Chapter 1

Basic principles

Viruses and bacteria have ensured their survival over millions of years by using a variety of techniques to make, break and join deoxynucleic acid (DNA) and ribonucleic acid (RNA). In recent years, molecular biologists have adapted and exploited these naturally occurring processes, leading to remarkable breakthroughs in understanding the molecular basis of human disease. Methods for manipulating the nucleic acids DNA and RNA are extensive, and protein methodology is developing rapidly. This chapter provides an overview of the molecules involved, with an emphasis on understanding some of the fundamental principles that underlie the rapidly evolving techniques used to study disease. A better grasp of the fundamental technology will help in appreciating the possibilities and identifying the limits of translating new discoveries in molecular medicine into diagnostic tests and therapies for patients.

Organisms are made of cells

Living organisms are composed of cells. Some organisms, including bacteria, algae and yeasts, exist as single cells, whereas plants and animals consist of collections of cells. New cells, required for growth of an existing organism or the formation of new organisms, arise by division of existing cells.

Cell functions depend on proteins

All cellular functions depend on proteins, which consist of chains of amino acids. Only 20 different amino acids are commonly found in the proteins of all organisms. The links in a chain of amino acids are termed peptide bonds and the chains themselves are called polypeptides. Proteins contain one or more polypeptides, and the structure and function of each protein depends on the sequence of amino acids making up the polypeptide chains.

Proteins have many diverse functions. They maintain cell structure and provide motility, act as intra- and extracellular messengers, and bind and transport molecules, including oxygen, lipids and other proteins. Many proteins are enzymes which catalyse (accelerate) chemical reactions. Almost all chemical reactions, including those involved in the synthesis of fats and carbohydrates, are catalysed by enzymes.

Some proteins, for example, the enzymes involved in glucose metabolism, are present in most cells. In contrast, cells in multicellular organisms may become specialized and produce certain proteins that provide them with highly specific functions. Cells that produce particular proteins are often grouped together to form complex tissues or organs. For example, muscle cells produce proteins, including tropomyosin and myosin, which are involved in the formation of muscle filaments, islet cells of the pancreas synthesize the polypeptide hormone insulin, and liver cells contain enzymes found exclusively in the liver, such as those required for the conjugation of bilirubin into water-soluble forms.

DNA contains the information needed to encode proteins

Cells therefore need:

- the information to produce proteins in a regulated fashion;
- the ability to convey this information to daughter cells during cell division.

The key to these requirements is provided by the *DNA double helix*, which contains two strands of DNA held together by weak chemical interactions.

Basic DNA Structure

Each strand of DNA has a backbone of sugars and phosphates, with a nitrogen-containing base attached to each sugar. Four different bases are found in DNA. Cytosine (C) and thymine (T) are pyrimidines which contain one nitrogenous ring, whereas adenine (A) and guanine (G) are purines which contain two. The bases from each strand are linked together to form the 'rungs' inside the helix in such a way that A can only pair with T, and C can only pair with G (Figure 1.1). In RNA the sugar is ribose, uracil replaces T and the resulting nucleic acid is single stranded.

The strands complement each other – the sequences of bases on one strand can be determined from the sequence of the other strand. During cell division each strand independently forms a new complementary strand and the DNA helix is able to direct its own duplication.

The sequence of bases in a DNA molecule carries the information that specifies the order of amino acids along a polypeptide chain. Each of the 20 amino acids is encoded by coding units, or codons, which consist of three consecutive bases. Reading this code and translating it into protein requires RNA.

A segment of DNA that carries the information needed to encode a specific polypeptide is known as a gene. To retrieve this information a single-stranded messenger RNA (mRNA) copy of the gene is made and the sequence of bases in the mRNA is then translated into a linear sequence of amino acids, composing a polypeptide. Genetic information is therefore stored in cells in DNA. During the expression of a gene a segment of DNA is first transcribed into RNA and then translated from RNA into protein. During cell division DNA replicates itself to form two identical DNA helices.

DNA

DNA is composed of three principal components:

1 bases;
2 sugars;
3 phosphates.

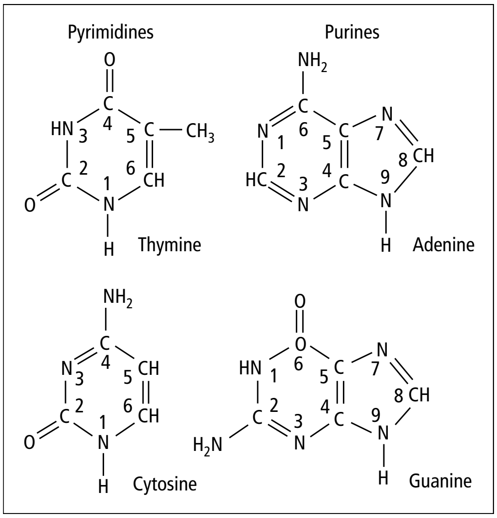
These are kept together by three principal types of linkage:

1 covalent bond;
2 hydrogen bond;
3 ester link.

The players

Bases

A base is a molecule that can combine with hydrogen ions in solution. The bases in DNA are nitrogen-containing rings (the nitrogen makes these molecules basic). Pyrimidines (C, T) have one ring, while purines (A, G) have two.



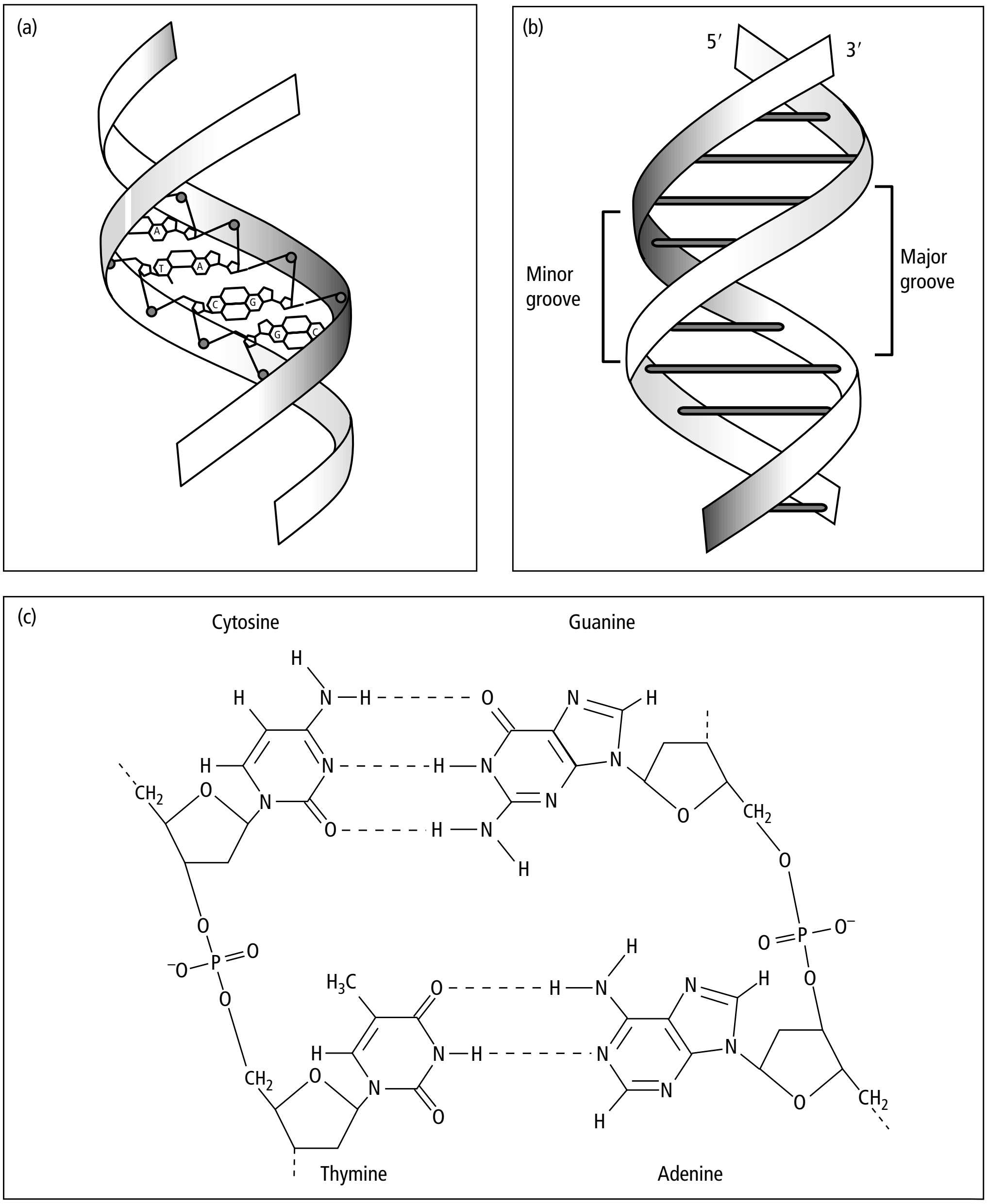
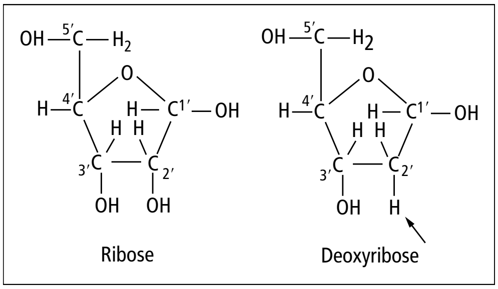


Figure 1.1 (a) Diagrammatic representation of the DNA helix. (b) The major and minor grooves of the DNA helix. (c) CG/TA base pairing.

Sugars

The sugars in DNA are pentoses (sugar molecules containing five carbon atoms). In DNA the pentose is always deoxyribose, indicating that it lacks an oxygen-containing hydroxyl group that is present in ribose, the parent compound. Ribose could not fit into a DNA helix as there is insufficient room for the 2′-OH group.

By convention the carbon atoms in ribose and deoxyribose are labelled by primed numbers (1′ to 5′) when part of a nucleotide. This labelling is important in understanding how the DNA molecule is assembled.

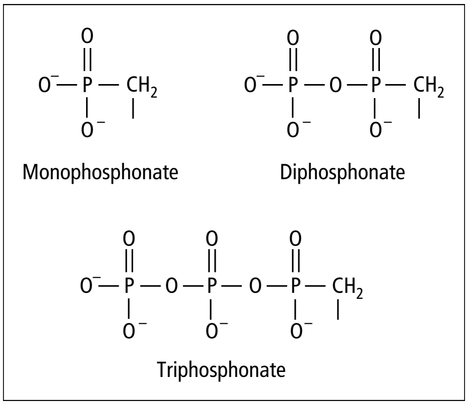


Ribose

Deoxyribose

Phosphates

The phosphates in DNA are either mono- or di- or triphosphates. The acidic character of nucleic acid is due to the presence of phosphate esters, which are relatively strong acids (molecules that release a hydrogen ion in solution). At neutral pH they dissociate from hydrogen ions and are thus normally referred to in their ionized form:



Monophosphonate

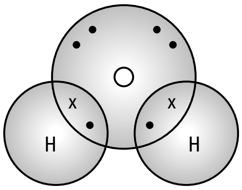
Diphosphonate

Triphosphonate

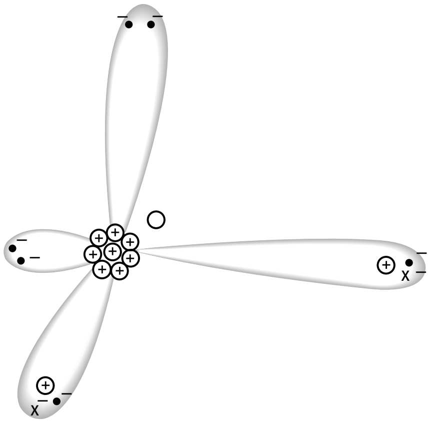
The ties that bind

Covalent bonds

A covalent bond exists between atoms that share electrons in their outermost shell. The bonding electrons move freely around both nuclei, which are held close together in a strong bond – energy is released when the bonds are formed, and the same amount of energy is required to break the bond.



In water each hydrogen atom shares its electron (x) with an electron from the outer shell of the oxygen atom (•).
The hydrogen atom thus fills its outer (first) shell with two electrons, and the oxygen atom fills its outer (second) shell with eight



Outer orbitals have characteristic shapes. The four outer orbitals of an oxygen atom in water point outwards, forming the corners of a tetrahedron

Hydrogen bond

A hydrogen atom can usually form only one covalent bond with another atom. A covalently bonded (*electron-depleted*) hydrogen atom can, however, form a weak electrostatic interaction (*hydrogen bond*) with an electronegative (*electron-rich*) atom (usually nitrogen or oxygen), for example,

N —— H --- O

Covalent bond   Hydrogen bond

Ester linkage

An ester link involves covalent bonding. It is formed when an alcohol and an acid unite with elimination of water.

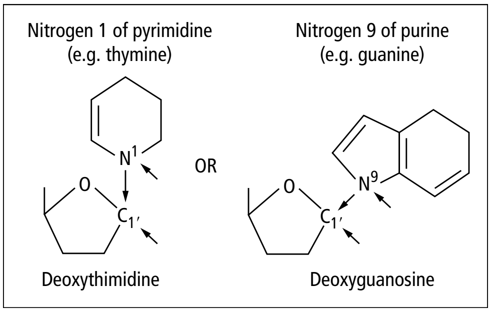
Bond Strength

The strength of the bonds is important in understanding the stability of different parts of the final DNA molecule. Strong covalent bonds link nucleic acids in a single DNA strand, whereas weaker hydrogen bonds hold the two complementary DNA strands together.

The formation of DNA

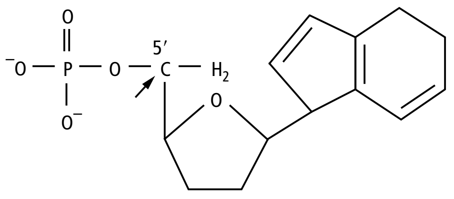
Base + ***s***ugar = nucleo***s***ide

The 1′ carbon of pentose ring is attached to nitrogen 1 of pyrimidine or nitrogen 9 of purine.



Base + sugar + phospha***t***e = nucleo***t***ide

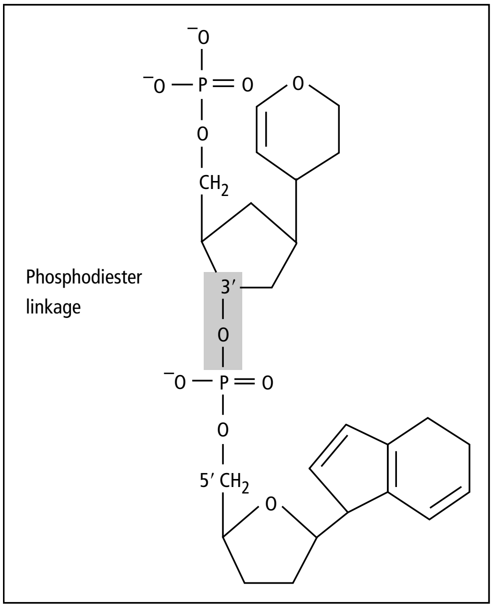
Phosphate is attached to the 5′-carbon of the pentose ring.



Deoxyguanosine 5′-phosphate

Nucleotides as Energy Stores

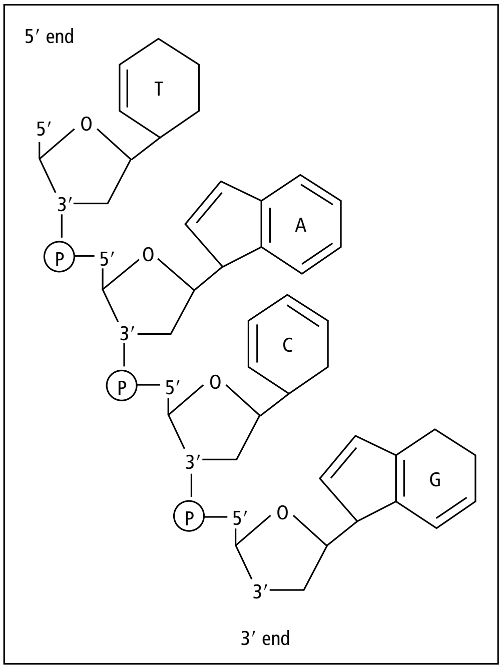
Nucleotides may have either one or two or three phosphates attached. In addition to forming the building blocks of DNA, the nucleotide di- and triphosphates are important stores of chemical energy; cleavage of the terminal diphosphate releases energy which is used to drive cell functions. Adenosine triphosphate (ATP) is the most widely used energy carrier in the cell.



Nucleotides join together to form nucleic acid

The hydroxyl group attached to the 3′-pentose carbon of one nucleotide forms an ester link with the phosphate of another molecule, eliminating a water molecule. The link between nucleotides is known as a phosphodiester link.

Thus, one end of a DNA strand has a sugar residue in which the 5′-carbon is not linked to another sugar residue (the 5′ end), whereas at the other end the 3′-carbon lacks a phosphodiester link (the 3′ end). This simple terminology is fundamental to understanding descriptions of how DNA replicates and is expressed.



DNA structure

The DNA helix

In the 1950s X-ray diffraction data suggested that DNA is helical (Figure 1.1(a)). In addition, biochemical data showed that the amount of A in DNA always equalled that of T, whilst the amount of G equalled that of C. These observations led Watson and Crick to propose the double helical structure of DNA, which could account for the physical properties of DNA and its replication in the cell.

The 'backbone' on the outside of the helix consists of alternating sugars and phosphates. The bases are attached to the sugars and form the 'rungs' of the helix.

As the distance between the sugar–phosphate backbone is fixed by the diameter of the helix, only two types of base pairs (bp) (AT or CG) can fit, explaining the constant regularity in the ratios between base pairs (A = T and G = C) (Figure 1.1(c)).

The strands are antiparallel (their 5′,3′-phosphodiester links run in opposite directions) and complementary (because of base pairing the chains complement each other). The sequences of bases on one strand can thus be deduced from the sequences of bases on the other, and each strand independently carries the information needed to form a double helix.

The DNA helix can take on several conformations. The most common form is *B-DNA*, in which the helix is right handed and has just over 10 bp per helical turn. There are two unequal grooves, the major and minor grooves (Figure 1.1(b)).

A-DNA is a right-handed helix which is shorter and wider than B-DNA. The phosphate groups bind fewer water molecules, and its formation is thus favoured by dehydration. The transition from B-DNA to dehydrated A-DNA was observed during the first X-ray studies of DNA over 50 years ago. The existence of the A-form in many protein–DNA complexes suggests that reversible B–A transition may be important for processing genetic information *in vivo*.

Z-DNA is a left-handed helix in which alternating purines and pyrimidines give rise to a zigzag appearance to the helix. Limited segments of Z-DNA occur *in vivo*. For example, sequences of

Cleavage of DNA Bonds

The relative weakness of the hydrogen bonds holding the complementary bp together is demonstrated by 'melting' the DNA. At increased temperatures the two strands separate: the DNA melts. The bonds holding the backbone of the helix together are stronger and do not melt but can be cleaved by enzymes derived from bacteria, which cut the backbone at specific sites. Bacteria use these enzymes as protective devices to degrade foreign DNA. They restrict the growth of viruses which infect bacteria (bacteriophages) and are known as restriction enzymes. A bacteriophage is a virus that infects bacteria sometimes referred to as a phage.

Describing a DNA Sequence

It is conventional to describe a DNA sequence by writing the sequence of bases in one strand only, and in the 5′ → 3′ direction. When identifying just two neighbouring bases in a sequence it is *usual to insert 'p'* between them to denote an intervening phosphodiester link (e.g. ApT). This is distinct from AT which indicates a hydrogen-bonded base pair on complementary strands.

Z-DNA can be induced at the 5′ ends of genes by transcription, where the Z-form may play a role in RNA processing.

DNA in eukaryotic organisms is organized into a nucleus

Living things may be divided into prokaryotes and eukaryotes. Prokaryotic organisms are simple, single-cell life forms that lack a distinct nucleus. Examples include bacteria and certain algae. Eukaryotes may be single-cell life forms such as yeasts or complex multicellular organisms such as plants and animals. The cells in eukaryotic organisms contain nuclei. DNA within the nucleus of eukaryotes is organized into chromosomes. Each chromosome contains an extensively folded DNA double helix.

Chromatin

The total length of all the strands of DNA in a human cell is ~2 m, all of which needs to be packed into a nucleus a few micrometers in diameter. This is achieved by the formation of a nucleoprotein complex called *chromatin* in which acidic phosphates in the backbone of DNA enable it to form ionic bonds with basic lysine- and arginine-rich proteins known as histones. Coiling of DNA around histone proteins allows long strands to be tightly packed into chromatin.

The core of the nucleosome contains two copies each of histone proteins named H2A, H2B, H3 and H4. A fifth histone, H1, protects the DNA linking the nucleosomes together. Histones can be modified by the addition of an acetyl, methyl or phosphate group in a manner that will alter chromatin structure and function. Variants of these histones encoded by different genes have also been described and shown to be important in functions such as DNA repair.

DNA is first packaged into a *nucleosome* (Figure 1.2), which consists of eight histone proteins around which a strand of DNA containing 146 bp is wound one and three-quarter times.

The histone protein H1 binds to DNA just next to each nucleosome and is involved in coiling DNA into chromatin fibres of 30 nm in diameter (Figure 1.3).

DNA in Prokaryotes

In prokaryotes all the DNA exists in a single molecule, which is circular. There are no 5′ or 3′ ends and no histones and there is no nucleus. The DNA can, however, be induced to supercoil into a compact structure around DNA-binding proteins by the enzyme DNA gyrase.

Chromosomes

During cell division chromatin becomes more condensed and can be recognized in the form of *chromosomes* by light microscopy. During metaphase in mitosis (Figure 1.9) each chromosome consists of two symmetrical *chromatids*, each containing DNA in which the chromatin fibres are folded in loops around a central scaffold of non-histone proteins.

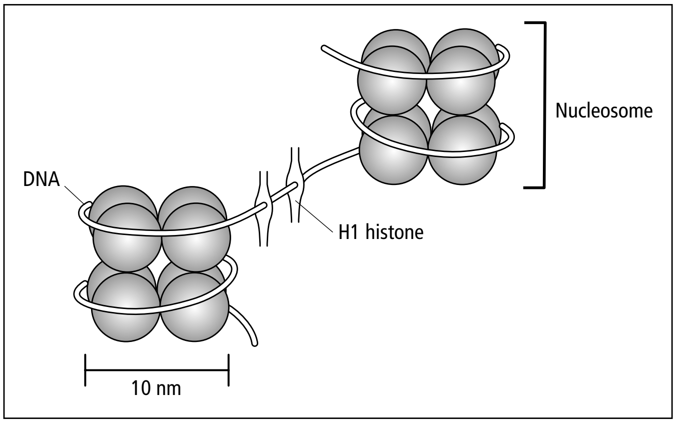


Figure 1.2 A nucleosome.

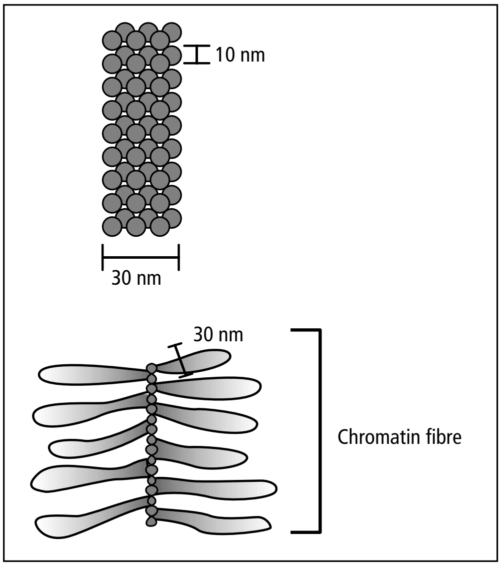


Figure 1.3 A chromatin 30 nm fibre.

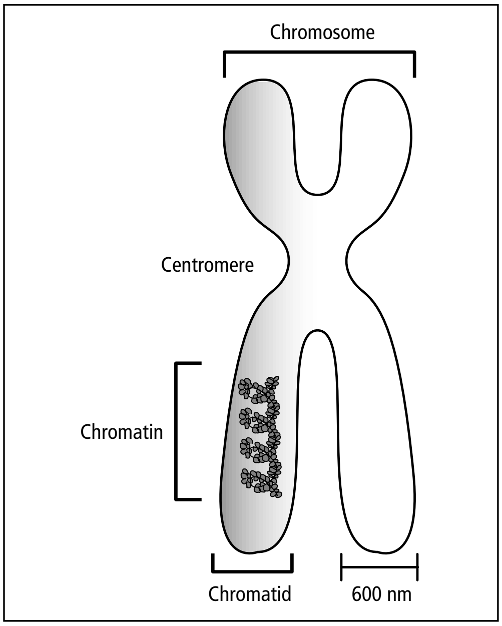


Figure 1.4 Chromosomes, chromatids and chromatin.

The chromatids are attached to each other at the centromere (Figure 1.4).

Condensed metaphase chromosomes can be visualized under the light microscope by various treatments which cause the appearance of light and dark bands. For example, staining with Giemsa gives rise to alternating dark- and pale-staining *G bands* (Figure 1.5). Such banding allows classification of sites on the chromosome according to their location on the short arm (p for petit), or long arm (q), and their position relative to the centromere. For example, the gene that encodes the β-globin chain of haemoglobin (which is abnormal in β thalassaemia), has been localized to the short arm of chromosome 11, in region 1, band 5, sub-band 5: written as 11p15.5 (Figure 1.5).

Chromosome Banding

G-banding with Giemsa stain yields the familiar pattern of ~500 light- and dark-stained bands at metaphase (Figure 1.6); Q-banding with *q*uinacrine, a fluorochrome, yields a fluorescent pattern very similar to that seen in G-banding.

Giemsa developed his staining technique in the early twentieth century. Fluorescent chromosome banding was introduced in 1969 by Caspersson and Zech. The banding patterns of chromosomes from humans, chimpanzees, gorillas and orangutans are remarkably similar.

Euchromatin and Heterochromatin

Euchromatin is genetically active and stains lightly with basic dyes. Heterochromatin is the darker staining condensed region of chromosomes that is characterized by the presence of highly repetitive sequences and relatively low gene density.

Centromeres, Telomeres and Arrays

The *centromere* is the site at which chromosomes constrict during metaphase. It separates the long and short arms of the chromosome.

The *telomere* forms the end of the chromosome.

In *tandemly repeated arrays* identical DNA sequences appear one after the other along a stretch of DNA.

Karyotype

Every species has a specific number and arrangement of chromosomes, which is referred to as a *karyotype*. Human cells contain 46 chromosomes, of which two are sex chromosomes (two X chromosomes in females, an X and a Y chromosome in males), and 44 are autosomes (22 matching pairs numbered 1–22) (Figure 1.6).

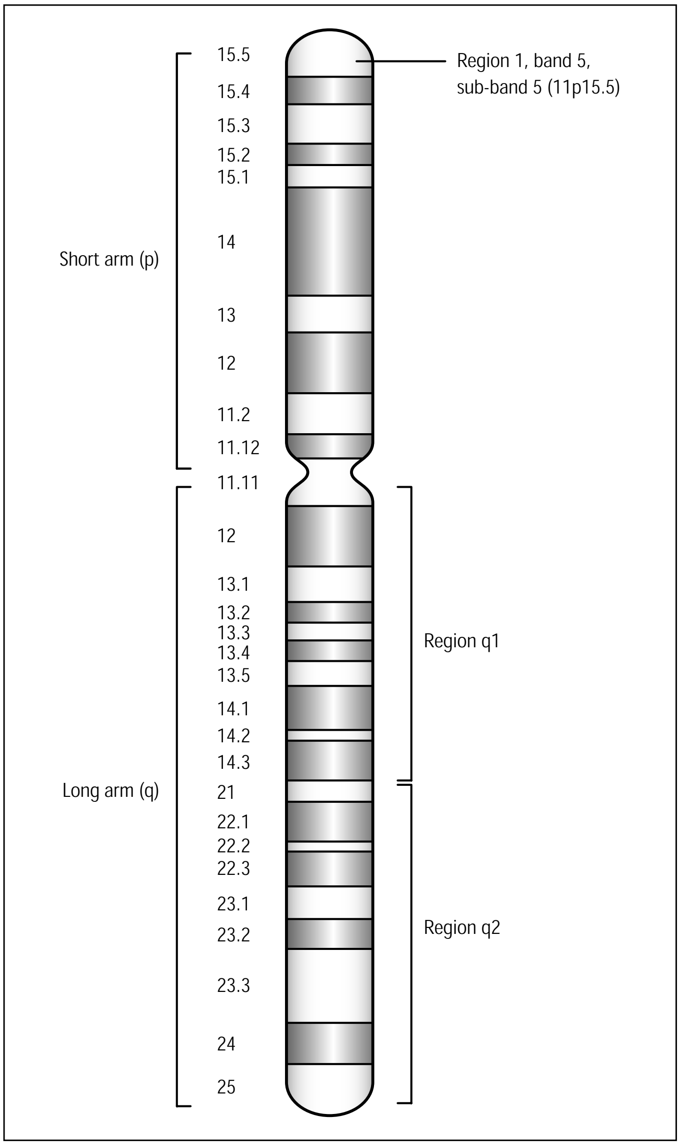


Figure 1.5 Chromosome 11 with G bands.

Genome

The complete genetic make-up of an individual is referred to as their *genome*. Thus, in human cells the genome is composed of 23 pairs of chromosomes within the nucleus, each chromosome containing a single, linear, double-helical strand of DNA. The human genome contains approximately 3×10^9 bp and is thought to contain about 23 000 different genes, most of which encode polypeptides. A small minority of genes encode RNA molecules that are not translated into proteins.

In addition to the nuclear genome, eukaryotic cells also contain a small mitochondrial genome which one tends to inherit from the mother. This is because, unlike sperm, eggs have a considerable amount of cytoplasm which contains mitochondria. In humans the mitochondrial genome consists of a 16 569-bp

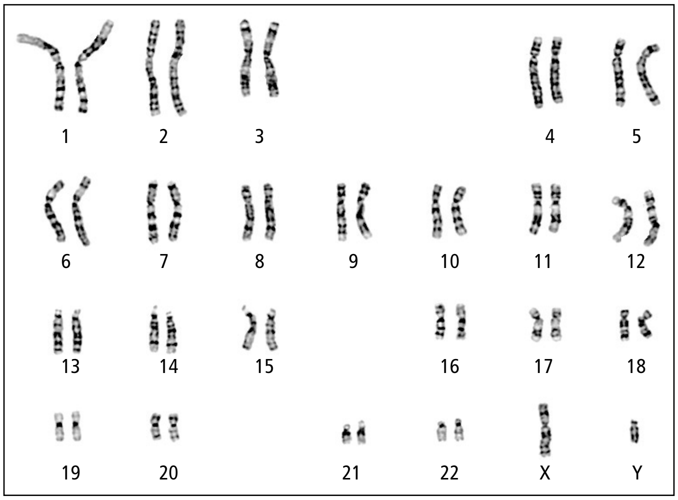


Figure 1.6 A normal human male karyotype stained with Giemsa.

circular DNA molecule which encodes proteins essential for mitochondrial structure and function, including oxidative enzymes, together with RNA molecules involved in mitochondrial protein synthesis (*see* Chapter 2). Although mitochondria possess their own genome, the majority of mitochondrial proteins are encoded by nuclear genes.

DNA replication

The double helical structure of DNA provides a mechanism by which nucleic acids can accurately replicate. Each DNA strand serves as a template for the synthesis of a new complementary strand of DNA. DNA replication is *semiconservative* – one strand (half of the original DNA) is retained ('conserved') in the new DNA molecule.

Much of our understanding of how DNA replicates has come from the study of the bacterium *Escherichia coli*, in which DNA is present as a single circular molecule. Replication starts at a specific site (the origin of replication of the *E. coli* chromosome is termed *oriC*), and proceeds sequentially in opposite directions by formation of discontinuous fragments, which are then joined together by the enzyme DNA ligase. The enzyme polymerase assembles the nucleotides on the new DNA strand and the enzyme exonuclease edits and corrects the process by removing unwanted nucleotides.

During DNA replication the following processes occur (Figure 1.7):

- The double-stranded helix must first unwind and each strand then acts as a template. DNA helicase stimulates separation of the two strands and DNA gyrase aids unwinding. DNA binding proteins then bind to and stabilize the single-stranded structure. This exposes the DNA *template* containing a region of single-stranded DNA from which a complementary copy is made.
- A short strand of RNA, known as a *primer*, is synthesized by an enzyme known as *RNA polymerase* or *Primase* on the DNA template at the start of replication, and removed at the end – *RNA primes the synthesis of DNA*.
- *DNA polymerases* (Pols) catalyse the addition of nucleotides to the primer RNA forming a new strand of DNA. DNA Pols produce a link between the inner phosphorus of the nucleotide and the 3′-OH group of the primer – elongation occurs in the 5′–3′ direction.
- Replication starts at a specific site and proceeds sequentially in opposite directions, even though synthesis can only occur in the 5′–3′ direction. This apparent paradox was resolved by the demonstration that synthesis of one strand occurs continuously, whereas the other strand is synthesized in short 5′ → 3′ fragments, known as Okazaki fragments.