Chapter 2 summary

Aims

- To illustrate how living systems are organised at the genetic level
- To describe the structure and function of DNA and RNA
- To outline the organisation of genes and genomes
- To describe the processes involved in gene expression and its regulation

Chapter summary/learning outcomes

When you have completed this chapter you will have knowledge of:

- The organisation of living systems and the concept of emergent properties
- The chemistry of living systems
- The genetic code and the flow of genetic information
- The structure of DNA and RNA
- Gene structure and organisation
- Transcription and translation
- The need for regulation of gene expression
- Genome organisation
- The transcriptome and the proteome

Key words

Structure, function, emergent properties, covalent bond, atom, molecule, macromolecule, lipid, carbohydrate, protein, nucleic acid, dehydration synthesis, hydrolysis, cell membrane, plasma membrane, cell wall, genetic material, prokaryotic, eukaryotic, nucleus, enzyme, heteropolymer, codon, minimum coding requirement, genetic code, redundancy, wobble, mRNA, transcription, translation, Central Dogma, replication, reverse transcriptase, mutation, nucleotide, polynucleotide, antiparallel, purines, pyrimidines, double helix, rRNA, tRNA, ribosomes, gene, chromosome, locus, homologous pair, allele, coding strand, promoter, transcriptional unit, operon, operator, polycistronic, cistron, intron, exon, RNA processing, RNA polymerase, anticodon, adaptive regulation, differentiation, developmental regulation, housekeeping gene, constitutive gene, catabolic, inducible, repressor protein, genome, C-value, C-value paradox, inverted repeat, palindrome, foldback DNA, repetitive sequence, single-copy sequence, multigene family, transcriptome, proteome.

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

Chapter 2

Introducing molecular biology

This chapter presents a brief overview of the structure and function of DNA and its organisation within the genome (the total genetic complement of an organism). We will also have a look at how genes are expressed, and at the ways by which gene expression is regulated. The aim of the chapter is to provide the non-specialist reader with an introduction to the molecular biology of cells, but it should also act as a useful refresher for those who have some background knowledge of DNA. More extensive accounts of the topics presented here may be found in the sources described in Suggestions for Further Reading.

2.1 The way that living systems are organised

Living systems are organised hierarchically, with close interdependence of structure and function.

Before we look at the molecular biology of the cell, it may be useful to think a little about what cells are and how living systems are organised. Two premises are useful here. First, there is a very close link between **structure** and **function** in biological systems. Second, living systems provide an excellent example of the concept of **emergent properties**. This is rather like the statement 'the whole is greater than the sum of the parts', in that living systems are organised in a hierarchical way, with each level of organisation becoming more complex. New functional features emerge as components are put together in more complicated arrangements. One often-quoted example is the reactive metal sodium and the poisonous gas chlorine, which combine to give sodium chloride (common table salt), which is of course not poisonous (although it can be harmful if taken in excess!). Thus, it is often difficult or impossible to predict the properties of a more complex system by looking at its constituent parts, which is a general difficulty with the reductionist approach to experimental science.

Complex molecules (macromolecules) are made by joining smaller molecules together using dehydration synthesis.

The chemistry of living systems is based on the element carbon, which can form four **covalent bonds** with other **atoms**. By joining carbon atoms together, and incorporating other atoms, **molecules** can be built up, which in turn can be joined together to produce **macromolecules**. Biologists usually recognise four groups of macromolecules: **lipids**, **carbohydrates**, **proteins**, and **nucleic acids**. The synthesis

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

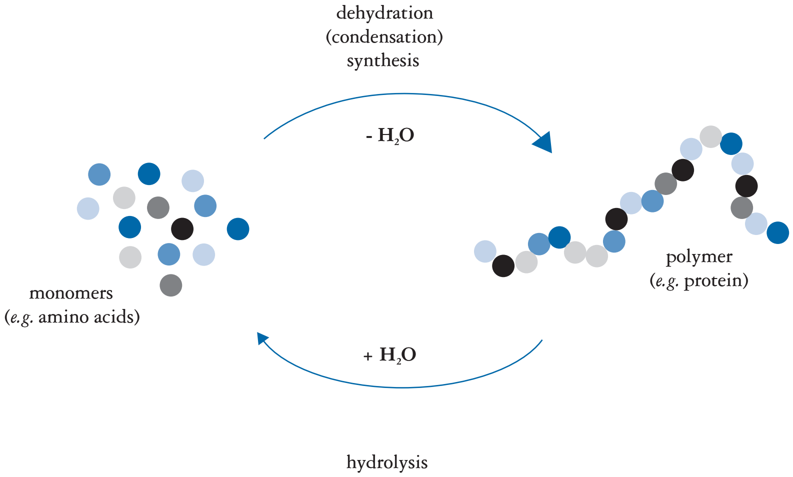


Fig. 2.1 The monomer–polymer cycle. In this example a representation of amino acids and proteins is shown. The amino acids (monomers) are joined together by removal of the elements of water (H_2O) during dehydration synthesis. When the protein is no longer required, it may be degraded by adding back the H_2O during hydrolysis. Although the cycle looks simple when presented like this, the synthesis of proteins requires many components whose functions are coordinated during the complex process of translation.

of macromolecules involves a condensation reaction between functional groups on the molecules to be joined together. This **dehydration synthesis** forms a covalent bond by removing the elements of water. In the case of the large polymeric macromolecules of the cell (polysaccharides, proteins, and nucleic acids) hundreds, thousands, or even millions of individual monomeric units may be joined together in this way. The polymers can be broken apart into their consituent monomers by adding the elements of water back to reconstitute the original groups. This is known as **hydrolysis** (literally *hydro lysis*, water breaking). The monomer/polymer cycle and dehydration/hydrolysis are illustrated in Fig. 2.1.

The cell is the basic unit of organisation in biological systems; prokaryotic cells have no nucleus, eukaryotic cells do.

The cell is the basic unit of organisation in biological systems. Although there are many different types of cell, there are some features that are present in all cells. There is a **cell membrane** (the **plasma membrane**) that is the interface between the cell contents and the external environment. Some cells, such as bacteria, yeasts, and plant cells, may also have a **cell wall** that provides additional structural support. Some sort of **genetic material** (almost always DNA) is required to provide the information for cells to function, and the organisation of this genetic information provides one way of classifying cells. In **prokaryotic** cells (*e.g.* bacteria) the DNA is not compartmentalised, whereas in **eukaryotic** cells the DNA is located within a membrane-bound **nucleus**. Eukaryotic cells also utilise membranes to provide additional internal structure. Prokaryotic cells are generally smaller in size than eukaryotic cells, but all cells have a maximum upper size limit. This is largely because of the limitations of diffusion as a mechanism for gas and nutrient exchange. Typical bacterial cells have a diameter of 1–10 μm, plant and animal cells 10–100 μm. In

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

multicellular eukaryotes, an increase in the size of the organism is achieved by using more cells rather than by making cells bigger.

2.2 The flow of genetic information

Life is directed by four nitrogenous bases: adenine (A), guanine (G), cytosine (C), and thymine (T).

To set the structure of nucleic acids in context, it is useful to think a little about what is required, in terms of genetic information, to enable a cell to carry out its various activities. It is a remarkable fact that an organism's characteristics are encoded by a four-letter alphabet, defining a language of three-letter words. The letters of this alphabet are the nitrogenous bases adenine (A), guanine (G), cytosine (C), and thymine (T). So how do these bases enable cells to function?

The expression of genetic information is achieved ultimately *via* proteins, particularly the **enzymes** that catalyse the reactions of metabolism. Proteins are condensation **heteropolymers** synthesised from amino acids, of which 20 are used in natural proteins. Given that a protein may consist of several hundred amino acid residues, the number of different proteins that may be made is essentially unlimited, assuming that the correct sequence of amino acids can be specified from the genetic information. As the bases are critical informatic components, we can calculate that using the bases singly would not provide enough scope (only 4 possible arrangements) to encode 20 amino acids, as there are only 4 possible code 'combinations' (A, G, C, and T). If the bases were arranged in pairs, that would give 4^2 or 16 possible combinations – still not enough. Triplet combinations provide 4^3 or 64 possible permutations, which is more than sufficient. Thus, great diversity of protein form and function can be achieved using an elegantly simple coding system, with sets of three nucleotides (**codons**) specifying the amino acids. Thus, a protein of 300 amino acids would have a **minimum coding requirement** of 900 nucleotides on a strand of DNA. The **genetic code** or 'dictionary' is one part of molecular biology that, like the double helix, has become something of a biological icon. Although there are more possible codons that are required (64 as opposed to 20), three of these are 'STOP' codons. Several amino acids are specified by more than one codon, which accounts for the remainder, a feature that is known as **redundancy** of the code. An alternative term for this, where the first two bases in a codon are often critical with the third less so, is known as **wobble**. These features can be seen in the standard presentation of the genetic code shown in Table 2.1.

The flow of genetic information is from DNA to RNA to protein, via the processes of transcription (T_C) and translation (T_L). This concept is known as the Central Dogma of molecular biology.

The flow of genetic information is unidirectional, from DNA to protein, with **messenger RNA** (**mRNA**) as an intermediate. The copying of DNA-encoded genetic information into RNA is known as **transcription** (T_C), with the further conversion into protein being termed **translation** (T_L). This concept of information flow is known as the **Central Dogma** of molecular biology and is an underlying theme in all studies of gene expression.

Two further aspects of information flow may be added to this basic model to complete the picture. First, duplication of the genetic

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

Table 2.1. The genetic code

First base (5′ end)	Second base				Third base (3′ end)
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Note: Codons read $5'\rightarrow 3'$; thus, AUG specifies Met. The three-letter abbreviations for the amino acids are as follows: Ala, Alanine; Arg, Arginine; Asn, Asparagine; Asp, Aspartic acid; Cys, Cysteine; Gln, Glutamine; Glu, Glutamic acid; Gly, Glycine; His, Histidine; Ile, Isoleucine; Leu, Leucine; Lys, Lysine; Met, Methionine; Phe, Phenylalanine; Pro, Proline; Ser, Serine; Thr, Threonine; Trp, Tryptophan; Tyr, Tyrosine; Val, Valine. The three codons UAA, UAG, and UGA specify no amino acid and terminate translation.

material prior to cell division represents a DNA–DNA transfer, known as DNA **replication**. A second addition, with important consequences for the genetic engineer, stems from the fact that some viruses have RNA instead of DNA as their genetic material. These viruses (chiefly members of the retrovirus group) have an enzyme called **reverse transcriptase** (an RNA-dependent DNA polymerase) that produces a double-stranded DNA molecule from the single-stranded RNA genome. Thus, in these cases the flow of genetic information is reversed with respect to the normal convention. The Central Dogma is summarised in Fig. 2.2.

2.3 The structure of DNA and RNA

In most organisms, the primary genetic material is double-stranded DNA. What is required of this molecule? First, it has to be stable, as genetic information may need to function in a living organism for up

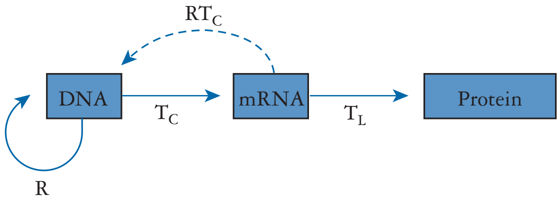


Fig. 2.2 The Central Dogma states that information flow is unidirectional, from DNA to mRNA to protein. The processes of transcription (T$_C$), translation (T$_L$), and DNA replication (R) obey this rule. An exception is found in retroviruses (RNA viruses), which have an RNA genome and carry out a process known as reverse transcription (RT$_C$) to produce a DNA copy of the genome following infection of the host cell.

to 100 years or more. Second, the molecule must be capable of replication, to permit dissemination of genetic information as new cells are formed during growth and development. Third, there should be the potential for limited alteration to the genetic material (**mutation**), to enable evolutionary pressures to exert their effects. The DNA molecule fulfils these criteria of stability, replicability, and mutability, and when considered with RNA provides an excellent example of the premises that we considered earlier – the very close relationship between structure and function, and the concept of emergent properties.

Nucleic acids are heteropolymers composed of monomers known as **nucleotides**; a nucleic acid chain is therefore often called a **polynucleotide**. The monomers are themselves made up of three components: a sugar, a phosphate group, and a nitrogenous base. The two types of nucleic acid (DNA and RNA) are named according to the sugar component of the nucleotide, with DNA having 2′-deoxyribose as the sugar (hence **D**eoxyribo**N**ucleic**A**cid) and RNA having ribose (hence **R**ibo**N**ucleic**A**cid). The sugar/phosphate components of a nucleotide are important in determining the structural characteristics of polynucleotides, and the nitrogenous bases determine their information storage and transmission characteristics. The structure of a nucleotide is summarised in Fig. 2.3.

Nucleic acids are polymers composed of nucleotides; DNA is deoxyribonucleic acid, RNA is ribonucleic acid.

Nucleotides can be joined together by a 5′→3′ phosphodiester linkage, which confers directionality on the polynucleotide. Thus, the 5′ end of the molecule will have a free phosphate group, and

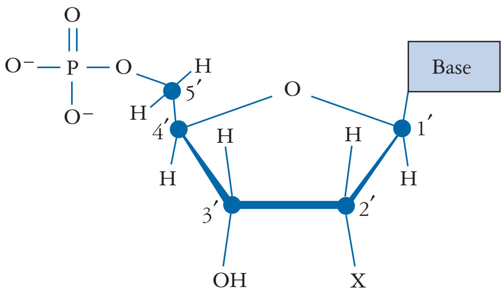
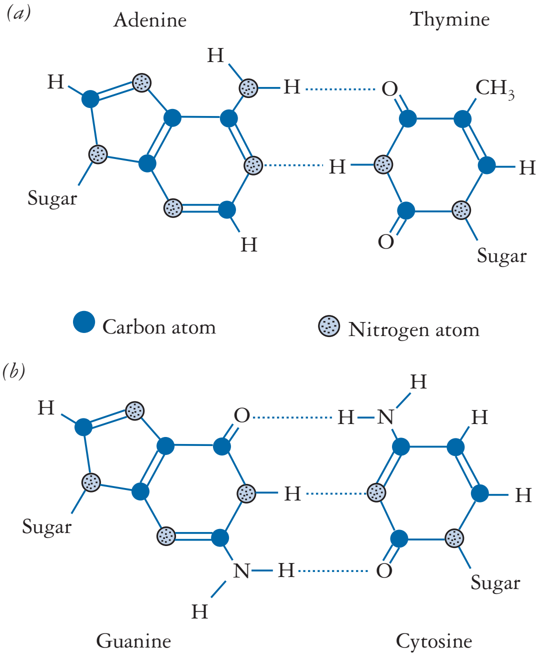


Fig. 2.3 The structure of a nucleotide. Carbon atoms are represented by solid circles, numbered 1′ to 5′. In DNA the sugar is deoxyribose, with a hydrogen atom at position X. In RNA the sugar is ribose, which has a hydroxyl group at position X. The base can be A, G, C, or T in DNA, and A, G, C, or U in RNA.

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

Fig. 2.4 Base-pairing arrangements in DNA. (*a*) An A · T base pair. The bases are linked by two hydrogen bonds (dotted lines). (*b*) A G · C base pair, with three hydrogen bonds.



In DNA the bases pair A · T and G · C; this complementary base pairing is the key to information storage, transfer, and use.

the 3′ end a free hydroxyl group; this has important consequences for the structure, function, and manipulation of nucleic acids. In a double-stranded molecule such as DNA, the sugar–phosphate chains are found in an **antiparallel** arrangement, with the two strands running in different directions.

The nitrogenous bases are the important components of nucleic acids in terms of their coding function. In DNA the bases are as listed in Section 2.1, namely adenine (A), guanine (G), cytosine (C), and thymine (T). In RNA the base thymine is replaced by uracil (U), which is functionally equivalent. Chemically adenine and guanine are **purines**, which have a double-ring structure, whereas cytosine and thymine (and uracil) are **pyrimidines**, which have a single-ring structure. In DNA the bases are paired, A with T and G with C. This pairing is determined both by the bonding arrangements of the atoms in the bases and by the spatial constraints of the DNA molecule, the only satisfactory arrangement being a purine–pyrimidine base pair. The bases are held together by hydrogen bonds, two in the case of an A · T base pair and three in the case of a G · C base pair. The structure and base-pairing arrangement of the four DNA bases is shown in Fig. 2.4.

The DNA molecule *in vivo* usually exists as a right-handed **double helix** called the *B*-form. This is the structure proposed by Watson and Crick in 1953. Alternative forms of DNA include the *A*-form (right-handed helix) and the *Z*-form (left-handed helix). Although DNA structure is a complex topic, particularly when the higher-order arrangements of DNA are considered, a simple representation will suffice here, as shown in Fig. 2.5.

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

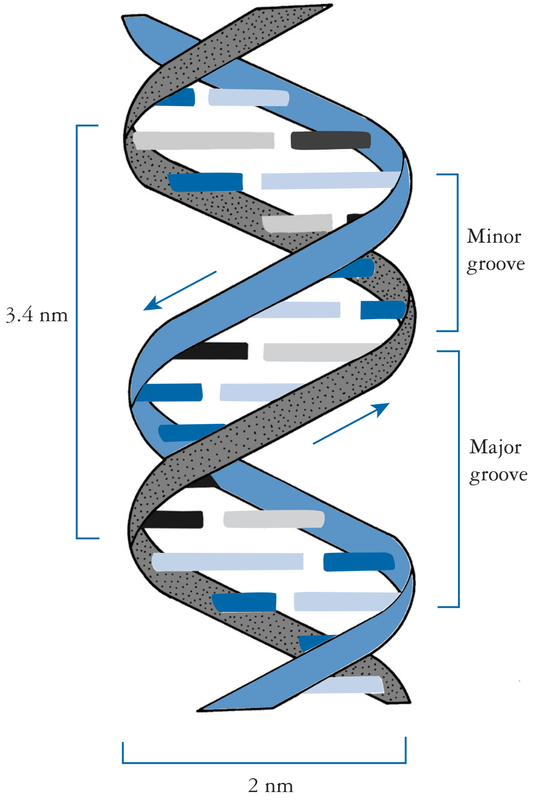


Fig. 2.5 The double helix. This is DNA in the commonly found *B*-form. The right-handed helix has a diameter of 2 nm and a pitch of 3.4 nm, with 10 base pairs per turn. The sugar–phosphate 'backbones' are antiparallel (arrowed) with respect to their 5′→3′ orientations. One of the sugar–phosphate chains has been shaded for clarity. The purine–pyrimidine base pairs are formed across the axis of the helix.

The structure of RNA is similar to that of DNA; the main chemical differences are the presence of ribose instead of 2′-deoxyribose and uracil instead of thymine. RNA is also most commonly single-stranded, although short stretches of double-stranded RNA may be found in self-complementary regions. There are three main types of RNA molecule found in cells: messenger RNA (mRNA), ribosomal RNA (**rRNA**), and transfer RNA (**tRNA**). Ribosomal RNA is the most abundant class of RNA molecule, making up some 85% of total cellular RNA. It is associated with **ribosomes**, which are an essential part of the translational machinery. Transfer RNAs make up about 10% of total RNA and provide the essential specificity that enables the insertion of the correct amino acid into the protein that is being synthesised. Messenger RNA, as the name suggests, acts as the carrier of genetic information from the DNA to the translational machinery and usually makes up less than 5% of total cellular RNA.

Three important types of RNA are ribosomal RNA (rRNA), messenger RNA (mRNA), and transfer RNA (tRNA).

2.4 Gene organisation

The **gene** can be considered the basic unit of genetic information. Genes have been studied since the turn of the century, when genetics became established. Before the advent of molecular biology and

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

the realisation that genes were made of DNA, study of the gene was largely indirect; the effects of genes were observed in phenotypes and the 'behaviour' of genes was analysed. Despite the apparent limitations of this approach, a vast amount of information about how genes functioned was obtained, and the basic tenets of transmission genetics were formulated.

As the gene was studied in greater detail, the terminology associated with this area of genetics became more extensive, and the ideas about genes were modified to take developments into account.

The gene is the basic unit of genetic information. Genes are located on chromosomes at a particular genetic locus. Different forms of the same gene are known as alleles.

The term 'gene' is usually taken to represent the genetic information transcribed into a single RNA molecule, which is in turn translated into a single protein. Exceptions are genes for RNA molecules (such as rRNA and tRNA), which are not translated. In addition, the nomenclature used for prokaryotic cells is slightly different because of the way that their genes are organised. Genes are located on **chromosomes**, and the region of the chromosome where a particular gene is found is called the **locus** of that gene. In diploid organisms, which have their chromosomes arranged as **homologous pairs**, different forms of the same gene are known as **alleles**.

2.4.1 The anatomy of a gene

Although there is no such thing as a 'typical' gene, there are certain basic requirements for any gene to function. The most obvious is that the gene has to encode the information for the particular protein (or RNA molecule). The double-stranded DNA molecule has the potential to store genetic information in either strand, although in most organisms only one strand is used to encode any particular gene. There is the potential for confusion with the nomenclature of the two DNA strands, which may be called coding/non-coding, sense/antisense, plus/minus, transcribed/non-transcribed, or template/non-template. In some cases different authors use the same terms in different ways, which adds to the confusion. Recommendations from the International Union of Biochemistry and the International Union of Pure and Applied Chemistry favour the terms coding/non-coding, with the **coding strand** of DNA taken to be the mRNA-like strand. This convention will be used in this book where coding function is specified. The terms template and non-template will be used to describe DNA strands when there is not necessarily any coding function involved, as in the copying of DNA strands during cloning procedures. Thus, genetic information is expressed by transcription of the non-coding strand of DNA, which produces an mRNA molecule that has the same sequence as the coding strand of DNA (although the RNA has uracil substituted for thymine; see Fig. 2.9(*a*)). The sequence of the coding strand is usually reported when dealing with DNA sequence data, as this permits easy reference to the sequence of the RNA.

In addition to the sequence of bases that specifies the codons in a protein-coding gene, there are other important regulatory sequences associated with genes (Fig. 2.6). A site for starting transcription is required, and this encompasses a region that binds RNA polymerase, known as the **promoter** (P), and a specific start point for transcription

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

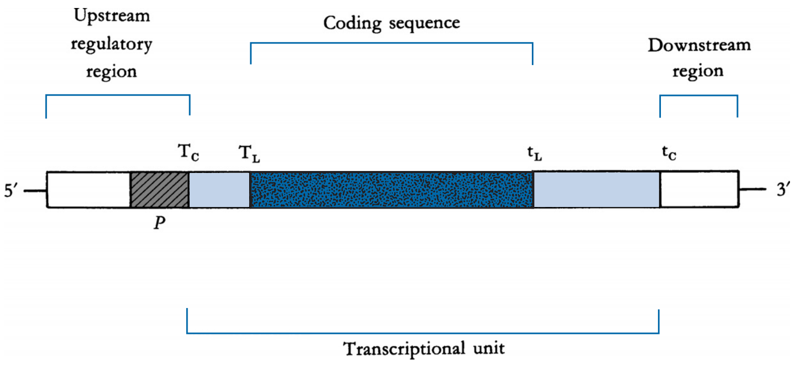


Fig. 2.6 Gene organisation. The transcriptional unit produces the RNA molecule and is defined by the transcription start site (T_C) and stop site (t_C). Within the transcriptional unit lies the coding sequence, from the translation start site (T_L) to the stop site (t_L). The upstream regulatory region may have controlling elements such as enhancers or operators in addition to the promoter (P), which is the RNA polymerase binding site.

(T_C). A stop site for transcription (t_C) is also required. From T_C start to t_C stop is sometimes called the **transcriptional unit**, that is, the DNA region that is copied into RNA. Within this transcriptional unit there may be regulatory sites for translation, namely a start site (T_L) and a stop signal (t_L). Other sequences involved in the control of gene expression may be present either upstream or downstream from the gene itself.

Genes have several important regions. A promoter is necessary for RNA polymerase binding, with the transcription start and stop sites defining the transcriptional unit.

2.4.2 Gene structure in prokaryotes

In prokaryotic cells such as bacteria, genes are usually found grouped together in **operons**. The operon is a cluster of genes that are related (often coding for enzymes in a metabolic pathway) and that are under the control of a single promoter/regulatory region. Perhaps the best known example of this arrangement is the *lac* operon (Fig. 2.7), which codes for the enzymes responsible for lactose catabolism. Within the

Genes in prokaryotes tend to be grouped together in operons, with several genes under the control of a single regulatory region.

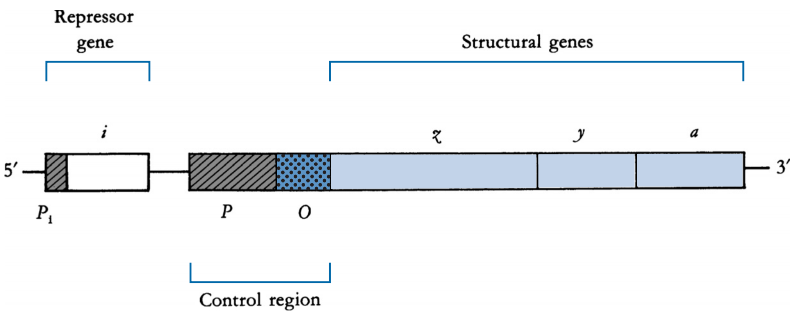


Fig. 2.7 The *lac* operon. The structural genes *lacZ*, *lacY*, and *lacA* (noted as *z*, *y*, and *a*) encode ß-galactosidase, galactoside permease, and a transacetylase, respectively. The cluster is controlled by a promoter (*P*) and an operator region (*O*). The operator is the binding site for the repressor protein, encoded by the *lacI* gene (*i*). The repressor gene lies outside the operon itself and is controlled by its own promoter, P_i.

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

operon there are three genes that code for proteins (termed structural genes) and an upstream control region encompassing the promoter and a regulatory site called the **operator**. In this control region there is also a site that binds a complex of cAMP (cyclic adenosine monophosphate) and CRP (cAMP receptor protein), which is important in positive regulation (stimulation) of transcription. Lying outside the operon itself is the repressor gene, which codes for a protein (the Lac repressor) that binds to the operator site and is responsible for negative control of the operon by blocking the binding of RNA polymerase.

The fact that structural genes in prokaryotes are often grouped together means that the transcribed mRNA may contain information for more than one protein. Such a molecule is known as a **polycistronic** mRNA, with the term **cistron** equating to the 'gene' as we have defined it (*i.e.* encoding one protein). Thus, much of the genetic information in bacteria is expressed *via* polycistronic mRNAs whose synthesis is regulated in accordance with the needs of the cell at any given time. This system is flexible and efficient, and it enables the cell to adapt quickly to changing environmental conditions.

2.4.3 Gene structure in eukaryotes

A major defining feature of eukaryotic cells is the presence of a membrane-bound nucleus, within which the DNA is stored in the form of chromosomes. Transcription therefore occurs within the nucleus and is separated from the site of translation, which is in the cytoplasm. The picture is complicated further by the presence of genetic information in mitochondria (plant and animal cells) and chloroplasts (plant cells only), which have their own separate genomes that specify many of the components required by these organelles. This compartmentalisation has important consequences for regulation, both genetic and metabolic, and thus gene structure and function in eukaryotes are more complex than in prokaryotes.

The most startling discovery concerning eukaryotic genes was made in 1977, when it became clear that eukaryotic genes contained 'extra' pieces of DNA that did not appear in the mRNA that the gene encoded. These sequences are known as intervening sequences or **introns**, with the sequences that will make up the mRNA being called **exons**. In many cases the number and total length of the introns exceed that of the exons, as in the chicken ovalbumin gene, which has a total of seven introns making up more than 75% of the gene. As our knowledge has developed, it has become clear that eukaryotic genes are often extremely complex, and may be very large indeed. Some examples of human gene complexity are shown in Table 2.2. This illustrates the tremendous range of sizes for human genes, the smallest of which may be only a few hundred base pairs in length. At the other end of the scale, the dystrophin gene is spread over 2.4 Mb of DNA on the X chromosome, with the 79 exons representing only 0.6% of this length of DNA.

Eukaryotic genes tend to be more complex than prokaryotic genes and often contain intervening sequences (introns). The introns form part of the primary transcript, which is converted to the mature mRNA by RNA processing.

The presence of introns obviously has important implications for the expression of genetic information in eukaryotes, in that the introns must be removed before the mRNA can be translated. This

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

Table 2.2. Size and structure of some human genes

Gene	Gene size (kbp)	Number of exons	% exon
Insulin	1.4	3	33
β-globin	1.6	3	38
Serum albumin	18	14	12
Blood clotting factor VIII	186	26	3
CFTR (cystic fibrosis)	230	27	2.4
Dystrophin (muscular dystrophy)	2400	79	0.6

Note: Gene sizes are given in kilobase pairs (kbp). The number of exons is shown, and the percentage of the gene that is represented by these exons is given in the final column.

is carried out in the nucleus, where the introns are spliced out of the primary transcript. Further intranuclear modification includes the addition of a ‘cap’ at the 5′ terminus and a ‘tail’ of adenine residues at the 3′ terminus. These modifications are part of what is known as **RNA processing**, and the end product is a fully functional mRNA that is ready for export to the cytoplasm for translation. The structures of the mammalian ß-globin gene and its processed mRNA are outlined in Fig. 2.8 to illustrate eukaryotic gene structure and RNA processing.

2.5 Gene expression

As shown in Fig. 2.2, the flow of genetic information is from DNA to protein. Whilst a detailed knowledge of gene expression is not

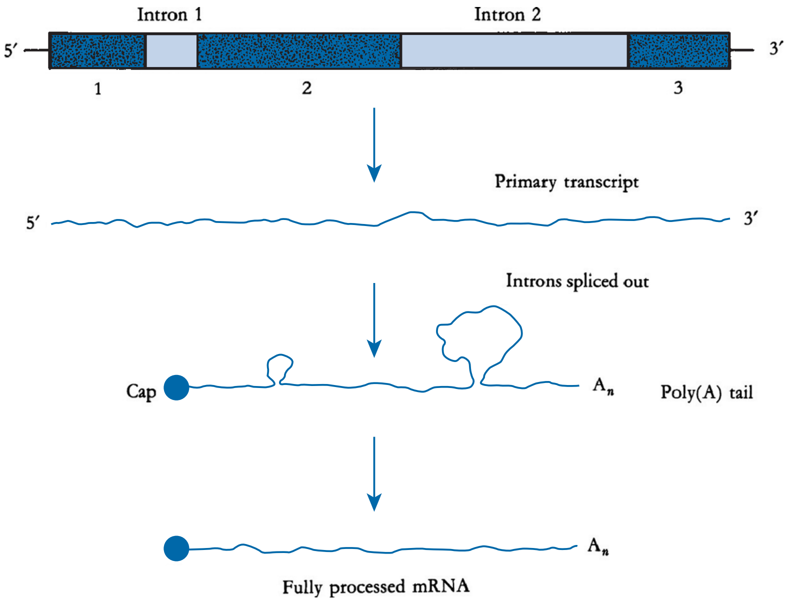


Fig. 2.8 Structure and expression of the mammalian ß-globin gene. The gene contains two intervening sequences or introns. The expressed sequences (exons) are shaded and numbered. The primary transcript is processed by capping, polyadenylation, and splicing to yield the fully functional mRNA.

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

required in order to understand the principles of genetic engineering, it is useful to be familiar with the main features of transcription and translation and to have some knowledge of how gene expression is controlled.

2.5.1 From genes to proteins

Genetic information is perhaps better thought of as a recipe than as a blueprint.

At this point it may be useful to introduce an analogy that I find helpful in thinking about the role of genes in determining cell structure and function. You may hear the term 'genetic blueprint' used to describe the genome. However, this is a little too simplistic, and I prefer to use the analogy of a recipe to describe how genes and proteins work. Let's consider making a cake – the recipe (gene) would be found in a particular book (chromosome), on a particular page (locus), and would contain information in the form of words (codons). One part of the recipe might read 'add 400 g of sugar and beat well', which is fairly clear and unambiguous. When put together with all the other ingredients and baked, the result is a cake in which you cannot see the sugar as an identifiable component. On the other hand, currants or blueberries would appear as identifiable parts of the cake. In a similar way many of the characteristics of an organism are determined by multiple genes, with no particular single gene product being identifiable. Conversely, in single-gene traits the effect of a particular gene may be easily identified as a phenotypic characteristic.

The effects of mutations can be mild or severe, depending on the type of mutation and its location in the gene sequence.

Mutation can also be considered in the recipe context, to give some idea of the relative severity of effect that different mutations can have. If we go back to our sugar example, what would be the effect of the last 0 of 400 being replaced by a 1, giving 401 g as opposed to 400 g? This change would almost certainly remain undetected. However, if the 4 of 400 changed to a 9, or if an additional 0 was added to 400, then things would be very different (and much sweeter!). Thus, mutations in non-critical parts of genes may be of no consequence, whereas mutation in a critical part of a gene can have extremely serious consequences. In some cases a single base insertion or substitution can have a major effect. (Think of adding a 'k' in front of the 'g' in 400 g!)

The recipe analogy is a useful one, in that it defines the role of the recipe itself (specifying the components to be put together) and also illustrates that the information is only part of the story. If the cake is not mixed or baked properly, even with the correct proportions of ingredients, it will not turn out to be a success. Genes provide the information to specify the proteins, but the whole process must be controlled and regulated if the cell is to function effectively.

2.5.2 Transcription and translation

These two processes are the critical steps involved in producing functional proteins in the cell. Transcription involves synthesis of an RNA from the DNA template provided by the non-coding strand of the transcriptional unit in question. The enzyme responsible is **RNA polymerase** (DNA-dependent RNA polymerase). In prokaryotes there is a

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

single RNA polymerase enzyme, but in eukaryotes there are three types of RNA polymerase (I, II, and III). These synthesise ribosomal, messenger, and transfer/5 S ribosomal RNAs, respectively. All RNA polymerases are large multisubunit proteins with relative molecular masses of around 500 000.

Transcription has several component stages: (1) DNA/RNA polymerase binding, (2) chain initiation, (3) chain elongation, and (4) chain termination and release of the RNA. Promoter structure is important in determining the binding of RNA polymerase but will not be dealt with here. When the RNA molecule is released, it may be immediately available for translation (as in prokaryotes) or it may be processed and exported to the cytoplasm (as in eukaryotes) before translation occurs.

Translation requires an mRNA molecule, a supply of charged tRNAs (tRNA molecules with their associated amino acid residues), and ribosomes (composed of rRNA and ribosomal proteins). The ribosomes are the sites where protein synthesis occurs; in prokaryotes, ribosomes are composed of three rRNAs and some 52 different ribosomal proteins. The ribosome is a complex structure that essentially acts as a 'jig' that holds the mRNA in place so that the codons may be matched up with the appropriate **anticodon** on the tRNA, thus ensuring that the correct amino acid is inserted into the growing polypeptide chain. The mRNA molecule is translated in a 5′→3′ direction, corresponding to polypeptide elongation from N terminus to C terminus.

The codon/anticodon recognition event marks the link between nucleic acid and protein.

Although transcription and translation are complex processes, the essential features (with respect to information flow) may be summarised as shown in Fig. 2.9. In conjunction with the brief descriptions presented earlier, this should provide enough background information about gene structure and expression to enable subsequent sections of the text to be linked to these processes where necessary.

2.5.3 Regulation of gene expression

Transcription and translation provide the mechanisms by which genes are expressed. However, it is vital that gene expression is controlled so that the correct gene products are produced in the cell at the right time. Why is this so important? Let's consider two types of cell – a bacterial cell and a human cell. Bacterial cells need to be able to cope with wide variations in environmental conditions and, thus, need to keep all their genetic material 'at the ready' in case particular gene products are needed. By keeping their genomes in this state of readiness, bacteria conserve energy (by not making proteins wastefully) and can respond quickly to any opportune changes in nutrient availability. This is an example of **adaptive regulation** of gene expression.

Prokaryotic genes are often regulated in response to external signals such as nutrient availability.

In contrast to bacteria, human cells (usually) experience a very different set of environmental conditions. Cells may be highly specialised and **differentiated**, and their external environment is usually stable and controlled by homeostatic mechanisms to ensure that no wide fluctuations occur. Thus, cell specialisation brings more complex

Eukaryotic genes are often regulated in response to signals generated from within the organism.

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

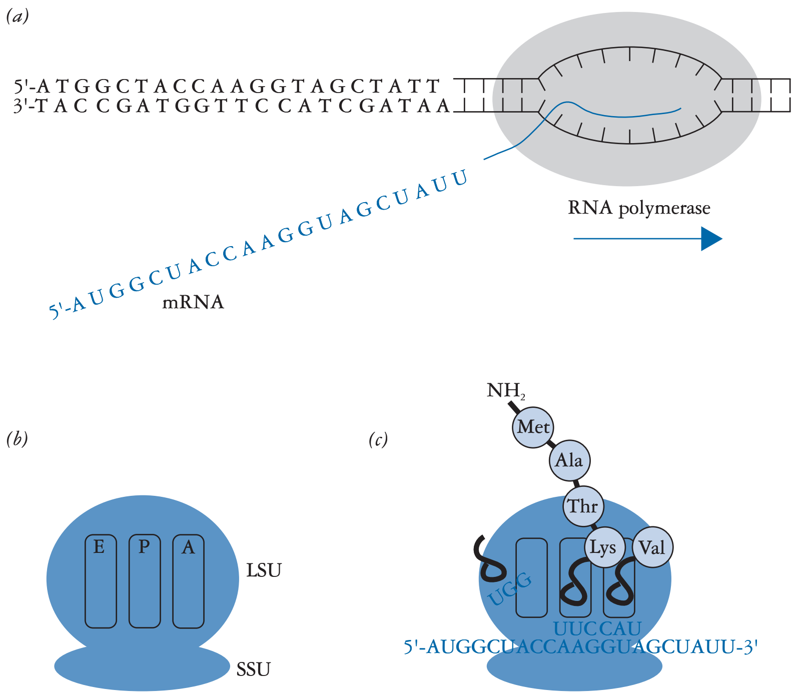

(a)
5'-A T G G C T A C C A A G G T A G C T A T T
3'-T A C C G A T G G T T C C A T C G A T A A
5'-A U G G C U A C C A A G G U A G C U A U U
mRNA
RNA polymerase
(b)
E
P
A
LSU
SSU
(c)
NH$_2$
Met
Ala
Thr
Lys
Val
UGG
UUCCAU
5'-AUGGCUACCAAGGUAGCUAUU-3'

Fig. 2.9 Transcription and translation. (*a*) Transcription involves synthesis of mRNA by RNA polymerase. Part of the DNA/mRNA sequence is given. The mRNA has the same sequence as the coding strand in the DNA (the non-template strand), apart from U being substituted for T. (*b*) The ribosome is the site of translation and is made up of the large subunit (LSU) and the small subunit (SSU), each made up of ribosomal RNA molecules and many different proteins. There are three sites within the ribosome. The A (aminoacyl) and P (peptidyl) sites are involved in insertion of the correct tRNA–amino acid complex in the growing polypeptide chain. The E (exit) site facilitates the release of the tRNA after peptide bond formation has removed its amino acid. (*c*) The mRNA is being translated. The amino acid residue is inserted into the protein in response to the codon/anticodon recognition event in the ribosome. The first amino acid residue is encoded by AUG in the mRNA (tRNA anticodon TAC), which specifies methionine (see Table 2.1 for the genetic code). The remainder of the sequence is translated in a similar way. The ribosome translates the mRNA in a 5′→3′ direction, with the polypeptide growing from its N terminus. The residues in the polypeptide chain are joined together by peptide bonds.

function but requires more controlled conditions. Differentiation is a function of development and, thus, genes in multicellular eukaryotes are often **developmentally regulated**. Gene regulation during the development and life cycle of a complex organism is, as you would expect, complex.

In addition to genes that are controlled and regulated, there are many examples of gene products that are needed at all times during a cell's life. Such genes are sometimes called **housekeeping genes** or **constitutive genes**, in that they are essentially unregulated and encode proteins that are essential at all times (such as enzymes for primary catabolic pathways).

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

Although a detailed discussion of the control of gene expression is outside the scope of this book, the basic principles can be illustrated by considering how bacterial operons are regulated. A bacterial cell (living outside the laboratory) will experience a wide range of environmental conditions. In particular, there will be fluctuations in the availability of nutrients. If the cell is to survive, it must conserve energy resources, which means that wasteful synthesis of non-required proteins should be prevented. Thus, bacterial cells have mechanisms that enable operons to be controlled with a high degree of sensitivity. An operon that encodes proteins involved in a **catabolic** pathway (one that breaks down materials to release energy) is often regulated by being switched 'on' only when the substance beceomes available in the extracellular medium. Thus, when the substance is absent, there are systems that keep catabolic operons switched 'off'. These are said to be **inducible** operons and are usually controlled by a negative control mechanism involving a **repressor protein** that prevents access to the promoter by RNA polymerase. The classic example of a catabolic operon is the *lac* operon (the structure of which is shown in Fig 2.7). When lactose is absent, the repressor protein binds to the operator and the system is 'off'. The system is a little 'leaky', however, and thus the proteins encoded by the operon (β-galactosidase, permease, and transacetylase) will be present in the cell at low levels. When lactose becomes available, it is transported into the cell by the permease and binds to the repressor protein, causing a conformational (shape) change so that the repressor is unable to bind to the operator. Thus, the negative control is removed, and the operon is accessible by RNA polymerase. A second level of control, based on the level of cAMP, ensures that full activity is only attained when lactose is present and energy levels are low. This dual-control mechanism is a very effective way of regulating gene expression, enabling a range of levels of expression that is a bit like a dimmer switch rather than an on/off swich. In the case of catabolic operons like the *lac* system, this ensures that the enzymes are only synthesised at maximum rate when they are really required.

Gene regulation in bacteria enables a range of levels of gene expression to be attained, rather than a simple on/off switch.

2.6 Genes and genomes

When techniques for the examination of DNA became established, gene structure was naturally one of the first areas where efforts were concentrated. However, genes do not exist in isolation, but as part of the **genome** of an organism. Over the past few years the emphasis in molecular biology has shifted slightly, and today we are much more likely to consider the genome as a whole – almost as a type of cellular organelle – rather than just a collection of genes. The Human Genome Project (considered in Chapter 10) is a good example of the development of the field of bioinformatics, which is one of the most active research areas in modern molecular biology.

The genome is the total complement of DNA in the cell.

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

Table 2.3. Genome size in some organisms

Organism	Genome size (Mb)
Escherichia coli (bacterium)	4.6
Saccharomyces cerevisiae (yeast)	12.1
Drosophila melanogaster (fruit fly)	150
Homo sapiens (man)	3000
Mus musculus (mouse)	3300
Nicotiana tabacum (tobacco)	4500
Triticum aestivum (wheat)	17000

Note: Genome sizes are given in megabase pairs (1 megabase = 1×10^6 bases).

2.6.1 Genome size and complexity

The amount of DNA in the haploid genome is known as the **C-value**. It would seem reasonable to assume that genome size should increase with increasing complexity of organisms, reflecting the greater number of genes required to facilitate this complexity. The data shown in Table 2.3 show that, as expected, genome size does tend to increase with organismal complexity. Thus, bacteria, yeast, fruit fly, and human genomes fit this pattern. However, mouse, tobacco, and wheat have much larger genomes than humans – this seems rather strange, as intuitively we might assume that a wheat plant is not as complex as a human being. Also, as *E. coli* has around 4000 genes, it appears that the tobacco plant genome has the capacity to encode 4000000 genes, and this is certainly not the case, even allowing for the increased size and complexity of eukaryotic genes. This anomaly is sometimes called the **C-value paradox**.

Eukaryotic genomes may have a range of different types of repetitive sequences.

In addition to the size of the genome, genome complexity also tends to increase with more complex organisation. One way of studying complexity involves examining the renaturation of DNA samples. If a DNA duplex is denatured by heating the solution until the strands separate, the complementary strands will renature on cooling (Fig. 2.10). This feature can be used to provide information about the sequence complexity of the DNA in question, since sequences that are present as multiple copies in the genome will renature faster than sequences that are present as single copies only. By performing this type of analysis, eukaryotic DNA can be shown to be composed of four different abundance classes. First, some DNA will form duplex structures almost instantly, because the denatured strands have regions such as **inverted repeats** or **palindromes**, which fold back on each other to give a hairpin loop structure. This class is commonly known as **foldback DNA**. The second fastest to re-anneal are highly **repetitive sequences**, which occur many times in the genome. Following these are moderately repetitive sequences, and finally there are the **unique** or **single-copy sequences**, which rarely re-anneal under the conditions used for this type of analysis. We will consider how repetitive

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

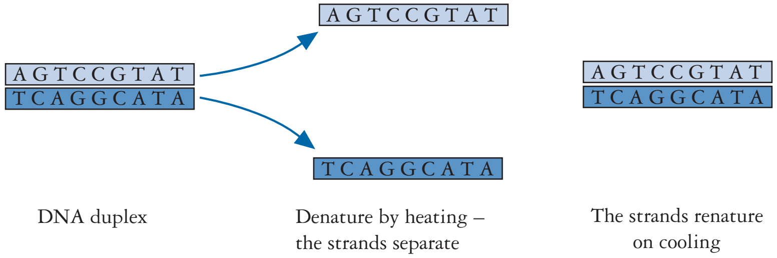

AGTCCGTAT
TCAGGCATA
AGTCCGTAT
TCAGGCATA
AGTCCGTAT
TCAGGCATA
DNA duplex
Denature by heating – the strands separate
The strands renature on cooling

Fig. 2.10 The principle of nucleic acid hybridisation. This feature of DNA molecules is a critical part of many of the procedures involved in gene manipulation and is also an essential feature of life itself. Thus, the simple G · C and A · T base pairing (see Fig. 2.4) has profound implications for living systems and for the applications of recombinant DNA technology.

DNA sequence elements can be used in genome mapping and DNA profiling in Chapters 10 and 12.

2.6.2 Genome organisation

The C-value paradox and the sequence complexity of eukaryotic genomes raise questions about how genomes are organised. Viral and bacterial genomes tend to show very efficient use of DNA for encoding their genes, which is a consequence of (and explanation for) their small genome size. However, in the human genome, only about 3% of the total amount of DNA is actually coding sequence. Even when the introns and control sequences are added, the majority of the DNA has no obvious function. This is sometimes termed ‘junk’ DNA, although this is perhaps the wrong way to think about this apparently redundant DNA.

Most of the human genome is not involved in coding for proteins.

Estimating the number of genes in a particular organism is not an exact science, and a number of different methods may be used. When the full genome sequence is determined, this obviously makes gene identification much easier, although there are many cases where gene coding sequences are recognised, but the protein products are unknown in terms of their biological function.

Many genes in eukaryotes are single copy genes, and tend to be dispersed across the multiple chromosomes found in eukaryotic cell nuclei. Other genes may be part of **multigene families**, and may be grouped at a particular chromosomal location or may be dispersed. When studying gene organisation in the context of the genome itself, features such as gene density, gene size, mRNA size, intergenic distance, and intron/exon sizes are important indicators. Early analysis of human DNA indicated that the ‘average’ size of a coding region is around 1 500 base pairs, and the average size of a gene is 10–15 kbp. Gene density is about one gene per 40–45 kbp, and the intergenic distance is around 25–30 kbp. However, as we have already seen, gene structure in eukaryotes can be very complex, and thus using ‘average’ estimates is a little misleading. What is clear is that genomes are now yielding much new information about gene structure and function as

Genome sequencing has greatly improved our understanding of how genomes work.

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

genome sequencing projects generate more data and we enter what is sometimes called the 'post-genomic era'.

2.6.3 The transcriptome and proteome

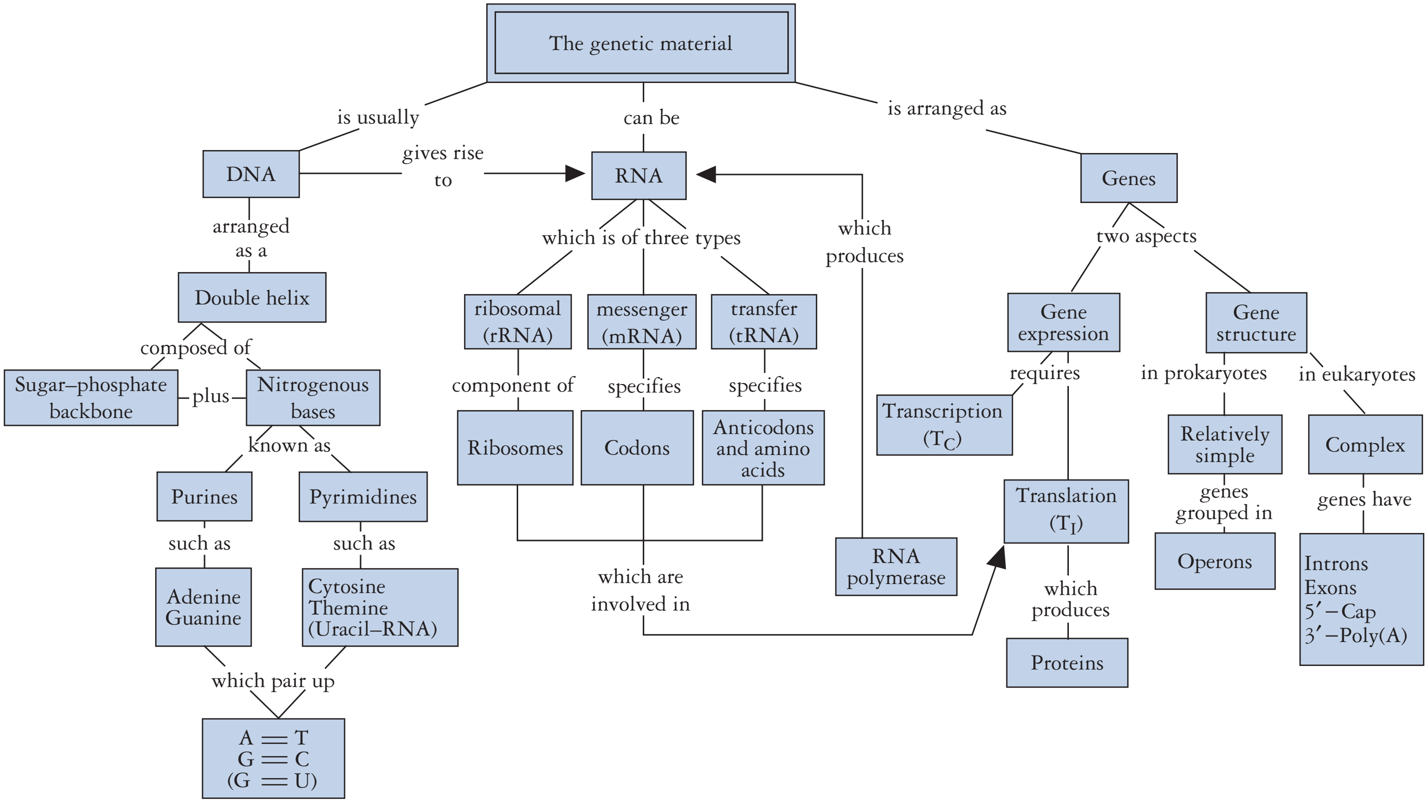
We finish this look at molecular biology by introducing two more '-omes' to complement the 'genome'. These terms have become widely used as researchers begin to delve into the bioinformatics of cells. The **transcriptome** refers to the population of transcripts at any given point in a cell's life. This expressed subset of the genomic information will be determnined by many factors affecting the status of the cell. There will be general 'housekeeping' genes for basic maintenance of cell function, but there may also be tissue-specific genes being expressed, or perhaps developmentally regulated genes will be 'on' at that particular point. Analysis of the transcriptome therefore gives a good snapshot of what the cell is engaged in at that point in time.

Analysis of the transcriptome and proteome provides useful information about which genes a cell is expressing at any given time.

The **proteome** is a logical extension to the genome and transcriptome in that it represents the population of proteins in the cell. The proteome will reflect the transcriptome to a much greater extent than the transcriptome reflects the genome, although there will be some transcripts that may not be translated efficiently, and there may be proteins that persist in the cell when their transcripts have been removed from circulation. Many biologists now accept that an understanding of the proteome is critical in developing a full understanding of how cells work. Some even consider the proteome as the 'holy grail' of cell biology, comparing it with the search for the unifying theory in physics. The argument is that, if we understand how all the proteins of a cell work, then surely we have a complete understanding of cell structure and function? As with most things in biology, this is unlikely to be a simple process, although the next few years will provide much excitement for biologists as the secrets of gene expression are revealed in more detail.

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press

https://doi.org/10.1017/CBO9780511800986.004 Published online by Cambridge University Press



Concept map 2