8

Protein synthesis, processing and turnover

Uncovering the cellular mechanisms resulting in sequential transfer of information from DNA (our genes) to RNA and then to protein represents one of major achievements of biochemistry in the 20th century. Beginning with the discovery of the structure of DNA in 1953 (Figure 8.1) by James Watson, Francis Crick, Maurice Wilkins and Rosalind Franklin this area of biodiversity has expanded dramatically in importance. The information required for synthesis of proteins resides in the genetic material of cells. In most cases this genetic material is double-stranded DNA and the underlying principles of replication, transcription and translation remain the same throughout all living systems.

The information content of DNA resides in the order of the four nucleotides (adenine, cytosine, guanine and thymine) along the strands making up the famous double helix. The structure of DNA is one of the most famous images of biochemistry and although this chapter will focus on the reactions occurring during or after protein synthesis it is worth remembering that replication and transcription involve the concerted action of many diverse proteins. The transfer of information from DNA to protein can be divided conveniently into three major steps: replication, transcription and translation, with secondary steps involving the processing of RNA and the modification and degradation of proteins (Figure 8.2).

Cell cycle

Cell division is one part of an integrated series of reactions called appropriately the cell cycle. The cell cycle involves replication, transcription and translation and is recognized as a universal property of cells. Bacteria such as *Escherichia coli* replicate and divide within 45 minutes whilst eukaryotic cell replication is slower and certainly more complicated, with rates of division varying dramatically. Typically, eukaryotic cells pass through distinct phases that combine to make the cell cycle and lead to mitosis. The first phase is G_1, the first gap phase, and occurs after cell division when a normal diploid chromosome content is present. G_1 is followed by a synthetic S phase when DNA is replicated, and at the end of this synthetic period the cell has twice the normal DNA content and enters a second gap or G_2 phase. The combined G_1, S and G_2 phase make up the interphase, a period first recognized by cytogeneticists studying cell division. After the G_2 phase cells enter mitosis, with division giving two daughter cells each with a normal diploid level of DNA and the whole process can begin again (Figure 8.3).

Mitosis is accompanied by the condensation of chromosomes into dense opaque bodies, the disintegration of the nuclear membrane and the formation of a specialized mitotic spindle apparatus containing

Proteins: Structure and Function by David Whitford

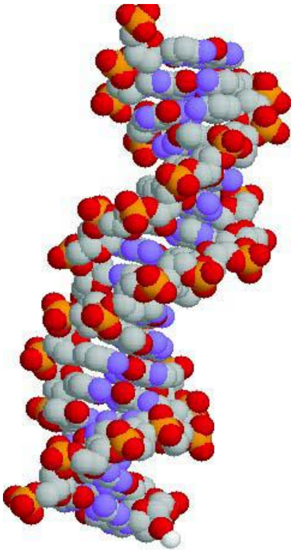


Figure 8.1 The structure of DNA. The bases lie horizontally between the two sugar-phosphate chains. The positions of the major and minor grooves are defined by the outline of phosphate/oxygen atoms (orange/red) arranged along the sugar-phosphate backbone. The pitch of the helix is ~3.4 nm and represents the distance taken for each chain to complete a turn of 360°

microtubule elements that directs chromosome movements so that daughter cells receive one copy of each chromosome pair. After mitotic division the cell cycle is complete and normally a new G_1 phase starts, although in some cells G_1 arrest is observed and a state called G_0 is recognized.

Proteins with significant roles in the cell cycle were discovered from cyclical variations in their concentration during cell division and called cyclins (Figure 8.4). The first cyclin molecule was discovered in 1982 from sea urchin (*Arbacia*) eggs, but subsequently at least eight cyclins have been identified in human cells.

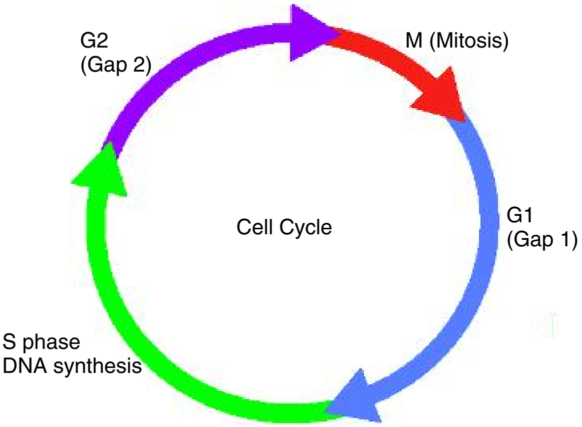


Figure 8.3 Cell cycle diagram showing the G_1, S, G_2 phases in a normal mitotic cell cycle

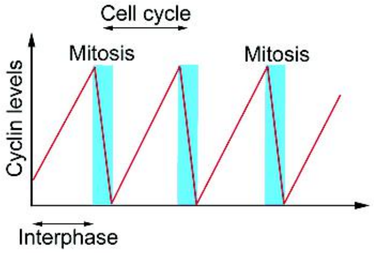


Figure 8.4 The levels of protein vary according to the stage of the cell cycle and gave rise to the name of this protein. Cyclins are abundant during the stage of the cell cycle in which they act and are then degraded

Cyclins have diverse sequences, but functionally important regions called cyclin boxes show high levels of homology. This homology has facilitated domain identification in other proteins such as transcription factors. Cyclin boxes confer binding specificity towards protein partners. They are found with other

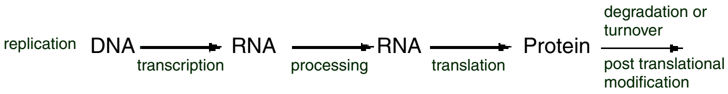


Figure 8.2 From DNA to protein

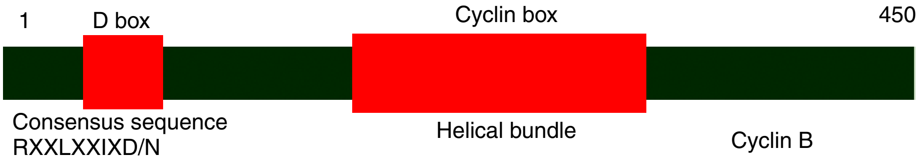


Figure 8.5 Organization of cyclins showing D box and cyclin box segments

characteristic motifs such as the cyclin destruction box with a minimal consensus sequence of Arg-Xaa-Xaa-Leu-Xaa-Xaa-Ile-Xaa-Asn/Asp (Figure 8.5).

The first 'box' structure was determined for a truncated construct of cyclin A representing the final 200 residues of a 450 residue protein. The structure has two domains each with a central core of five regular α helices that is preserved between cyclins despite low levels of sequence identity. Five charged, invariant, residues (Lys266, Glu295, Leu255, Asp240, and Arg21) form a binding site for protein partners such as cyclin-dependent kinases.

Cyclins regulate the activity of cyclin-dependent kinases (Cdks) by promoting phosphorylation of a single Thr side chain. Cdks were originally identified via genetic analysis of yeast cell cycles and purification of extracts stimulating the mitotic phase of frog and marine invertebrate oocytes. However, Cdks are found in all eukaryotic cell cycles where they control the major transitions of the cell cycle. Studies of the cell cycle were aided by the identification of mutant strains of fission yeast (*Schizosaccharomyces pombe*) and budding yeast (*Saccharomyces cerevisiae*) containing genetic defects (called lesions) within key controlling genes. In *S. cerevisiae* Leland Hartwell identified many genes regulating the cell cycle and coined the term 'cell division cycle' or *cdc* gene (Figure 8.6). One gene, *cdc28*, controlled the important first step of the cell cycle – the 'start' gene. Cells normally go through a number of 'checkpoints' to ensure cell division proceeded correctly; at each of these points mutants were identified in cell cycle genes. These studies allowed a picture of the cell cycle in terms of genes and cells would not advance further in the cycle until the 'block' was overcome by repairing defects. Cancerous cells turn out to be cells that manage to evade these 'checkpoints' and divide uncontrollably.

Using the fission yeast, *S. pombe*, Paul Nurse identified a gene (*cdc2*) whose product controlled many reactions in the cell cycle and was identified by homology to cyclin-dependent kinase 1 (Cdk1) found in human cells. The human gene could substitute for the yeast gene in Cdk$^-$ mutants. The genes *cdc2* and *cdc28* from *S. pombe* and *S. cerevisiae* encode homologous

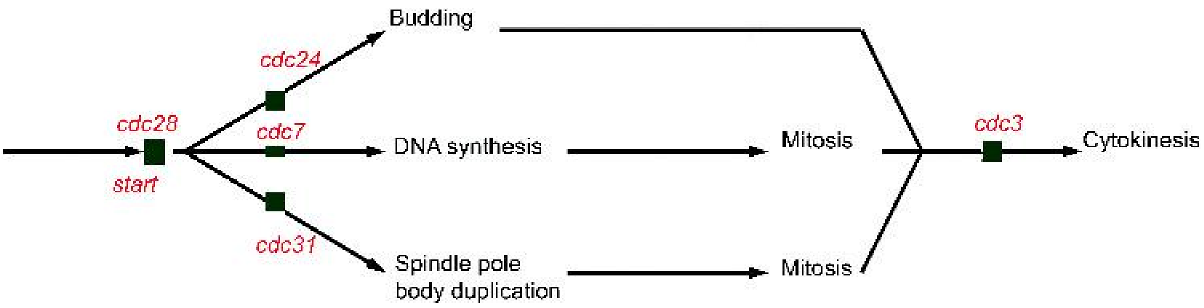


Figure 8.6 The role of mutants and the identification of key genes in the cell cycle of yeast. The genes *cdc28*, *cdc24*, *cdc31* and *cdc7* were subsequently shown to code for cdk2, a guanine nucleotide exchange factor, centrin – a Ca-binding protein, and another cyclin-dependent kinase

cyclin-dependent protein kinases occupying the same pivotal role within the cell cycle.

Cdks are identified in all cells and retain activity within a single polypeptide chain ($M_r \sim 35$ kDa). This subunit contains a catalytic core with sequence similarity to other protein kinases but requires cyclin binding for activation together with the phosphorylation of a specific threonine residue. Molecular details of this activity were uncovered by crystallographic studies of Cdks, cyclins and the complex formed between the two protein partners. Cdks catalyse phosphorylation (and hence activation) of many proteins at different stages in the cell cycle: including histones, as part of chromosomal DNA unpackaging; lamins, proteins forming the nuclear envelope that must disintegrate prior to mitosis; oncoproteins that are often transcription factors; and proteins involved in mitotic spindle formation. Cdks control reactions occurring within the cell by the combined action of a succession of phosphorylated Cdk–cyclin complexes that elicit activation and deactivation of enzymes. It is desirable that these systems represent an 'all or nothing' signal in which the cell is compelled to proceed in one, irreversible, direction. The reasons for this are fairly obvious in that once cell division has commenced it would be extremely hazardous to attempt to reverse the procedure. Cdks trigger cell cycle events but also enhance activity of the next cyclin–Cdk complex with a cascade of cyclin–Cdk complexes operating during the G_1/S/G_2 phases in yeast and human cells (Figure 8.7).

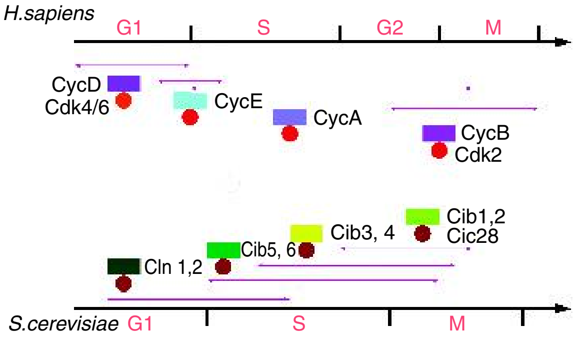


Figure 8.7 Major Cdk-cyclin complexes involved in cell cycle control in humans and budding yeast. Purple arrows indicate approximate timing of activation and duration of the Cdk–cyclin complexes. Cyclins are shown as coloured rectangles with the cognate Cdk shown by circles. The S and M phases overlap in *S. cerevisiae*

The structure of Cdk and its role in the cell cycle

Understanding the structure of Cdks uncovered the mechanism of action of these enzymes with the first structure obtained in 1993 for Cdk2 in a complex with ATP (Figure 8.8). A bilobal protein comprising a small lobe at the N terminal together with a much larger C-domain resembled other protein kinases, such as cAMP-dependent protein kinase. The comparatively small size reflected a 'minimal' enzyme. Within this small structure two regions of Cdk2 stood out as 'different' from other kinases. A single helical segment, called the PSTAIRE helix from the sequence of conserved residues Pro-Ser-Thr-Ala-Ile-Arg-Glu-, was located in the N-terminal lobe and was unique to Cdks. The second distinctive region contained the site of phosphorylation (Thr160) in Cdk and was comparable to other regulatory elements seen in protein kinases. This region was called the T or activation loop.

The smaller N-terminal lobe has the PSTAIRE helix and a five stranded β sheet whilst the larger lobe has six α helices and a small section of two-stranded β sheet. ATP binding occurs in the cleft between lobes with the adenine ring located between the β sheet of the small lobe and the L7 loop between strands 2 and 5. A glycine-rich region interacts with the adenine ring, whilst Lys33, Asp145, and amides in the backbone of the glycine-rich loop interact with phosphate groups (Figure 8.9).

The basic fold of Cdk2 contains the PSTAIRE helix and T loop positioned close to the catalytic cleft, but is an inactive form of the enzyme. Activation by a factor of ~10 000 occurs upon cyclin binding and is generally assessed via the ability of Cdk2 to phosphorylate protein substrates such as histone. Further enhancement of activity occurs with phosphorylation of Thr160 in Cdk2 and suggested that conformational changes might modulate biological activity. The origin and extent of these conformational changes were determined by comparing the structures of Cdk2–cyclin A complexes in phosphorylated (fully active) and unphosphorylated

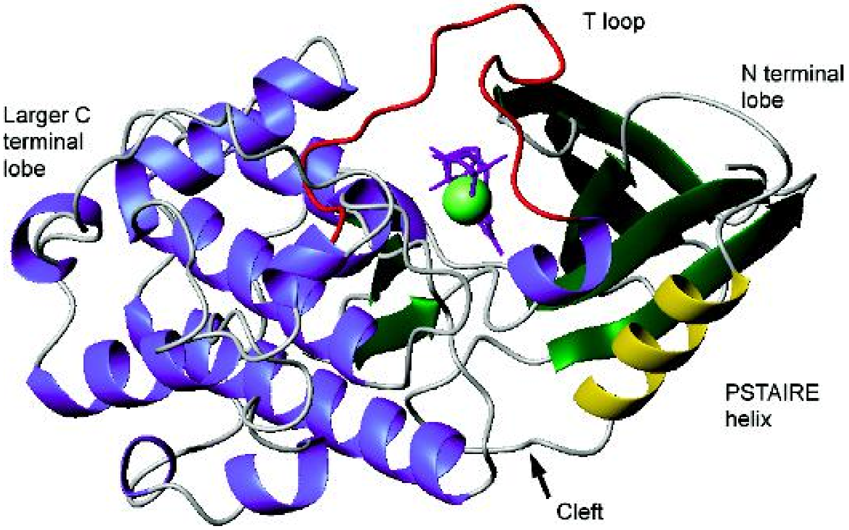


Figure 8.8 Structure of isolated Cdk2 and its interaction with Mg^{2+}-ATP. The ATP is shown in magenta, the Mg^{2+} ion in light green. The secondary structure elements are shown in blue and green with the PSTAIRE helix shown in yellow and the T loop in red. The L12 helix can be seen adjacent to the PSTAIRE helix in the active site of Cdk2

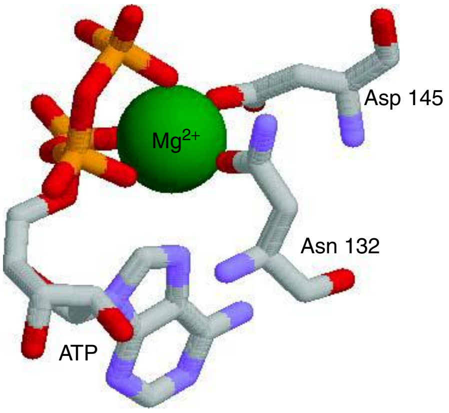


Figure 8.9 Binding of Mg^{2+}-ATP to monomeric Cdk2. Magnesium bound in an octahedral environment with six ligands provided by phosphates of ATP, Asp145, Asn132 and a bound water molecule

(partially active) states together with the isolated (inactive) Cdk2 subunit.

In the Cdk2–cyclin A complex conformational changes were absent in cyclin A with structural perturbations confined to Cdk2. The binding site between cyclin A and Cdk2 involved the cyclin A box containing Lys266, Glu295, Leu255, Asp240, and Arg211 and a binding site located close to the catalytic cleft. In proximity to the cleft hydrogen bonds with Glu8, Lys9, Asp36, Asp38, Glu40 and Glu42 were important to complex formation as well as hydrophobic interactions that underpinned movement of the ‘PSTAIRE’ helix and T loop. Large conformational changes upon cyclin binding occur for the PSTAIRE helix and the T loop with the helix rotated by ~90° and the loop region displaced. In the absence of cyclin the T loop is positioned above the catalytic cleft blocking the active site ATP to polypeptide substrates. Cyclin binding moves the T loop opening the catalytic cleft and exposing the phosphorylation site at Thr160. Figures 8.10 and 8.11 show the structure of the Cdk2–cyclin A–ATP complex.

Threonine phosphorylation in the T loop is performed by a second kinase called CAK (Cdk-activating kinase), and in an activated state the complex phosphorylates downstream proteins modifying serine or threonine residues located within target (S/T-P-X-K/R) sequences.

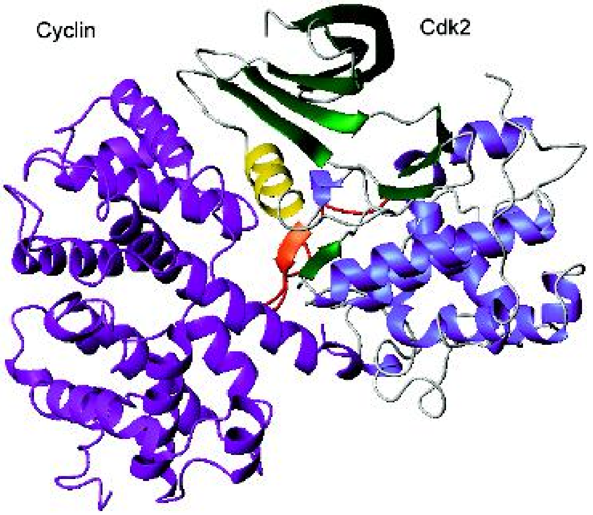


Figure 8.10 Structure of the Cdk2–cyclin A–ATP complex. The structure shows the binding of cyclin A (purple, left) to the Cdk2 protein (green, right). The colour scheme of Figure 8.8 is used with the exception that the β-9 strand of Cdk (formerly the L12 helix) is shown in orange

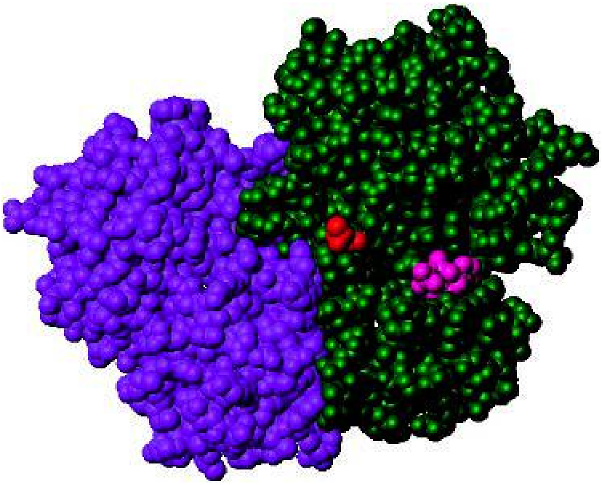


Figure 8.11 Space-filling representation of the cyclinA (purple) and Cdk2 (green) interaction. Phosphorylated Thr160 (red) of Cdk2 in the T loop together with ATP (magenta) are shown

Cdk–cyclin complex regulation

Cyclin binding offers one level of regulation of Cdk activity but other layers of regulatory control exist within the cell. Phosphorylation of Cdk at Thr160 by CAK is a major route of activation but further phosphorylation sites found on the surface of Cdk lead to enzyme inhibition. In Cdk1 the most important sites are Thr14 and Tyr15. Further complexity arises with the observation of specialized Cdk inhibitors that regulate most kinase activity in the form of two families of inhibitory proteins: the INK4 family bind to free Cdk preventing association with cyclins whilst the CIP family act on Cdk–cyclin complexes (Figure 8.12). The INK4 inhibitors are specific for the G_1 phase Cdks whereas CIP inhibitors show a broader preference. Regulation has become an important issue with the observation that certain tumour lines show altered patterns of Cdk regulation as a result of inhibitor mutation. One inhibitor of CDK4 and CDK6 is p16INK4a, and in about one third of all cancer cases this tumour suppressor is mutated. Another CDK inhibitor, p27Cip2, is present at low levels in cancers that show poor clinical prognosis whilst mutation of Cdk4 in melanoma cells leads to a loss of inhibition by Ink4 proteins.

Cdks are protein kinases transferring the γ-phosphate of ATP onto Ser/Thr side chains in target proteins. This type of phosphorylation is a common regulatory event

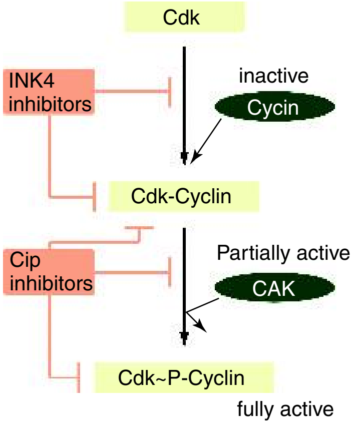


Figure 8.12 The interaction of cyclin–Cdk complexes with INK4 and CIP inhibitors

within cells and underpins signal transduction mediated by cytokines, kinases and growth factors as well as the events within the cell cycle.

In yeast the cell cycle is controlled by one essential kinase known as Cdk1 whilst multicellular eukaryotes have Cdk1 and Cdk2 operating in the M phase and S phase, respectively, along with two additional Cdks (Cdk4 and Cdk6) regulating G_1 entry and exit. Despite the presence of additional Cdks in higher eukaryotes the molecular mechanisms controlling cell cycle events show evolutionary conservation operating similarly in yeasts, insects, plants, and vertebrates, including humans. Research aimed at understanding the cell cycle underpins the development of cancer since the failure to control these processes accurately is a major molecular event during carcinogenesis. It is likely that discoveries in this area will have an enormous impact on molecular medicine. Knowledge of the structure of Cdks together with the structures of Cdk–cyclin–inhibitor complexes will be translated into the design of drugs that modulate Cdk activity and eliminate unwanted proliferative reactions.

DNA replication

A major function of the cell cycle is to integrate DNA synthesis within cell division. The double helix structure of DNA suggested conceptually simple methods of replication; a conservative model where a new double-stranded helix is synthesized whilst the original duplex is preserved, or a semi-conservative scheme in which the duplex unfolds and each strand forms half of a new double-stranded helix.

The semi-conservative' model (Figure 8.13) was demonstrated conclusively by Matthew Meselson and Franklin Stahl in 1958 by growing bacteria (*E. coli*) on media containing the heavy isotope of nitrogen (^{15}N) for several generations to show that the DNA was measurably heavier than that from cells grown on the normal ^{14}N isotope. Small differences in density between these two forms of DNA were demonstrated experimentally by density gradient centrifugation. Isolated DNA from bacteria grown on ^{15}N for many generations gives a single band corresponding to a density of 1.724 g ml^{-1} whilst DNA from bacteria grown on normal (^{14}N) nitrogenous sources yielded a lighter density single band of 1.710 g ml^{-1}. Comparing DNA isolated from bacteria grown on heavy (^{15}N) nitrogenous media and then transferred to ^{14}N-containing media for a single generation led to the observation of a single DNA species of intermediate density (1.717 g ml^{-1}). This was only consistent with the semi-conservative model.

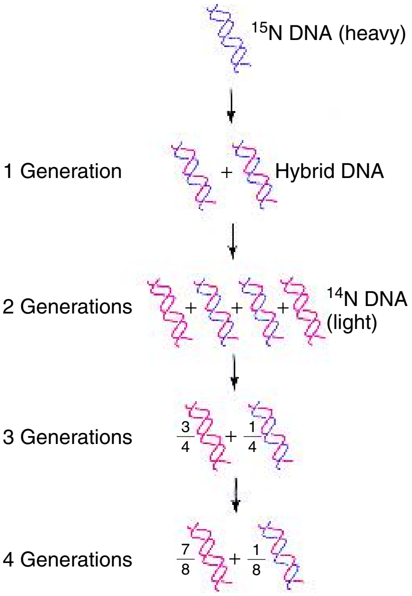


Figure 8.13 Semi-conservative model of DNA replication (reproduced with permission from Voet, D Voet, J.G & Pratt, C.W. *Fundamentals of Biochemistry*. John Wiley & Sons Inc, Chichester, 1999)

The semi-conservative mechanism suggested that each strand of DNA acted as a template for a new chain. A surprisingly large number of enzymes are involved in these reactions including:

1. *Helicases.* These proteins bind double-stranded DNA catalysing strand separation prior to synthesis of new daughter strands.
2. *Single-stranded binding proteins.* Tetrameric proteins bind DNA stabilizing single-stranded structure and enhancing rates of replication.
3. *Topoisomerases.* This class, sometimes called DNA gyrases, assist unwinding before new DNA synthesis.
4. *Polymerases.* DNA Polymerase I was the first enzyme discovered with polymerase activity. It

is not the primary enzyme involved in bacterial DNA replication, a reaction catalysed by DNA polymerase III. DNA polymerase I has exonuclease activities useful in correcting mistakes or repairing defective DNA. DNA polymerase can be isolated as a fragment exhibiting the 5′–3′ polymerase and the 3′–5′ exonuclease activity but lacking the 5′–3′ exonuclease of the parent molecule. This fragment, known as the Klenow fragment, was widely used in molecular biology prior to the discovery of thermostable polymerases.

5. *Primase*. DNA synthesis proceeds by the formation of short RNA primers that lead to sections of DNA known as Okazaki fragments. The initial priming reaction requires a free 3′ hydroxyl group to allow continued synthesis from these primers. Specific enzymes known as RNA primases catalyse this process.
6. *Ligase*. Nicks or breaks in DNA strands occur continuously during replication and these gaps are joined by the action of DNA ligases.

The synthesis of RNA primers in *E. coli* requires the concerted action of helicases (*DnaB*) and primases (*DnaG*) in the primosome complex (see Table 8.1; $M_r \sim 600\,000$) whilst a second complex, the replisome, contains two DNA polymerase III enzymes engaged in DNA synthesis in a 5′–3′ direction. Figure 8.14 summarizes DNA replication in *E. coli*.

Table 8.1 Primosome proteins

Protein	Organization	Subunit mass
PriA	Monomer	76
PriB	Dimer	11.5
PriC	Monomer	23
DnaT	Trimer	22
DnaB	Hexamer	50
DnaC	Monomer	29
DnaG (primase)	Monomer	60

The complex lacking DnaG is called a pre-primosome. Derived from Kornberg, A. & Baker, T.A. *DNA Replication*, 2nd edn. Freeman, New York, 1992.

Transcription

The first step in the flow of information from DNA to protein is transcription. Transcription forms complementary messenger RNA (mRNA) from DNA through catalysis by RNA polymerase. RNA polymerase copies the DNA coding strand adding the base uracil (U) in place of thymine (T) into mRNA. Most studies of transcription focused on prokaryotes where the absence of a nucleus allows RNA and protein synthesis to occur rapidly, in quick succession and catalysed by a single RNA polymerase. In eukaryotes multiple RNA polymerases exist and transcription is more complex involving larger numbers of accessory proteins.

Structure of RNA polymerase

Most prokaryotic RNA polymerases have five subunits (α, β, β′, σ and ω) with two copies of the α subunit found together with single copies of the remaining subunits (Table 8.2). From reconstitution studies the ω subunit is not required for holoenzyme assembly or function. Similarly the σ subunit is easily dissociated to leave a ‘core’ enzyme retaining catalytic activity. A decrease in DNA binding identified a role for the σ subunit in promoter recognition where it reduces non-specific binding by a factor of $\sim 10^4$.

Table 8.2 Subunit composition of RNA polymerase of *E. coli*; the σ subunit is one member of a group of alternative subunits that have interchangeable roles

Subunit	M_r	Stoichiometry	Function
α	36 500	2	Chain initiation. Interaction with transcription factors and upstream promoter elements
β	151 000	1	Chain initiation and elongation
β′	155 000	1	DNA binding
σ	70 000	1	Promoter recognition
ω	11 000	1	Unknown

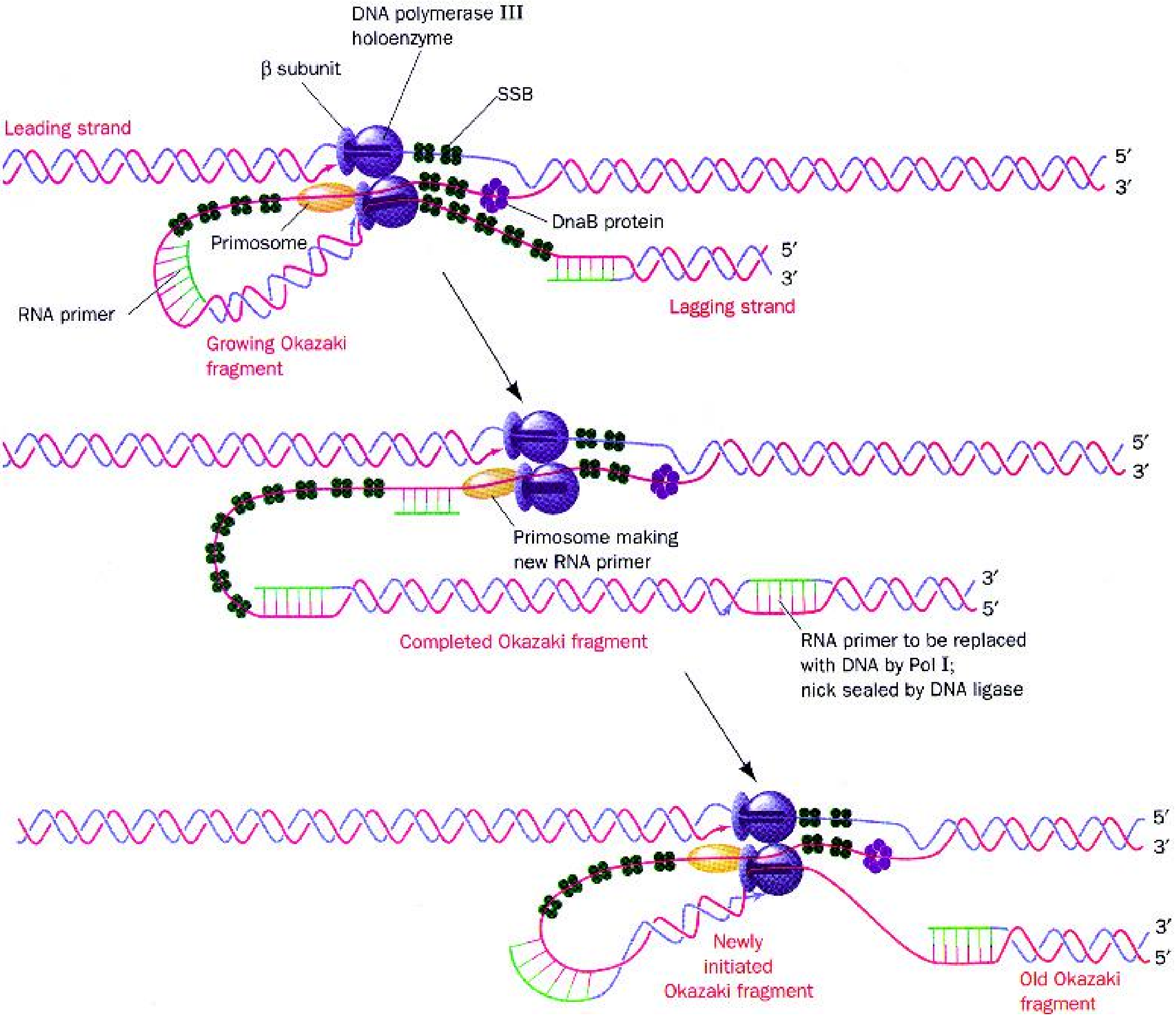


Figure 8.14 The integrated and complex nature of DNA replication in *E. coli*. The replisome contains two DNA polymerase III molecules and synthesizes leading and lagging strands. The lagging strand template must loop around to permit holoenzyme catalysed extension of primed lagging strands. DNA polymerase III releases the lagging strand template when it encounters previously synthesized Okazaki fragments. This signals the initiation of synthesis of a lagging strand RNA fragment. DNA polymerase rebinds the lagging strand extending the primer forming new Okazaki fragments. In this model leading strand synthesis is always ahead of the lagging strand synthesis (reproduced with permission from Voet, D., Voet, J.G & Pratt, C.W. *Fundamentals of Biochemistry*. John Wiley & Sons Inc, Chichester, 1999)

One of the best-characterized polymerases is from bacteriophage T7 (Figure 8.15) and departs from this organization by containing a single polypeptide chain (883 residues) arranged as several domains. It is organized around a cleft that is sufficient to accommodate the template of double-stranded DNA and has been likened to the open right hand divided into palm, thumb and finger regions. The palm, fingers and thumb regions define the substrate (DNA) binding and catalytic sites.

The thumb domain is a helical extension from the palm on one side of the binding cleft and stabilizes the ternary complex formed during transcription by wrapping around template DNA. Mutant T7 RNA polymerases with 'shortened thumbs' have less activity as a result of lower template affinity. The palm domain consisting of residues 412–565 and 785–883 is located at the base of the deep cleft, an integral part of T7 RNA polymerase, and is surrounded by the fingers and the thumb. It contains three β strands, a structurally conserved motif in all polymerases with conserved Asp residues (537 and 812) participating directly in catalysis. The finger domain extends from residues 566–784 and contains a specificity loop involved in direct interactions with specific bases located in the major groove of the promoter DNA sequence. This region also has residues that form part of the active site such as Tyr639 and Lys631. Finally the N-terminal domain forms the front wall of the catalytic cleft and leads to the enzyme's characteristic concave appearance. This domain interacts with upstream regions of promoter DNA and the nascent RNA transcript.

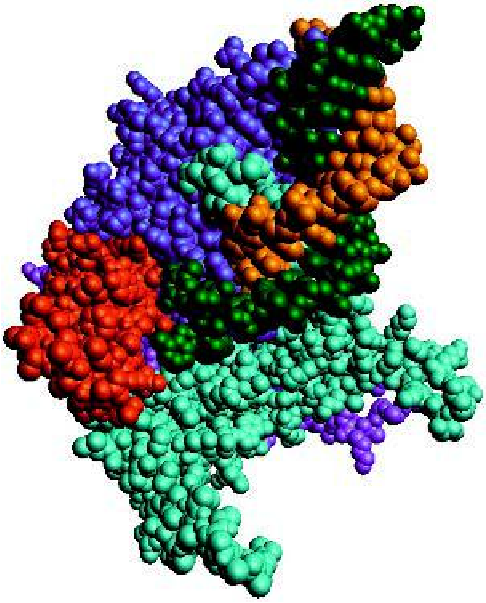


Figure 8.15 The structure of T7 RNA polymerase. A space-filling representation with synthetic DNA (shown in orange and green) has a thumb region represented by residues 325–411 (red), a palm region underneath the DNA and largely obscured by other domains (purple) whilst fingers surround the DNA and are shown in cyan to the right. The N-terminal domain is shown in dark blue (PDB: 1CEZ)

Consensus sequences exist for DNA promoter regions (Figure 8.16) and David Pribnow identified a region rich in the bases A and T approximately −10 bp upstream of the transcriptional start site. A second region approximately 35 bp upstream of the transcriptional start site was subsequently identified with the space between these two sites having an optimal size of 16 or 17 bases in length. Although consensus sequences for the −35 and −10 regions have been identified no natural promoter possesses exactly these bases although all show strong sequence conservation.

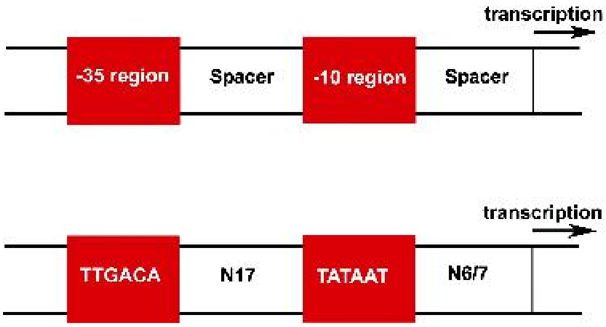


Figure 8.16 The organization of the promoter region of genes in *E. coli* for recognition by RNA polymerase. The consensus sequence TATAAT at a position ∼10 bp upstream of the transcription start site is the Pribnow box

Eukaryotic RNA polymerases

In eukaryotes polymerase binding occurs at the TATA box and is more complex. Three nuclear RNA polymerases designated as RNA polymerase I, II and III, together with polymerases located in the mitochondria and chloroplasts are known. The nuclear polymerases are multi-subunit proteins whose activities frequently require the presence of accessory proteins called transcription factors (Table 8.3).

In eukaryotes RNA polymerases I and III transcribe genes coding for ribosomal (rRNA) and transfer RNA (tRNA) precursors. RNA polymerase I produces a large transcript – the 45S pre-rRNA – consisting of tandemly arranged genes that after synthesis in the nucleolus are processed to give 18S, 5.8S and 28S rRNA molecules (Figure 8.17). The 18S rRNA contributes to the structure of the small ribosomal subunit with the 28S and 5.8S rRNAs associating with proteins to form the 60S subunit. RNA polymerase III

Table 8.3 The role of eukaryotic RNA polymerases and their location

RNA Polymerase	Location	Role
I	Nuclear (nucleolus)	Pre rRNA except 5S rRNA
II	Nuclear	Pre mRNA and some small nuclear RNAs
III	Nuclear	Pre tRNA, 5S RNA and other small RNAs
Mitochondrial	Organelle (matrix)	Mitochondrial RNA
Chloroplast	Organelle (stroma)	Chloroplast RNA

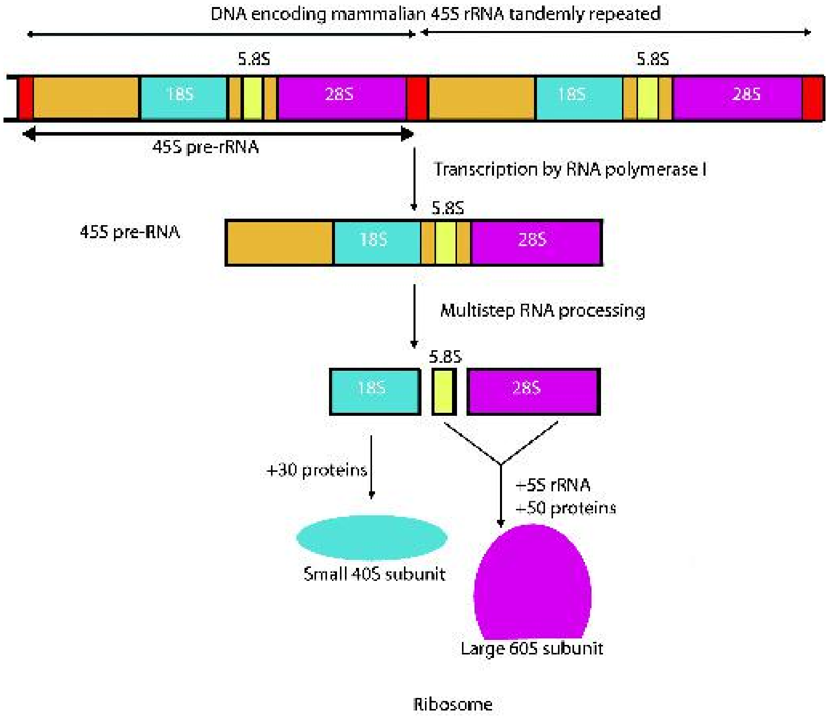


Figure 8.17 The genes for rRNAs exist as tandemly repeated copies separated by non-transcribed regions. The 45S transcript is processed to remove the regions shown in tan to give the three rRNAs

Table 8.4 Basal transcription factors required by eukaryotic RNA polymerase II

Factor	Subunits	Mass (kDa)	Function
TFIID - TBP	1	38	TATA box recognition and TFIIB recruitment.
TAFs	12	15–250	Core promoter recognition
TFIIA	3	12,19,35	Stabilization of TBP binding
TFIIB	1	35	Start site selection.
TFIIE	2	34,57	Recruitment and modulation of TFIIH activity
TFIIF	2	30,74	Promotor targeting of polymerase and destabilization of non specific RNA polymerase-DNA interactions
TFIIH	9	35–90	Promoter melting via helicase activity

Adapted from Roeder, R.G. *Trends Biochem. Sci.* 1996, **21**, 327–335. Subunit composition and mass are those of human cells but homologues are known for rat, *Drosophila* and yeast. TBP, TATA binding protein; TAFs, TATA binding associated factors.

synthesizes the precursor of the 5S rRNA, the tRNAs as well as a variety of other small nuclear and cytosolic RNAs.

RNA polymerase II transcribes DNA into mRNA and is the enzyme responsible for structural gene transcription. However, effective catalysis requires additional proteins with at least six basal factors involved in the formation of a ‘pre-initiation’ complex with RNA polymerase II (Table 8.4). These transcription factors are called TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH. The Roman numeral ‘II’ indicates their involvement in reactions catalysed by RNA polymerase II.

The assembly of the pre-initiation complex begins with TFIID (TBP) identifying the TATA consensus sequence (T-A-T-A-A/T-A-A/T) and is followed by the coordinated accretion of TFIIB, the non-phosphorylated form of RNA polymerase II, TFIIF, TFIIE, and TFIIH (Figure 8.18). Before elongation commences RNA polymerase is phosphorylated and remains in this state until termination when a phosphatase recycles the polymerase back to its initial form. The enzyme can then rebind TATA sequences allowing further transcriptional initiation.

Basal levels of transcription are achieved with TBP, TFIIB, TFIIF, TFIIE, TFIIH, RNA polymerase and the core promoter sequence, and this system has been used to demonstrate minimal requirements for initiation and complex assembly (Figure 8.19). A cycle of efficient re-initiation of transcription is achieved when RNA polymerase II re-enters the pre-initiation complex before TFIID dissociates from the core promoter. Non-basal rates of transcription require proximal and distal enhancer regions of the promoter and proteins that regulate polymerase efficiency (TAFs).

TBP (and TFIID) binding to the TATA box is a slow step but yields a stable protein–DNA complex that has been structurally characterized from several systems. The plant, yeast, and human TBP in complexes with TATA element DNA share similar structures (Figure 8.20) suggesting conserved mechanisms of molecular recognition during transcription. The three-dimensional structure of the conserved portion of TBP is strikingly similar to a saddle; an observation that correlates precisely with function where protein ‘sits’ on DNA creating a stable binding platform for other transcription factors. DNA binding occurs on the concave underside of the saddle with the upper surface (seat of the saddle) binding transcriptional components.

DNA binding is mediated by the curved, antiparallel β sheet whose distinctive topology provides a large concave surface for interaction with the minor DNA

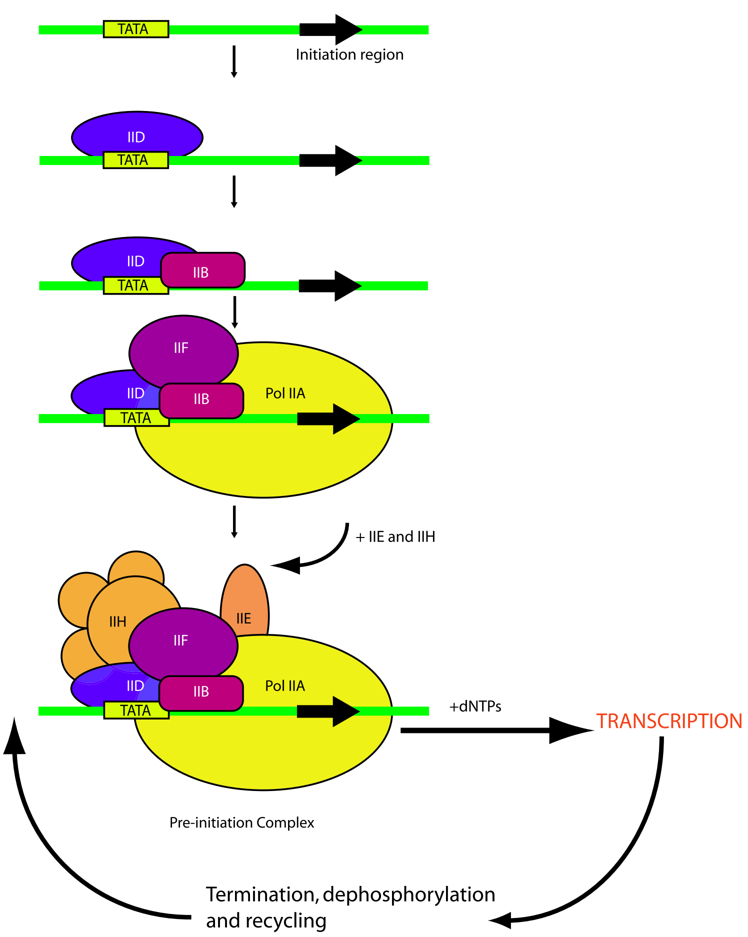


Figure 8.18 The formation of pre-initiation complexes between RNA polymerase II plus TATA sequence DNA

groove of the 8-bp TATA element. The 5′ end of DNA enters the underside of the molecular saddle where TBP causes an abrupt transition from the normal B form configuration to a partially unwound form by the insertion of two Phe residues between the first T:A base step. The widened minor groove adopts a conformation comparable to the underside of the molecular saddle allowing direct interactions between protein side chains and the minor groove edges of the central 6 bp. A second distortion induced by insertion of another two Phe residues into a region between the last two base pairs of the TATA element mediates an equally abrupt return back to the normal B-form DNA.

TFIIB is the next protein to enter the pre-initiation complex and its interaction with TBP was sufficiently strong to allow the structure of TFIIB-TBP-TATA complexes to be determined by Ronald Roeder and Stephen Burley in 1995 (Figure 8.21). A core region based around the C-terminal part of the molecule (cTFIIB) shows homology to cyclin A and binds to an acidic C-terminal ‘stirrup’ of TBP using a basic cleft on its own surface. The protein also interacts with the DNA sugar-phosphate backbone upstream and downstream of the TATA element. The first domain of cTFIIB forms the downstream surface of the cTFIIB–TBP–DNA ternary complex and in conjunction with the N-terminal domain of TFIIB this

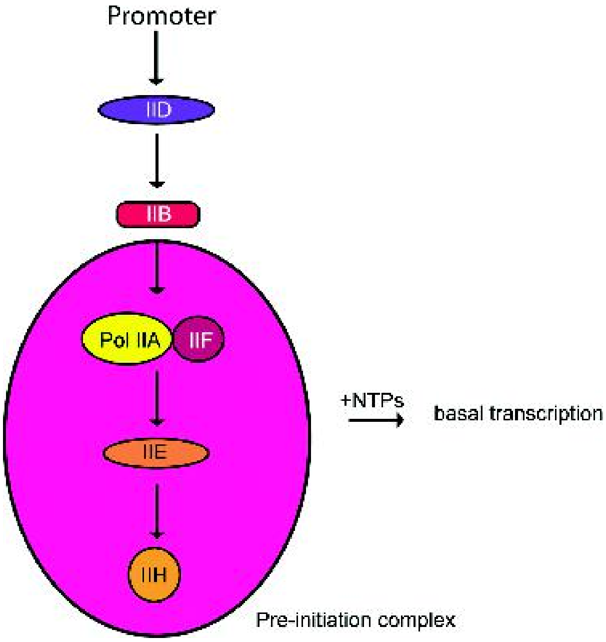


Figure 8.19 Representation of functional interactions modulating basal transcription. The basal factors TBP, TFIIB, TFIIF, TFIIE, and TFIIH together with RNA polymerase II are shown

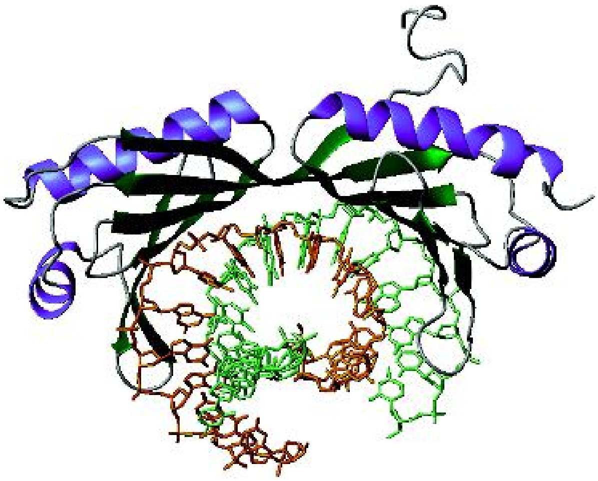


Figure 8.20 The co-complex of TATA element DNA plus TBP showing the remarkable molecular saddle shape. The DNA distorted helix lies on the concave surface of TBP

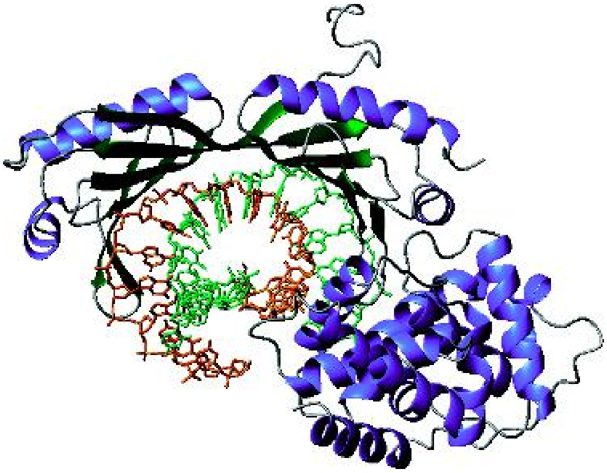


Figure 8.21 The TBP–TFIIB–TATA element complex. TFIIB, shown lower right, is an α helical protein

complex could act between TBP and RNA polymerase II to fix the transcription start site. The rest of the complex reveals solvent accessible surfaces that allow further binding of other transcription factors. The TBP–TATA complex is unaltered in conformation as a result of TFIIB binding and it is likely that a major role of TFIIB is stabilization of the initial protein–DNA complex. After formation of the TFIIB–TFIID–DNA complex three other general transcription factors and RNA polymerase II complete the assembly of the pre-initiation complex.

TFIIA is a co-activator that supports regulation of RNA polymerase II directed transcription. At an early stage in the assembly of the pre-initiation complex TFIIA associates with the TFIID–DNA or the TFIIB–TFIID–DNA complexes recognizing an N-terminal stirrup of TBP and the phosphoribose backbone upstream of the TATA element. This occurs on the opposite face of the double helix to cTFIIB binding and TFIIA preferentially binds to the preformed TBP–DNA complex enhancing its stability.

By comparing structures of the complexes formed by cTFIIB–TBP–DNA with those of TFIIA–TBP–DNA it is possible to create a model of the

TFIIA–TFIIB–TBP–DNA quaternary complex where the mechanism of synergistic action between TFIIB and TFIIA is rationalized by the observation that the two proteins do not interact directly but instead use their basic surfaces to make contact with the phosphoribose backbone on opposite surfaces of the double helix upstream of the TATA element.

Eukaryotic RNA polymerases interact with a number of transcription factors before binding to the promoter and transcription commences. Thus although the fundamental principles of transcription are similar in prokaryotes and eukaryotes the latter process is characterized by the presence of multiple RNA polymerases and additional control processes modulated by a variety of transcription factors.

Structurally characterized transcription factors

The prokaryotic transcription factors cro and λ are examples of proteins containing helix-turn-helix (HTH) motifs that function as DNA binding proteins. The basis for highly specific interactions relies on precise DNA–protein interactions that involve side chains found in the recognition helix of HTH motifs and the bases and sugar-phosphate regions of DNA. A substantial departure in the use of the HTH motif to bind DNA is seen in the *met* repressor of *E. coli.* A homodimer binds to the major groove via a pair of symmetrically related β strands, one from each monomer, that form an antiparallel sheet composed of just two strands. Other repressor proteins such as the arc repressor from a Salmonella bacteriophage (P22) were subsequently shown to use this DNA recognition motif.

Eukaryotic transcription factors: variation on a 'basic' theme

Eukaryotic cells require a greater range of transcription factors, and because genes are often selectively expressed in a tissue-specific manner these processes require an efficient regulatory mechanism. This is achieved in part through a diverse class of proteins. Included in this large and structurally diverse group of proteins are Zn finger DNA binding motifs, leucine zippers, helix loop helix (HLH) motifs (reminiscent of the HTH domains of prokaryotes) and proteins containing a β scaffold for DNA binding. Eukaryotic transcription factors contain distinct domains each with a specific function; an activation domain binds RNA polymerase or other transcription factors whilst a DNA binding domain recognizes specific bases near the transcription start site. There will also be nuclear localization domains that direct the protein towards the nucleus after synthesis in the cytosol. The following sections describe briefly the structural properties of some of the important DNA binding motifs found in eukaryotes.

Zn finger DNA binding motifs

The first Zn finger domain characterized was transcription factor IIIA (TFIIIA) from *Xenopus laevis* oocytes (Figure 8.22). Complexed with 5S RNA was a protein of M_r 40 000 required for transcription *in vitro* that yielded a series of similar sized fragments ($M_r \sim 3000$) containing repetitive primary sequence after limited proteolysis. Sequence analysis confirmed nine repeat units of ~25 residues each containing two cysteine and two histidine residues. Each domain contained a single Zn^{2+} ion.

The proteolytic fragments were folded in the presence of Zn which was ligated in a tetrahedral geometry to the side chains of two invariant Cys and two invariant His residues. In subsequently characterized domains it was found that the Cys_2His_2 ligation pattern varied with sometimes four Cys residues used to bind a single Zn^{2+} ion (Cys_4) or 6 Cys residues used to bind two Zn^{2+} ions. The Zn-binding domains formed small, compact, autonomously folding structures lacking extensive hydrophobic cores, and were christened Zn fingers.

The structures of simple Zn finger domains (Figure 8.23) reveal two Cys ligands located in a short strand or turn region followed by a regular α helix containing two His ligands. From the large number of repeating Zn finger domains and their intrinsic structures it is clear that the ends of the polypeptide chain are widely separated and that complexes with DNA might involve multiple Zn finger domains wrapping around the double helix. This arrangement was confirmed

```
                  *        20         *        40         *        60         *
TFIIIA : MAAKVASTSSEEAEGSLVTEGEMGEKALPVVYKRYICSFADCGAAYNKNWKLQAHLCKHTGEKPFPCKEEGCEKG

            80         *       100         *       120         *       140         *
TFIIIA : FTSLHHLTRHSLTHTGEKNFTCDSDGCDLRFTTKANMKKHFNRFHNIKICVYVCHFENCGKAFKKHNQLKVHQFS

                160         *       180         *       200         *       220
TFIIIA : HTQQLPYECPHEGCDKRFSLPSRLKRHEKVHAGYPCKKDDSCSFVGKTWTLYLKHVAECHQDLAVCDVCNRKFRH

             *       240         *       260         *       280         *       300
TFIIIA : KDYLRDHQKTHEKERTVYLCPRDGCDRSYTTAFNLRSHIQSFHEEQRPFVCEHAGCGKCFAMKKSLERHSVVHDP

                  *       320         *       340         *       360
TFIIIA : EKRKLKEKCPRPKRSLASRLTGYIPPKSKEKNASVSGTEKTDSLVKNKPSGTETNGSLVLDKLTIQ
```

Figure 8.22 The primary sequence of transcription factor TFIIIA from *Xenopus laevi* oocytes. The regions highlighted in yellow are units of 25 residues containing invariant Cys and His residues. Zn binding domains have a characteristic motif $CX_{(2-3)}CX_{(12)}HX_{(3-4)}H$

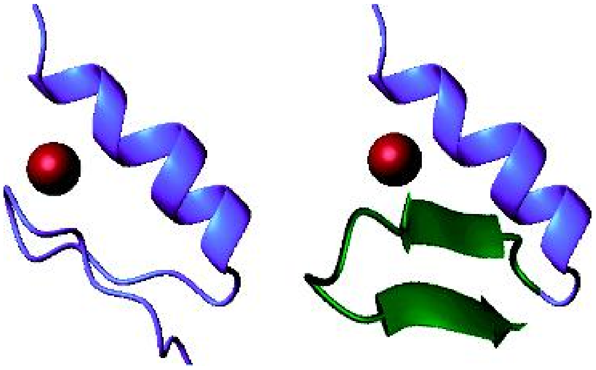


Figure 8.23 A Zn finger domain with helix and either an antiparallel β strand or a loop region

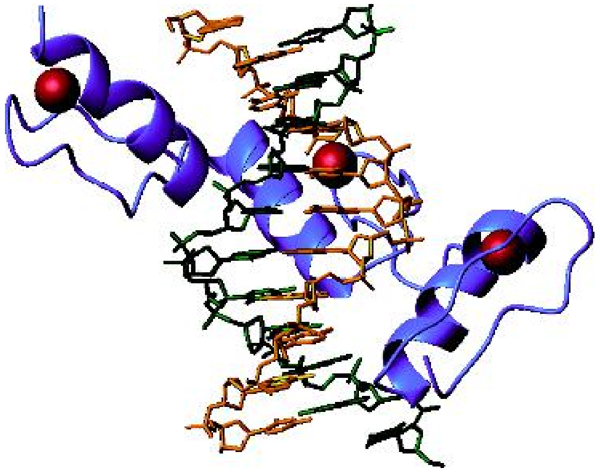


Figure 8.24 The structure of the Zn finger domains of mouse transcription factor Zif268 in a complex with DNA (PDB:1AIH). Three base pairs form hydrogen bonds with the N terminal region of the helix in one strand of the double helix. The periodicity of positively charged side chains allows interactions with phosphate groups

by crystallographic studies of the mouse transcription factor Zif268 where three repeated Zn finger domains complex with DNA by wrapping around the double helix with the β strands of each finger positioning the α helix in the major groove.

Zn finger domains are widely distributed throughout all eukaryotic cells and in the human genome there may be over 500 genes encoding Zn finger domains. Zif268, shown in Figure 8.24, has three such repeats while the homologous *Drosophila* developmental control proteins known as Hunchback and Kruppel have four and five Zn finger domains, respectively. TFIIIA has a comparatively large number of domains (nine) although increasingly proteins are being found with much greater numbers of ‘fingers’. Binuclear Zn clusters exist as Cys_6 Zn fingers. In GAL4, a yeast transcriptional activator of galactose (Figure 8.25), an N terminal

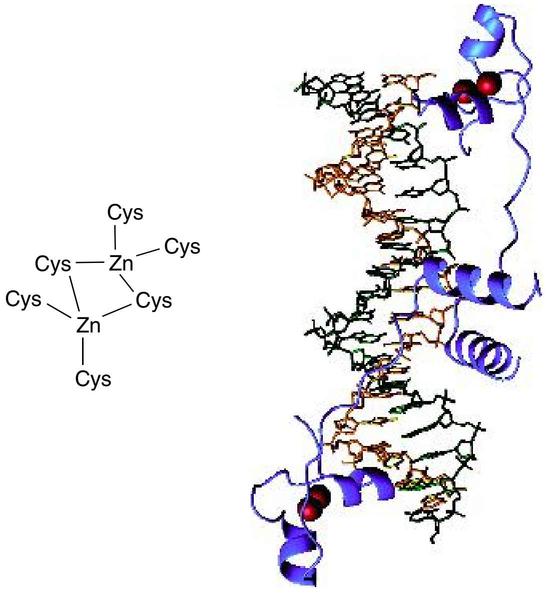


Figure 8.25 The GAL4 DNA binding domain. Tetrahedral geometry formed by six cysteine side chains binds two zinc ions

region containing six Cys residues coordinates two Zn ions. Each Zn cation binds to four Cys side chains in a tetrahedral environment with the central two cysteines ligating both Zn ions.

Leucine zippers

Many transcription factors contain sequences with leucine residues occurring every seventh position. This should strike a note of familiarity because coiled coils such as keratin have similar arrangements called heptad repeats. The heptad repeat is a unit of seven residues represented as $(a\text{-}b\text{-}c\text{-}d\text{-}e\text{-}f\text{-}g)_n$ in which residues *a* and *d* are frequently hydrophobic. The sequences promote the formation of 'coiled coils' by the interaction of residues '*a*' and '*d*' in neighbouring helices. The observation that many DNA binding proteins contained heptad repeats suggested that these regions assist in dimerization – a view supported by the structure of the 'leucine zipper' region of the yeast transcription factor GCN4 (Figure 8.26). The hydrophobic regions do not bind DNA but instead promote the association of subunits containing DNA binding motifs into suitable dimeric structures.

Close inspection of the interaction between helices reveals that the term zipper is inappropriate. The side chains do not interdigitate but show a simpler 'rungs along a ladder' arrangement with stability enhanced by hydrophobic interactions between residues at positions '*a*' and '*d*'. Although misleading in terms of organizational structure the name 'leucine zipper' remains widely used in the literature. Dimerization of the leucine zipper domains allows the N-terminal regions of GCN4, rich in basic residues, to lie in the major groove of DNA. The combination of basic and zipper regions yields a class of proteins with bZIP domains.

Helix-loop-helix motifs

The HLH motif is similar in both name and function to the HTH motif seen in prokaryotic repressors such as cro. The HLH motif is frequently found in transcription factors along with leucine zipper and basic (DNA binding) motifs (see Figure 8.27). As the name implies the structure consists of two α helices often arranged as segments of different lengths – one short and the other long – linked by a flexible loop of between 12 and 28 residues that allows the helices to fold back and pack against each other. The effect of folding is that each helix lies in parallel plane to the other and is amphipathic, presenting a hydrophobic surface of residues on one side with a series of charged residues on the other.

Eukaryotic HTH motifs

The fruit fly *Drosophila melanogaster* is a popular experimental tool for studying gene expression and regulation via changes in developmental patterns. Developmental genes control the pattern of structural gene expression and many *Drosophila* genes share common sequences known as homeodomains or homeoboxes (Figure 8.28). These sequences occur in other animal genomes with an increase in the number of *Hox* genes from one cluster in nematodes, two clusters in *Drosophila*, with the human genome having 39 homeotic genes organized into four clusters. The *Drosophila engrailed* gene encodes a polypeptide that binds to DNA sequences upstream of the transcription start site and imposes its pattern of regulation on other genes as a transcription factor. Crystallization of a 61-residue homeodomain in a complex with DNA showed a HTH motif comparable to the HTH domains of prokaryote repressors such as λ.

```
                      *        20         *        40         *        60
Yeast_GCN4 : KPNSVVKKSHHVGKDDESRLDHLGVVAYNRKQRSIPLSPIVPESSDPAALKRARNTEAAR : 60

                      *        80         *       100
Yeast_GCN4 : RSRARKLQRMKQLEDKVEELLSKNYHLENEVARLKKLVGER : 101
```

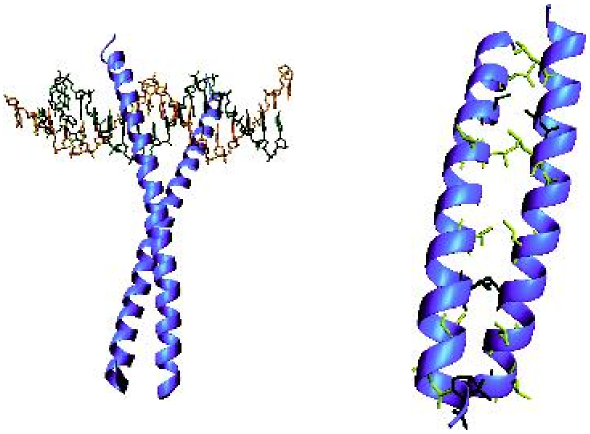


Figure 8.26 A leucine zipper motif such as that found in GCN4 shows a characteristic sequence known as the heptad repeat where every seventh side chain is hydrophobic and is frequently a leucine residue. Interaction between two helices forms a coiled coil and facilitates dimerization of the proteins. The structure of the leucine zipper region of GCN4 complexed with DNA and a detailed view of the interaction between helices (derived from PDB:1YSA and 2TZA)

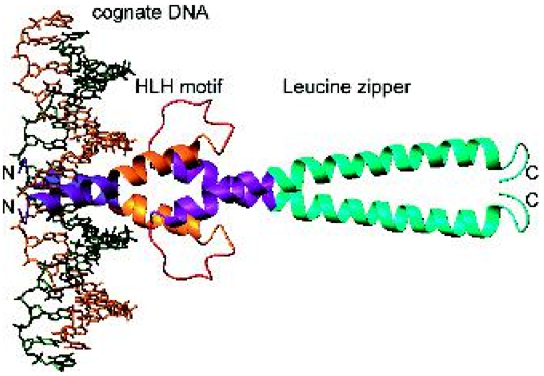


Figure 8.27 The structure determined for the Max–DNA complex. A basic domain (blue) at the N terminal interacts with DNA. The first helix of the HLH motif (orange) is followed by the loop region (red) and second helix (purple). The second helix flows into the leucine zipper (cyan) or dimerization region. This class of transcription factor are often called bHLHZ protein reflecting the three different types of domains namely basic, HLH and zipper (PDB: 1AN2)

Homeodomains are variations of HTH motifs containing three α helices, where the last two constitute a DNA binding domain. Two distinct regions make contact with TAAT sequences in complexes with cognate DNA. An N-terminal arm fits into the minor groove of the double helix with the side chains of Arg3 and Arg5 making contacts near the 5′ end of this ‘core consensus’ binding site. The second contact site involves an α helix that fits into the major groove with the side chains of Ile47 and Asn51 interacting with the base pairs near the 3′ end of the TAAT site. The ‘recognition helix’ is part of a structurally conserved HTH unit, but when compared with the structure of the λ repressor the helices are longer and the relationship between the HTH unit and the DNA is significantly different.

It has been estimated that mammals contain over 200 different cell types. These cell lines arise from highly regulated and specific mechanisms of gene expression involving the concerted action of many transcription factors. The human genome project has identified thousands of transcription factors and many carry exotic family names such as Fos and Jun, nuclear factor-κB,

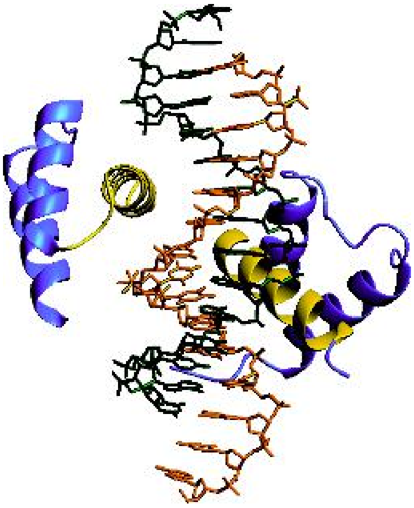


Figure 8.28 A typical homeobox domain showing the HTH motif in contact with its cognate DNA. The structure shown is that of a fragment of the *ubx* or *ultrabithorax* gene product showing the HTH motif. The recognition helix is shown in yellow lying in the major groove (PDB: 1B8I). Similar structures exist in homeodomain proteins such as *antennapedia* (*antp*) or *engrailed* (*eng*). All play a role in the developmental programme of *Drosophila* and each of these proteins acts an a transcriptional activator, binding to upstream regulatory sites and interacting with the RNA polymerase II/TFII complex

Pax and Hox. Growing importance is attached to the study of transcription factors with the recognition that mutations occur in many cancer cells. Wilm's tumour is one of the most common kidney tumours and is prevalent in children as a result of mutation in the WT-1 gene, a gene coding for a transcription factor containing Zn finger motifs. p53 is another transcription factor and is found in a mutated state in over 50 percent of all human cancers. These examples emphasize the link between transcription factors and human disease, where structural characterization has elucidated the basis of DNA–protein interaction.

The spliceosome and its role in transcription

Eukaryotic DNA contains introns or non-coding regions whose function remains enigmatic despite widespread occurrence within chromosomal DNA. Initial mRNA transcripts are processed to remove introns in a series of reactions that occur on the spliceosome. Spliceosomes are ribonucleoprotein complexes with sedimentation coefficients between 50 and 60S and a size and complexity comparable to the ribosome.

Before splicing many eukaryotic RNAs are modified at the 5′ and 3′ ends by capping with GTP and adding polyA tails. GTP is added in a reversed orientation compared to the rest of the polynucleotide chain and is often accompanied by methylation at the N_7 position of the G base. Together with the first two nucleotides the added G base forms a 5′-cap structure (Figure 8.29). Further modifications remove nucleotides from the 3′ end adding 'tails' of 50 to 200 adenine nucleotides. The mechanistic reasons for modification are unclear but involve mRNA stability, mRNA export from the nucleus to the cytoplasm and initiation of translation.

The spliceosome associates with the primary RNA transcript in the nucleus and consists of four RNA–protein complexes called the U1, U2, U4/U6 and U5 or small nuclear ribonucleoproteins (snRNPs, read as 'snurps'). The snRNPs are named after their RNA components so that U1 snRNP contains U1 small nuclear RNA. The U signifies uridine-rich RNAs, a feature found in all small nuclear RNA molecules. The U4 and U6 snRNAs are found extensively base paired and are therefore referred to collectively as the U4/U6 snRNP. The snRNPs pre-assemble in an ordered pathway to form a complex that processes mRNA transcripts.

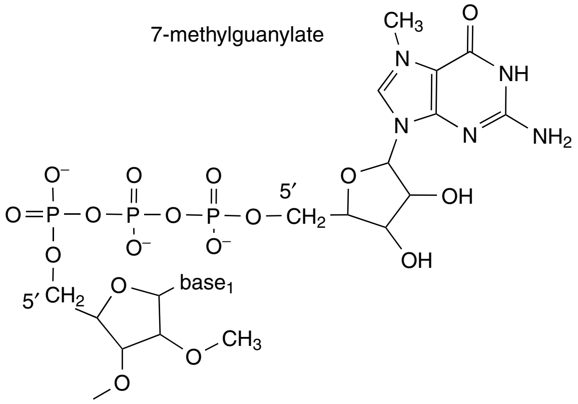


Figure 8.29 Methylation and G-capping of mRNA was discovered with the purification of mRNA

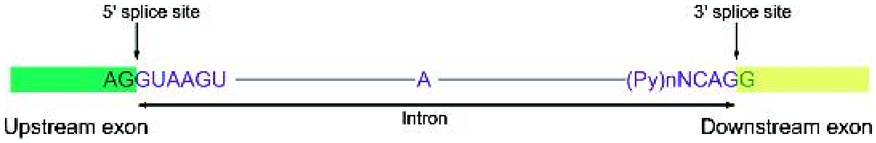


Figure 8.30 Splice sites are specific mRNA sequences and assist intron removal

The first requirement for effective mRNA splicing is to distinguish exons from introns. This is achieved by specific base sequences that signal exon–intron boundaries to the spliceosomes and define precisely the 5′ and 3′ regions to be removed from mRNA. Many introns begin with −5′ GU... and end with −3′ AG and in vertebrates the consensus sequence at the 5′ splice site of introns is AGGUAAGU (Figure 8.30). This sequence base pairs with complementary spliceosomal RNA aligning pre-mRNA and assisting removal of the intron. A second important signal called the branch point is found in the intron sequence as a block of pyrimidine bases together with a special adenine base approximately 50 bases upstream from the 3′ splice site.

Using this sequence information the spliceosome performs complex catalytic reactions to excise the introns in two major reactions. The first step involves the 2′ OH at the branch site and a nucleophilic attack on the phosphate group at the 5′ end of the intron to be removed. For this reaction to occur the mRNA is rearranged to bring two distant sites into close proximity. This movement breaks the pre mRNA at the 5′ end of the intron to leave, at this stage, a branch site with three phosphodiester bonds. The reaction creates a free OH group at the 3′ end with exon 1 no longer attached to the intron although it is retained by the spliceosome. The free OH group at the 3′ end of exon 1 attacks the phosphodiester bond between exon 2 and the intron, with the result that both exons are joined together in the correct reading frame (Figure 8.31). The resulting lasso-like entity is called a lariat structure and is removed from the mRNA (Figure 8.31). The processed mRNA is then exported from the nucleus and into the cytoplasm for translation.

Translation

Conversion of genetic information into protein sequences by ribosomes is called translation and involves the use of a genetic code residing in the order of nucleotide bases. In theory, four nucleotide bases could be used in a variety of different coding systems, but theoretical considerations proposed by George Gamow in the 1950s led to the idea that a triplet of bases is the minimum requirement to code for 20 different amino acids. With the realization that a triplet of bases was involved two experiments critically defined the nature of the genetic code.

The first showed that the genetic code was not ‘punctuated’ or ‘overlapping’ but involved a continuous message. Using the bacteriophage T4 Francis Crick and Sydney Brenner showed that insertion or deletion of one or two bases resulted in the formation of nonsense proteins. However, when three bases were added or removed a protein was formed with one residue deleted or added. This demonstrated a code that was an unpunctuated series of triplets arising from a fixed starting point. The second experiment identified codons corresponding to individual amino acids through the use of cell-free extracts, capable of protein synthesis, containing ribosomes, tRNA, amino acids and amino acyl synthetases. When synthetic polynucleotide templates such as polyU were added a polymer of phenylalanine residues was produced. It was deduced that UUU coded for Phe. Similar experiments involving polyC and polyA sequences pointed to the respective codons for proline and lysine. Refinement of this experimental approach allowed the synthesis of polyribonucleotide chains containing repeating sequences that varied according to the frame in which the sequence is read. For example AAGAAGAAGAAGAAG, i.e. $(AAG)_n$ can be read as three different codons (Figure 8.32).

In a variation of this last experiment Philip Leder and Marshall Nirenberg showed that synthetic triplets would bind to ribosomes triggering the binding of specific tRNAs. UUU and UUC lead to Phe-tRNAs

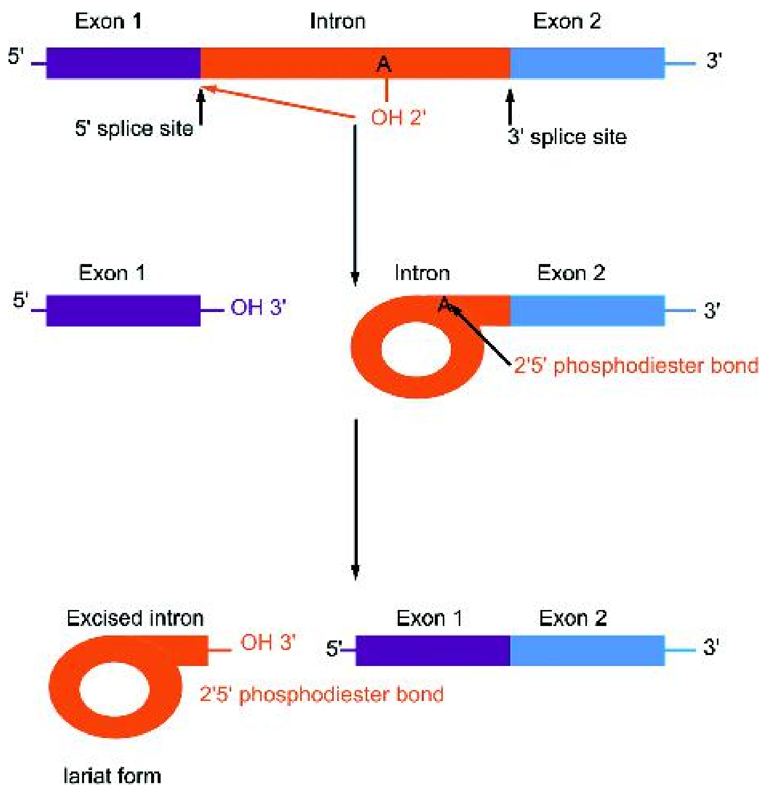


Figure 8.31 Splicing of two exons by intron removal. Exons are drawn in different shades of blue, intron in orange

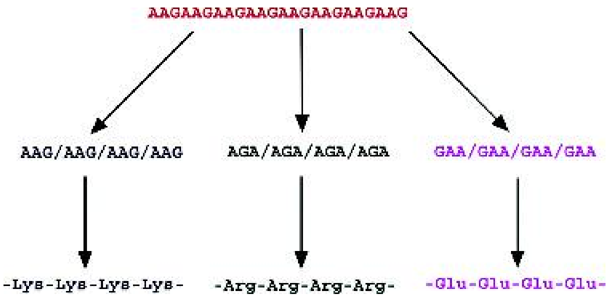


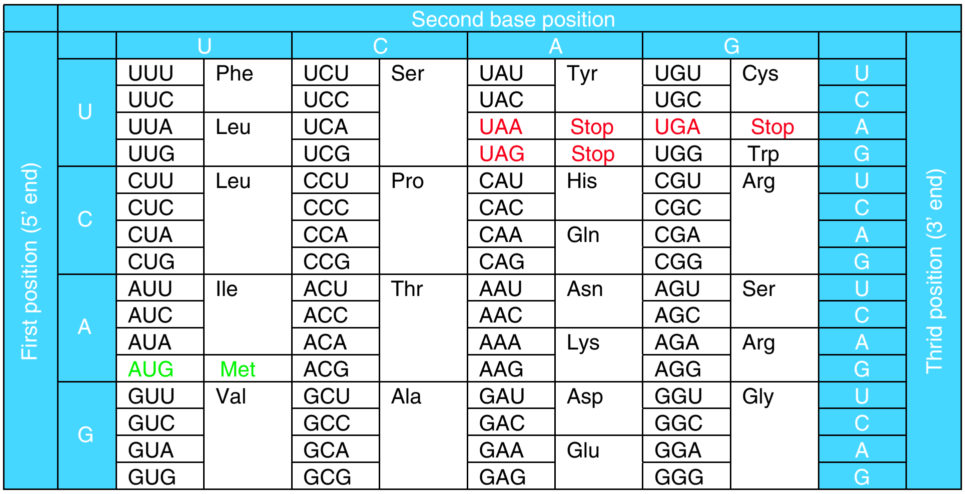
Figure 8.32 Translation of the synthetic polyribonucleotide $(AAG)_n$ leads to the production of three different amino acids in cell free translation systems. The translation of $(AAG)_n$ leads to three possible sequences depending on the 'frame' used to read the message and can yield polylysine, polyarginine or polyglutamate

binding to ribosomes and this approach deciphered the remaining codons of the genetic code confirming degeneracy in the 64 possible triplets.

The genetic code (Figure 8.33) is usually described as universal with the same codons used for the same amino acid in every organism. The ability of bacteria to produce eukaryotic proteins derived from the translation of eukaryotic mRNAs testifies to this universality. However, a few rare exceptions to this universality are documented (Table 8.5) and include alternative use of codons by the mitochondria, organelles with their own transcription and translation systems.

Transfer RNA (tRNA)

In the 1950s Francis Crick postulated that adaptor molecules should contain enzymatically appended amino acids and recognize mRNA codons. The



		Second base position							
First position (5' end)	U		C		A		G		Thrid position (3' end)
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 8.33 The genetic code. The series of triplets found in mRNA coding for the twenty amino acids are shown with the chain termination signals highlighted in red and the start codon in green

Table 8.5 Rare modifications to the genetic code

Codon	Normal use	Alternative use	Organism
AGA, AGG	Arg	Stop, Ser	Some animal mitochondria; some protozoans
AUA	Ile	Met	Mitochondria
CGG	Arg	Trp	Plant mitochondria
CUX	Leu	Thr	Yeast mitochondria
AUU	Ile	Start codon	Prokaryotic cells
GUG	Val		
UUG	Leu		

‘adaptor’ molecule was small soluble RNA now called transfer RNA (tRNA). The structure of yeast alanyl-tRNA was reported in 1965 by Robert Holley and contained 76 nucleotides. Base pairing of nucleotides led to a cloverleaf pattern and all tRNAs have lengths between 60 and 95 nucleotides showing comparable structure and function. The cloverleaf structure has distinct regions represented by arms, stems and loops (Figures 8.34 and 8.35).

At the 5′ end the short acceptor stem 7bp in length forms non-Watson–Crick base pairs with bases at the 3′ end of the chain and links via a short stem of 3–4 nucleotides to a region known as the D arm because it contains a modified base, dihydrouridine.[1] A 5bp stem from the D arm leads to the anticodon arm containing a sequence of three bases complementary to the mRNA codon. A variable loop from 3 to 21 nucleotides in length leads to a final arm containing a sequence of bases TψC where ψ represents a modified base called pseudouridine. This region is the T or TψC arm and rejoins an acceptor stem terminated by the sequence -CCA to which amino acids are added by amino acyl tRNA synthetases.

Covalent coupling of amino acids to the tRNA acceptor stem involves activation by reaction with ATP forming an amino adenylate complex, followed by coupling to tRNA in a reaction catalysed by amino acyl tRNA synthetases.

$$\text{Aminoacyl-AMP} + \text{tRNA} \rightleftharpoons \text{aminoacyl-tRNA} + \text{AMP}$$

[1] Non-Watson–Crick base pairs involve, for example, the pairing of G with U.

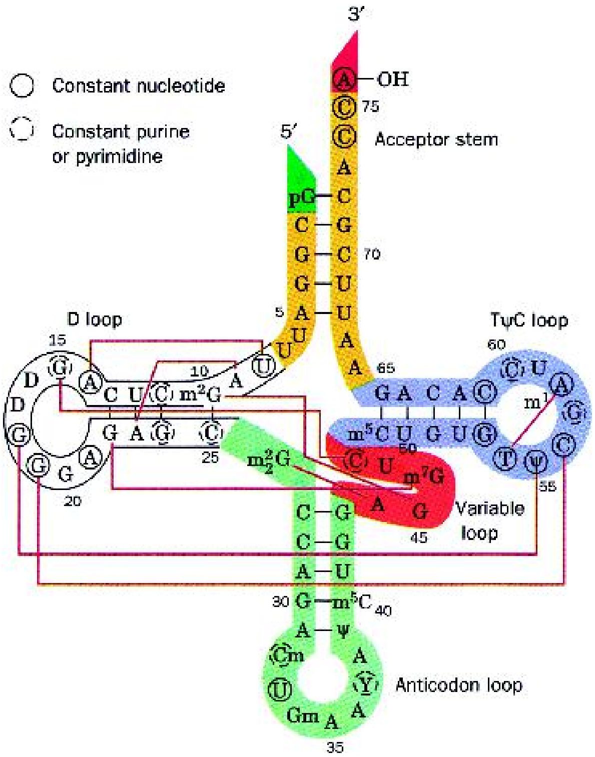


Figure 8.34 The structure of yeast tRNA with the base sequence drawn as a cloverleaf structure highlighting the acceptor stem and four arms; the TψC arm, the variable arm, the anticodon arm and the D arm. Red lines indicate base pairings in the tertiary structure with conserved and semi-conserved nucleotides indicated by solid and dashed circles respectively. The 3′ end is shown in red, the 5′ end in green with the acceptor stem in yellow. The anticodon stem is shown in light green, the D loop is white with the variable loop shown in orange and the TψC arm in cyan

Two classes of synthetases exist: class I enzymes are usually monomeric and attach the carboxyl group of their target amino acid to the 2′ OH of A76 in the tRNA molecule; class II enzymes are dimeric or tetrameric and attach amino acids to the 3′ OH of cognate tRNA (Figure 8.36) with the exception of Phe-tRNA synthetase which uses the 2′ OH. Enzymes within the same class (Table 8.6) show considerable variations in structure and subunit composition.

Table 8.6 Classification of amino acyl tRNA synthetases

Class 1			Class 2		
Amino Acid	Subunit structure and number of residues		Amino Acid	Subunit structure and number of residues	
Arg	α	577	Ala	α_4	875
Cys	α	461	Asn	α_2	467
Gln	α	551	Asp	α_2	590
Gln	α	471	Gly	$\alpha_2\beta_2$	303/689
Ile	α	939	His	α_2	424
Leu	α	860	Lys	α_2	505
Met	α_2	676	Pro	α_2	572
Trp	α_2	325	Phe	$\alpha_2\beta_2$	327/795
Tyr	α_2	424	Ser	α_2	430
Val	α	951	Thr	α_2	642

The composition of prokaryotic and eukaryotic ribosomes

In *E. coli* ribosomes may account for 25 percent of the dry mass and are relatively easily isolated from cells to reveal large structures 25 nm in diameter containing RNA and protein in a 60:40 ratio. The ribosome is always found as two subunits, one with approximately twice the mass of the other.

Prokaryotic ribosomes are described by sedimentation coefficients of 70S with two unequal subunits having individual S values of 50S (large subunit) and 30S (small subunit). Three ribosomal RNA components are identified on the basis of sedimentation coefficients – the 5S, 16S and 23S rRNAs. The small subunit has a single 16S rRNA containing ~1500 nucleotides and the large subunit contains two rRNA molecules – a 23S rRNA of about 2900 nucleotides and a much smaller 5S rRNA 120 nucleotides length. When combined with over 50 different proteins the prokaryotic ribosome has a mass of ~2.5 MDa.

Eukaryotic ribosomes are larger than their prokaryotic counterparts containing more RNA and protein. They have greater sedimentation coefficients (80S vs 70S) with subunits of 60S and 40S. The large subunit

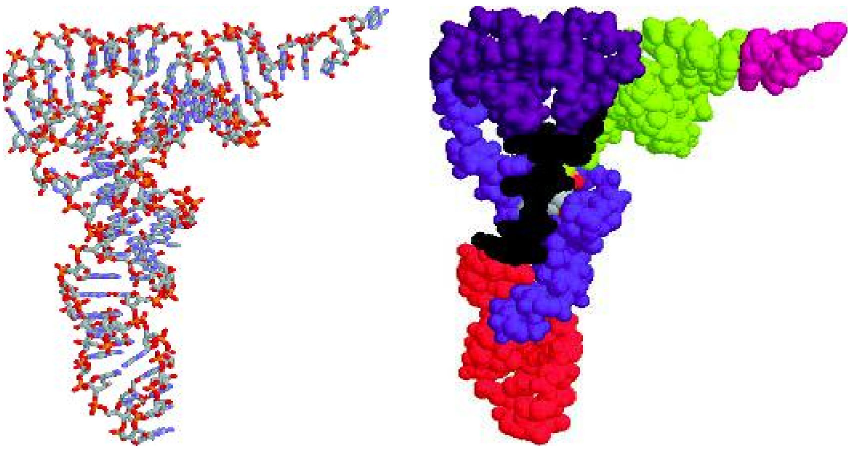


Figure 8.35 The structure of tRNA showing wireframe and spacefilling models for tRNA (PDB:6TNA). The arms are shown in different colours: acceptor stem (yellow), CCA region including A76 (cerise), variable loop (dark green), D arm (blue), the TψC arm (purple) and the anti-codon arm (red)

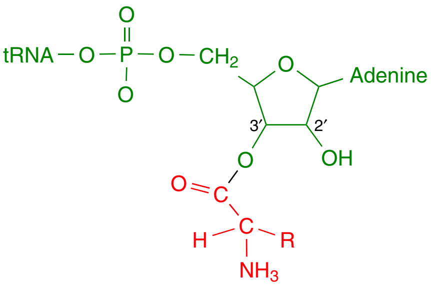


Figure 8.36 Esterification of an amino acid to the 3′ OH group of A76 in tRNA

contains 5S, 5.8S and 28S rRNA composed of 120,156 and 4700 nucleotides respectively and about 50 proteins. In contrast the small subunit has a single 18S rRNA about 1900 nucleotides in length and 32 proteins. Despite increased complexity their architecture is fundamentally the same.

By 1984 the sequences of 52 different polypeptide chains were identified in *E. coli* ribosomes with those from the large subunit designated as L1, L2 . . . etc. and those from the small subunit proteins S1 . . . S21 (Figure 8.37). The number of proteins found in ribosomes from different species varies and their homology is much lower than the rRNA components. The *E. coli* large subunit contains 31 different polypeptides each occurring once, with the exception of L7 where four copies are found. Subunit nomenclature refers to migration patterns in two-dimensional electrophoresis and as a result of partial acetylation L7 can have different mobility and was originally thought to be a distinct subunit (L12). The L7/L12 subunits associate with the L10 subunit forming a stable complex that was also mistakenly 'identified' as a separate protein (L8).

The small subunit contains 21 polypeptides, leading to a total of 52 different ribosomal proteins varying in size from small fragments with <100 residues to polypeptide chains containing in excess of 550 residues. Isolated ribosomal proteins were shown to have a high percentage of lysine and arginine residues, a low aromatic content and a topology based around antiparallel β sheets where the strands are connected by regions of α helix. This structure known as the RNA recognition motif (RRM) is seen in L7 and L30 but was also observed in other RNA binding proteins such as spliceosomal U1-snRNP (Figure 8.38).

Recognition motifs are widely found in RNA binding proteins in prokaryotes and eukaryotes suggesting

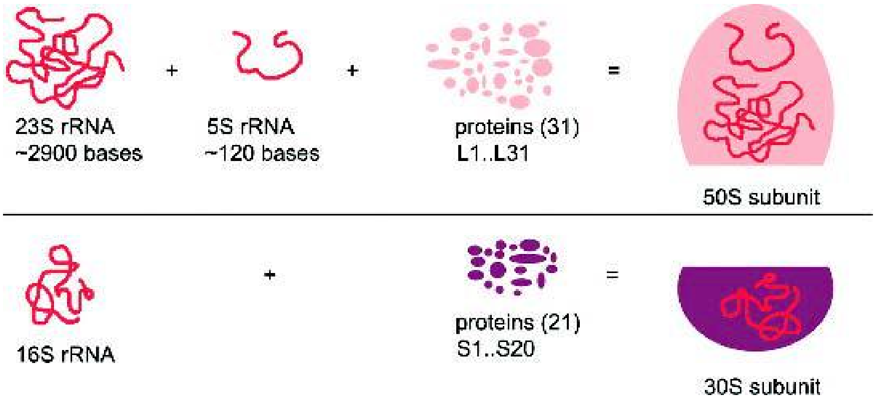


Figure 8.37 Assembly of the prokaryotic ribosome. In prokaryotes the 50 and 30S subunits combine to form a functional 70S ribosome. Association is favoured by increasing concentrations of Mg^{2+} $30S + 50S \rightleftharpoons 70S$. *In vivo* a substantial proportion of the ribosomes are dissociated and this is important during initiation of protein synthesis

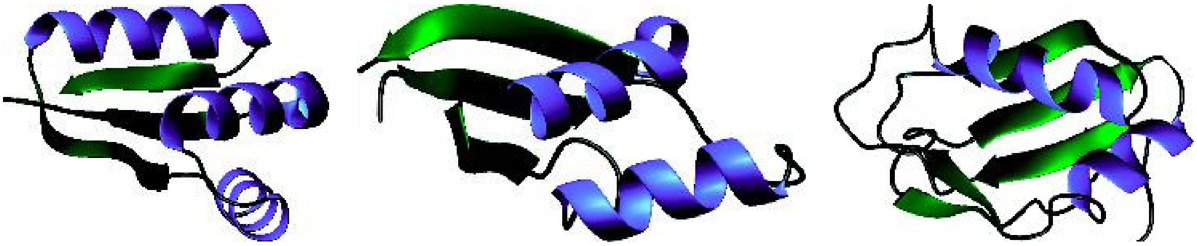


Figure 8.38 The structure of two ribosomal proteins and the U1-snRNP protein showing the homology based around three β strands and two α helices. The structures show the C terminal domain of *E. coli* ribosomal subunit L7/L12 (PDB: 1CTF), the L30 domain of the large ribosomal subunit of *T. thermophilus* (PDB:1BXY) and the N-terminal domain of the spliceosome RNA binding protein U1A (PDB:1FHT)

evolution from an ancestral RNA binding protein. All RRMs consist of a domain of 80–90 residues where the four-stranded antiparallel β sheet and helices occur in the order β-α-β-β-α-β. RNA binds on the flat surface presented by the β sheet a region carrying many conserved Arg/Lys residues on its edge to facilitate interactions with the RNA sugar-phosphate backbone. Groups of non-polar, often aromatic, residues exposed in the sheet region interact directly with purine and pyrimidine bases and in strand 1 a conserved sequence of K/R-G-F/Y-G/A-F-V-x-F/Y is found with alternate residues exposed (blue). Similarly in strand 3 a consensus hexameric sequence L/I-F/Y-V/I-G/K-N/G-L/M is observed.

Low-resolution studies of the ribosome

Despite differences in RNA and protein composition the architecture of all ribosomes is similar with a shape revealed by electron microscopy to contain lobes, ridges, protuberances and even the suggestion of channels as ‘anatomical’ features (Figure 8.39). The ribosome contains single copies of each protein and allowed antibodies to be raised against exposed epitopes. In combination with electron microscopy the distribution of proteins through the large and small subunits was mapped to enhance the picture of the ribosome (Figure 8.40).

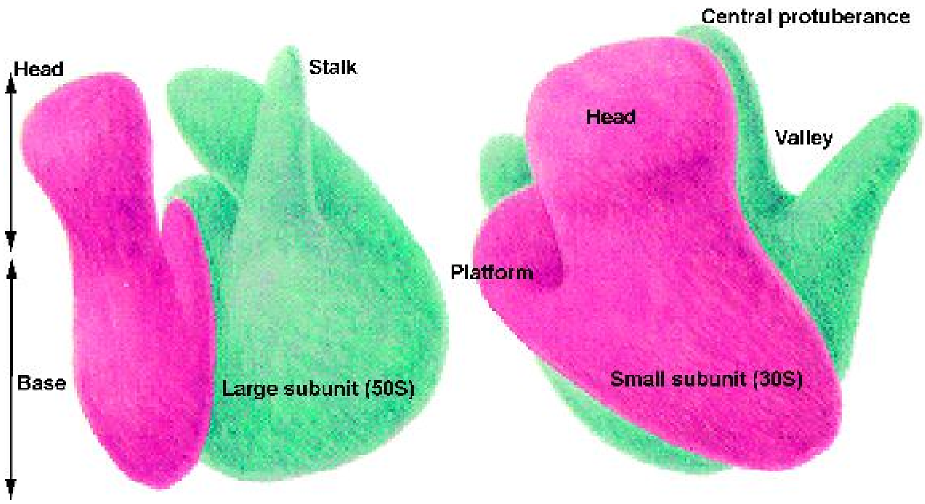


Figure 8.39 A three-dimensional model of the *E. coli* ribosome derived from electron microscopy. (Reproduced with permission from Voet, D., Voet, J.G & Pratt, C.W. *Fundamentals of Biochemistry*. John Wiley & Sons Inc, Chichester, 1999)

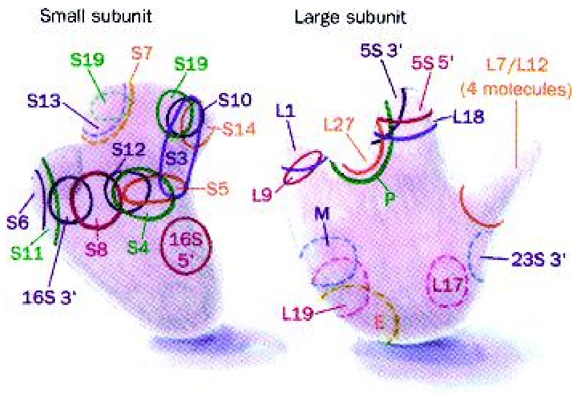


Figure 8.40 A positional map of surface epitopes for large and small *E. coli* ribosomal subunits. The small subunit is shown on the left and dashed lines indicate the positions for subunits lying of the reverse surface. The symbols 16S 3′ and 16S 5′ mark the two ends of the 16S rRNA molecule. In the large subunit P indicates the peptidyl transferase site; E marks the site of emergence for the nascent polypeptide from the 50S subunit and M specifies the ribosome's membrane anchor site. The 3′ and 5′ sites are indicated for 5S rRNA. (Reproduced with permission from Voet, D., Voet, J.G., and Pratt, C.W. *Fundamentals of Biochemistry*. John Wiley & Sons, Chicheter, 1999)

However, delineation of the catalytic sites of protein synthesis required higher levels of resolution – a major experimental problem – for an assembly with over 50 different proteins, several RNA molecules and a mass in excess of 2.5 MDa. Ribosome crystals were produced in the 1960s but yielded two-dimensional arrays of poor diffraction quality. Improved crystal quality stemmed from use of ribosomes derived from thermophiles (*Thermus thermophilus* and *Haloarcula marismortui*) where protein and nucleic acid stability was enhanced when compared with mesophiles and from advances in technology that introduced synchrotron radiation, high efficiency area detectors and crystal freezing techniques that limited radiation damage.

A structural basis for protein synthesis

The first structures for the ribosome were published in 2000 by Tom Steitz and Peter Moore and represented the culmination of decades of intensive effort. The structure of the large subunit was followed by structures for the small (30S) subunits from both *E. coli* and

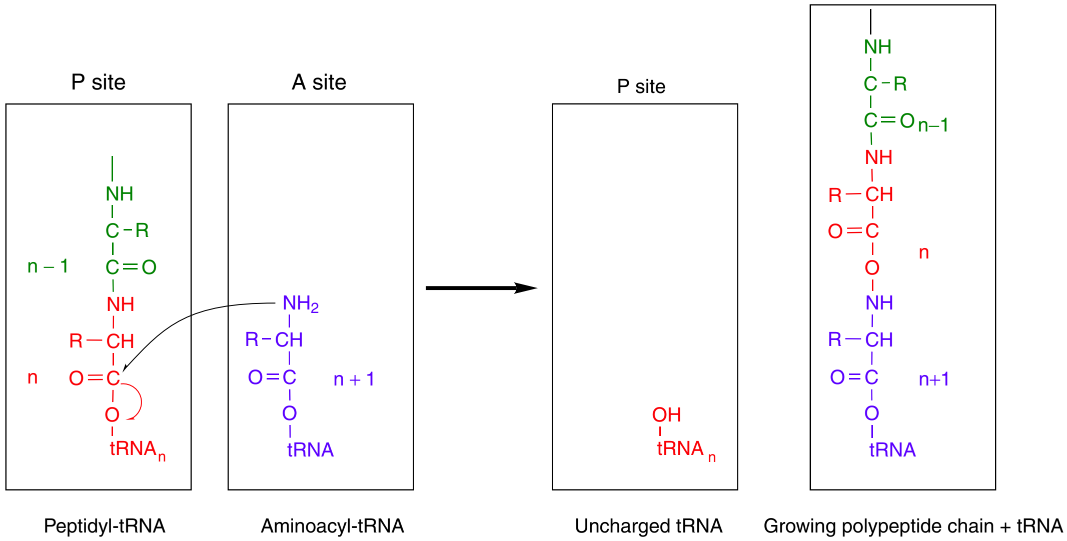


Figure 8.41 The peptidyl transferase reaction at the A and P sites

T. thermophilus. However, a broad understanding of the functional properties of prokaryotic ribosomes arose from many years of biochemical studies long before structural data was available. These studies established the self-assembly of ribosomes, inactivation of protein synthesis by antibiotics, the involvement of accessory proteins in initiation, elongation and termination, a role for GTP in protein synthesis, the nature of tRNA binding via codon/anticodon recognition and the presence of discrete binding sites within the ribosome.

An outline of protein synthesis

Translation of mRNA consists of three stages – initiation, elongation and termination – with the most important process being elongation where nascent polypeptides are extended by the addition of amino acids. Protein synthesis results in the addition of amino acid residues to a growing polypeptide chain in a direction that proceeds from the N to C terminals as mRNA is read in a 5′–3′ direction. Synthesis proceeds at surprisingly high rates for a complex reaction with up to 20 residues per second added to a growing chain.

During elongation the peptidyl transferase reaction catalyses movement and covalent linkage of an amino acid onto a growing polypeptide chain in a process involving two specific sites on the ribosome. The A site is the location for amino acyl tRNA binding to the ribosome and is proximal to the peptidyl (P) site. Displacement of extending polypeptide chains from the ribosome by high salt treatment leaves a peptidyl-tRNA species as the final 'residue'. This results from transfer of the peptidyl tRNA to the incoming amino acyl tRNA to form a peptidyl tRNA with one additional residue located in the A site. The next stage is translocation to create P site peptidyl tRNA, a vacant A site with uncharged tRNA transferred to the E or exit site (Figure 8.41).

Chain initiation

Methionine, a relatively uncommon residue, is frequently found as the first residue of a polypeptide chain modified by the addition of formic acid. In most cases the *N*-formylmethionine group is removed (Figure 8.42), but occasionally the reaction does proceed efficiently and methionine appears to be the first residue.

The initiation site is not solely defined by the AUG or start codon because a mRNA sequence will normally contain other AUG triplets for internal methionine residues. Additional structural elements such as the Shine–Dalgarno sequence located ∼10 nucleotides upstream of the start codon are complementary to a

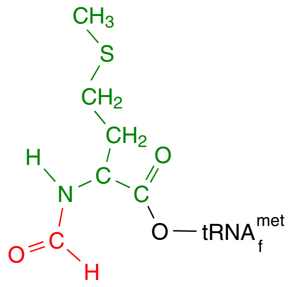


Figure 8.42 *N*-formylmethionine complexed to $tRNA_f^{met}$

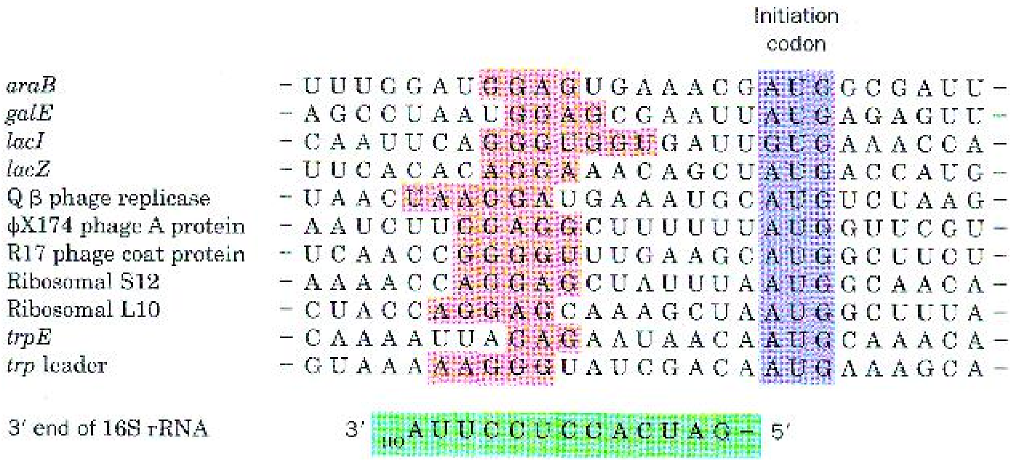
pyrimidine-rich sequence found at the 3′ end of 16S rRNA of the 30S subunit, and facilitates ribosomal selection of the correct AUG codon (Figure 8.43).

Protein synthesis requires additional proteins known as initiation factors (IF) that exhibit dynamic and transient association with the ribosome (Table 8.7). Initiation starts after completion of polypeptide synthesis with the ribosome as an inactive 70S complex. IF-3 binding to the 30S subunit promotes complex dissociation and is assisted by IF-1 (Figure 8.44). In a dissociated state mRNA, IF-2, GTP and fMet-tRNA bind to the small subunit. IF-2 is a G protein and is obligatory for binding fMet-tRNA to the 30S subunit in a reaction that is unique in not requiring mRNA and interactions between codon and anticodon. In this state mRNA binds to the small subunit completing priming reactions and the complex is capable of binding the 50S subunit. The association of the 50S subunit results in conformational change, GTP hydrolysis by IF-2 and the release of all initiation factors. The result is a ribosome primed with fMet-tRNA (P site) and a vacant A site poised to accept the next amino acyl tRNA specified by the second codon in an event that marks the start of chain elongation.

Chain elongation

Elongation involves the addition of a new residue to the terminal carboxyl group of an existing polypeptide chain. Elongation is divided into three key events: aminoacyl tRNA binding, the peptidyl transferase reaction, and translocation. Elongation factors assist in many stages of these reactions (Figure 8.45).

Charged amino acyl tRNA is escorted to the vacant A site by an elongation factor -EF-Tu. EF-Tu is yet another G protein involved in protein synthesis and after depositing amino acyl tRNA at the A site hydrolysis of GTP releases the EF-Tu–GDP complex



araB	– U U U G G A U G G A G U G A A A C G A U G G C G A U U –
galE	– A G C C U A A U G G A G C G A A U U A U G A G A G U U –
lacI	– C A A U U C A G G G U G G U G A U U G U G A A A C C A –
lacZ	– U U C A C A C A G G A A A C A G C U A U G A C C A U G –
Q β phage replicase	– U A A C U A A G G A U G A A A U G C A U G U C U A A G –
ϕX174 phage A protein	– A A U C U U G G A G G C U U U U U U A U G G U U C G U –
R17 phage coat protein	– U C A A C C G G G G U U U G A A G C A U G G C U U C U –
Ribosomal S12	– A A A A C C A G G A G C U A U U U A A U G G C A A C A –
Ribosomal L10	– C U A C C A G G A G C A A A G C U A A U G G C U U U A –
trpE	– C A A A A U U A G A G A A U A A C A A U G C A A A C A –
trp leader	– G U A A A A A G G G U A U C G A C A A U G A A A G C A –
3′ end of 16S rRNA	3′ HO A U U C C U C C A C U A G – 5′

Figure 8.43 Translation initiation sequences aligned with the start codon recognized by *E. coli* ribosomes. The RNA sequences are aligned at the start (AUG) codon highlighted in blue. The Shine–Dalgarno sequence is highlighted in red and is complementary to a region at the 3′ end of the 16S rRNA. This sequence is shown in green and involves G-U pairing. (Reproduced with permission from Voet *et al.*, John Wiley & Sons, Ltd, Chichester, 1998)

Table 8.7 Soluble protein factors involved in *E. coli* protein synthesis

Factor	Mass (kDa)	Role
Initiation		
IF-1	9	Assist in IF-3 binding
IF-2	97	Binds initiator tRNA and GTP
IF-3	22	Dissociates 30S subunit from inactive ribosome and aids mRNA binding
Elongation		
EF-Tu	43	Binds amino acyl tRNA and GTP
EF-Ts	74	Displaces GTP from EF-Tu
EF-G	77	Promotes translocation by binding GTP to ribosome. Molecular mimic of EF-Tu + tRNA.
Termination		
RF-1	36	Recognizes UAA and UAG stop codons
RF-2	38	Recognizes UAA and UGA stop codons
RF-3	46	G protein (binds GTP) and enhances RF-1 and RF-2 binding

from the ribosome. At this stage the amino acyl tRNA is checked in a process known as 'proof-reading' and if incorrect a new tRNA is reloaded to the A site. The activity of EF-Tu is aided by EF-Ts – a protein that binds to EF-Tu as a binary complex to facilitate further rounds of GTP binding and further cycles of elongation (Figure 8.46).

The structure of Ef-Tu determined by crystallography in a complex with GDP and a slowly hydrolysing analogue of GTP known as guanosine-5′-(β, γ-imido)-triphosphate (GDPNP) reveals three domains connected by short flexible peptide linkers. The 210 residue N-terminal domain with a GTP/GDP binding site undergoes structural reorganization when GTP is hydrolysed, changing orientation with respect to the remaining two domains by ~90°. The Phe-tRNA–EF-Tu/GDPNP complex is stable and shows two macromolecules arranged into the shape of a 'corkscrew' (Figure 8.47). The 'handle' consists of EF-Tu and the acceptor region (-CCA) of Phe-tRNA. The remaining part of the Phe-tRNA molecule forms the 'screw' and by comparing individual tRNA and EF-Tu structures with the complex it is clear that structural perturbations are small. In the complex formed with tRNA most interactions involve the amino acyl region of the tRNA – the CCA arm – and side chains of EF-Tu. These interactions also explain the observation of poor binding between uncharged tRNA molecules and EF-Tu.

The peptidyl transferase reaction is the second phase of elongation and involves nucleophilic reactions between uncharged amino groups of A site amino acyl-tRNA and the carbonyl carbon of the final residue of P site tRNA arranged in close proximity in the 50S subunit. Evidence points towards a site and mechanism of protein synthesis performed entirely by RNA. Ribosomes consist predominantly of highly conserved rRNA (Figure 8.48),[1] the proteins of large and small subunits show little conservation of sequence, almost all of the proteins of the large subunit of the ribosome of *T. aquaticus* could be removed without loss of activity, and mutations conferring antibiotic resistance were localized to ribosomal RNA genes.

In the final phase of elongation (translocation), uncharged tRNA is moved to the E site whilst recently extended tRNA in the A site is moved to the P site. The A site becomes vacant and will accept the next charged tRNA. Ribosomes move along the mRNA chain in the 3′ direction so that a new codon occupies the A site. This reaction involves an additional G protein, EF-G, whose structure reveals a remarkable similarity to the amino acyl-tRNA/EF-Tu/GTP complex despite an absence of sequence homology and RNA. This is not

[1] The sequence of rRNA is so highly conserved that it is used in taxonomic studies to identify new species or to highlight evolutionary lineage.

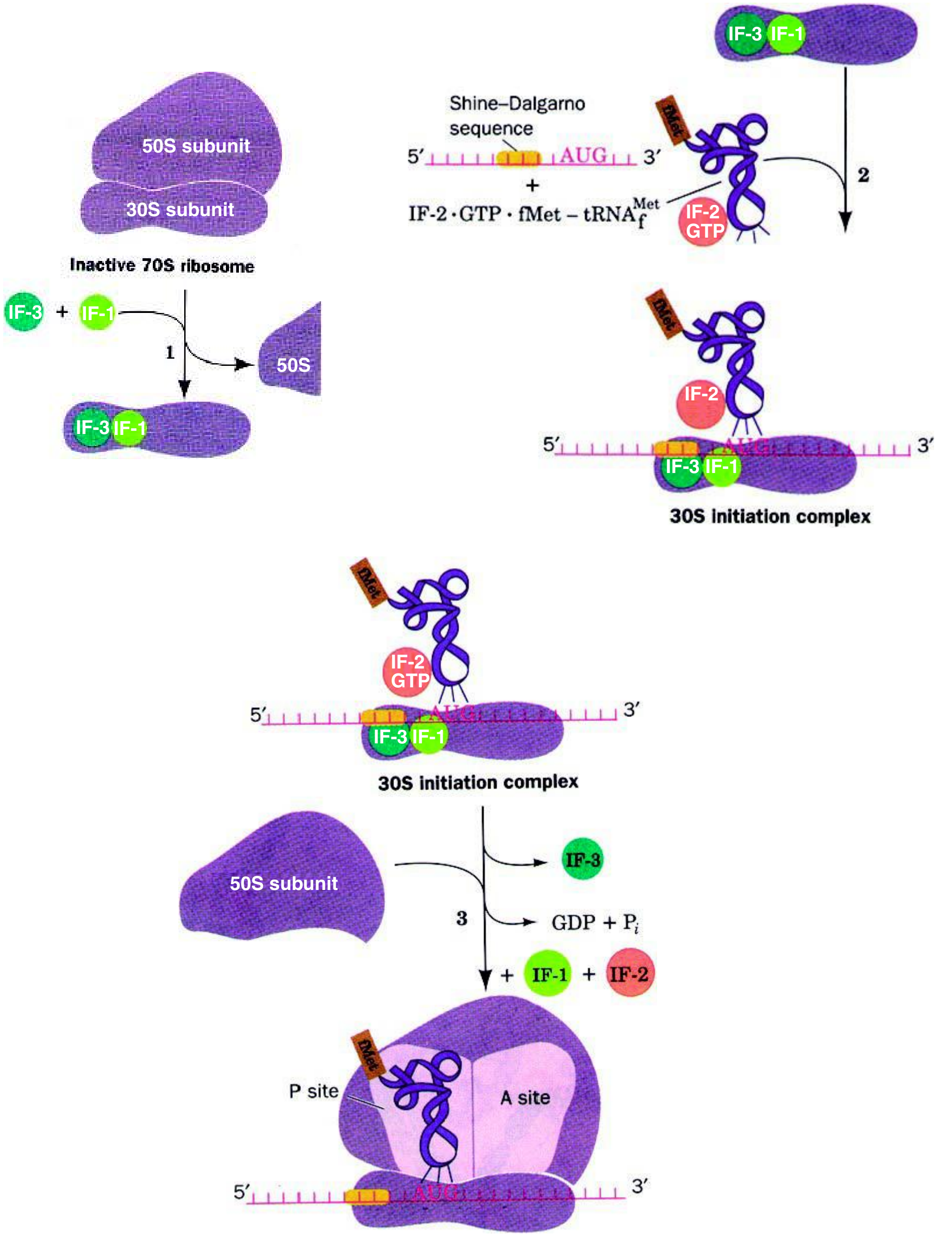


Figure 8.44 The initiation pathway of translation in *E. coli* ribosomes (reproduced with permission from Voet, D., Voet, J.G & Pratt, C.W. *Fundamentals of Biochemistry*. John Wiley & Sons, Ltd, Chichester, 1999)

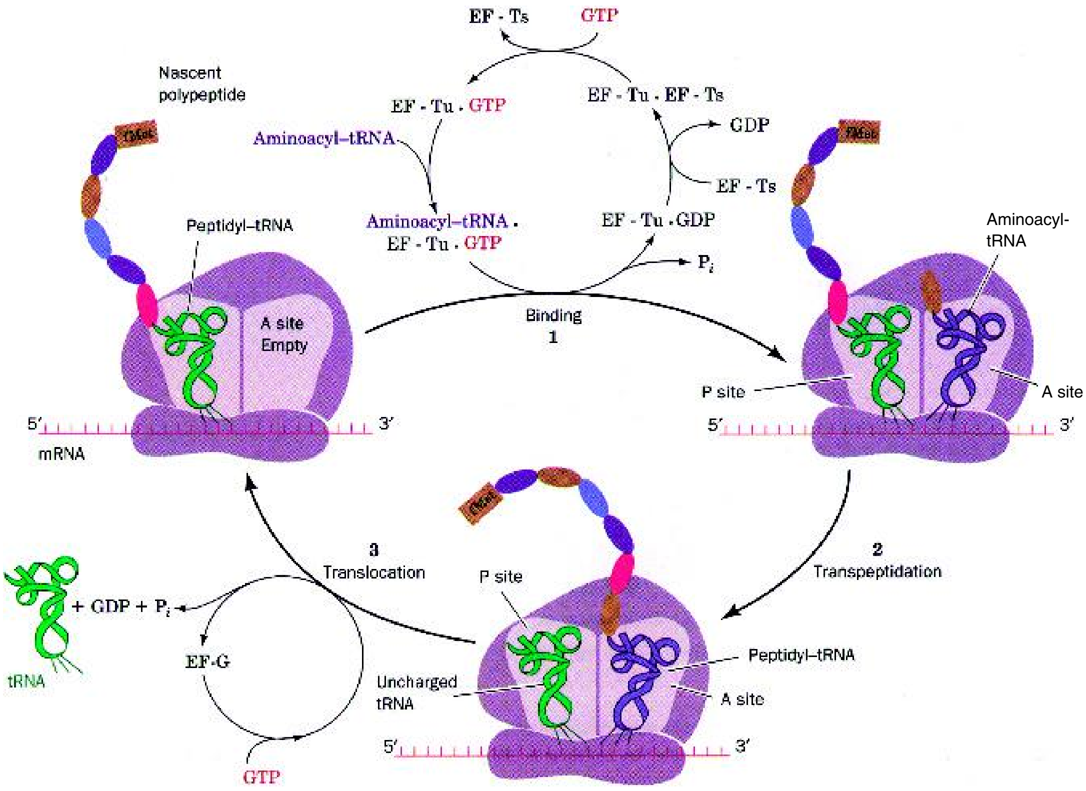


Figure 8.45 The elongation phase of translation. Only the A and P sites are shown in each stage of the elongation cycle (reproduced with permission from Voet, D., Voet, J.G & Pratt, C.W. *Fundamentals of Biochemistry*. John Wiley & Sons, Ltd, Chichester, 1999)

a coincidence but a strategy that involves competition between EF-G and the EF-Tu complex for the A site.

EF-G has five domains (Figure 8.49); domains 1 and 2 resemble the EF-Tu complex and the remaining three domains fold in a similar arrangement to the anticodon stem of tRNA. These features assist in promoting conformational changes in the ribosome by displacing peptidyl-tRNA from the A site and switching the A site to a low affinity for amino acyl tRNAs. It is an example of molecular mimicry and the 'switch' completes translocation.

Termination

Termination marks the end of protein synthesis and is mediated by three codons, UAA, UAG and UGA.

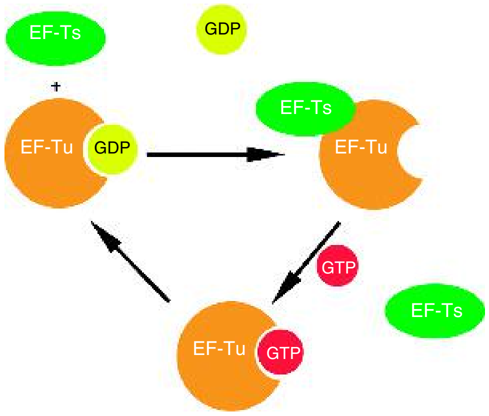


Figure 8.46 Regeneration of EF-Tu-GTP through the action of EF-Ts binding

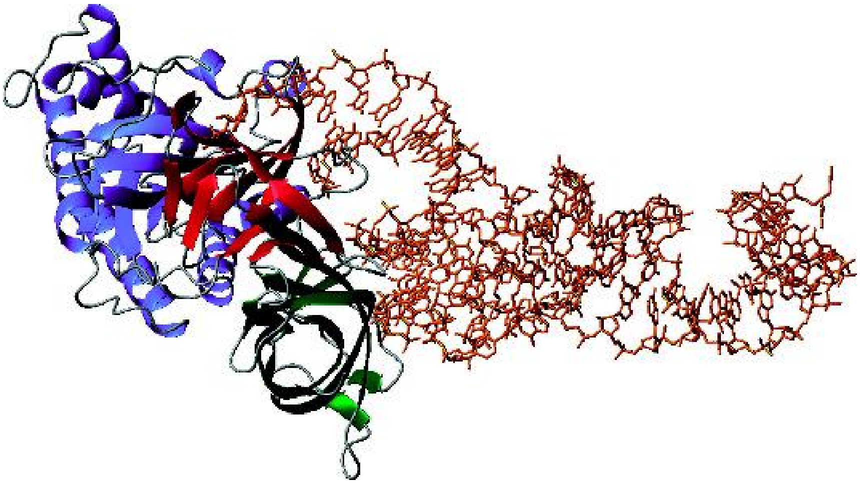


Figure 8.47 The structure of EF-Tu in a complex with Phe-tRNA (PDB: 1TTT). Domain 1 is shown in blue, domain 2 in red and domain 3 in green. The CCA arm of the Phe-tRNA is bound between the blue and red subunits

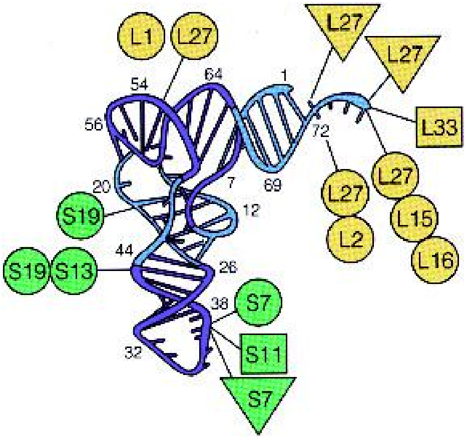


Figure 8.48 tRNA environment within the ribosome defined by cross-linking experiments. The cross links were from defined nucleotide positions within a tRNA to ribosomal proteins in the large and small subunits (denoted by prefix L and S). The A site was defined by triangular symbols, the P site by circles and the E site by squares. (Reproduced and adapted from Wower, J. *Biochimie* 1994, **76**, 1235–1246)

There are no tRNAs for termination codons but protein release factors RF-1, RF-2 and RF-3 identify the codons. RF-1 recognizes UAA and UAG whereas RF-2 identifies UAA and UGA. The third protein RF-3 is a G protein and in the presence of GTP stimulates RF-1 and RF-2 binding to the large subunit. With RF-3 and either RF-1 and RF-2 bound to the A site the peptidyl transferase reaction adds water to the end of the growing chain releasing the peptide into the cytosol along with an uncharged tRNA.

Antibiotics provide insight into protein synthesis

Some antibiotics block protein synthesis and are used to probe initiation, elongation and termination. Their effectiveness in halting prokaryotic protein synthesis coupled with their relative ineffectiveness in eukaryotic systems has allowed their use in medicines. Amongst the antibiotics used to interfere with protein synthesis in microbes are tetracycline, erythromycin, puromycin, streptomycin and chloramphenicol. These reagents block different parts of protein synthesis and from analysis of their chemical structures their effects on ribosome function can be rationalized. Chloramphenicol (Figure 8.50), for example, blocks the peptidyl transferase reaction by acting as a competitive inhibitor in which the secondary

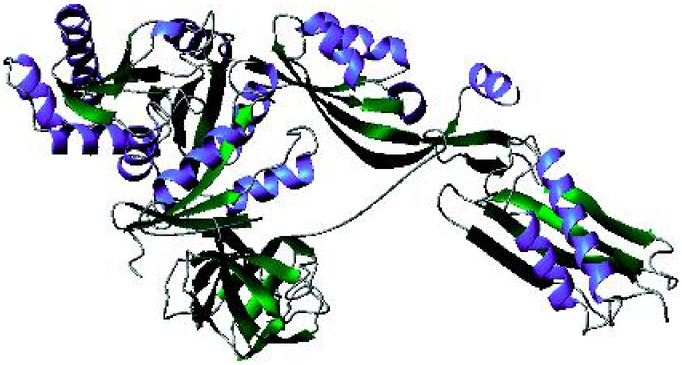


Figure 8.49 The five domains of EF-G. Domains 3–5 mimic the conformation of tRNA complexed to EF-Tu (PDB: 2EFG)

amide resembles the normal peptide bond. Similarly puromycin (Figure 8.50) contains within its structure a portion resembling the 3′ end of the amino acyl tRNA. As a result it enters the A site and is transferred to extending peptide chains causing premature release from the ribosome.

Affinity labelling and RNA ‘footprinting’

Chemical probes such as dimethyl sulfate or carbodiimides attack accessible bases of rRNA but in the presence of ligands such as antibiotics, mRNA or tRNA some regions will remain unmodified and leave a ‘footprint’ when analyzed by electrophoresis. In 23S RNA A2451and A2439 were protected by the acyl region of tRNA at the A and P sites whilst the 3′ terminal protected G2252 and G2253. These bases were universally conserved (or very nearly so) across the phylogenetic spectrum and the results point to functional importance. Similarly regions of the 16S rRNA known as helix 44, the 530 loop and helix 34 are involved in the 30S decoding site. Genetic studies further identified two bases A1492 and A1493 from their universal conservation and requirement for viability. A few ‘hot spots’ or critical bases effectively modulate protein synthesis and these experiments allowed crystallographers to assess structures on the basis of the known importance of these bases.

Structural studies of the ribosome

The structure of the large ribosomal subunit of *Haloarcula marismortui* defined sites of protein synthesis and highlighted potential mechanisms for the peptidyl

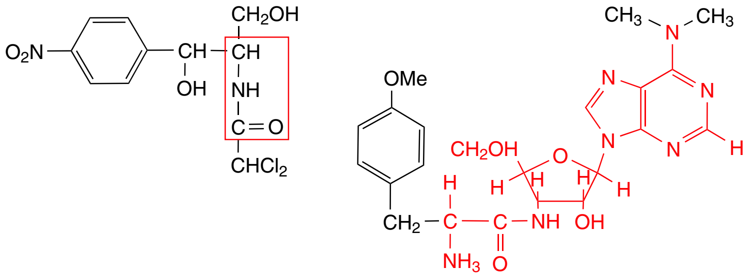


Figure 8.50 The structures of chloramphenicol and puromycin

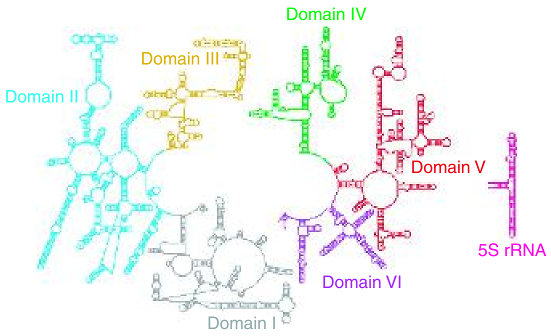


Figure 8.51 Computer-based secondary structure prediction for 23S and 5S rRNA sequences. (Reproduced with permission from Ramakrishnan, V. & Moore, P.B. *Curr. Opin. Struct. Biol.* 2001, **11**, 144–154. Elsevier)

transferase reaction with coordinates for RNA and protein representing over 90 percent of the subunit.

The secondary structure of 23S rRNA (Figure 8.51) shows regions of helix formed by base pairing as well as extended regions but provides little information on how the RNA folds to give the structure adopted by the large subunit. The 23S rRNA has six helical rich domains linked via extended loops.

The tertiary structure of 23S rRNA (Figure 8.52) showed domains that are *not* widely separated but are interwoven via helical interactions to form a compact monolithic structure. The 23S rRNA is a highly convoluted structure dictating the overall shape of the ribosome as well as defining the peptidyl transferase site.

Large subunit proteins (Figures 8.53 and 8.54) are located towards the ribosome's surface, often exposed to solvent, or on the periphery between the two subunits. Many proteins have multiple RNA binding sites and a principal function is directed towards stabilization of RNA folds. One surprising observation was that some proteins were not globular domains and for those located in crevices between rRNA helices were often extended structures. Many were basic proteins – an unsurprising observation in view of the large number of phosphate groups present within RNA.

Catalytic mechanisms within the large subunit

Using substrate analogues of the -CCA region of tRNA the location of the peptidyl-transferase site was shown to lie at the bottom of a deep cleft at the interface between large and small subunits. The peptidyl transferase reaction involves the bimolecular reaction between two substrates, namely the A site amino acyl-tRNA and the P site peptidyl-tRNA, and to enhance reactivity the molecules are constrained and in close proximity.

Transition state analogues based on puromycin pinpointed the location of the CCA arms of the A and P site tRNA molecules on the large subunit. In the presence of CCdA-phosphate-puromycin (the Yarus inhibitor, Figure 8.55) the peptidyl transferase reaction was inhibited after transfer of peptide chains from P site bound tRNA to the α amino group of puromycin. The peptidyl transferase site was based around nucleotides belonging to the central loop of 23*S* rRNA domain V a region called the 'peptidyl transferase loop' and known to bind CCdA-phosphate-puromycin. No proteins were located close to the puromycin complex and catalysis was mediated entirely by rRNA.

To facilitate formation of a tetrahedral intermediate an active functional group is required. The nearest suitable residue was the N3 atom of A2486 (A2451 in *E. coli*) located ∼3 Å from the phosphoramide oxygen of the CCdA-phosphate-puromycin. No other functional group was within 5 Å of this reaction site and more significantly there were no atoms derived from polypeptide chains within 18 Å of this centre. This residue had previously been implicated in peptidyl transferase activity from genetic, footprinting and crosslinking studies.

Under normal conditions the N1 atom of adenine monophosphate has a pK of ∼3.5 with the N3 centre observed to be an even weaker base, with a pK two units lower. The crystal structure (Figure 8.56) suggested a higher pK for the N3 atom of A2486 because the distance between the N atom and the oxygen is 3 Å and appears as a formal hydrogen bond – an interaction that would only occur if the N atom is protonated. The crystallization performed at pH 5.8 suggested a value for the p*K* of the N3 group above pH 6.0 to account for N3 protonation.

Many details of the catalytic process remain to be clarified but a charge relay network is believed to elevate the p*K* of the N3 atom of A2486 via a network of hydrogen bonds with G2482 and G2102

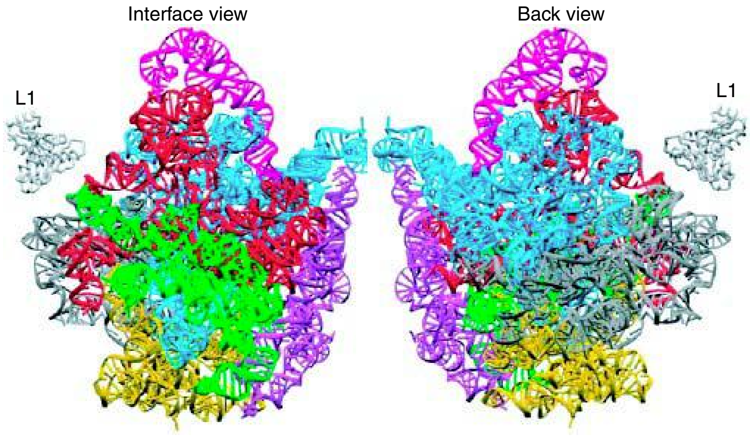


Figure 8.52 The tertiary structure of the 23S rRNA sequence. The colours utilize those shown in Figure 8.51 for the secondary structure. (Reproduced with permission from Ramakrishnan, V. & Moore, P.B. *Curr. Opin. Struct. Biol.* 2001, **11**, 144–154. Elsevier)

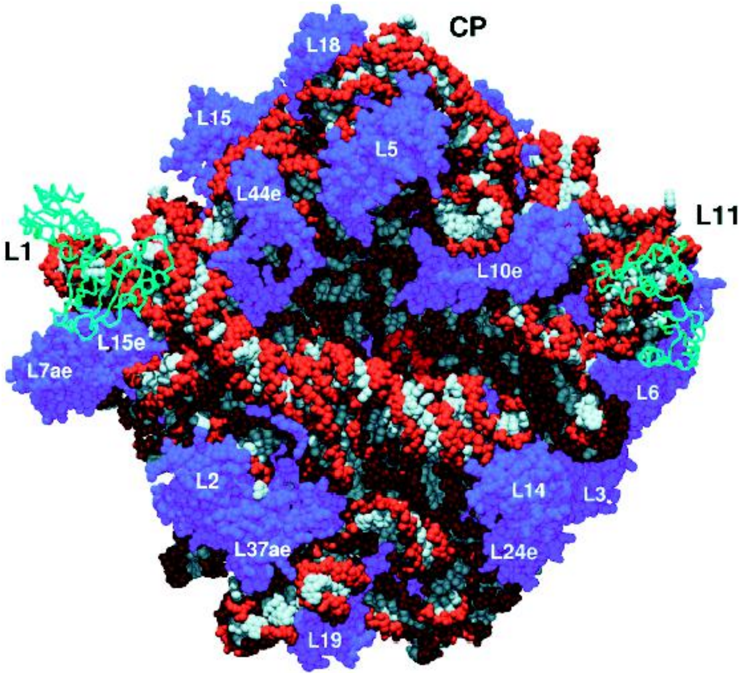


Figure 8.53 The structure of the large subunit of the ribosome of *H. marismortui*. A space-filling model of the 23*S* and 5*S* rRNA together with the proteins is shown looking down on the active site cleft. The bases are white and the sugar phosphate backbones are orange. The numbered proteins are blue with the backbone traces shown for the L1 and L11 proteins. The central protuberance is labelled CP. (Reproduced with permission from Nissen, P. *et al. Science* 2000, **289**, 920–930)

(Figure 8.57). A2486 and G2102 are conserved in the 23S rRNA sequences analysed from all three kingdoms.[1]

With this level of structural characterization it was proposed that peptidyl transferase involved the abstraction of a proton from the α amino group of the amino acyl tRNA by the N3 atom of A2486. In turn the NH_2 group of amino acyl tRNA acts as a nucleophile attacking the carbonyl group of peptidyl-tRNA. The protonated N3 stabilizes the tetrahedral carbon intermediate by hydrogen bonding to the oxyanion centre with subsequent proton transfer from the N3 to the peptidyl tRNA 3′ OH occurring as the newly formed peptide deacylates. The reaction mechanism described is not firmly established and future work will undoubtedly seek to experimentally test the proposed mechanism and role of A2486; a possible mechanism is shown in Figure 8.58.

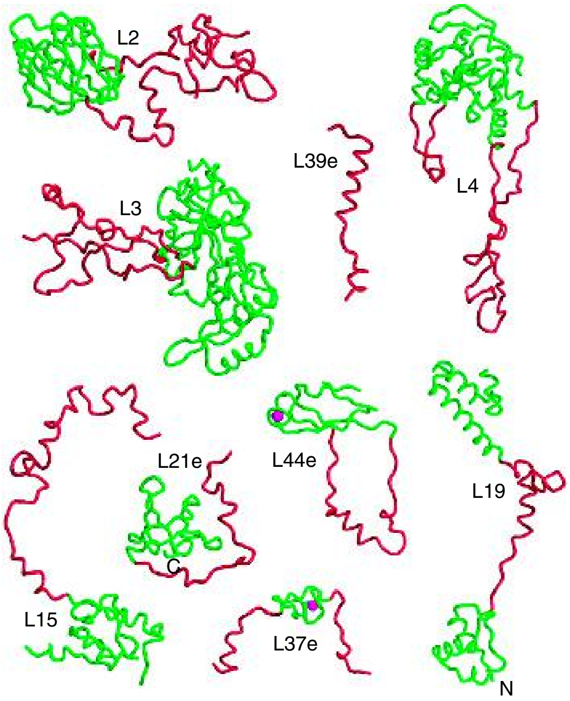


Figure 8.54 The globular domains and extended regions of proteins found in the large subunit. The long extensions would probably prove destabilizing in isolated proteins. Globular domains are shown in green with the extended region in red. (Reproduced with permission from Ramakrishnan, V. & Moore, P.B. *Curr. Opin. Struct. Biol.* 2001, **11**, 144–154. Elsevier)

The structure and function of the 30S subunit

Although the A and P sites are located on the large subunit the small subunit plays a crucial and obligatory role in protein synthesis. Protein synthesis is

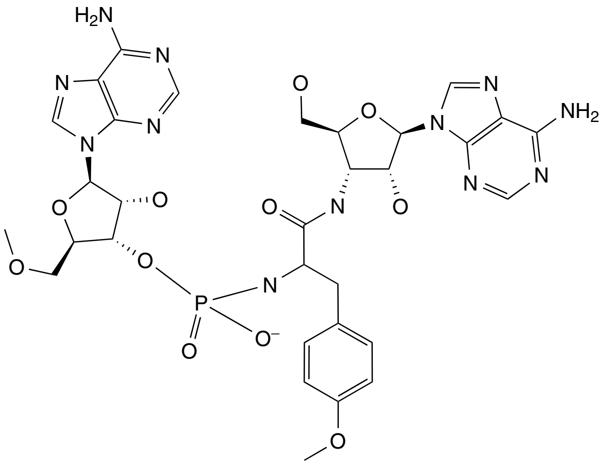


Figure 8.55 The transition state analogue CCdA-phosphate-puromycin

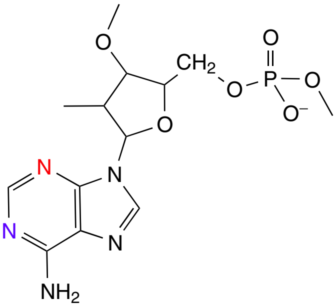


Figure 8.56 The N1 and N3 of adenine monophosphate; the N1 is shown in blue and the N3 in red

[1] G2482 is conserved at a level of ~98 percent in all sequences but is replaced by an A in some archae sequences and deleted altogether in some eubacterial sequences.

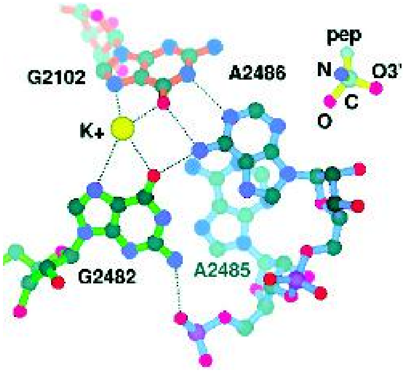


Figure 8.57 A skeletal representation with dashed hydrogen bonds showing G2482, G2102, and A2486, as well as the buried phosphate that may result in a charge relay through G2482 to the N3 of A2486. (Reproduced with permission from Nissen, P. *et al. Science* 2000, **289**, 920–930)

a multi-step process and starts with IF-3 binding to the small subunit. The 30S subunit plays a direct role in 'decoding' mRNA by facilitating base pairing between mRNA codon and the anticodon of relevant tRNAs.

The small subunit from *T. thermophilus* was crystallized in the 1980s but poor crystal diffraction properties limited use for many years until improvements in resolution occurred with removal of the S1 subunit from the ribosome prior to crystallization. Structures, initially at a resolution of 5.5 Å were followed by resolutions below ∼3 Å and detailed studies of the 30S ribosomal subunit were largely the results of two groups headed by Ada Yonath and V. Ramakrishnan. The structure defined ordered regions of 16S rRNA and confirmed the general shape of the small subunit deduced previously using microscopy. All morphological features were derived from RNA and not protein.

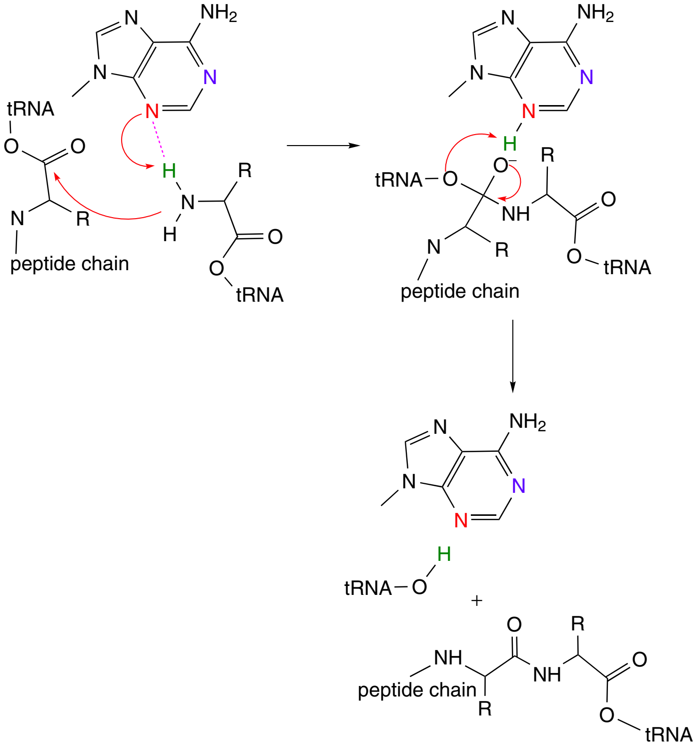


Figure 8.58 A possible mechanism for the peptidyl transferase involving the N3 atom of A2486 (adapted from Nissen, P. *et al. Science* 2000, **289**, 920–930)

The 16S RNA structure

The 16S rRNA contributes a significant proportion of the mass and volume of the 30S subunit and consists of approximately 50 elements of double stranded helix interspersed with irregular single stranded loops. The 16S RNA fold consists of four domains based on inter-helical packing and interactions with proteins (Figure 8.59). Anatomical features define the 30S structure with a head region containing a beak that points away from the large subunit. This structure is on top of a shoulder region whilst at the bottom a spur or projection is observed with a main interface defined by body and platform areas (Figure 8.60).

The 5′ domain of 16S rRNA forms the body region, the central domain and most of the platform of the 30S subunit. In contrast, the 3′ major domain constitutes the bulk of the head region whilst the 3′ minor domain forms part of the body at the subunit interface. The four domains of the 16S rRNA secondary structure radiate from a central point in the neck region of the subunit and are closely associated in this functionally important region of the 30S subunit.

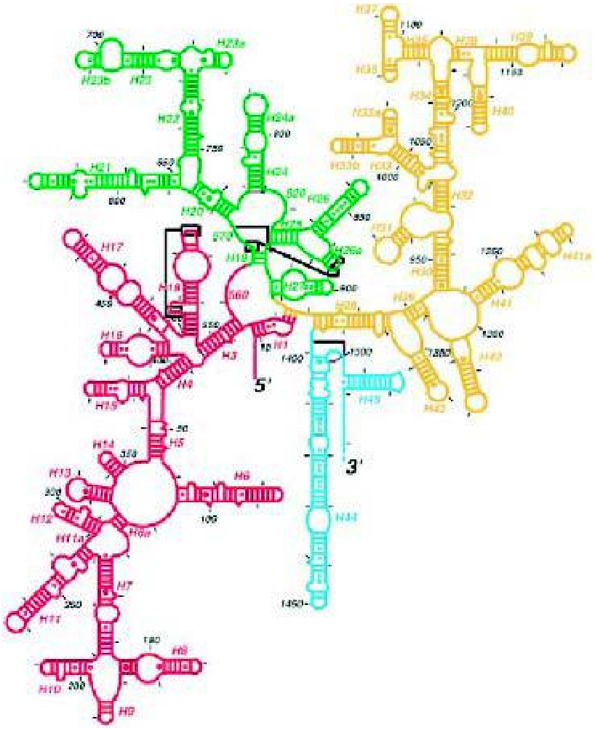


Figure 8.59 Secondary structure diagram of 16S RNA showing the various helical elements. The 5′ domain is shown in red; a central domain in green; a major domain of the 3′ region in orange–yellow and a minor domain in this region in cyan

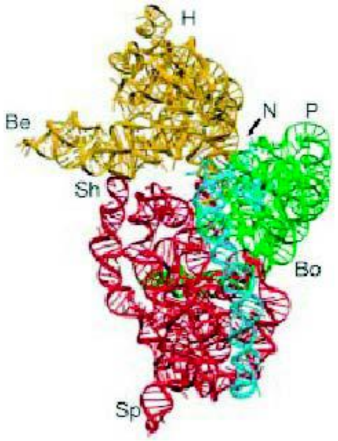


Figure 8.60 The tertiary structure of 16S RNA with the same colour scheme for the domains as in Figure 8.59. The model shows H, head; Be, beak; N, neck; P, platform; Sh, shoulder; Sp, spur; Bo, body regions of the 30S subunit from the perspective of the 50S subunit. (Reproduced with permission from Ramakrishnan, V. & Moore, P.B. *Curr. Opin. Struct. Biol.* 144–154. 2001, **11**, Elsevier)

Proteins of the small subunit

Small differences in composition of 30S subunits were noticed between *T. thermophilus* and the previously dissected *E. coli* ribosome. *Thermus* lacks subunit S21 but contains an additional short 26 residue peptide fragment. As a result of S1 removal the 30S structures contain only subunits S2–S20 along with the short 26 residue peptide fragment. Many proteins are located at junctions between helices of the RNA. For example, the S4 subunit binds to a junction formed by five helices in the 5′ domain whilst S7 binds tightly to a junction formed from four rRNA helices in the 3′ major domain. Both proteins are important in the

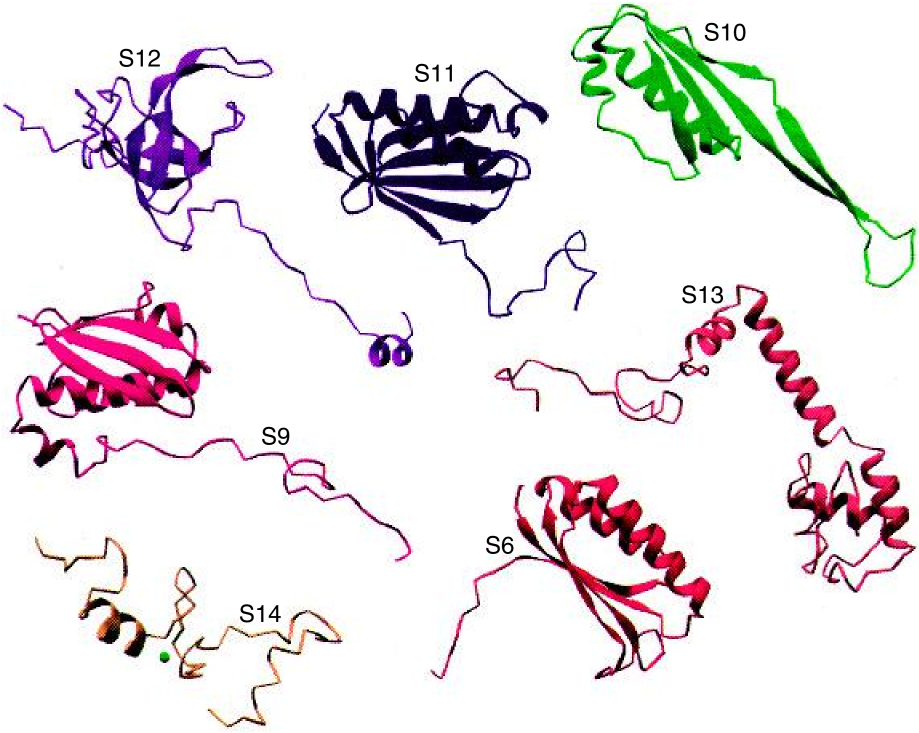


Figure 8.61 The globular domains and extended regions of some of the proteins found in the 30S subunit. Long extensions are probably destabilizing in isolated proteins. S14 is a Zn-binding protein, the cation is shown by a green sphere. (Reproduced with permission from Brodersen, D.E. *et al. Cold Spring Harbor Symp.* 2001, **66**, 17–32. CSHL Press)

assembly of the 30S subunit forming parts of the body and head, respectively. Table 8.8 summarizes the polypeptide structures of the small subunit.

Almost all of the proteins contain one or more globular domains and common topologies such as the β barrel are found in subunit S12 and S17. The packing of α-helices against an extended β-sheet is observed in several proteins such as S3, S10, S6 and S11. However, a defining characteristic is the relatively long extended regions found in subunits. These extensions are ordered, occur at the N or C terminal and may contain helical regions such as hairpin structures (S2) or C-terminal helices (S13). The extensions include loops or long β hairpins (S10 and S17) or 'tails' to proteins such as S4, S9, S11, S12, S13 and S19 (Figure 8.61). In almost all cases their role is to stabilize the RNA fold. In the 30S subunit the extensions reach far into cavities surrounded by RNA and make contact with several RNA elements. The extensions are well suited to this role since they are narrow, allowing close approach to different RNA elements, and they have basic patches to counter the highly charged sugar-phosphate backbone of RNA. Although protein–RNA interactions are obviously important, some of the protein subunits interact with each other via hydrophobic contacts. S3, S10 and S14 form a tight cluster held together by hydrophobic interactions. Other subunits interact via electrostatic and hydrogen bonding interactions and these include S4, S5 and S8.

Functional activity in the 30S subunit

The major function of the small subunit is decoding and matching the anticodon of tRNA with the mRNA

Table 8.8 Summary of the structural properties of polypeptide chains in the 30S subunit. (Adapted from Wimberley, B.T. *et al. Nature* 2000, **407**, 327–339. Macmillan).

Protein	Residues	No. of domains	Secondary structure in domains	Protein interaction	Unusual features
S2	256	2	α_2, $\alpha_1\beta_5\alpha_3$	None	Extended α hairpin
S3	239	2	$\alpha_2 b_3$, $\alpha_2\beta_4$	S10, S14	N-terminal tail
S4	209	3	Zn finger, α_4, $\alpha_3\beta_4$	S5	N-terminal Zn finger
S5	154	2	$\alpha_1\beta_3$, $\alpha_2\beta_4$	S4, S8	Extended β hairpin
S6	101	1	$\alpha_2\beta_4$	S18	C-terminal tail
S7	156	1	$\alpha_6\beta_2$	S9, S11	Extended β hairpin
S8	138	2	$\beta_2\alpha_3$, $\alpha_1\beta_3$	S5, S12, S17	
S9	128	1	$\beta\alpha_3\beta_4$	S7	Long C-terminal tail
S10	105	1	$\alpha_2\beta_4$	S3, S14	Long β hairpin
S11	129	1	$\alpha_2\beta_5$	S18, S7	Long N-terminal tail
S12	135	1	$\alpha_1\beta_5$	S8, S17	Long N-terminal tail and extended β-hairpin loops
S13	126	1	α_3	S19	Long C-terminal tail
S14	61	1	none	S3, S10	Zn module mostly extended
S15	89	1	α_4	none	
S16	88	1	$\beta_1\alpha_2\beta_4$	none	C terminal tail
S17	105	1	β_5	S12	C-terminal helix + β hairpin loops
S18	88	1	α_4	S6	β strand extends S11 sheet
S19	93	1	$\alpha_1\beta_3$	S13	
S20	105	1	α_3	none	

Partial structural information was available for S4–S8, and S15–S19 and aided their identification in crystals of 30S subunits.

codon. Within the 30S subunit the A, P and E sites were defined by RNA elements derived from different domains (45 different helical domains are found in 16S rRNA designated as H1–H45). The A and P sites are defined predominantly by RNA with helix 44, helix 34 and the 530 loop together with S12 forming part of the A site. In addition, the extended polypeptide chains of S9 and S13 intrude into the tRNA binding sites and may form interactions with tRNA. In contrast, the E site is largely defined by protein.

A combination of molecular genetics, sequence analysis and biochemical studies highlighted A1492 as an important base to the decoding process. The structure of the small subunit confirmed the importance of A1492 along with its neighbour and conserved base A1493. The crux of the decoding process is the ability to discriminate between cognate and near cognate tRNAs and this activity lies in the unique conformation of bases found at the A site. Discrimination against non-cognate tRNA results in two or three mismatches in base pairing at the codon–anticodon level with the high energetic cost making the process unfavourable.

The structure of the small subunit is split into domains, unlike the 50S subunit, with each possessing independent mobility and an involvement in conformational changes that influence ribosomal activity. The E site is predominantly protein and formed by the S7 and S11 subunits, a small interface between subunits forms the anticodon stem loop binding site whilst an extended β hairpin structure in S7 plays a role in dissociation of the vacant tRNA molecule from the ribosome.

Strong sequence conservation of rRNA suggests that RNA topology is likely to be maintained within the eukaryotic ribosomal subunits although there is clearly an insertion of elements with the formation of 5.8S, 18S and 28S rRNA. Yeast has a 18S rRNA 256 nucleotides longer than the 16S rRNA (*E. coli*) although the pattern of stem loops and major domains seen in 16S rRNA is mirrored by rRNA of yeast and other eukaryotes. However, the 40S and 60S subunits possess increased numbers of polypeptides relative to their *E. coli* counterparts, there are functional differences in the mechanism of initiation, elongation involves a greater variety and number of accessory proteins whilst the mechanisms of antibiotic inhibition are not shared with prokaryotes.

The remarkable confirmation that the ribosome is a ribozyme defined a new era in structural biology. The structures produced for the 50 and 30S subunits assimilates and unifies four decades of biochemical data on the ribosome and provides a wealth of new information about RNA and protein structure, their respective interactions and ribosome assembly. In 50 years our knowledge has progressed from the initial elucidation of the atomic structure of DNA to a near complete description of replication, transcription and translation.

Post-translational modification of proteins

The initial translation product may not represent the final, mature, form of the protein with some polypeptide chains undergoing additional reactions called post-translational modification. These reactions include processing to remove sequences normally at the ends of the molecule, the addition of new groups such as phosphate or sugars, and the modification of existing groups such as the oxidation of thiol groups. In almost all cases these modifications are vital to protein structure and function.

Proteolytic processing

The digestive enzymes chymotrypsin, trypsin and pepsin function in the alimentary tracts of animals to degrade proteins. These enzymes are initially translated as inactive forms called zymogens (chymotrypsinogen, trypsinogen and pepsinogen) to prevent unwanted degradation. Zymogens are converted into active enzyme by removal of a short ‘pro’ sequences. Such processes are important in the production of pancreatic proteases and mechanisms such as the blood clotting cascade (see Chapter 7).

Peptide hormones as examples of processing

Peptide hormones are frequently synthesized as longer derivatives. Angiotensin, a hormone involved in the control of vasoconstriction and insulin, with a role in the maintenance of blood glucose concentrations are ‘processed’ hormones. The ‘pro’ sequences may also have additional sequences located at the N-terminal to facilitate targeting to specific intracellular compartments. For example insulin, is synthesized as preproinsulin (Figure 8.62).

The ‘pre’ sequence transfers insulin across the endoplasmic reticulum (ER) membrane into the lumen where cleavage by proteases leaves the ‘pro’ insulin molecule (Figure 8.63). In this state folding starts with disulfide bridge formation stabilizing tertiary structure. A fragment known as the C peptide is released (Figure 8.64) and persists in secretory granules seen in pancreatic cells. The insulin remains inactive as a hormone and in high concentrations in the pancreas. The chain requires further modification and a second proteolytic reaction cuts the primary

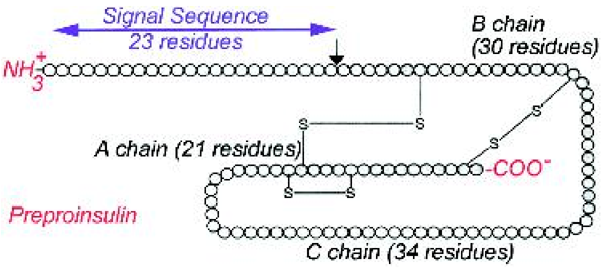


Figure 8.62 Preproinsulin contains 110 residues within a single polypeptide. Each circle represents a single amino acid residue

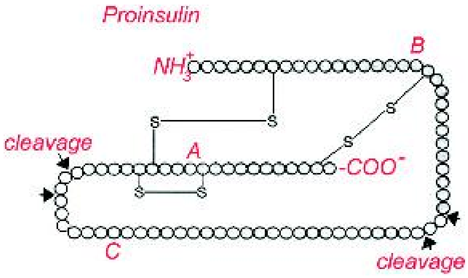


Figure 8.63 Signal sequence removal yields proinsulin. The A chain has an intramolecular disulfide bridge as well as two intermolecular disulfide bridges with the B chain

sequence at pairs of Lys-Arg and Arg-Arg residues between the A and B regions creating two separate polypeptides.

A very dramatic example of processing is the peptide hormone opiomelanocortin (POMC). The POMC gene is expressed in the anterior and intermediate lobes of the pituitary gland where a 285-residue precursor undergoes differential processing to yield at least eight hormones (Figure 8.65).

Disulfide bond formation

The formation of disulfide bonds between cysteine residues is a common post-translational modification resulting in strong covalent bonds. The covalent bond restricts conformational mobility in a protein and normally occurs between residues widely separated in the primary sequence. For example, in ribonuclease four disulfide bridges form between cysteine residues 26–84, 40–95, 58–110 and 55–72.

Disulfide bond formation occurs post-translationally in the ER of eukaryotes or across the plasma membrane of bacteria in proteins destined for secretion. The formation of disulfide bonds is often a rate-limiting step during folding, and sometimes the 'wrong' disulfide bond forms where there is a potential for multiple bridges. In all cells enzymes catalyse efficient formation of disulfide bonds and correct 'wrong' disulfide bonds by rapid isomerization. These enzymes are collectively called disulfide oxidoreductases.

Eukaryotes and prokaryotes catalyse disulfide formation with different groups of enzymes. In prokaryotes the Dsb family of enzymes are identified in Gram-negative bacteria and carry names such as DsbA, DsbB, DsbD etc. In *E. coli* DsbA contains a consensus motif Cys-Xaa-Xaa-Cys and shows a similar fold to thioredoxin despite a low sequence identity (<10 percent). In the cytoplasm of *E. coli* thioredoxin reduces disulfides whilst in the periplasm DsbA oxidizes thiol groups. Since both thioredoxin and DsbA contain the Cys-Xaa-Xaa-Cys motif the ability to oxidize disulfide bonds centres around the properties of this catalytic centre. In DsbA the first cysteine (Cys30) has a low pK of 3.5, compared with a normal value around 8.5, and the sidechain forms a reduced thiolate anion S^- representing a highly oxidizing centre that allows DsbA to catalyse disulfide formation.

In eukaryotes comparable reactions are performed by protein disulfide isomerase (PDI). Again the activity of PDI depends on the diagnostic motif Cys-Xaa-Xaa-Cys and when the active-site cysteines are present as a

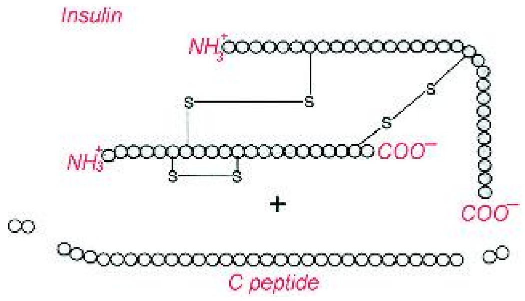


Figure 8.64 The processed (active) form of insulin plus proteolytic fragments

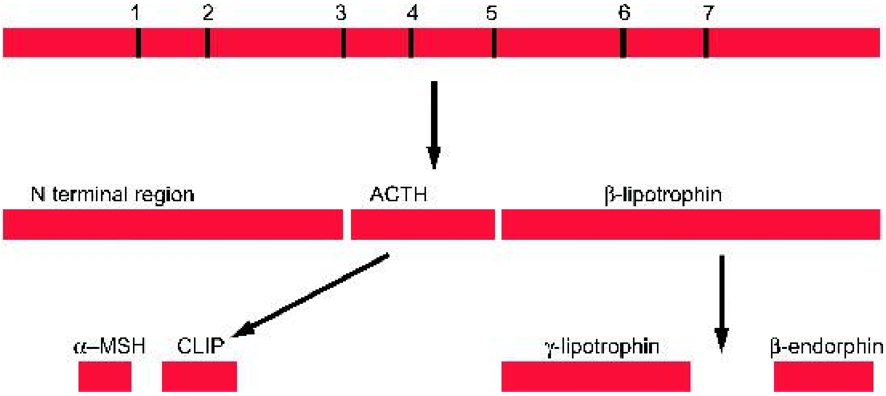


Figure 8.65 The processing of pro-opiomelanocortin yields additional bioactive peptides. Initial processing of pro-opiomelanocortin in the anterior and intermediate lobes of the pituitary yields an N-terminal fragment; ACTH, adrenocorticotrophic hormone (39 residues); and β-lipotrophin. Further processing of ACTH in the intermediate lobes yields a melanocyte stimulating hormone – a 13-residue peptide acetylated at the N-terminal and amidated at the C-terminal residue together with CLIP, a corticotropin-like intermediate lobe peptide of 21 residues. Further processing of β lipotrophin yields γ-lipotrophin and β endorphin. The peptides are 59 and 26 residues in length respectively with the β endorphin acetylated at the N terminal

disulfide the enzyme transfers disulfide bonds directly to substrate proteins acting as a dithiol oxidase. Under more reducing conditions with thiols in the active-site the enzyme reshuffles disulfides (isomerase) in target proteins.

Hydroxylation

Hydroxylation is another example of a post-translational modification. It is particularly important in the maturation of collagen where hydroxylation of proline and lysine residues found in the Gly-Xaa-Xaa motif occurs in procollagen in the ER as part of the normal secretory pathway. The reaction maintains structural rigidity in collagen enabling a role as a biological scaffold or framework. Proline is hydroxylated most commonly at the γ or fourth carbon by the enzyme prolyl-4-hydroxylase in reactions requiring oxygen, ascorbate (vitamin C) and α-ketoglutarate (Figure 8.66).

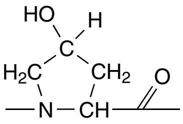


Figure 8.66 4-Hydroxyproline, a common post-translational modification observed in collagen

The requirement of vitamin C for effective folding and stability of collagen emphasizes the physiological consequences of scurvy. Scurvy is a disease arising from a lack of ascorbate in the diet, and was frequently experienced by sailors in the 16th and 17th centuries during long maritime voyages with the absence of fresh fruit or vegetables containing high levels of vitamin C. Scurvy can also occur under conditions of malnourishment and is sometimes seen in the more disadvantaged regions of the world where famine is prevalent. Hydroxylation of proline residues stabilizes collagen by favouring additional hydrogen bonding and in its absence the microfibrils are weaker. One observed effect is a weakening of connective tissue surrounding the teeth, a common symptom of scurvy. Proline residues are also hydroxylated at the β or C3 position with lower frequency.

In an entirely analogous manner lysine residues of collagen are hydroxylated in a reaction conferring increased stability on triple helices by allowing the

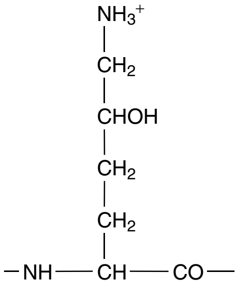


Figure 8.67 The hydroxylation of lysine at the C5 or δ position

subsequent attachment of glycosyl groups that form cross links between microfibrils. Lysine residues are hydroxylated at the δ carbon (C5) by the enzyme lysyl-5-hydroxylase (Figure 8.67).

Phosphorylation

The phosphorylation of the side chains of specific serine, threonine or tyrosine residues is a general phenomenon known to be important to functional activity of proteins involved in intracellular signalling. The addition of phosphoryl groups occurs through the action of specific enzymes (kinases) that utilize ATP as a donor whilst the removal of these groups is controlled by specific phosphatases. A generalized scheme for phosphorylation is

$$\text{Protein} + \text{ATP} \rightleftharpoons \text{protein-P} + \text{ADP}$$

and results in the products phosphoserine, phosphotyrosine, and phosphothreonine (Figure 8.68). More rarely other residues are phosphorylated such as His, Asp and Lys.

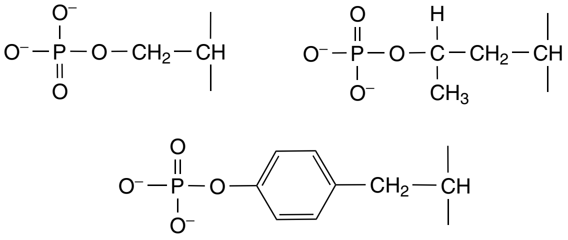


Figure 8.68 The phosphorylation of serine, tyrosine and threonine residues

Glycosylation

Many proteins found at the cell surface are anchored to the lipid membrane by a complex series of glycosyl-phosphatidyl inositol (GPI) groups. These groups are termed GPI anchors and they consist of an array of mannose, galactose, galactosamine, ethanolamine and phosphatidyl inositol groups. All eukaryotic cells contain cell-surface proteins anchored by GPI groups to the membrane. These proteins have diverse functions ranging from cell-surface receptors to adhesion molecules but are always located on exterior surfaces.

The term GPI was first introduced for the membrane anchor of the variant surface glycoprotein (VSG) of *Trypanosoma brucei*, a protozoan parasite that causes sleeping sickness in humans (a disease that is fatal if left untreated and remains of great relevance to sub-Saharan Africa). GPI anchors are particularly abundant in protozoan parasites. The overall biosynthetic pathway of GPI precursor is now well understood with signals in the form of consensus sequences occurring within a polypeptide chain as sites for covalent attachment of GPI anchors. This site is often a region of ~20 residues located at the C-terminal, and specific proteases cleave the signal sequence whilst transamidases catalyse the addition of the GPI anchor to the newly generated C-terminal residue. The outline organization of GPI anchors (Figure 8.69) involves an ethanolamine group linked to a complex glycan containing mannose, galactose and galactosamine and an inositol phosphate before linking to fatty acid chains embedded in the lipid bilayer.

Alongside GPI anchors the principal post-translational modification involving glycosylation involves the addition of complex sugars to asparagine side chains (N-linked) or threonine/serine side chains (O-linked) in the form of glycosidic bonds. In N-linked glycosylation an oligosaccharide is linked to the side chain N of asparagine in the sequence Asn-X-Thr or Asn-X-Ser, where X is any residue except proline. The first sugar to be attached is invariably *N*-acetylglucosamine and strictly occurs co-translationally as

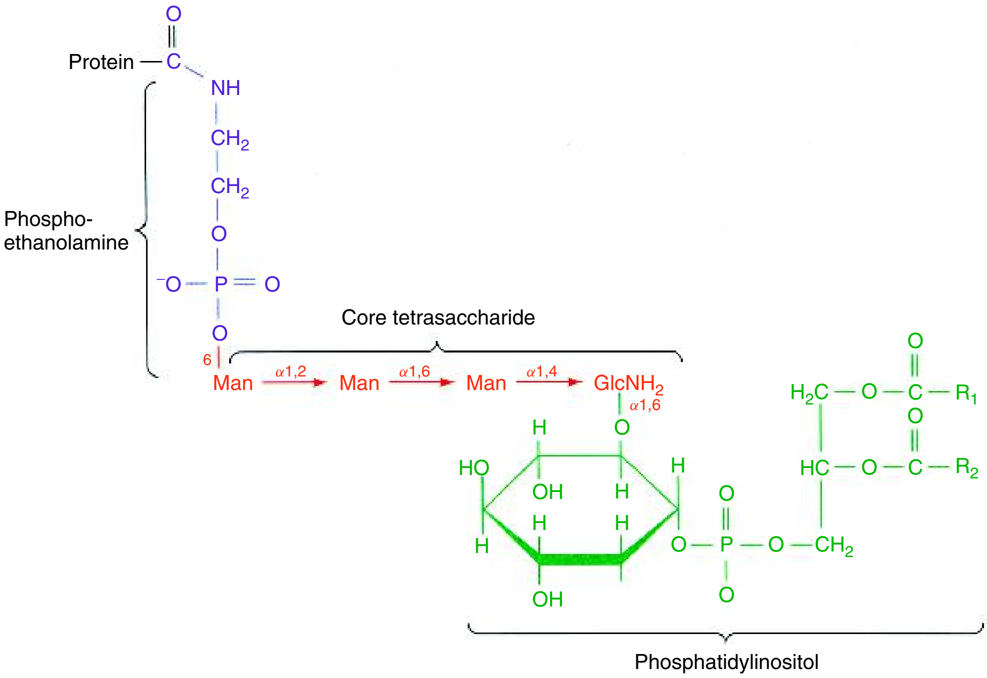


Figure 8.69 The attachment of GPI anchors to proteins together with the molecular organization of these complex oligosaccharides. GPI anchors are present in many proteins including enzymes, cell adhesion molecules, receptors and antigens. They are found in virtually all mammalian cell types and share a core structure of phosphatidylinositol glycosidically linked to non-acetylated glucosamine (GlcN). Glucosamine is usually found acetylated or sulfated form and thus the presence of non-acetylated glucosamine is an indication of a GPI anchor

the polypeptide chain is synthesized. This attachment is the prelude for further glycosylation which can involve nine mannose groups, three glucose groups and two *N*-acetyl-glucosamine groups being covalently linked to a single Asn side chain. Further processing removes some of these sugar groups in the ER and in the Golgi through the action of specific glucosidases and mannosidases, whilst in some instances fucose and sialic acid groups can be added by glucosyl transferases. These reactions lead to considerable heterogeneity and diversity in glycosylation reactions occurring within cells. However, all N-linked oligosaccharides have a common core structure (Figure 8.70).

In O-linked glycosylation the most common modification involves a disaccharide core of β -galactosyl(1 → 3)-α-*N*-acetylgalactosamine forming a covalent bond with the side chain O of Thr or Ser. Less frequently, galactose, mannose and xylose form O-linked glycosidic bonds. Whilst N-linked glycosylation proceeds through recognition of Asn-containing motifs, within primary sequences O-linked glycosylation has proved more difficult to identify from specific Ser/Thr-containing sequences and appears to depend more on tertiary structure and a surface accessible site.

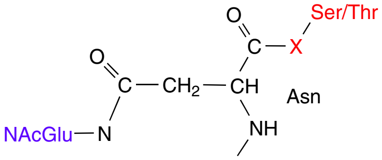


Figure 8.70 Common core structure for N-linked oligosaccharides

A single protein may contain N- and O-linked glycosylation at multiple sites and the effect of linking large numbers of oligosaccharide units is to increase molecular mass dramatically whilst also increasing the surface polarity or hydrophilicity. The function of added oligosaccharide groups has proved difficult to delineate and in some instances proteins can function perfectly well without glycosylated surfaces. Generally, many cell-surface proteins contain glycosylation sites and this has raised the possibility that oligosaccharides function in molecular recognition events although mechanistic roles have yet to be uncovered.

N- and C-terminal modifications

Post-translational modification of the N-terminal of proteins includes acetylation, myristoylation – the addition of a short fatty acid chain containing 14 carbon atoms, and the attachment of farnesyl groups. Many eukaryotic proteins are modified by acetylation of the first residue and the donor atoms are provided by acetyl-CoA and involve *N*-acetyltransferase enzymes. Acetylation is a common covalent modification at the N-terminal but other small groups are attached to the free amino group including formyl, acyl and methyl groups. The precise reasons for these modifications are unclear but may be correlated with protein degradation since chains with modified residues are frequently more resistant to turnover than unmodified proteins.

In contrast the attachment of myristoyl units has a clearly defined structural role. As a long aliphatic and hydrophobic chain with 14 carbon units it causes proteins to associate with membranes although these proteins become soluble when this ‘anchor’ is removed. Myristoylated proteins have properties typical of globular proteins and remain distinct from integral membrane proteins with which they are often associated. Important examples of myristoylated proteins include small GTPases that function in many intracellular signalling pathways.

Myristoylation is actually a co-translational modification, since the enzyme *N*-myristoyltransferase is often bound to ribosomes modifying nascent polypeptide as they emerge. The substrate is myristoyl-CoA with glycine being the preferential N-terminal amino group. When the residue after glycine contains Asn, Gln, Ser, Val or Leu myristoylation is enhanced, whilst Asp, Phe, or Tyr are inhibitory. The sequence specificity appears efficient with all proteins containing N-terminal glycine followed by a stimulatory residue observed to be myristoylated. Palmitoylation follows many of the patterns of myristoylation with the exception that a C16 unit is attached to proteins post-translationally in the cytoplasm. The reaction is catalysed by palmitoyl transferase and recognizes a sequence motif of Cys-Aliphatic-Aliphatic-Xaa where Xaa can be any amino acid. As progressively more proteins are purified it is clear that the range of modifications extents to iodination and bromination of tyrosine side chains, adenylation, methylation, sulfation, amidation and side chain decarboxylation.

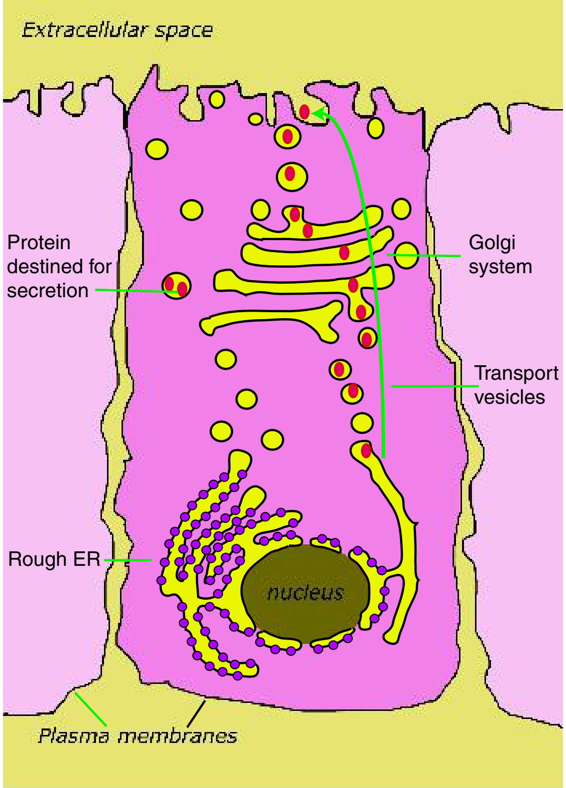


Figure 8.71 The movement of vesicles as part of the normal route for transfer of proteins from the ER to the Golgi apparatus and on to other destinations. Transfer through the Golgi occurs with protein processing before further vesicle budding from the trans Golgi region. Vesicles from the trans Golgi are either directed to the plasma membrane, for secretion or to lysosomes.

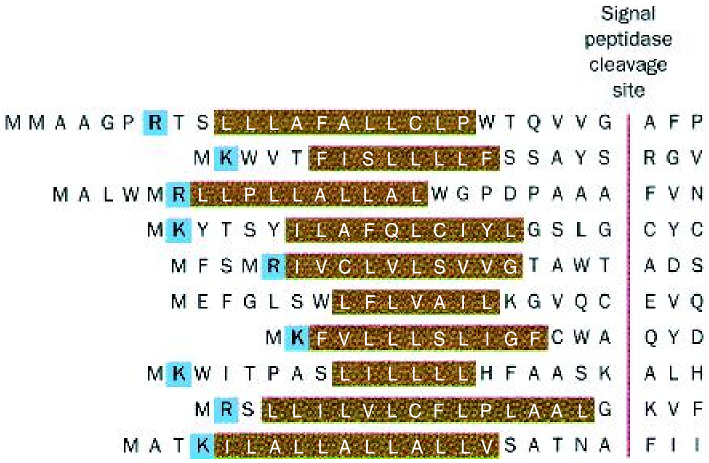

Signal peptidase cleavage site
MMAAGPRTSLLLAFALLCLPWTQVVG | AFP
MKWVTFISLLLLFSSAYS | RGV
MALWMRLLPLLALLALWGPDPAAA | FVN
MKYTSYILAFQLCIYLGSLG | CYC
MFSMRIVCLVLSVVGTAWT | ADS
MEFGLSWLFLVAILKGVQC | EVQ
MKFVLLLSLIGFCWA | QYD
MKWITPASLILLLLHFAASK | ALH
MRSLLILVLCFLPLAALG | KVF
MATKILALLALLALLVSATNA | FII

Figure 8.72 The N-terminal sequence of secretory preproteins from eukaryotes. The sequences show basic residues in blue and the hydrophobic cores in brown. The signal peptidase cleavage site is indicated

Protein sorting or targeting

The cell sorts thousands of proteins into specific locations through the use of signal sequences located normally at the N-terminal. The hypothesis that signal sequences direct proteins towards cell compartments was proposed by Gunter Blobel and David Sabatini in the early 1970s to explain the journey of secreted proteins from cytoplasmic sites of synthesis through the ER to the cell's exterior (Figure 8.71). Signal sequences direct proteins first to the ER membrane in a process that is best described as a co-translational event.

A fundamental part of sorting is the interaction of the ribosome with a macromolecular complex known as the signal recognition particle (SRP). The SRP identifies a signal peptide sequence on nascent polypeptide chains emerging from the ribosome and by association temporarily 'halts' protein synthesis. The whole ensemble remains 'halted' until the SRP–ribosome complex binds to further receptors (SRP receptors) in the ER membrane. Chain elongation resumes and the growing polypeptide is directed towards the lumen.

Proteins with signal sequences recognized by the SRP have a limited number of destinations that includes insertion of membrane proteins directly into the bilayer, transfer to the ER lumen as a soluble protein confined to this compartment, transfer via the ER lumen to the Golgi apparatus, lysosomal targeting, or secretion from the cell. Protein targeting to the nucleus, mitochondrion and chloroplast is based on a SRP-independent pathway.

The SRP-mediated pathway

The presence of additional residues at the N terminus of nascent polypeptide chains not found in the mature form of the protein provided a clue that signal sequences existed (Figure 8.72). Edman sequencing methods showed that the primary sequences of signal motifs lacked homology but shared physicochemical properties. The overall length of the sequence ranged from 10 to 40 residues and was divided into regions with different properties. A charged region between two and five residues in length occurs immediately after the N-terminal methionine residue. The basic region is followed by a series of residues that are predominantly hydrophobic such as Ala, Val, Leu, Ile, and Phe and are succeeded by a block of about five hydrophilic or polar residues.

Protein synthesis of ~80–100 residues exposes the signal peptide through the large subunit channel and results in SRP binding and the cessation of further extension until the ribosomal–mRNA–peptide–SRP complex associates with specific receptors in the

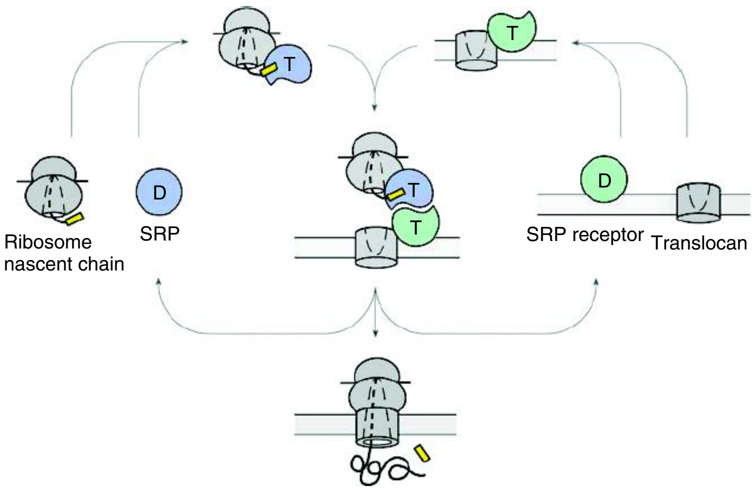


Figure 8.73 Co-translational targeting by SRP. Cycles of GTP binding and hydrolysis occur during the SRP cycle. GTPase activity is associated with SRP54 and the SRP receptor subunits. Activity is modulated by the ribosome and the translocon. The SRP complex is shown in blue and the receptor in green. SRP D reflects GDP binding, SRP T reflects a GTP-bound state. The signal sequence is shown in yellow for the nascent chain (reproduced with permission from Keenan, R.J. *et al. Ann. Rev. Biochem.* 2001, **70**, 755–75. Annual Reviews Inc.)

ER membrane. This is known as co-translational targeting (Figure 8.73) and is conveniently divided into two processes; recognition of a signal sequence and association with the target membrane.

The structure and organization of the SRP

SRPs are found in all three major kingdoms (archae, eubacteria and eukaryotes) with the mammalian SRP consisting of a 7S rRNA and six distinct polypeptides called SRP9, SRP14, SRP19, SRP54, SRP68 and SRP72 and named according to their respective molecular masses. With the exception of SRP72 all bind to the 7S rRNA and SRP54 has a critical role in binding RNA *and* signal peptide. The diversity of signal sequences plays a significant part in determining the mode of binding since introduction of a single charged residue destroys interaction with SRP54, whilst a block of hydrophobic residues remains critical to association. In SRP54 the M domain is important in molecular recognition of signal peptide. The name derives from a high percentage of methionine residues and is conserved in bacterial and mammalian homologues The homologous *E. coli* protein Ffh (Ffh = f̲ifty f̲our h̲omologue) has ∼16 percent of residues in the M domain as methionines – a frequency six times greater than its typical occurrence in proteins.

The *E coli* SRP has a simpler composition – the Ffh protein and a 4.5S RNA – and evolutionary conservation is emphasized by the ability of human SRP54 to bind 4.5S RNA. Bacterial SRPs are minimal homologues of the eukaryotic complex and are attractive systems for structural studies. The crystal structure derived for Ffh from *T. aquaticus* (Figure 8.74) revealed a C-terminal helical M domain forming a long hydrophobic groove that is exposed to solvent and is made from three helices together with a long flexible region known as the finger loop linking the first two helices. The dimensions of the groove are compatible with signal sequence binding in a helical conformation and, as predicted, the surface of the groove is lined entirely with phylogenetically conserved hydrophobic residues.[1]

Homologous domains from *E. coli* and mammalian SRP54 have similar structures (Figure 8.75),

[1] In *T. aquaticus*, a thermophile living above 70 °C, many of the conserved Met residues found in SRP54/Ffh are replaced by Leu, Val and Phe, possibly reflecting a need for increased flexibility to counter increased thermal mobility.

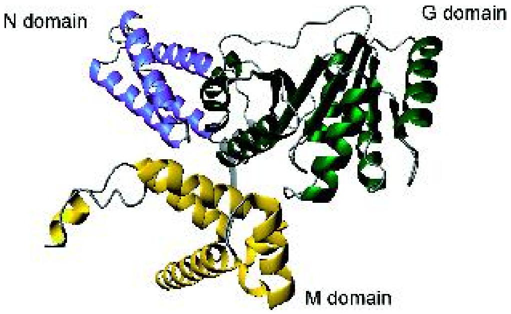


Figure 8.74 The structure of the full length FfH subunit from *Thermus aquaticus*. The M domain encompassing residues 319–418 is shown in yellow, the N domain, residues 1–86, in blue and the largest G domain, residues 87–307, in green (PDB: 2FFH)

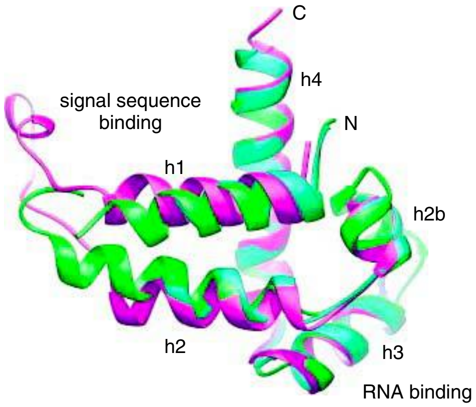


Figure 8.75 Superposition of *E. coli* (light blue), *T. aquaticus* (magenta), and human (green) M domains showing similar structures (reproduced with permission from Batey, R.T. *et al. J. Mol. Biol.* 2001, **307**, 229–246. Academic Press)

although the finger loop region adopts different conformations – a variation that reflects intrinsic flexibility. Additional N-terminal and GTPase domains interact with the SRP receptor.

SRP54 is a GTPase and the catalytic cycle involves GTP/GDP exchange and GTP hydrolysis although signal sequence binding does not require GTP hydrolysis. Ribosome association stimulates GTP binding by SRP54 and is followed by hydrolysis upon interaction with the membrane receptors. The SRP receptor contains two subunits designated α and β with the α subunit showing homology to the GTPase domain of SRP54 and interacting with the β subunit, a second GTPase, that shows less similarity to either SRP54 or the α subunit.

The SRP receptor structure and function

Mammalian SRP receptors are membrane protein complexes but the corresponding bacterial receptors are simpler containing a single polypeptide known as FtsY (Figure 8.76). FtsY is a GTPase loosely associated with the bacterial inner membrane that can substitute for the α subunit of the SRP receptor. FtsY is composed of N and G domains comprising the GTPase catalytic unit and adopting a classic GTPase fold based on four conserved motifs (I–IV) arranged around a nucleotide-binding site. Motif II is contained within a unique insertion known as the insertion box domain and this extends the central β sheet of the domain by two strands and is characteristic for the SRP GTPase subfamily.

The N and G domains of SRP54 (Ffh) interact with the α subunit of the SRP receptor (FtsY) through their respective, yet homologous, NG domains. The interaction occurs when each protein binds GTP and the interaction is weaker in the apo protein. Nucleotide dependent changes in conformation are the basis for the association between SRP and its cognate receptor and GTP hydrolysis causes SRP to dissociate from the receptor complex. SRP lacking an NG domain fails to target ribosome–nascent chain complexes to the membrane and conversely the NG domain of FtsY when fused to unrelated membrane proteins promotes association. The primary role of the SRP receptor is to ‘shuttle’ the ribosomal-mRNA–nascent chain from a complex with SRP to a new interaction with a membrane-bound complex known as the translocon. The translocon is a collection of three integral membrane proteins, Sec61α, Sec61β and Sec61γ, that form pores within the membrane allowing the passage of nascent polypeptide chains through to the ER lumen.

Signal peptidases have a critical role in protein targeting

After translocation across the ER membrane polypeptides reach the lumen where a peptidase removes the

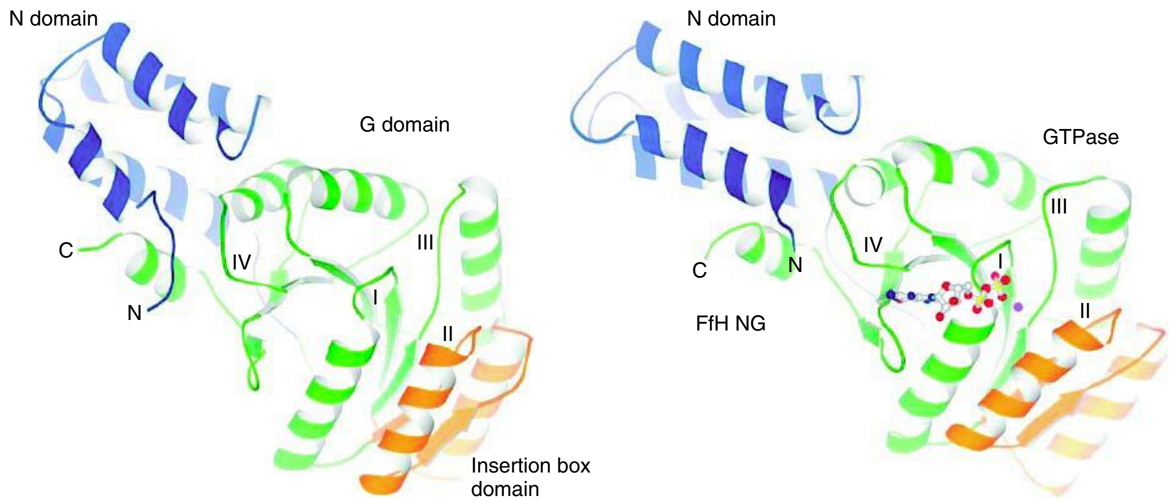


Figure 8.76 The structures of *E. coli* FtsY show homology to Ffh. The apo-form of the NG domain shows the N-terminal N domain (blue) packing tightly against the GTPase fold (green). The conserved insertion box domain (orange) is unique to the SRP family of GTPases. The four conserved GTPase sequence motifs are indicated (I–IV). (Reproduced with permission from Keenan, R.J. *et al. Ann. Rev. Biochem.* 2001, **70**, 755–75. Annual Reviews Inc.) The structure of Ffh is show for comparison in similar colours

'signal peptide' by cleavage at specific sites. Sites of cleavage are determined by local sequence composition after the 'polar' portion of the signal peptide. The ER signal peptidase shows a preference for residues with small side chains at positions −1 and −3 from the cleavage site. Consequently Gly, Ala, Ser, Thr and Cys are common residues whilst aromatic, basic or large side chains at the −3 position inhibits cleavage. Alanine is the most common residue found at the −1 and −3 positions and this has given rise to the Ala-X-Ala rule or −1, −3 rule for cleavage site identification.

Post-lumenal targeting is directed by additional sequence-dependent signals

When proteins reach the lumen other sequence dependent signals dictate targeting to additional sites or organelles. One signal pathway involves the KDEL sequence named after the order of residues found at the C terminal of many soluble ER proteins. In mammals proteins with the sequence Lys-Asp-Glu-Leu (i.e. KDEL) are marked for recovery from transport pathways. One example is PDI and altering this sequence causes the protein to be permanently secreted from the ER. The KDEL sequence binds to specific receptors located in the ER membrane and in small secretory vesicles that bud off from the ER en route for the Golgi. The regulation of this receptor is puzzling but it represents a very efficient mechanism for restraining and recovering ER proteins.

An interesting variation of this pattern of transport occurs in channel proteins such as the K/ATPase complex consisting of two different subunits each containing a KDEL motif. Individually these subunits are prevented from leaving the ER but their assembly into a functional channel protein leads to the occlusion of the KDEL motif from the receptor. Thus once assembled these subunits are exported from the ER and this represents an effective form of 'quality control' ensuring that only functional complexes enter the transport pathway.

After synthesis, partially processed proteins destined for the plasma membrane, lysosomes or secretion are observed in the Golgi apparatus – a series of flattened membrane sacs. Within the Golgi apparatus progressive processing of proteins is observed in the form of glycosylation and once completed proteins are sent to their final destination through the use of clathrin coated vesicles.

Protein targeting to the mitochondrion and chloroplast

Mitochondria and chloroplasts have small genomes and most proteins are nuclear coded. This requires that the majority of chloroplast and mitochondrial proteins are transferred to the organelle from cytoplasmic sites of synthesis. Both organelles have outer and inner membranes and targeting involves traversing several bilayers. The import of proteins by mitochondria and chloroplasts is similar to that operating in the ER with N-terminal sequences directing the protein to organelles where specific translocation pathways operate to assist movement across membranes. The mitochondria and chloroplast present unique targeting systems in view of the number of potential locations. For the mitochondrion this includes locations in the outer or inner membranes, a location in the inter-membrane space, or a location in the matrix. Unsurprisingly the number of potential locations even within a single organelle places stringent demands on the 'signal' directing nuclear coded proteins to specific sites.

Mitochondrial targeting

In the mitochondrion translocation of precursor proteins is an energy-dependent process requiring ATP utilization and complexes in the inner and outer membranes. For small proteins (<10 kDa) the outer membrane has pores that may allow entry but in most cases the targeting information for precursor proteins resides in an N-terminal sequence extension. Mitochondrial signal sequences vary in length and composition although they have a high content of basic residues and residues with hydroxyl side chains but lack acidic side chains. An important property of these signal sequences is the capacity to form amphipathic helices in solution and this is probably related to the requirement to interact with membranes, receptors *and* aqueous environments. The majority of mitochondrial proteins are synthesized in the cytosol and chaperones play a role in maintaining newly synthesized protein in an 'import competent' state that is not necessarily the native fold but a form that can be translocated across membranes.

Table 8.9 Proteins identified as components of the yeast outer membrane translocase system

Protein	Proposed function
Tom5	Component of GIP, transfer of preprotein from Tom20/22 to GIP
Tom6	Assembly of the GIP complex
Tom7	Dissociation of GIP complex
Tom20	Preprotein receptor, preference for presequence-containing preproteins
Tom22	Preprotein receptor, cooperation with Tom20, part of GIP complex
Tom37	Cooperation with Tom machinery (Tom70)
Tom40	Formation of outer membrane translocation channel (GIP)
Tom70	Preprotein receptor, preference for hydrophobic or membrane preproteins
Tom72	None (homologue of Tom70)

Adapted from Voos, W. *et al. Biochim. Biophys. Acta* 1999, **1422**, 235–254.

Translocation involves specific complexes in the membrane given the abbreviations Tom (translocation outer membrane) and Tim (translocation inner membrane). The first step of translocation involves precursor protein binding to a translocase system consisting of at least nine integral membrane proteins (see Table 8.9). Tom20, Tom22 and Tom70 are the principal proteins involved in import. Tom20 has a single transmembrane domain at the N-terminus with a large cytosolic domain recognizing precursor proteins from their targeting signals. Tom20 cooperates with Tom22 a protein with an exposed N terminal domain that also binds targeting sequences. This may indicate autonomous rather than cooperative function but import is driven by a combination of hydrophobic and electrostatic interactions between pre-sequence and receptor. Tom70, the other major receptor has a large cytosolic domain preferentially interacting with preproteins carrying 'internal' targeting information. Specific recognition by Tom 20, 22, and 70 leads to

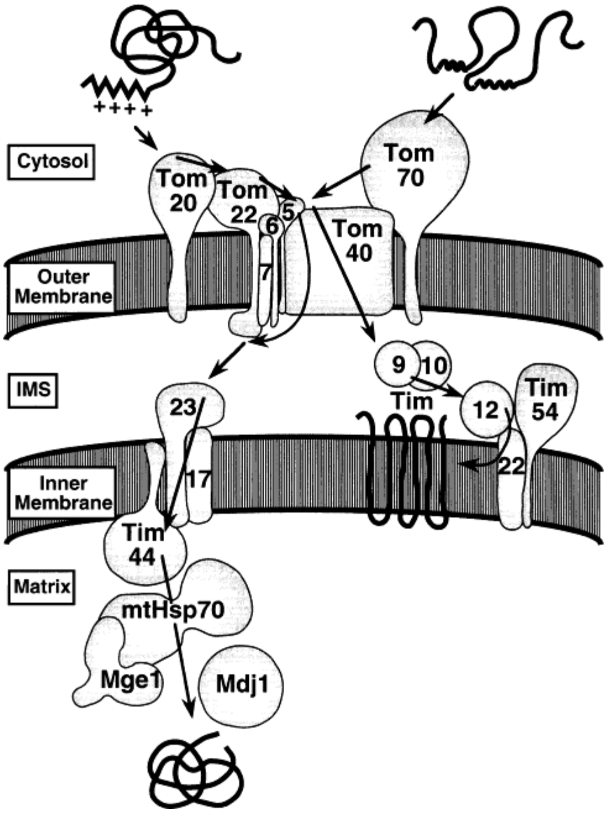


Figure 8.77 Schematic model of the mitochondrial protein import machinery of *S. cerevisiae*. (Reproduced with permission from Voos, W. *et al. Biochim. Biophys. Acta* 1999, **1422**, 235–254. Elsevier)

Table 8.10 Proteins identified as components of the yeast inner mitochondrial membrane translocase system

Protein	Possible function
Tim8	Cooperation with Tim9-Tim10
Tim9	Complex with Tim10, guides hydrophobic carrier proteins through inter membrane space.
Tim10	Complex with Tim9, guides hydrophobic carrier proteins through inter membrane space.
Tim12	Complex with Tim22 and Tim54, inner membrane insertion of carrier proteins.
Tim13	Cooperation with Tim9-Tim10
Tim17	Complex with Tim23, translocation of preproteins through inner membrane into matrix
Tim22	Inner membrane insertion of carrier proteins
Tim23	Complex with Tim17, translocation of preproteins through inner membrane into matrix
Tim44	Membrane anchor for mitochondrial Hsp70
Tim54	Complex with Tim22, inner membrane insertion of carrier proteins
Tim11	Unclear role.

Adapted from Voos, W. *et al*. *Biochim. Biophys. Acta* 1999, **1422**, 235–254.

a second step involving pre-sequence insertion into the Tom40 hydrophilic channel.

The inner mitochondrial membrane import pathway (Figure 8.77) completes the sequence of events started by the Tom assembly. The pathway is composed of Tim proteins (Table 8.10) with an inner membrane import channel formed by Tim23, Tim17 and Tim44. Tim23 is a membrane protein containing a small exposed acidic domain that may act as a potential binding site for sequences translocated through the outer membrane. Tim23 is hydrophobic and with Tim17 forms the central channel of the inner membrane import pathway. Tim44 is a peripheral protein that interacts with mitochondrial chaperones (Hsp70) to regulate protein folding and prevent aggregation. Since proteins are translocated in an extended conformation the role of chaperones is important to mediate controlled protein folding in the mitochondrial matrix.

Closely coupled to translocation across the inner membrane is removal of the N-terminal signal sequence by mitochondrial processing proteases (MPP) located in the matrix. An interesting, but unexplained, observation, and one that will be recalled by the careful reader, is that MPPs constituted the 'core' proteins of complex III found in mitochondrial respiratory chains.

The import pathway will target proteins into the matrix of mitochondria but to reach other destinations the basic import pathway is supplemented by additional sorting reactions.

The Rieske FeS protein is synthesized with an N-terminal matrix targeting sequence that is cleaved by the normal processing peptidase and *in vitro* this protein is detected in the matrix, a foreign location for this protein. *In vivo* additional sorting mechanisms target the protein back to the inner membrane. In contrast cytochrome c_1 has a dual targeting sequence. A typical *matrix* targeting sequence is supplemented by a second sorting signal that has a hydrophobic core and resembles bacterial export signals. For proteins in the inter-membrane space an incredibly diverse array of pathways exist. Cytochrome c is targeted to this location without obvious targeting sequences whilst yeast flavocytochrome b_2 (lactate dehydrogenase) is located in this compartment by a dual targeting sequence.

Targeting of proteins to the chloroplast

The chloroplast (Figure 8.78) uses similar principles of sorting and targeting to the mitochondrion. Targeting sequences contain significant numbers of basic residues, a high content of serine and threonine residues and are highly variable in length ranging from 20 to 120 residues. In direct analogy to their mitochondrial counterparts the translocating channel of the outer and inner chloroplast membranes are given the nomenclature Toc and Tic.

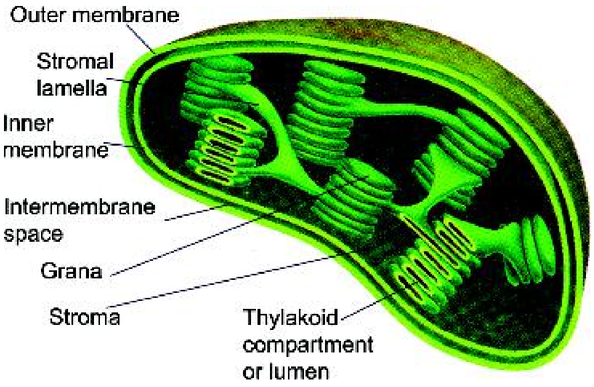


Figure 8.78 A schematic diagram of a chloroplast

Although *in vitro* the Toc and Tic translocation systems are separable the two complexes coordinate activities *in vivo* to direct proteins from the cytosol to the stroma. Two proteins, Toc159 and Toc75, act as receptors and form a conducting channel with a diameter of 0.8–0.9 nm that requires proteins to traverse outer membranes in an extended conformation. The Tic proteins involved in precursor import are Tic110, Tic55, Tic40, Tic22 and Tic20, but little is known about the mechanism of import or their structural features.

Stromal proteins such as ribulose bisphosphate carboxylase, ferredoxin or ferredoxin-NADP reductase do not require additional sorting pathways. They are synthesized with a single signal sequence that is cleaved by a stromal processing peptidase. Thylakoid proteins such as the light-harvesting chlorophyll complexes, reaction centres, and soluble proteins such as plastocyanin must utilize additional targeting pathways. Targeting is via a bipartite N-terminal signal sequence where the first portion ensures transfer into the stromal compartment whilst a second part directs the peptide to the thylakoid membrane where further proteolysis in the lumen removes the second signal sequence.

Nuclear targeting and nucleocytoplasmic transport

The final targeting system involves nuclear import and export. Nuclear proteins are required for unpacking, replication, synthesis and transcription of DNA as well as forming the structure of the nuclear envelope. There are no ribosomes in the nucleus so all proteins are imported from cytosolic sites of synthesis. This creates obvious problems with the nucleus being surrounded by an extensive double membrane. Proteins are imported through a series of large pores that perforate the nuclear membrane and result from a collection of proteins assembled into significant macromolecular complexes.

Specific nuclear import pathways were demonstrated with nucleoplasmin, a large pentameric protein complex ($M_r \sim 165$ kDa), which accumulates in the soluble phase of frog oocyte nuclei. Extraction of the protein followed by introduction into the cytosol of oocytes led to rapid accumulation in the nucleus. The rate of accumulation was far greater than expected on the basis of diffusion and involved active transport mechanisms and sequence dependent receptors. The sequence specific import was demonstrated by removing the tail regions from nucleoplasmin pentamers. This region contained a 'signal' directing nuclear import and in tail-less protein no accumulation within the nucleus was observed.

Nuclear proteins have import signals defined by short sequences of 4–10 residues that lack homology but show a preference for lysine, arginine and proline. Nuclear localization signals (NLS) are sensitive to mutation and linking sequences to cytoplasmic proteins directs import into the nucleus. Today it is possible to use computers to 'hunt' for NLS within primary sequences and their identification may indicate a potential role for a protein as well as a nuclear localization (Table 8.11). More complex bipartite sequences occur in proteins where short motifs of two to three basic residues are separated by a linker region of 10–12 residues from another basic segment.

Nucleoplasmin-coated gold particles defined the route of protein entry into the nucleus when electron-dense particles were detected by microscopy around pore complexes shortly after injection into the cytosol of oocytes. These observations suggested specific association between cytoplasmic-facing proteins of the nuclear pore complex and NLS containing protein. Nuclear transport is bi-directional (Figure 8.79) with RNA transcribed in the nucleus exported to the cytoplasm as ribonucleoprotein whilst the disassembly of the nuclear envelope during mitosis requires the re-import of all proteins. Many proteins shuttle continuously between the nucleus and cytoplasm and the average cell has an enormous level of nucleocytoplasmic traffic – some estimates suggest that over 10^3 macromolecules are transferred between the two compartments every second in a growing mammalian cell.

Table 8.11 Sequences identified to act as signals for localization to the cell nucleus

Protein	NLS sequence
SV40 T antigen	Pro Lys Lys Lys Arg Lys Val
Adenovirus E1a	Lys Arg Pro Arg Pro
Nucleoplasmin	Lys Arg Pro Ala Ala Thr Lys Lys Ala Gly Gln Ala Lys Lys Lys Lys

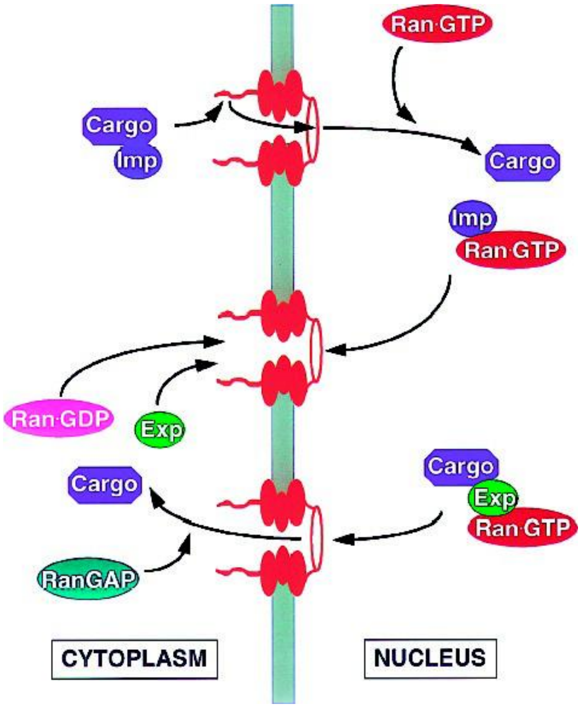


Figure 8.79 A schematic representation of nuclear import involving importins, cargo and Ran-GTP (reproduced with permission from Cullen, B.R. *Mol. Cell. Biol.* 2000, **20**, 4181–4187)

The first proteins discovered with a role in shuttling proteins from the cytosol to nucleus were identified from their ability to bind to proteins containing the NLS. The protein was importin-α (Figure 8.80) and homologues were cloned from other eukaryotes to define a family of importin-α-like proteins. Importin-α functions as a heterodimer with a second protein called importin-β. A typical scenario for nuclear transport involves cargo recognition via the NLS by importin-α and complex formation with importin-β followed by association and translocation through the nuclear pore complex in a reaction requiring GTP. In the

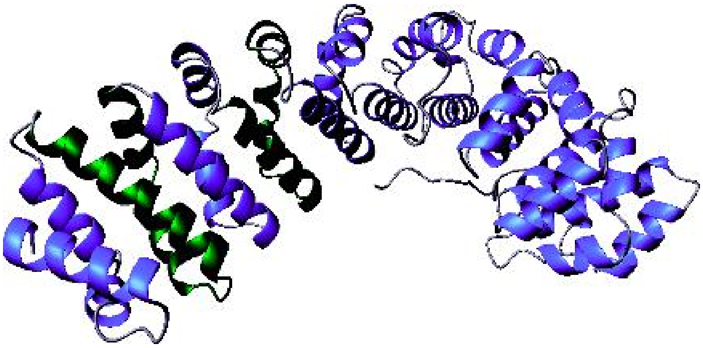


Figure 8.80 The super-helix formed by collections of armadillo motifs in mouse importin-α. A unit of three helices is shown in green (PDB: 1AIL)

nucleus the complex dissociates in reactions catalysed by monomeric G proteins such as Ran-GTP. Ran-GTP binds specifically to importin-β allowing it to be recycled back to the cytoplasm in events again linked to GTP hydrolysis. Importin-α and the cargo dissociate with the protein 'delivered' to its correct destination.

The structure of importin-α, determined in a complex with and without NLS sequence, reveals two domains within a polypeptide of ~60 kDa. A basic N-terminal domain binds importin-β and a much larger second domain composed of repeating structural units known as armadillo repeats (Figure 8.80). Each unit is composed of three α helices and the combined effect of each motif is a translation of ~0.8–1.0 nm and a rotation of ~30°, forming a right-handed super-helical structure.

Importin-β binds to the N-terminal of importin-α but also interacts with the nuclear pore complex and Ran GTPases. The interaction with Ran proteins is located at its N-terminal whilst interaction with importin-α is centred around the C-terminal region. Importin-β (Figure 8.81) also contains repeated structural elements – the HEAT motif (the proteins in which the motif was first identified are Huntingtin protein–Elongation factor (EF3)–A subunit of protein phosphatase and the TOR protein). The motif possesses a pattern of hydrophobic and hydrophilic residues and is predicted to form two helical elements of secondary structure. In a complex with the N-terminal fragment (residues 11 to 54) of the α subunit the HEAT motifs arrange in a convoluted snail-like appearance where the

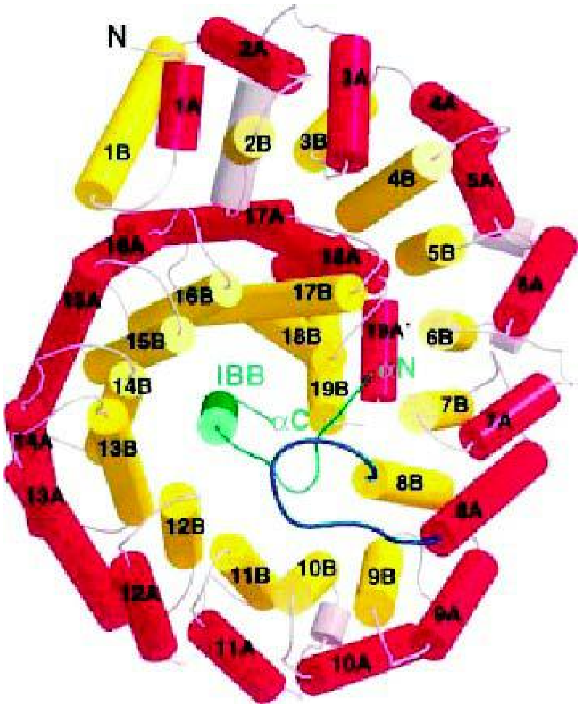


Figure 8.81 Structure of importin-β plus IBB domain from a view down the superhelical axis. A and B helices of each HEAT motif are shown in red and yellow respectively. The acidic loop with the DDDDDW motif in HR8 is shown in blue. (Reproduced with permission from Cingolani, G. *et al. Nature* 1999, **399**, 221–229. Macmillan)

extended IBB domain is located at the centre of the protein. Importin-β is composed of 19 HEAT repeats forming a right-handed super-helix where each HEAT motif, composed of A and B helices connected by a short loop, arranges into an outer layer of 'A' helices defining a convex surface and an inner concave layer of 'B' helices. The HEAT repeats vary in length from 32 to 61 residues with loop regions containing as many as 19 residues. HEAT repeats 7–19 bind importin-α with the N-terminal fragment of importin-α bound on the inner surface forming an extended chain from residues 11–23 together with an ordered helix from residues 24–51. The axis of this helix is approximately coincident to that of importin-β super-helix. The N-terminal region of the importin-α interacts specifically with a long acidic loop linking helices 8A and 8B that contains five aspartate residues followed by a conserved tryptophan residue. HEAT repeats 1–6 are implicated in the interaction with Ran.

The third protein with a critical role in nuclear–cyto plasmic transport is Ran – a GTPase essential for nuclear transport. Ran is a member of the Ras-like family of GTP binding proteins and switches between active GTP bound forms and an inactive state with GDP (Figure 8.82). Ran is similar to other monomeric GTPases but it also possesses a long C-terminal extension that is critical for nuclear transport function and is terminated by a sequence of acidic residues (DEDDDL).

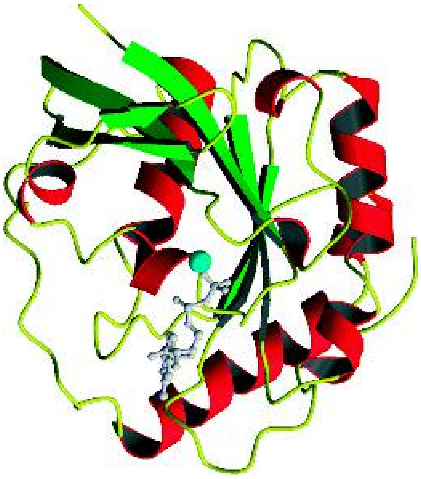


Figure 8.82 The structure of Ran-GDP

Ran consists of a six-stranded β sheet surrounded by five α helices. GDP and GTP are bound via the conserved NKXD motif with Lys123 and Asp125 interacting directly with the base, the ribose portion exposed to the solvent and the phosphates binding via a P loop motif and a large number of polar interactions with the sequence GDGGTGKT. The loop region acts as a switch through nucleotide-exchange induced conformational changes and dictates interactions with other import proteins such as importin-β.

Ran itself is 'regulated' by a cytosolic GTPase activating protein termed Ran GAP1 that causes GTP hydrolysis and inactivation. In the nucleus a chromatin-bound nucleotide exchange factor called RCC1 functions to exchange GDP/GTP. The interplay of factors controlling nucleotide exchange and hydrolysis generates gradients of Ran-GTP that direct nuclear–cytoplasmic transport.

Ran binds to the N terminal region of importin-β, but importantly the C-terminal of Ran is able to associate with other proteins largely through charged residues in the tail (DEDDDL). These charges mediate Ran-GTP/importin-β complexes interaction with Ran binding proteins in the nuclear pore complex but also offer a mechanism of dissociating complexes in the nucleus. Gradually many of the molecular details of protein import via nucleocytoplasmic transport machinery have become clear with structures for the above proteins and complexes formed between Ran and RCC1 as well as the ternary complex of Ran–RanBP1-RanGAP. Ran, importin-α and -β subunits are exported back to the cytosol for future rounds of nucleocytoplasmic transport.

Proteins identify themselves to transport machinery with two signals. In addition to the positively charged NLS some proteins have nuclear export signals (NES) based around leucine-rich domains.

The nuclear pore assembly

In higher eukaryotes the nuclear pore complex is a massive macromolecular assembly containing nearly 100 different proteins with a combined mass in excess of 120 MDa. In yeast this assembly is simpler with approximately 30 different polypeptides forming the

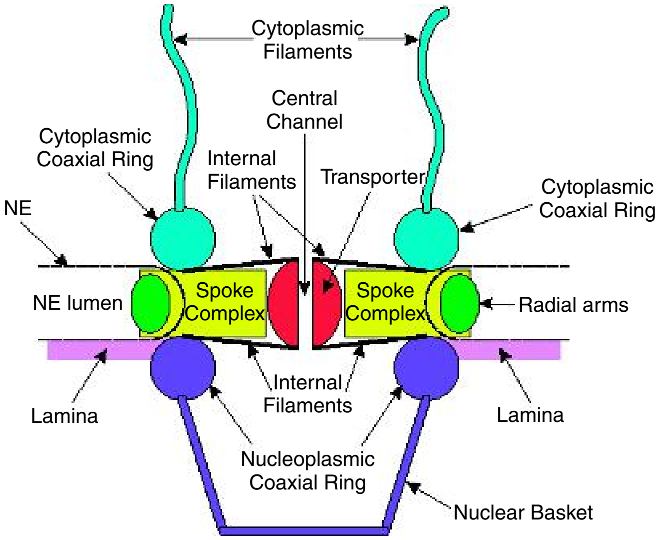


Figure 8.83 A cutaway representation of the nuclear pore complex (reproduced with permission from Allen, T.D. *et al. J. Cell Sci.* 2000, **113**, 1651–1659. Company of Biologists Ltd)

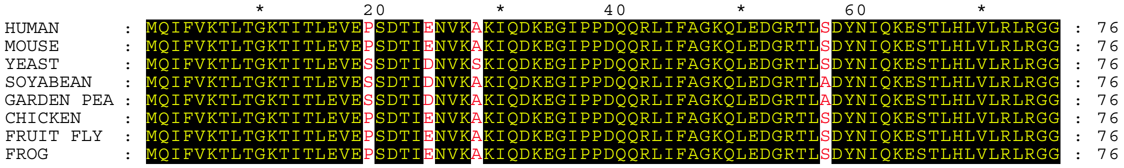
complex. The increased size of the vertebrate pore probably reflects complexity associated with metazoan evolution, but the basic structural similarities between the pore complexes of yeast and higher eukaryotes indicates a shared mechanism of translocation that can be discussed within a single framework.

The proteins of the nuclear pore complex are called nucleoporins or ‘nups’. As the role of Ran, importins and other import proteins were uncovered over the last decade emphasis has switched to the structure and organization of the nuclear pore complex. Electron microscopy reveals the pore complex as a symmetrical assembly containing a pore at the centre that presents a barrier to most proteins. The nucleoporin complex (NPC) spans the dual membrane of the nuclear envelope and is the universal gateway for macromolecular traffic between the cytoplasm and the nucleus. The basic framework of the NPC consists of a central core with a spoke structure (Figure 8.83). From this central ring long fibrils 50–100 nm in length extend into the nucleoplasm and the cytoplasm whilst the whole NPC is anchored within the envelope by the nuclear lamina.

Clarification of nuclear pore organization came from studies of the yeast complex where genetic malleability coupled with genome sequencing revealed 30 nucleoporins that are conserved across phyla and thus allowed identification of homologues in metazoan eukaryotes. Once identified nups were cloned and localized within the envelope using a combination of immunocytochemistry, electron microscopy and cross-linking studies. The yeast nuclear pore complex has been redefined in terms of nucleoporin distribution. Nup358 is equated with the RanBP2 protein described previously whilst p62, p58, p54 and p45 are proteins found at the centre of the nuclear pore complex and gp210 and POM121 form all or part of the spoke assembly.

Protein turnover

Within the cell sophisticated targeting systems direct newly translated proteins to their correct destinations. However, proteins have finite lifetimes and normal cell function requires specialized pathways for degradation and recycling. The half-life of most proteins is measured in minutes although some, such as actin or haemoglobin, are more stable with half-lives in excess of 50 days. At some point all proteins ‘age’ as a result of limited proteolysis, covalent modification or non-enzymatic reactions leading to a decline in activity.



```
                      *        20         *        40         *        60         *
HUMAN      : MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG : 76
MOUSE      : MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG : 76
YEAST      : MQIFVKTLTGKTITLEVESSDTIDNVKSKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG : 76
SOYABEAN   : MQIFVKTLTGKTITLEVESSDTIDNVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLADYNIQKESTLHLVLRLRGG : 76
GARDEN PEA : MQIFVKTLTGKTITLEVESSDTIDNVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLADYNIQKESTLHLVLRLRGG : 76
CHICKEN    : MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG : 76
FRUIT FLY  : MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG : 76
FROG       : MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG : 76
```

Figure 8.84 The primary sequence of ubiquitin from plants, animal and microorganisms shows only four changes in sequence at positions 19, 24, 28 and 57

These proteins are degraded into constituent amino acids and re-cycled for further synthetic reactions. The proteasome is a large multimeric complex designed specifically for controlled proteolysis found in all cells. In eukaryotes the ubiquitin system contributes to this pathway by identifying proteins destined for degradation whilst organelles such as the lysosome perform ATP-independent protein turnover.

The ubiquitin system and protein degradation

In eukaryotes ubiquitin (Figure 8.84) occupies a pivotal role in the pathway of protein degradation. It is a highly conserved protein found in all eukaryotic organisms but never in either eubacteria or archae. The level of homology exhibited by ubiquitin is high with only 3 differences in primary sequence between yeast and human protein. The lack of evolutionary divergence points to functional importance and suggests that most residues in ubiquitin have essential roles.

In view of its importance to protein degradation the structure of ubiquitin (Figure 8.85) is unspectacular containing five β strands together with a single α helix. The final three C-terminal residues are Arg-Gly-Gly and extend away from the globular domain into bulk solvent. This arrangement is critical to biological function as the C-terminal glycine forms a peptide bond with the ε-amino group of lysine in target proteins. Covalent modification by adding ubiquitin in a process analogous to phosphorylation has led to the terminology ubiquitination (also called ubiquitinylation). Multiple copies of ubiquitin attach to target proteins acting as 'signals' for degradation by the proteasome.

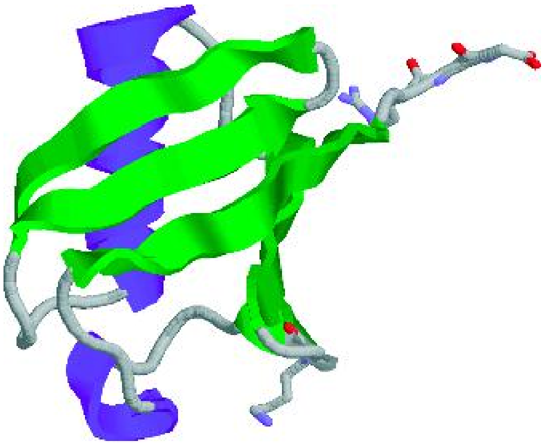


Figure 8.85 The structure of ubiquitin showing four significant β strands (1–7,10–17, 40–45, 64–72) and a single α helix (23–34) whilst shorter strand and helix regions exists from 48–50 and 56–59, respectively. The residues Arg 74, Gly75 and Gly76 are shown along with Lys48

Many enzymes catalyse ubiquitin addition to proteins and include E1 enzymes or ubiquitin-activating enzymes, E2 enzymes, known as ubiquitin-conjugating enzymes, and E3 enzymes, which act as ligases. Ubiquitin is activated by E1 in a reaction hydrolysing ATP and forming ubiquitin adenylate. The activated ubiquitin binds to a cysteine residue in the active site of E1 forming a thiol ester. E2, the proximal donor of ubiquitin to target proteins, transfers ubiquitin to the acceptor lysine forming a peptide bond, although E3 enzymes also participate by forming complexes with target protein and ubiquitin-loaded E2. Figure 8.86 shows the ubiquitin-mediated degradation of cytosolic proteins.

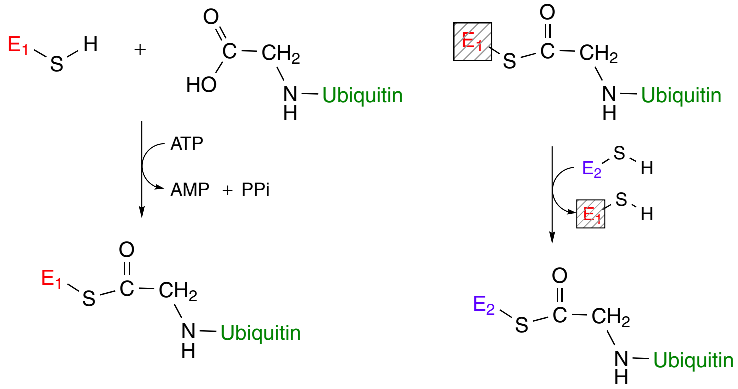


Figure 8.86 The first step in the ubiquitin-mediated degradation of cytosolic proteins is ATP dependent activation of E1 to form a thiol ester derivative with the C-terminal glycine of ubiquitin. It is followed by transfer of ubiquitin from E1 to the E2 enzyme, the proximal donor of ubiquitin to proteins

Multiple ubiquitination is common and involves peptide bond formation between the carboxyl group of Gly76 and the ε-amino group of Lys48 on a second ubiquitin molecule. Substitution of Lys48 with Cys results in a protein that does not support further ubiquitination and is incapable of targeting proteins for proteolysis. Four copies of ubiquitin are required to target proteins efficiently to the 26*S* proteasome with 'Ub_4' units exhibiting structural characteristics that enhance recognition. Ubiquitin does not degrade proteins but tags proteins for future degradation. A view of ubiquitin as a simple tag may be an over-simplification since a role for ubiquitin enhancing association between proteins and proteasome is implied by the observation that without ubiquitin proteins interact with the proteasome but quickly dissociate.

Alternative strategies exist for degrading proteins within cells. One strategy used by cells is to inactivate proteins by oxidizing susceptible side chains such as arginine, lysine and proline. This event is enough to 'mark' proteins for degradation by cytosolic proteases. A second method of protein turnover is identified with sequence information and 'short-lived' proteins contain sequences rich in proline, glutamate, serine and threonine. From the single letter code for these residues the sequences have become known as PEST sequences. Comparatively few long-lived proteins contain PEST-rich regions and the introduction of these residues into stable proteins increases turnover although the structural basis for increased ability is unclear along with the relationship to the ubiquitin pathway. A third mechanism of controlling degradation lies in the composition of N-terminal residues. A correlation between the half-life of a protein and the identity of the N-terminal residue gave rise to the 'N-end' rule. Semi-quantitative predictions of protein lifetime from the identity of the N-terminal residue suggests that proteins with Ser possess half-lives of 20 hours or greater whilst proteins with Asp have on average half-lives of ~3 min. The mechanism that couples recognition of the N-terminal residue and protein turnover is unknown but these correlations are also observed in bacterial systems lacking the ubiquitin pathway of degradation.

The proteasome

Intracellular proteolysis occurs via two pathways: a lysosomal pathway and a non-lysosomal, ATP-dependent, pathway. The latter pathway degrades most cell proteins and involves the proteasome first identified as an endopeptidase with multicatalytic activities from bovine pituitary cells. The multicatalytic protease was called the proteasome, reflecting its complex structure and proteolytic role.

The proteosome degrades proteins via peptide bond scission, and multiple catalytic functions are seen via

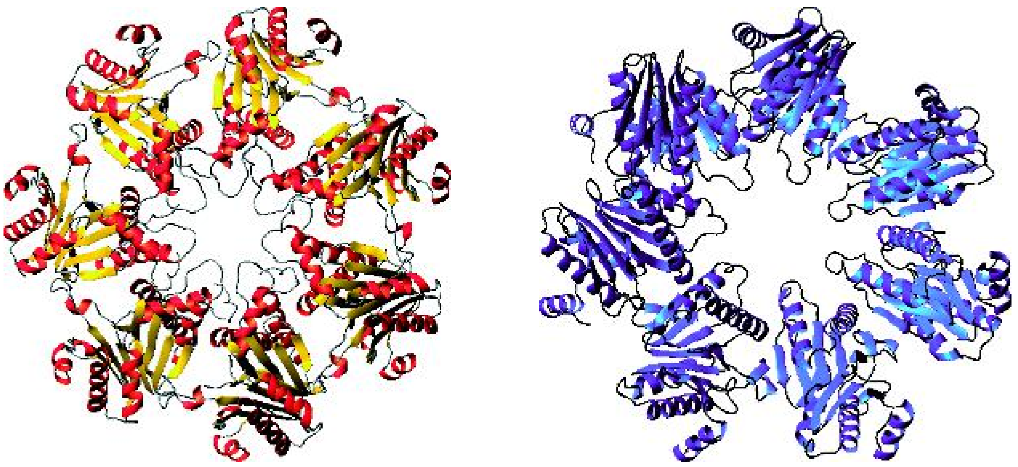


Figure 8.87 Superior views of α and β heptameric rings showing central cavity

the hydrolysis of many synthetic and natural substrates. The eukaryotic proteasome has activities described by terms such as ‘chymotryptic-like’ activity (preference for tyrosine or phenylalanine at the P1 position), ‘trypsin-like’ activity (preference for arginine or lysine at the P1 position) and ‘post-glutamyl’ hydrolysing activity (preference for glutamate or other acidic residues at the P1 position). The variety of catalytic activities created a confusing functional picture since proteasomes cleave bonds after hydrophobic, basic and acidic residues. In the proteasome the myriad of catalytic activities suggested a unique site. The proteasome was first identified in eukaryotes but comparable complexes occur in prokaryotes where the system can be deleted without inducing lethality. As usual, eukaryotic proteasomes have complex structures whilst those from prokaryotes are simpler.

The hyperthermophile *Thermoplasma acidophilum* proteasome contains two subunits of 25.8 kDa (α) and 22.3 kDa (β) arranged in a 20S proteasome containing 28 subunits: 14 α subunits and 14 β subunits arranged in four stacked rings. The two ends of the cylinder each consist of seven α subunits whilst the two inner rings had seven β subunits – each ring is a homoheptamer. The $\alpha_7\beta_7\beta_7\alpha_7$ assembly forms a three-chambered cylinder with two antechambers located on either sides of a central cavity (Figure 8.87). Sequence similarity exists between α and β subunits and there is a common fold based around an antiparallel array of β strands surrounded by five helices. Helices 1 and 2 are on one side of the β sandwich whilst helices 3, 4 and 5 are on the other. In the case of the α subunits an N-terminal extension of ∼35 residues leads to further helical structure that fits into a cleft in the β strand sandwich. In the β subunits this extension is absent and the cleft remains ‘open’ forming part of the active site. The crystal structure of the 20S complex (Figure 8.88) showed a cylindrical assembly (length ∼15 nm, diameter 11.3 nm) with the channel running the length of the assembly but widening to form three large internal cavities separated by narrow constrictions. The cavities between the α subunit and β subunit rings are the ‘antechambers’ (∼4 × 5 nm diameter) with the third cavity at the centre of the complex containing the active sites. Between the antechamber and the central cavity access is restricted to approximately 1.3 nm by a loop region containing a highly conserved RPXG motif derived from the α subunit. In this loop a Tyr residue protrudes furthest into the channel to restrict access. Further into the channel side chains derived from the β subunits exist at the entrance to the central cavity and restrict the width to ∼2.2 nm. Together, these systems form a gating system controlling entry of polypeptide chains to those that thread their way into the central cavity. A scheme of proteolysis involving protein unfolding and

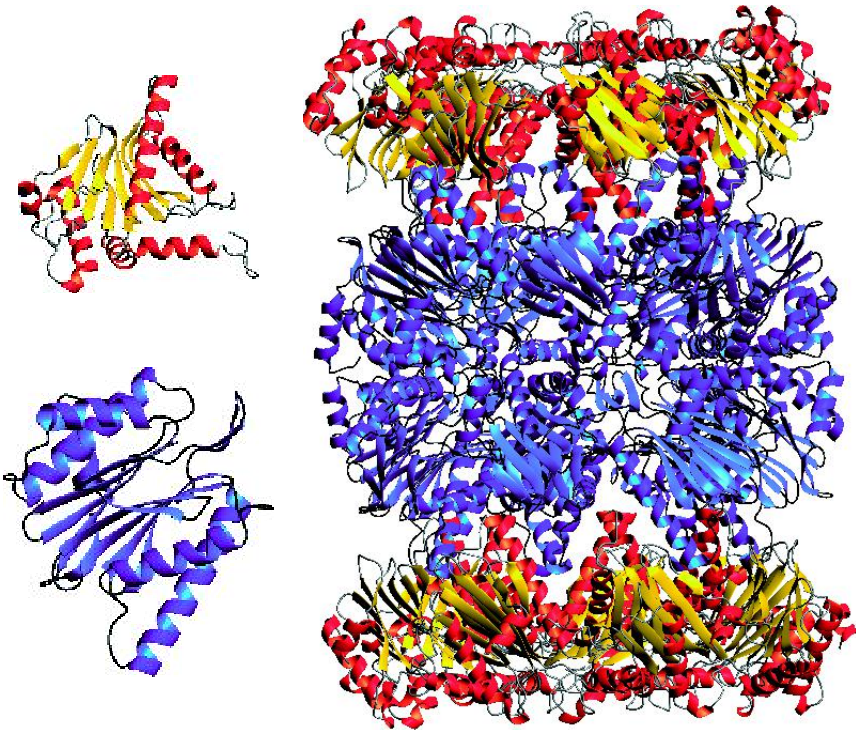


Figure 8.88 The three-dimensional structure of the 20S proteasome from the archaebacteria *T. acidophilum* (PDB: 1PMA). The structure – a 673 kDa protease complex has a barrel-shaped structure of four stacked rings. An individual α subunit is shown top left, an individual β subunit bottom left. The colour scheme is maintained in the whole complex and shows only the distribution of secondary structure for clarity

the formation of extended structure prior to degradation is likely from the organization of the proteasome.

The active site is located on the β subunit where the structure reveals a β sandwich with one side predominantly open. This side faces the inside pointing towards the central cavity. The α subunits possess a highly conserved N-terminal extension with residues 1–12 remaining invisible in the crystal structure possibly due to mobility whilst part of this extension is seen as a helix (residues 20–31) that occupies the cleft region in the β sandwich. Although the precise function of this extension is unknown its strategic location coupled with sequence conservation suggests an important role in translocation of substrate to the proteasome interior. The β subunits lack these N-terminal extensions but have pro-sequences of varying length that are cleaved during assembly of the proteasome. The most important function of this proteolytic processing is the generation of the active site residues.

The 14 identical catalytically active (β) subunits show highest activity for bond cleavage after hydrophobic residues but extensive mutagenesis involving all serines, two histidines, a single cysteine and two conserved aspartate residues in the β subunit failed to inhibit enzyme activity. The results suggested degradation was not associated with four classical forms of protease action, namely serine, cysteine, aspartyl and metalloproteinases. Further mutagenesis revealed that deletion of the N-terminal threonine or its replacement by alanine resulted in inactivation. Mechanistic studies showed that *N*-acetyl-Leu-Leu-norLeu-CHO (Figure 8.89) was a potent inhibitor of proteolysis, and crystallography of the proteasome-inhibitor adduct located the peptide bound via the

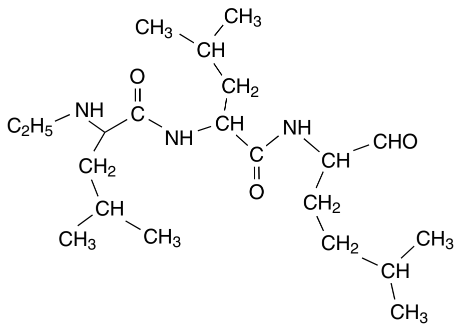


Figure 8.89 The potent inhibitor of proteasome activity *N*-acetyl-Leu-Leu-norLeu-CHO

aldehyde group to the –OH group of the N-terminal threonine.

Further insight into catalytic mechanism came with observations that the metabolite lactocystin derived from *Streptomyces* bound to the eukaryotic complex resulting in covalent modification of Thr1.[1] The results implied the involvement of N-terminal Thr residues in catalytic reactions. Initially the proteasome fold was thought to be a unique topology but gradually more members of this family of proteins have been uncovered with a common link being that they are all N-terminal nucleophile hydrolases or NTN hydrolases.

Fourteen genes contribute to the eukaryotic 20S proteasome (Figure 8.90). Each subunit has a similar structure but is coded by a separate gene. Crystallography and cryo-EM studies have shown that the patterns of subunit organization extend to eukaryotic proteasomes and when viewed in the electron microscope archaeal, bacterial, and eukaryotic 20S proteasomes form similar barrel-shaped particles ~15 nm in height and ~11 nm in diameter consisting of four heptameric rings (Figure 8.91).

The 20S assembly of the proteasome identified in eukaryotes, and first structurally characterized in *T. acidophilum*, is one component of a much larger complex found in eukaryotes. The 20S proteasome represents the core region of a larger assembly that is capped by a 19S regulatory complex. The attachment of a regulatory 'unit' to the 20S proteasome results in the formation of a 2.1 MDa complex called the 26S proteasome.

With two 19S regulatory complexes attached to the central 20S proteasome the 26S proteasomes from *Drosophila*, *Xenopus*, rat liver and spinach cells have similar shapes with an approximate length of 45 nm and a diameter of 20 nm. The significance of the 19S regulatory complex in association with the 20S assembly is that it confers ATP- and ubiquitin-dependent proteolysis on the proteasome; it is this pathway that performs most protein turnover within eukaryotic cells degrading over 90 percent of cellular proteins.

The organization of the 19S complex remains an active area of research and is relatively poorly understood in comparison to the 20S assembly. The complex from yeast contains at least 18 different proteins whose sequences are known and reversible interactions occur with proteins in the cytosol complicating compositional analysis. Within the group of 18 different subunits are six polypeptides hydrolysing ATP (ATPases) and showing homology to other members of the AAA superfamily. The AAA superfamily represents ATPases associated with various cellular activities and are found in all organisms where a defining motif is the presence of one or more modules of ~230 residues that includes an ATP-binding site and the consensus sequence A/G-x_4-G-K-S/T.

A general working model for the action of the proteasome envisages that the cap assembly binds ubiquitinated proteins, removes the ubiquitin at a later stage for recycling and unfolds the polypeptide prior to entry into the 20S core particle. The proteolytic active sites of the proteasome are found in the core particle and the cap assembly does not perform proteolysis. However, free 20S proteasome does not degrade ubiquitinated protein conjugates and shows lower activity suggesting that the 19S cap assembly is crucial to entry of substrate, activity and ubiquitin dependent hydrolysis.

Lysosomal degradation

Lysosomes contain hydrolytic enzymes including proteases, lipases, and nucleases. Any biomolecule trapped in a lysosome is degraded, and these organelles are viewed as non-specific degradative systems whilst the ubiquitin system is specific, selective and 'fine-tuned' to suit cellular demand. Lysosomes are often described as enzyme sacs and defects in lysosomal

[1] Surprisingly this inhibitor does not inactivate the *Thermoplasma* proteasome.

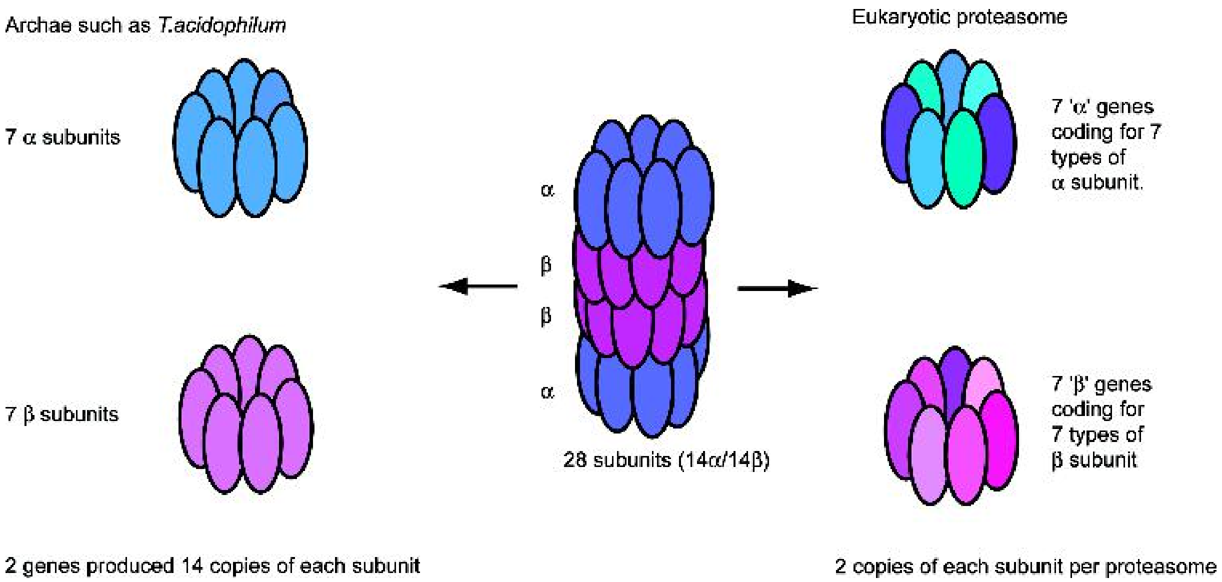

Archae such as *T.acidophilum*
7 α subunits
7 β subunits
2 genes produced 14 copies of each subunit
α
β
β
α
28 subunits (14α/14β)
Eukaryotic proteasome
7 'α' genes coding for 7 types of α subunit.
7 'β' genes coding for 7 types of β subunit
2 copies of each subunit per proteasome

Figure 8.90 The structural organization of the proteasome.

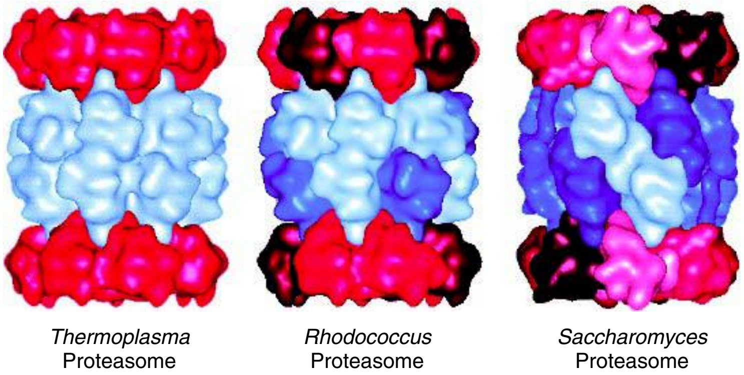


Figure 8.91 Cryo-EM derived maps of the assembly of 20S proteasomes from archaea, eubacteria and eukaryotic cells. The α type subunits are colored in red and the β type subunits in blue. The different shades of these colors indicate that two or seven distinct α and β-type subunits form the outer and inner rings of the *Rhodococcus* or *Saccharomyces* proteasome respectively (reproduced with permission from Vosges, D. *et al*. *Ann. Rev. Biochem*. 1998, **68**, 1015–1068. Annual Reviews Inc)

proteins contribute to over 40 recognizable disorders, collectively termed lysosomal storage diseases.

Individually, the disease states are rare but lead to progressive and severe impairment resulting from a deficiency in the activity of specific enzymes. The loss of activity impedes the normal degradative function of lysosomes and contributes to the accumulation of unwanted biomolecules. Included in the collection of identified lysosomal based diseases are: (i) Hurler syndrome, a deficiency in α-L-iduronidase required to degrade mucopolysaccharides; (ii) Gauchers' disease, a deficiency in β-glucocerebrosidase needed in the degradation of glycolipids; (iii) Fabry disease, caused by a deficiency in α-galactosidase, an enzyme degrading glycolipids; (iv) Pompe disease, a deficiency of α-glucosidase required to break down glycogen; and (v) Tay–Sachs disease, a lysosomal defect occurring as a result of a mutation in one subunit of the enzyme β-hexosaminidase A that leads to the accumulation of the GM2 ganglioside in neurones.

Apoptosis

Apoptosis is the process of programmed cell death. Whilst superficially this appears to be an unwanted process more careful consideration reveals that apoptosis is vital during development. A developing embryo goes through many changes in structure that can only be achieved by programmed cellular destruction. Similarly, the development of insects involves the reorganization of tissues between the larval and adult states. The concept of apoptosis is vital to normal development and although many details, such as control and initiation remain to be elucidated, the importance of a family of intracellular proteases called caspases is well documented.

Caspases are synthesized as zymogens with their activity inhibited by covalent modification, in which the active enzyme is synthesized as a precursor protein joined to additional death effector domains. This configuration is vital to achieve complete inhibition of activity since the cell is easily destroyed by the activity of caspases. Activation involves a large number of 'triggers' that excise the death effector domains, process the zymogen and form active caspases. Once activated caspases catalyse the formation of other caspases, leading to a proteolytic cascade of cellular destruction. Several methods of regulating caspase activity and hence apoptosis exist within the cell. Zymogen gene transcription is regulated whilst anti-apoptotic proteins such as those of the Bcl-2 family occur in cells and block activation of certain pro-caspases.

The first members of the caspase family identified were related to human interleukin-1 converting enzyme (ICE) and the product of the nematode cell-death gene *CED3*. To date 11 caspases have been identified in humans although mammals may possess 13 different caspases, *Drosophila melanogaster* contains seven and nematodes such as *Caenorhabditis elegans* contain three. Thus, there appears to have been an expansion in their number with increasing cellular complexity. The term caspase refers to the action of these proteins as cysteine-dependent aspartate-specific proteases. Their enzymatic properties are governed by specificity for substrates containing Asp and the use of a Cys285 sidechain for catalysing peptide-bond cleavage located within a conserved motif of five residues (QACRG). Whereas the activity of the proteasome governs the day to day turnover of proteins the activity of caspases signals the end of the cell with all proteins degraded.

Summary

Fifty years have passed since the discovery of the structure of DNA. This event marking the beginning of molecular biology expanded understanding of all events in the pathway from DNA to protein. This included structural and functional descriptions of proteins involved in the cell cycle, DNA replication, transcription, translation, post-translational events and protein turnover. In many instances determining structure has uncovered intricate details of their biological function.

The cell cycle is characterized by four distinct stages: a mitotic or M phase is followed by a G_1 (gap) phase that represents most of the cell cycle, a period of intense synthetic activity called the S phase, and finally a short G_2 phase as the cell prepares for mitosis.

Genetic studies of yeast identified mutants in which cell division events were inhibited. Control of the cell cycle is mediated by protein kinases known as Cdks where protein phosphorylation regulates cellular activity. Activity of Cdks is dependent on cyclin binding with optimal activity occurring for Cdk2 in a complex with cyclin A and phosphorylation of Thr160. The structure of this complex results in the critical movement of the T loop, a flexible region of Cdk2, governing accessibility to the catalytic cleft and the active site threonine.

Transcription is DNA-directed synthesis of RNA catalysed by RNA polymerase. Transcription proceeds from a specific sequence (promoter) in a 5′–3′ direction until a second site known as the transcriptional terminator is reached. In eukaryotes three nuclear RNA polymerases exist with clearly defined functions. RNA polymerase II is concerned with the synthesis of mRNA encoding structural genes.

Sequences associated with transcriptional elements have been identified in prokaryotes and eukaryotes. In eukaryotes the TATA box is located upstream of the transcriptional start site and governs formation of a pre-transcriptional initiation complex. The TATA box resembles the Pribnow box or −10 region found in prokaryotes.

Specific TATA binding proteins have been identified and structural studies reveal that basal transcription requires in addition to RNA polymerase the formation of pre-initiation complexes of TBP, TFIIB, TFIIE, TFIIF and TFIIH.

In eukaryotes transcription is followed by mRNA processing that involves addition of 5′ G-caps and 3′ polyA tails. Non-coding regions of mRNA known as introns are removed creating a coherent translation-effective mRNA. Introns are removed by the spliceosome.

The spliceosome contains snRNA complexed with specific proteins. Processing initial mRNA transcripts involves cutting at specific pyrimidine rich recognition sites followed by splicing them together to create mRNA that is exported from the nucleus for translation at the ribosome.

Translation converts mRNA into protein and occurs in ribonucleoprotein components known as the ribosomes. Ribosomes convert the genetic code, a series of three non-overlapping bases known as the codon, into a series of amino acids covalently linked together in a polypeptide chain.

All ribosomes are composed of large and small subunits based predominantly on highly conserved rRNA molecules together with over 50 different proteins. Biochemical studies identified two major sites known as the A and P sites. The P site (peptidyl) contains a growing polypeptide chain attached to tRNA. The A site (aminoacyl) contains charged tRNA species bearing a single amino acid that will be added to extending chains.

Protein synthesis is divided into initiation, elongation and termination. All stages involve accessory proteins such as IF1, IF2 and IF3 together with elongation and release factors such as EF1, EF2, EF3, RF1, RF2 and RF3.

Elongation is the most extensive process in protein synthesis and is divided into three steps. These processes are amino acyl tRNA binding at the A site, peptidyl transferase activity and translocation.

Structures for 50 and 30S subunits revolutionized understanding of ribosome function. The structure of the large subunit confirmed conclusively that the peptidyl transferase reaction is catalysed entirely by RNA; the ribosome is a ribozyme.

Initial translation products undergo post-translational modification that vary dramatically in type from oxidation of thiol groups to the addition of new covalent groups such as GPI anchors, oligosaccharides, myristic acid ‘tails’, inorganic groups such as phosphate or sulfate and larger organic skeletons such as heme.

The removal of peptide ‘leader’ sequences in the activation of zymogens is another important post-translational modification and converts inactive protein into an active form. Many enzymes such as proteolytic digestive enzymes, caspases and components of the blood clotting cascade are activated in this type of pathway.

N-terminal signal sequences share physicochemical properties and are recognized by a SRP. The SRP directs nascent chains to the ER membrane or cell membrane of prokaryotes. Signal sequences do not exhibit homology but have a basic N-terminal region followed by a hydrophobic core and a polar C-terminal region proximal to the cleavage site.

SRPs are found in the archae, eubacteria and eukaryotes. Mammalian SRP is the most extensively characterized system consisting of rRNA and six distinct polypeptides. The SRP directs polypeptide chains to the ER membrane and the translocon, a membrane bound protein-conducting channel.

Other forms of intracellular protein sorting exist within eukaryotes. Proteins destined for the mitochondria, chloroplast and nucleus all possess signals within their polypeptide chains. For the nucleus protein import requires the presence of a basic stretch of residues arranged either as a single block or as a bipartite structure anywhere within the primary sequence.

NLS are recognized by specific proteins (importins) that bind the target protein and shuttle the ‘cargo’ towards the nuclear pore complex. The formation of importin–cargo complexes is controlled by G proteins such as Ran.

Proteins are not immortal – they are degraded with turnover rates varying from minutes to weeks. Turnover is controlled by a complex pathway involving ubiquitin labelling.

Ubiquitin is a signal for destruction by the proteasome. The proteasome has multiple catalytic activities in a core unit based around four heptameric rings. The 20S proteasome from *T. acidophilum* has an $\alpha_7\beta_7\beta_7\alpha_7$ assembly forming a central channel guarded by two antechambers. The central chamber catalyses proteolysis based around the N terminal threonine residue (Thr1) where the side chain acts as a nucleophile attacking the carbonyl carbon of peptide bonds.

In prokaryotes the proteasome degrades proteins in ubiquitin independent pathways. In eukaryotes the 20S proteasome catalyses ubiquitin-dependent proteolysis only in the presence of a cap assembly.

Further pathways of protein degradation exist in the lysosome where defects are responsible for known metabolic disorders, such as Tay–Sachs disease, and by caspases that promote programmed cell death (apoptosis).

Problems

1. Using knowledge about the genetic code describe the products of translation of the following synthetic nucleotide ......AUAUAUAUAUAUA.... in a cell free system. Explain the results and describe the critical role of alternating oligonucleotides in determining the genetic code. Using the data presented in Table 8.6 how would this translation product differ using a system reconstituted from mitochondria.
2. Arrange the following macromolecules in decreasing order of molecular mass: tRNA, subunit L23, tetracycline, 23S rRNA, the large ribosomal subunit, 5S RNA, spliceosome protein U1A, *E. coli* DNA polymerase I, *E. coli* RNA polymerase. Obtain information from databases or information given within Chapter 8.
3. Recover the structure of the cyclinA: cdk2 binary complex from a protein database. Using any molecular graphics package highlight the following residues Lys266, Glu295, Leu255, Asp240, and Arg211 of cyclin A and Glu 8, Lys 9, Asp36, Asp38 Glu40 and Glu42 of cdk2. Describe the arrangement of these residues.
4. Puromycin binds to the A site and prevents further chain elongation. From the structure presented in Figure 8.54 account more fully for this observation.
5. The Yarus inhibitor is described as a transition state analogue that mimics the tetrahedral intermediate generated when the α amino group of the A site bound amino acyl tRNA attacks the carbonyl group of the ester linking a peptide to the peptidyl tRNA at the P site. Draw the normal assumed tetrahedral intermediate most closely linked to this inhibitor highlighting the tetrahedral carbon centre.
6. An N-terminal photo-affinity label has been used to map the tunnel of the large ribosomal subunit. Describe how you might perform this experiment and what results might you expect to obtain from such a procedure?
7. Unlike the mitochondria and chloroplast signal sequence nuclear localization signals are not removed. What might be the reasons underlying this observation?
8. Caspases are proteolytic enzymes with an active site cysteine. Discuss possible enzyme mechanisms in view of the known activity of serine proteases.