second codon. With both sites occupied by charged tRNAs, the attached amino acids are placed in close contact and a peptide bond can form between the carboxyl group of the methionine and the amino group of the second amino acid. The reaction is catalyzed by a complex enzyme called peptidyl transferase which probably contains several different ribosomal proteins. Peptidyl transferase works in conjunction with another enzyme, tRNA deacylase, which breaks the link between methionine and its tRNA after formation of the peptide bond.

A number of accessory proteins are required for elongation. In prokaryotes two elongation factors, EF-Tu and EF-Ts, are involved. EF-Tu is associated with entry of a tRNA into the A site. EF-Tu binds charged tRNAs in association with GTP. Following entry to the A site, the GTP is hydrolyzed and EF-Tu is released bound to guanosine diphosphate (GDP\textsuperscript{3}). Before another tRNA can bind, EF-Tu
must be regenerated with the help of EF-Ts. First, EF-Ts displaces GDP by binding to EF-Tu; a new molecule of GTP then replaces EF-Ts (Fig. 4). In eukaryotes, a complex protein called eEF-1 brings the tRNA to the A site. Again, the reaction is associated with hydrolysis of GTP. The details of how eEF-1 is regenerated are unknown but components equivalent to EF-Tu and EF-Ts may be present.
Following peptide bond formation translocation occurs and the ribosome moves on to the next codon. The newly formed dipeptide which is bound to the second tRNA (aa-aa-tRNA) moves into the P site expelling the uncharged tRNA and the A site becomes vacant. A third charged tRNA enters the A site and the elongation cycle is repeated. Following each addition of an amino acid to the growing polypeptide chain the ribosome translocates to the next codon. In bacteria, translocation is mediated by the elongation factor EF-G which binds to the ribosome in a complex with GTP which is then hydrolyzed providing energy for translocation. Binding of EF-G and EF-Tu to the ribosome is mutually exclusive. This ensures that translocation is completed before the next round of elongation begins. The eukaryotic equivalent of EF-G is eEF-2 which also requires GTP and functions in a similar way. As translation proceeds the ribosome moves along the mRNA away from the initiation site which is then free to bind a new ribosome. Thus mRNAs can be translated by several ribosomes at once forming a structure known as a polysome.

**Termination**

Translation ends when a termination codon enters the A site (Fig. 5). There are no tRNAs able to bind to the termination codons. Instead, proteins known as release factors enter the A site and cause the completed polypeptide to be released. In *E. coli* two release factors, RF1 and RF2, perform this function. RF1 recognizes UAA and UAG stop codons and RF2 recognizes UAA and UGA. A third release factor, RF3, plays an ancillary role in the process. The release factors cause peptidyl transferase to transfer the polypeptide to a water molecule rather than to another aminoacyl tRNA.

In eukaryotes a single protein, eRF, is involved which requires GTP for ribosome binding. This is subsequently hydrolyzed and eRF is released from the ribosome. Following release of the polypeptide, the ribosome releases the mRNA and dissociates; it then joins the pool of ribosomal subunits before becoming involved in further translation.

**Post-translational modifications**

Following translation, newly synthesized polypeptides may undergo a range of modifications before becoming functional proteins. These involve mainly the covalent attachment of chemical groups and cleavage of the polypeptide chain. Many different chemical modifications of the side chains of amino acids or the amino and carboxyl termini of proteins are found. Modifications may involve addition of small groups, such as methylation, phosphorylation, acetylation, and hydroxylation as well as the addition of larger molecular structures such as lipids and oligosaccharides (glycosylation). Some modifications such as phosphorylation regulate enzyme activity. Cleavage of polypeptide chains is a very common modification; this may involve trimming of amino acids from the termini of the
proteins, removal of internal peptides, removal of amino-terminal signal peptide sequences from secreted proteins and the cleavage of polyproteins into smaller peptides. Cleavage of proteins is sometimes associated with the activation of inactive precursor proteins.
DNA replication

This is the process by which a cell copies its DNA. Replication is necessary so that the genetic information present in cells can be passed on to daughter cells following cell division. The DNA is copied by enzymes called DNA polymerases. These act on single-stranded DNA synthesizing a new strand complementary to the original strand. DNA synthesis always occurs in the 5′→3′ direction.
Replication is said to be semi-conservative (Fig. 1). This means that each copied DNA molecule contains one strand derived from the parent molecule and one newly synthesized strand.

The mechanism of DNA replication is very similar in most organisms. Differences exist only with respect to the enzymes and proteins involved. In prokaryotes such as *E. coli*, two enzymes, DNA polymerases I and III, are responsible for DNA synthesis. In eukaryotes, DNA is replicated by five DNA polymerases (α, β, γ, δ, ε). Replication needs to be very accurate because even a small error rate would result in the loss of important genetic information after just a few cell divisions. Accuracy is ensured by the ability of the DNA polymerases to check that the correct bases have been inserted in the newly synthesized strand. This is achieved through the reverse (3′→5′) exonuclease activity of the enzymes which allows them to remove incorrectly inserted bases from newly synthesized DNA and replace them with the correct base. This is referred to as proofreading ability. It is estimated that just one base in five billion is inserted incorrectly.

During DNA replication the double helix of a cell’s entire DNA is progressively unwound producing segments of single-stranded DNA which can be copied by DNA polymerases. Unwinding of the double helix begins at a distinct position called the replication origin and gradually progresses along the molecule, usually in both directions. Replication origins usually contain sequences rich in weak A-T base pairs. The region where the helix unwinds and new DNA is synthesized is called the replication fork (Fig. 2). At the replication fork a number of distinct events occur:

- **Separation of the double helix.** This is achieved by the action of a helicase enzyme. Following separation of the strands, single-strand binding (SSB) protein attaches to the DNA and prevents the double helix from reforming.

- **Synthesis of leading and lagging strands.** Synthesis of DNA by DNA polymerases occurs only in the 5′→3′ direction. As the two strands of the double helix run in opposite directions (one strand runs 5′→3′ and the other 3′→5′) slightly different mechanisms are required to replicate each. One strand, called the leading strand, is copied in the same direction as the unwinding helix and so can be synthesized continuously (Fig. 3a). The other strand, known as the lagging strand, is synthesized in the opposite direction and must be copied discontinuously. The lagging strand is synthesized as a series of segments known as Okazaki fragments (Fig. 3b).

- **Priming.** DNA polymerases require a short double-stranded region to initiate or prime DNA synthesis. This is produced by an RNA polymerase, called primase, which is able to initiate synthesis on single-stranded DNA. The
Fig. 2. Replication origin and replication forks.

(a) Direction of unwinding of helix

DNA is synthesized continuously on the leading strand as the helix unwinds

(b) Direction of unwinding of helix

Helix unwinds further

DNA is synthesized in segments on lagging strand

Fig. 3. (a) Replication of leading strand. (b) Replication of lagging strand.
primase synthesizes a short RNA primer sequence on the DNA template creating a short double-stranded region. In *E. coli*, DNA polymerase III then synthesizes DNA beginning at the RNA primer. On the lagging strand, synthesis ends when the next RNA primer is encountered. At this point DNA polymerase I takes over and removes the RNA primer replacing it with DNA. In eukaryotes the situation is different. DNA polymerase α, which has integral primase activity, is responsible for initiating DNA synthesis. DNA is replicated by DNA polymerases α and δ with α synthesizing the lagging strand and δ synthesizing the leading strand. The other polymerases have ancillary roles. DNA polymerase ε is involved in DNA repair and DNA polymerase γ replicates mitochondrial DNA.

- **Ligation.** The final step required to complete synthesis of the lagging strand is for the Okazaki fragments to be joined together by phosphodiester bonds. This is carried out by a DNA ligase enzyme (Fig. 4).

**Teratogenic DNA replication**

Although the mechanism of DNA replication is similar in all organisms, the overall process varies depending on the nature of the DNA molecule being copied. Different strategies are required for replication of the circular DNA molecules which occur in bacteria and for the linear chromosomal DNA molecules present in eukaryotes. The simplest and most common form of replication for circular DNA involves a single origin of replication from which two replication forks progress in opposite directions. This results in an intermediate θ form (Fig. 5). The replication forks eventually meet and fuse and replication terminates.

The replication of DNA molecules requires unwinding of the DNA double helix. This causes the helix ahead of the replication fork to rotate. For circular DNA molecules that do not have free ends, this produces supercoiling of the DNA preventing the replication fork from progressing. This problem is overcome by the action of enzymes called topoisomerases of which there are two types. DNA topoisomerase I produces a transient break in the polynucleotide backbone of one of the DNA strands a short distance ahead of the replication fork allowing the DNA to rotate freely around the other intact strand removing the supercoiling. The enzyme then rejoins the ends of the broken strand. When replication of a bacterial chromo-
Fig. 5. Replication of circular DNA molecules via theta intermediate.

DNA replication in eukaryotes

Before a cell can divide it must replicate its DNA. Cell division in eukaryotes is highly regulated and occurs as a series of phases known as the cell cycle (Fig. 6). The length of the cell cycle varies but is typically several hours. The longest phase is \( G_1 \) during which the cells prepare for division. \( G_1 \) is followed by the \( S \) phase, in which replication of the DNA occurs. A second short gap, \( G_2 \), is next and is followed by the \( M \) phase during which the cells undergo mitosis involving separation of the chromosomes and cell division. After \( M \) phase, proliferating cells enter the \( G_1 \) phase of the next cell cycle. Alternatively, cells may exit the cell cycle by entering the \( G_0 \) phase where they remain quiescent for extended periods. Some cells such as neurons stop dividing completely and are permanently in \( G_0 \) phase.

Due to the extreme length of eukaryotic chromosomes, DNA replication must be initiated at multiple origins to ensure that the process is completed within a reasonable time span. Replication forks proceed in either direction from each replication origin forming replication bubbles which eventually meet and merge. DNA replicated from a single origin is called a replicon. A typical mammalian cell has 50–100 000 replicons, each of which replicates 40–200 kb of DNA. Not all the DNA is replicated at once. Clusters of about 50 replicons initiate
simultaneously at defined points during S phase. Areas containing transcriptionally active genes are replicated first with nonactive regions replicated later. The DNA in eukaryotic chromosomes is packaged as DNA–protein complexes called **nucleosomes**. As the replication fork progresses DNA must unwind the nucleosome for replication to occur. This slows the progress of the replication forks and may explain the short length of the Okazaki fragments on the lagging strand in eukaryotes (100–200 bases) compared with prokaryotes (1000–2000 bases). After the replication fork has passed the nucleosomes re-form.

Replication of linear eukaryotic chromosomes poses a problem not encountered with circular bacterial chromosomes in that the extreme 5' end of the lagging strand cannot be replicated because there is no room for an RNA primer to initiate replication. This creates the potential for chromosomes to shorten after each round of replication leading to a loss of genetic information. The problem is overcome by a specialized structure at the end of the chromosome known as the **telomere** which contains tandem (side-by-side) repeats of a simple noncoding sequence. In humans this is 5' TTAGGG 3'. In addition, the 3' end of the leading strand extends beyond the 5' end of the lagging strand. The enzyme **telomerase** contains an RNA molecule which partly overlaps with the ends of the leading strand extends beyond the 5' end of the lagging strand. The enzyme then extends the leading strand using the RNA as a template. The telomerase then dissociates and binds to the repeat sequence on the leading strand. The enzyme then extends the leading strand again. This process of extension may occur hundreds of times before the telomerase finally dissociates. The newly extended leading strand then acts as a template for replication of the 5' end of the lagging strand (Fig. 7). The two processes whereby the DNA is shortened during normal replication and lengthened by the action of the telomerase are roughly balanced so the overall length of the chromosome remains approximately the same.

**Fig. 7.** Replication of DNA at telomeres.
**Regulation of Gene Expression in Prokaryotes**

**Key Notes**

**Regulation of gene expression**
Bacteria regulate the activity of their genes so that only gene products necessary for the cell's functions are produced. This allows bacteria to respond to environmental changes. Alterations in the amount of gene product may potentially be achieved by varying rates of transcription, mRNA turnover, mRNA processing and translation. Of these, the mechanisms that alter gene transcription are the best characterized.

**Organization of bacterial genes**
Many bacterial genes are arranged as coordinately regulated operons that encode proteins with related functions. Inducible operons such as the lac operon encode enzymes involved in metabolic pathways and are induced by the substrate for the pathway. Repressible operons such as the trp operon encode enzymes involved in biosynthetic pathways. Gene expression is regulated by the pathway end product or by attenuation.

**The lac operon**
This operon contains three genes (lac Z, Y, A) encoding enzymes required by *E. coli* to metabolize lactose. The genes are transcribed from a single promoter and their expression is induced by lactose. In the presence of lactose, allolactose binds the lac repressor preventing it from binding the lac operator and allowing the operon to be transcribed. When lactose has been used up the lac repressor regains its ability to bind the lac operator and transcription is blocked.

**Catabolite repression**
This regulatory mechanism allows *E. coli* to repress the lac operon in the presence of glucose. Catabolite activator protein (CAP) binds cAMP and stimulates transcription of the lac operon by binding upstream of the lac promoter. Levels of cAMP are regulated by glucose which inhibits adenylate cyclase. When glucose is available, cAMP levels are low, CAP fails to bind the lac promoter and the operon is transcribed at a low level. When glucose levels are low, cAMP levels rise, CAP binds the lac promoter and stimulates transcription of the operon. Catabolite repression ensures that when glucose and lactose are both available, glucose is used first.

**The trp operon**
This operon contains five genes transcribed from a single promoter encoding enzymes required for the biosynthesis of tryptophan. The trp repressor binds the trp operator in the presence of tryptophan and blocks transcription of the operon. In the absence of tryptophan, the trp repressor fails to bind and transcription of the operon proceeds.
This regulatory mechanism allows fine adjustment of expression of the \textit{trp} operon and other operons. DNA sequences between the \textit{trp} promoter and the first \textit{trp} operon gene are capable of forming either a large stem–loop structure that does not influence transcription or a smaller terminator loop. A short coding region upstream contains tryptophan codons. When tryptophan levels are adequate RNA polymerase transcribes the region closely followed by a ribosome which prevents formation of the larger stem–loop, allowing the terminator loop to form ending transcription. If tryptophan is lacking, the ribosome is stalled, the RNA polymerase moves ahead and the large stem–loop forms. Formation of the terminator loop is blocked and transcription of the operon proceeds.

This mechanism is used to make major alterations to gene expression in response to environmental changes. Alternative $\sigma$ factors alter the specificity of bacterial RNA polymerase allowing it to recognize different gene promoters. Alternative $\sigma$ factors activate gene transcription in \textit{E. coli} in response to heat shock and in \textit{Bacillus subtilis} during sporulation. Bacteriophages synthesize $\sigma$ factors that direct the transcription of bacteriophage genes.

**Related topics**
- Genes (A2)
- Bacteriophages (B8)
- Prokaryotic genomes (B3)

For a bacterium to function it is not necessary that all of its genes are transcribed at all times. To conserve energy and resources bacteria regulate the activity of their genes so that only those gene products necessary for the cell’s functions are produced. For example, it would be wasteful for a bacterium to produce enzymes required to synthesize an amino acid that was already available to it from its environment. Regulation of gene expression allows bacteria to respond to changes in their environment, typically to the presence or absence of nutrients.

Bacteria regulate expression of their genes in order to control the amount of gene product present. The steady state concentration of a gene product is determined by the balance between the rate of synthesis and the rate of degradation of the expressed protein. In practice, changes in the rate of synthesis are what alter the amount of gene product. The rate of synthesis could potentially be altered by a number of factors:

- changes in the rate of transcription;
- changes in mRNA turnover time;
- changes in the rate of translation.

In practice, all three mechanisms probably influence gene expression but the best understood examples are those involving the regulation of gene transcription.

An important feature which determines how gene transcription is regulated in bacteria is the organization of the genes as operons. These are transcriptional units in which several genes, usually encoding proteins with related functions, are regulated together. Other genes also occur which encode regulatory proteins that
control gene expression in operons. Many different operons have been identified in *E. coli*. Most contain genes that encode proteins involved in the biosynthesis of amino acids or the metabolism of nutrients. Operons are classified as inducible or repressible. Inducible operons contain genes that encode enzymes involved in metabolic pathways. Expression of the genes is controlled by a substrate of the pathway. An example of an inducible operon is the lac operon which encodes enzymes required for the metabolism of lactose. Repressible operons contain genes that encode enzymes involved in biosynthetic pathways and gene expression is controlled by the end product of the pathway which may repress expression of the operon or control it by an alternative mechanism called attenuation. An example of a repressible operon is the trp operon which encodes enzymes involved in the biosynthesis of tryptophan.

**The lac operon**

This operon contains three genes encoding enzymes required by the *E. coli* bacterium for the utilization of the disaccharide sugar lactose. These are lactose permease, which transports lactose into the cell, β-galactosidase which hydrolyzes lactose into its component sugars (glucose and galactose) and β-galactoside transacetylase which is also involved in the hydrolysis of lactose. These enzymes are normally present in *E. coli* at very low levels but in the presence of lactose their levels rise rapidly. The three genes in the lac operon are known as lac Z, Y and A and encode β-galactosidase, lactose permease and β-galactoside transacetylase, respectively. The genes are sequential and are transcribed as a single mRNA from a single promoter. Another regulatory gene, lac I, which is expressed separately, lies upstream of the operon and encodes a protein called the lac repressor which regulates the expression of the lac Z, Y and A genes. In the absence of lactose the lac repressor binds to a DNA sequence called the operator positioned between the lac promoter and the beginning of the lac Z gene. When bound to the operator, the lac repressor blocks the path of the RNA polymerase bound to the lac promoter upstream of it and prevents transcription of the lac genes. When the cell encounters lactose a few molecules of the lac enzymes present in the cell allow lactose to be taken up and metabolized. Allolactose, an isomer of lactose produced as an intermediate during the metabolism of lactose, acts as an inducer. It binds to the lactose repressor and changes its conformation such that it can no longer bind to the operator. The path of the RNA polymerase is no longer blocked and the operon is transcribed. Large numbers of enzyme molecules are produced which take up lactose and metabolize it. The presence of lactose thus induces the expression of the enzymes needed to metabolize it. When the lactose is used up the lac repressor returns to its original conformation and again, binds the lac operator preventing transcription and switching off the operon (Fig. 1).

**Catabolite repression**

This term describes an additional regulatory mechanism which allows the lac operon to sense the presence of glucose, an alternative and preferred energy source to lactose. If glucose and lactose are both present, cells will use up the glucose first and will not expend energy splitting lactose into its component sugars. The presence of glucose in the cell switches off the lac operon by a mechanism called catabolite repression which involves a regulatory protein called the catabolite activator protein (CAP) (Fig. 2). CAP binds to a DNA sequence upstream of the lac promoter and enhances binding of the RNA polymerase leading to enhanced transcription of the operon. However CAP only binds in the presence of a derivative of ATP called cyclic adenosine monophosphate (cAMP).
Fig. 1. Regulation of lac operon (a) in the absence and (b) in the presence of inducer.

Fig. 2. Catabolite repression of lac operon.
whose levels are influenced by glucose. The enzyme adenylate cyclase catalyzes the formation of cAMP and is inhibited by glucose. When glucose is available to the cell adenylate cyclase is inhibited and cAMP levels are low. Under these conditions CAP does not bind upstream of the promoter and the lac operon is transcribed at a very low level. Conversely, when glucose is low adenylate cyclase is not inhibited, cAMP is higher and CAP binds increasing the level of transcription from the operon. If glucose and lactose are present together the lac operon will only be transcribed at a low level. However when the glucose is used up catabolite repression will end and transcription from the lac operon increases allowing the available lactose to be used up.

The trp operon

This operon contains five genes encoding enzymes involved in biosynthesis of the amino acid tryptophan. The genes are expressed as a single mRNA transcribed from an upstream promoter. Expression of the operon is regulated by the level of tryptophan in the cell (Fig. 3). A regulatory gene upstream of the trp operon encodes a protein called the trp repressor. This protein binds a DNA sequence called the trp operator which lies just downstream of the trp promoter partly overlapping it. When tryptophan is present in the cell it binds to the trp repressor protein enabling it to bind the trp operator sequence, obstructing binding of the RNA polymerase to the trp promoter and preventing transcription of the operon. In the absence of tryptophan the trp repressor is incapable of binding the trp operator and transcription of the operon proceeds. Tryptophan, the end product of the enzymes encoded by the trp operon, thus acts as a corepressor with the trp repressor protein and inhibits its own synthesis by end product inhibition.

---

**Fig. 3. The trp operon.**

The trp operon makes use of an alternative strategy for controlling transcription called attenuation which can finely tune expression levels (Fig. 4). The transcribed mRNA sequence between the trp promoter and the first trp gene is capable of forming two stem-loop structures. The relative positions of the sequences mean that both stem-loops cannot form at once: just one or the other may be present at any time. The larger, more stable structure does not influence transcription and occurs upstream of the smaller stem-loop which acts as
Fig. 4. Transcriptional attenuation in the trp operon.

a transcription terminator. If this structure forms it will terminate transcription before the first gene is reached eliminating gene expression. Attenuation depends on the fact that transcription and translation are linked in bacteria: ribosomes attach to mRNAs as they are being synthesized and begin translating them into protein. An mRNA that is being transcribed may already have one or more ribosomes attached to it. Binding of ribosomes to the trp mRNA influences which of the two stem-loop structures can form and so determines whether or not termination occurs. Immediately upstream of the stem-loop region is a short open reading frame containing 14 codons followed by a stop codon which is translated before the structural genes; two out of these 14 codons are for tryptophan. If levels of tryptophan are adequate the ribosome will translate the coding region following closely behind the RNA polymerase. In these circumstances the presence of the ribosome prevents formation of the larger stem-loop allowing the terminator loop to form and transcription ends. If tryptophan is lacking, the ribosome will be stalled as it translates the coding region.
The RNA polymerase will move ahead and the first stem-loop will be free to form. Formation of the terminator loop is then blocked and transcription of the operon can proceed. The speed at which the ribosome translates the coding region will not be the same for each transcript. When tryptophan is present in the medium at intermediate levels some transcripts will terminate and others will not, thus allowing fine adjustments in the levels of transcription of the operon. Overall, the \textit{trp} repressor determines whether the operon is switched on or off and attenuation determines how efficiently it is transcribed, both depend on the level of tryptophan in the cell. Attenuation allows the cell to synthesize tryptophan according to its exact requirements. Attenuation is not restricted to the \textit{trp} operon and occurs in at least six other operons that encode amino acid biosynthetic enzymes. Some operons such as the \textit{trp} and \textit{phe} operons are regulated by repressors and attenuation and others such as the \textit{his}, \textit{leu} and \textit{thr} operons rely only on attenuation.

Bacterial RNA polymerase is composed of five individual polypeptide subunits \((\alpha_2, \beta, \beta', \sigma)\). One of the subunits called \textit{sigma} (\(\sigma\) factor) is responsible for initiating transcription by recognizing bacterial promoter DNA sequences. Bacteria, including \textit{E. coli}, make alternative sigma factors that recognize different sets of promoters and cause the RNA polymerase to transcribe different sets of genes. This is used as a way of regulating gene expression where environmental conditions dictate major alterations in the pattern of gene expression. The \(\sigma^70\) factor is the most common factor used by \textit{E. coli}. Alternative \(\sigma\) factors come into play in a variety of situations:

- **Heat shock.** When exposed to increased temperature \textit{E. coli} transcribes a set of 17 proteins which help the cell to adapt to the altered environmental conditions. An alternative factor \(\sigma^{32}\) is produced that recognizes promoters of heat shock genes.

- **Sporulation in \textit{Bacillus subtilis}.** This bacterium undergoes sporulation in response to adverse environmental conditions. Sporulation requires drastic changes in gene expression involving a shut down of most protein synthesis and the production of proteins needed for resumption of protein synthesis when the spore germinates. \textit{B. subtilis} achieves this change by the use of alternative \(\sigma\) factors.

- **Bacteriophage \(\sigma\) factors.** Some phages make use of the host cell RNA polymerase and supply it with \(\sigma\) factors to instruct it to transcribe phage genes preferentially (see Topic B8). Other phages produce a series of \(\sigma\) factors that allow temporal control of their own gene expression with early genes transcribed by the host polymerase containing a \(\sigma\) factor that directs transcription of later genes.
A11 REGULATION OF GENE EXPRESSION IN EUKARYOTES

Key Notes

Gene expression

Genes in eukaryotic cells are subject to complex patterns of regulation. Cells express only about 15% of their genes with different genes expressed by different cell types. The pattern of gene expression determines the characteristics of a cell and its role in the organism. Changes in the pattern of gene expression drive cell differentiation. Abnormal patterns of gene expression are associated with the development of tumors.

Regulation of transcription

Eukaryotic cells regulate gene expression mostly by varying the rate of gene transcription. Interactions between RNA polymerase II and basal transcription factors lead to the formation of the transcription initiation complex (TIC) at the TATA box. Other transcription factors change the rate of transcription initiation by binding to promoter sequences and influencing the stability of the TIC. Distant sequences called enhancers and silencers also influence the rate of transcription.

Transcription factors

Gene promoters have multiple binding sites for transcription factors each of which can influence transcription. The overall effect on transcription depends on the complement of transcription factors bound. Transcription factors have a modular structure containing DNA binding, dimerization and transactivation domains with characteristic structural motifs. DNA binding domains contain three motifs: helix-turn-helix, zinc fingers and basic domains which occur in combination with dimerization domains. Dimerization domains contain two motifs: leucine zippers and helix-loop-helix. Dimerization allows the formation of homo- and heterodimers creating transcription factors with diverse functions. Transactivation domains do not have recognizable motifs but are often enriched with acidic amino acids, glutaminines or prolines. Transactivation domains probably interact with a variety of proteins in the TIC and at different stages during transcription. Transcription factors can also repress transcription by direct or indirect mechanisms.

Regulation of gene expression by hormones and cytokines

Hormones and cytokines influence target cells by activating gene transcription. Steroid hormones enter cells and bind steroid hormone receptor protein releasing it from an inhibitory protein. The receptor dimerizes and is translocated to the nucleus where it binds target gene promoters activating transcription. Polypeptide hormones and cytokines bind receptor proteins on the surface of target cells. Gene activation is triggered by signal transduction in which a network of proteins is sequentially activated by protein phosphorylation.
The human genome is estimated to contain as many as 100,000 genes which are subject to complex patterns of regulation that are, as yet, incompletely understood. In eukaryotic cells not all of the genes present in the genome are active. Cells express about 15% of their genes; the rest remain inactive. In multicellular organisms the genes that are active vary between cell types. The genes that are active in a particular cell type may be very different from those in another type of cell. The active genes determine which proteins and enzymes are present in a cell and are responsible for determining the characteristics of the cell and its role in the organism. For example, in lymphocytes which produce antibodies to fight infection the genes that encode the polypeptides that make up the antibodies are expressed at a high level. The pattern of genes expressed can change during the lifetime of a cell. For example, blood cells develop by differentiation from primitive progenitor cells. The changes that occur in the cell’s characteristics as it differentiates are a result of changes in the pattern of gene expression. Understanding how gene expression is regulated is important for understanding diseases such as cancer in which abnormal expression of genes leads to uncontrolled cell division and formation of a tumor.

Eukaryotic cells regulate the expression of their genes largely by determining the rate at which they are transcribed into mRNA. This can vary greatly with abundant proteins transcribed at very high rates and rare proteins at much lower rates. Regulation of gene expression is achieved by the interaction of gene promoters and DNA binding proteins called transcription factors (Fig. 1). Transcription of the gene by the RNA polymerase is initiated at the promoter and the efficiency of transcription initiation can be varied by the interaction between short regulatory DNA sequences present in the promoter that are recognized and bound by transcription factor proteins. The regulatory sequences present in the promoter are parallel with the coding sequence and are said to be cis-acting.

In eukaryotic cells protein coding genes are all transcribed by RNA polymerase II. Transcription is initiated by the formation of the transcription initiation complex (TIC) which involves binding of RNA polymerase II and a number of associated proteins called basal transcription factors to the DNA.

Fig. 1. Regulation of gene expression in eukaryotes.
of the promoter at a characteristic sequence known as the TATA box (see Topic A4). This has the sequence 5' TATAA/T A.A/T 3' and is present in most but not all eukaryotic genes located approximately 25 bp upstream of the transcription initiation site. Its function is to locate the RNA polymerase in the correct position to initiate transcription. Some genes, especially those expressed only in specific tissues or cells, do not have a TATA box but instead have an initiator sequence usually located over the transcription start site. Other genes, usually those expressed at low levels, have neither a TATA box nor an initiator element.

The efficiency of transcription initiation and hence the amount of mRNA produced is influenced by additional transcription factors that bind other DNA sequences present in the promoter and can interact with the proteins of the TIC affecting its stability. Transcription factors can increase or decrease the rate of transcription. Many different transcription factors exist each of which recognizes and binds a DNA sequence in the promoter. The sequence recognized can vary between promoters and the binding site is usually described as a consensus sequence which incorporates possible variations. The transcription factors are synthesized in the cytoplasm of the cell but exert their effects in the nucleus. As such, they are often referred to as trans-acting factors.

The rate of transcription of a gene can also be influenced by sequence elements called enhancers that may be located thousands of base pairs distant from the transcription start site. Enhancers are typically 100-200 bp long and contain sequences that bind transcription factors and can stimulate transcription of the linked gene. The position of the enhancer relative to the gene it influences can vary, and may be upstream or downstream. Enhancers work independently of their orientation and are equally effective facing either forward or reverse. Interaction between the enhancer and its promoter occurs by looping of the intervening DNA to bring the two into close proximity. Some enhancers contain sequences that bind transcription factors that influence transcription negatively. These are known as silencers and they may be responsible for restricting expression to specific cell types. Other distant sequences called locus control elements exist which influence expression of entire families of genes by controlling access of transcription proteins to the DNA. An example is the locus control region that regulates expression of the globin gene family.

These are a large family of proteins that regulate the expression of protein coding genes. They are distinct from the basal transcription factors that interact with the RNA polymerase II to form the TIC. Transcription factors have varied patterns of expression: some occur only in specific cell types whereas others occur in all cell types. Each transcription factor bound by a gene promoter can regulate the transcription of the gene either positively or negatively. The overall effect on transcription depends on the complement of transcription factors bound. The ability of transcription factors to form dimers with themselves and other transcription factors adds further to the possibilities for regulation of gene expression.

Transcription factors have a modular structure composed of discrete protein domains with specific functions. Three types of domain occur commonly:

- DNA binding domains;
- dimerization domains;
- transactivation domains.
The DNA binding and dimerization domains contain characteristic protein structures called motifs that identify them (Fig. 2).

**DNA binding domains**

Three types of DNA binding domain have been identified on the basis of the motifs present:

- **Helix–turn–helix.** This motif is composed of two α helices separated by a β turn. One of the helices, called the recognition helix, binds to the DNA by making contact with the major groove of the double helix. An example of proteins containing this motif are the homeodomain family of transcription factors encoded by the highly conserved homeobox genes that play an important role in embryonic development.

- **Zinc fingers.** This DNA binding motif occurs in two forms called C_{2}H_{2} and C_{4}. The C_{2}H_{2} form has a loop of 12 amino acids anchored at the base by two cysteines and two histidines that tetrahedrally coordinate a zinc ion. The motif forms a compact structure composed of two β strands and an α helix that contains basic amino acids that interact with the DNA via the major groove of the double helix. The zinc finger motif is repeated multiple times in the DNA binding domain and usually three or more are required for binding. An example of this motif is found in the ubiquitously expressed transcription factor Sp1. The C_{4} motif has a similar structure but has four cysteines coordinated to zinc. An example of this type of domain is found in steroid hormone receptor transcription factors.

- **Basic domains.** This DNA binding domain is usually found in association with one of two dimerization domains called the leucine zipper or the
helix-loop-helix (HLH) which give rise to basic leucine zipper (bZIP) and basic HLH proteins. Binding to DNA requires the presence of two basic domains which are brought together by dimerization.

**Dimerization domains**
Two types of motif are found in dimerization domains:

- **The leucine zipper** motif is usually present on the carboxyl terminal side of a DNA binding domain. It contains an α helix in which every seventh amino acid is leucine such that a leucine is present on the same side of the helix every second turn creating a hydrophobic face. Dimerization is achieved by the interaction between the hydrophobic faces of two leucine zippers and results in the basic DNA binding domains of the two proteins being brought into close proximity. The two DNA binding domains face in opposite directions allowing them to bind a DNA sequence which has inverted symmetry.

- **The HLH dimerization domain** contains two α helices separated by a nonhelical loop. Dimerization is achieved by interaction between hydrophobic amino acids present on one side of the carboxyl terminal helix. Dimerization can occur not only between two molecules of the same transcription factor (homodimers) but between different transcription factors with the same dimerization domain (heterodimers). The formation of heterodimers creates transcription factors with new functions and increases the possibilities for regulating expression of target genes. An example of this is the MyoD family of transcription factors which form homo- and heterodimers that regulate gene expression in developing muscle cells.

**Activation domains**
Unlike DNA binding and dimerization domains, no motifs have been identified which characterize activation domains. Analysis of amino acid sequences has shown only that activation domains are often enriched for certain amino acids. Specifically, activation domains have been identified that are rich in acidic amino acids (for example, in yeast Gal4 transcription factor), glutamines (for example, in Sp1 transcription factor), or prolines (for example, in c-Jun Ap2 and Oct-2 transcription factors).

Transcription factors regulate the expression of target genes by binding to the TIC and changing its stability. It is probable that activation domains interact with different proteins in the TIC and at a variety of stages during transcription initiation and elongation.

Some transcription factors can repress transcription. This may be achieved in a number of ways. Some may repress transcription directly by interacting with the transcription initiation complex. Other may act indirectly in a number of ways including: (i) blocking the DNA binding site of an activating transcription factor; (ii) formation of a dimer that lacks a DNA binding domain; or (iii) binding of a repressor protein to the activation domain of another transcription factor.

**Regulation of the expression of hormones and cytokines**
Hormones are agents produced by cells that act on other cells influencing their characteristics and functions by activating transcription of specific genes. Hormones may be small molecules, often steroids such as estrogens and glucocorticoids, or polypeptides such as insulin. Cytokines are proteins that act in a similar way to hormones, often with blood cells as their targets. Hormones and
cytokines modulate gene expression in target cells in a number of ways. Steroid hormones are lipid-soluble and so can pass through the cell membrane into the cytoplasm where they bind to a transcription factor called the **steroid hormone receptor**. Binding causes the steroid hormone receptor to be released from an inhibitory protein. It then dimerizes and is translocated to the nucleus where it activates transcription of target genes by binding to promoter sequences. Polypeptide hormones and cytokines act in a different way to steroid hormones by binding to receptor proteins on the surface of the target cell. Gene activation is triggered by a process called **signal transduction** in which a network of proteins is sequentially activated by protein phosphorylation. Ultimately this leads to stimulation of transcription of target genes by binding of transcription factors to gene promoter sequences.
**B1 CHROMOSOMES**

**Key Notes**

**Prokaryote and eukaryote chromosomes**

Prokaryote chromosomes consist of a single DNA molecule, that is usually circular, with only a small amount of associated protein. Each chromosome has a single origin of DNA replication. Eukaryotes have several linear chromosomes, and the DNA is tightly associated with large amounts of protein. Each eukaryote chromosome has multiple origins of DNA replication.

**Chromosome morphology**

Eukaryote chromosomes are visible by light microscopy only during cell division, after they have replicated. Chromosomes are placed into morphological groups, metacentric, submetacentric, acrocentric and telocentric, according to the position of their centromere. Within a species each autosome is given a specific number in ascending order from the largest to the smallest. Chromosome banding aids chromosome identification and gives some information as to the underlying organization of the chromosome. G-banding gives a series of light and dark bands along the length of the chromosome. C-banding produces dark bands in regions of constitutive heterochromatin.

**Specialized chromosome structures**

Centromeres are points on chromosomes to which the spindle fibers are attached. This is mediated through specialized protein structures known as kinetochores. Centromeres are composed largely of highly repeated satellite DNA sequences. Specialized structures at chromosome ends are known as telomeres. These are also composed of short repeated DNA sequences. The number of repeats decreases with age in somatic cells, but is maintained in germ cells and tumor cells by the enzyme telomerase. Telomeres prevent recombination between the ends of chromosomes.

Nucleolar organizer regions (NORs) contain tandem repeats of the major ribosomal RNA genes and are located in secondary constrictions. When the region of the chromosome distal to the NOR is small it is referred to as a chromosomal satellite.

**Molecular structure of chromosomes**

Chromatin is the term given to the association of DNA and proteins that composes chromosomes. It contains basic proteins, histones, and nonhistone acidic proteins. The histones form nucleosomes around which the DNA is wound. Nucleosomes consist of two discs containing histones H2a, H2b, H3 and H4; 146 base pairs of DNA are associated with each nucleosome and linker DNA leads to the next nucleosome. A single molecule of histone H1 attaches outside the core. This molecule is responsible for further folding of the nucleosomes into solenoids and more complex structures. Acidic proteins are involved in the chromosomal scaffold and in gene regulation.

**Functional and nonfunctional chromatin**

Heterochromatin is inactive chromatin, whereas euchromatin is actively involved in RNA transcription. Heterochromatin appears denser than euchromatin under the electron microscope and stains darker under the light
microscope. Some chromatin can exist as either hetero- or euchromatin. This is called facultative heterochromatin. One of the two X chromosomes in cells of female mammals is converted to heterochromatin. It forms a small dark body attached to the nuclear membrane. Chromatin that is permanently heterochromatic is called constitutive heterochromatin, and can be identified by C-banding.

Polyploidy is where the altered chromosome number is a multiple of the haploid chromosome number. This is rare in animals but important in plants. Small changes in chromosome number are classed as aneuploidy. In humans up to 4% of conceptuses are aneuploid, but very few of these survive to birth. Those that do survive tend to involve smaller chromosomes or alterations of the sex chromosomes. Aneuploidy arises by nondisjunction of homologous chromosomes or chromatids. It is more common in older mothers. Translocations can cause inheritance of trisomy 21 within families. Loss of a chromosome has more severe effects than gain of an extra chromosome. Some aneuploidies are often found as mosaics.

### Related topics
- Cell division (B2)
- Meiosis and gametogenesis (C3)
- The human genome (B4)
- Polyploidy (D8)

### Prokaryote and eukaryote chromosomes

All cellular life-forms have structures carrying genes, encoded in DNA, that are referred to as chromosomes. There are, however, major differences between these structures in prokaryotes and even the most simple of the eukaryotes. In prokaryotes the chromosome consists of a single DNA double helix, that is usually circular and has relatively few proteins associated with it. DNA replication proceeds from a single origin of replication (see Topic A9). Eukaryotes have, in almost all cases, a number of different chromosomes which are linear and which are contained within a membrane-bound organelle, the nucleus. The DNA molecules are intimately associated with large amounts of specific proteins. These may have functional or structural roles. The amount of DNA per chromosome is much greater in eukaryotes and because of this there are multiple origins of replication on each chromosome.

### Chromosome morphology

Eukaryote chromosomes are usually only visible when a cell is in the process of dividing (see Topics B2 and C3), after the chromosome has been replicated into identical double structures known as chromatids (daughter chromosomes). Chromosomes are classified on the basis of their morphology. This is determined by the position of the centromere (primary constriction). Fig. I shows four typical morphologies for chromosomes. In metacentric (median) chromosomes the centromere is close to the midpoint of the chromosome. This divides the chromosome into two roughly equal halves (arms). Where the centromere is sufficiently far away from the midpoint for a long arm (q arm) and a short arm (p arm) to be distinguished the chromosome is referred to as submetacentric. The other two morphological classes relate to chromosomes in which the centromere is close to one end of the chromosome. In telocentric chromosomes the centromere is at the end of the chromosome and there is only one arm. If the centromere is so close to the end of the chromosome that the
short arm is only just discernible then the chromosome is termed acrocentric.
Students often confuse submetacentric and acrocentric morphologies. The differences between the two are easily recognized in the chromosomes of humans (see Fig. 2).

In any species the complete diploid set of chromosomes is referred to as the karyotype. The autosomes are numbered in order of decreasing size, and the sex chromosomes are referred to as X or Y (see Topic C8). A chromosome preparation can be photographed, the homologous chromosomes paired, and set out in order. This is known as an ideogram, and is the conventional way to display karyotypes.

By grouping chromosomes by relative size and morphology it is usually possible to individually identify each chromosome in a species. This process was made much simpler by the development of treatments which, when applied to chromosomes prior to staining, produce a pattern of dark and light bands unique to each chromosome. Although there are several different banding techniques available only the two major processes, G-banding and C-banding, will be described here.

G-banding patterns can be produced by a wide range of different treatments but the most commonly used is a mild pretreatment of chromosome preparation slides with protease enzymes such as trypsin. The treated slides are then stained with Giemsa (hence the term G-banding). The process causes chromosomes to stain as a series of dark G-bands and pale interbands. The pattern is unique to each homologous chromosome pair and greatly aids the process of chromosome identification. These patterns have been used to create cytological maps of each chromosome in many species so that subchromosomal region can be accurately identified. This is important in various processes such as gen mapping (see Topic F5) and medical genetics (see Topic F1).

Fig. 2 sets out the G band pattern for each of the human chromosome diagrammatically. The pattern of bands allows each individual chromosome to be divided into a series of regions and subregions. Using chromosome 1 as a
Fig. 2. The human karyotype. This shows one member of each pair of homologous chromosomes, drawn to show their individual pattern of G-bands (G-banding is described later in this topic). The chromosomes are arranged in order starting with the longest, chromosome 1. The X chromosome is placed in order of its length between autosomes numbers 7 and 8. The Y chromosome is placed at the end. Each chromosome arm is subdivided into regions. Only chromosome 2 and chromosome 19 appear to be metacentric. Chromosomes 13, 14, 15, 21, 22 and Y are acrocentric. The remainder are submetacentric, showing a variety of different arm ratios. Chromosomes 13, 14, 15, 21 and 22 are drawn to show chromosomal satellites distal to the region of ribosomal DNA. Reproduced from Human Molecular Genetics (Strachan and Read), BIOS Scientific Publishers,
example the first division is between the long (q) and the short (p) arm. Fig. 2 shows that the q arm is further divided into regions 1, 2, 3 and 4. There are two further levels of subdivision so that a specific region may be accurately defined as 1q42.1. Note that not all chromosomes nor chromosome regions are equally subdivided. These regions are arbitrary, and are decided by international conventions of cytogeneticists. As techniques have allowed the banding of longer, less condensed chromosomes so the number of subregions has increased. G-banding not only provides us with a convenient method for identifying chromosomes, but it also gives us information as to the overall organization of DNA and genes within eukaryote chromosomes. In general the darkly stained G bands are rich in the bases adenine and thymine whereas the pale staining interbands are richer in guanine and cytosine. More genes are located in the interbands than in the G bands. C-banding also gives us an insight into the organization of chromosomes. C-banding produces a number of dark bands. These are largely confined to areas around centromeres. These indicate regions of constitutive heterochromatin and are discussed in greater detail later in this topic.

Specialized chromosomal structures

All eukaryote chromosomes contain two different areas which have specific structural importance. These are the centromeres and telomeres. In addition some chromosomes contain nucleolar organizer regions (NORs). Centromeres are the sites at which the spindle attaches during cell division and functional centromeres are essential to this process. Any chromosome fragment which loses its connection to a centromere will not segregate to daughter cells at the end of cell division. The best studied centromeres are those of yeast where some are as short as 200 bp. Most centromeres are much larger than this. Normally the centromere consists of highly repeated satellite DNA (see Topic B4). In humans different chromosomes can be distinguished by the presence of specific alphoid satellite DNAs within their centromeres. Connection of the chromosome to the microtubular spindle fibers is effected by proteins that attach to the centromere forming a multilayered structure known as a kinetochore.

Telomeres are not simply the ends of chromosomes and DNA molecules, but are specialized structures. They contain multiple repeats of simple, short DNA sequences. In humans the repeat sequence is TTAGGG, but there is little variation between eukaryotes: similar sequences are found in plant and protist species. Specific proteins bind to the telomere region and the resulting nucleo-protein structures are thought to prevent recombination between the ends of different chromosomes. The number of repeats per telomere is high in germ cells but decreases with age in somatic tissues; this is a molecular marker of the aging process. Telomere length is maintained by the enzyme telomerase, a protein that contains RNA complementary to the telomere repeat DNA sequence, that acts as a template for extension of the telomere. Telomerase is absent from somatic cells but reappears in tumor cells, where telomere length is stabilized.

NORs are usually found at secondary constrictions. They consist of tandemly repeated 5S, 18S and 28S rRNA genes (see Topic A5). In most species the 5S rRNA genes are clustered elsewhere in the genome. In humans NORs are found on the short arms of all acrocentric chromosomes except the Y chromosome. Each NOR consists of approximately 80–100 repeats. During interphase the NOR decondenses and a nucleolus forms around it; NORs from different chromosomes can be incorporated into a single nucleolus. When the cell enters