G1 RNA STRUCTURE

Key Notes

Covalent structure

RNA is a polymer chain of ribonucleotides joined by 3′5′ phosphodiester bonds. The covalent structure is very similar to that for DNA except that uracil replaces thymine and ribose replaces deoxyribose.

RNA secondary structure

RNA molecules are largely single-stranded but there are regions of self-complementarity where the RNA chain forms internal double-stranded regions.

Related topics

DNA structure (F1)
Transcription in prokaryotes (G2)
Operons (G3)
Transcription in eukaryotes: an overview (G4)
Transcription of protein-coding genes in eukaryotes (G5)
Regulation of transcription by RNA Pol II (G6)

Covalent structure

Like DNA (see Topic F1), RNA is a long polymer consisting of nucleotides joined by 3′5′ phosphodiester bonds. However, there are some differences:

- The bases in RNA are adenine (abbreviated A), guanine (G), uracil (U) and cytosine (C). Thus thymine in DNA is replaced by **uracil** in RNA, a different pyrimidine (*Fig. 1a*). However, like thymine (see Topic F1), uracil can form base pairs with adenine.
- The sugar in RNA is **ribose** rather than deoxyribose as in DNA (*Fig. 1b*).

The corresponding **ribonucleosides** are **adenosine**, **guanosine**, **cytidine** and **uridine**. The corresponding **ribonucleotides** are **adenosine 5′-triphosphate** (ATP), **guanosine 5′-triphosphate** (GTP), **cytidine 5′-triphosphate** (CTP) and **uridine 5′-triphosphate** (UTP).

As with DNA, the nucleotide sequence of RNA is also written as a base sequence in the 5′ → 3′ direction. Thus GUCAAGCCGGAC is the sequence of one short RNA molecule.

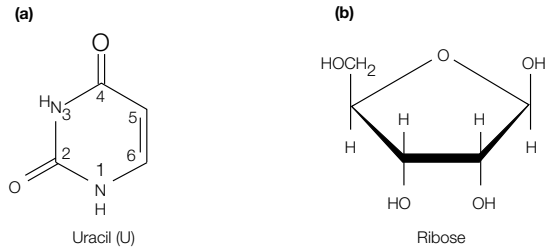

Fig. 1. (a) Uracil, (b) ribose.

RNA secondary structure

Most RNA molecules are single-stranded but an RNA molecule may contain regions which can form complementary base pairing where the RNA strand loops back on itself (*Fig. 2*). If so, the RNA will have some double-stranded regions. Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) (see Topics G8 and G9, respectively) exhibit substantial secondary structure, as do some messenger RNAs (mRNAs).

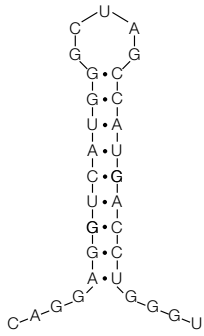

Fig. 2. An example of self-complementarity in RNA, forming an internal double-stranded region; hydrogen bonding between bases is shown by the symbol •.

G2 TRANSCRIPTION IN PROKARYOTES

Key Notes

Three phases of transcription

Transcription by *E. coli* RNA polymerase occurs in three phases; initiation, elongation and termination. Initiation involves binding of the enzyme to a promoter upstream of the gene. During elongation, the antisense DNA strand is used as the template so that the RNA made has the same base sequence as the sense (coding) strand, except that U replaces T. A termination signal is eventually encountered that halts synthesis and causes release of the completed RNA.

Promoters and initiation

RNA polymerase holoenzyme (containing $\alpha_2\beta\beta'\omega\sigma$ subunits) initiates transcription by binding to a 40–60 bp region that contains two conserved promoter elements, the –10 sequence (Pribnow box) with the consensus TATAAT and the –35 sequence with the consensus TTGACA. The σ factor is essential for initiation. No primer is required. Promoters vary up to 1000-fold in their efficiency of initiation which depends on the exact sequence of the key promoter elements as well as flanking sequences.

Elongation

Following initiation, the σ subunit dissociates from RNA polymerase to leave the core enzyme ($\alpha_2\beta\beta'\omega$) that continues RNA synthesis in a 5′ → 3′ direction using the four ribonucleoside 5′-triphosphates as precursors. The DNA double helix is unwound for transcription, forming a transcription bubble, and is then rewound after the transcription complex has passed.

Termination

A common termination signal is a hairpin structure formed by a palindromic GC-rich region, followed by an AT-rich sequence. Other signals are also used which require the assistance of rho (ρ) protein for effective termination.

RNA processing

Messenger RNA transcripts of protein-coding genes in prokaryotes require little or no modification before translation. Ribosomal RNAs and transfer RNAs are synthesized as precursor molecules that require processing by specific ribonucleases to release the mature RNA molecules.

Related topics

DNA structure (F1)
RNA structure (G1)
Operons (G3)
Transcription in eukaryotes: an overview (G4)
Transcription of protein-coding genes in eukaryotes (G5)
Regulation of transcription by RNA Pol II (G6)
Processing of eukaryotic pre-mRNA (G7)
Ribosomal RNA (G8)
Transfer RNA (G9)

Three phases of transcription

Gene transcription by *E. coli* RNA polymerase takes place in three phases: **initiation**, **elongation** and **termination**. During initiation, RNA polymerase recognizes a specific site on the DNA, upstream from the gene that will be transcribed, called a **promoter site** and then unwinds the DNA locally. During elongation the RNA polymerase uses the **antisense (−) strand** of DNA as template and synthesizes a complementary RNA molecule using ribonucleoside 5′-triphosphates as precursors. The RNA produced has the same sequence as the nontemplate strand, called the **sense (+) strand** (or **coding strand**) except that the RNA contains U instead of T. At different locations on the bacterial chromosome, sometimes one strand is used as template, sometimes the other, depending on which strand is the coding strand for the gene in question. The correct strand to be used as template is identified for the RNA polymerase by the presence of the promoter site. Finally, the RNA polymerase encounters a termination signal and ceases transcription, releasing the RNA transcript and dissociating from the DNA.

Promoters and initiation

In *E. coli*, all genes are transcribed by a single large RNA polymerase with the subunit structure $\alpha_2 \beta\beta'\omega\sigma$. This complete enzyme, called the **holoenzyme**, is needed to initiate transcription since the σ factor is essential for recognition of the promoter; it decreases the affinity of the core enzyme for nonspecific DNA binding sites and increases its affinity for the promoter. It is common for prokaryotes to have several σ factors that recognize different types of promoter (in *E. coli*, the most common σ factor is σ^{70}).

The holoenzyme binds to a promoter region about 40–60 bp in size and then initiates transcription a short distance downstream (i.e. 3′ to the promoter). Within the promoter lie two 6-bp sequences that are particularly important for promoter function and which are therefore highly conserved between species. Using the convention of calling the first nucleotide of a transcribed sequence as +1, these two **promoter elements** lie at positions –10 and –35, that is about 10 and 35 bp, respectively, upstream of where transcription will begin (*Fig. 1*).

- The **–10 sequence** has the consensus TATAAT. Because this element was discovered by Pribnow, it is also known as the **Pribnow box**. It is an important recognition site that interacts with the σ factor of RNA polymerase.
- The **–35 sequence** has the consensus TTGACA and is important in DNA unwinding during transcriptional initiation.

The actual sequence between the –10 sequence and the –35 sequence is not conserved (i.e. it varies from promoter to promoter) but the distance between these two sites is extremely important for correct functioning of the promoter.

Promoters differ by up to 1000-fold in their efficiency of initiation of transcription so that genes with strong promoters are transcribed very frequently

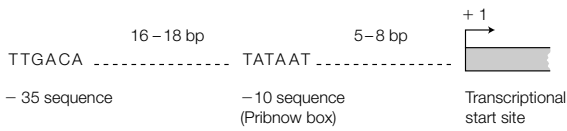

Fig. 1. Prokaryotic promoter showing the –10 sequence and –35 sequence. By convention, the first nucleotide of the template DNA that is transcribed into RNA is denoted +1, the transcriptional start site.

whereas genes with weak promoters are transcribed far less often. The –10 and –35 sequences of strong promoters correspond well with the consensus sequences shown in *Fig. 1* whereas weaker promoters may have sequences that differ from these at one or more nucleotides. The nature of the sequences around the transcriptional start site can also influence the efficiency of initiation. RNA polymerase does not need a primer to begin transcription (cf. DNA polymerases, Topics F3 and F4); having bound to the promoter site, the RNA polymerase begins transcription directly.

Elongation

After transcription initiation, the σ factor is released from the transcriptional complex to leave the **core enzyme** ($\alpha_2 \beta\beta'\omega$) which continues elongation of the RNA transcript. Thus the core enzyme contains the catalytic site for polymerization, probably within the β subunit. The first nucleotide in the RNA transcript is usually pppG or pppA. The RNA polymerase then synthesizes RNA in the 5′ → 3′ direction, using the four ribonucleoside 5′-triphosphates (ATP, CTP, GTP, UTP) as precursors. The 3′-OH at the end of the growing RNA chain attacks the α phosphate group of the incoming ribonucleoside 5′-triphosphate to form a 3′5′ phosphodiester bond (*Fig. 2*). The complex of RNA polymerase, DNA template and new RNA transcript is called a **ternary complex** (i.e. three components) and the region of unwound DNA that is undergoing transcription is called the **transcription bubble** (*Fig. 3*). The RNA transcript forms a transient RNA–DNA hybrid helix with its template strand but then peels away from the DNA as transcription proceeds. The DNA is unwound ahead of the transcription bubble and after the transcription complex has passed, the DNA rewinds.

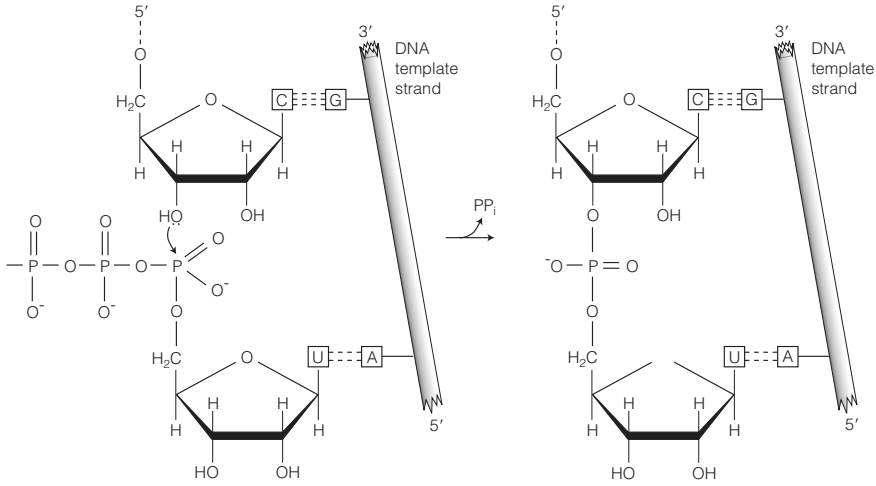

Fig. 2. Transcription by RNA polymerase. In each step the incoming ribonucleotide selected is that which can base pair with the next base of the DNA template strand. In the diagram, the incoming nucleotide is rUTP to base pair with the A residue of the template DNA. A 3′5′ phosphodiester bond is formed, extending the RNA chain by one nucleotide, and pyrophosphate is released. Overall the RNA molecule grows in a 5′ to 3′ direction.

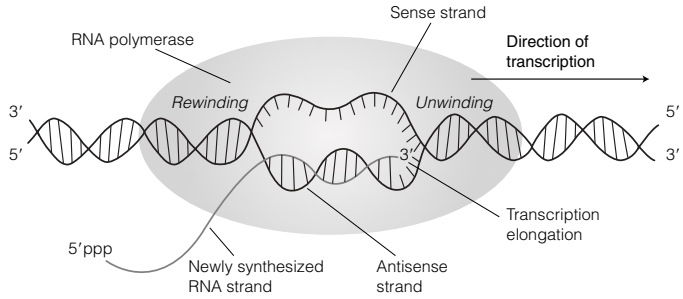

Fig. 3. A transcription bubble. The DNA double helix is unwound and RNA polymerase then synthesizes an RNA copy of the DNA template strand. The nascent RNA transiently forms an RNA–DNA hybrid helix but then peels away from the DNA which is subsequently rewound into a helix once more.

Termination

Transcription continues until a termination sequence is reached. The most common termination signal is a GC-rich region that is a **palindrome**, followed by an AT-rich sequence. The RNA made from the DNA palindrome is self-complementary and so base pairs internally to form a **hairpin structure** rich in GC base pairs followed by four or more U residues (*Fig. 4*). However, not all termination sites have this hairpin structure. Those that lack such a structure require an additional protein, called **rho** (ρ), to help recognize the termination site and stop transcription.

RNA processing

In prokaryotes, RNA transcribed from protein-coding genes (**messenger RNA**, **mRNA**), requires little or no modification prior to translation. In fact, many mRNA molecules begin to be translated even before RNA synthesis has finished. However, **ribosomal RNA** (**rRNA**) and **transfer RNA** (**tRNA**) are synthesized as precursor molecules that do require post-transcriptional processing (see Topics G8 and G9, respectively).

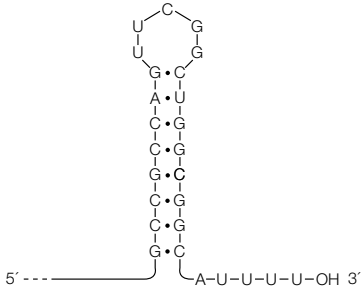

Fig.4 A typical hairpin structure formed by the 3′ end of an RNA molecule during termination of transcription.

G3 OPERONS

Key Notes

Operons: an overview

Operons are clusters of structural genes under the control of a single operator site and regulator gene which ensures that expression of the structural genes is coordinately controlled.

The *lac* operon

The *lac* operon contains *lacZ*, *lacY* and *lacA* genes encoding β-galactosidase, galactose permease, and thiogalactoside transacetylase, respectively, preceded by an operator site (O_{lac}) and a promoter (P_{lac}). The operon is transcribed by RNA polymerase to produce a single polycistronic mRNA that is then translated to produce all three enzymes. When lactose is present, the background level of β-galactosidase converts some lactose to allolactose which then acts as an inducer and turns on transcription of the *lac* operon. IPTG can also act as an inducer. Transcription of the operon is controlled by the lac repressor protein encoded by the *lacI* gene.

The lac repressor

The *lac*I gene has its own promoter (P_{lacI}) to which RNA polymerase binds and initiates transcription. In the absence of an inducer, the *lacI* gene is transcribed, producing lac repressor protein which binds to the *lac* operator site, O_{lac}, and prevents transcription of the *lac* operon. In the presence of an inducer (such as allolactose or IPTG), the inducer binds to the repressor and changes its conformation, reducing its affinity for the *lac* operator. Thus the repressor now dissociates and allows RNA polymerase to transcribe the *lac* operon.

CRP/CAP

Catabolite activator protein, CAP (also called cAMP receptor protein, CRP) is required for high level transcription of the *lac* operon. It associates with 3′5′ cyclic AMP to form a CRP–cAMP complex. CRP–cAMP binds to the *lac* promoter and increases the binding of RNA polymerase, stimulating transcription of the *lac* operon. When glucose is present, the intracellular level of cAMP falls, CRP alone cannot bind to the *lac* promoter and the *lac* operon is only weakly transcribed. When glucose is absent, the level of intracellular cAMP rises, the CRP–cAMP complex is formed and stimulates transcription of the *lac* operon, allowing lactose to be used as an alternative carbon source.

Positive and negative regulation

In negative regulation of prokaryotic gene expression, bound repressor prevents transcription of the structural genes. In positive regulation of gene expression, an activator binds to DNA and increases the rate of transcription. The *lac* operon is subject to both negative and positive control.

The *trp* operon

The *trp* operon contains five structural genes encoding enzymes for tryptophan biosynthesis, a *trp* promoter (P_{trp}) and a *trp* operator sequence (O_{trp}). The operon is transcribed only when tryptophan is scarce.

The trp repressor

When tryptophan is lacking, a trp repressor protein (encoded by the *trpR* operon) is synthesized but cannot bind to the *trp* operator and so the *trp* operon is transcribed to produce the enzymes that then synthesize tryptophan for the cell. When tryptophan is present, it binds to the repressor and activates it so that the repressor now binds to the *trp* operator and stops transcription of the *trp* operon.

Attenuation

The *trp* operon is also controlled by attenuation. A leader sequence in the polycistronic mRNA can form several possible stem-loop secondary structures, one of which can act as a transcription terminator whilst a different stem-loop can act as an anti-terminator. In the presence of tryptophan, ribosomes bind to the *trp* polycistronic mRNA that is being transcribed, following closely behind the RNA polymerase, and begin to translate the leader sequence. The position of the bound ribosomes prevents formation of the anti-terminator stem-loop but allows the terminator loop to form which then inhibits further transcription of the *trp* operon. If tryptophan is scarce, the ribosome pauses when attempting to translate the two trp codons in the leader sequence, which leaves the leader sequence available to form the antiterminator stem-loop. Transcription of the *trp* operon is then allowed to continue.

Attenuation vs. repression

The *trp* operon is regulated by both repression (which determines whether transcription will occur or not) and attenuation (which fine tunes transcription). Other operons for amino acid biosynthetic pathways may be regulated by both repression and attenuation or only by attenuation.

Related topics

DNA structure (F1)
RNA structure (G1)
Transcription in prokaryotes (G2)
Transcription in eukaryotes: an overview (G4)
Transcription of protein-coding genes in eukaryotes (G5)
Regulation of transcription by RNA Pol II (G6)
Processing of eukaryotic pre-RNA (G7)
Ribosomal RNA (G8)
Transfer RNA (G9)

Operons: an overview

Many protein-coding genes in bacteria are clustered together in **operons** which serve as transcriptional units that are coordinately regulated. It was Jacob and Monod in 1961 who proposed the operon model for the regulation of transcription. The operon model proposes three elements:

- a set of **structural genes** (i.e. genes encoding the proteins to be regulated);
- an **operator site**, which is a DNA sequence that regulates transcription of the structural genes;
- a **regulator gene** which encodes a protein that recognizes the operator sequence.

The *lac* operon

One of the most studied operons is the ***lac* operon** in *E. coli*. This codes for key enzymes involved in lactose metabolism: **galactoside permease** (also known as **lactose permease**; it transports lactose into the cell across the cell membrane) and **β-galactosidase** (which hydrolyzes lactose to glucose and galactose). It also codes for a third enzyme, **thiogalactoside transacetylase**. Normally *E. coli* cells

make very little of any of these three proteins but when lactose is available it causes a large and coordinated increase in the amount of each enzyme. Thus each enzyme is an **inducible enzyme** and the process is called **induction**. The mechanism is that the few molecules of β-galactosidase in the cell before induction convert the lactose to allolactose which then turns on transcription of these three genes in the *lac* operon. Thus allolactose is an **inducer**. Another inducer of the *lac* operon is **isopropylthiogalactoside** (IPTG). Unlike allolactose, this inducer is not metabolized by *E. coli* and so is useful for experimental studies of induction.

In the *lac* operon (*Fig. 1*), the structural genes are the *lacZ*, *lacY* and *lacA* genes encoding β-galactosidase, the permease and the transacetylase, respectively. They are transcribed to yield a single **polycistronic mRNA** that is then translated to produce all three enzymes (*Fig. 1*). The existence of a polycistronic mRNA ensures that the amounts of all three gene products are regulated coordinately. Transcription occurs from a single promoter (P_{lac}) that lies upstream of these structural genes (*Fig. 1*) and binds RNA polymerase (see Topic G2). However, also present are an operator site (O_{lac}) between the promoter and the structural genes, and a *lac*I gene that codes for the **lac repressor** protein.

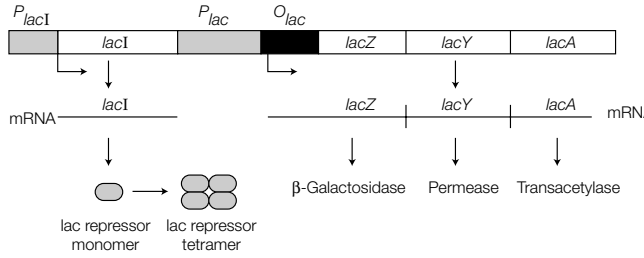

Fig. 1. Structure of the lac *operon.*

The lac repressor

The *lac*I gene has its own promoter (P_{lacI}) that binds RNA polymerase and leads to transcription of lac repressor mRNA and hence production of lac repressor protein monomers. Four identical repressor monomers come together to form the active tetramer which can bind tightly to the *lac* operator site, O_{lac}. The O_{lac} sequence is **palindromic**, that is it has the same DNA sequence when one strand is read 5′ to 3′ and the complementary strand is read 5′ to 3′. This symmetry of the operator site is matched by the symmetry of the repressor tetramer.

In the absence of an inducer such as allolactose or IPTG, the *lac*I gene is transcribed and the resulting repressor protein binds to the operator site of the *lac* operon, O_{lac}, and prevents transcription of the *lacZ*, *lacY* and *lacA* genes (*Fig. 2*). During induction, the inducer binds to the repressor. This causes a change in conformation of the repressor that greatly reduces its affinity for the *lac* operator site. The lac repressor now dissociates from the operator site and allows the RNA polymerase (already in place on the adjacent promoter site) to begin transcribing the *lacZ*, *lacY* and *lacA* genes (*Fig. 3*). This yields many copies of the polycistronic mRNA and, after translation, large amounts of all three enzymes.

If inducer is removed, the lac repressor rapidly binds to the *lac* operator site and transcription is inhibited almost immediately. The *lacZYA* RNA transcript is

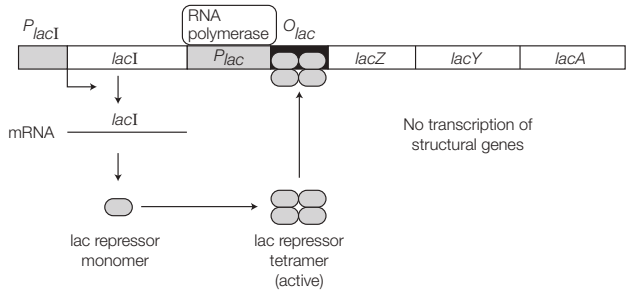

Fig. 2. Repression of transcription by the lac repressor in the absence of inducer.

very unstable and so degrades quickly such that further synthesis of the β-galactosidase, permease and transacetylase ceases.

CRP/CAP

High level transcription of the *lac* operon requires the presence of a specific activator protein called **catabolite activator protein** (**CAP**), also called **cAMP receptor protein** (**CRP**). This protein, which is a dimer, cannot bind to DNA unless it is complexed with 3′5′ cyclic AMP (cAMP). The CRP–cAMP complex binds to the *lac* promoter just upstream from the binding site for RNA polymerase. It increases the binding of RNA polymerase and so stimulates transcription of the *lac* operon.

Whether or not the CRP protein is able to bind to the *lac* promoter depends on the carbon source available to the bacterium (*Fig. 2*). When glucose is present, *E. coli* does not need to use lactose as a carbon source and so the *lac* operon does not need to be active. Thus the system has evolved to be responsive to glucose. Glucose inhibits **adenylate cyclase**, the enzyme that synthesizes cAMP from ATP. Thus, in the presence of glucose the intracellular level of cAMP falls, so CRP cannot bind to the *lac* promoter, and the *lac* operon is only weakly active (even in the presence of lactose). When glucose is absent, adenylate cyclase is not inhibited, the level of intracellular cAMP rises and binds to CRP. Therefore, when glucose is absent but lactose is present, the CRP–cAMP complex stimulates

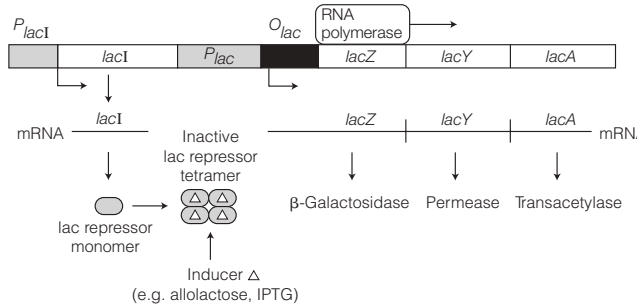

Fig. 3. Inducer inactivates the lac repressor and so allows transcription of the structural genes.

transcription of the *lac* operon and allows the lactose to be used as an alternative carbon source. In the absence of lactose, the lac repressor of course ensures that the *lac* operon remains inactive. These combined controls ensure that the *lacZ*, *lacY* and *lacA* genes are transcribed strongly only if glucose is absent and lactose is present.

Positive and negative regulation

The *lac* operon is a good example of **negative control** (**negative regulation**) of gene expression in that bound repressor prevents transcription of the structural genes. **Positive control** (**positive regulation**) of gene expression is when the regulatory protein binds to DNA and increases the rate of transcription. In this case the regulatory protein is called an activator. The CAP/CRP involved in regulating the *lac* operon is a good example of an activator. Thus the *lac* operon is subject to both negative and positive control.

The *trp* operon

The tryptophan (*trp*) operon (*Fig. 4*) contains five structural genes encoding enzymes for tryptophan biosynthesis with an upstream *trp* promoter (P_{trp}) and *trp* operator sequence (O_{trp}). The *trp* operator region partly overlaps the *trp* promoter. The operon is regulated such that transcription occurs when tryptophan in the cell is in short supply.

The trp repressor

In the absence of tryptophan (*Fig. 4a*), a trp repressor protein encoded by a separate operon, *trpR*, is synthesized and forms a dimer. However, this is inactive and so is unable to bind to the *trp* operator and the structural genes of the *trp* operon are transcribed. When tryptophan is present (*Fig. 4b*), the enzymes for

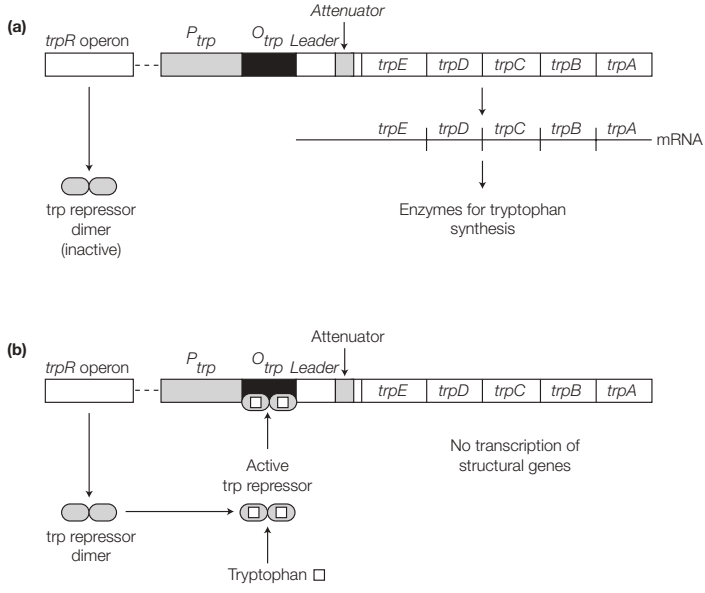

Fig. 4. Regulation of the trp *operon (a) transcription in the absence of tryptophan (b) no transcription in the presence of tryptophan.*

tryptophan biosynthesis are not needed and so expression of these genes is turned off. This is achieved by tryptophan binding to the repressor to activate it so that it now binds to the operator and stops transcription of the structural genes. In this role, tryptophan is said to be a **co-repressor**. This is negative control, because the bound repressor prevents transcription, but note that the *lac* operon and *trp* operon show two ways in which negative control can be achieved; either (as in the *lac* operon) by having an active bound repressor that is inactivated by a bound ligand (the inducer) or (as in the *trp* operon) by having a repressor that is inactive normally but activated by binding the ligand. As in the case of the *lac* operator, the core binding site for the *trp* repressor in the *trp* operator is palindromic.

Attenuation

A second mechanism, called **attenuation**, is also used to control expression of the *trp* operon. The 5′ end of the polycistronic mRNA transcribed from the *trp* operon has a **leader sequence** upstream of the coding region of the *trpE* structural gene (*Fig. 4*). This leader sequence encodes a 14 amino acid **leader peptide** containing two tryptophan residues.

The function of the leader sequence is to fine tune expression of the *trp* operon based on the availability of tryptophan inside the cell. It does this as follows. The leader sequence contains four regions (*Fig. 5*, numbered 1–4) that can form a variety of base paired **stem-loop** (‘**hairpin**’) secondary structures. Now consider the two extreme situations: the presence or absence of tryptophan. Attenuation depends on the fact that, in bacteria, ribosomes attach to mRNA as it is being

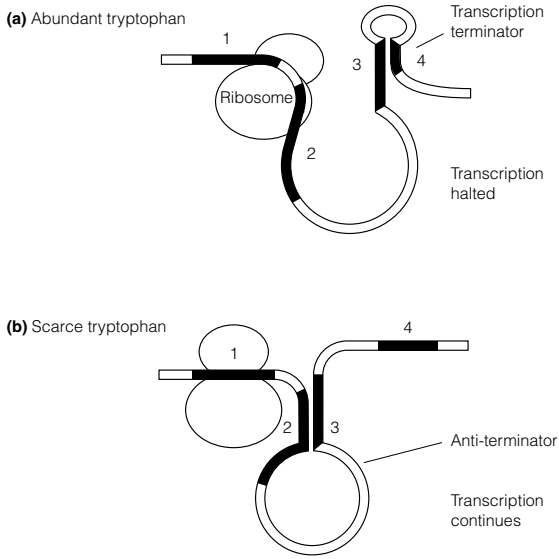

Fig. 5. Attenuation of the trp operon. (a) When tryptophan is plentiful, sequences 3 and 4 base pair to form a 3:4 structure that stops transcription (b) when tryptophan is in short supply, the ribosome stalls at the trp codons in sequence 1, leaving sequence 2 available to interact with sequence 3. Thus a 3:4 transcription terminator structure cannot form and transcription continues.

synthesized and so translation starts even before transcription of the whole mRNA is complete. When tryptophan is abundant (*Fig. 5a*), ribosomes bind to the *trp* polycistronic mRNA that is being transcribed and begin to translate the leader sequence. Now, the two trp codons for the leader peptide lie within sequence 1, and the translational Stop codon (see Topic H1) lies between sequence 1 and 2. During translation, the ribosomes follow very closely behind the RNA polymerase and synthesize the leader peptide, with translation stopping eventually between sequences 1 and 2. At this point, the position of the ribosome prevents sequence 2 from interacting with sequence 3. Instead sequence 3 base pairs with sequence 4 to form a 3:4 stem loop which acts as a **transcription terminator.** Therefore, when tryptophan is present, further transcription of the *trp* operon is prevented. If, however, tryptophan is in short supply (*Fig. 5b*), the ribosome will pause at the two *trp* codons contained within sequence 1. This leaves sequence 2 free to base pair with sequence 3 to form a 2:3 structure (also called the **anti-terminator**), so the 3:4 structure cannot form and transcription continues to the end of the *trp* operon. Hence the availability of tryptophan controls whether transcription of this operon will stop early (attenuation) or continue to synthesize a complete polycistronic mRNA.

Historically, attenuation was discovered when it was noticed that deletion of a short sequence of DNA between the operator and the first structural gene, *trpE*, increased the level of transcription. This region was named the **attenuator** (see *Fig. 4*) and is the DNA that encodes that part of the leader sequence that forms the transcription terminator stem-loop.

Attenuation vs. repression

Overall, for the *trp* operon, repression via the *trp* repressor determines whether transcription will occur or not and attenuation then fine tunes transcription. Attenuation occurs in at least six other operons that encode enzymes for amino acid biosynthetic pathways. In some cases, such as the *trp* operon, both repression and attenuation operate to regulate expression. In contrast, for some other operons such as the *his*, *thr* and *leu* operons, transcription is regulated only by attenuation.

G4 TRANSCRIPTION IN EUKARYOTES: AN OVERVIEW

Key Notes

Three RNA polymerases

In eukaryotes, RNA is synthesized by three RNA polymerases: RNA Pol I is a nucleolar enzyme that transcribes rRNAs, RNA Pol II is located in the nucleoplasm and transcribes mRNAs, snoRNAs and most snRNAs, RNA Pol III is also nucleoplasmic and transcribes tRNA and 5S rRNA, as well as U6 snRNA and the 7S RNA of the signal recognition particle (SRP).

RNA synthesis

Each RNA polymerase transcribes only one strand, the antisense (−) strand, of a double-stranded DNA template, directed by a promoter. Synthesis occurs 5′ → 3′ and does not require a primer.

RNA polymerase subunits

Each of the three RNA polymerases contains 12 or more subunits, some of which are similar to those of *E. coli* RNA polymerase. However, four to seven subunits in each enzyme are unique to that enzyme.

Related topics

DNA structure (F1)
RNA structure (G1)
Transcription in prokaryotes (G2)
Operons (G3)
Transcription of protein-coding genes in eukaryotes (G5)
Regulation of transcription by RNA Pol II (G6)
Processing of eukaryotic pre-mRNA (G7)
Ribosomal RNA (G8)
Transfer RNA (G9)

Three RNA polymerases

Unlike prokaryotes where all RNA is synthesized by a single RNA polymerase, the nucleus of a eukaryotic cell has three RNA polymerases responsible for transcribing different types of RNA.

- **RNA polymerase I** (**RNA Pol I**) is located in the nucleolus and transcribes the **28S**, **18S** and **5.8S rRNA** genes.
- **RNA polymerase II** (**RNA Pol II**) is located in the nucleoplasm and transcribes **protein-coding genes**, to yield pre-mRNA, and also the genes encoding **small nucleolar RNAs (snoRNAs)** involved in rRNA processing (see Topic G8) and **small nuclear RNAs** (**snRNAs**) involved in mRNA processing (see Topic G7), except for U6 snRNA.
- **RNA polymerase III** (**RNA Pol III**) is also located in the nucleoplasm. It transcribes the genes for **tRNA**, **5S rRNA, U6 snRNA**, and the **7S RNA** associated with the signal recognition particle (SRP) involved in the translocation of proteins across the endoplasmic reticulum membrane (see Topic H4).

RNA synthesis

The basic mechanism of RNA synthesis by these eukaryotic RNA polymerases is the same as for the prokaryotic enzyme (see Topic G2), that is:

- the initiation of RNA synthesis by RNA polymerase is directed by the presence of a promoter site on the 5′ side of the transcriptional start site;
- the RNA polymerase transcribes one strand, the **antisense** (−) **strand**, of the DNA template;
- RNA synthesis does not require a primer;
- RNA synthesis occurs in the 5′ → 3′ direction with the RNA polymerase catalyzing a nucleophilic attack by the 3′-OH of the growing RNA chain on the α phosphorus atom on an incoming ribonucleoside 5′-triphosphate.

RNA polymerase subunits

Each of the three eukaryotic RNA polymerases contains 12 or more subunits and so these are large complex enzymes. The genes encoding some of the subunits of each eukaryotic enzyme show DNA sequence similarities to genes encoding subunits of the core enzyme ($\alpha_2 \beta\beta'\omega$) of *E. coli* RNA polymerase (see Topic G2). However, four to seven other subunits of each eukaryotic RNA polymerase are unique in that they show no similarity either with bacterial RNA polymerase subunits or with the subunits of other eukaryotic RNA polymerases.

G5 TRANSCRIPTION OF PROTEIN-CODING GENES IN EUKARYOTES

Key Notes

Gene organization

Most protein-coding genes in eukaryotes consist of coding sequences called exons interrupted by noncoding sequences called introns. The number of introns and their size varies from gene to gene. The primary transcript (pre-mRNA) undergoes processing reactions to yield mature mRNA.

Initiation of transcription

Most promoter sites for RNA polymerase II have a TATA box located about 25 bp upstream of the transcriptional start site. RNA polymerase binding to the promoter requires the formation of a transcription initiation complex involving several general (basal) transcription factors that assemble in a strict order. Some protein-coding genes lack a TATA box and have an initiator element instead, centered around the transcriptional start site. The initiation of transcription of these genes requires an additional protein to recognize the initiator element and facilitate formation of the transcription initiation complex; many of the same transcription factors for initiation of TATA box promoters are also involved here. Yet other promoters lack either a TATA box or an initiator element and transcription starts within a broad region of DNA rather than at a defined location.

Elongation, termination and RNA processing

After TFIIH phosphorylates the C-terminal domain (CTD) of RNA polymerase II, this enzyme starts moving along the DNA template synthesizing RNA. Elongation continues until transcription comes to a halt at varying RNA processing distances downstream of the gene, releasing the primary RNA transcript, pre-mRNA. This molecule then undergoes processing reactions to yield mRNA.

Related topics

- DNA structure (F1)
- RNA structure (G1)
- Transcription in prokaryotes (G2)
- Operons (G3)
- Transcription in eukaryotes: an overview (G4)
- Regulation of transcription by RNA Pol II (G6)
- Processing of eukaryotic pre-mRNA (G7)
- Ribosomal RNA (G8)
- Transfer RNA (G9)

Gene organization

In marked contrast to prokaryotic genes where proteins are encoded by a continuous sequence of triplet codons, the vast majority of protein-coding genes in eukaryotes are **discontinuous**. The coding sections of the gene (called **exons**) are interrupted by noncoding sections of DNA (called **introns**; *Fig. 1*). Nevertheless, the triplet codons within the exons and the order of exons themselves in the gene

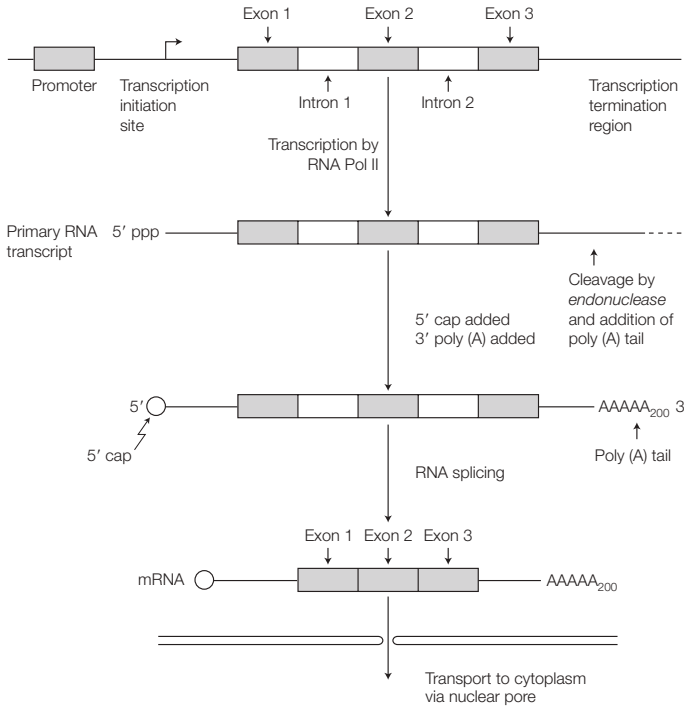

Fig. 1. Structure and expression of a protein coding gene in a eukaryote.

is still **colinear** with the amino acid sequence of the encoded polypeptide. The number of introns in a protein-coding gene varies and they range in size from about 80 bp to over 10 000 bp. The primary transcript is a **pre-mRNA** molecule which must be processed to yield mature mRNA ready for translation. During RNA processing, the pre-mRNA receives a 5′ cap and (usually but not always) a poly(A) tail of about 200 A residues, and the intron sequences are removed by RNA splicing. These RNA processing reactions are covered in detail in Topic G7.

Initiation of transcription

Most promoter sites for RNA polymerase II include a highly conserved sequence located about 25–35 bp upstream (i.e. to the 5′ side) of the start site which has the consensus TATA(A/T)A(A/T) and is called the **TATA box** (*Fig. 2*). Since the start site is denoted as position +1, the TATA box position is said to be located at about position −25. The TATA box sequence resembles the −10 sequence (see Topic G2) in prokaryotes (TATAAT) except that it is located further upstream. Both elements have essentially the same function, namely recognition by the RNA polymerase in order to position the enzyme at the correct location to initiate transcription. The sequence around the TATA box is also important in that it influences the efficiency of initiation. Transcription is also regulated by **upstream control elements** that lie 5′ to the TATA box (*Fig. 2* and Topic G6).

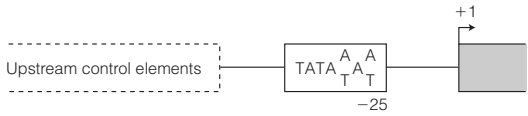

Fig. 2. A typical promoter for RNA Pol II. The TATA box is located approximately 25 bp upstream of the transcriptional start site (denoted as +1).

Some eukaryotic protein-coding genes lack a TATA box and have an **initiator element** instead, centered around the transcriptional initiation site. This does not have a strong consensus between genes but often includes a C at position −1 and an A at position +1. Yet other promoters have neither a TATA box nor an initiator element; these genes tend to be transcribed at low rates and initiate transcription somewhere within a broad region of DNA (about 200 bp or so) rather than at a defined transcriptional start site.

In order to initiate transcription, RNA polymerase II requires the assistance of several other proteins or protein complexes, called **general** (or **basal**) **transcription factors**, which must assemble into a complex on the promoter in order for RNA polymerase to bind and start transcription (*Fig. 3*). These all have the generic name of **TFII** (for Transcription Factor for RNA polymerase II). The first event in initiation is the binding of the **transcription factor IID** (**TFIID**) protein complex to the TATA box via one its subunits called **TBP** (**TATA box binding protein**). As soon as the TFIID complex has bound, **TFIIA** binds and stabilizes the TFIID-TATA box interaction. Next, **TFIIB** binds to TFIID. However, TFIIB can also bind to RNA polymerase II and so acts as a bridging protein. Thus, RNA polymerase II, which has already complexed with **TFIIF**, now binds. This is followed by the binding of **TFIIE** and **H**. This final protein complex contains at least 40 polypeptides and is called the **transcription initiation complex**.

Those protein-coding genes that have an initiator element instead of a TATA box (see above) appear to need another protein(s) that binds to the initiator element. The other transcription factors then bind to form the transcription initiation complex in a similar manner to that described above for genes possessing a TATA box promoter.

Elongation, termination and RNA processing

TFIIH has two functions. It is a helicase, which means that it can use ATP to unwind the DNA helix, allowing transcription to begin. In addition, it phosphorylates RNA polymerase II which causes this enzyme to change its conformation and dissociate from other proteins in the initiation complex. RNA polymerase II now starts moving along the DNA template, synthesizing RNA, that is, the process enters the **elongation phase**. The key phosphorylation occurs on a long C-terminal tail called the **C-terminal domain** (**CTD**) of the RNA polymerase II molecule. Interestingly, only RNA polymerase II that has a nonphosphorylated CTD can initiate transcription but only an RNA polymerase II with a phosphorylated CTD can elongate RNA.

The RNA molecule made from a protein-coding gene by RNA polymerase II is called a **primary transcript**. Unlike the situation in prokaryotes, the primary transcript from a eukaryotic protein-coding gene is a precursor molecule, **pre-mRNA**, that needs extensive RNA processing in order to yield mature mRNA ready for translation. Several **RNA processing** reactions are involved: capping, 3′ cleavage and polyadenylation, and RNA splicing (see *Fig. 1* and

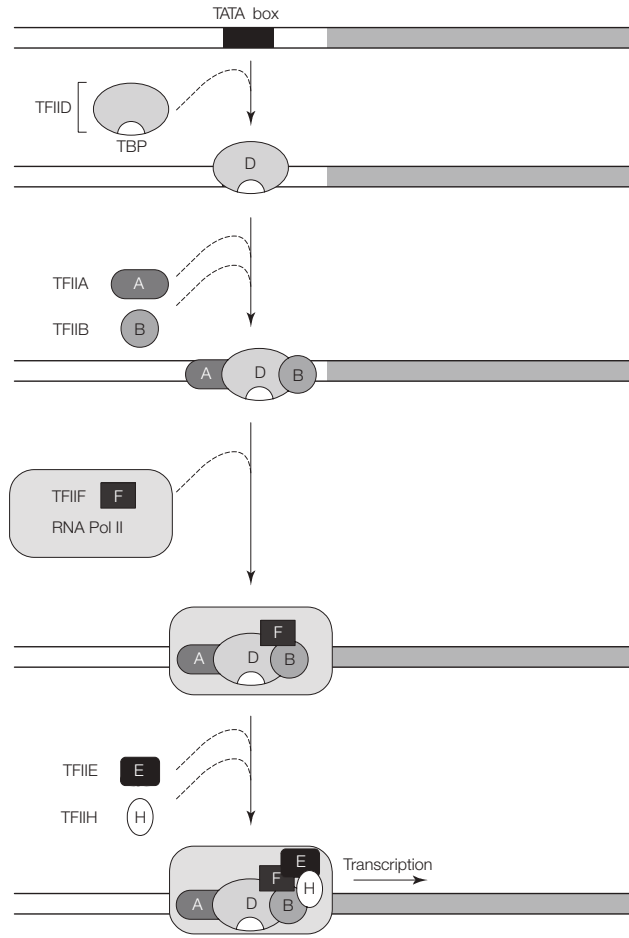

Fig. 3. Initiation of transcription by RNA polymerase II. TFIID binds to the TATA box followed in order by the binding of TFIIA, TFIIB and a pre-formed complex of TFIIF.RNA polymerase II. Subsequently TFIIE and TFIIH bind in order and transcription then starts about 25 bp downstream from the TATA box. Note that the placement of the various factors in this diagram is arbitrary; their exact positions in the complex are not known.

Topic G7). A number of the enzymes involved in the processing steps become associated with the RNA transcript through initial interaction with the CTD of RNA polymerase II and then transfer to the growing RNA molecule, so the CTD is also essential for triggering RNA processing. Thus, in eukaryotes, elongation by RNA polymerase II is tightly coupled to RNA processing.

Elongation of the RNA chain continues until termination occurs. Unlike RNA polymerase in prokaryotes, RNA polymerase II does not terminate transcription at a specific site but rather transcription can stop at varying distances downstream of the gene.

G6 Regulation of transcription by RNA Pol II

Key Notes

Mechanism of regulation

Many genes are active in all cells but some are transcribed only in specific cell types, at specific times and/or only in response to specific external stimuli. Transcriptional regulation occurs via transcription factors that bind to short control elements associated with the target genes and then interact with each other and with the transcription initiation complex to increase or decrease the rate of transcription of the target gene.

Regulatory elements

Many transcription factors bind to control elements located upstream within a few hundred base pairs of the protein-coding gene. The SP1 box and CAAT box are examples of such regulatory elements found upstream of most protein-coding genes, but some regulatory elements are associated with only a few genes and are responsible for gene-specific transcriptional regulation (e.g. hormone response elements).

Enhancers

Enhancers are positive transcriptional control elements typically 100–200 bp long that can be located either upstream or downstream of the target gene, are active in either orientation, and can activate transcription from the target gene even when located a long distance away (sometimes 10–50 kb). The transcription factors bound to these long-distance elements interact with the transcription initiation complex by looping out of the DNA.

Transcription factors have multiple domains

Transcription factors that increase the rate of transcription usually have at least two domains of protein structure, a DNA-binding domain that recognizes the specific DNA control element to bind to, and an activation domain that interacts with other transcription factors or the RNA polymerase. Many transcription factors operate as dimers (homodimers or heterodimers) held together via dimerization domains. Some transcription factors interact with small ligands via a ligand-binding domain.

DNA binding domains

DNA binding domains contain characteristic protein motifs. The helix-turn-helix motif contains two α-helices separated by a short β-turn. When the transcription factor binds to DNA, the recognition helix lies in the major groove of the DNA double helix. The second type of motif, the zinc finger, consists of a peptide loop with either two cysteines and two histidines (the C_2H_2 finger) or four cysteines (the C_4 finger) at the base of the loop that tetrahedrally coordinate a zinc ion. The zinc finger secondary structure is two β-strands and one α-helix. Transcription factors often contain several zinc fingers; in each case the α-helix binds in the major groove of the DNA double helix. Some transcription factors (e.g. bZIP proteins, basic HLH proteins) contain basic domains that interact with the target DNA.

Dimerization domains

A leucine zipper has a leucine every seventh amino acid and forms an α-helix with the leucines presented on the same side of the helix every second turn, giving a hydrophobic surface. Two transcription factor monomers can interact via the hydrophobic faces of their leucine zipper motifs to form a dimer. The helix-loop-helix (HLH) motif contains two α-helices separated by a nonhelical loop. The C-terminal α-helix has a hydrophobic face; two transcription factor monomers, each with an HLH motif, can dimerize by interaction between the hydrophobic faces of the two C-terminal α-helices.

Activation domains

No common structural motifs are known for the activation domains of transcription factors. Activation domains that are rich in acidic amino acids, glutamines or prolines have been reported.

Repressors

Repressor proteins that inhibit the transcription of specific genes in eukaryotes may bind either to control elements near to the target gene or to silencers that may be located a long distance away. The repressor may inhibit transcription of the target gene directly or may do so by interfering with the function of an activator protein required for efficient gene transcription.

Related topics

DNA structure (F1)
RNA structure (G1)
Transcription in prokaryotes (G2)
Operons (G3)
Transcription in eukaryotes: an overview (G4)
Transcription of protein-coding genes in eukaryotes (G5)
Processing of eukaryotic pre-mRNA (G7)
Ribosomal RNA (G8)
Transfer RNA (G9)

Mechanism of regulation

A number of protein-coding genes are active in all cells and are required for so-called 'house-keeping' functions, such as the enzymes of glycolysis (Topic J3), the citric acid cycle (Topic L1) and the proteins of the electron transport chain (Topic L2). However, some genes are active only in specific cell types and are responsible for defining the specific characteristics and function of those cells; for example immunoglobulin genes in lymphocytes, myosin in muscle cells. In addition, the proteins expressed by any given cell may change over time (for example during early development) or in response to external stimuli, such as hormones. Eukaryotic cells can regulate the expression of protein-coding genes at a number of levels but a prime site of regulation is transcription.

Transcriptional regulation in a eukaryotic cell (i.e. which genes are transcribed and at what rate) is mediated by **transcription factors** (other than the general transcription factors; see Topic G5) which recognize and bind to short regulatory DNA sequences (**control elements**) associated with the gene. These sequences are also called ***cis*-acting elements** (or simply ***cis*-elements**) since they are on the same DNA molecule as the gene being controlled (*cis* is Latin for 'on this side'). The protein transcription factors that bind to these elements are also known as ***trans*-acting factors** (or simply ***trans*-factors**) in that the genes encoding them can be on different DNA molecules (i.e. on different chromosomes). The transcription factors which regulate specific gene transcription do so by interacting with the proteins of the transcription initiation complex and

may either increase (activate) or decrease (repress) the rate of transcription of the target gene. Typically each protein-coding gene in a eukaryotic cell has several control elements in its promoter (*Fig. 1*) and hence is under the control of several transcription factors which interact with each other and with the transcription initiation complex by protein:protein interaction to determine the rate of transcription of that gene.

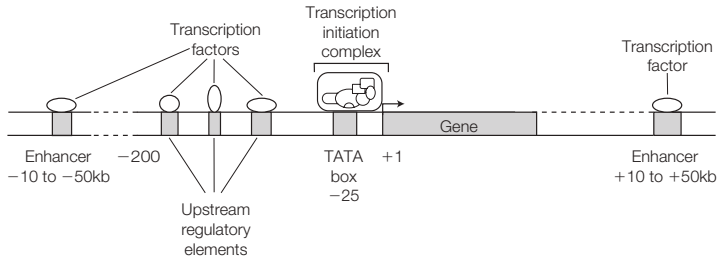

Fig. 1. Control regions that regulate transcription of a typical eukaryotic protein-coding gene. Although shown as distinct entities here for clarity, in vivo *the different regulatory proteins bound to the control elements and distant enhancers interact with each other and with the general transcription factors of the transcription initiation complex to modulate the rate of transcriptional initiation.*

Regulatory elements

Many transcription factors bind to control sequences (**regulatory elements**) within a few hundred base pairs of the protein-coding gene being regulated. Positive control elements that lie upstream of the gene, usually within 200 bp of the transcriptional start site (*Fig. 1*), are often called **upstream regulatory elements** (**UREs**) and function to increase the transcriptional activity of the gene well above that of the basal promoter. Some of these elements, for example the **SP1 box** and the **CAAT box**, are found in the promoters of many eukaryotic protein-coding genes; indeed genes often have several copies of one or both elements. The SP1 box has the core sequence GGGCGG, and binds **transcription factor SP1** which then interacts with general transcription factor TFIID (see Topic G5). In contrast, some upstream regulatory elements are associated only with a few specific genes and are responsible for limiting the transcription of those genes to certain tissues or in response to certain stimuli such as steroid hormones. For example, steroid hormones control metabolism by entering the target cell and binding to specific **steroid hormone receptors** in the cytoplasm. The binding of the hormone releases the receptor from an inhibitor protein that normally keeps the receptor in the cytoplasm. The hormone–receptor complex, now free of inhibitor, dimerizes and travels to the nucleus where it binds to a transcriptional control element, called a **hormone response element,** in the promoters of target genes. Then, like other transcription factors, the bound hormone–receptor complex interacts with the transcription initiation complex to increase the rate of transcription of the gene. The result is a hormone-specific transcription of a subset of genes in target cells that contain the appropriate steroid hormone receptor. Here, the hormone receptor is itself a transcription factor that is activated by binding the hormone ligand. Unlike steroid hormones, **polypeptide hormones**, such as **insulin** and **cytokines**, do not enter the target cell but instead bind to protein receptors located at the cell surface. The binding reaction triggers a cascade of protein activations, often involving protein phos-

phorylation, which relay the signal inside the cell (**signal transduction**). Again the response may be that specific transcription factors are activated and stimulate the transcription of selected genes, but here the activation is mediated via the signal transduction pathway and does not involve direct binding of the hormone or cytokine to the transcription factor. Many additional examples of transcriptional activation of specific genes by transcription factors exist in eukaryotes.

Enhancers

Although many positive control elements lie close to the gene they regulate, others can be located long distances away (sometimes 10–50 kb) either upstream or downstream of the gene (*Fig. 1*). A long-distance positive control sequence of this kind is called an **enhancer** if the transcription factor(s) that binds to it increases the rate of transcription. An enhancer is typically 100–200 bp long and contains several sequence elements that act together to give the overall enhancer activity. When they were first discovered, enhancers were viewed as a distinct class of control element in that they:

- can activate transcription over long distances;
- can be located upstream or downstream of the gene being controlled;
- are active in either orientation with respect to the gene.

However, it is now clear that some upstream promoter elements and enhancers show strong similarities physically and functionally so that the distinction is not as clear as was once thought. For enhancers located a long distance away from the gene being controlled, interaction between transcription factors bound to the enhancer and to promoter elements near the gene may occur by looping out of the DNA between the two sets of elements (*Fig. 2*).

Transcription factors have multiple domains

In most cases, the transcription factors in eukaryotes that bind to enhancer or promoter sequences are activator proteins that induce transcription. These proteins usually have at least two distinct domains of protein structure, a **DNA-binding domain** that recognizes the specific DNA sequence to bind to, and an **activation domain** responsible for bringing about the transcriptional activation by interaction with other transcription factors and/or the RNA polymerase molecule. Many transcription factors operate as dimers, either **homodimers** (identical subunits) or **heterodimers** (dissimilar subunits) with the subunits held together via **dimerization domains**. DNA binding domains and dimerization domains contain characteristic protein structures (**motifs**) that are described

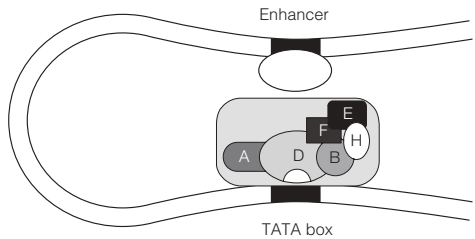

Fig. 2. Looping out of DNA allowing the interaction of enhancer-bound factor(s) with the transcription initiation complex.

below. Finally, some transcription factors (e.g. steroid hormone receptors) are responsive to specific small molecules (**ligands**) which regulate the activity of the transcription factor. In these cases, the ligand binds at a **ligand-binding domain**.

DNA binding domains

Helix-turn-helix

This motif consists of two α-helices separated by a short (four-amino acid) peptide sequence that forms a **β-turn** (*Fig. 3a*). When the transcription factor binds to DNA, one of the helices, called the **recognition helix**, lies in the major groove of the DNA double helix (*Fig. 3b*). The helix-turn-helix motif was originally discovered in certain transcription factors that play major roles in *Drosophila* early development. These proteins each contain a 60-amino acid DNA-binding region called a **homeodomain** (encoded by a DNA sequence called a **homeobox**). The homeodomain has four α-helices in which helices II and III are the classic helix-turn-helix motif. Since the original discovery, the helix-turn-helix motif has been found in a wide range of transcription factors, including many that have no role in development.

Zinc finger

Several types of zinc finger have been reported, two of which are the **C_2H_2 finger** and the **C_4 finger**. The C_2H_2 zinc finger is a loop of 12 amino acids with two cysteines and two histidines at the base of the loop that tetrahedrally coordinate a zinc ion (*Fig. 4a*). This forms a compact structure of two β-strands and one α-helix (*Fig. 4b*). The α-helix contains a number of conserved basic amino acids and interacts directly with the DNA, binding in the major groove of the double helix. Transcription factors that contain zinc fingers often contain several such motifs, arranged such that α-helix of each contacts the DNA. Indeed RNA polymerase III transcription factor A (**TFIIIA**; see Topic G8) contains nine zinc fingers! The SP1 transcription factor, which binds to the SP1 box, has three zinc fingers.

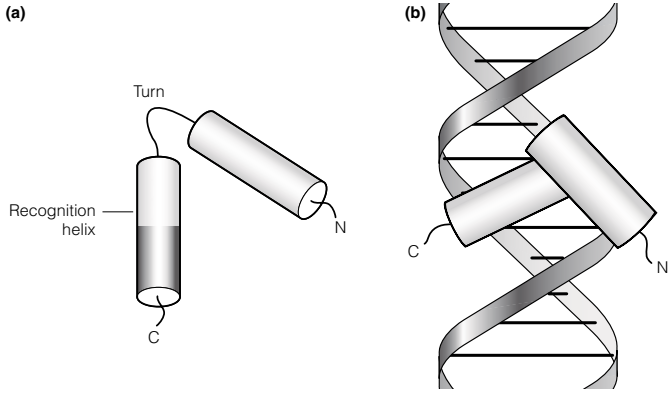

Fig. 3. (a) Helix-turn-helix motif of a DNA-binding protein; (b) binding of the helix-turn-helix to target DNA showing the recognition helix lying in the major groove of the DNA.

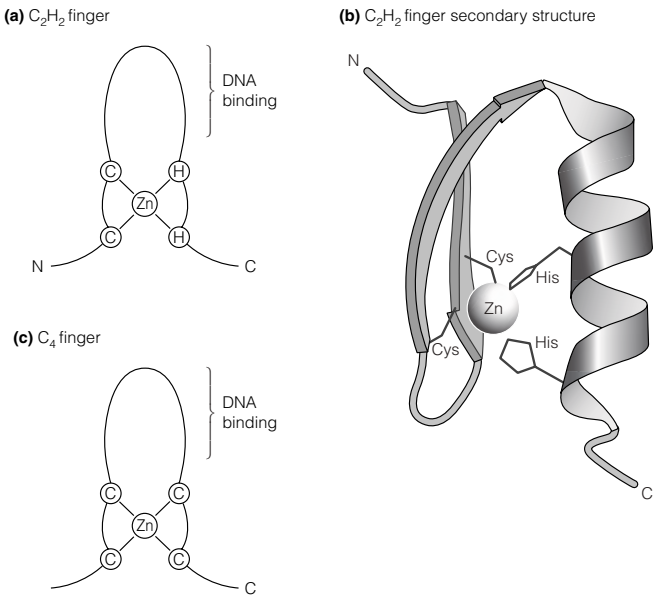

Fig. 4. (a) A C_2H_2 zinc finger; (b) C_2H_2 zinc finger secondary structure. From A. Travers, DNA–Protein Interactions*, Chapman & Hall, 1993. Reprinted with permission of A. Travers. (c) a C_4 zinc finger.*

The C_4 zinc finger is also found in a number of transcription factors, including steroid hormone receptor proteins. This motif forms a similar structure to that of C_2H_2 zinc finger but has four cysteines coordinated to the zinc ion instead of two cysteines and two histidines (see *Fig. 4c*).

Basic domains

DNA binding domains called **basic domains** (rich in basic amino acids), occur in transcription factors in combination with leucine zipper or helix-loop-helix (HLH) dimerization domains (see below). The combination of basic domain and dimerization domain gives these proteins their names of **basic leucine zipper proteins** (**bZIP**) or **basic HLH proteins**, respectively. In each case the dimerization means that two basic domains (one from each monomer) interact with the target DNA.

Dimerization domains

Leucine zippers

The leucine zipper motif contains a leucine every seventh amino acid in the primary sequence and forms an α-helix with the leucines presented on the same side of the helix every second turn, giving a hydrophobic surface. The transcription factor dimer is formed by the two monomers interacting via the hydrophobic faces of their leucine zipper motifs (*Fig. 5a*). In the case of bZIP proteins, each monomer also has a basic DNA binding domain located N-terminal to the leucine zipper. Thus the bZIP protein dimer has two basic domains. These actually face in opposite directions which allows them to bind to

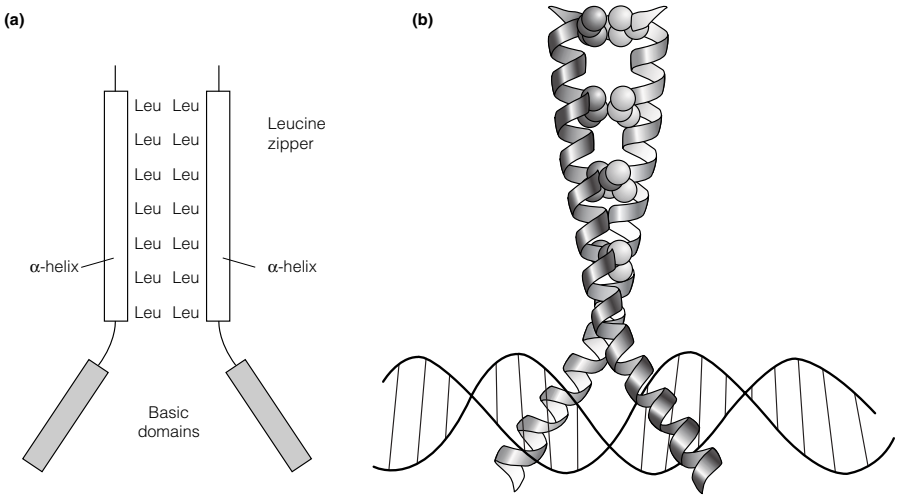

Fig. 5. (a) A bZIP protein dimer showing the leucine zipper dimerization domain and the two basic domains; (b) folded structure of a bZIP protein showing the basic domains binding in the major groove of the target DNA. Reprinted from A. Travers, DNA–Protein Interactions, *Chapman & Hall, 1993. With permission from A. Travers.*

DNA sequences that have inverted symmetry. They bind in the major groove of the target DNA (*Fig. 5b*). The leucine zipper domain also acts as the dimerization domain in transcription factors that use DNA binding domains other than the basic domain. For example, some homeodomain proteins, containing the helix-turn-helix motif for DNA binding, have leucine zipper dimerization domains. In all cases, the dimers that form may be homodimers or heterodimers.

Helix-loop-helix motif

The helix-loop-helix (HLH) dimerization domain is quite distinct from the helix-turn-helix motif described above (which is involved in DNA binding *not* dimerization) and must not be confused with it. The HLH domain consists of two α-helices separated by a nonhelical loop. The C-terminal α-helix has hydrophobic amino acids on one face. Thus two transcription factor monomers, each with an HLH motif, can dimerize by interaction between the hydrophobic faces of the two C-terminal α-helices. Like the leucine zipper (see above), the HLH motif is often found in transcription factors that contain basic DNA binding domains. Again, like the leucine zipper, the HLH motif can dimerize transcription factor monomers to form either homodimers or heterodimers. This ability to form heterodimers markedly increases the variety of active transcription factors that are possible and so increases the potential for gene regulation.

Activation domains

Unlike DNA binding domains and dimerization domains, no common structural motifs have yet been identified in the activation domains of diverse transcription factors. Some types of activation domain are as follows:

- **acidic activation domains** are rich in acidic amino acids (aspartic and glutamic acids). For example, mammalian glucocorticoid receptor proteins contain this type of activation domain;
- **glutamine-rich domains** (e.g. as in SP1 transcription factor);
- **proline-rich domains** (e.g. as in *c-jun* transcription factor).

Repressors

Gene repressor proteins that inhibit the transcription of specific genes in eukaryotes also exist. They may act by binding either to control elements within the promoter region near the gene or at sites located a long distance away from the gene, called **silencers**. The repressor protein may inhibit transcription directly. One example is the **mammalian thyroid hormone receptor** which, in the absence of thyroid hormone, represses transcription of the target genes. However, other repressors inhibit transcription by blocking activation. This can be achieved in one of several ways: by blocking the DNA binding site for an activator protein, by binding to and masking the activation domain of the activator factor, or by forming a non-DNA binding complex with the activator protein. Several examples of each mode of action are known.

G7 PROCESSING OF EUKARYOTIC PRE-mRNA

Key Notes

Overview

The primary RNA transcript from a protein-coding gene in a eukaryotic cell must be modified by several RNA processing reactions in order to become a functional mRNA molecule. The 5′ end is modified to form a 5′ cap structure. Most pre-mRNAs are then cleaved near the 3′ end and a poly(A) tail is added. Intron sequences are removed by RNA splicing.

5′ processing: capping

Immediately after transcription, the 5′ phosphate is removed, guanosyl transferase adds a G residue linked via a 5′–5′ covalent bond, and this is methylated to form a 7-methylguanosine (m^7G) cap (methylated in N–7 position of the base). The ribose residues of either the adjacent one or two nucleotides may also be methylated by methyl group addition to the 2′ OH of the sugar. The cap protects the 5′ end of the mRNA against ribonuclease degradation and also functions in the initiation of protein synthesis.

RNA splicing

Intron sequences are removed by RNA splicing that cleaves the RNA at exon–intron boundaries and ligates the ends of the exon sequences together. The cleavage sites are marked by consensus sequences that are evolutionarily conserved. In most cases the intron starts with GU and ends with AG, a polypyrimidine tract lies upstream of the AG, and a conserved branchpoint sequence is located about 20–50 nt upstream of the 3′ splice site. The splicing reaction involves two transesterification steps which ligate the exons together and release the intron as a branched lariat structure containing a 2′5′ bond with a conserved A residue in the branchpoint sequence. The RNA splicing reactions require snRNPs and accessory proteins that assemble into a spliceosome at the intron to be removed. The RNA components of the snRNPs are complementary to the 5′ and 3′ splice site sequences and to other conserved sequences in the intron and so can base pair with them. Some introns start with AU and end with AC, instead of GU and AG respectively. The splicing of these 'AT-AC introns' requires a different set of snRNPs than those used for splicing of the major form of intron, except both classes of intron use U5 snRNP. A few organisms can splice together exons from two different RNA molecules: trans-splicing. Some introns are self-splicing; the intron RNA sequence catalyzes its own excision without the involvement of a spliceosome.

3′ processing: cleavage and polyadenylation

Most pre-mRNA transcripts are cleaved post-transcriptionally near the 3′ end between a polyadenylation signal (5′-AAUAAA-3′) and a GU-rich (or U-rich) sequence further downstream. Specific proteins bind to these sequence elements to form a complex. One of the bound proteins, poly(A) polymerase, then adds a poly(A) tail of about 200 A residues to the new 3′ end of the RNA molecule and poly(A) binding protein molecules bind to this. The poly(A) tail protects the 3′ end of the final mRNA against nuclease degradation and also increases translational efficiency of the mRNA. Some pre-mRNAs (e.g. histone pre-mRNAs) are cleaved near the 3′ end but no poly(A) tail is added.

Alternative processing

Some pre-mRNAs contain more than one set of sites for 3′ end cleavage and polyadenylation, such that the use of alternative sites can lead to mRNA products that contain different 3′ noncoding regions (which may influence the lifetime of the mRNA) or have different coding capacities. Alternative splice pathways also exist whereby the exons that are retained in the final mRNA depends upon the pathway chosen, allowing several different proteins to be synthesized from a single gene.

RNA editing

The sequence of an mRNA molecule may be changed after synthesis and processing by RNA editing. Individual nucleotides may be substituted, added or deleted. In human liver, apolipoprotein B pre-mRNA does not undergo editing and subsequent translation yields apolipoprotein B100. In cells of the small intestine, RNA editing converts a single C residue in apolipoprotein B pre-mRNA to U, changing a codon for glutamine (CAA) to a termination codon (UAA). Translation of the edited mRNA yields the much shorter protein, apolipoprotein B48, with a restricted function in that it lacks a protein domain for receptor binding. Many other examples of editing occur, including trypanosome mitochondrial mRNAs, where RNA editing results in over half of the uridines in the final mRNA being acquired through the editing process.

Related topics

DNA structure (F1)
RNA structure (G1)
Transcription in prokaryotes (G2)
Operons (G3)
Transcription in eukaryotes: an overview (G4)
Transcription of protein-coding genes in eukaryotes (G5)
Regulation of transcription by RNA Pol II (G6)
Ribosomal RNA (G8)
Transfer RNA (G9)

Overview

In eukaryotes, the product of transcription of a protein-coding gene is pre-mRNA (see Topic G5) which requires processing to generate functional mRNA. Several processing reactions occur. Very soon after it has been synthesized by RNA polymerase II, the 5′ end of the **primary RNA transcript**, **pre-mRNA**, is modified by the addition of a **5′ cap** (a process known as **capping**). The primary RNA transcript that continues to be synthesized includes both coding (exon) and noncoding (intron) regions (see Topic G5, *Fig. 1*). The latter need to be removed and the exon sequences joined together by RNA splicing (Topic G5, *Fig. 1*) to generate a continuous RNA coding sequence ready for translation. The RNA splicing reactions are catalyzed by a large RNA–protein complex called a **spliceosome** (see below) that assembles on the primary RNA transcript as it is being synthesized, so RNA splicing occurs soon after RNA synthesis. Finally the 3′ ends of most (but not all) pre-mRNAs are modified by cleavage and the addition of about 200 A residues to form a **poly (A) tail** (a process called **polyadenylation**).

As described earlier (Topic G5), the C-terminal domain (CTD) of RNA polymerase II plays a role in transferring the 5′ capping, RNA splicing and key polyadenylation components to the growing RNA chain. Thus transcription and RNA processing in eukaryotes are tightly coupled events.

5′ processing: capping

Capping of pre-mRNA occurs immediately after synthesis and involves the addition of **7-methylguanosine** (**m^7G**) to the 5′ end (*Fig. 1*). To achieve this, the terminal 5′ phosphate is first removed by a phosphatase. **Guanosyl transferase**

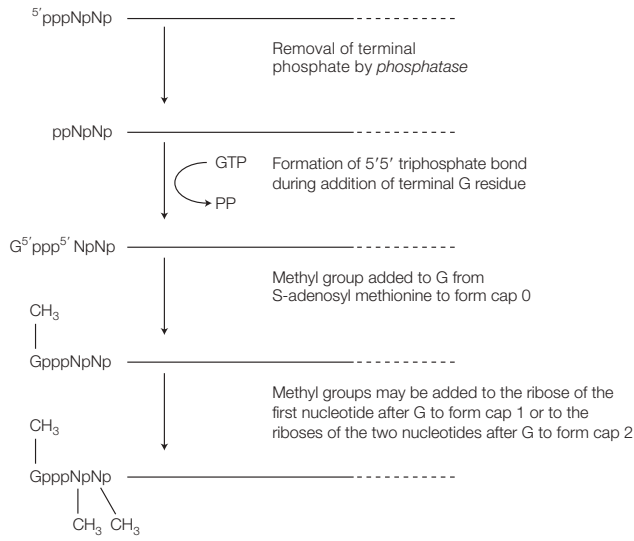

Fig. 1. Steps involved in the formation of the 5′ cap.

then catalyzes a reaction whereby the resulting diphosphate 5′ end attacks the α phosphorus atom of a GTP molecule to add a G residue in an unusual 5′5′ triphosphate link (*Fig. 1*). The G residue is then methylated by a **methyl transferase** adding a methyl group to the N–7 position of the guanine ring, using **S-adenosyl methionine** as methyl donor. This structure, with just the m^7G in position, is called a **cap 0 structure**. The ribose of the adjacent nucleotide (nucleotide 2 in the RNA chain) or the riboses of both nucleotides 2 and 3 may also be methylated to give **cap 1** or **cap 2** structures respectively. In these cases, the methyl groups are added to the 2′-OH groups of the ribose sugars (*Fig. 1*).

The cap protects the 5′ end of the primary transcript against attack by ribonucleases that have specificity for 3′5′ phosphodiester bonds and so cannot hydrolyze the 5′5′ bond in the cap structure. In addition, the cap plays a role in the initiation step of protein synthesis in eukaryotes. Only RNA transcripts from eukaryotic protein-coding genes become capped; prokaryotic mRNA and eukaryotic rRNA and tRNAs are uncapped.

RNA splicing

A key step in RNA processing is the precise removal of intron sequences and joining the ends of neighboring exons to produce a functional mRNA molecule, a process called **RNA splicing**. The exon–intron boundaries are marked by specific sequences (*Fig. 2*). In most cases, at the 5′ boundary between the exon and the intron (**the 5′ splice site**), the intron starts with the sequence GU and at the 3′ exon–intron boundary (**the 3′ splice site)** the intron ends with the sequence AG. Each of these two sequences lies within a longer consensus sequence. A **polypyrimidine tract** (a conserved stretch of about 11 pyrimidines) lies upstream of the AG at the 3′ splice site (*Fig. 2*). A key signal sequence is the **branchpoint sequence** which is located about 20–50 nt upstream of the 3′ splice

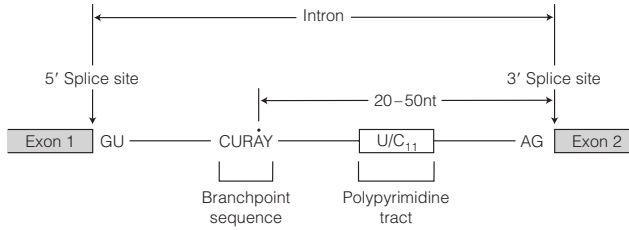

Fig. 2. Conserved sequences for RNA splicing. The residue marked as Å in the branchpoint sequence is the site of formation of the 2′5′ branch.

site. In vertebrates this sequence is 5′-CURAY-3′ where R = purine and Y = pyrimidine (in yeast this sequence is 5′-UACUAAC-3′).

RNA splicing occurs in two steps (*Fig. 3)*. In the first step, the 2′-OH of the A residue at the branch site (indicated as Å in *Fig. 2*) attacks the 3′5′ phosphodiester bond at the 5′ splice site causing that bond to break and the 5′ end of the intron to loop round and form an unusual 2′5′ bond with the A residue in the branchpoint sequence. Because this A residue already has 3′5′ bonds with its neighbors in the RNA chain, the intron becomes branched at this point to form what is known as a **lariat** intermediate (named as such since it resembles a cowboy's lasso). The new 3′-OH end of exon 1 now attacks the phosphodiester bond at the 3′ splice site causing the two exons to join and release the intron, still as a lariat. In each of the two splicing reactions, one phosphate–ester bond is exchanged for another (i.e. these are two **transesterification reactions**). Since the number of phosphate–ester bonds is unchanged, no energy (ATP) is consumed.

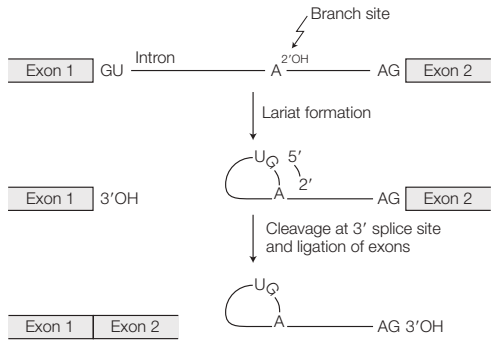

Fig. 3. The two steps of RNA splicing.

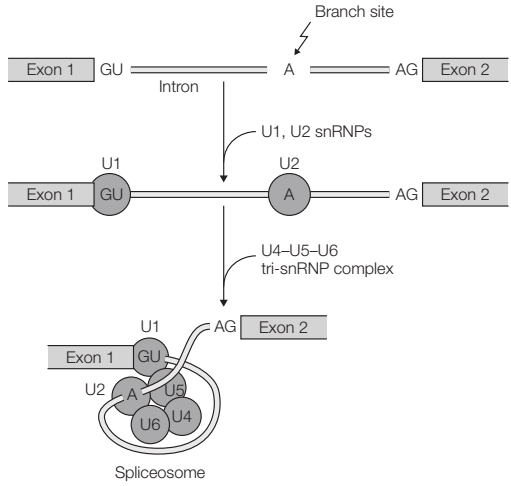

Fig. 4. Formation of the spliceosome.

RNA splicing requires the involvement of several **small nuclear RNAs** (**snRNAs**) each of which is associated with several proteins to form a small nuclear ribonucleoprotein particle or **snRNP** (pronounced '**snurp**'). Because snRNAs are rich in U residues, they are named U1, U2, etc. The RNA components of the snRNPs have regions that are complementary to the 5′ and 3′ splice site sequences and to other conserved sequences in the intron and so can base pair with them. The U1 snRNP binds to the 5′ splice site and U2 snRNP binds to the branchpoint sequence (*Fig. 4*). A **tri-snRNP complex** of U4, U5 and U6 snRNPs then binds, as do other accessory proteins, so that a multicomponent complex (called a **spliceosome**) is formed at the intron to be removed and causes the intron to be looped out (*Fig. 4*). Thus, through interactions between the snRNAs and the pre-mRNA, the spliceosome brings the upstream and downstream exons together ready for splicing. The spliceosome next catalyzes the two-step splicing reaction to remove the intron and ligate together the two exons. The spliceosome then dissociates and the released snRNPs can take part in further splicing reactions at other sites on the pre-mRNA.

Although the vast majority of pre-mRNA introns start with GU at the 5′ splice site and end with AG at the 3′ splice site, some introns (possibly as many as 1%) have different splice site consensus sequences. In these cases, the intron starts with AU and ends with AC instead of GU and AG, respectively (*Fig. 5*). Since RNA splicing involves recognition of the splice site consensus sequences by key

Fig. 5. Comparison of the conserved splice site sequences of the majority of introns (top diagram) with those for AT-AC introns (bottom diagram).

snRNPs (see above), and since these sequences are different in the minor intron class, U1, U2, U4, U6 snRNPs do not take part in splicing these so-called '**AT-AC introns**' (the AT-AC refers of course to the corresponding DNA sequence). Instead, U11, U12, $U4_{atac}$ and $U6_{atac}$ snRNPs are involved, replacing the roles of U1, U2, U4 and U6 respectively, and assemble to form the '**AT-AC spliceosome**'. U5 snRNP is required for splicing both classes of intron.

A few organisms, such as nematodes and trypanosomes, are able to splice together exons from two different RNA molecules, a process called **trans-splicing**. In this context, the more usual splicing together of two exons in the same RNA molecule would be **cis-splicing**.

Some **self-splicing introns** are also known, for example, *Tetrahymena* rRNA (see Topic G8) and some mitochondrial and chloroplast mRNAs. In these cases, the intron RNA sequence catalyzes its own cleavage out of the RNA precursor without the need for a spliceosome. Such catalytic RNA molecules are called **ribozymes** (a name that is clearly fashioned on 'enzymes', i.e. protein catalysts). The chemical similarity of some of these self-splicing reactions with the reactions that occur during spliceosome-mediated splicing has led to a realization of the central role of RNA catalysis in the latter. Spliceosome-mediated splicing probably evolved from self-splicing entities, with snRNAs having roles not only in recognition of splice sites but also in the catalytic reactions of spliceosome-mediated splicing. In particular, the structure formed by U2 snRNA base paired to U6 snRNA probably forms the catalytic center of the spliceosome.

3′ processing: cleavage and polyadenylation

Most eukaryotic pre-mRNAs undergo polyadenylation which involves cleavage of the RNA at its 3′ end and the addition of about 200A residues to form a poly(A) tail. The cleavage and polyadenylation reactions require the existence of a **polyadenylation signal sequence** (**5′-AAUAAA-3′**) located near the 3′ end of the pre-mRNA followed by a sequence **5′-YA-3′** (where Y = a pyrimidine), often 5′- CA-3′, in the next 11–20 nt (*Fig. 6*). A **GU-rich sequence** (or U-rich sequence) is also usually present further downstream. After these sequence elements have been synthesized, two multisubunit proteins called **CPSF (cleavage and polyadenylation specificity factor)** and **CStF (cleavage stimulation factor F)** are transferred from the CTD of RNA polymerase II to the RNA molecule and bind to the sequence elements. A protein complex is formed which includes additional **cleavage factors** and an enzyme called **poly(A) polymerase (PAP)**. This complex cleaves the RNA between the AAUAAA sequence and the GU-rich sequence (*Fig. 6*). Poly(A) polymerase then adds about 200A residues to the new 3′ end of the RNA molecule using ATP as precursor. As it is made, the poly(A) tail immediately binds multiple copies of a poly(A) binding protein. The poly(A) tail protects the 3′ end of the final mRNA against ribonuclease digestion and hence stabilizes the mRNA. In addition, it increases the efficiency of translation of the mRNA. However, some mRNAs, notably histone pre-mRNAs, lack a

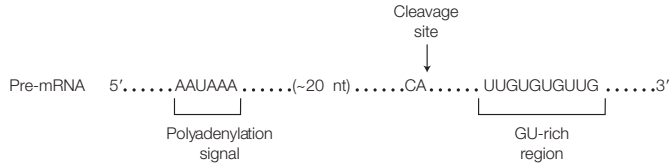

Fig. 6. Conserved sequences for polyadenylation.

poly(A) tail. Nevertheless, histone pre-mRNA is still subject to 3′ processing. It is cleaved near the 3′ end by a protein complex that recognizes specific signals, one of which is a stem-loop structure, to generate the 3′ end of the mature mRNA molecule.

Alternative processing

Alternative polyadenylation sites

Certain pre-mRNAs contain more than one set of signal sequences for 3′ end cleavage and polyadenylation. In some cases, the location of the alternative polyadenylation sites is such that, depending on the site chosen, particular exons may be lost or retained in the subsequent splicing reactions (*Fig. 7*). Here the effect is to change the coding capacity of the final mRNA so that different proteins are produced depending on the polyadenylation site used. In other cases, the alternative sites both lie within the 3′ noncoding region of the pre-mRNA so that the same coding sequences are included in the final mRNA irrespective of which site is used but the 3′ noncoding region can vary. Since the 3′ noncoding sequence may contain signals to control mRNA stability, the choice of polyadenylation site in this situation can affect the lifetime of the resulting mRNA.

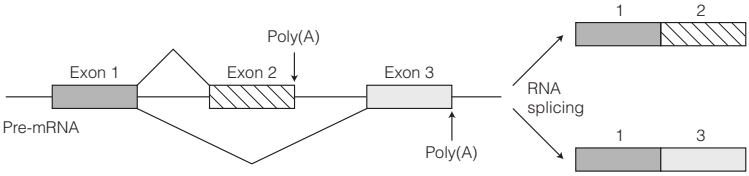

Fig. 7. Use of alternative polyadenylation sites.

Alternative splicing

Many cases are now known where different tissues splice the primary RNA transcript of a single gene by alternative pathways, where the exons that are lost and those that are retained in the final mRNA depend upon the pathway chosen (*Fig. 8*). Presumably some tissues contain regulatory proteins that promote or suppress the use of certain splice sites to direct the splicing pathway selected. These **alternative splicing pathways** are very important since they allow cells to synthesize a range of functionally distinct proteins from the primary transcript of a single gene.

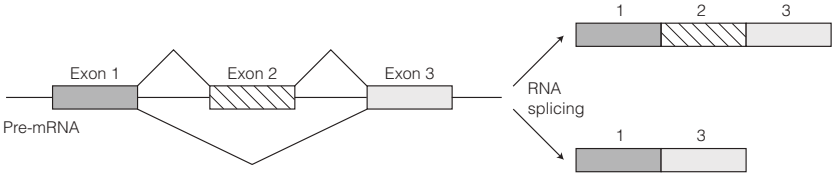

Fig. 8. Alternative RNA splicing pathways. In the simple example shown, the transcript can be spliced by alternative pathways leading to two mRNAs with different coding capacities, i.e. exons 1, 2 and 3 or just exons 1 and 3. For genes containing many exons, a substantial number of alternative splice pathways may exist which are capable of generating many possible mRNAs from the single gene.

RNA editing

RNA editing is the name given to several reactions whereby the nucleotide sequence on an mRNA molecule may be changed by mechanisms other than RNA splicing. Individual nucleotides within the mRNA may be changed to other nucleotides, deleted entirely or additional nucleotides inserted. The effect of RNA editing is to change the coding capacity of the mRNA so that it encodes a different polypeptide than that originally encoded by the gene. An example of RNA editing in humans is **apolipoprotein B mRNA**. In liver, the mRNA does not undergo editing and the protein produced after translation is called **apolipoprotein B100** (*Fig. 9a*). In cells of the small intestine, RNA editing (*Fig. 9b*) causes the conversion of a single C residue in the mRNA to U and, in so doing, changes a codon for glutamine (CAA) to a termination codon (UAA). Subsequent translation of the edited mRNA yields the much shorter **apolipoprotein B48** (48% of the size of apolipoprotein B100). This is not a trivial change; apolipoprotein B48 lacks a protein domain needed for receptor binding which apolipoprotein B100 possesses and hence the functional activities of the two proteins are different. Many other cases of RNA editing are also known. Trypanosome mitochondrial mRNAs, for example, undergo extensive RNA editing which results in over half of the uridines in the final mRNA being acquired through the editing process.

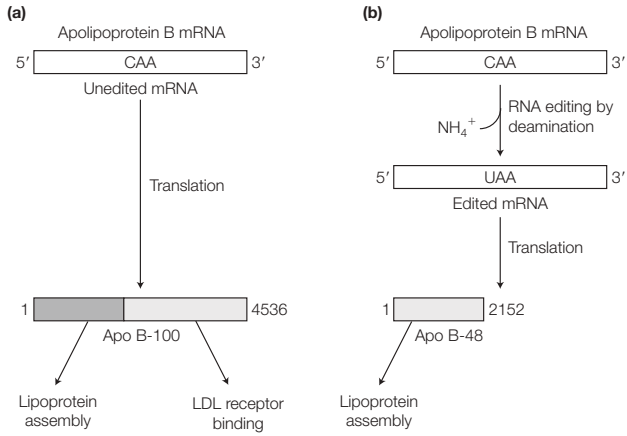

Fig. 9. RNA editing. (a) Unedited apolipoprotein B mRNA is translated to yield ApoB-100, a 4536-amino acid long polypeptide with structural domains for lipoprotein assembly and receptor binding functions; (b) translation of the edited mRNA yields the shorter ApoB-48 which lacks the receptor binding domain.

G8 RIBOSOMAL RNA

Key Notes

Ribosomes

A prokaryotic 70S ribosome comprises two subunits (50S and 30S). The 50S subunit has 23S and 5S rRNAs complexed with 34 polypeptides whereas the 30S subunit contains 16S rRNA and 21 polypeptides. A eukaryotic 80S ribosome comprises two subunits (60S and 40S). The 60S subunit has 28S, 5.8S and 5S rRNAs complexed with approx. 49 polypeptides whereas the 40S subunit contains 18S rRNA and about 33 polypeptides. The 3D structure of bacterial ribosomes shows that it is the rRNAs which fold and base pair with each other to form the overall structure of the ribosome, with the proteins located peripherally. In addition, the A, P and E sites, and even the catalytic site for peptide bond formation, are formed by rRNAs which therefore have a catalytic function as well as structural roles.

Transcription and processing of prokaryotic rRNA

E. coli has seven rRNA transcription units, each containing one copy each of the 23S, 16S and 5S rRNA genes as well as one to four tRNA genes. Transcription produces a 30S pre-rRNA transcript. This folds up to form stem-loop structures, ribosomal proteins bind, and a number of nucleotides become methylated. The modified pre-rRNA transcript is then cleaved at specific sites by RNase III and the ends are trimmed by ribonucleases M5, M6 and M23 to release the mature rRNAs.

Synthesis of eukaryotic 28S, 18S and 5.8S rRNA

The 28S, 18S and 5.8S rRNA genes are present as multiple copies clustered together as tandem repeats. These rRNA transcription units are transcribed, in the nucleolus, by RNA Pol I. The promoter contains a core element that straddles the transcriptional start site and an upstream control element (UCE) about 50–80 bp in size, located at about position −100. Transcription factors, one of which is TATA binding protein (TBP), bind to these control elements and, together with RNA Pol I, form a transcription initiation complex. Transcription produces a 45S pre-rRNA which has external transcribed spacers (ETSs) at the 5′ and 3′ ends and internal transcribed spacers (ITSs) internally separating rRNA sequences. The pre-rRNA folds up to form a defined secondary structure with stem-loops, ribosomal proteins bind to selected sequences, and multiple methylation and isomerization reactions (of uridine to pseudouridine) occur at specific sites, guided by interaction of the pre-rRNA with snoRNAs (as snoRNPs). The 45S pre-rRNA molecule is then cleaved, releasing 32S and 20S precursor rRNAs that are processed further to generate mature 28S, 18S and 5.8S rRNAs.

Ribozymes

In *Tetrahymena*, the pre-rRNA molecule contains an intron that is removed by self-splicing (in the presence of guanosine, GMP, GDP or GTP) without the need for involvement of any protein. This was the first ribozyme discovered but many have since been reported.

Synthesis of eukaryotic 5S rRNA

Eukaryotic cells contain multiple copies of the 5S rRNA gene. Unlike other eukaryotic rRNA genes, the 5S rRNA genes are transcribed by RNA Pol III. Two control elements, an A box and a C box, lie downstream of the transcriptional start site. The C box binds TFIIIA which then recruits TFIIIC. TFIIIB now binds and interacts with RNA Pol III to form the transcription initiation complex. Transcription produces a mature 5S rRNA that requires no processing.

Related topics

DNA structure (F1)
RNA structure (G1)
Transcription in prokaryotes (G2)
Operons (G3)
Transcription in eukaryotes: an overview (G4)
Transcription of protein-coding genes in eukaryotes (G5)
Regulation of transcription by RNA Pol II (G6)
Processing of eukaryotic pre-mRNA (G7)
Transfer RNA (G9)

Ribosomes

Each ribosome consists of two subunits, a small subunit and a large subunit, each of which is a multicomponent complex of **ribosomal RNAs** (**rRNAs**) and **ribosomal proteins** (*Fig. 1*). One way of distinguishing between particles such as ribosomes and ribosomal subunits is to place the sample in a tube within a centrifuge rotor and spin this at very high speed. This causes the particles to sediment to the tube bottom. Particles that differ in mass, shape and/or density sediment at different velocities (sedimentation velocities). Thus a particle with twice the mass of another will always sediment faster provided both particles have the same shape and density. The sedimentation velocity of any given particle is also directly proportional to the gravitational forces (the centrifugal field) experienced during the centrifugation, which can be increased simply by spinning the rotor at a higher speed. However, it is possible to define a **sedimentation coefficient** that depends solely on the size, shape and density of the particle and is independent of the centrifugal field. Sedimentation coefficients are usually measured in **Svedberg units** (**S**). A prokaryotic ribosome has a sedimentation coefficient of 70S whereas the large and small subunits have sedimentation coefficients of 50S and 30S, respectively (note that S values are not additive). The 50S subunit contains two rRNAs (23S and 5S) complexed with 34 polypeptides whereas the 30S subunit contains 16S rRNA and 21 polypeptides (*Fig. 1*). In eukaryotes the ribosomes are larger and more complex; the ribosome monomer is 80S and consists of 60S and 40S subunits. The 60S subunit contains three rRNAs (28S, 5.8S and 5S) and about 49 polypeptides and the 40S subunit has 18S rRNA and about 33 polypeptides (*Fig. 1*). However, despite this extra complexity, the overall structure and function of eukaryotic ribosomes is very similar to those from bacteria. In each case, about two thirds of the structure is rRNA and one third is protein.

A wide range of studies have built up a detailed picture of the fine structure of ribosomes, mapping the location of the various RNA and protein components and their interactions. The overall shape of a 70S ribosome, gained through electron microscopy studies, is shown in *Fig. 2*. The three-dimensional structure of bacterial ribosomal subunits, determined only a few years ago, shows that the various rRNA are tightly folded and base pair extensively with each other to form the core of the ribosomal subunits with the proteins restricted to the

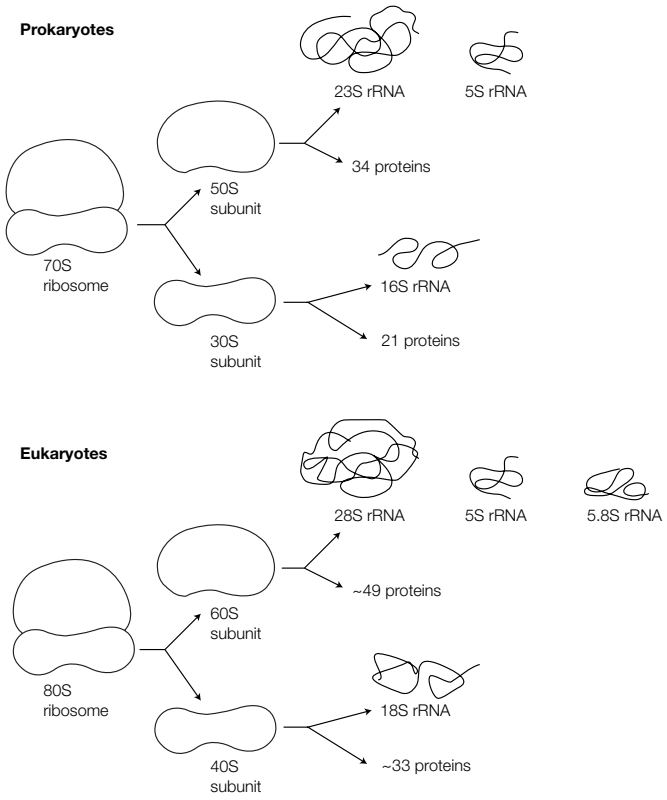

Fig. 1. Composition of ribosomes in prokaryotic and eukaryotic cells.

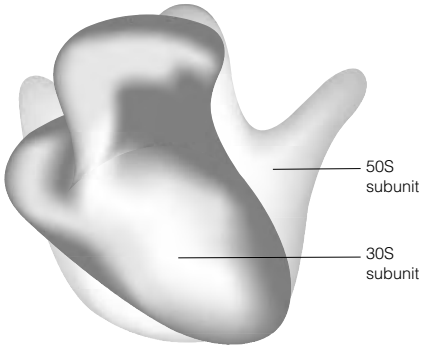

Fig. 2. The prokaryotic 70S ribosome.

surface and filling gaps between RNA folds. Not only are the rRNAs largely responsible for the overall structure of the ribosome but they also form the three main binding sites (the A, P and E sites) involved in protein synthesis (see Topic H2). In addition, the 23S rRNA, rather than a protein, forms the catalytic site for peptide bond formation (see Topic H2). Therefore, protein synthesis involves rRNAs acting as ribozymes.

Transcription and processing of prokaryotic rRNA

In *E. coli* there are seven rRNA transcription units scattered throughout the genome, each of which contains one copy of each of the 23S, 16S and 5S rRNA genes and one to four copies of various tRNA genes (*Fig. 3*). This gene assembly is transcribed by the single prokaryotic RNA polymerase to yield a single **30S pre-rRNA transcript** (about 6000 nt in size). This arrangement ensures that stoichiometric amounts of the various rRNAs are synthesized for ribosome assembly. Following transcription, the 30S pre-rRNA molecule forms internal base paired regions to give a series of stem-loop structures and ribosomal proteins bind to form a **ribonucleoprotein** (**RNP) complex**. A number of the nucleotides in the folded pre-rRNA molecule are now methylated, on the ribose moieties, using S-adenosylmethionine as the methyl donor. Next the pre-rRNA molecule is cleaved at specific sites by **RNase III** to release precursors of the 23S, 16S and 5S rRNAs. The precursors are then trimmed at their 5′ and 3′ ends by ribonucleases **M5**, **M16** and **M23** (which act on the 5S, 16S and 23S precursor rRNAs respectively) to generate the mature rRNAs.

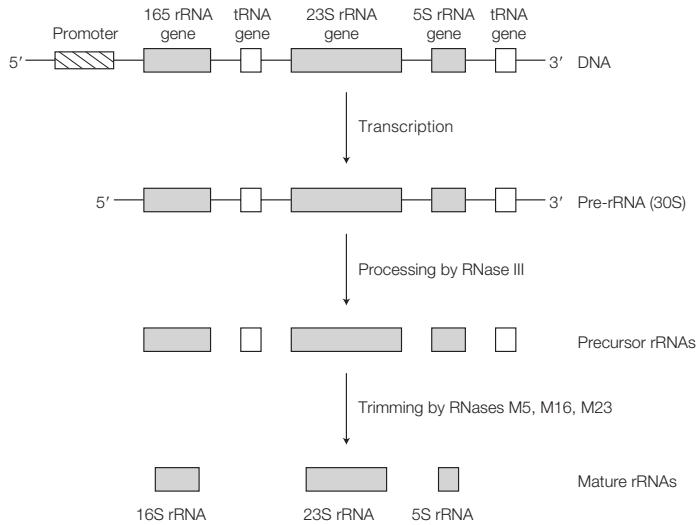

Fig. 3. Transcription and processing of prokaryotic rRNA.

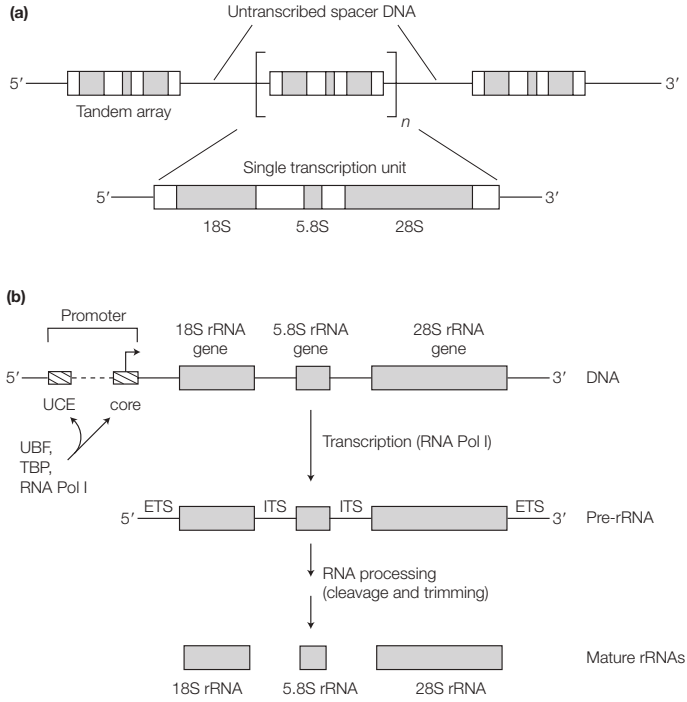

Fig. 4. (a) rRNA transcription units; (b) transcription of a single transcription unit by RNA Pol I and processing of pre-rRNA.

Synthesis of eukaryotic 28S, 18S and 5.8S rRNA

In eukaryotes, the genes for 28S, 18S and 5.8S rRNA are typically **clustered** together and **tandemly repeated** in that one copy each of 18S, 5.8S and then 28S genes occur, followed by untranscribed spacer DNA, then another set of 18S, 5.8S and 28S genes occur and so on (*Fig. 4a*). In humans, there are about 200 copies of this rRNA transcription unit arranged as five clusters of about 40 copies on separate chromosomes. These rRNA transcription units are transcribed by **RNA polymerase I** (**RNA Pol I**) in a region of the nucleus known as the **nucleolus** (see Topic A2). The nucleolus contains loops of DNA extending from each of the rRNA gene clusters on the various chromosomes and hence each cluster is called a **nucleolar organizer**.

The rRNA promoter consists of a **core element** which straddles the transcriptional start site (designated as position +1) from residues −31 to +6 plus an **upstream control element** (**UCE**) about 50–80 bp in size and located about 100 bp upstream from the start site (i.e. at position –100; *Fig. 4b*). A transcription factor called **upstream binding factor** (**UBF**) binds both to the UCE as well as to a region next to and overlapping with the core element. Interestingly, **TATA box binding protein** (**TBP**; see Topic G5), also binds to the rRNA promoter (in fact, TBP is required for initiation by all three eukaryotic RNA polymerases). The UBF and TBP transcription factors interact with each other and with RNA Pol I to form a **transcription initiation complex**. The RNA Pol I then transcribes

the whole transcription unit of 28S, 18S and 5.8S genes to synthesize a single large pre-rRNA molecule (*Fig. 4b*).

In humans, the product of transcription is a **45S pre-rRNA** which has non-rRNA **external transcribed spacers** (**ETSs**) at the 5′ and 3′ ends and non-rRNA **internal transcribed spacers** (**ITSs**) internally separating the rRNA sequences (*Fig. 4b*). This 45S molecule folds up to form a defined secondary structure with stem-loops, ribosomal proteins bind to selected sequences, methylation of ribose moieties occurs (at over 100 nucleotides) and more than 100 uridine residues are modified to pseudouridine (ψ). The 45S pre-rRNA molecule is then cleaved, first in the ETSs and then in the ITSs, to release precursor rRNAs which are cleaved further and trimmed to release the mature 28S, 18S and 5.8S rRNAs (*Fig. 4b*).

In eukaryotes, selection of the sites in pre-rRNA that will be methylated depends upon small RNAs found in the nucleolus called **small nucleolar RNAs** (**snoRNAs**) that exist in ribonucleoprotein complexes called **snoRNPs**. The snoRNAs contain long regions (10–21 nt) that are complementary to specific regions of the pre-rRNA molecule, form base pairs with the pre-rRNA at these sites and then guide where methylation of specific ribosome residues (2′O methylation) and isomerization of uridine to pseudouridine will occur.

Ribozymes

In at least one eukaryote, *Tetrahymena*, the pre-rRNA molecule contains an intron. Removal of the intron during processing of the pre-rRNA does not require the assistance of any protein! Instead, in the presence of guanosine, GMP, GDP or GTP, the intron excises itself, a phenomenon known as **self-splicing**. This was the first demonstration of **ribozymes**, that is, **catalytic RNA** molecules that catalyze specific reactions. Self-splicing introns have also been discovered in some eukaryotic mRNAs and even peptidyl transferase, a key activity in protein synthesis, is now known to be a ribozyme (see Topic H2).

Synthesis of eukaryotic 5S rRNA

In eukaryotes, the 5S rRNA gene is also present in multiple copies (2000 in human cells, all clustered together at one chromosomal site). Unlike other eukaryotic rRNA genes, the 5S rRNA genes are transcribed by **RNA polymerase III** (**RNA Pol III**). The promoters of tRNA genes, which are also transcribed by RNA Pol III, contain control elements called the A box and B box located downstream of the transcriptional start site (see Topic G9). A similar situation exists for 5S rRNA genes in that the promoter has two control elements located downstream of the transcriptional start site, an **A box** and a **C box** (*Fig. 5*). The C box binds **transcription factor IIIA** (**TFIIIA**) which then in turn interacts with **TFIIIC** to cause it to bind, a process which probably also involves recognition of the A box. Once TFIIIC has bound, **TFIIIB** binds and interacts with RNA Pol III, causing that to bind also to form the **transcription initiation complex**. One of the three subunits of TFIIB is **TATA box binding protein** (**TBP**; see Topic G5), the transcription factor required for transcription by all three eukaryotic RNA polymerases. Following transcription, the 5S rRNA transcript requires no processing. It migrates to the nucleolus and is recruited into ribosome assembly.

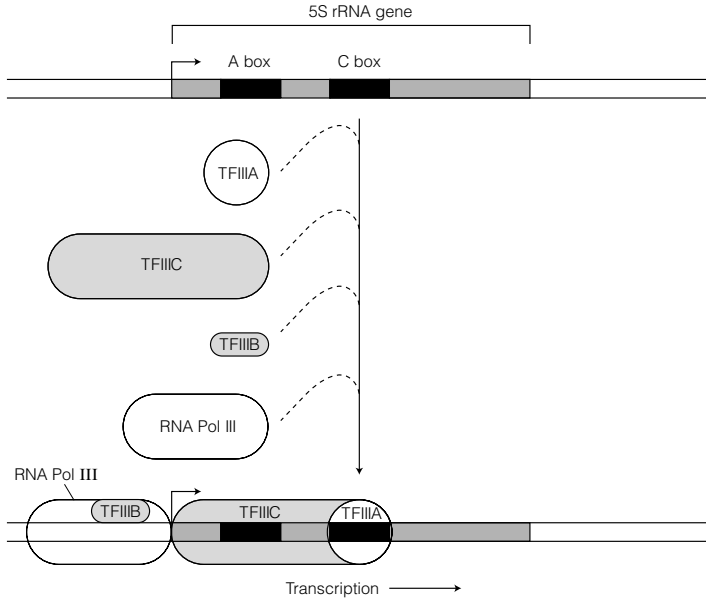

Fig. 5. Initiation of transcription of a 5S rRNA gene by RNA Pol III.

G9 TRANSFER RNA

Key Notes

tRNA structure

Each tRNA has a cloverleaf secondary structure containing an anticodon arm, a D (or DHU) arm, a T or TΨC arm, and an amino acid acceptor stem to which the relevant amino acid becomes covalently bound, at the 3′ OH group. Some tRNAs also have a variable (or optional) arm. The three-dimensional structure is more complex because of additional interactions between the nucleotides.

Transcription and processing of tRNA in prokaryotes

E. coli contains clusters of up to seven tRNA genes separated by spacer regions, as well as tRNA genes within ribosomal RNA transcription units. Following transcription, the primary RNA transcript folds up into specific stem-loop structures and is then processed by ribonucleases D, E, F and P in an ordered series of reactions to release the individual tRNA molecules.

Transcription and processing of tRNA in eukaryotes

In eukaryotes, tRNA genes are present as multiple copies and are transcribed by RNA Pol III. Several tRNA genes may be transcribed to yield a single pre-tRNA that is then processed to release individual tRNAs. The tRNA promoter includes two control elements, called the A box and the B box, located within the tRNA gene itself and hence downstream of the transcriptional start site. Transcription initiation requires transcription factor IIIC (TFIIIC), which binds to the A and B boxes and TFIIIB that binds upstream of the A box. The primary RNA transcript folds up into stem-loop structures and non-tRNA sequence is removed by ribonuclease action. Unlike prokaryotes, in eukaryotes the CCA sequence at the 3′ end of the tRNA is added after the trimming reactions (by tRNA nucleotidyl transferase). Unlike prokaryotes, pre-tRNA molecules in eukaryotes may also contain a short intron in the loop of the anticodon arm. The intron is removed by tRNA splicing reactions involving endonuclease cleavage at both ends of the intron and then ligation of the cut ends of the tRNA.

Modification of tRNA

Following synthesis, nucleotides in the tRNA molecule may undergo modification to create unusual nucleotides such as 1-methylguanosine (m^1G), pseudouridine (ψ), dihydrouridine (D), inosine (I) and 4-thiouridine (S^4U).

Related topics

DNA structure (F1)
RNA structure (G1)
Transcription in prokaryotes (G2)
Operons (G3)
Transcription in eukaryotes: an overview (G4)
Transcription of protein-coding genes in eukaryotes (G5)
Regulation of transcription by RNA Pol II (G6)
Processing of eukaryotic pre-mRNA (G7)
Ribosomal RNA (G8)

tRNA structure

Transfer RNA (tRNA) molecules play an important role in protein synthesis (Topics H2 and H3). Each tRNA becomes covalently bonded to a specific amino acid to form **aminoacyl-tRNA** which recognizes the corresponding codon in mRNA and ensures that the correct amino acid is added to the growing polypeptide chain. The tRNAs are small molecules, only 74–95 nt long, which form distinctive **cloverleaf** secondary structures (*Fig. 1a*) by internal base pairing. The stem-loops of the cloverleaf are known as **arms**:

- the **anticodon arm** contains in its **loop** the three nucleotides of the **anticodon** which will form base pairs with the complementary codon in mRNA during translation;
- the **D** or **DHU arm** (with its **D loop**) contains **dihydrouracil**, an unusual pyrimidine;
- the **T** or **TΨC arm** (with its **T loop**) contains another unusual base, **pseudouracil** (denoted Ψ) in the sequence TΨC;
- Some tRNAs also have **a variable arm** (**optional arm**) which is 3–21 nt in size.

The other notable feature is the **amino acid acceptor stem**. This is where the amino acid becomes attached, at the 3′ OH group of the 3′-CCA sequence.

The three-dimensional structure of tRNA (*Fig. 1b*) is even more complex because of additional interactions between the various units of secondary structure.

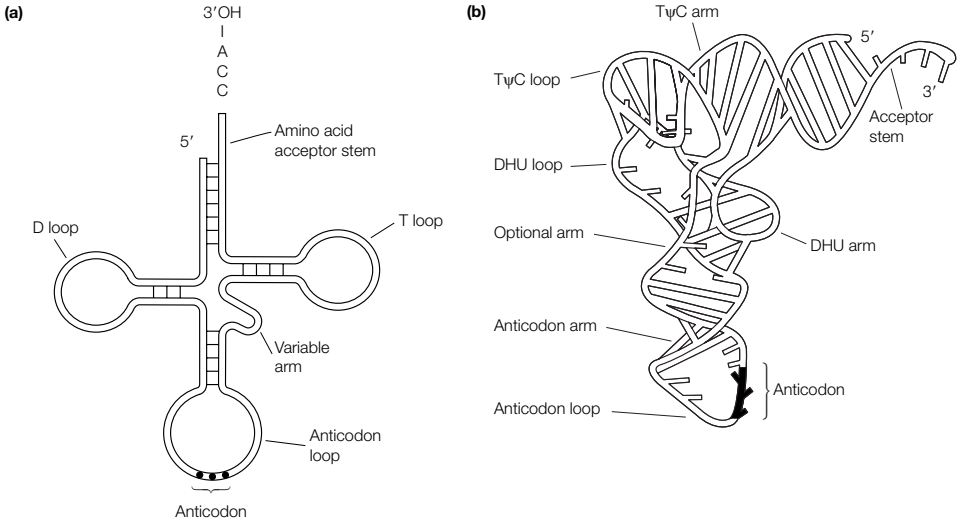

Fig. 1. (a) Cloverleaf secondary structure of tRNA; (b) tertiary structure of tRNA (from Genetics: a Molecular Approach, *second edition, T.A. Brown, Kluwer Academic Publishers, with permission).*

Transcription and processing of tRNA in prokaryotes

The rRNA transcription units in *E. coli* contain some tRNA genes that are transcribed and processed at the time of rRNA transcription (Topic G8). Other tRNA genes occur in clusters of up to seven tRNA sequences separated by spacer regions. Following transcription by the single prokaryotic RNA polymerase, the

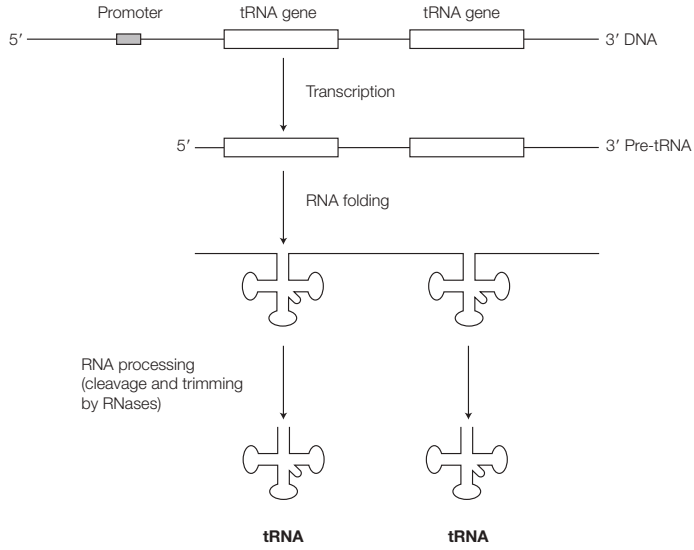

Fig. 2. Transcription and processing of tRNA in prokaryotes.

primary RNA transcript folds up into the characteristic stem-loop structures (*Fig. 2*) and is then processed in an ordered series of cleavages by ribonucleases (RNases) which release and trim the tRNAs to their final lengths. The cleavage and trimming reactions at the 5′ and 3′ ends of the precursor tRNAs involves **RNases D**, **E**, **F** and **P**. RNases E, F and P are **endonucleases**, cutting the RNA internally, whilst RNase D is an **exonuclease**, trimming the ends of the tRNA molecules.

Transcription and processing of tRNA in eukaryotes

In eukaryotes, the tRNA genes exist as multiple copies and are transcribed by **RNA polymerase III** (**RNA Pol III**). As in prokaryotes, several tRNAs may be transcribed together to yield a single **pre-tRNA molecule** that is then processed to release the mature tRNAs. The promoters of eukaryotic tRNA genes are unusual in that the transcriptional control elements are located downstream (i.e. on the 3′ side) of the transcriptional start site (at position +1). In fact they lie within the gene itself. Two such elements have been identified, called the **A box** and **B box** (*Fig. 3*). Transcription of the tRNA genes by RNA Pol III requires **transcription factor IIIC** (**TFIIIC**) as well as **TFIIIB**. TFIIIC binds to the A and B boxes whilst TFIIIB binds upstream of the A box. TFIIIB contains three subunits, one of which is **TBP** (**TATA binding protein**), the polypeptide required by all three eukaryotic RNA polymerases.

After synthesis, the pre-tRNA molecule folds up into the characteristic stem-loops structures (*Fig. 1*) and non-tRNA sequence is cleaved from the 5′ and 3′ ends by ribonucleases. In prokaryotes, the CCA sequence at the 3′ end of the tRNA (which is the site of bonding to the amino acid) is enclosed by the tRNA gene but this is not the case in eukaryotes. Instead, the CCA is added to the 3′ end after the trimming reactions by **tRNA nucleotidyl transferase**. Another

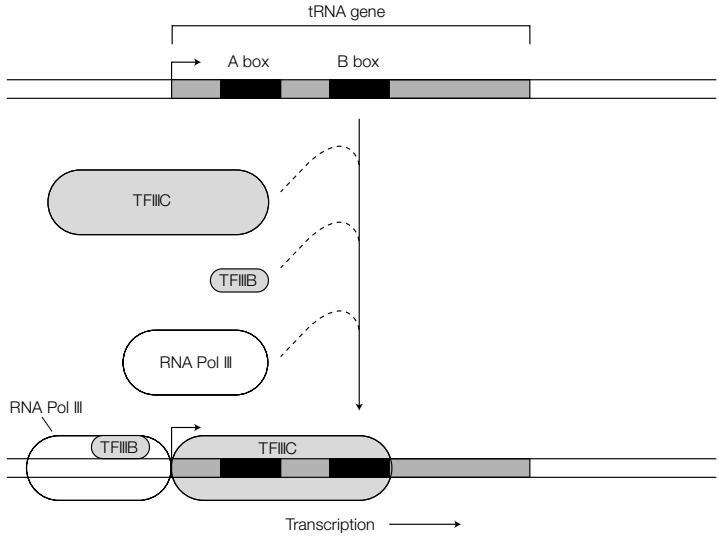

Fig. 3. Initiation of transcription of a tRNA gene by RNA Pol III.

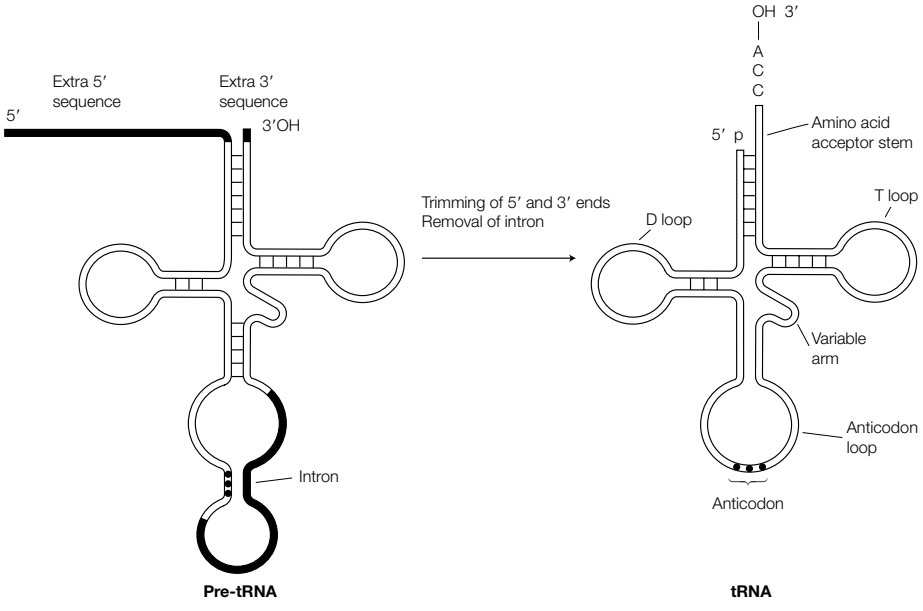

Fig. 4. Processing of a typical eukaryotic pre-tRNA molecule.

difference between prokaryotes and eukaryotes is that eukaryotic pre-tRNA molecules often contain a short **intron** in the loop of the anticodon arm (*Fig. 4*). This intron must be removed in order to create a functional tRNA molecule. Its removal occurs by cleavage by a **tRNA splicing endonuclease** at each end of the intron and then ligation together of the tRNA ends by **tRNA ligase**. This RNA splicing pathway for intron removal is totally different from that used to remove introns from pre-mRNA molecules in eukaryotes (Topic G7) and must have evolved independently.

Modification of tRNA

Transfer RNA molecules are notable for containing unusual nucleotides (*Fig. 5*) such as **1-methylguanosine** (**m^1G**), **pseudouridine** (ψ), **dihydrouridine** (**D**), **inosine** (**I**) and **4-thiouridine** (**S^4U**). These are created by modification of guanosine and uridine after tRNA synthesis. For example, inosine is generated by deamination of guanosine.

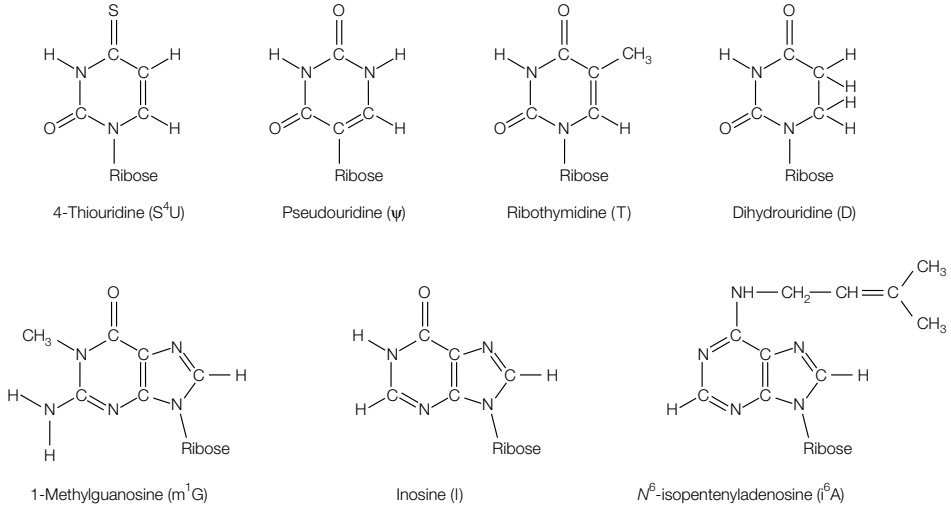

Fig. 5. Some modified nucleosides found in tRNA molecules.