1

Molecular and cell biology

Julian R. E. Davis

Key points

- The human genome functions through transcription of coding regions (exons) of DNA to messenger RNA (mRNA) and translation of mRNA into protein.
- Normal cell function and growth are controlled by intracellular signalling systems that couple external stimuli to cellular responses.
- Gene expression can be studied *in vitro* by cloning of DNA, sequence analysis, analysis of DNA–protein interactions *in vivo* and by gene transfer, including the use of transgenic animals.
- Such studies have led to the identification of genetic defects in diseases such as cystic fibrosis and the genetic events underlying cancer formation; they have also directed the first applications of gene therapy.
- Molecular and cell biology, as techniques used to elucidate normal cellular mechanisms, are themes that run through each chapter of this book as mechanisms of disease are explored.

Molecular biology

The human genome and gene structure

The genome comprises all the inherited material passed on from one generation to the next. In humans, it consists of the 23 pairs of chromosomes in the nucleus together with a small amount of mitochondrial DNA. Chromosomal DNA is a tightly packaged array of genes (the units of inheritance) together with long tracts of intergenic DNA whose function is still largely unknown. The overall organisation and structure of the human genome is under intense study at present, and this chapter focuses only on small parts of this overall structure in order to describe some of the essential elements of molecular biology involved in mechanisms of disease. In Chapter 3, changes in chromosomes and genes relating to hereditary diseases are described. Later sections of this chapter outline intracellular signalling systems, especially as they relate to the control of gene expression and regulation of growth. Finally, some of the methods of analysis currently used in molecular and cell biology are reviewed.

The basic chemical composition of chromosomes – DNA and protein – was generally understood long before it was clear which functioned as genes. Work by Griffith and Avery showed that DNA was the most likely candidate, and experiments by Hershey and Chase in 1952 showed that only DNA from bacteriophages entered the host cell and initiated the production of viral particles. The double helix structure of DNA, defined by Watson and Crick, comprising two antiparallel strands with the sugar phosphate backbones on the outside and the purine–pyrimidine base pairs (bps) on the inside, forms the basis of our concepts of the replication and utilisation of genes. A consequence of this hydrogen-bonded pairing of the two DNA strands is that the strands can be separated by conditions that break hydrogen bonds (heat or extremes of pH) and then allowed to rejoin or anneal under less stringent conditions. Because of the obligatory complementary binding, strands that are complementary in sequence will bond to form a double helix. This is the basis of many of the experimental studies that allow related genes to be identified and probes of known sequence to be 'annealed' or bound to sections of DNA (see below).

The linear sequence of nucleotides forms a genetic code in which triplets of nucleotides code for each amino acid. The possible permutations of nucleotides allow some degeneracy (more than one code for an amino acid) and for start and stop codons.

Chromosomes are generally organised in pairs – one inherited from each parent – and in each pair, genetic material undergoes rearrangement by the crossing over of paired segments during meiosis in paternal or maternal gametes. Thus, each chromosome contains newly arranged genetic material, but with conserved overall structure in homologous pairs of chromosomes. Most genes will be represented by a maternal and a paternal 'allele', (alternative forms of the same gene), which in turn can undergo pairing in meiosis to form the next generation of gametes.

Genes contain structural information that ultimately dictates the sequence of a protein; for example, a peptide hormone, an enzyme, or a structural protein. The linear sequence of deoxynucleotides determines the properties of the gene and its protein product. However, this protein-coding information is not an unbroken stretch of DNA but instead consists (in most mammalian genes) of separate coding regions, exons interspersed with non-coding introns. Each gene also contains characteristic flanking sequences both upstream and downstream of the coding region (Fig 1.1).

The function of much of the intergenic DNA is unknown: it occupies a large percentage of the genome, proportionately more in humans than in simpler organisms. Some of this material has obvious structural functions; for example, several megabases of DNA in the centromere of most chromosomes are involved in the formation of the mitotic spindle in cell division. Other parts of chromosomes contain long stretches of repetitive non-coding sequences, such as tandem repeats, whose function is still unknown.

Repetitive sequences have, in some cases, been found to have major significance, as illustrated by the discovery of genetic alterations in repetitive DNA in a number of disorders, exemplified by Huntington's disease. The Huntington's disease gene on chromosome 4 contains an expanded unstable region of DNA comprising a series of CAG repeats whose overall length changes during gamete formation: this repetitive DNA stretch is longer in patients with Huntington's disease than in non-affected people, and the more CAG repeats that occur, the earlier the disease develops. Such alterations in the length of this DNA region may favour the assembly of *nucleosomes*. These are complexes of chromosomal DNA with histone proteins that hinder the access of regulatory proteins to control elements of genes; hence

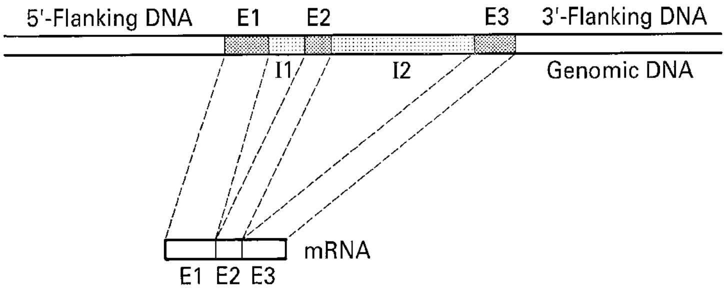

Figure 1.1 Gene structure. Exons (E1, E2 and E3) contain the sequences that code for proteins. Exons are separated by regions of non-coding DNA, introns (I1 and II2).

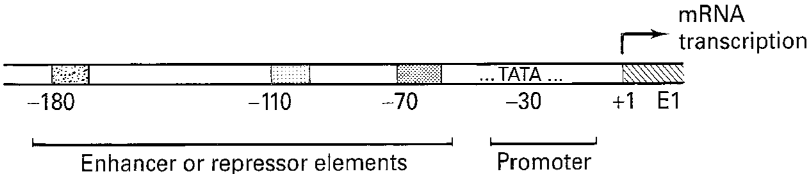

Figure 1.2 Structure of the 5′-flanking DNA. Boxes represent DNA elements to which diffusible protein factors can bind. These occur at different points upstream of the transcriptional start site (right-angled arrow), which is known as nucleotide +1. Transcription of the exon occurs from +1. The non-coding region is numbered back from the start site with nucleotides numbered –1upwards.

they will mediate general repression of gene transcription.

The non-coding DNA that is closely associated with genes themselves, upstream and downstream flanking regions, are now known to be the major regulatory elements that modulate both gene transcription and the stability of mRNA; this is considered in more detail in the following section.

Gene transcription

The upstream flanking DNA in most genes contains sequences involved in regulation of transcription. This stretch of DNA contains a number of specific sequences, known as *cis* elements because they can affect only adjacent genes that bind diffusible nuclear proteins (*trans* elements) involved in transcription. The minimal upstream element necessary for transcription to occur is in many cases a clearly defined region of less than 30 base pairs of DNA and is termed the gene's *promoter*. Additional upstream elements may stimulate or inhibit the process of transcription and are termed *enhancers* or *repressors* (Fig. 1.2). The transcription of mRNA from the genomic DNA occurs via the enzyme RNA polymerase II, which starts the process at a point on the gene determined by specific sequences in the promoter region, such as a TATA motif (the TATA box). A complex of protein transcription factors (designated TFIIA, TFIIB, etc.) is established at this transcriptional start site, initiated by the binding of the factor TFIID, which binds to the TATA element itself. The cluster of TF proteins thus built up then allows RNA polymerase II to bind tightly as part of this transcriptional initiation complex (Fig. 1.3)

Upstream of the minimal promoter, enhancer elements may be extensive. These are often several thousand base pairs distant and contain a large number of characteristic sequences that serve as recognition motifs for other transcription factors. Enhancers have the ability to influence the rate of gene transcription at a variable distance and in either orientation relative to the promoter. Their transcription factors modulate the rate of transcription of the gene: some are specific to the cell type whereas others are ubiquitous but tightly regulated by intracellular signalling systems. The nature of their interaction with the transcription initiation complex is not fully understood, but it probably involves looping of DNA in order to bring distant enhancer elements into proximity with the promoter to modulate the function of the transcriptional machinery (Fig. 1.3).

Transcription factors

A series of families of transcription factors have been described with distinct regions of the protein (domains) involved in DNA binding or in transcriptional activation. Several different classes of DNA-binding domains are now recognised, including 'zinc fingers', 'leucine zippers' and helix-turn-helix motifs; more are being identified.

Zinc fingers

These are found in many transcription factors, including steroid receptors, and consist of peptide

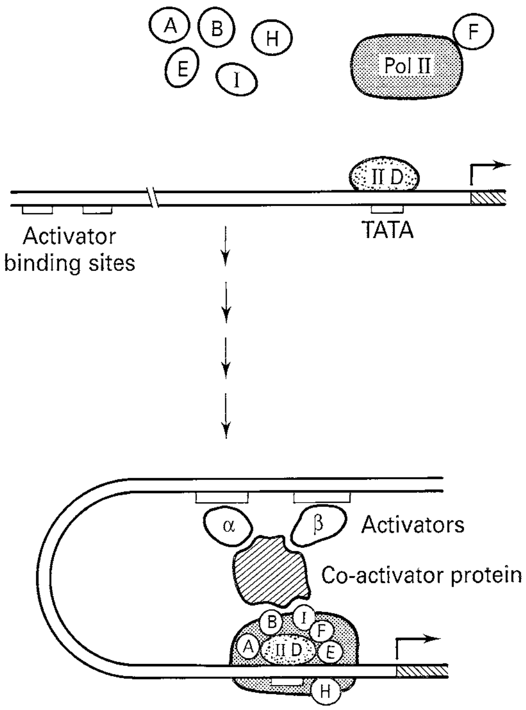

Figure 1.3 Simplified model of the transcriptional initiation complex. Transcription factors TFIID, TFIIA, TFIIB, etc. sequentially bind at the TATA box and stabilize RNA polymerase II (Pol II) attachment. Upstream DNA –bound transcription factors (*α*, *β*) may interact directly or indirectly (via a co-activator) with the transcriptional complex. This probably involves some form of DNA looping.

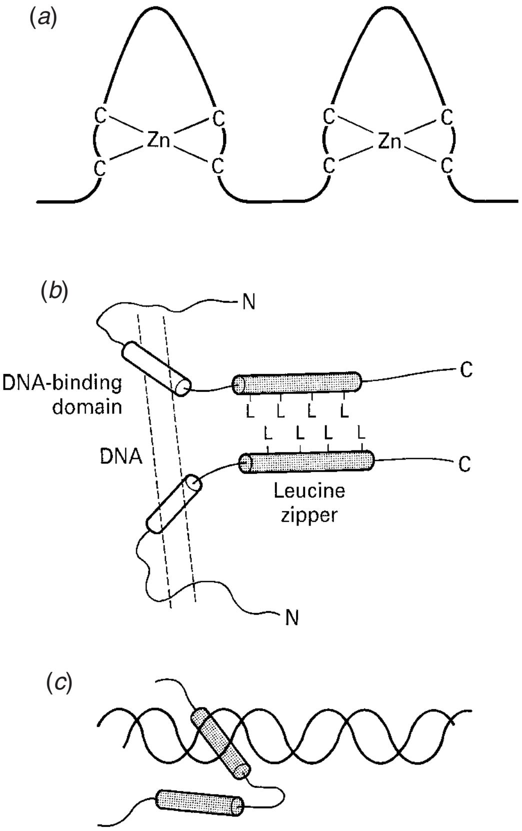

Figure 1.4 Transcription factor structures. (*a*) Zinc fingers. (*b*) Leucine zipper, allowing dimerisation of two factors. (*c*) Helix-turn-helix motif orientated against a DNA helix.

loops in which an atom of zinc is tetrahedrally co-ordinated by cysteine and histidine residues at the base of the finger (Fig. 1.4*a*). Usually, there are several zinc fingers in transcription factor proteins, and the tips of the fingers (containing basic amino acids) are thought to contact the acidic DNA by poking into the major groove of the double helix.

Leucine zippers

These domains have been identified in several transcription factors (e.g. Jun, Fos and Myc) and are regions in which every seventh amino acid is leucine. In an *α*-helical structure, the leucines occur every second turn, and their long side chains can interdigitate with those of an analogous helix in a second protein, like a zipper, allowing dimerisation of the two proteins (Fig. 1.4*b*). Leucine zippers are important not only for transcription factor dimerisation but also for DNA binding; they allow the formation of either homo- or heterodimers among related proteins, for example, Jun–Jun and Jun–Fos.

Helix-turn-helix motifs

These comprise two *α* helices separated by a *β* turn. One of the helices, the ‘recognition helix’, lies in the major groove of the DNA and provides the DNA sequence specificity of binding, the second lies across the major groove and probably stabilises the

DNA contact (Fig. 1.4*c*). An example of this type of transcription factor is the pituitary-specific factor Pit-1/GHF-1.

Activation domains

These are less clearly defined in transcription factors than the DNA-binding structures but may contain characteristic acidic domains or proline- or glutamine-rich domains. Their function has been confirmed by 'domain-swap' experiments in which chimaeric factors are constructed with the DNA-binding region of one factor linked to the activation domain of another. The mechanism of transcriptional activation is still not well understood but may involve direct contact between the activation domains and the components of the transcription initiation complex (e.g. TFIID, TFIIB, or RNA polymerase II itself) or, in some cases, indirect contact via intermediate adaptor proteins.

Repression

Although most transcription factors seem to be activators, in some cases they can *repress* gene transcription, and several mechanisms are possible. For example, a negatively acting factor can simply interfere with the effect of an activator by occupying the activator protein's binding site (or a closely adjacent site) on the target DNA. In other cases, two factors may interact such that a positively acting factor is sequestered by dimerisation. For example, the glucocorticoid receptor can be prevented from *trans*-activating target genes by becoming complexed with the factor AP-1.

Regulation of gene transcription by intracellular signalling systems is an important aspect of transcriptional control and this is one of the best characterised systems. Most of the steps between an external stimulus and a cellular response have been defined. This is discussed in more detail later in this chapter.

Control of transcription: tissue specificity

Differentiation of tissues with a variety of distinct phenotypes requires the expression of particular genes in a cell type–specific manner; a number of tissue-specific transcription factors have recently been identified in addition to those that are ubiquitous. For example, the factor MyoD is a transcription factor expressed only in differentiating myoblast cells, and artificial expression of MyoD alone in undifferentiated fibroblast cell lines can induce this differentiation process. MyoD can either form transcriptionally active homodimers or it can heterodimerise by helix-loop-helix interaction with other proteins with varying effects. One such protein, Id, is a negative regulator that lacks a DNA-binding domain and so prevents MyoD from binding to DNA. Levels of Id decline during differentiation; therefore the overall effect of the tissue-specific factor MyoD will depend on its interaction with changing levels of other factors, such as Id, that determine its transcriptional activity.

Another tissue-specific transcription factor is Pit-1/GHF-1, which is expressed in the differentiating fetal pituitary gland. This factor is necessary not only for pituitary-specific expression of the peptide hormones prolactin and growth hormone but also for the development of the respective lactotroph and somatotroph cell types. Recently, a number of cases have been described of loss-of-function mutations of Pit-1/GHF-1 in which patients are hypopituitary, with pituitary hypoplasia as well as prolactin and growth hormone deficiency.

mRNA and protein synthesis

The process of gene transcription occurs by complementary base pairing from the genomic DNA template to form a primary transcript of heterogeneous nuclear RNA (hnRNA), which contains both exonic and intronic sequences. This RNA forms the precursor to mRNA, whose message is carried in the genetic code of nucleotide triplets (codons), each specifying one of the usual 20 amino acids. The hnRNA is then processed to form mature messenger RNA (mRNA) by a series of steps (Fig. 1.5).

1. A methylated guanosine residue (m^7Gppp) is added at the 5′ end (the 5′ cap) at the start of the first exon's untranslated leader sequence; the

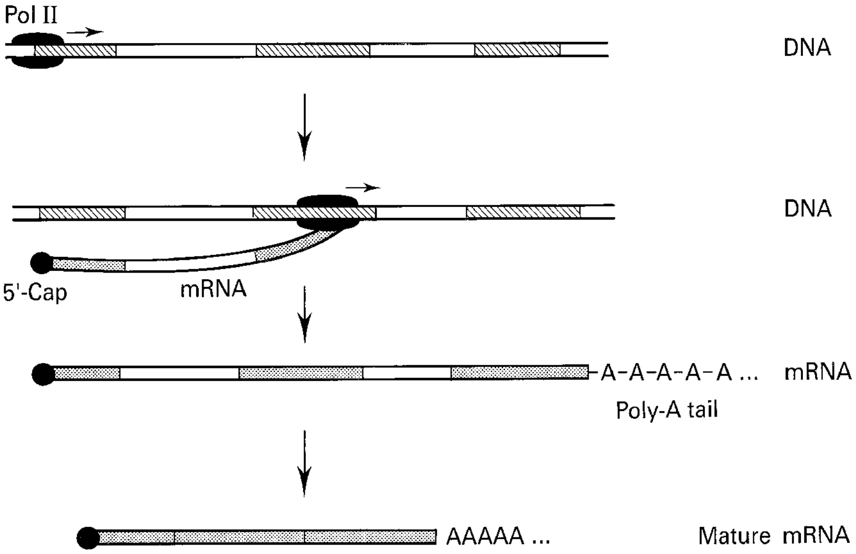

Figure 1.5 Formation of mature mRNA. Double-stranded DNA is transcribed by RNA polymerase II (Pol II) and the mRNA is modified at its 5′-end by the addition of a G residue with a triphosphate bond (Gppp) that is methylated (5′-cap). Splicing removes intronic sequences and a 'poly-A tail' is added.

end of this untranslated region is marked by an AUG initiation codon encoding the first methionine residue of the protein.

2. Intronic regions are removed and the exons spliced together by a 'spliceosome' complex of small ribonucleoproteins, the splice sites being marked by characteristic GU and AG splice donor and acceptor sites at the beginning and end of intron transcripts.
3. A long tract of 100 or more A residues (the 'poly-A tail') is added at the 3′ end, signalled by characteristic polyadenylation sequences (such as AAUAAA) downstream of the stop codon (UGA, UAA, or UAG) that terminates protein translation.

These features of mature mRNA are important for stability and translocation. In particular, the structure of the 5′ and 3′ untranslated regions appears to be significant, with secondary structures such as hairpin and cruciate loops affecting the rate of peptide translation in the ribosome. The process of splicing also appears to have a significant function, allowing the cell to select which exons will be represented. Therefore alternative splicing of the primary transcript can generate alternative gene products, as in the case of calcitonin and the calcitonin gene–related peptide (CGRP), which are encoded by one gene. The gene encoding the transcriptional repressor CREM (cyclic AMP-response element modulator; see pages 11–12) similarly can be alternatively spliced to produce two forms of the factor with different DNA-binding domains.

Protein synthesis: mRNA translation

mRNA is translated into protein in the cytoplasm by a complex process of matching nucleotide sequences to amino acids. These are polymerised to form the polypeptide chain at the ribosomes.

Ribosomes

These are large multimolecular complexes of many different proteins associated with several structural

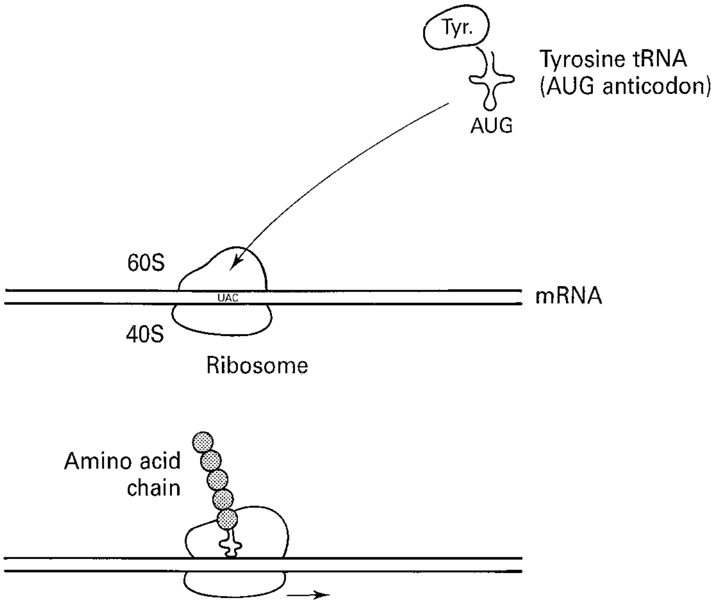

Figure 1.6 Translation of mRNA. tRNAs enter the ribosome and bind to the mRNA by matching of their anticodons. Peptide bonds are formed between the amino acids to give a growing amino acid chain with its sequence defined by the nucleotide sequence of the mRNA.

RNA (ribosomal RNA, rRNA) molecules. These complexes act to position transfer RNA (tRNA) molecules sequentially to match the triplet code of mRNA. Each ribosome comprises two subunits, one large and one small. The small 40S subunit contains a single (18S) rRNA molecule with over 30 proteins, and the large 60S subunit contains three different rRNAs and over 40 proteins. The overall assembly has a molecular weight of 4.5 million daltons (Da) and forms a particle visible by electron microscopy. The three-dimensional structure allows this molecular machine to engage both a strand of mRNA and a growing peptide chain (Fig. 1.6).

tRNA

Molecules of tRNA are essentially adaptors that recognise both a mRNA nucleotide sequence *and* an amino acid sequence. They are single polynucleotide chains, 70 to 90 bases in length, that undergo internal base pairing to form a complex with exposed nucleotide loops. One such loop contains the 'anticodon' that can base-pair with the corresponding codon in mRNA, while the exposed 3′ end of the tRNA molecule is attached covalently to a specific amino acid (Fig. 1.6).

Translation

The process of translation is rapid, a single ribosome taking only 1 min to polymerise over 1000 amino acids. The ribosome binds to a specific site on the mRNA, allowing the first 'initiator tRNA' to bind to the AUG initiation codon and start the peptide chain with the initial methionine residue. Subsequently, the ribosome moves along the mRNA translating codon by codon with a series of tRNAs adding amino acids to the growing peptide chain. When the end of the message is reached at the stop codon, the ribosome subunits are released along with the newly made peptide.

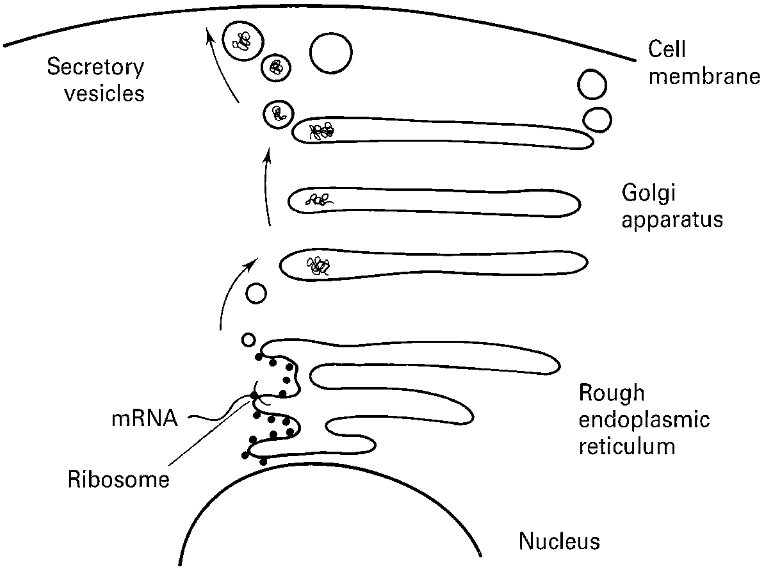

Figure 1.7 Cellular organelles involved in protein synthesis. Ribosomes attached to rough endoplasmic reticulum synthesise peptide chains that contain signals allowing entry to the endoplasmic reticulum. These leader sequences are later cleaved off and the protein is routed through the Golgi apparatus to secretory vesicles or other cellular destinations.

Protein secretion

The fate of the protein product of mRNA translation depends upon the nature of the protein and the cell type. For proteins such as peptide hormones that are exported by the cell into the extracellular fluid, specialised secretory processes are involved. Cells can secrete hormones 'constitutively', in a continuous manner unaffected by external stimuli and dependent only on the rate of transcription and translation, or via a 'regulated' pathway, using secretory granules to package and store hormone until an external or internal stimulus causes exocytosis.

Secreted proteins are synthesised on ribosomes attached to rough endoplasmic reticulum. They then enter the lumen of the endoplasmic reticulum by the binding of a hydrophobic leader sequence of the peptide with a 'signal recognition particle' to a docking protein on the endoplasmic reticulum surface. The proteins are then transported to the Golgi complex, where they undergo post-translational modification, such as glycosylation, before being concentrated into granules which are pinched off from the Golgi membrane (Fig. 1.7). Finally, under the influence of secretory stimuli, the granules 'marginate' and fuse with the cell membrane, allowing their contents to be released into the extracellular fluid, a process known as exocytosis.

Post-translational processing

Post-translational processing of proteins is an important regulating process that modifies the biological activity of many proteins. It occurs largely in the endoplasmic reticulum, the Golgi apparatus, and the cytoplasm, but it can also occur within the secretory granule.

An important initial modification is the *three-dimensional folding* of the new polypeptide chain, which is largely determined by the array of hydrophobic or hydrophilic amino acid side chains.

Particular patterns of protein folding have been confirmed by X-ray crystallography to occur in many different proteins, namely the α *helix*, a rigid cylinder formed by a spiral of amino acid residues, and the β *sheet*, formed by alignment of antiparallel or parallel straight chains of amino acids.

The folded protein conformation may be stabilised by the formation of covalent bonds between or within chains by *disulphide bridges* between nearby cysteine-SH groups.

Further covalent post-translational modifications include *phosphorylation*, catalysed by protein kinases that transfer a high-energy phosphate group from ATP to specific amino acid sequences in proteins, and *glycosylation*, the addition of complex carbohydrates to particular residues, often asparagine (N-linked oligosaccharides) or the hydroxy groups of serine or threonine (O-linked oligosaccharides). Other modifications include the aggregration of protein subunits to form multimers, the attachment of co-enzymes such as biotin to some enzymes, and acetylation and hydroxylation of certain amino acids.

Cellular signalling and growth regulation

Cells respond to a series of extracellular stimuli such as hormones, growth factors, and neurotransmitters. Some agents, such as steroid and thyroid hormones, are able to enter the cell and bind to intracellular receptors that in turn bind to DNA as transcription factors, directly altering the transcription of target genes. However, many other factors (such as peptide hormones) are unable to enter the cell and instead must stimulate a receptor on the cell membrane to trigger an intracellular ‘second messenger’. This then generates a cellular response. This process is termed signal transduction, and a variety of intracellular signalling processes have been discovered.

The systems are complex and can be viewed as molecular cascades comprising receptors, transducing proteins (G proteins), effector proteins, second-messenger molecules, protein kinases and kinase substrates. The complexity of these systems allows for great amplification within the cell of an initial extracellular signal, and also for interaction and co-regulation of parallel signalling pathways. The corollary is that the genes for some of the many proteins involved are subject to mutations that result in human disease, including tumour formation. Indeed these genes are in many cases known as ‘proto-oncogenes’, the normal cellular homologues of viral ‘oncogenes’ that cause cancers (see pages 18–21).

Membrane receptors

Peptide hormones, catecholamines, growth factors and neurotransmitters bind to specific cell surface receptors, which are coupled to intracellular signalling pathways in a variety of ways.

G protein–linked receptors

A very large number of membrane receptors are coupled to second messenger–generating systems via intermediate *transducers*, ‘G proteins’ (see below), which in turn are linked to *effector* molecules that generate the intracellular second messenger. Molecular cloning has shown that these G protein–linked receptors belong to a superfamily of proteins that have similar structures. They are characterised by seven hydrophobic α helices traversing the membrane, with an extracellular amino-terminal and an intracellular carboxy-terminal, three intracellular loops thought to couple to the G proteins, and three extracellular loops involved in ligand binding (Fig. 1.8). The G protein–linked receptors in general operate to initiate the generation of diffusible small molecules such as cyclic adenosine monophosphate (cAMP), which in turn activate protein kinases.

G protein–independent receptors

Not all receptors are linked to G proteins, and a number of transmembrane receptors possess intrinsic intracellular effector domains without intermediate transducing proteins. Some such receptors, such as the epidermal growth factor (EGF) receptor,

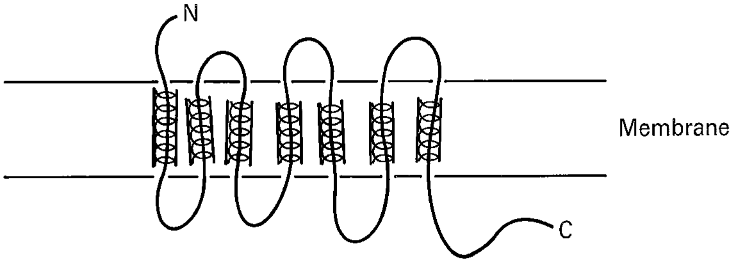

Figure 1.8 A membrane receptor with seven α-helical transmembrane domains.

have a single polypeptide chain, while others, such as the insulin receptor, have linked α and β subunits. Some of these receptors possess intrinsic tyrosine kinase activity, allowing them to phosphorylate (and hence activate) target intracellular proteins, while others are closely associated with separate tyrosine kinase proteins – for example, the 'Janus kinases' such as Jak-2, linked to the erythropoietin and growth hormone receptors. Thus, these receptors are able directly or indirectly to initiate a cascade of protein phosphorylation as their mechanism of action without necessarily generating intermediate second messengers. A typical phosphorylation cascade of this sort is illustrated in Fig. 1.9.

Nuclear receptors

Steroid and thyroid hormones, vitamin D and retinoic acid are small lipophilic molecules that are membrane-soluble and interact directly with intracellular receptor proteins. These receptors exist in the cytoplasm complexed with 'chaperone' molecules (for example, heat-shock protein 90, or hsp-90), from which they dissociate on activation by the hormonal ligand. After dissociation, they change conformation and translocate to the nucleus.

Again, molecular cloning has shown that there is a large superfamily of nuclear receptors that function as ligand-activated transcription factors. Some of these receptors have no identifiable ligand and have been termed 'orphan receptors'. Nonetheless, despite having widely differing ligands, the nuclear receptors have remarkable structural similarity, with six identifiable domains (A to F), including conserved DNA-binding domains with two zinc-finger motifs (see above) and a hormone-binding domain (Fig. 1.10).

The receptors for oestrogen and glucocorticoid activate gene transcription as homodimers bound to short, palindromic DNA response elements (for example, a typical oestrogen response element would be 5′-GGTCAnnnTGACC-3′, the palindrome being apparent on the complementary strand in the reverse direction). The other members of the family form heterodimers with a different protein, the retinoid X receptor (RXR), whose ligand is 9-*cis*-retinoic acid; the nature of the complexes that form on DNA is determined by the arrangement of the response elements in the enhancer regions of gene promoters, usually as short direct-repeat or inverted-repeat sequences with variable spacing of one to five nucleotides.

G proteins

G proteins are a large family of membrane-associated transducing proteins that are linked to transmembrane receptors, as described above. G proteins are so named because they bind guanosine triphosphate (GTP) and are themselves members of a larger superfamily of GTP-binding proteins (which includes the *ras* proto-oncogene product p21) whose general function in cells is proposed to be one of molecular switching. The receptor-associated G proteins fulfil exactly this role, conveying an 'on' signal from a newly occupied receptor to switch an intracellular effector protein into its activated state.