UNIT-I

SYLLABUS

Biosynthesis of RNA in prokaryotes :RNA polymerases, transcription cycle in bacteria, sigma factor, bacterial promoters, identification of DNA binding sites by DNA footprinting, the three stages of RNA synthesis, initiation, elongation and termination, rho-dependent and rho-independent termination. Inhibitors of transcription and applications as anti-microbial drugs.

What is RNA polymerase?

RNA polymerase in simple word means an enzyme which helps in the production of RNA in the cell. These polymerase enzymes are very essential for existence and moreover found in all organisms ranging from bacteria to viruses and in eukaryotic organism too……

RNA polymerase was discovered by Samuel B Weiss and Jerard Hurwitz in 1960. First, whereas all genes are transcribed by a single RNA polymerase in bacteria, eukaryotic cells contain multiple different RNA polymerases that transcribe distinct classes of genes.

Products of RNA polymerase

- Messenger RNA (mRNA): template for the synthesis of proteins by ribosomes.
- Transfer RNA (tRNA): transfers specific amino acids to growing polypeptide chains at the ribosomal site of protein synthesis during translation
- Ribosomal RNA (rRNA): a component of ribosomes
- Micro RNA: regulates gene activity
- Catalytic RNA (Ribozyme): enzymatically active RNA molecules

Types of RNA polymerase

RNA polymerase I: it is the enzyme that copies DNA To rRNA. It almost account for over 50% of RNA synthesis. It synthesizes RNA for large subunit of ribosomes. Molecular mass is of around 500KD. The rate of transcription by it is slower than RNA polymerase II, it is only of 20 nucleotide. Termination by it involves DNA binding protein. TTF1 in mice and REB1P Yeast attach with the DNA at recognized site at 12-20 downstream of termination point.

RNA polymerase II: It is an enzyme found in eukaryotic cells. It catalyzes the transcription of DNA to synthesize precursors of mRNA, most snRNA and microRNA. A 550 kDa complex of 12 subunits, RNAP II is the most studied type of RNA polymerase. A wide range of transcription factors are required for it to bind to upstream gene promoters and begin transcription. It has 10 - 12 subunits (RBP1-12)

RNA polymerase III: In eukaryote cells, RNA polymerase III (also called Pol III) transcribes DNA to synthesize ribosomal 5S rRNA, tRNA and other small RNAs. This enzyme complex has a more limited role than the Pol III in prokaryote cells. The genes transcribed by RNA Pol III fall in the category of "Housekeeping" genes whose expression is required in all cell types and most environmental conditions. Therefore the regulation of Pol III transcription is primarily tied to the regulation of cell growth and the cell cycle, thus requiring fewer regulatory proteins than RNA polymerase II.

Transcription

Transcription is a fundamental cellular process: RNA polymerases "transcribe" the genetic information on DNA into RNA strands. All cells have RNA polymerases (RNAP). The RNA polymerases increase in complexity as you go from viruses (example, T7 RNA polymerase is made up of a single protein), to bacterial systems (one RNA polymerase made up of the proteins - beta, beta', 2 x alpha, omega and the sigma factor), and finally to eukaryotic systems (Three RNA polymerases - Pol I, Pol II, and Pol III, each with ten or more subunits).

While the RNA polymerases have become increasing complex as life evolved, their overall structure (as evidenced by crystallographic structures of bacterial RNA polymerase and Pol II) show remarkable similarity. There is also sequence similarity between the bacterial polymerase protein subunits and the proteins that make up the eukaryotic polymerases.

Transcription of any gene usually involves three distinct stages:

1. First, the RNA polymerase has to find the start site of a gene. The "holoenzyme" form of the polymerase does this by looking for the "promoter" site that exists just upstream of the

gene start site. This process is termed "transcription initiation". This is followed by opening up (melting) of the duplex DNA to form an "open complex".

2. This is followed by a rapid change into the "elongation" phase of transcription where the "core polymerase" part of the RNA polymerase rapidly transcribes an RNA strand that is complementary to the "template" strand of the DNA. The change into the elongation phase usually occurs after a few bases of RNA have been transcribed (typically about 8-9 bases of RNA in bacterial system which form a RNA-DNA hybrid with the template strand), and involves a "clamping down" on the DNA to prevent the polymerase from falling off the DNA.

3. The final stage of the transcription of a gene is "termination", after the stop codon of the gene. The process of termination usually involves sequences where the polymerase slows down or stalls, and the polymerase-RNA-DNA complex (often proteins such as rho and NusA are involved in bacterial systems).

Genes have to be transcribed to mRNAs before they can be translated into proteins; more or less mRNA from a particular gene equals more or less of the protein encoded by the gene. Transcription is, thus, an important point in the control of gene "expression". Most genes are controlled transcriptionally, usually by regulation of the level of transcriptional initiation. For example, if a gene has a strong promoter, it will be more highly expressed when compared to another gene with a weak promoter site. Similarly if a regulatory protein can bind the promoter site of a gene (and prevent transcription initiation), then it can turn off the expression of that gene. Transcriptional control of genetic expression is vital for cellular functions, and many diseases and cancers are results of defects in the transcriptional control of essentials genes.

DNA foot printing or EMSA (electrophoretic mobility shift assay)

DNA foot printing also referred as gel shift assay, band shift assay or gel retardation assay is a common affinity electrophoresis technique used to study protein-DNA and protein-RNA interaction. The gel retardation assay tests the ability of a protein to bind a radiolabel DNA fragment as it migrate through a non-denaturing gel under the influence of an electric current. Binding of protein will reduce the mobility of the DNA.

Variations of DNA foot printing -

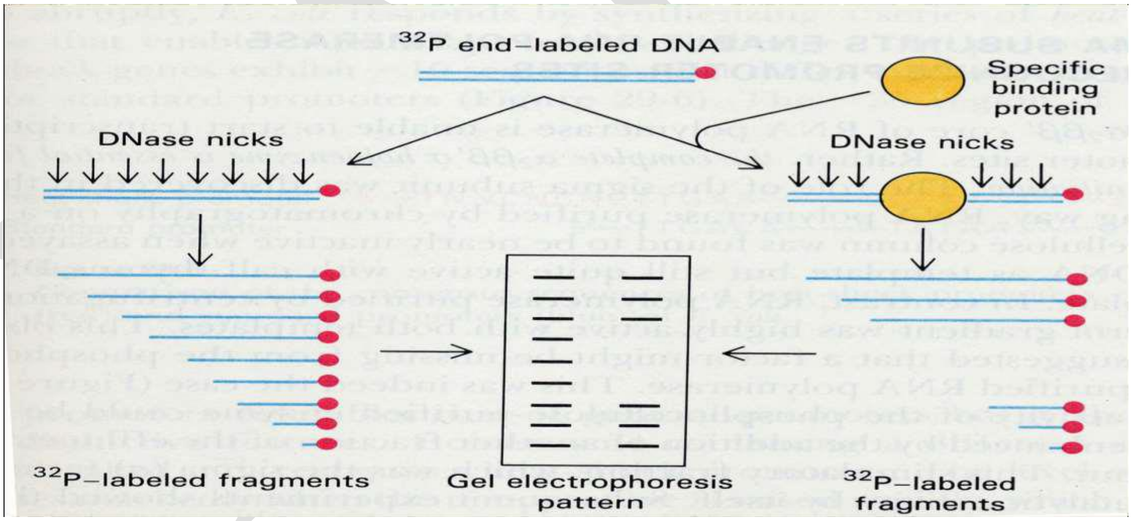
A. Nuclease protection foot printing

B. Modification protection foot printing

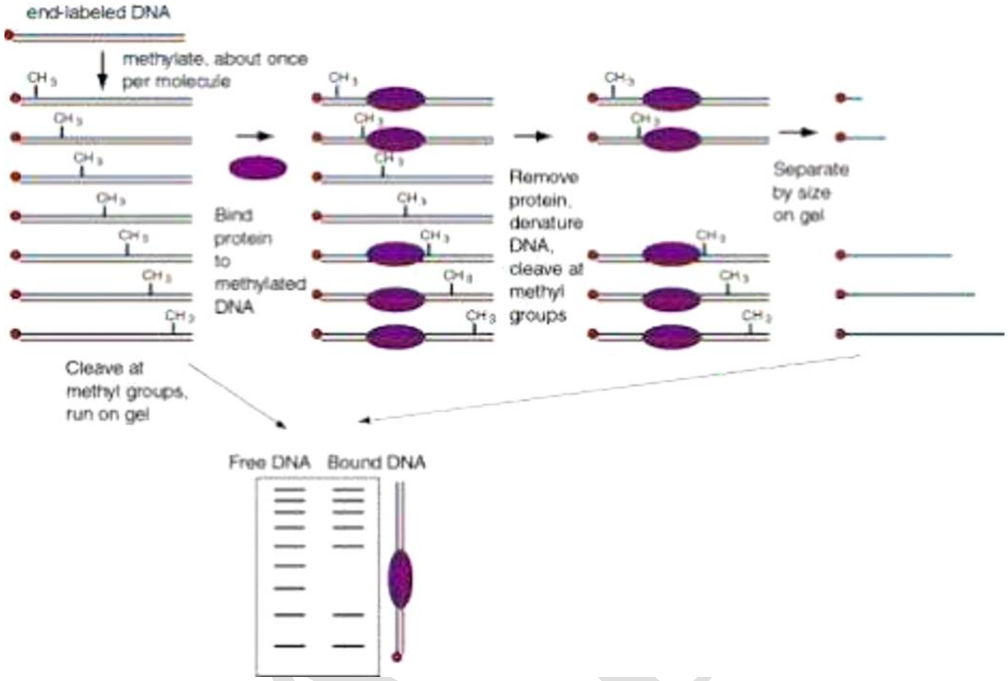
Principle of EMSA

A mobility assay is electrophoretic separation of a protein- DNA & protein-RNA mixture on a polyacrylamide or agarose gel for a short period (about 1.5-2 hr for a 15 to 20 cm gel). The speed at which different molecules (and combination thereof) move through the gel is determined by their size and charge. The control lane (DNA probe without protein present) will contain a single band corresponding to the unbound DNA or RNA fragment.

A. **Nuclease protection foot printing:** The DNaseI foot printing assay measures the ability of a protein to protect a radiolabel DNA fragment against digestion by DNaseI. If a protein X binds the DNA at a particular site, then this protects that site against digestion by DNaseI, as a consequence, a gap or footprint appears in the ladder of DNAmolecule.


^{32}P end–labeled DNA
Specific binding protein
DNase nicks
DNase nicks
^{32}P–labeled fragments
Gel electrophoresis pattern
^{32}P–labeled fragments

B. **Modification protection foot printing:** DNA fragments are treated with limited amounts of dimethyl sulfate (DMS) so that a single G base is metylated in each fragment. If a protein X binds the DNA at particular site, then this protects guanines at that site from the action from dimethyl sulfate. Guanines that are protected by the bound protein X cannot be modified. After removal of the protein, the DNA is treated with piperidine, which cuts at the modified nucleotide positions. Piperidine only cuts the strand that is modified.


end-labeled DNA
methylate, about once per molecule
CH3
Bind protein to methylated DNA
Remove protein, denature DNA, cleave at methyl groups
Separate by size on gel
Cleave at methyl groups, run on gel
Free DNA
Bound DNA

Applications of DNA foot printing

1. To analyze sequence-specific recognition of nucleic acid by proteins and the management of cell growth and behavior.
2. In study of regulation of gene expression.
3. For studies on transcriptional regulation in bacteria.
4. In the study of AU-rich element (ARE) and a member of human Hu family RNA-binding protein (HuR) interaction.
5. EMSAs using near-infrared fluorescence technology are used to study of RNA processing, DNA replication, DNA repair mechanism, protein-DNA interaction and protein-RNA interaction.

Transcription initiation: Transcription initiation in bacteria (prokaryotes) involves sigma factors. The sigma factor combines with the core RNA polymerase to form a holoenzyme that is competent for promoter binding. The core RNA polymerase, by itself, cannot bind the promoter site. The sigma factor can be thought of as the specificity factor in the RNA polymerase. Each bacteria has several different sigma factors that recognize slightly different promoter sequences. Predominant among these is the sigma70 (70 kDa in E. coli; also called rpoD), which initiates the transcription of most genes in exponentially growing cells. There are two general classes of sigma factors - sigma 70 class and sigma 54 class:

1. The sigma 70 class of sigma factors share extensive sequence homology, and bind to two conserved sequences upstream of the gene start site (the -35 box and the -10 box). Each of these sigma factors recognizes slight variations in these conserved sequence boxes. Sigma70 binds promoter DNA as a part of the holoenzyme, and binds DNA very poorly in absence of the core enzyme.
2. The sigma54 class of genes (54 kDa in E. coli; also called rpoN) controls a much smaller set of genes than sigma70. It recognizes different conserved sequences (the -12 and -24 boxes). Unlike Sigma70, sigma54 can bind DNA even in the absence of the core RNA polymerase. It however lacks the ability to melt promoter DNA on its own - for this it needs to interact with other activator proteins that bind further upstream of the promoter site, as well as the core RNAP.

   RNA polymerase (enzyme that catalyzes the synthesis of RNA from a DNA template).

   a) **Core enzyme** = 3 different types of subunits (2α; 1β; 1β')

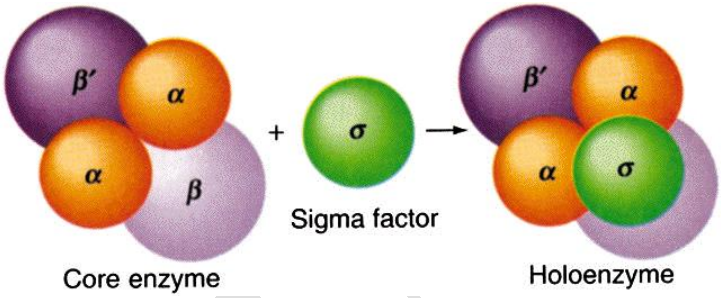
   (1) β - binds incoming nucleotides

   (2) β' – binds DNA

   (3) α - helps with enzyme assembly; interacts with other transcriptional activator proteins; recent work demonstrated that a also interacts with some DNA sequences

   b) **Holoenzyme** = core + σ factor (recognizes the promoter)

c) **σ factors** – Initially, people thought that there was only one σ factor that functioned to direct RNAP to the promoters of genes. Later, different classes of s factors were found. Each s factor directs RNAP to a different type of promoter (differentiated by a specific DNA sequence in the promoter).

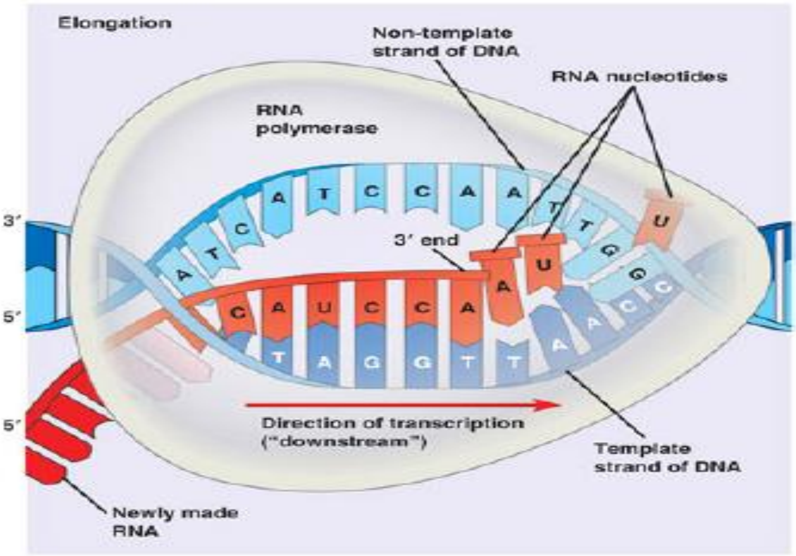


In eukaryotic systems, transcription initiation is very different. There are no sigma factors. Instead, the central protein in forming the "pre-initiation complex " (PIC) is the TATA binding protein (TBP), that binds to the TATA box, a conserved sequence just upstream of the initiation region. A large number of other general transcription factors such as TFIIB, TFIIE, TFIIF, TFIIH (TF stands for transcription factor; II stands for Pol II; there are similar factors for Pol I and Pol II) and others assemble to form the multisubunit TFIID complex. This PIC then recruits the RNA polymerase (Pol I, II, or III in eukaryotic cells) to initiate transcription. The PIC often remains at the promoter site, and is then available to initiate another round of transcription. TBP is a universal transcription factor, and is seen in all eukaryotes and archaea. It sharply bends DNA at the TATA box.

Transcription elongation

Once initiation is complete and the open complex forms, the RNA polymerase begins to read the template strand and add corresponding RNA nucleotides. This process is not always efficient, and the polymerase may make several passes at this. After a long enough RNA-DNA hybrid is made, the polymerase clears the promoter region and moves rapidly downstream. This is preceeded by a large conformational change in the polymerase core enzyme, as it clamps down on the DNA and becomes quite processive.

While transcription elongation is quite rapid, the polymerase does not transcribe all sequences with equal efficieny. The elongation rate is not uniform, and there can be pausing or stalling. Elongation factors (GreA/GreB in bacterial systems; TFIIS in eukaryotes) act to help the polymerase along by stimulating backtracking and cleavage of the newly formed RNA (from the 3' end). Other factors are involved in the elongation cycle - for example, in eukaryotic Pol II there are the elongins and ELL proteins that increase the elongation rate, as well as factors to remodel chromatin.



Transcription elongation

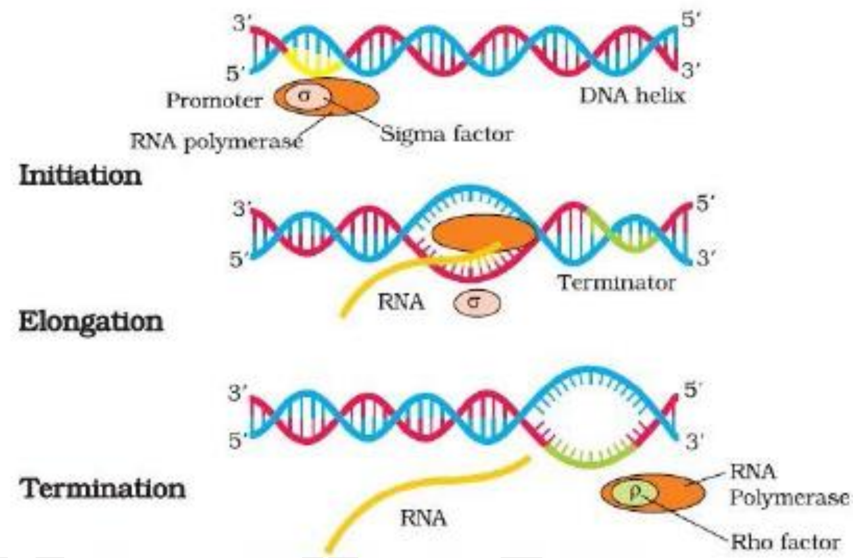
Transcription termination

Termination in bacterial system can be broadly classified as "rho independent", and "rho dependent". In rho independent termination, there is formation of a stable GC rich stem-loop in the newly synthesized RNA followed by a string of U's (A's in the template strand) spaced about 20 bases downstream (these sites are often called intrinsic terminators). The stem loop "snares" the polymerase, slowing or stalling it. This pause, coupled with the low stability of the RNA-DNA hybrid at the active site (run of A=U basepairs) allows the RNA polymerase to fall off the template DNA and terminates the RNA transcription for that gene.

In rho dependent termination, rho binds the newly formed RNA (as hexamers), and stalls the RNA polymerase by interacting with it. In some models, the Rho hexamer translocates on the

RNA. Rho termination activity is stimulated by ATP hydrolysis. This activity is greatly enhanced by Nus factors.

In eukaryotic cells, transcription termination involves cleavage of the elongating RNA chain by specific endonucleases which recognize particular sequences (AAUAAA) in the newly formed RNA. Once this happens, the RNA elongation complex is destabilized, and falls off the DNA. It is then available to attach to another nearby PIC, and start transcribing again.



Process of transcription in bacteria

Inhibitors of transcription and applications as anti-microbial drugs

Bacterial transcription Inhibitors

RNA polymerase ensures the transcription of the information contained in DNA from DNA to mRNA. The gene, fragment of DNA, is the informational unit. Transcription of DNA into mRNA is done in the presence of ATP, UTP, GTP and CTP through RNA-polymerase-DNA-dependent, which catalyzes the elongation of the chain. The activity of RNA polymerase is controlled by activator and repressor proteins.

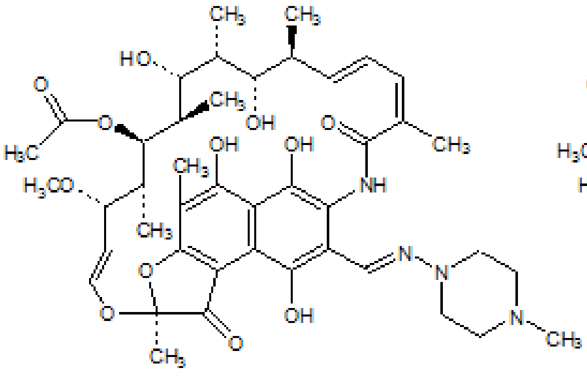
$$\text{DNA} \xrightarrow[\text{en présence d'ATP, UTP, GTP, CTP}]{\textit{RNA-polymérase-DNA-dépendante}} \text{RNA}$$

The reading for transcription begins at the promoter, area of regulation of the coding domain of DNA. The RNA-polymerase identifies the promoter, binds to it. Initiation elicits the spacing of the two strands which constitute a bubble where transcription is carried out. mRNA, resulting from the transcript, undergoes modifications or maturation such as intron splicing, to be converted into a functional structure .

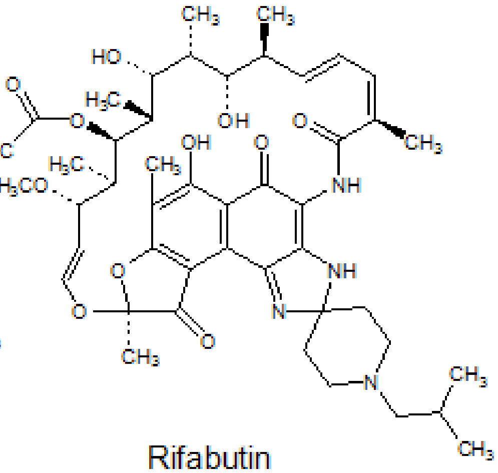
Rifamycins, macrocyclic antibiotics produced by *Streptomyces mediterranei*, inhibit the bacterial RNA polymerase, by binding to the beta subunit, which is one of the five subunits of the enzyme: They have little action on the human RNA polymerase. This group of antibiotics includes rifampicin, rifabutin and rifamycine SV.

Rifampin: It is also called rifampicin, has a bactericidal activity against a wide range of microorganisms, of which *Mycobacterium tuberculosis* and *Mycobacterium lepræ* as well as staphylococci, streptococci, *Neisseria*, *Listeria monocytogenes*, *Brucella*… It is used as anti-tuberculous drug, always combined to two or three other drugs to avoid the emergence of resistance and as anti-leprous drug. Its other clinical uses are brucellosis and the prophylaxis of meningococcal meningitis. A pharmacokinetic characteristic of rifampin is to be a microsomal enzyme inducer of most cytochrome P450 isoforms. It can accelerate the inactivation of many drugs among which oral contraceptives, or cause attacks of porphyria in porphyric patients. Among its other potential adverse effects, it can cause a flu-like syndrome. Resistance to rifampin is explained in certain cases by its ribosylation i.e. by the binding of a sugar via the ADP-ribosyl transferase produced by the resistant microorganism. Rifampicin (Rifadin*, Rimactan*) is marketed alone and in combination with isoniazid (Rifinah*) and with isoniazid and pyrazinamid (Rifater*)

Rifabutin: It has an antibacterial activity quite similar to that of rifampin, it is active against mycobacteria such as Mycobacterium tuberculosis and Mycobacterium avium complex. It is also active against several gram-positive bacteria. Rifabutin (Mycobutin*) is used for the curative treatment of multidrug-resistant tuberculosis and for the prophylactic treatment of Mycobacterium avium complex infection in immunocompromised patients. Rifabutin is a less potent microsomal enzyme inducer than rifampin and can be preferred in patients taking other drugs. Rifampin and rifabutin can elicit a rise in hepatic transaminases and thrombocytopenia and neutropenia. They give an orange color to the urine. Rifabutin can cause uveitis. Rifamycine S.V is used in the form of ophthalmic solution. Rifapentine is a rifampin analog used in certain countries for tuberculosis therapy.



Rifampin or rifampicin



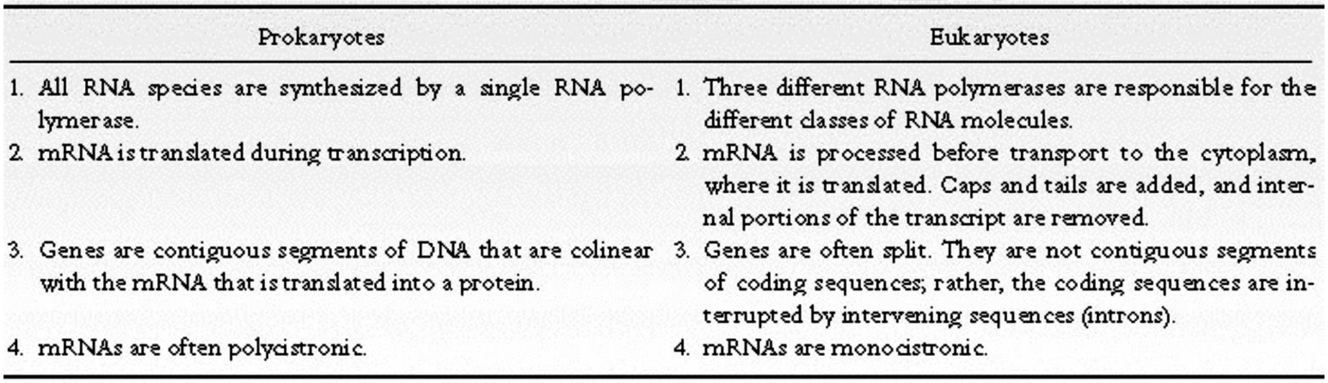
Rifabutin

UNIT-II

SYLLABUS

Biosynthesis of RNA in eukaryotes: Comparison between prokaryotic and eukaryotic transcription. Transcription by RNA polymerase II, RNA polymerase II core promoters, general transcription factors, various types of RNA processing, transcription by RNA polymerase I and III. Inhibitors of eukaryotic transcription and their applications. Comparison of fidelity of transcription and replication.**RNA splicing-** Chemistry of RNA splicing, the spliceosome machinery, splicing pathways, group I and group II introns, alternative splicing, exon shuffling, RNA editing.

Comparison between prokaryotic and eukaryotic transcription



Prokaryotes	Eukaryotes
1. All RNA species are synthesized by a single RNA polymerase.	1. Three different RNA polymerases are responsible for the different classes of RNA molecules.
2. mRNA is translated during transcription.	2. mRNA is processed before transport to the cytoplasm, where it is translated. Caps and tails are added, and internal portions of the transcript are removed.
3. Genes are contiguous segments of DNA that are colinear with the mRNA that is translated into a protein.	3. Genes are often split. They are not contiguous segments of coding sequences; rather, the coding sequences are interrupted by intervening sequences (introns).
4. mRNAs are often polycistronic.	4. mRNAs are monocistronic.

Transcription by RNA Polymerase:

• A single RNA polymerase is responsible for transcribing all types of RNA in prokaryotic system.

• However, eukaryotes have three different RNA polymerases, which have been found to specialize in the synthesis of various types ofRNA:

• RNA polymerase I (Pol I) -transcribes rRNA (ribosomal RNA) genes.

• RNA polymerase II (Pol II) - transcribes protein-coding genes or mRNA (messenger RNA).

• RNA polymerase III (Pol III) - transcribes other functional RNA genes (e.g., tRNA).

• In eukaryotes, transcription occurs inside the nucleus. All the enzymes responsible for translation are present in the cytosol therefore the transcripts formed then move out of the nucleus through nuclear pores into the cytosol (the liquid phase of the cytoplasm), where translation occurs. Organelle genomes (like the mitochondrial genome) are transcribed within the organelle (the mitochondria) and translation is also within the organelle (the mitochondria).

• Since prokaryotes have no nucleus, the step involving the movement of transcripts from nucleus to cytoplasm does not take place, and translation can take place immediately in the cytoplasm, directly on the growing transcript.

Promoters

A promoter is a region of DNA that initiates transcription of a particular gene. Promoters are located near the transcription start sites of genes, on the same strand and upstream on the DNA (towards the 5' region of the antisense strand). Promoters can be about 100–1000 base pairs long.

For transcription to take place, the enzyme that synthesizes RNA, known as RNA polymerase, must attach to the DNA near a gene. Promoters contain specific DNA sequences such as response elements that provide a secure initial binding site for RNA polymerase and for proteins called transcription factors that recruit RNA polymerase. These transcription factors have specific activator or repressor sequences of corresponding nucleotides that attach to specific promoters and regulate gene expression.

In bacteria

The promoter is recognized by RNA polymerase and an associated sigma factor, which in turn are often brought to the promoter DNA by an activator protein's binding to its own DNA binding site nearby.

In eukaryotes

The process is more complicated, and at least seven different factors are necessary for the binding of an RNA polymerase II to the promoter.

Promoters represent critical elements that can work in concert with other regulatory regions (enhancers, silencers, boundary elements/insulators) to direct the level of transcription of a given gene.

Promoter elements

Core promoter – the minimal portion of the promoter required to properly initiate transcription[3]

- Includes the transcription start site (TSS) and elements directly upstream
- A binding site for RNA polymerase
- RNA polymerase I: transcribes genes encoding ribosomal RNA
- RNA polymerase II: transcribes genes encoding messenger RNA and certain small nuclear RNAs and microRNA
- RNA polymerase III: transcribes genes encoding transfer RNAs and other small RNAs
- General transcription factor binding sites, e.g. TATA box
- Proximal promoter – the proximal sequence upstream of the gene that tends to contain primary regulatory elements
- Approximately 250 base pairs upstream of the start site
- Specific transcription factor binding sites
- Distal promoter – the distal sequence upstream of the gene that may contain additional regulatory elements, often with a weaker influence than the proximal promoter
- Anything further upstream (but not an enhancer or other regulatory region whose influence is positional/orientation independent)
- Specific transcription factor binding sites

Operator

- an **operator** is a segment of DNA to which a transcription factorbinds to regulate gene expression. The transcription factor is a repressor, which can bind to the operator to prevent transcription.

- The main operator (O2) in the classically defined *lac* operon is located slightly downstream of the promoter. Two additional operators, O1 and O3 are located at -82 and +412, respectively.

Mechanism

- The repressor protein physically obstructs the RNA polymerase fromtranscribing the genes.
- An inducer (small molecule) can displace a repressor (protein) from the operator site (DNA), resulting in an uninhibited operon.
- Alternatively, a corepressor can bind to the repressor to allow its binding to the operator site. A good example of this type of regulation is seen for the trp operon.

Terminator

transcription terminator is a section of nucleic acid sequence that marks the end of a gene or operon in genomic DNA during transcription. This sequence mediates transcriptional termination by providing signals in the newly synthesized mRNA that trigger processes which release the mRNA from the transcriptional complex. These processes include the direct interaction of the mRNA secondary structure with the complex and/or the indirect activities of recruitedtermination factors. Release of the transcriptional complex frees RNA polymerase and related transcriptional machinery to begin transcription of new mRNAs.

Two classes of transcription terminators, Rho-dependent and Rho-independent, have been identified throughout prokaryotic genomes. These widely distributed sequences are responsible for triggering the end of transcription upon normal completion of gene or operon transcription, mediating early termination of transcripts as a means of regulation such as that observed in transcriptional attenuation, and to ensure the termination of runaway transcriptional complexes that manage to escape earlier terminators by chance, which prevents unnecessary energy expenditure for the cell.

Rho-dependent terminators

Rho-dependent transcription terminators require a protein called Rho factor, which exhibits RNA helicase activity, to disrupt the mRNA-DNA-RNA polymerase transcriptional complex. Rho-dependent terminators are found in bacteria andphage. The Rho-dependent terminator occurs downstream of translational stop codons and consists of an unstructured, cytosine-rich sequence on the mRNA known as a Rho utilization site (*rut*) for which a consensus sequence has not been identified, and a downstream transcription stop point (*tsp*). The *rut* serves as a mRNA loading site and as an activator for Rho; activation enables Rho to efficiently hydrolyze ATP and translocate down the mRNA while it maintains contact with the rut site. Rho is able to catch up with the RNA polymerase, which is stalled at the downstream *tsp* sites. Contact between Rho and the RNA polymerase complex stimulates dissociation of the transcriptional complex through a mechanism involvingallosteric effects of Rho on RNA polymerase.

Rho-independent terminators

Intrinsic transcription terminators or Rho-independent terminators require the formation of a self-annealing hairpin structure on the elongating transcript, which results in the disruption of the mRNA-DNA-RNA polymerase ternary complex. The terminator sequence contains a 20 basepair GC-rich region of dyad symmetry followed by a short poly-T tract or "T stretch" which is transcribed to RNA to form the terminating hairpin and a 7-9 nucleotide "U tract" respectively. The mechanism of termination is hypothesized to occur through a combination of direct promotion of dissociation through allosteric effects of hairpin binding interactions with the RNA polymerase and "competitive kinetics". The hairpin formation causes RNA polymerase stalling and destabilization, leading to a greater likelihood that dissociation of the complex will occur at that location due to an increased time spent paused at that site and reduced stability of the complex. Additionally, the elongation protein factor NusA interacts with the RNA polymerase and the hairpin structure to stimulate transcriptional termination.

Termination in eukaryotes

In eukaryotic transcription of mRNAs, terminator signals are recognized by protein factors that are associated with the RNA polymerase II and which trigger the termination process. Once the poly-A signals are transcribed into the mRNA, the proteinscleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CstF) transfer from the carboxyl terminal domain of RNA polymerase II to the poly-A signal. These two factors then recruit other proteins to the site to cleave the transcript, freeing the mRNA from the transcription complex, and add a string of about 200 A-repeats to the 3' end of the mRNA in a process known as polyadenylation. During these processing steps, the RNA polymerase continues to transcribe for several kilobases and eventually dissociates from the DNA and downstream transcript through an unclear mechanism; there are two basic models for this event known as the torpedo and allosteric models.

Torpedo model

After the mRNA is completed, the residual RNA strand remains in association with the DNA template and the RNA polymerase II, continuing to be transcribed. XRN2 (5'-3' Exoribonuclease 2), a RNase, attaches to the carboxyl terminal domain of RNA polymerase II and proceeds to degrade the uncapped residual RNA from 5’ to 3’ until it reaches the RNA pol II. 5' cap refers to a modified guanine added to the front of mRNA for protection from RNase. 3' poly(A) tail is added to the end of a mRNA strand for protection from exonucleases. Similar to Rho-dependent termination, XRN2 triggers dissociation of RNA polymerase II by either pushing the polymerase off of the DNA template or pulling the template out of the RNA polymerase. The entire mechanism remains unclear.

Allosteric model

RNA polymerase normally is capable of transcribing DNA into single-stranded mRNA efficiently. However, upon transcribing over the poly-A signals on the DNA template, a conformational shift is induced in the RNA polymerase from the proposed loss of associated proteins from its carboxyl terminal domain. This change of conformation reduces RNA polymerase'sprocessivity making the enzyme more prone to dissociating from its DNA-RNA substrate. In this case, termination is not completed by degradation of mRNA but instead is mediated by limiting the elongation efficiency of RNA polymerase and thus increasing the likelihood that the polymerase will dissociate and end its current cycle of transcription.

Enhancer

An enhancer is a short region of DNA that can be bound with proteins (namely, the trans-acting factors, much like a set of transcription factors) to enhance transcription levels of genes (hence the name) in a gene cluster. While enhancers are usually cis-acting, an enhancer does not need to be particularly close to the genes it acts on, and need not be located on the same chromosome.

In eukaryotic cells the structure of the chromatin complex of DNA is folded in a way that functionally mimics the supercoiled state characteristic of prokaryotic DNA, so that although the enhancer DNA is far from the gene in regard to the number of nucleotides, it is geometrically close to the promoter and gene. This allows it to interact with the general transcription factors and RNA polymerase II. An enhancer may be located upstream or downstream of the gene that it regulates.

Furthermore, an enhancer does not need to be located near to the transcription initiation site to affect the transcription of a gene, as some have been found to bind several hundred thousand base pairs upstream or downstream of the start site.**Enhancers do not act on the promoter region itself, but are bound by activator proteins**. These activator proteins interact with the mediator complex, which recruits polymerase II and the general transcription factors which then begin transcribing the genes. Enhancers can also be found within introns. An enhancer's orientation may even be reversed without affecting its function. Additionally, an enhancer may be excised and inserted elsewhere in the chromosome, and still affect gene transcription. That is the reason that intron polymorphisms are checked though they are not translated.

Silencers

In genetics, a **silencer** is a DNA sequence capable of binding transcription regulation factors, called repressors. DNA contains genes and provides the template to produce messenger

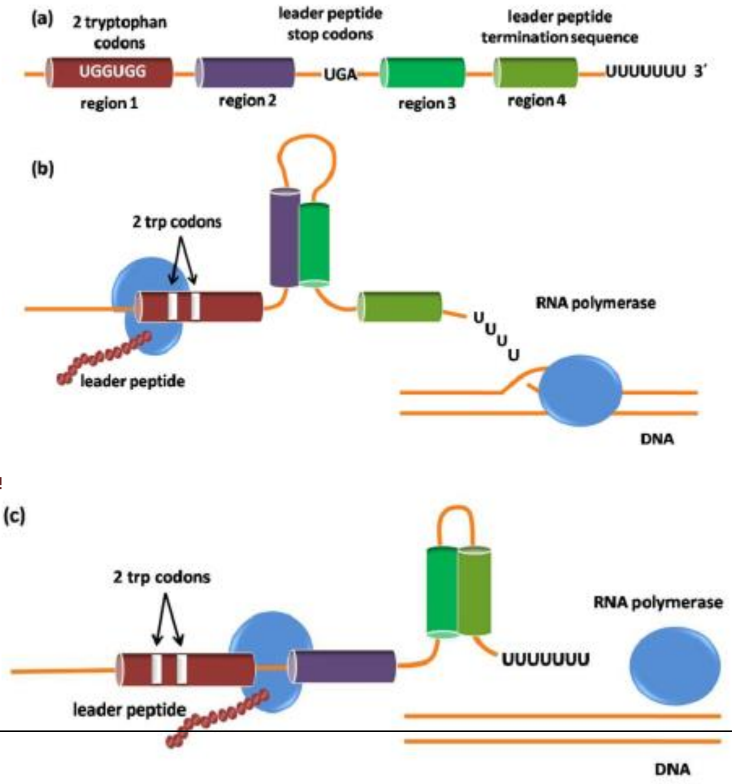
RNA (mRNA). That mRNA is then translated into proteins that activate or inactivate gene expression in cells. When a repressor protein binds to the silencer region of DNA, RNA polymerase is prevented from transcribing the DNA sequence into RNA. With transcription blocked, the translation of RNA into proteins is impossible. Thus, silencers prevent genes from being expressed as proteins.

RNA polymerase, a DNA-dependent enzyme, transcribes the DNA sequences, called nucleotides, in the 3' to 5' direction while the complementary RNA is synthesized in the 5' to 3' direction. RNA is similar to DNA, except that RNA contains uracil, instead of thymine, which forms a base pair with adenine. An important region for the activity of gene repression and expression found in RNA is the 3' untranslated region. This is a region on the 3' terminus of RNA that will not be translated to protein but includes many regulatory regions.

Not much is yet known about silencers but scientists continue to study in hopes to classify more types, locations in the genome, and diseases associated with silencers.

Attenuation

Premature termination of primary transcript in the leader region i.e. before the first structural genes is called attenuation. Attenuation is carried out by attenuator, a sequence within leader region of the tryptophan operon, (Fig 2). At this site, choice is made by RNA polymerase either to terminate or continue transcription. Mutants with small deletions in this region produce tryptophan synthesizing enzymes even in the presence of tryptophan.

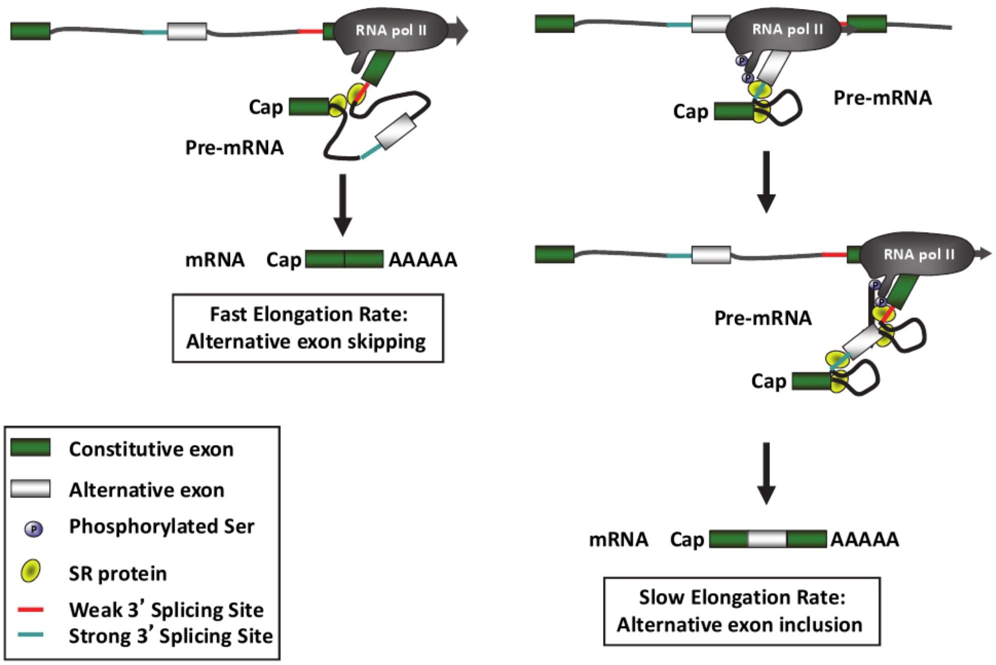


Termination of Transcription regulated by attenuation (a) Stem-loop structures of the trp operon in the mRNA; (b) Low level of trp full length mRNA made; (c) High level transcription of the trp operon is prematurely halted

RNA Splicing

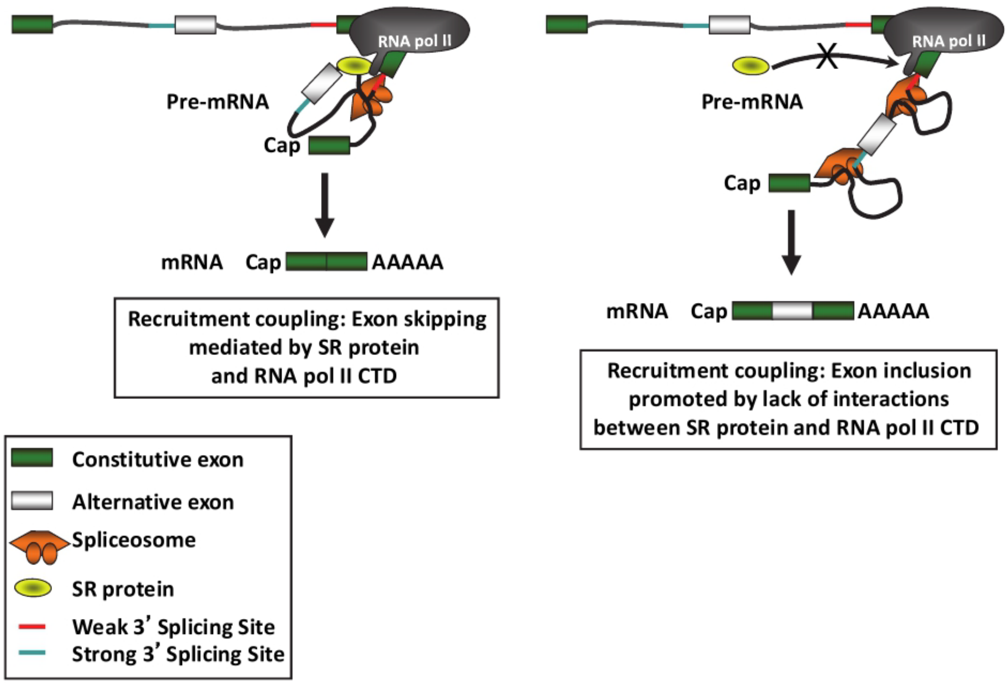
RNA splicing, catalyzed by the spliceosome, a large RNA-protein complex composed of five small nuclear ribonucleoproteins (snRNPs), provides the cell with an additional level of phenotypic complexity without the need for additional transcript generation. Control of splicing can occur in “cis” through regulatory sequences in pre-mRNA, as well as “trans” by factors that bind and act upon these sequences. An example of these factors is the SR proteins which act in the control of splice site recognition by affecting spliceosome assembly. It is the control of splice site recognition which provides the major mechanism by which RNA splicing is regulated. Splice sites within introns have been found to have differing “strengths” which affect their ability to be recognized and acted upon by components of the splicing machinery. This form of splicing regulation is directly related to the control of transcription elongation, both through the kinetic and recruitment models mentioned earlier. Therefore, “co-transcriptional“ splicing provides the cell with the advantages of increased efficiency of transcript generation and processing, preventing mRNA degradation and back-hybridization with DNA.

The kinetic model of co-transcriptional splicing revolves around the concept that the rate of RNA pol II elongation directly affects splice site recognition and spliceosome assembly. The rate by which RNA pol II transcribes along the length of a gene can be affected by two factors: the phosphorylation level of Ser5 and Ser2 on the RNA pol II CTD, as well as the chromatin structure which encapsulates the gene being transcribed. In a nutshell, fast elongation, which occurs when the RNA pol II CTD is hyperphosphorylated and/or the chromatin of the gene being transcribed has a low nucleosome density, favors the inclusion of downstream exons with “strong” splice sites. In contrast, when the RNA pol II CTD is hypophosphorylated and/or the nucleosome density of the transcribed gene is increased, a slow elongation rate allows enough temporal flexibility for the splicing machinery to assemble on upstream, “weaker” splice sites. Initial experiments supporting this concept showed that using “slow” RNA pol II mutants or inserting pausing The regulation of alternative splicing is modulated by the rate of elongation of RNA pol II. Fast elongation rate (Left panel) results in more frequent exon skipping. The rate of elongation can be influenced by the level of CTD Ser 2 and Ser 5 phosphorylation. The weak 3’ splicing site is indicated in blue and the strong 3’ splicing site in red. Slower elongation (Right panel) results in inclusion of the alternative exon (Grey rectangle) between the two constitutive exons (Green rectangles). Cap: 7’methyl guanosine; AAAAA: poly A tail elements in reporter minigenes favors “weak” exon inclusion in the fibronectin and fibroblast growth factor receptor 2 (FGFR2) genes. The fact that there are 46 Ser2 and 51 Ser5 residues in mammalian CTDs provide a sort of “gas pedal” mechanism for the control of elongation rate, and therefore splicing decisions. In an intriguing example, the chromatin remodeling factor SWI/SNF which interacts with RNA pol II, splicing factors, and spliceosome-associated proteins, can cause inclusion of a block of exons in the middle of the CD44 gene by stalling RNA pol II through a phosphorylation status switch from phospho-Ser2 to phospho-Ser5. Further evidence for this intragenic “brake” control mechanism comes from the transient accumulation of phospho-serine 5 on the RNA pol II CTD around the 3’ end of yeast introns. This pausing before an exon is suggestive of a splicing-dependent transcriptional checkpoint which holds any further transcription until spliceosome assembly is accomplished.



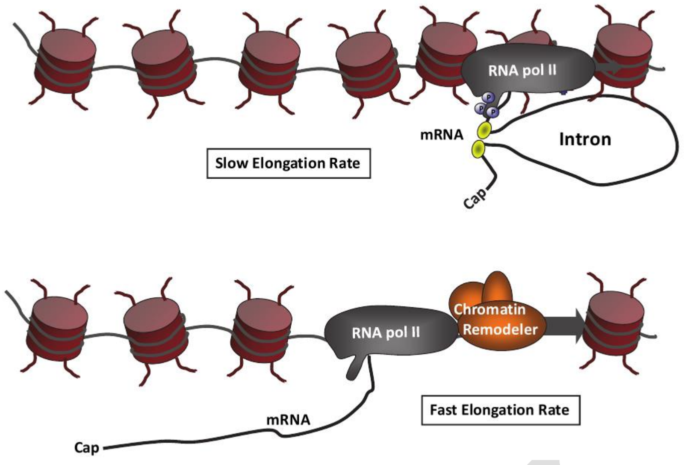
RNA pol II Kinetic Model for Alternative Splicing

In terms of chromatin structure altering elongation rate and splicing, on genes regulated by the chromatin-remodeler SWI/SNF, the ATPase subunit Brahma (Brm) has been shown to contribute to transcription-splicing crosstalk by decreasing the elongation rate (through alterations in nucleosome density patterns) and facilitating recruitment of the splicing machinery to variant exons with suboptimal splice sites. Conversely, treatment with the histone deacetylase inhibitor, Trichostatin A (TSA) facilitates a more “open” chromatin conformation, stimulating elongation rates and causing inhibition of the fibronectin exon EDI inclusion. Much more evidence exists that relates chromatin structure and composition to the regulation of both transcription and splicing, independent of elongation rate and the kinetic model of “co-transcriptionality”. These concepts, including chromatin as a recruiter of both transcription and splicing factors, nucleosome positioning in delineating critical transcription and splice sites, and the involvement of chromatin modifications and modifiers in both transcription and splicing, will be discussed later in detail in this chapter.


RNA pol II
Pre-mRNA
Cap
mRNA Cap AAAAA
Recruitment coupling: Exon skipping mediated by SR protein and RNA pol II CTD
RNA pol II
Pre-mRNA
Cap
mRNA Cap AAAAA
Recruitment coupling: Exon inclusion promoted by lack of interactions between SR protein and RNA pol II CTD
Constitutive exon
Alternative exon
Spliceosome
SR protein
Weak 3' Splicing Site
Strong 3' Splicing Site

Recruitment Coupling Model for Alternative Splicing

The recruitment model of co-transcriptional splicing is similar to the kinetic model in the sense that it revolves around the RNA pol II CTD. Specifically, the recruitment model involves the allosteric regulation of splicing decisions through interactions with the elongation machinery mediated by the RNA pol II CTD ([23]). The most clear-cut example of the recruitment model involves the RNA pol II CTD, the SR protein SRp20, and the alternative exon EDI of the fibronectin gene. SRp20 has an inhibitory effect on the In this model, the recruitment of splicing factors SR protein(s) (yellow ovals) is mediated by interactions with the RNA pol II CTD. Binding of SR proteins, such as SRp20, to the CTD favors alternative exon skipping (Left panel). Loss of SR protein interactions with the CTD prevents formation of a proper ternary complex and leads to alternative exon inclusion (right panel). This model is based on work by de la Mata et al.. inclusion of the fibronectin EDI exon, and this effect is mediated by the RNA pol II CTD. When a RNA pol II mutant lacking the CTD is present, SRp20 is not recruited to the site, and EDI inclusion is greatly enhanced. Long before this evidence was presented, a trend connecting



the recruitment of splicing factors with transcript elongation mediated by the RNA pol II CTD was emerging. When genes are placed under the control of RNA polymerase I, III, or bacteriophage T7 RNA polymerase promoters, transcription occurs, but splicing is greatly affected. Logically, the recruitment of splicing factors to proper slice sites is dependent on the RNA pol II CTD ([28]) and deletion of the CTD affects all aspects of RNA processing in the β-globin gene. The most obvious factor(s) playing a role in the recruitment model of co-transcriptional splicing are the SR proteins. Virtually all members of the SR family are known to interact with RNA pol II, as well as other splicing factors and components of the spliceosome. Additional factors which interact with both the pol II CTD (phorphorylated or un-phosphorylated), as well as integral components of the spliceosome include the elongation factors CA150 and SPT6, as well as the transcriptional regulators TRAP150/Med23, TFII H PSF/p54nrb, and EWS-Fli and NOR1. Many factors that affect chromatin structure and composition also interact with members of both the elongation and splicing machinery, consequently playing a role in the recruitment model of co-transcriptionality as well. These factors primarily include chromatin remodelers and histone modifiers. Evidence for these chromatin-associated factors acting as “adaptor” molecules, bridging both processes and playing roles in both proposed models for the coupling of transcription and splicing is overwhelming. Therefore, additional sections in this chapter have been included to explore their multi-faceted activity in full detail.

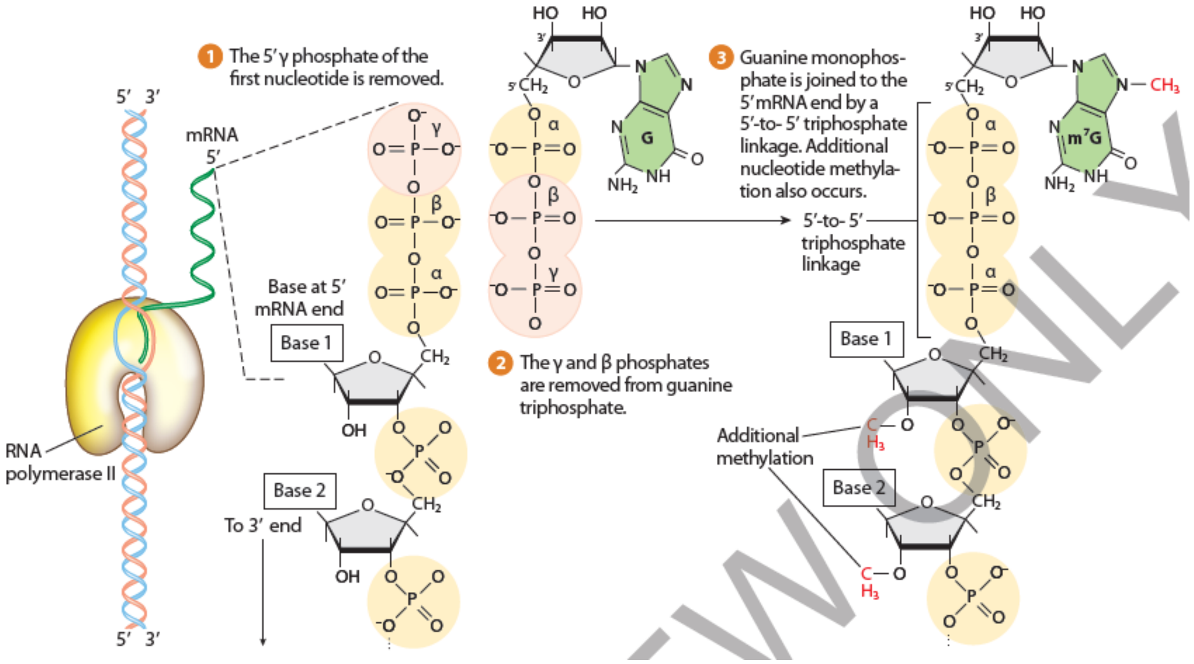
Post-Transcriptional Processing Modifies RNA Molecules

Bacterial, archaeal, and eukaryotic transcripts differ in several ways. For example, eukaryotic transcripts are more stable than bacterial and archaeal transcripts. The half-life of a typical eukaryotic mRNA is measured in hours to days, whereas bacterial mRNAs have an average half-life measured in seconds to minutes. A second difference is the separation, in time and in location, between transcription

and translation. Recall that in bacteria the lack of a nucleus leads to coupling of transcription and translation. Similarly, archaea lack a nucleus, leading to the possibility of synchrony between transcription and translation. In eukaryotic cells, on the other hand, transcription takes place in the nucleus, and translation occurs later at free ribosomes or at those attached to the rough endoplasmic reticulum in the cytoplasm. A third difference is the presence of introns in eukaryotic genes that are absent from most bacterial and archaeal genes. Each of these differences comes into play as we consider post-transcriptional modifications of mRNA in eukaryotic cells, which is the focus of this section. In discussing post-transcriptional processing, we highlight three processing steps that are coordinated during transcription to modify the initial eukaryotic gene mRNA transcript, called **pre-mRNA,** into **mature mRNA,** the fully processed mRNA that migrates out of the nucleus to the cytoplasm for translation. These modification steps are (1) **5′ capping,** the addition of a modified nucleotide to the 5′ end of mRNA; (2) **3′ polyadenylation,** cleavage at the 3′ end of mRNA and addition of a tail of multiple adenines to form the **poly-A tail**; and (3) **intron splicing,** RNA splicing to remove introns and ligate exons. We conclude the section with a discussion of the mechanisms directing alternative splicing and self-splicing RNAs.

Capping 5′ mRNA

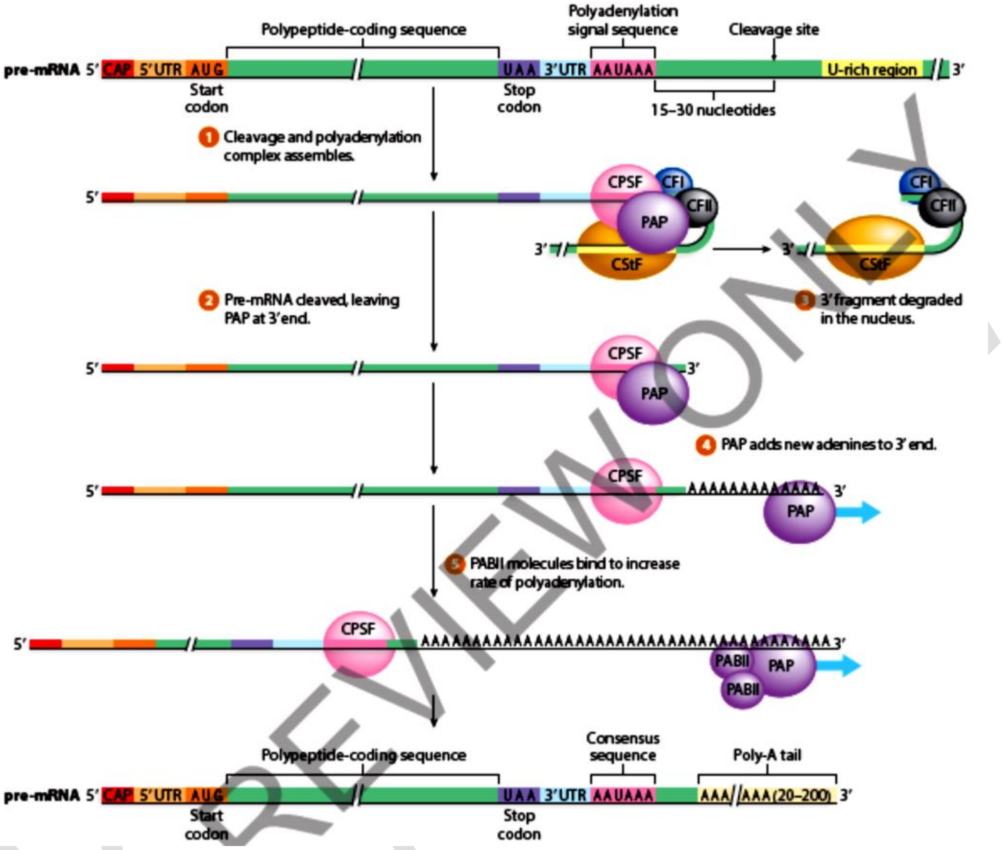
After RNA pol II has synthesized the first 20 to 30 nucleotides of the mRNA transcript, a specialized enzyme, guanylyl transferase, adds a guanine to the 5′ end of the pre-mRNA, producing an unusual 5′-to-5′ bond that forms a triphosphate linkage. Additional enzymatic action then methylates the newly added guanine and may also methylate the next one or more nucleotides of the transcript. This addition of guanine to the transcript and the subsequent methylation is known as 5′ capping. Guanylyl transferase initiates 5′ capping in three steps depicted in **Figure**. Before capping, the terminal 5′ nucleotide of mRNA contains three phosphate groups, labeled α, β, and γ in Figure 8.17. Guanylyl transferase first removes the γ phosphate, leaving two phosphates on the 5′ terminal nucleotide **1** . The guanine triphosphate containing the guanine that is to be added loses two phosphates (γ and β) to form a guanine monophosphate **2** . Then, guanylyl transferase joins the guanine monophosphate to the mRNA terminal nucleotide to form the 5′-to-5′ triphosphate linkage **3** . Methyl transferase enzyme then adds a methyl (CH3) group to the 7-nitrogen of the new guanine, forming 7-methylguanosine (m7G). Methyl transferase may also add methyl groups to 2′–OH of nearby nucleotides of mRNA. The 5′ cap has several functions, including (1) protecting mRNA from rapid degradation, (2) facilitating mRNA transport across the nuclear membrane, (3) facilitating subsequent intron splicing, and (4) enhancing translation efficiency by orienting the ribosome on mRNA.


1 The 5′ γ phosphate of the first nucleotide is removed.
5′ 3′
mRNA
5′
γ
β
α
Base at 5′ mRNA end
Base 1
Base 2
OH
CH₂
To 3′ end
RNA polymerase II
5′ 3′
HO HO
G
NH₂
NH
2 The γ and β phosphates are removed from guanine triphosphate.
3 Guanine monophosphate is joined to the 5′ mRNA end by a 5′-to-5′ triphosphate linkage. Additional nucleotide methylation also occurs.
5′-to-5′ triphosphate linkage
N—CH₃
m⁷G
Additional methylation
CH₃

Capping the 5′ end of eukaryotic pre-mRNA

Polyadenylation of 3′ Pre-mRNA

Termination of transcription by RNA pol II is not fully understood, but it appears likely to be tied to the processing and polyadenylation of the 3′ end of pre-mRNA. It is clear that the 3′ end of mRNA is not generated by transcriptional termination. Rather, the 3′ end of the premRNA is created by enzymatic action that removes a segment from the 3′ end of the transcript and replaces it with a string of adenine nucleotides, the poly-A tail. This step of pre-mRNA processing is thought to be associated with subsequent termination of transcription. Polyadenylation begins with the binding of a factor called cleavage and polyadenylation specificity factor (CPSF) near a six-nucleotide mRNA sequence, AAUAAA, that is downstream of the stop codon and thus not part of the coding sequence of the gene. This six-nucleotide sequence is known as the **polyadenylation signal sequence.** The binding of cleavagestimulating factor (CStF) to a uracil-rich sequence several dozen nucleotides downstream of the polyadenylation signal sequence quickly follows, and the binding of two other cleavage factors, CFI and CFII, and polyadenylate polymerase (PAP) enlarges the complex **1** . The premRNA is then cleaved 15 to 30 nucleotides downstream of the polyadenylation signal sequence **2** . The cleavage releases a transcript fragment bound by CFI, CFII, and CStF, which is later degraded **3** . The 3′ end of the cut pre-mRNA then undergoes the enzymatic addition of 20 to 200 adenine nucleotides that form the 3′ poly-A tail through the action of CPSF and PAP **4** . After addition of the first 10 adenines, molecules of poly-A-binding protein II (PABII) join the elongating poly-A tail and increase the rate of adenine addition **5** . The 3′ poly-A tail has several functions, including (1) facilitating transport of mature mRNA across the nuclear membrane, (2) protecting mRNA from degradation, and (3) enhancing translation by enabling ribosomal recognition of messenger RNA. Certain eukaryotic mRNA transcripts do not undergo polyadenylation. The most prominent of these are transcripts of genes producing *histone proteins,* which are key components of *chromatin,* the DNA–protein complex that makes up eukaryotic chromosomes. On these and other “tailless” mRNAs, the 3′ end contains a short stem-loop structure reminiscent of the ones seen in the intrinsic transcription termination mechanism of bacteria. There may be an evolutionary connection between bacterial transcription termination and stem-loop formation on "tailless" eukaryotic mRNAs.


Polypeptide-coding sequence
Polyadenylation signal sequence
Cleavage site
pre-mRNA 5′
CAP
5′UTR
AUG
UAA
3′UTR
AAUAAA
U-rich region
3′
Start codon
Stop codon
15–30 nucleotides
1 Cleavage and polyadenylation complex assembles.
5′
CPSF
CFI
CFII
PAP
CStF
3′
2 Pre-mRNA cleaved, leaving PAP at 3′ end.
3 3′ fragment degraded in the nucleus.
4 PAP adds new adenines to 3′ end.
AAAAAAAAAAAA 3′
5 PABII molecules bind to increase rate of polyadenylation.
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA 3′
PABII
Polypeptide-coding sequence
Consensus sequence
Poly-A tail
AAA//AAA(20–200)

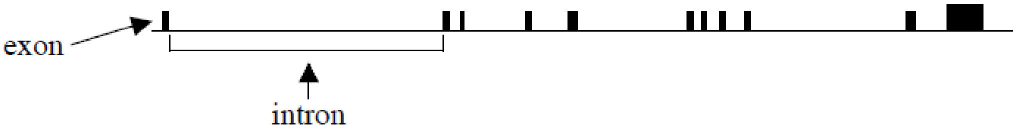
Polyadenylation of the 3′ end of eukaryotic pre-mRNA.

Alternative Splicing

Introduction

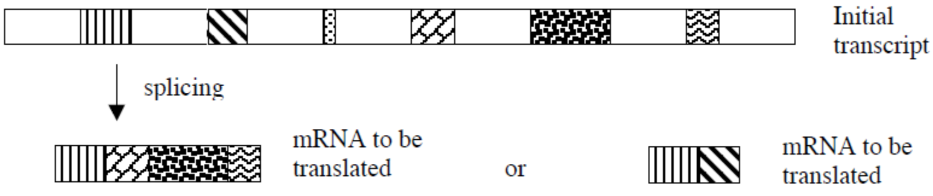
One of the ways in which the vertebrate genome is more complex than that of other organisms is by increased use of alternative splicing. In alternative splicing, more than one protein product is made from one gene. This explains how vertebrates are able to make 5 times as many proteins as flies or worms, with only 2 times as many genes. Alternative splicing occurs frequently in vertebrates. Alternative splicing is estimated to occur for 59% of vertebrate transcripts compared to 22% of nematode transcripts. Vertebrates have an estimated 3.2 different final transcripts per gene compared to 1.34 for nematodes.

Gene Structure



Eucaryotic genes are composed of interspersed exons and introns. Exons (**ex**pressed sequences) contain the coding regions of the protein. Introns (**in**terspersed sequences) are removed after transcription. In vertebrates usually only 5% of the gene is made up of exons. Below is pictured what the gene structure of a gene looks like on the NCBI website. Black boxes are exons. Introns are the open areas between the boxes.

Most genes have 7 or 8 exons, with an average length of 145 bp. Introns have an average length of 3365 bp. In alternative splicing, the same gene is processed in two or more ways. When the gene is spliced, different exons may be included or excluded from the final transcript. For example, the initial transcript shown below could be spliced to produce several different mRNA products, two of which are shown below.

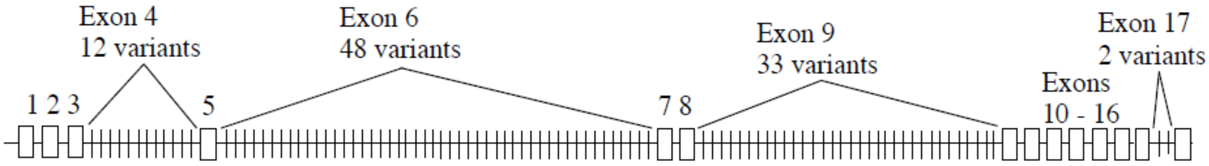


This same initial transcript could be spliced in many additional ways not shown above. This changes the “one gene, one protein” theory. Now we see that many proteins can be made from one gene. Humans have only 30,000 genes. Earlier predictions of gene number based on the number of proteins humans were estimated to produce was 100,000. Vertebrates have an average of 3.2 different final transcripts per gene, compared to 1.34 for nematodes. Thus the human genome with only 30,000 genes can produce 100,000 different proteins with only 30,000 genes. About 59% of our genes are spliced into more than one different mRNA, compared to 22% of nematode transcripts. Alternative splicing explains how vertebrates have 5 times as many proteins as flies and worms but only two times as many genes.

Examples of Alternative Splicing

Drosophila melanogaster – axon guidance. In *Drosophila*, the dscam (Down syndrome cell adhesion molecule) gene encodes an axon guidance receptor which is involved in the migration and connection of neurons during embryogenesis.

The gene contains 24 exons (exons 1 – 18 are depicted below). Four of the exons have different variants. Note: this method of numbering is different from the standard numbering. By standard numbering the 12 different variants of exon 4 would be numbered exons 5 – 17. Researchers chose to number these exons differently because of the large number of exons involved and due to the way in which the exons are included in final transcripts.



Only one of each exon is included in the final transcript. Thus the gene is able to encode 12 x 48 x 33 x 2 = 38,016 possible proteins. This is 2 – 3 times the total number of genes drosophila has! It is thought that these many different variants of dscam are responsible for helping neuronal axons find their correct destinations during embryogenesis.

Ways in which alternative splicing can be used

This is just a preview, you will explore some of these different uses in more detail in the activity that follows. In the example above, alternative splicing was used to include only one of several versions of an exon into a final protein product. This allows many slightly different versions of the same protein to be made without repeating the whole gene. Alternative splicing can also be used to make 2 completely different proteins by inclusion of mutually exclusive coding exons in the final transcript. In this case, alternative splicing allows the cell to use the same cellular localization and transport exons for both proteins. Alternative splicing can be used to convert a transcriptional repressor into an activator by including or excluding exons responsible for activating transcription. The same protein can be targeted to different places in the cell by including different exons encoding different transport signals in the final transcript.

Exon Shuffling

Before going on, it may be useful for me to define certain key terms and concepts. I will be referring frequently to “exons” and “introns.” Exons are sections of genes that code for proteins; whereas introns are sections of genes that don’t code for proteins. Proteins have multiple structural levels. Primary structure refers to the linear sequence of amino acids comprising the protein chain. When segments within this chain fold into structures such as helices and loops, this is referred to as secondary structure. Common units of secondary structure include α-helices and β-strands. Tertiary structure is the biologically active form of the protein, and refers to the packing of secondary structural elements into domains. Since a protein’s tertiary structure optimizes the forces of attraction between amino acids, it is the most stable form of the protein. When multiple folded domains are arranged in a multi-subunit complex, it is referred to as a quaternary structure.

A further concept is domain shuffling. This is the hypothesis that fundamentally new protein folds can be created by recombining already-existing domains. This is thought to be accomplished by moving exons from one part of the genome to another (exon shuffling). There are various ways in which exon shuffling might be achieved, and it is to this subject that I now turn.

The Mechanisms of Exon Shuffling

There are several ways in which exon shuffling may occur. Exon shuffling can be transposon-mediated, or it can occur as a result of crossover during meiosis and recombination between non-homologous or (less frequently) short homologous DNA sequences. Alternative splicing is also thought to play a role in facilitating exon shuffling.

When domain shuffling occurs as a result of crossover during sexual recombination, it is hypothesized that it takes place in three stages (called the “modularization hypothesis”). First, introns are gained at positions that correspond to domain boundaries, forming a “protomodule.” Introns are typically longer than exons, and thus the majority of crossover events take place in the noncoding regions. Second, within the inserted introns, the newly formed protomodule undergoes tandem duplication. Third, intronic recombination facilitates the movement of the protomodule to a different, non-homologous, gene.

Another hypothesized mechanism for domain shuffling involves transposable elements such as LINE-1 retroelements and Helitron transposons, as well as LTR retroelements. LINE-1 elements are transcribed into an mRNA that specifies proteins called ORF1 and ORF2, both of which are essential for the process of transposition. LINE-1 frequently associates with 3′ flanking DNA, transporting the flanking sequence to a new locus somewhere else on the genome. This association can happen if the weak polyadenylation signal of the LINE-1 element is bypassed during transcription, causing downstream exons to be included on the RNA transcript. Since LINE-1’s are “copy-and-paste” elements (i.e. they transpose via an RNA intermediate), the donor sequence remains unaltered.

Long-terminal repeat (LTR) retrotransposons have also been established to facilitate exon shuffling, notably in rice. LTR retrotransposons possess a *gag* and a *pol* gene. The *pol* gene translates into a polyprotein composed of an aspartic protease (which cleaves the polyprotein), and various other enzymes including reverse transcriptase (which reverse transcribes RNA into DNA), integrase (used for integrating the element into the host genome), and Rnase H (which serves to degrade the RNA strand of the RNA-DNA hybrid, resulting in single-stranded DNA). Like LINE-1 elements, LTR retrotransposons transpose in a “copy-and-paste” fashion via an

RNA intermediate. There are a number of subfamilies of LTR retrotransposons, including endogenous retroviruses, Bel/Pao, Ty1/copia, and Ty3/gypsy.

Alternative splicing by exon skipping is also believed to play a role in exon shuffling. Alternative splicing allows the exons of a pre-mRNA transcript to be spliced into a number of different isoforms to produce multiple proteins from the same transcript. This is facilitated by the joining of a 5′ donor site of one intron to the 3′ site of another intron downstream, resulting in the "skipping" of exons that lie in between. This process may result in introns flanking exons. If this genomic structure is reinserted somewhere else in the genome, the result is exon shuffling.There are of course other mechanisms that are hypothesized to play a role in exon shuffling. But this will suffice for our present purposes. Next, we will look at the evidence for and against domain shuffling as an explanation for the origin of new protein folds.

Introns Early vs. Introns Late

It was hypothesized fairly early, after the discovery of introns in vertebrate genes, that they could have contributed to the evolution of proteins. In a 1978 article in *Nature*, Walter Gilbert first proposed that exons could be independently assorted by recombination within introns. Gilbert also hypothesized that introns are in fact relics of the original RNA world. According to the "exons early" hypothesis, all protein-coding genes were created from exon modules — coding for secondary structural elements (such as α-helices, β-sheets, signal peptides, or transmembrane helices) or folding domains — by a process of intron-mediated recombination.

The alternative "introns late" scenario proposed that introns only appeared much later in the genes of eukaryotes. Such a scenario renders exon shuffling moot in accounting for the origins of the most ancient proteins. The "introns early" hypothesis was the dominant view in the 1980s. The frequently cited evidence for this was the then widespread belief in the general correspondence between exon-intron structure and protein secondary structure. From the mid 1980s, this view became increasingly untenable, however, as new information came to light that raised doubts about a general correlation between protein structure and intron-exon structure. Such a correspondence is not borne out in many ancient protein-coding genes. Moreover, the apparently clearest examples of exon shuffling all took place fairly late in the evolution of eukaryotes, becoming significant only at the time of the emergence of the first multicellular animals.

In addition, analysis of intron splicing junctions suggested a similar pattern of late-arising exon shuffling. The location where introns are inserted and interrupt the protein's reading frame determines whether exons can be recombined, duplicated or deleted by intronic recombination without altering the downstream reading frame of the modified protein. Introns can be grouped according to three "phases": Phase 0 introns insert between two consecutive codons; phase 1 introns insert between the first and second nucleotide of a codon; and phase 2 introns insert between the second and third nucleotide. Thus, if exon shuffling played a major role in protein evolution, we should expect a characteristic intron phase distribution. But the hypothetical modules of ancient proteins do not conform to such expectations. It is clear, then, that exon shuffling (at the very least) is unlikely to explain the origins of the most ancient proteins that have emerged in the history of life. But is this mechanism adequate to explain the origins of later proteins such as those that arise in the evolution of eukaryotes? I now turn to evaluate the evidence pro-and-con for the role of exon shuffling in protein origins.

The Case for Exon Shuffling

What, then, are the best arguments for exon shuffling? If the thesis is correct, a prediction would be that exon boundaries should correlate strongly with protein domains. In other words, one exon should code for a single protein domain. One argument, therefore, points to the fact that there is a statistically significant correlation between exon boundaries and protein domains. However, there are many, many examples where this correspondence does not hold. In many cases, single exons code for multiple domains. For instance, *protocadhedrin* genes typically involve large exons coding for multiple domains. In other cases, multiple exons are required to specify a single domain.

A further argument for the role of exon shuffling in protein evolution is the intron phase distributions found in the exons coding for protein domains in humans. In 2002, Henrik Kaessmann and colleagues reported that “introns at the boundaries of domains show high excess of symmetrical phase combinations (i.e., 0-0, 1-1, and 2-2), whereas nonboundary introns show no excess symmetry”. Their conclusion was thus that “exon shuffling has primarily involved rearrangement of structural and functional domains as a whole.” They also performed a similar analysis on the nematode worm *Caenorhabditis elegans*, finding that “Although the *C. elegans* data generally concur with the human patterns, we identified fewer intron-bounded domains in this organism, consistent with the lower complexity of *C. elegans* genes.”

Another line of evidence relates to genes that appear to be chimeras of parent genes. These are typically associated with signs indicative of its mode of origin. One famous example is the *jingwei*gene in *Drosophila*, which may have arisen when “the sequence of the processed *Adh* [alcohol dehydrogenase] messenger RNA became part of a new functional gene by capturing several upstream exons and introns of an unrelated gene”. We must take care, however, not to confuse the observed pattern of intron phase distribution, or exon/domain mapping, with proof that exon shuffling is actually the process by which this pattern arose.Perhaps common ancestry is the cause, but this must be demonstrated and not assumed. It is the biologist’s duty to determine whether unintelligent chance-based mechanisms actually can produce novel genes in this manner. It is to this question that I now turn.

The Problems with Domain Shuffling as an Explanation for Protein Folds

While the hypothesis of exon shuffling does, taken at face value, have some attractive elements, it suffers from a number of problems. For one thing, the model at its core *presupposes* the prior existence of protein domains. A protein’s lower-level secondary structures (α-helices and β-strands) exist stably only in the context of the tertiary structures in which they are found. In other words, the domain level is the lowest level at which self-contained stable structural modules exist. This leaves the origins of these domains in the first place unaccounted for. But stable and functional protein domains are demonstrably rare within amino-acid sequence space.

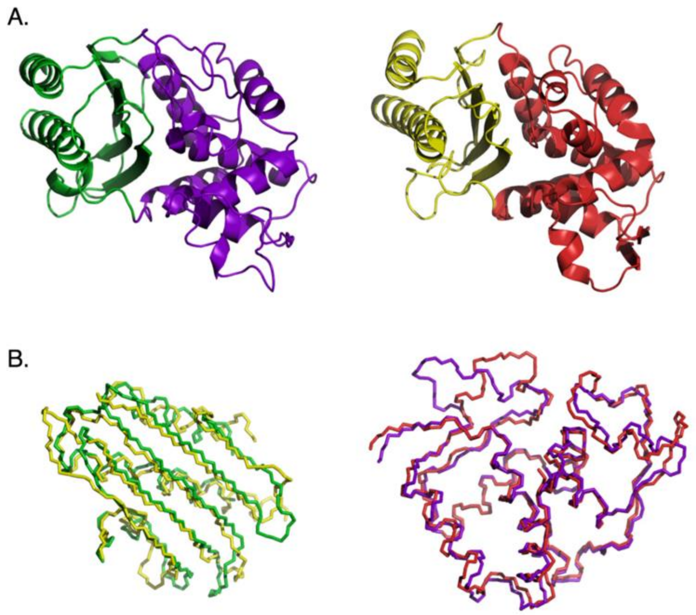
A fairly recent study examined many different combinations of *E. coli* secondary structural elements (α-helices, β-strands and loops), assembling them “semirandomly into sequences comprised of as many as 800 amino acid residues”. The researchers screened 108 variants for features that might suggest folded structure. They failed, however, to find any folded protein structures. Reporting on this study, writes:

“After a definitive demonstration that the most promising candidates were not properly folded, the authors concluded that “the selected clones should therefore not be viewed as ‘native-like’ proteins but rather ‘molten-globule-like'”, by which they mean that secondary structure is present only transiently flickering in and out of existence along a compact but mobile chain. This contrasts with native-like structure, where secondary structure is locked-in to form a well defined and stable tertiary fold. Their finding accords well with what we should expect in view of the above considerations. Indeed, it would be very puzzling if secondary structure were modular.”

“For those elements to work as robust modules,” explains Axe, “their structure would have to be effectively context-independent, allowing them to be combined in any number of ways to form new folds.” In the case of protein secondary structure, however, this requirement is not met. The model also seems to require that the diversity and disparity of functions carried out by proteins in the cell can in principle originate by mixing and matching prior existing domains. But this presupposes the ability of blind evolutionary processes to account for a specific “toolbox” of domains that can be recombined in various ways to yield new functions. This seems unlikely, especially in light of the estimation that “1000 to 7000 exons were needed to construct all proteins” . In other words, a primordial toolkit of thousands of diverse protein domains needs to be constructed before the exon shuffling hypothesis even becomes a possibility. And even then there are severe problems.

A further issue relates to interface compatibility. The domain shuffling hypothesis in many cases requires the formation of new binding interfaces. Since amino acids that comprise polypeptide chains are distinguished from one another by the specificity of their side-chains, however, the binding interfaces that allow units of secondary structure (i.e. α-helices and β-strands) to come together to form elements of tertiary structure is dependent upon the specific sequence of amino acids. That is to say, it is non-generic in the sense that it is strictly dependent upon the particulars of the components. Domains that must bind and interact with one another can't simply be pieced together like jenga tiles.

In his 2010 paper in the journal *BIO-Complexity* Douglas Axe reports on an experiment conducted using β-lactamase enzymes which illustrates this difficulty. Take a look at the following figure, excerpted from the paper:



The top half of the figure (labeled "A") reveals the ribbon structure of the TEM-1 β-lactamase (left) and the PER-1 β-lactamase (right). The bottom half of the figure (labeled "B") reveals the backbone alignments for the two corresponding domains in the two proteins. Note the high level of structural similarity between the two enzymes. Axe attempted to recombine sections of the two genes to produce a chimeric protein from the domains colored green and red. Since the two parent enzymes exhibit extremely high levels of structural and functional similarity, this should be expected to work. No detectable function was identified in the chimeric construct, though, presumably as a consequence of the substantial dissimilarity between the respective amino-acid sequences and the interface incompatibility between the two domains.

This isn't by any means the only study demonstrating the difficulty of shuffling domains to form new functional proteins. Another study described "a set of hybrid sequences" from "the 50%-identical TEM-1 and *Proteus mirabilis* β-lactamases," which were created such that the "hybrids match[ed] the TEM-1 sequence except for a region at the C-terminal end, where they [were] random composites of the two parents." The results? "All of these hybrids are biologically inactive."

In fact, in the few cases where protein chimeras *do* possess detectable function, it only works for the precise reason that the researchers used an algorithm to carefully select the sections of a protein structure that possess the fewest side-chain interactions with the rest of the fold, and chose parent proteins with relatively high sequence identity. This only serves to underscore the problem. Even in the Voigt study, the success rate was quite low, even with highly favorable circumstances, with only one in five chimeras possessing discernible functionality. To conclude, although there is some indirect inferential evidence for the role of exon shuffling in protein evolution, a consideration of how such a process might work in reality reveals that the hypothesis itself is fraught with severe difficulties.

RNA Editing

Occasionally researches encounter a gene with a sequence of nucleotides that does **not match exactly** that in its RNA product:

- messenger RNA (mRNA) or
- ribosomal RNA (rRNA) or
- transfer RNA (tRNA) and even
- microRNA (miRNA)

If the product is **mRNA**, some of the codons in the open reading frame (ORF) of the gene specify different amino acids from those in the protein translated from the mRNA of the gene.

The reason is **RNA editing**: the alteration of the sequence of nucleotides in the RNA

- after it has been transcribed from DNA but
- before it is translated into protein

RNA editing occurs by two distinct mechanisms:

- **Substitution Editing**: chemical alteration of individual nucleotides (the equivalent of point mutations).

  These alterations are catalyzed by **enzymes** that recognize a specific target sequence of nucleotides (much like restriction enzymes):

  - cytidine deaminases that convert a C in the RNA to uracil (U);
  - adenosine deaminases that convert an A to inosine (I), which the ribosome translates as a G. Thus a CAG codon (for Gln) can be converted to a CGG codon (for Arg).
- **Insertion/Deletion Editing**: insertion or deletion of nucleotides in the RNA.

  These alterations are mediated by **guide RNA** molecules that

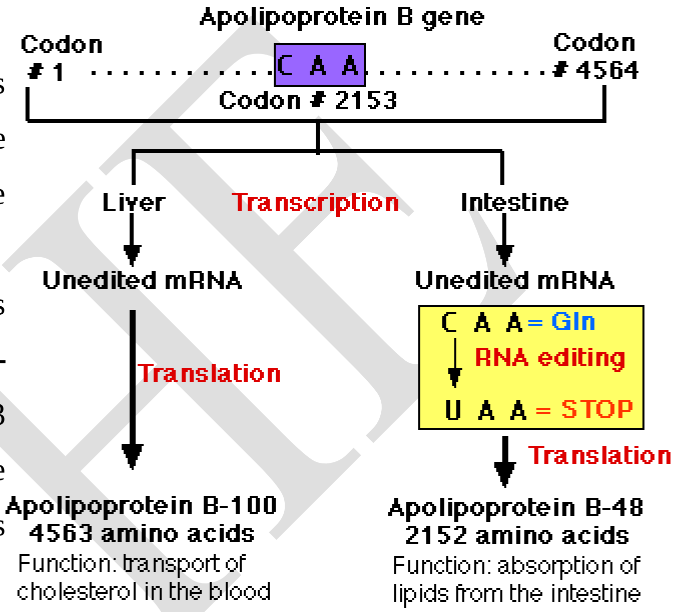
  - base-pair as best they can with the RNA to be edited and
  - serve as a **template** for the addition (or removal) of nucleotides in the target

Substitution Editing

Example: the human *APOB* gene

Humans have a single locus encoding the ***APOB*** gene.

- It contains 29 exons (separated by 28 introns).
- The exons contain a total of 4564 codons.
- Codon 2153 is CAA, which is a codon for the amino acid glutamine (Gln).
- The gene is expressed in cells of both the liver and the intestine.
- In both locations, transcription produces a **pre-messenger RNA** that must be spliced to produce the mRNA to be translated into protein.
- In the **Liver**. Here the process occurs normally producing **apolipoprotein B-100** — a protein containing 4,563 amino acids — that is essential for the transport of cholesterol and other lipids in the blood.



- In the **Intestine**.
  - In the cells of the intestine, an additional step of pre-mRNA processing occurs: the chemical modification of the C nucleotide in Codon 2153 (CAA) into a U.
  - This **RNA editing** changes the codon from one encoding the amino acid glutamine (Gln) to a STOP codon (UAA)
  - The modification is catalyzed by the enzyme **cytidine deaminase** that
    - recognizes the sequence of the RNA at that one place in the molecule and
    - catalyzes the deamination of C thus forming U.
  - Translation of the mRNA stops at codon #2153 forming **apolipoprotein B-48** — a protein containing 2152 amino acids — that aids in the absorption of dietary lipids from the contents of the intestine.

DNA can also be edited. B cells express another **cytidine deaminase** (called **a**ctivation-**i**nduced **d**eaminase or **AID**) that is essential for both class switch recombination (**CSR**) and somatic hypermutation(**SHM**) of antibody genes. Humans with disabling mutations in the gene for this enzyme produce only IgM antibodies. However, here the enzyme is acting on DNA, not RNA. In attempting to repair the mismatch formed (dC•dG converted to dU•dG), the normal DNA repair machinery of the cell produces CSR or SHM as the situation warrants. (This process is also responsible for the occasional aberrant translocation of the heavy-chain gene segments to a proto-oncogene. The result is a B-cell cancer — a lymphoma or leukemia.)

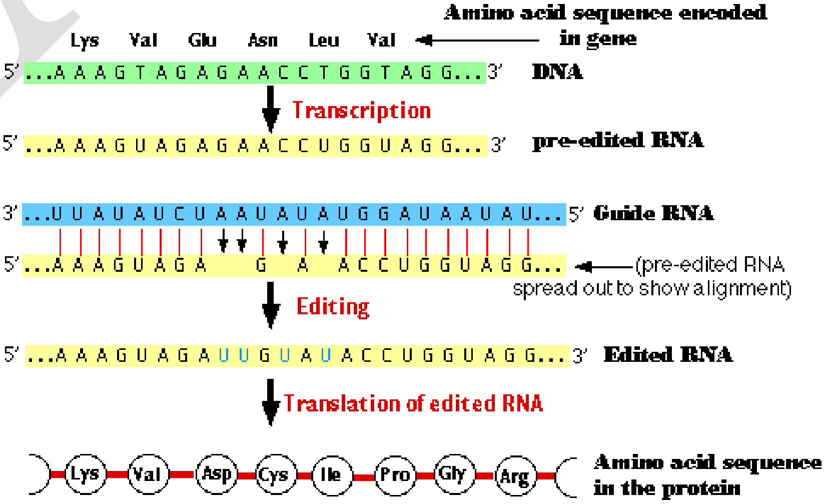
Some other examples of substitution editing

- Some mRNAs, tRNAs, and rRNAs in both the mitochondria and chloroplasts of **plants**;
- mRNAs encoding subunits of some receptors of neurotransmitters in the mammalian brain, e.g.,
  - the AMPA receptor for Glu. [Link]
  - a serotonin receptor
- a tRNA in the mitochondria of the duckbill platypus.

Insertion/Deletion Editing

Example: the gene (in the mitochondria of *Trypanosoma brucei*) for one of the subunits of cytochrome c oxidase

Several genes encoded in the mitochondrial DNA of this species (the cause of sleeping sickness in humans)


Amino acid sequence encoded in gene
Lys Val Glu Asn Leu Val
5′ ...A A A G T A G A G A A C C T G G T A G G... 3′
DNA
Transcription
5′ ...A A A G U A G A G A A C C U G G U A G G... 3′
pre-edited RNA
3′ ...U U A U A U C U A A U A U A U G G A U A A U A U... 5′
Guide RNA
5′ ...A A A G U A G A G A A C C U G G U A G G...
(pre-edited RNA spread out to show alignment)
Editing
5′ ...A A A G U A G A U U G U A U A C C U G G U A G G... 3′
Edited RNA
Translation of edited RNA
Lys Val Asp Cys Ile Pro Gly Arg
Amino acid sequence in the protein

encode transcripts that must be edited to make the mRNA molecules that will be translated into protein.

Editing requires a special class of RNA molecules called **guide RNA** (**gRNA**).

These small molecules have sequences that are complementary to the region around the site to be edited. The guide RNA base-pairs — as best it can — with this region. (Note that in addition to the usual purine-pyrimidine pairing of **C-G** and **A-U**, **G-U** base-pairing can also occur.)
Because of the lack of precise sequence complementarity, bulges occur either

- in the guide RNA where, usually, there are **A**s not found in the transcript to be edited (as shown here) or
- in the transcript to be edited.

The bulges are eliminated by cutting the backbone of the shorter molecule and inserting complementary bases.

- In the first case (shown here) this produces **insertions** (here of **U**s)
- In the second case (not shown) this produces **deletions**.

Note that in the example shown here, the insertion of 4 nucleotides has created a frameshift so that the amino acids encoded downstream (after **Val**) in the edited RNA are entirely different from those specified by the gene itself.

Some other examples of insertion/deletion editing
Insertion/deletion editing has also been found occur with

- mRNA, rRNA, and tRNA transcripts in the mitochondria of the slime mold **Physarum polycephalum**.
- in measles virus transcripts.

Why RNA Editing?

Good question. Some possibilities:

- Perhaps — like alternative splicing — it is a mechanism to increase the number of different proteins available without the need to increase the number of genes in the genome. (The human genome is not that much larger than those of Drosophila and *C. elegans*, but our proteome is much larger.)
- So it can create proteins with slightly different functions to use in specialized circumstances.

○ The ability to synthesize two versions (with different functions) of **apolipoprotein B** from a single gene — as shown above — is an example. Note that in this case, RNA editing has accomplished the same result as alternative splicing.

○ There is evidence that **Drosophila** (and humans) use editing to create subtle differences in the properties of some voltage-gated ion channels and some receptors of neurotransmitters in different regions of the brain.

○ In two species of octopus — one tropical and one in the Antarctic — their **gene** encoding a voltage-gated potassium (K^+) channel differs at only four nucleotides, and these have no effect on the electrical properties of the channels. However, the **messenger RNAs** for the channel are extensively edited in each species to produce channel proteins with different electrical properties. In each case, the channel protein enables rapid firing of action potentials at the temperature of their environment (−1.8°C in the Antarctic and ~30°C in the waters off Puerto Rico). See the report by Sandra Garrett and J. J. C. Rosenthal in the 17 February 2012 issue of **Science**.

Still unanswered: do these evolutionary adaptations arise during the life of the individual or did they arise earlier? In either case, while the different properties of the two channel proteins do not arise from differences in the encoding gene, what genetic differences establish the different pattern of editing in the two species?

- RNA (as well as DNA) editing may help protect the genome against
  - some viruses (and against which they often evolve counter-measures)
  - damage by retrotransposons
- Other possibilities: The untranslated regions at either end of many messenger RNA (mRNA) molecules — the 5'-UTRs and 3'-UTRs — often contain sequences of nucleotides that permit **intra**molecular base-pairing resulting in stretches of double-stranded RNA (dsRNA). (View another example.) But dsRNA in the cell runs the risk of being destroyed by **RNA interference** (RNAi). [Link to discussion]. Perhaps RNA editing in these areas protects against that risk. There is also evidence that RNA editing (converting As to Is in the 3'-UTR) of precursor mRNAs is a signal to retain them within the nucleus ready to be quickly exported if needed by the cell.
- Or perhaps it is simply the legacy of a system that was first used to correct RNA transcripts for harmful mutations in the DNA encoding them. Now with no strong evolutionary pressure to correct the DNA, RNA editing persists.

So RNA editing appears to be here to stay. In fact, defects in RNA editing are associated with some human cancers as well as with amyotrophic lateral sclerosis (ALS — "Lou Gehrig's disease").

UNIT-III

SYLLABUS

Biosynthesis of proteins - The genetic code-Degeneracy of the genetic code, wobble in the anticodon, features of the genetic code, nearly universal code. **Biosynthesis of proteins-** Messenger RNA, transfer RNA, attachment of amino acids to tRNA, the ribosome - initiation, elongation and termination of translation, regulation of translation. Comparison of prokaryotic and eukaryotic protein synthesis. Use of antibiotics in understanding protein synthesis and applications in medicine. **Protein targeting and degradation -** Post translational modifications, glycosylation, signal sequences for nuclear transport, bacterial signal sequences, import of proteins by receptor mediated endocytosis, specialized systems for protein degradation.

Genetic Code

The structure of DNA provides a mechanism for self-replication. The structure of DNA also reveals the mechanism for storing the genetic information that determines what a cell is and how it functions. This information is known as the genetic code. In this section, we will look at how the information stored in molecules of DNA is used to direct the synthesis of protein. It is also important to know how DNA is regulated in organisms, so that the appropriate DNA is expressed in each cell. Later on we will look at some of the ways in which gene expression is regulated.

Learning What a Gene Does

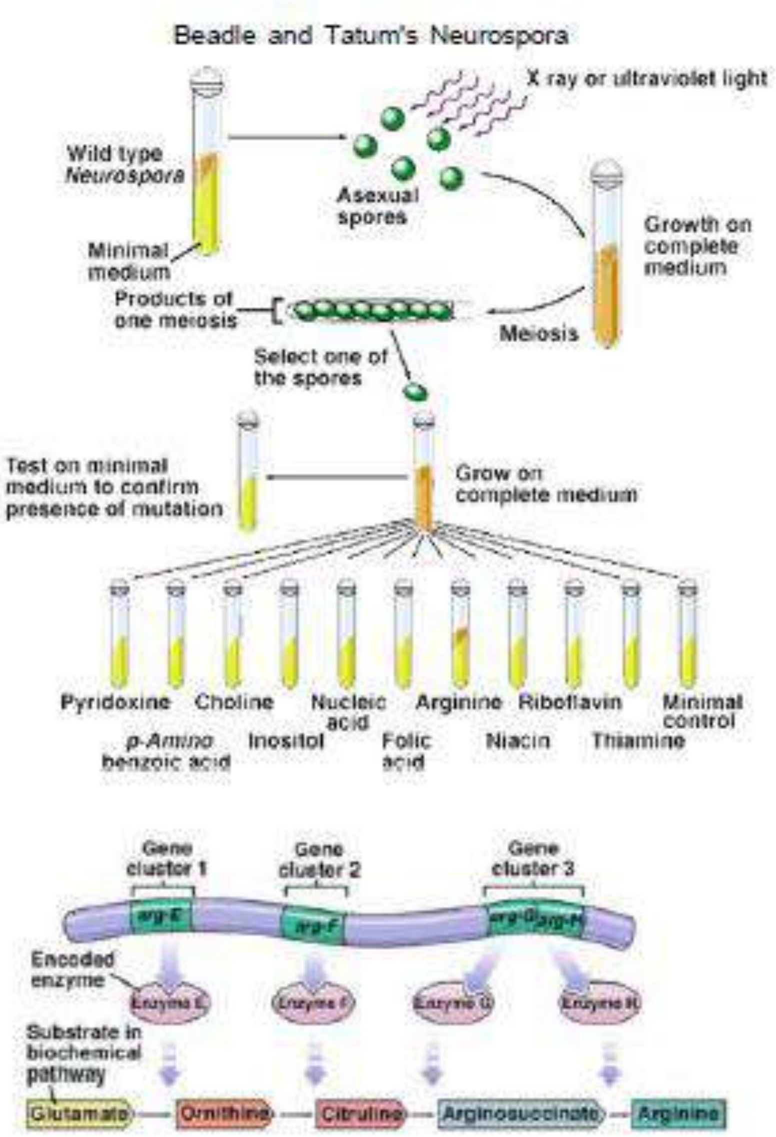
Before we discuss how DNA does its job of storing and using genetic information, let's look a bit at one bit of research that led to the conclusion that a **gene** is a piece of DNA that specifies the amino acid sequence in a polypeptide (or protein).

Garrod's contribution

In 1909 Archibald Garrod postulated that inherited diseases were caused by the inability of the individual to synthesize a particular enzyme. He was correct; many of our metabolic disorders are caused by not having a specific enzyme. However, it took decades of research to "prove" that a gene's function is to provide instructions on how to synthesize a specific protein, and that metabolic pathways for synthesis and degradation of molecules within a cell (and organism) are catalyzed by specific enzymes at each step of the pathway.

Beadle and Ephrussi and Mutations in Fruit Fly Eye Color

In the 1930's, Beadle and Ephrussi proposed that mutations in eye color in the fruit fly prevented synthesis of the pigment by blocking the synthesis of an enzyme along some step in the metabolic pathway. As is common in science, the pathway for eye pigment production was not known at the time, so they could not pursue their hypothesis.



Beadle and Tatum and Pink Bread Mold

In the 1940's Beadle and Tatum induced mutations (changes in the genetic code) in *Neurospora*, a pink mold common on bread, and tracked the metabolism of the mutant strains. They mapped chromosome locations of the mutant strains, and then related their chromosome maps to the presence of absence of specific enzymes needed in Neurospora's metabolic pathway for the synthesis or arginine. From their research, Beadle and Tatum postulated the **one gene-one enzyme** theory. Eventually, we also learned that not all genes must code for enzymes; some code for structural proteins or functional proteins (many of which we discussed in our first unit). Furthermore, quaternary proteins are composed of more than one polypeptide, so the concept of gene has been further refined to be that a **gene codes for a polypeptide**.

The expression of DNA in the genetic control of the cell - Preview

• DNA contains the instructions for each cell to function, precisely coded in its four-letter "alphabet", A, T, C, and G. DNA also has regions of no apparent function and regions of genetic gibberish.

• These instructions are used to direct the synthesis of polypeptides (the primary structures of proteins) for the cell. Many of these proteins become functional enzymes catalyzing the metabolic activities of cells. Others are the structural proteins of cells.

• DNA is not used directly as a template for protein synthesis, a process that occurs in the cytoplasm at ribosomes. DNA molecules never leave the nucleus of the cell. To carry the information stored in DNA to the cytoplasm, we use the molecule, **RNA (ribose nucleic acid)**.

• DNA is used as a template to build a set of RNA instructions, a process called **transcription**. This occurs in the nucleus.

• RNA molecules travel from the nucleus to the ribosomes in the cytoplasm of the cell.

• At the ribosomes, a second process, called **translation** occurs. During translation, the information carried by the messenger RNA molecules is used to direct the assembly of specific amino acids into proteins. Two other types of RNA, **transfer RNA** and **ribosomal RNA** are also needed for the process of translation.

However, before we go further in our discussion of transcription and translation we need to look a little more closely at: • some specifics about the different types of RNA • the genetic code and how it is read

RNA – The molecule that uses the information stored in DNA

Structure of RNA Molecules

RNA is composed of

• Phosphate

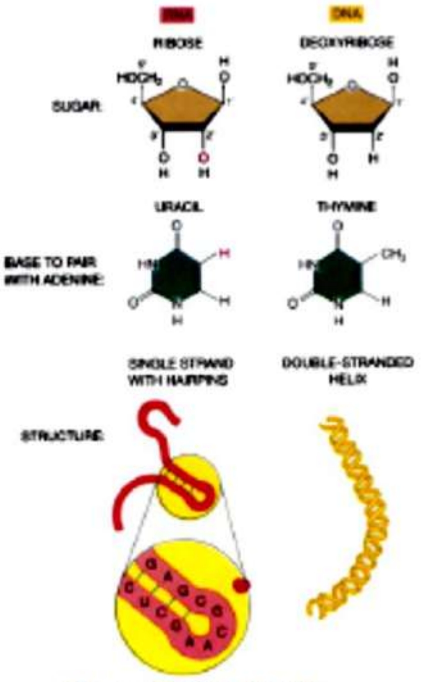
• Ribose sugar

Ribose has -OH groups on both the 2nd, 3rd and 5th carbons Deoxyribose lacks an -OH group on just the 2nd carbon)

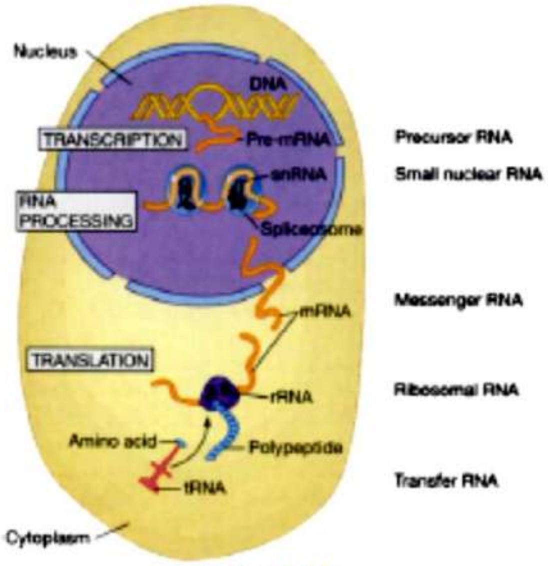
• Four nucleotides

- Adenine
- Guanine
- Cytosine
- Uracil

Replaces the thymine found in DNA. Uracil bonds to Adenine. Uracil has a single -H on the ring at the position where Thymine has a - CH3 group. Molecules of RNA are single-stranded. However, some RNA molecules fold back on themselves at places forming complementary base pair bonds. The double stranded sections are called **hairpins**, and provide stability to the structure of the RNA molecule for its function in protein synthesis and provide resistance to the RNA hydrolyzing enzymes in the cytosol.



Comparing RNA with DNA



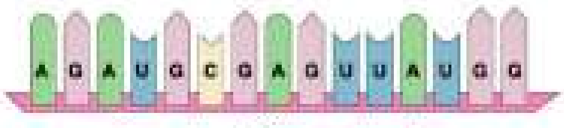
Types of RNA

The Types of RNA

1 . Messenger RNA (mRNA)

- The unique blueprint, or transcript for each protein to be assembled.
- mRNA is manufactured by transcription on demand; that is when a specific protein is needed in the cell.

• A specific mRNA migrates from the nucleus to ribosomes for the process of translation.



Sample mRNA molecule

2. Transfer RNA (tRNA)

• There are a variety of tRNA molecules in the cytoplasm of the cell.

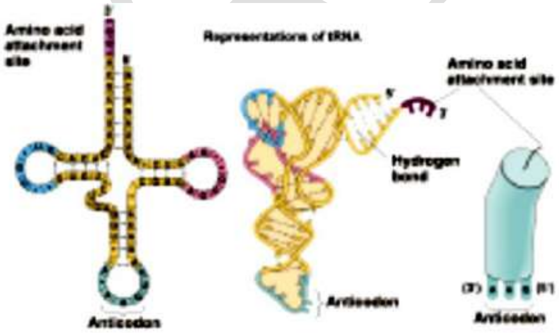
• Each tRNA has 3 hairpin loops in which the RNA is folded back on itself and makes hydrogen bonds. A fourth "leg" also base-paired where the ends of the tRNA are adjacent to each other.

• Each different type of tRNA has two important pieces:

1. An amino acid attachment site at the 3' end, which can attach to one specific amino acid

2. A tRNA triplet sequence which pairs with one specific mRNA triplet sequence found on the 2nd hairpin loop. This triplet specifies the precise amino acid that attaches to the attachment site of the tRNA.

• tRNA is the critical connection between the information carried on the DNA and the amino acids which will be assembled into proteins



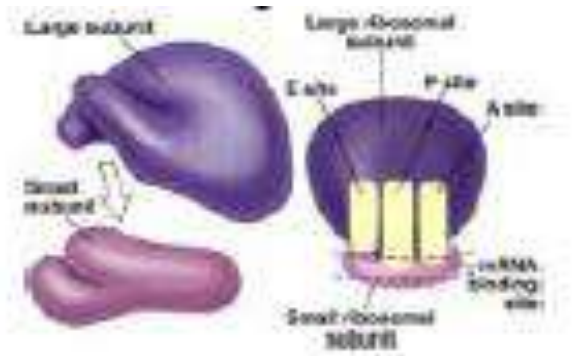
Ribosomal RNA (rRNA)

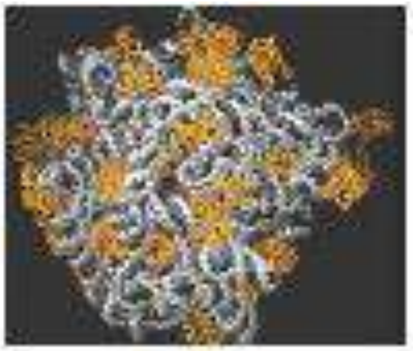
• Component, with protein, of the ribosomes.

• A ribosome is composed of 2 subunits, a small subunit containing RNA molecules plus proteins, and a larger subunit containing RNA plus proteins and the enzymes needed for protein synthesis.

• The ribosomal subunits are manufactured in the nucleolus, but the complete ribosome is found in the cytoplasm, frequently attached to rough endoplasmic reticulum.

• The small rRNA subunit has a binding site for mRNA molecules during protein synthesis. The larger subunit has three attachment sites for tRNA molecules, the P site (Peptidyl-tRNA site), the A site (AminoacyltRNA site) and E site (Exit site). During protein synthesis the two subunits bind together.





Other kinds of RNA (which we will discuss during the details of transcription and translation):

• **pre-RNA**, which is the RNA transcribed from DNA prior to processing to a functional RNA molecule

• **snRNA** (small nuclear RNA) is a component of **snRNP** ("snurps") or small nuclear ribonucleoproteins, which function in mRNA processing. snRNPs, along with additional proteins, form **spliceosomes**

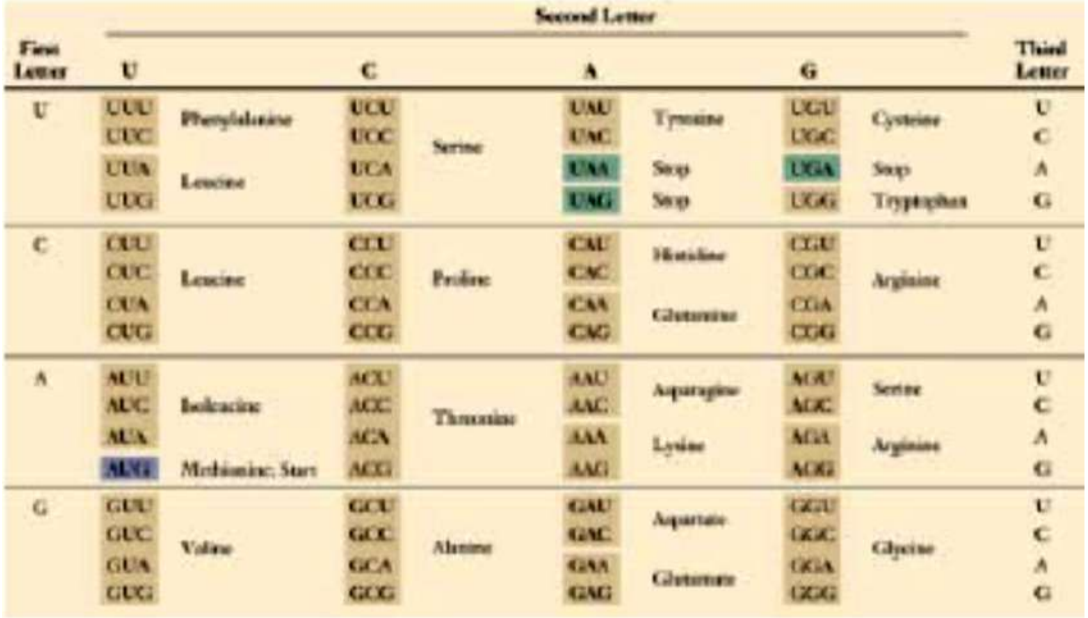
• **SRP RNA**, (Signal Recognition Particle RNA) is part of the signal-recognition particle that helps move ribosomes to the ER from the cytosol.

The Genetic Code: DNA and RNA at Work - Overview

The information of DNA is coded into three-nucleotide long sequences (the triplet code). Each triplet sequence of nucleotides in a DNA molecule is a "code word" for one specific amino acid. DNA molecules contain a linear sequence of triplets that will specify which amino acids a protein will contain, and the sequence, or order, in which these amino acids will peptide bond to form a polypeptide. Moreover, the code is non-overlapping and lacks separators, or punctuation, between the triplets. Much of the work of figuring out the code was done by Francis Crick and colleagues of his in 1961. They studied the impact of deleting one, two or three nucleotides on gene transcription. With one or two deletions, the portion of the gene transcribed past the point of the deletion was nonsense. With three deletions, correct gene transcription was restored. This is known as **frame-shift** alterations of the genetic code. They also looked at the possibility that each code word would be separated by a "punctuation mark". They found this not to be so using the same process.

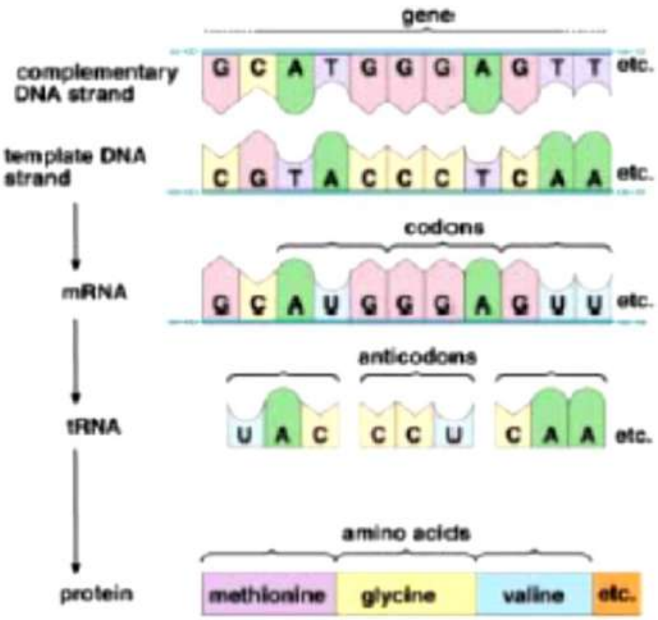
The DNA sequence for a single protein will have 3 times the number of nucleotides as the number of amino acids in the protein for which the sequence codes. In addition, there will be start and stop regions of the DNA associated with these instructions for protein synthesis, and regions within the DNA molecule that do not code and are removed by RNA processing after transcription. Although there can be 64 different DNA code words, three of them are "nonsense" and do not code for specific amino acids. The three "nonsense" code words specify the end of a polypeptide coding. A mRNA nucleotide triplet (synthesized from a DNA template) that codes for a specific amino acid is a complementary (rather than identical) triplet **codon** to the DNA**.** Synthesis of RNA follows the same nitrogen base pair rules that dictate DNA replication.

The codons for each of the amino acids are known, as well as specific codons that are used as start and stop messages. This was accomplished in the 1960's by Nirenberg, who synthesized artificial RNA molecules comprised of just one nucleotide. Nirenberg used first poly-U. In a mixture of ribosomes, amino acids and protein-synthesizing enzymes, a polypeptide would be formed consisting of just phenylalanine and no other amino acid. Poly-A resulted in a polypeptide of just lysine. Eventually codes for all of the amino acids were determined. When you look at Codon tables, you will see that some amino acids are coded for by more than one codon. Often, only the first two nucleotides of the triplet are essential; the third is redundant. (e.g., CCU, CCC, CCA and CCG all code for the amino acid, proline, and UCU, UCC, UCA and UCG all code for the amino acid, serine.) But the codon always contains the entire triplet. The reverse is not true. One codon can never code for more than one specific amino acid. UCU codes for serine. UCU can never code for any other amino acid.



	Second Letter								
First Letter	U		C		A		G		Third Letter
U	UUU	Phenylalanine	UCU	Serine	UAU	Tyrosine	UGU	Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	Leucine	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Tryptophan	G
C	CUU	Leucine	CCU	Proline	CAU	Histidine	CGU	Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Isoleucine	ACU	Threonine	AAU	Asparagine	AGU	Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lysine	AGA	Arginine	A
	AUG	Methionine; Start	ACG		AAG		AGG		G
G	GUU	Valine	GCU	Alanine	GAU	Aspartate	GGU	Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glutamate	GGA		A
	GUG		GCG		GAG		GGG		G

Although we say that the DNA triplet code is universal, there are exceptions, notably in mitochondria and chloroplast DNA. Some mitochondria codons are read differently – and you would need a special codon table that have codons that code for amino acids that are different than the typical amino acids. The stop codes are also different. It may have been a protective thing for endosymbionts early on in evolution. The process of **translation** requires triplet sequences of tRNA that match the mRNA for the amino acid assembly. The only way to match nucleotides is by base pairs, which are complements to each other, so the tRNA triplet that codes for and attaches to a specific amino acid is often called the **anticodon.** Each tRNA has an amino acid binding site along its stem, which can attach to its specific amino acid. Specific enzymes do this. These attachment sites are also phosphorylated to provide the energy for protein synthesis. Codon-anticodon (mRNA-tRNA) matches occur at ribosomes where the amino acids, which are attached to the tRNA molecules, can be joined by peptide bonds to form polypeptides. Several ribosomes can function at one time so that several copies of a polypeptide can be made at one time.



How it works: Details of Process of Transcription

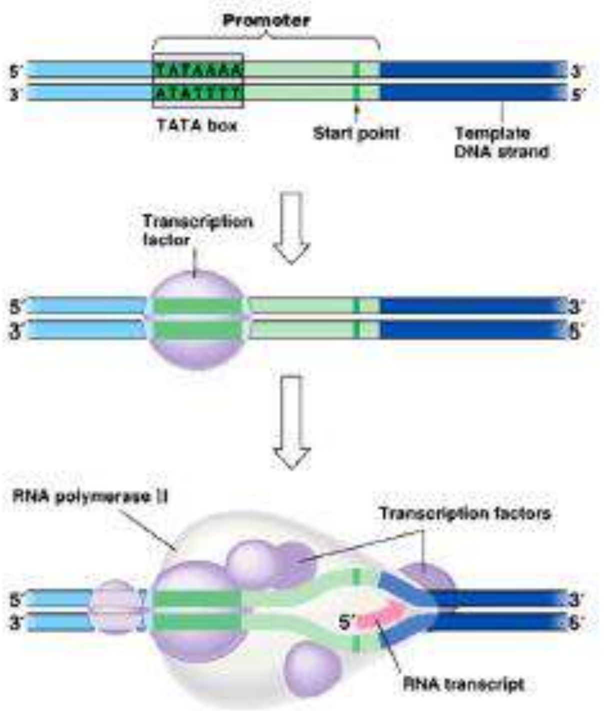
RNA synthesis uses DNA as a template, and occurs in the nucleus. There are three stages in transcription: **Initiation, Elongation** and **Termination.**

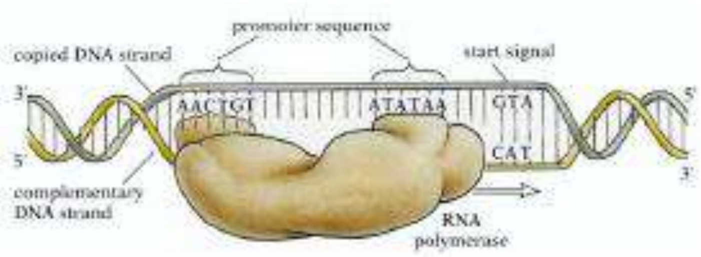
Initiation - • The region of DNA which codes for the specific gene to be transcribed, called the **transcription unit**, starts to unwind using the enzyme, **RNA polymerase**, to **initiate** the process. There are three forms of RNA polymerase, one for each kind of RNA molecule. RNA polymerase II transcribes mRNA. RNA polymerase is one of the most complex enzymes known. It has binding sites for: • regulatory proteins, • the DNA template strand, • the RNA nucleotide subunits, • the promoter region of the gene to be transcribed

• Transcription is started at regions of the DNA molecule called **promoters,** specific DNA base sequences at the 3' of each gene. A promoter determines the template strand of the DNA and where transcription will start. Special proteins, called **transcription factors**, help RNA polymerase find the promoter regions on the DNA. You will recall that many of the signal transduction pathways discussed earlier result in the production of transcription factors.

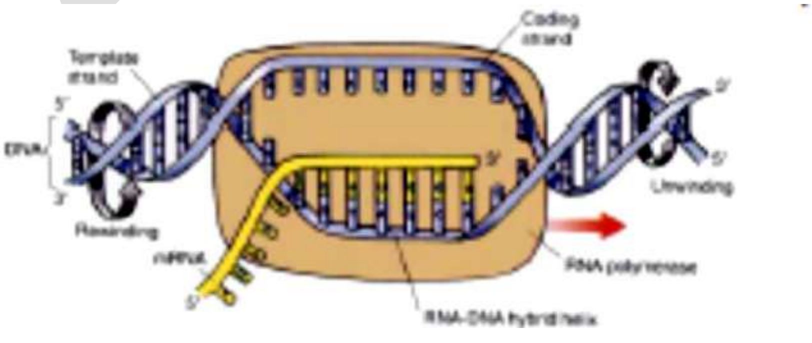
The promoter region DNA-RNA polymerase-transcription factors complex is called the **transcription initiation complex.** One example of a promoter that includes a specific DNA sequence is the **TATA box**, comprised of TATAAA. Binding of RNA polymerase to the transcription factors at the TATA box is crucial to RNA synthesis.

Rate of transcription is related to the efficiency of promoters. Some bacterial promoters facilitate very rapid transcription (every 2 seconds) while weaker promoters may transcribe every 10 minutes. Weaker promoters may have alterations in their recognition sequences (the TATA box region).





Elongation - • RNA polymerase will move in the 3' to 5' direction along the DNA template during what is now called the **elongation** process of transcription. Like DNA, RNA is synthesized in the 5' to 3' direction from the 3' to 5' DNA template.



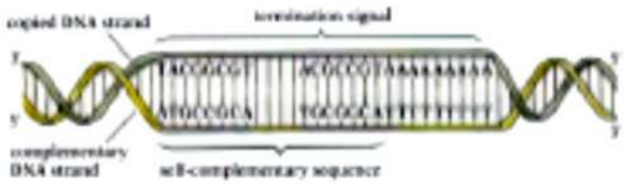
• RNA Nucleotides are added to the chain according to the complementary base pairing; that is: RNA A - DNA T, RNA U - DNA A, RNA C - DNA G, RNA G - DNA C

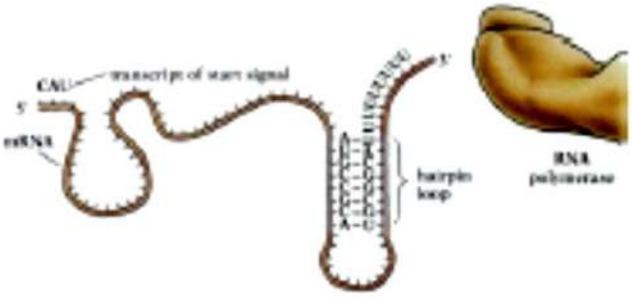
• Several RNA enzymes can be present so that several RNA transcripts can be made of the gene (DNA sequence) at one time. As one mRNA is being transcribed, a new RNA polymerase molecule attaches to its transcription factors at the promoter and starts transcribing a second. As the second starts elongating, a third RNA polymerase can attach, until many, many mRNA molecules are being synthesized along the DNA template.

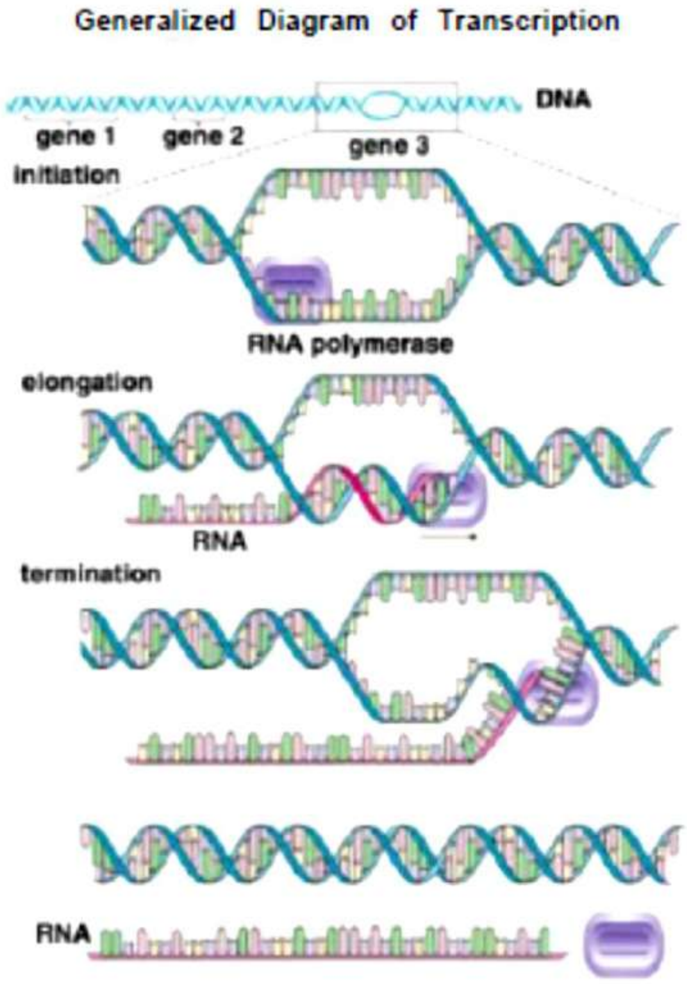
• The DNA strand that will be read by RNA polymerase is called the **template strand**. The opposite strand, which is not read, is called the **coding strand**, because its sequence will be the same as the mRNA transcript being formed, with the exception of T on the DNA coding strand, and U on the RNA transcript.

Termination - • There is a **terminator** sequence that tells the RNA polymerase to stop. The actual termination and separation of the RNA polymerase molecule is several nucleotides beyond the nucleotide stop sequence in eukaryote organisms. A typical terminator sequence is one that forms a hairpin loop in the mRNA

followed by a poly-U sequence, a weak pairing partner for the DNA template. The hairpin forces RNA polymerase to pause, and the weak U-A bonds dissociate freeing the RNA polymerase.





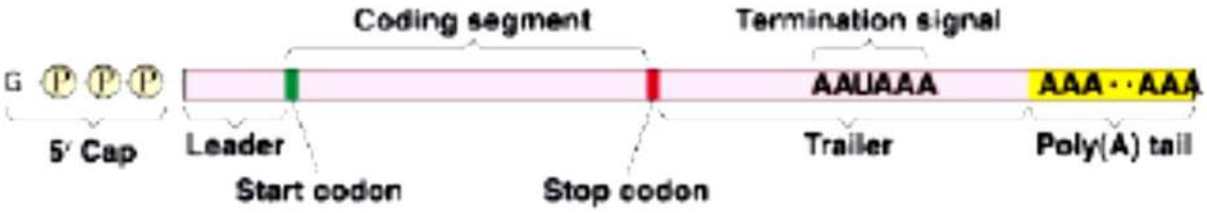


Modifying the pre-RNA for use – RNA Processing Cap and Tail

• After transcription, a cap (made of GTP) is attached to the 5' end of the RNA molecule (the end made first). The cap is called the 5' cap.

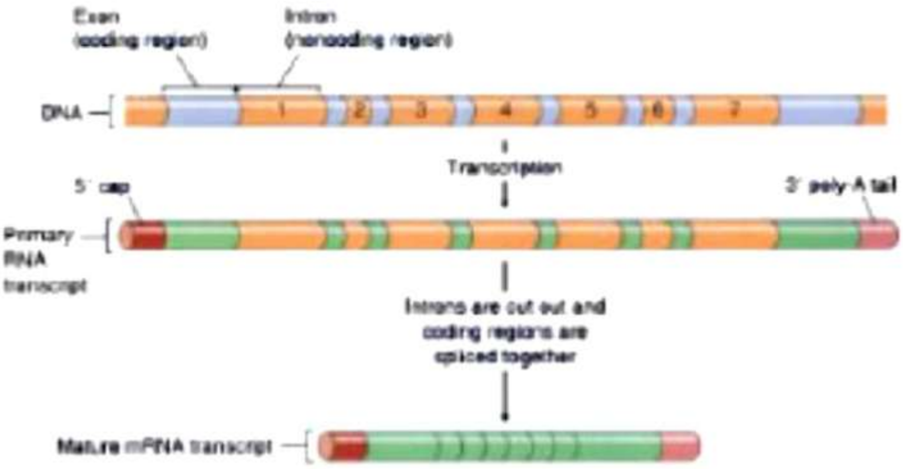
• Often, in eukaryotes, the 3' end of a transcript is removed and a tail of 30 - 300 poly-A (adenine) nucleotides is attached to the 3' end of the mRNA transcript. This tail is called the **3' poly-A tail**.

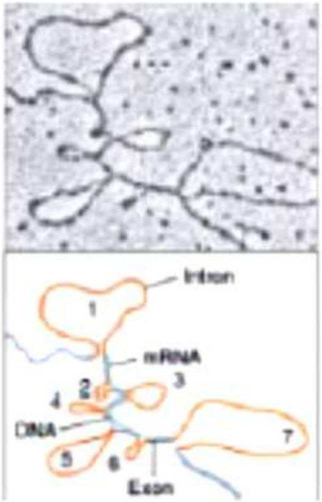
• Both cap and tail help the mRNA attach to the ribosome for translation, and also inhibit enzyme degradation of the mRNA transcript.



RNA Splicing – Introns and Exons

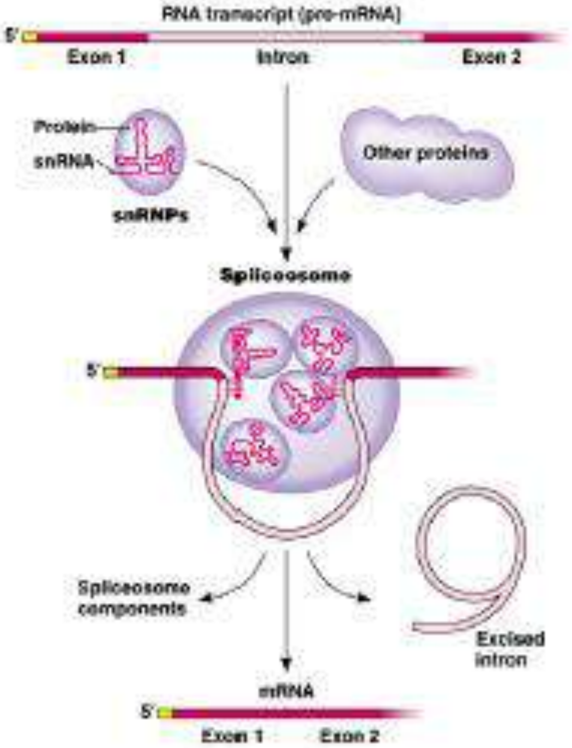
The RNA molecule transcribed consists of far more nucleotides than are actually used in protein synthesis. Some parts of the transcribed gene, called **introns**, do not code for amino acids. (No one knows why.) The name introns is derived from the fact that the introns are intervening segments that interrupt the message. Those regions that do code are called **exons** (because they are expressed). As much as 90% of the pre-mRNA transcript may be non-protein coding nucleotides. And only about 25% of the entire genome is coded at all. Exons may comprise as little as 1% of our DNA. Prior to use, introns must be removed from the pre-mRNA transcript, which is done during the RNA processing stage. This process is called **RNA splicing**, because the introns are cut out and the remaining exons get spliced together. (Richard Roberts and Phillip Sharp won the Nobel prize in 1993 for this discovery.)





• The process of gene splicing takes place at **spliceosomes** in the nucleus. Small pieces of RNA-protein complexes, called **snRNPs** or "snurps" (for small nuclear ribonucleoprotein) recognize specific codes on the pre-mRNA at the ends of introns. Several snRNPs aggregate with additional protein to form the splicesome bodies.

Splicesomes are often thought to be similar to ribosomes, small granules of RNA and protein that reside in the nucleus, into which the pre-mRNA fits for intron removal.



It is believed that snRNAs in the snRNPS serve as the catalysts for the intron excision and splicing of the exons, as well as in the formation of spliceosomes and the exon splice sites. This is unusual, since our familiar catalysts in living organisms are all enzymes. It is not unheard of, however. Some RNA molecules do function as catalysts. Such RNA molecules are called **ribozymes**. Ribozymes have been shown to process pre-RNA for ribosomal RNA (rRNA) in some organisms. In these cases, the intron itself is the ribozyme, and the introns are removed by "self-splicing". Introns found in prokaryotes are self-splicing. (Introns are rare or absent in prokaryote DNA.)

The completed mRNA (with a precise linear sequence of nucleotides) moves from the nucleus to the cytoplasm for translation. It should be noted that with the number of modifications that can occur to a mRNA transcript, that one "gene" can, and some cases, does code for more than one polypeptide. This may be why our 30,000 genes may code for as many as 120,000 different proteins. It should also be noted that not all biological catalysts are proteins, since snRNA can function as a catalyst in RNA transcript processing.

Why Introns?

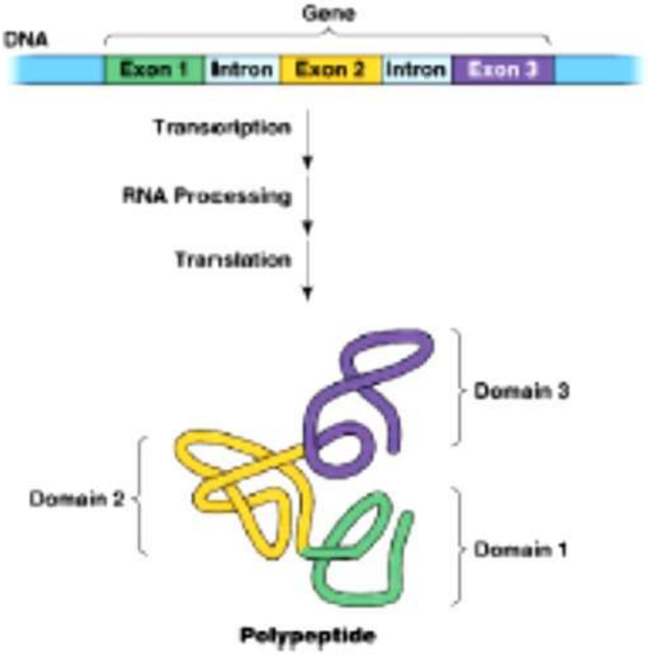
One area of interest is why so much of the DNA molecule contains introns. There is much research in this field.

• As mentioned above, one pre-mRNA can be used to code for more than one protein, depending on what is determined to be introns in the processing stage. The proteins that determine gender development in fruit flies have been shown to share a common pre-mRNA.

• Some introns may regulate transcription. There are some genes which are inhibited when RNA binds to the DNA molecule.

• Introns may help regulate the movement of the RNA from the nucleus to the cytoplasm of the cell.

• Introns may help in modifying protein shape. Different introns can affect the location of an active site for an enzyme or the attachment site for a membrane protein. This may affect change in proteins through time and may be involved with protein **domains**. For example, the active site is one protein domain. The attachment site for a co-factor is a second domain, and attaching the enzyme to a membrane surface a third domain. Introns can be used to modify domains without having to have the remainder of the protein's instructions recoded.



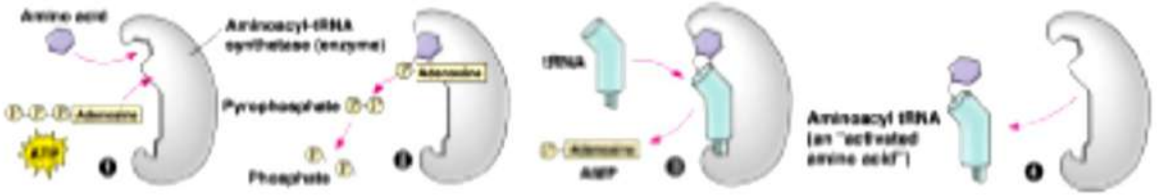
• It is also believed that length of introns affects rate of recombination. Long introns give more room for crossing over, hence more genetic variation.

Protein Synthesis – The Process of Translation

Once we have a final mRNA transcript, the mRNA is moved from the nucleus of the cell to the cytoplasm where is attaches to a small subunit of a ribosome and we are ready for the process of **translation**. Translation is where the information coded in DNA molecules is interpreted and translated to direct the actual synthesis of proteins. Translation also involves **Initiation, Elongation** and **Termination.**

Amino Acid Attachment

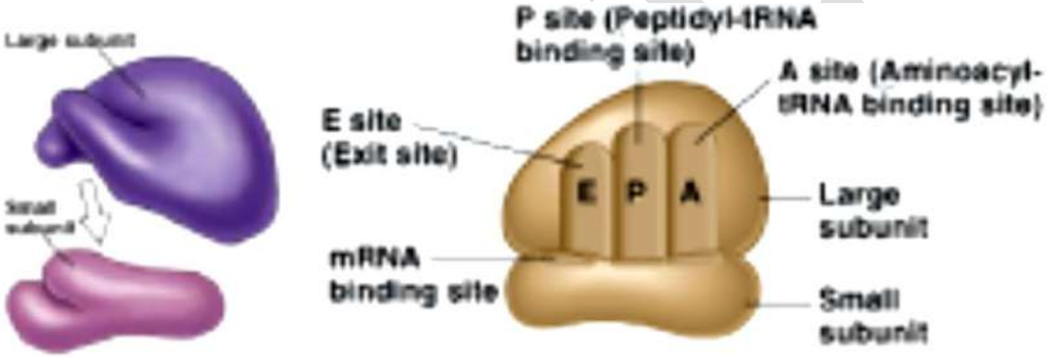
Prior to translation, one additional activity must occur: amino acids must be attached to their appropriate tRNA molecules. The process of **Amino acid attachment** involves ATP and a set of **aminoacyl-tRNA synthetase** enzymes. ATP loses two of its phosphates in the process and AMP is complexed to the amino acid-tRNA. Both the shape and charge of the tRNA molecules and the amino acids are important for the correct recognition and attachment of its specific aminoacyl-tRNA synthase enzyme. It is the job of the amino acid activating enzymes to correctly interpret the genetic code and attach the appropriate amino acid to its corresponding tRNA. There are 20 different activating enzymes, one for each of the different amino acids.



Although theoretically there should be 64 different tRNAs, one for each triplet code word other plus the 3 stop triplets, there are only about 45. As mentioned, the triplet code for DNA - amino acids is redundant. Often the third nucleotide is not crucial. Some tRNA molecules have a modified adenine as the third base, called **inosine**, that can base pair with any of the other nucleotide bases, so that these tRNA molecules can recognize more than one mRNA codon. The tRNA molecule that attaches to the amino acid, leucine, for example, recognizes a codon that is CU_. The third nucleotide on the mRNA codon does not matter. This tRNA "flexibility" is called **wobble**.

Initiation

• The small rRNA subunit has a binding site for mRNA molecules during protein synthesis and the initiator tRNA. The larger rRNA subunit has three attachment sites for tRNA molecules, the P site (Peptidyl-tRNA site), the A site (Aminoacyl-tRNA site) and E site (Exit site). During protein synthesis the two subunits bind together.

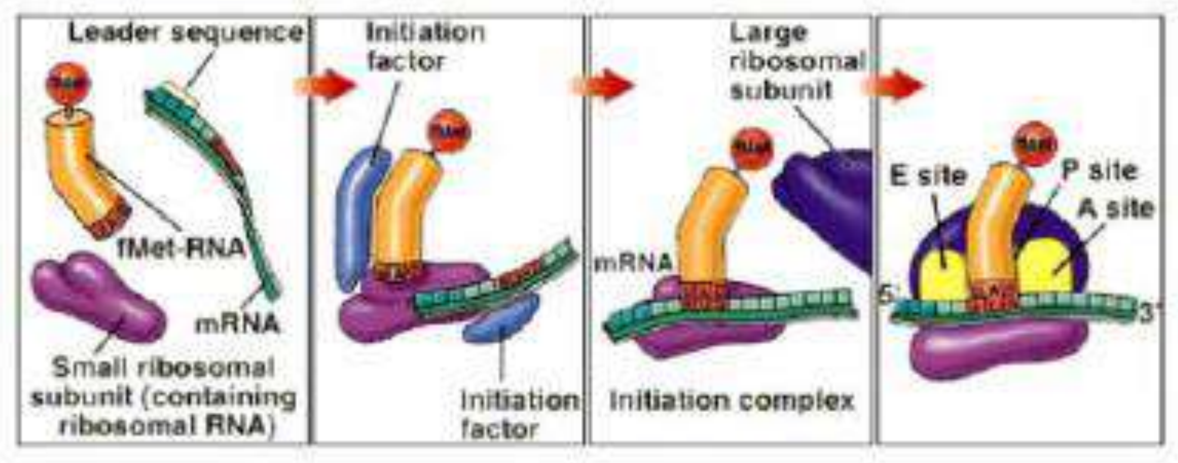


• To initiate protein synthesis, protein initiation factors bind the 5' leader of the mRNA to the small ribosome. The tRNA that has the "initiator" anticodon, UAC and its amino acid, N-formyl-methionine (fMet) in prokaryotes or methionine in eukaryotes, attach to the start codon of the mRNA that is located some distance beyond the 5' cap.

• The large ribosomal subunit binds to the small subunit, and the initiator tRNA attaches to the **P site** of the large subunit of the ribosome with the assistance of **protein initiation factors**, bringing the complex together and forming a functional ribosome. **GTP** provides the energy for the initiation process.

• A functioning ribosome is large enough to hold three mRNA codons. As stated, the first tRNA with its amino acid attaches to the P site. The A site of the larger subunit will be available for the 2nd tRNA molecule's anticodon to bind to the 2nd mRNA codon during elongation. The third codon site is the exit site.

• Note: Polypeptide synthesis is initiated at the amino end of the chain. Amino acids can only be added to the carboxyl end of an amino acid on the ribosome.



Elongation

Elongation involves three activities: **codon recognition** by tRNA molecules, **peptide bonding**, and **translocation** of the ribosome.

• **Codon Recognition**

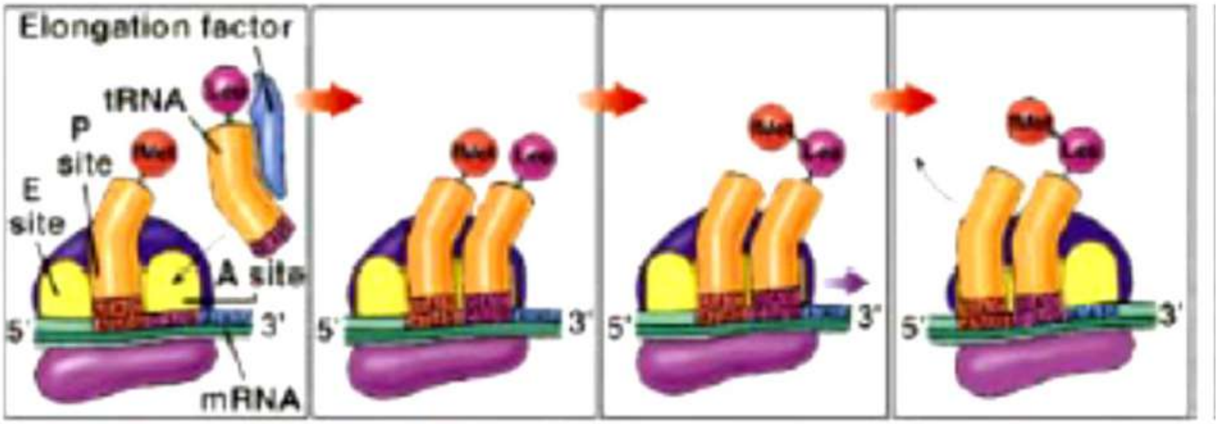
The next tRNA molecule, with its attached amino acid, is brought into place at the ribosome's **A site** with the assistance of proteins called **elongation factors** and **GTP**, which provides energy, as determined by the mRNA codon message. The tRNA anticodon hydrogen bonds to the mRNA codon at this time.

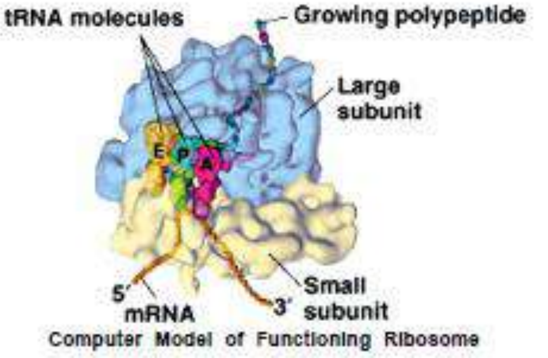
• **Peptide Bonding**

The positioning of the two tRNA molecules (each with its proper amino acid) at the P and A sites is such that a peptide bond can be formed between the two amino acids that are attached to their respective tRNAs. rRNA functions as a ribozyme to catalyze the peptide bond between the amino acid from the P site to the amino acid at the A site at the peptide bonding site on the ribosome. The peptide bond forms between the carboxyl end of the P site amino acid and the amino end of the A site amino acid. This process detaches the P site amino acid from its tRNA; the first amino acid attaches to the second amino acid at the A site. The polypeptide chain always elongates at the A site.

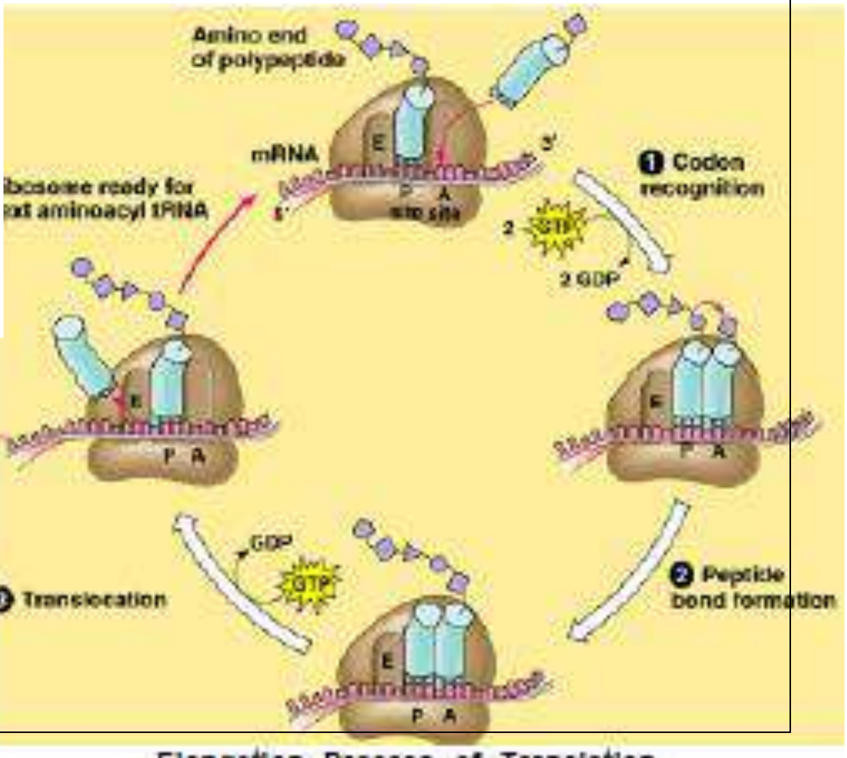
• **Translocation**

Once the peptide bond is formed, the ribosome will shift the A-site tRNA to the P site and the P site tRNA will shift to the E site to be dislodged from the ribosome (which is why the E site is called the exit site). **GTP** is required for this process. Because the mRNA is still attached to the tRNA on P site, the mRNA is moved along with the tRNA molecules locating the 3rd mRNA codon at the now vacant A site. The 3rd tRNA will be brought into the A site by elongation factor proteins. The codon-anticodon binding, peptide bonding, detachment of tRNA and shifting continues until all of the codons of the mRNA have been matched by tRNA anticodons. Note that the mRNA moves along the ribosome with its 5' end leading. mRNA moves only in one direction. Ribosomes and mRNA move relative to each other, codon by codon, unidirectionally.





Computer Model of Functioning Ribosome

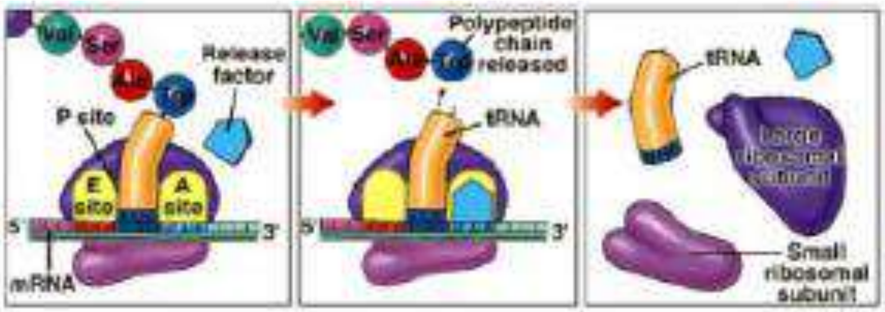


Elongation Process of Translation

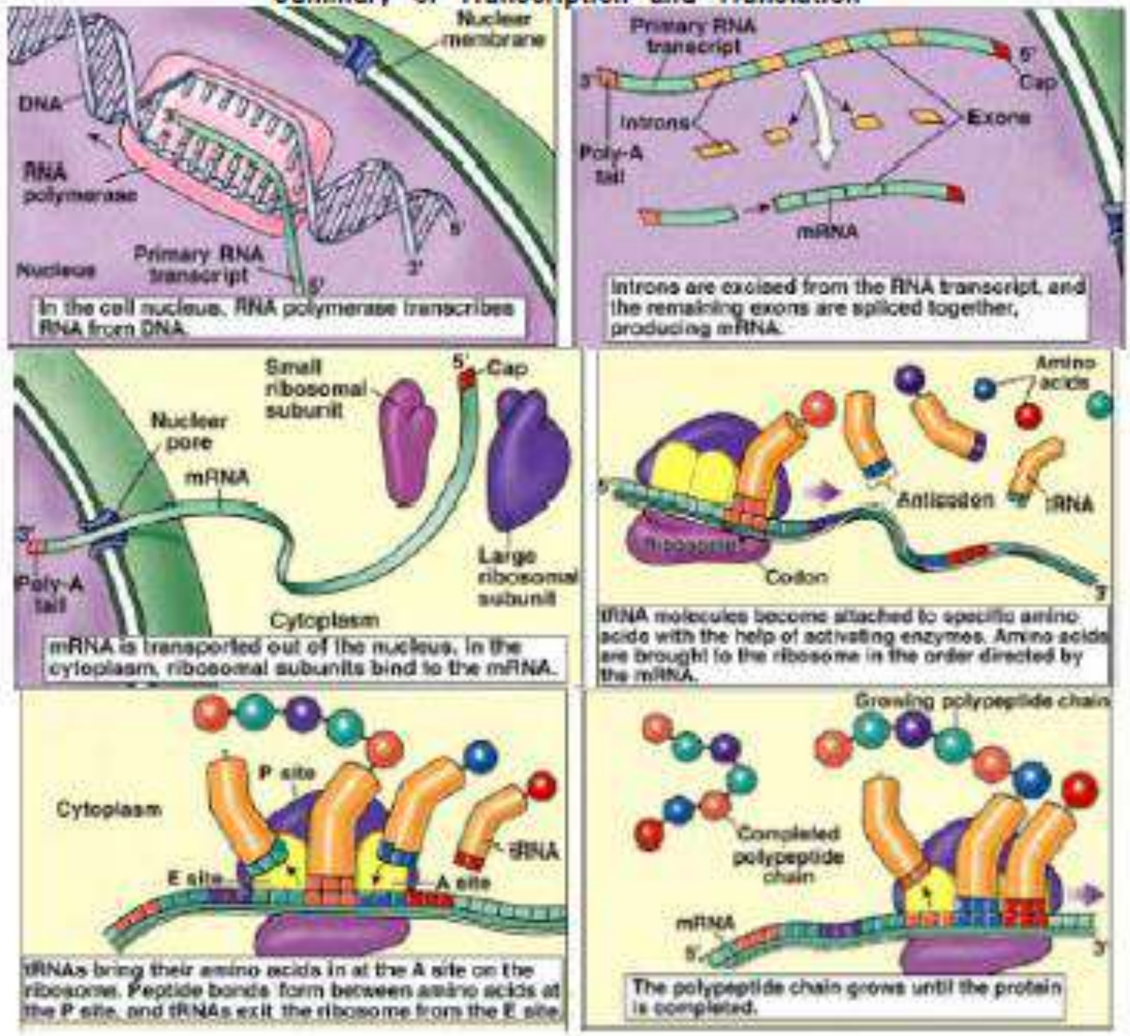
Termination

• The mRNA has a stop codon (UAA, UAG or UGA) which prevents any more tRNA from attaching to the A site. A **releasing factor protein** attaches instead and hydrolyzes the polypeptide causing it to be released from the ribosome.

• The ribosomal subunits and related proteins dissociate.



Summary of Transcription and Translation



In the cell nucleus, RNA polymerase transcribes RNA from DNA.

Introns are excised from the RNA transcript, and the remaining exons are spliced together, producing mRNA.

mRNA is transported out of the nucleus. In the cytoplasm, ribosomal subunits bind to the mRNA.

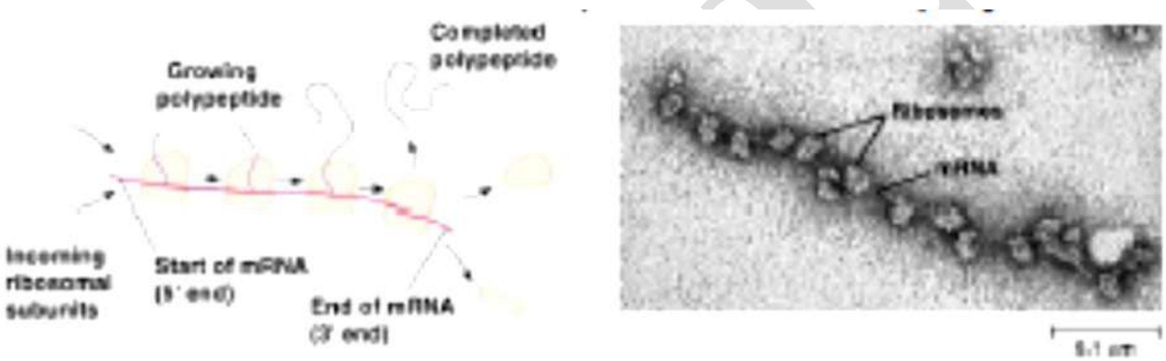
tRNA molecules become attached to specific amino acids with the help of activating enzymes. Amino acids are brought to the ribosome in the order directed by the mRNA.

tRNAs bring their amino acids in at the A site on the ribosome. Peptide bonds form between amino acids at the P site, and tRNAs exit the ribosome from the E site.

The polypeptide chain grows until the protein is completed.

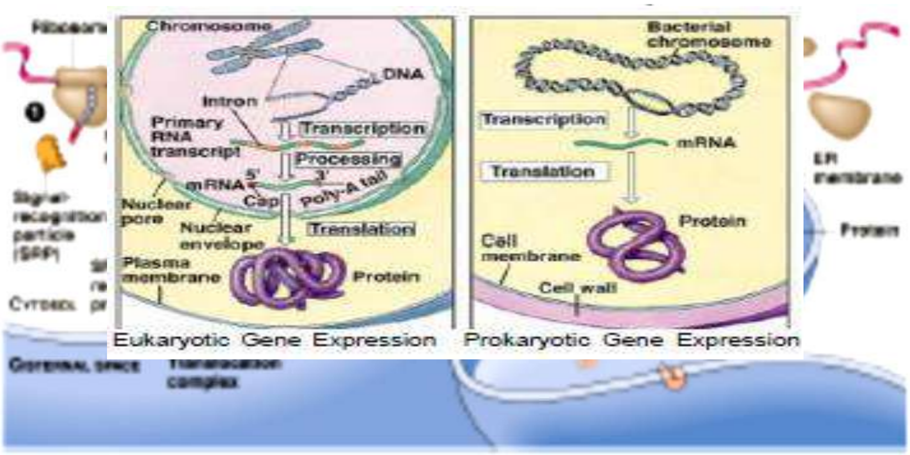
Rate of polypeptide synthesis

A polypeptide is generally synthesized in about a minute. However, it is typical of mRNA to be working along many ribosomes at a time to direct the synthesis of many polypeptide molecules in sequence. As soon as the 5' end of a mRNA leaves one ribosome it will attach to the small subunit of an adjacent ribosome to initiate protein synthesis at that ribosome. It is common for one mRNA to have many ribosomes associated at once. Such complexes are called **polyribosomes**.



Modifying Polypeptides into Functional Proteins

As discussed in earlier units, the secondary, tertiary and quaternary structure of proteins follows the polypeptide synthesis to obtain a functional protein conformation. Other modifications are typically made as well, adding other molecules, removing some amino acids from the chain, etc. mRNA initially attaches to free ribosomes (and polyribosomes) in the cytosol. Proteins that will function in the cytosol are synthesized at free ribosomes. However, at some point in the polypeptide synthesis proteins that are associated with the endomembrane system, or destined for export from the cell, will associate with the ER for completion. These growing polypeptides have a **signal peptide sequence** (of about 20 nucleotides) located near the start of the growing polypeptide that is recognized by **signal-recognition particles (SRPs).** SRPs are RNA-protein particles. The SRP brings the ribosome to a receptor protein site on the ER. As protein synthesis continues on the attached ribosome, elongating polypeptides move through ER pores into the ER cisternal spaces. Similar SRPs move completed polypeptides to mitochondria, chloroplasts, nuclei and other organelles after translation has been completed on free ribosomes.



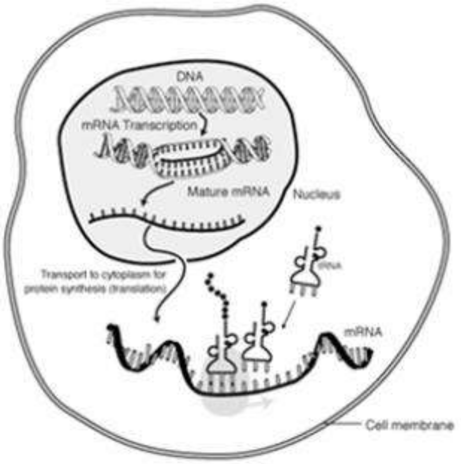
Gene Expression in Prokaryotes and Eukaryotes

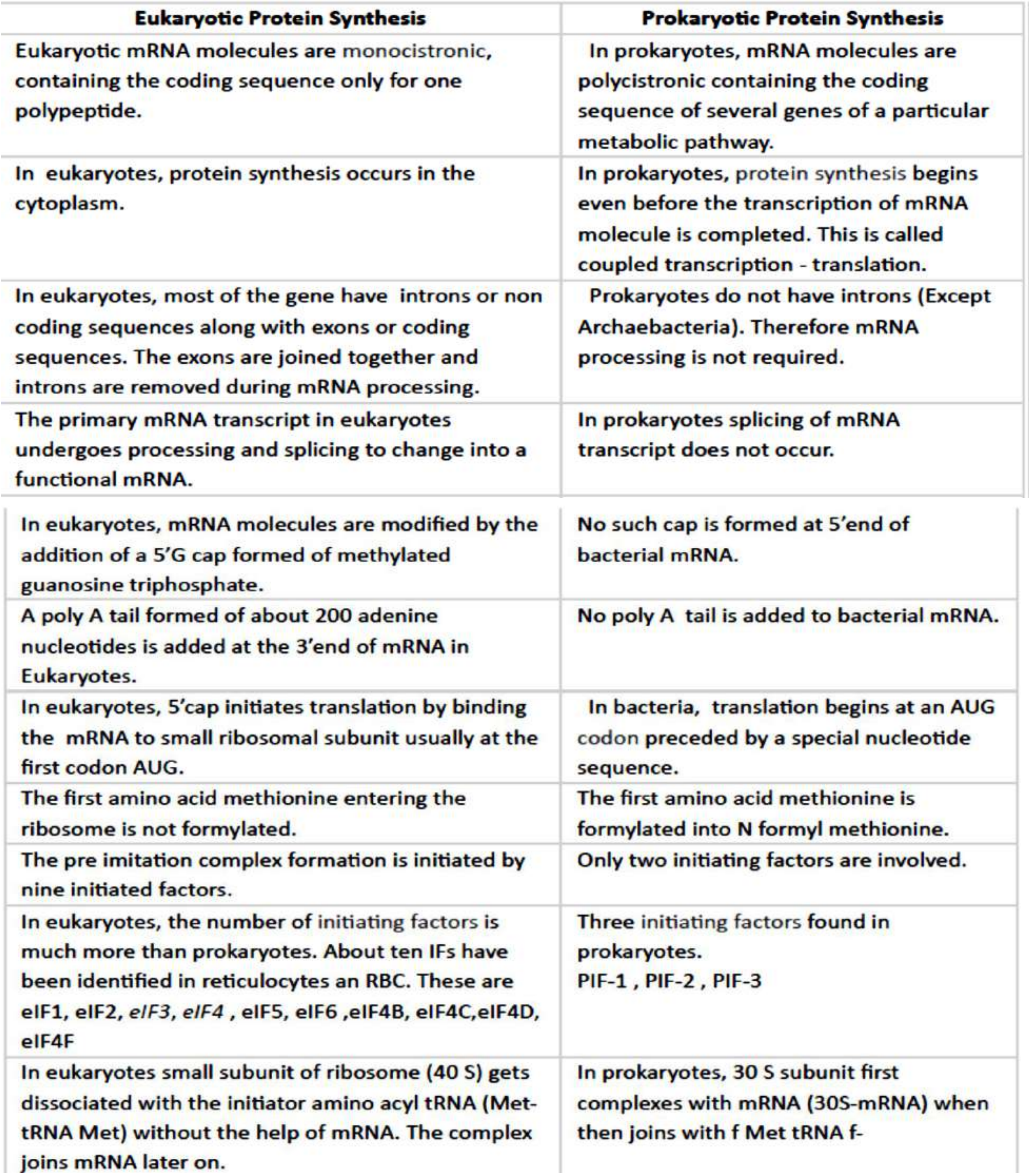
Although we have been discussing protein synthesis as a general concept, most of what we have addressed applies specifically to eukaryotes. There are some differences between protein synthesis in eukaryotes and in prokaryotes. • The prokaryotic RNA polymerase is different in structure from the eukaryotic RNA polymerase and does not depend on transcription factors. • As stated earlier, introns, if present in prokaryotes, are removed during transcription so virtually no mRNA processing is required prior to translation. • Prokaryote mRNA molecules may carry code for many proteins, not just one polypeptide as found in eukaryotes. • Because prokaryotes have no separation of nucleus and cytosol, translation can be initiated as soon as the growing mRNA transcript is freed from the DNA. • All prokaryote translation begins with the AUG codon, which follows a special non-translated nucleotide sequence. Prokaryotes do not add a 5' GTP cap as is found in the eukaryotes. (Makes sense since prokaryotes have no posttranscription processing of the mRNA)

Comparison of prokaryotic and Eukaryotic protein synthesis

Protein synthesis is the process in which cells build proteins. The term is somemes used to refer only to protein translaon but more oen it refers to a mulstep process in which cells follow a very systemac procedure that first transcribes DNA into mRNA and then translates the mRNA into chains of amino acids. The amino acid chain then folds to form funconal proteins.

Eukaryotic Protein Synthesis vs Prokaryotic Protein Synthesis





Eukaryotic Protein Synthesis	Prokaryotic Protein Synthesis
Eukaryotic mRNA molecules are monocistronic, containing the coding sequence only for one polypeptide.	In prokaryotes, mRNA molecules are polycistronic containing the coding sequence of several genes of a particular metabolic pathway.
In eukaryotes, protein synthesis occurs in the cytoplasm.	In prokaryotes, protein synthesis begins even before the transcription of mRNA molecule is completed. This is called coupled transcription - translation.
In eukaryotes, most of the gene have introns or non coding sequences along with exons or coding sequences. The exons are joined together and introns are removed during mRNA processing.	Prokaryotes do not have introns (Except Archaebacteria). Therefore mRNA processing is not required.
The primary mRNA transcript in eukaryotes undergoes processing and splicing to change into a functional mRNA.	In prokaryotes splicing of mRNA transcript does not occur.
In eukaryotes, mRNA molecules are modified by the addition of a 5'G cap formed of methylated guanosine triphosphate.	No such cap is formed at 5'end of bacterial mRNA.
A poly A tail formed of about 200 adenine nucleotides is added at the 3'end of mRNA in Eukaryotes.	No poly A tail is added to bacterial mRNA.
In eukaryotes, 5'cap initiates translation by binding the mRNA to small ribosomal subunit usually at the first codon AUG.	In bacteria, translation begins at an AUG codon preceded by a special nucleotide sequence.
The first amino acid methionine entering the ribosome is not formylated.	The first amino acid methionine is formylated into N formyl methionine.
The pre imitation complex formation is initiated by nine initiated factors.	Only two initiating factors are involved.
In eukaryotes, the number of initiating factors is much more than prokaryotes. About ten IFs have been identified in reticulocytes an RBC. These are eIF1, eIF2, *eIF3*, *eIF4* , eIF5, eIF6 ,eIF4B, eIF4C,eIF4D, eIF4F	Three initiating factors found in prokaryotes. PIF-1 , PIF-2 , PIF-3
In eukaryotes small subunit of ribosome (40 S) gets dissociated with the initiator amino acyl tRNA (Met-tRNA Met) without the help of mRNA. The complex joins mRNA later on.	In prokaryotes, 30 S subunit first complexes with mRNA (30S-mRNA) when then joins with f Met tRNA f-

Protein Targeting and Degradation

Protein turnover – that is synthesis and degradation – occurs constantly in eukaryotic cells but it is a highly selective process with different rates of turnover for various proteins. Turnover of proteins can control the level of certain enzymes, furnish amino acids in times of need and degrade faulty or damaged proteins that are generated during synthesis or arise from deleterious activities in the cell. Nascent proteins contain signals that determine their ultimate destination. A newly synthesized protein in the prokaryotic *Escherichia coli* cell, for example, can stay in the cytosol or it can be sent to the plasma membrane, the outer membrane, the space between them, or the extracellular medium.

The eukaryotic cell is made up of many structures, compartments and organelles, each with specific functions requiring different types of proteins and enzymes. The synthesis of most of these proteins begins on free ribosomes in the cytosol. Therefore, eukaryotic cells must direct proteins to internal sites such as lysosomes, mitochondria, chloroplasts, nucleus etc. How then is sorting accomplished? In eukaryotes, a key choice is made soon after the synthesis of a protein begins. A ribosome remains free in the cytosol unless it is directed to the endoplasmic reticulum (ER) by a signal sequence in the protein being synthesized. Nascent polypeptide chains formed by ribosomes are translocated across the ER membrane. In the lumen of the ER, many of them are glycosylated and modified in other ways. These are then transported to Golgi complex where they are further modified. Finally, they are sorted for delivery to lysosomes, secretory vesicles, and the plasma membrane. Transported proteins are carried by vesicles that bulge out from donor compartments and fuse with target compartments. *The signals used to target eukaryotic proteins for transfer across the ER membrane are ancient,* for bacteria also use similar sequences or signals for sending proteins to their plasma membrane and to secrete them. The transported proteins must reach their assigned cellular locations. This is of utmost importance because mistakes in transport can severely affect cellular metabolism and the cumulative effect may prove fatal to an organism. For instance, *I-cell disease*, a rare human disease, is characterized by export from the cell of at least 8 enzymes which should be transported to lysosomes.

SIGNAL HYPOTHESIS

The pathways by which proteins are sorted and transported to their proper cellular locations are referred to as *protein targeting pathways*. A characteristic feature of these targeting pathways (with the exception of cytosolic and nuclear proteins) is the presence of a short amino acid sequence at the amino terminus of a newly synthesized polypeptide called the **signal sequence** or **signal peptide**. In many cases, the targeting capacity of particular signal sequences has been confirmed by fusing the signal sequence from one protein, say protein A, to a different protein B, and showing that the signal directs protein B to the location where protein A is normally found. The signal sequence, whose function was first postulated by David Sabatini and Günter Blobel (1970), directs a protein to its proper location in the cell and is removed by a *signal peptidase* during transport or when the protein reaches its final destination. Obviously, the signal sequence is absent in the protein once secreted.

At present, the signal sequences of more than 100 secretory proteins from a wide variety of eukaryotic species have been determined; some of which have been presented in Figure. A well defined consensus sequence such as the TATA box guiding the initiation of transcription, is not evident. However, signal sequences do exhibit certain common characteristics;

1. They range in length from 13 to 36 amino acid residues.
2. The amino terminal part of the signal contains at least one or more positively charged amino acid residues, preceding the hydrophobic sequence.
3. A sequence of highly hydrophobic amino acids (10 to 15 residues long) forms the centre of the signal sequence. Ala, Val Leu, Ile and Phe residues are common in this region.
4. There is present a region of more polar short sequence (of about 5 residues) at the carboxyl terminus, upstream the cleavage site. The amino acid residues having short side chains (*esp*, Ala) predominate in this region at positions closest to the cleavage site. However, in certain secretory and plasma membrane proteins, the signal sequence is not situated at the amino terminus. These proteins contain an **internal signal sequence** that serves the same role. For example, in the case of ovalbumin, the sequence is located between residues 22 and 41 and is critical for the transfer of nascent albumin across the ER membrane. In 1975, George Palade, at the Rockfeller Institute in New York, demonstrated that proteins with these signal sequences are synthesized on ribosomes attached to the ER membrane. The overall pathway, summarized in Fig. 29–3, proceeds in following 8 steps:

1. First of all, the ribosomal subunits assemble in an initiation complex at the initiation codon and begin protein synthesis.

2. Later, a proper signal sequence appears early in the synthetic process because it is at the amino terminus of the nascent polypeptide.

3. Then, this signal sequence and the ribosome itself are rapidly bound by a large rod-shaped complex called **signal recognition particle (SRP).** This binding event halts elongation and the signal sequence has completely emerged from the ribosome. The SRP receptor isa heterodimer of α (M*r* 69,000) and β (M*r* 30,000) subunits and consists of a 305- nucleotide RNA (called 7 SL-RNA) and 6 different proteins, with a combined molecular weight of 3,25,000. One protein subunits binds directly to the signal sequence, inhibiting elongation by sterically blocking entry of aminoacyl-tRNAs and inhibiting peptidyl transferase.

4. The ribosome-SRP complex with the incomplete polypeptide is bound by two receptors (ribosome receptor and SRP receptor) present on the cytosolic face of the ER. For transport of a polypeptide into the ER lumen, the signal sequence attaches to the **SRP receptor**. The hydrophobicity of the signal sequence is postulated to be the molecular key for the polypeptide's interaction with the ER membrane, which is also a hydrophobic structure. The second recognition site, **ribosome receptor**, serves to anchor the organelle (ribosome) to the ER membrane. The interaction between the signal sequence and the ER membrane is believed to open a channel in the membrane through which the polypeptide is transported into the ER lumen. Thus, the molecular instructions for transport into the ER (in the form of a hydrophobic sequence) are furnished by the polypeptide.

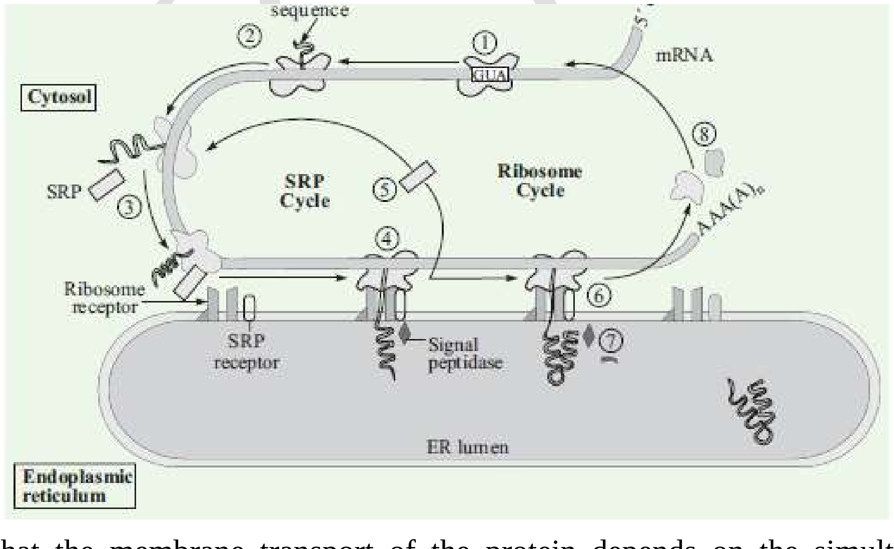
5. The SRP dissociates and is recycled.

6. Protein synthesis then resumes, along with translocation of the polypeptide chain into the lumen of the ER. The nascent polypeptide is delivered to a **peptide translocation complex** in the ER. The translocation complex feeds the growing polypeptide chain into the lumen of the ER in a reaction driven by the energy of ATP.

7. The signal sequence is cleaved by a membrane enzyme, **signal peptidase** which is located on the lumenal side of the ER.

8. Once the complete protein has been synthesized, the ribosome dissociates from the ER and is recycled.

The proteins to be secreted and the lysosomal proteins completely pass through the membrane of the ER. On the contrary, other proteins must form part of a membrane. Such proteins, in the lumen of the ER, are modified in several ways. Besides the removal of signal sequences, polypeptide chains fold and disulfide bonds form. Many proteins are also glycosylated. As a result of about 20 years of strenuous work, Günter Blobel formulated in 1980 general principles for the sorting and targeting of proteins to particular cell compartments. Each protein carries in its structure the information needed to specify its proper location in the cell. Specific amino acid sequences (**topogenic signals**) determine whether a protein will pass through a membrane into a particular organelle, become integrated into the membrane, or be exported out of the cell. In essence, the **signal hypothesis** may be summarized below : Proteins which are to be exported out of the cell are synthesized by ribosomes, associated with the ER. The genetic information from DNA is transferred *via* RNA. This information determines how the amino acids build up the proteins. First, a signal peptide is formed as a part of the protein. With the help of binding proteins, the signal peptide directs the ribosome to a channel in the ER. The growing protein chain penetrates the channel, the signal peptide is cleaved and the completed protein is released into the lumen of ER. The protein is subsequently transported out of the cell.Infact, the signal hypothesis explains how new polypeptides scheduled for intracellular transport are routed into the ER lumen. The signal hypothesis was originally proposed for the transport of secretory proteins. But it is also applicable to storage proteins. An important feature of this hypothesis is that the membrane transport of the protein depends on the simultaneous protein synthesis by the



membrane-bound ribosomes, thus causing the polypeptide to migrate through the tunnel in the endoplasmic reticular membrane. Thus, it may be emphasized that the signal hypothesis is both correct and universal, since the various processes associated with it operate in the same way in yeast, plant and animal cells.

The SRP cycle and nascent polypeptide translocation and cleavage. The signal sequence of a nascent polypeptide chain is recognized by SRP. The complex consisting of SRP, the nascent peptide chain, and the ribosome binds to the SRP receptor in the ER membrane. The ribosome is then transferred to the translocation machinery, which actively threads the polypeptide chain across the ER membrane. SRP released from its receptor is free to bind another emerging signal sequence. The steps conform to the description in the text.

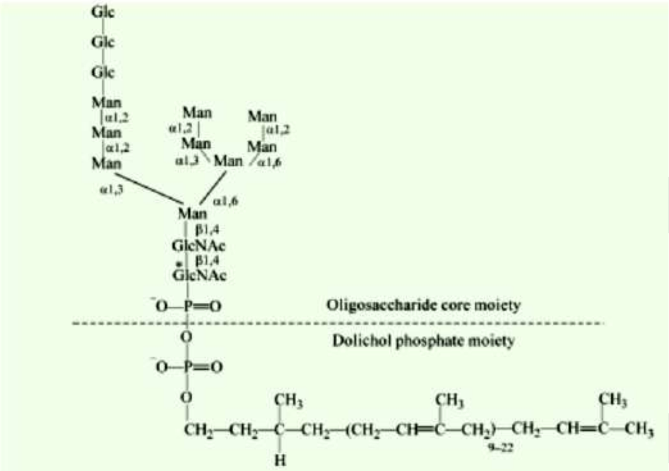
GLYCOSYLATION OF PROTEINS AT THE LEVEL OF ER

While the individual polypeptides are in the ER lumen, the biochemical processes start for their cellular distribution. Whereas hydrophobicity provides the first molecular instructions for intracellular transport, it is **glycosylation** (*i.e.*, addition of carbohydrates) that establishes the molecular patterns acquired by polypeptides to continue their intracellular routing (Armstrong, 1989). The addition of oligosaccharide units, which convert polypeptides into glycoproteins, commences in the ER lumen and continues when they are transported from the ER to the Golgi apparatus. The particular oligosaccharide unit, attached to a glycoprotein, furnishes the molecular instructions for its cellular destination. Acquisition of oligosaccharide units by polypeptides may be compared to the assignment of Pin Codes to mailing addresses with each type of oligosaccharide, representing a distinct Pin Code.

Glycosylation begins soon after a nascent polypeptide enters the ER lumen. Carbohydrates bind to either the amide group of an asparagine (Asn) or the hydroxyl group of a serine (Ser) or threonine (Thr). Oligosaccharides attached to asparaginyl residues are referred to as **N-linked** and those to seryl or threonyl residues as **O-linked.** The following discussion will make it amply clear that, in the case of N-linked glycosylation, the molecular instructions dictating which oligosaccharide unit a protein will attain reside in the sequence and composition of the protein. For example, which asparaginyl residues will be glycosylated and which of the diverse oligosaccharides it will bear.

A. Core Glycosylation

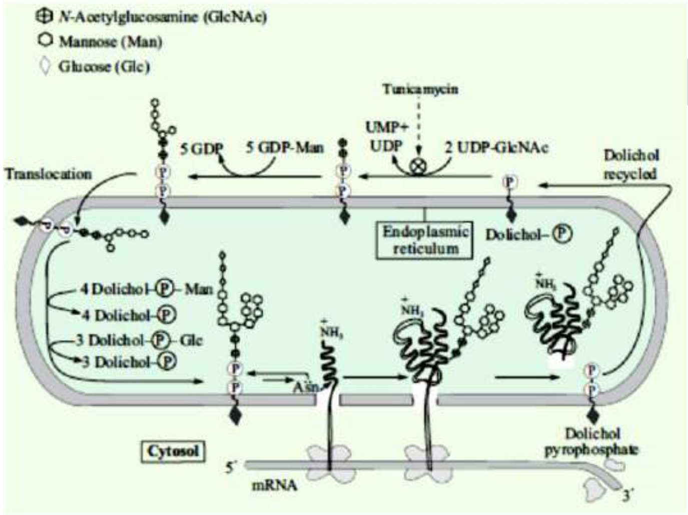
Carbohydrate processing in the ER is called **core glycosylation** to distinguish it from *terminal glycosylation* (described in the subsequent Section), which takes place in the Golgi complex. In the ER lumen, an N-linked oligosaccharide is not added to a polypeptide by a series of one-carbohydrate addition, but instead as an intact unit, called the *common oligosaccharide* core, consisting of 14 residues (2 N-acetylglucosamine + 9 mannose + 3 glucose residues.) However, this oligosaccharide core is constructed by the successive addition of single monosaccharide units to dolichol phosphate (Fig 29–4). Dolichol is an unusually long-chain lipid, containing from 9 to 22 isoprene units.



Activated oligosaccharide core
Dolichol phosphate is a highly hydrophobic lipid carrier, whose terminal phosphate group is the site of attachment of the activated oligosaccharide. Note that the first carbohydrate, *N*-acetylglucosamine, GlcNAc (indicated by an asterisk) is added to the dolichol phosphate moiety as a phosphorylated derivative.

Phosphorylation of dolichol at the nonolefinic end produces dolichol phosphate. Dolichol phosphate is used to carry activated sugars in the membrane-associated synthesis of glycoproteins and some polysaccharides. When the oligosaccharide core is completely synthesized on dolichol phosphate moiety, the whole structure is now called the *activated* oligosaccharide core. The synthesis of activated oligosaccharide core and its transfer to the protein in the ER is depicted in Figure. Once this oligosaccharide core is completely synthesized, it is enzymatically transferred en bloc from dolichol phosphate to a specific asparagine residue of the growing polypeptide chain. The enzyme *transferase* is located on the lumenal face of the ER and thus does not catalyze glycosylation of cytosolic proteins. An asparagine residue can accept the oligosaccharide only if it is a part of an Asn-X-Ser or Asn-X-Thr sequence, and if it is sterically accessible to the transferase. Dolichol pyrophosphate, released in the transfer of the oligosaccharide to the protein, is recycled to dolichol phosphate by the action of a *phosphatase*. After the transfer, the oligosaccharide core is trimmed (*i.e.*, carbohydrates removed) in the ER but all linked oligosaccharides retain a pentasaccharide core derived from the original 14-residue oligosaccharide. Trimming continues when the polypeptides are transferred to the Golgi apparatus. For some polypeptides, trimming produces the required oligosaccharide units; for others, trimming and subsequent addition of new carbohydrates are needed for these polypeptides to acquire their characteristic glycosylated patterns. It is in the Golgi apparatus that most of the final trimming and additions take place (called terminal glycosylation). Several antibiotics interfere with one or more steps in the core glycosylation process. The best-characterized is **tunicamycin**, which blocks the first step (*i.e.*, addition of *N*-acetylglucosamine to dolichol phosphate). Tunicamycin (Fig. 29–6) is a hydrophobic analogue of UDP-*N*acetylglucosamine which blocks the fixation of N-acetylglucosamine on dolichol phosphate, and therefore prevents the glycosylation of proteins. Tunicamycin, thus, mimics UDP-*N*acetylglucosamine. Another antibiotic, **bacitracin** blocks the hydrolysis of dolichol pyrophosphate to dolichol phosphate by a phosphatase.

Synthesis of the oligosaccharide core of glycoproteins and its transfer to the protein in the endoplasmic reticulum

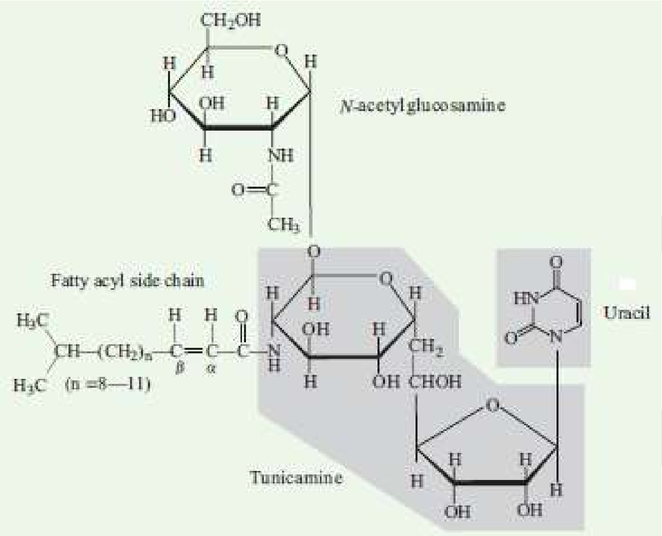


The oligosaccharide core is built up in a series of steps as shown. The first few steps occur on the cytosolic face of the ER. Completion occurs within the lumen of the ER after a translocation step (upper left) in which the incomplete oligosaccharide is moved across the membrane. The synthetic precursors that provide additional mannose and glucose residues to the growing oligosaccharide in the ER lumen are themselves dolichol phosphate derivatives. Dolichol-phosphate-mannose and dolichol-phosphate-glucose are synthesized from dolichol phosphate and GDP-mannose or UDP-glucose, respectively. After it is transferred to the protein, the oligosaccharide core is further modified in the ER and the Golgi complex in pathways that differ for different proteins. The released dolichol pyrophosphate is recycled. The 5-sugar residues (shown in a dotted enclosure on lower right side) are retained in the final structure of all N-linked oligosaccharides.

B. Terminal Glycosylation

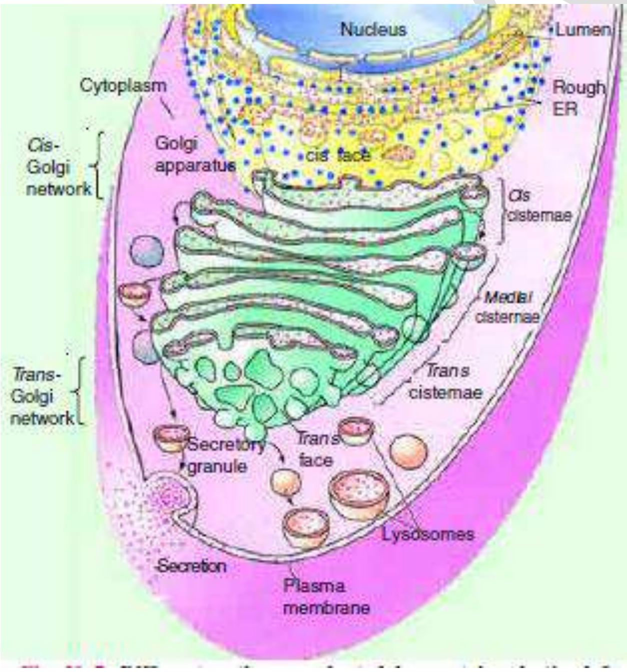
Proteins are transported from the ER to the Golgi complex in transport vesicles. Golgi complex is an asymmetric stack of flattened membranous sacs called **cisternae**. A typical mammalian cell has 3 or 4 cisternae, whereas many plant cells usually have about twenty. The Golgi (*pronounced* as GOAL-gee) is differentiated into (1) a *cis* compartment, the receiving end, which is closed to the ER; (2) *medial* compartments; and (3) a *trans* compartment, which exports proteins to various destinations. These compartments contain different enzymes and carry out distinctive functions. Different vesicles transfer proteins from one Golgi compartment to another and then to lysosomes, secretory vesicles, and the plasma membrane. The transport of proteins between the ER and Golgi, and between the Golgi and subsequent destinations is mediated by small (~ 50 to 100 nm in diameter) membrane-bound compartments called *transport vesicles* (or *transfer vesicles*). The Golgi complex performs two main roles. First, carbohydrate units of glycoproteins are altered and elaborated in the Golgi. O-linked sugar units are trimmed there, and N-linked ones are modified in many different ways. Second, the Golgi is the major sorting and packaging center of the cell. It sends proteins to lysosomes, secretory granules, or the plasma membrane according to signals encoded by their 3-dimensional structures The carbohydrate moieties of glycoproteins are modified in each of the compartments of the Golgi complex. In the *cis* compartment, 3 mannoses are removed from the oligosaccharide chains of proteins destined for secretion or for insertion in the plasma membrane. The carbohydrate moieties of glycoproteins targeted to the lysosomal lumen are modified differently (described later). In the medial compartments, 2 or more mannoses are removed, and 2 *N*-acetylglucosamines and a fucose are added. Finally, in the *trans* compartment, another N-acetylglucosamine is added, followed by galactose and sialic acid, to form a complex Although the biochemical mechanisms involved in “sorting and packaging” is not fully understood, however, with respect to *N*-linked oligosaccharides, a unified concept about the types of units attached is developing. As a rule, *N*-linked oligosaccharides have the same *inner core* which is the branched pentasaccharide containing 3 mannose and 2 acetylglucosamine. Apparently, trimming of the common oligosaccharide core can proceed to the level of the inner core.

Structure of tunicamycin -Tunicamycin is actually a family of antibiotics produced by (and isolated as a mixture from) *Streptomyces lysosuperficens*. They all contain uracil-*N*-acetylglucosmine, an 11-carbon aminodialdose called tunicamine and a fatty acyl side chain. The structure of the fatty acyl side chain varies in the different compounds within the family. In addition to the variation in length of the fatty acyl side chain, some homologues lack the isopropyl group at the end and/or α, β-unsaturation.



The *N*-linked oligosaccharides generally fall into one of the following two categories;

(*a*) **Simple mannose-rich units:** These possess the inner core either with short or long mannose oligosaccharides attached (chicken albumin) or with one or few carbohydrates attached (human immunoglobulin M, IgM).



(*b*) **Complex N-acetyllactosamine units:** These are oligosaccharides with *N*-acetylgalactosamine (disaccharide unit of galactose and *N*-acetylglucosamine) linked to the mannosyl residues of the inner core, since they generally have additional sialate (NAN) residues bonded to their galactosyl residues. The two common examples are the oligosaccharide units of human transferrin and immunoglobulin G, IgG, which, unlike the simple units, have been found only in animals.

Different pathways adopted by proteins destined for lysosomes, the plasma membrane or secretion

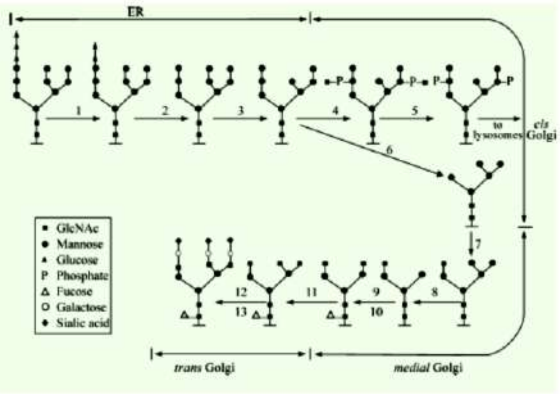
Most proteins destined for secretion of insertion into a membrane are synthesized by ribosomes (*blue dots*) attached to the rough endoplasmic reticulum (rough ER; *top*). As they are synthesized, the proteins (*red dots*) are either injected into the lumen of the endoplasmic reticulum or inserted into its membrane. After initial processing, the proteins are

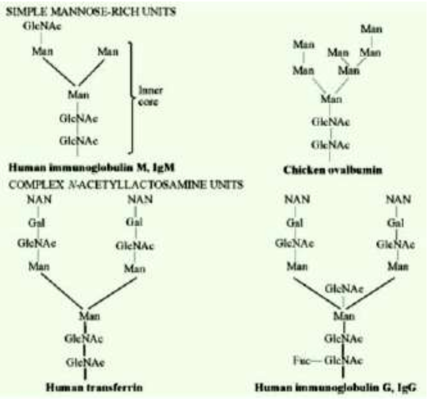
encapsulated in vesicles formed from endoplasmic reticulum membrane, which subsequently fuse with the *cis* Golgi network. The proteins are progressively processed according to their cellular destinations, in the *cis*, *medial,* and *trans* cisternae of the Golgi, between which they are transported by other membranous vesicles, Finally, in the *trans* Golgi network (*bottom*), the completed glycoproteins are sorted for delivery to the final destinations, for example, lysosomes, the plasma membrane, or secretory granules, to which they are transported by yet other vesicles.

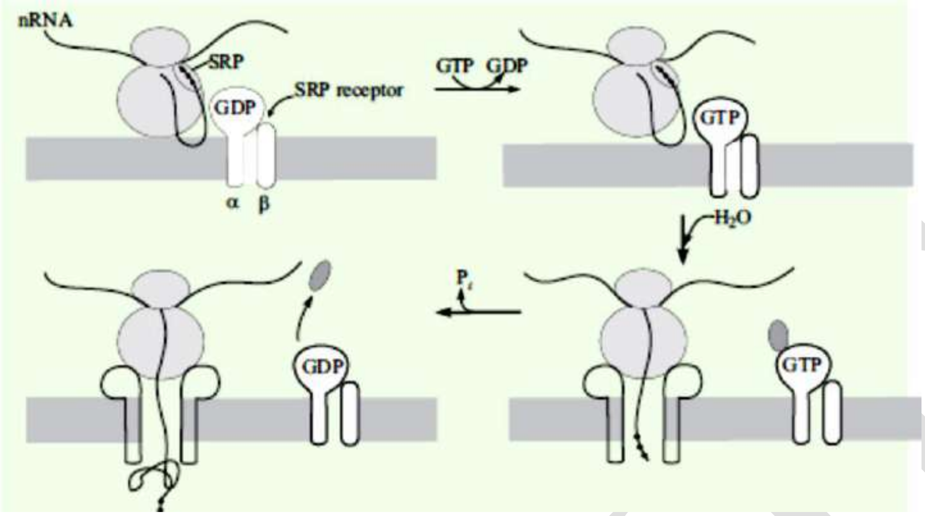
Processing of asparagine-linked (or N-linked) oligosaccharides in the ER and the 3 compartments of the Golgi complex

Steps 4 and 5 apply only to proteins destined for delivery to lysosomes. The enzymes catalyzing all the 13 steps are: 1. glucosidase I 7. GlcNAc transferase I 2. glucosidase II 8. mannosidase II 3. ER -1, 2-mannosidase 9. GlcNAc transferase II 4. N-acetylglucosaminyl- 10. fucosyl transferase phosphotransferase 11. GlcNAc transferase IV 5. phosphodiester glycosidase 12. galactosyltransferase 6. Golgi mannosidase I 13. sialyltransferase

Four oligosaccharide units of glycoproteins







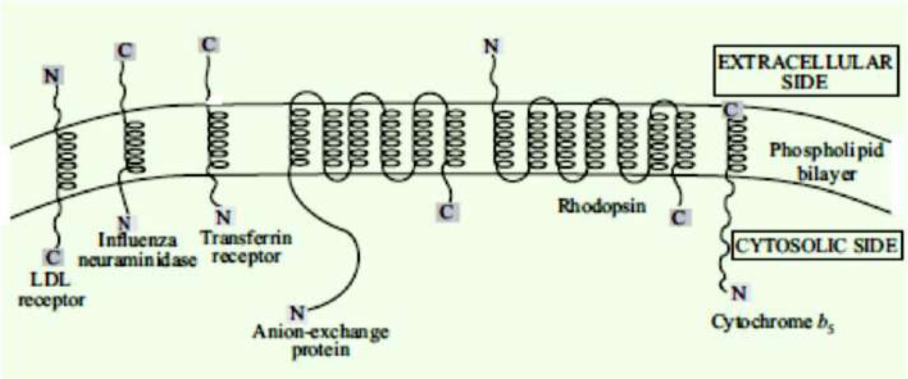
GTP-GDP cycle of the SRP receptor

The cycle drives the delivery of the signal sequence to the translocation machinery of the ER membrane.

GTP-GDP Cycle and the Signal Sequence

The signal sequence on the nascent polypeptide is protected rather sequestered by SRP until it is delivered to the translocation machinery on the ER membrane. The exact timing of the release of the polypeptide by SRP is achieved by a GTP-GDP cycle in the SRP receptor, which is an integral membrane protein consisting of two submits, α (68 kd) and β (30 kd). The binding of SRP-signal peptide to the receptor triggers the replacement of GDP (bound to the α subunit) by GTP. The GTP form of the receptor firmly binds SRP, which loses its grip on the signal peptide. The released signal peptide quickly binds to the *translocon*, a multisubunit assembly of integral and peripheral membrane proteins, which act as translocation machinery. The α subunit of the receptor then hydrolyzes its bound GTP to GDP, which releases SRP. The delay in GTP hydrolysis gives the signal peptide enough time to find its new partner so that the signal peptide is not recaptured by SRP. *Ribosomes bearing signal sequences are targeted to the ER membrane because of the unidirectionality of the GTP-GDP cycle*.

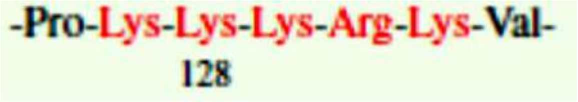
Although elongation of a polypeptide and its translocation across the ER membrane are two separate processes, yet they do occur simultaneously. This is because the synthesized proteins become folded and cannot be efficiently translocated as they do not fit in the protein conducting channel. *Unfolded polypeptide chains are the optimal substrates for translocation across the ER membrane*. Also, binding of SRP to ribosomes arrests elongation so that the premature folding of the nascent chain is prevented. Moreover, the ribosomes keep the nascent polypeptide chain fully stretched out in the narrow tunnel of the large subunit. The translocation process for integral membrane proteins is more complex than for secretory and lysosomal proteins which are threaded through in entirety. The integral membrane proteins have either one or many membrane-spanning helices. Moreover, the amino and carboxy termini can be on either side of the membrane in such proteins. The translocation machinery acts restlessly unless stopped by a specific instruction, which in this case is a **stop transfer sequence** (also called **a membrane anchor sequence**) present on the nascent polypeptide chain. A second signal sequence is also required to start another round of translocation of a chain that spans the membrane more than once. Furthermore, the translocation machinery must be able to thread the nascent chains in the reverse direction also. All this is yet unexplored and needs investigation.



Different topological arrangements of integral membrane proteins

Protein Transport to Nucleus

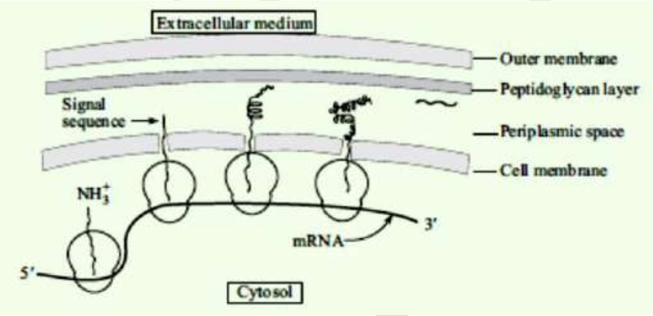
All nuclear proteins (*esp*, histones, DNA polymerases, RNA polymerases and all proteins participating in the replication of DNA and transcription) are synthesized in the cystosol by free ribosomes and must pass through the nuclear envelope of eukaryotes comprising an outer membrane and an inner membrane. This transport, for the small proteins (*e.g*, histones), seems to take place through nuclear pores of 70 Å diameter; but for larger proteins (> 90 kd), a short peptide sequence (or signal sequence) appears to be necessary. For example, the T antigen of SV 40 virus is a protein of molecular weight 92,000 daltons (or 92 kd) that regulates the replication and transcription of viral DNA. And studies have shown that the transport of this large protein, depends on the presence of a **nuclear localization sequence**, containing five consecutive positively-charged residues (shown in red):


-Pro-Lys-Lys-Lys-Arg-Lys-Val-
128

A change of even a single amino acid residue can render this sequence inactive. As an instance, T antigen containing Thr or Asn in place of Lys at residue 128 stays in the cystol and not transported to the nucleus. The nuclear localization sequences can also accelerate the entry of small proteins. The transport of large proteins into nuclei is powered by ATP hydrolysis. Interestingly, none of the nuclear localization signals are cleaved on entry into the nucleus. It was also possible to bring about the transport of proteins to the nucleus by grafting (at the DNA level) this heptapeptide sequence (shown above) on pyruvate kinase or on other cytosolic proteins. It is moteworthy that fully-folded proteins can be imported into nuclei but not into mitochondria or chloroplasts, which must maintain a tight permeability barrier to sustain a proton-motive force. No bilayer membrane is crossed on entering the nucleus– hence, unfolding is not essential. The nucleus can afford to be more relaxed about its border since the pH and ionic composition of the nucleoplasm is essentially the same as that of the cytosol.

BACTERIAL SIGNAL SEQUENCES AND PROTEIN TARGETING

Protein targeting is not confined to eukaryotes. Bacteria also target proteins to destinations encoded in their sequences. A Gram-negative microorganism such as *E.coli* can translocate nascent proteins to the inner and outer membranes, the periplasmic space between the membranes or (rarely) the extracellular medium (secretion) as depicted in Figure.

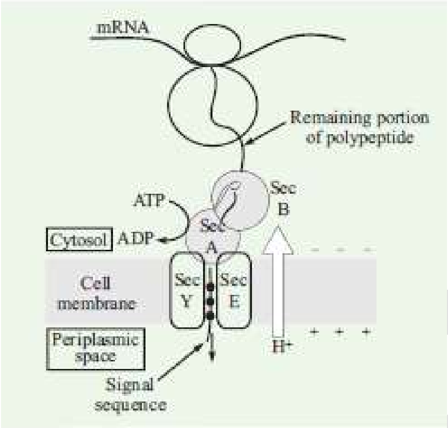


Schematic of the synthesis of noncytosolic proteins by ribosomes bound to the plasma membrane in prokaryotes

A signal sequence (shown by bold line) on the nascent chain directs the ribosomes to the plasma membrane and enables the protein to be translocated. The translocation machinery is not depicted in this schematic diagram.

As in eukaryotes, translocation is not mechanistically coupled to chain elongation. This targeting uses signal sequences (also called **leader sequences**) at the amino terminus of the proteins, much like those found on eukaryotic proteins targeted to the ER. These signal sequences are usually 16 to 26 residues long. Though diverse, the prokaryotic signal sequences have a positively-charged amino-terminal region, and a helix-breaking segment. Like eukaryotes, the signal sequence is usually cleaved by a signal peptidase at the helix-breaking site. Another similarity is that polypeptide chain elongation and translocation usually take place at about the same time but are not mechanistically coupled. Translocation of a polypeptide chain through the cell membrane of *E. coli* is catalyzed by a soluble chaperone and a membrane-bound, multisubunit translocase. The bacterium contains a major chaperone, *SecB protein* which keeps nascent chains in unfolded or partially folded state to enable them to traverse the membrane. SecB presents the nascent chain to Sec A, a peripheral membrane component of the translocase. SecA works in unison with SecY and SecE, the membrane-embedded portion of the translocase. Two forms of free energy (ATP and protonmotive force) drive protein translocation in *E. coli*. SecA is an ATPase– the ATP state has high affinity for the protein to be translocated, whereas the ADP state has low affinity. Portions of the nascent chain are successively handed from from SecA to SecY-SecE channel as a result of many cycles of ATP hydrolysis. The proton-motive force across the cell membrane then drives the threading of the nascent chain through the membrane.

Schematic diagram showing the interplay of Sec proteins in protein translocation across the cell membrane



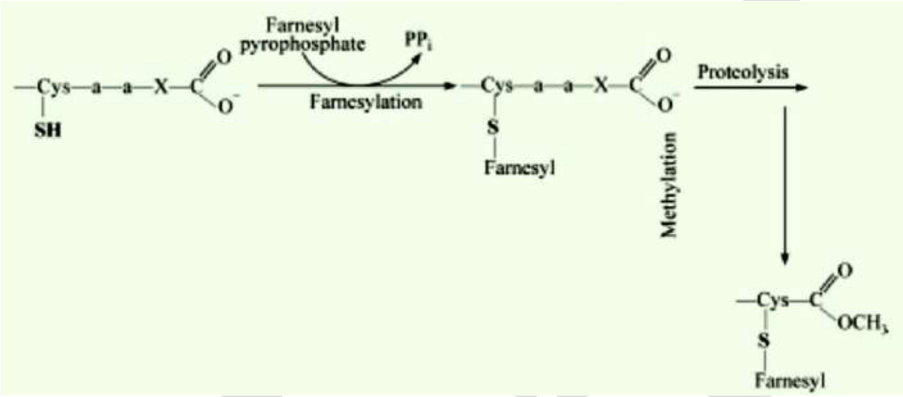
Proton-motive force powers the unidirectional translocation of the polypeptide from the cytosolic to the periplasmic side of the membrane Some proteins that are translocated

Some proteins that are translocated through one or more membranes to reach their final destinations are maintained in a distinct "*translocation-competent*" *conformation* until this process is complete. The functional conformation is assumed after translocation, and proteins purified in this final form are now longer capable of translocation. Available evidences indicate that the translocation conformation is stabilized by a specialized set of proteins in all bacterial cells. These bind to the protein to be translocated while it is being synthesized, preventing it from folding into its final 3-'D' structure. In *E. coli*, a protein called trigger factor (Mr 63,000) appears to facilitate the translocation of at least one outer membrane protein through the inner membrane.

PROTEIN IMPORT BY RECEPTOR-MEDIATED ENDOCYTOSIS

Specific proteins are imported into a cell by their binding to receptors in the plasma membrane and their inclusion into vesicles. Such a process is called **receptor-mediated endocytosis** and has a great number of biological applications:

Farnesylation of a cytosolic protein at the C terminus



Farnesylation is followed by trimming of 3 C-terminal residues and methylation of the terminal carboxylate.

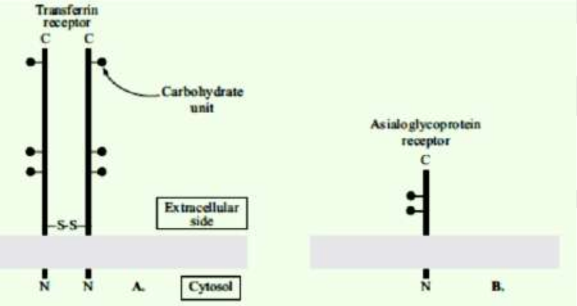
1. It is a means of delivering essential metabolites to cells. For instance, the low-density lipoprotein (LDL) carrying cholesterol is taken up by the LDL receptor in the plasma membrane and internalized.
2. Endocytosis regulates responses to many protein hormones and growth factors. Epidermal growth factor and nerve growth factor are taken into the cell and degraded together with their receptors.
3. Proteins destined for degradation are taken up and delivered to lysosomes for digestion. Phagocytes, for example, have receptors that enable them to take up antigen-antibody complexes.
4. Receptor-mediated endocytosis is employed by many viruses and toxins to gain entry into cells, as exemplified by the ingenious mode of entry and departure of Semliki Forest Virus (SFV), a membrane-enveloped virus.

5. Disorders of receptor-mediated uptake can lead to diseases, such as some forms of familial hypercholesterolemia.

Cell-surface Receptors and Clathrin

The **cell-surface receptors**, which mediate endocytosis, are transmembrane glycoproteins. They have a large extracellular domain and a small cytosolic region and contain either one (*e.g.*, asialoglycoproteins) or two (*e.g.*, transferrin) transmembrane helices. Many of the receptors are located in specialized regions of the plasma membrane called *coated pits*. The cytosolic side of these pits has a thick coat of *clathrin*, a protein designed to form lattices around membranous vesicles. Many receptors (such as those for LDL, transferrin, asialogycoproteins, insulin) congregate in coated pits; others (such as the receptor for epidermal growth factor) cluster there after binding their cognate protein.

Clathrin is a trimeric protein, consisting of 3 heavy chains (H ; M*r* = 180,000) and 3 light chains (L; M*r* = 35,000). The (HL)3 clathrin unit (8 S; M*r* = 650,000) is organized as a three-legged structure, called a *triskelion*. The carboxy termini of the 3 heavy chains (each about 500 Å long) come together at a vertex. A bend in the heavy chain divides it into a proximal arm, closest to the vertex, and a distal arm. Each of 3 the light chains is aligned with the proximal arm of a heavy chain. Many clathrin assemble into closed shells having a polyhedral structure. The polyhedra are made of both pentagons and hexagons. A single edge of a pentagon or hexagon is made of parts of four triskelions, 2 proximal arms and 2 distal arms. The flexibility of a triskelion is important in enabling it to fit into a pentagon or hexagon.

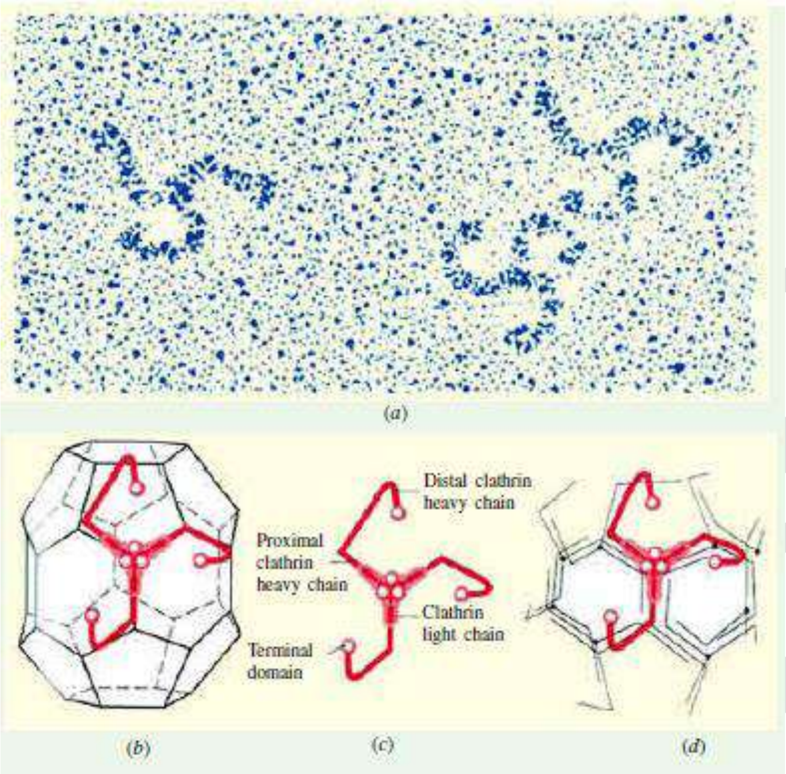


Schematic of two cell-surface receptors that are internalized at coated pits A. Transferrin receptor B. The asialoglycoprotein receptor. The short N-terminal tails of these receptors are critical for internalization.

Receptor-mediated Endocytosis

Receptor-mediated endocytosis (Fig. 29–27) begins with the binding of certain proteins (such as LDL, transferrin, peptide hormones etc) to receptors on the outer face of the plasma membrane. The receptors are concentrated in invaginations of the membrane called coated pits, which are coated on their cytosolic face with a lattice made up of the protein clathrin. The clathrin lattice grows up until a complete membrane-bound endocytic vesicle, with a diameter of about 80 nm, buds off the plasma membrane and moves into the cytosol. The endocytic vesicle then rapidly loses its clathrin shell by uncoating enzymes and fuses with an endosome. The endosomes, in turn, fuse with one another to form bigger vesicles, ranging between 200 and 600 nm. The membrane ATPases present in the endosomes lower the pH, so that the receptors dissociate from their target proteins. Proteins and receptors then follow separate paths, their fates varying according to the system (Table 29–2). The protein transferrin transports iron from sites of absorption and storage to sites of utilization. Two Fe3+ ions are bound to the protein which contains two similar domains. The protein devoid of iron is called *apotransferrin*. Transferrin, but not apotransferrin, binds to a dimeric receptor (Fig. 29–25). The low pH within the endosome causes dissociation of Fe3+ from transferrin. The acidity lowers the affinity of transferrin for Fe3+ more than a millionfold. However, apotransferrin remains bound to the receptor. Sorting then takes place : part of the vesicle bearing apotransferrin bound to the receptor pinches off and proceeds towards plasma membrane, whereas the remaining Fe3+ is stored in ferritin in the cytosol. When the pinched off vesicle fuses with the plasma membrane, apotransferrin is released from the receptor because of the sudden increase in pH. Apotransferrin has little affinity for the receptor at pH 7.4. Thus, *pH changes are used twice to drive the transferrin transport cycle*: first to release iron from transferrin in the endosome, and then to discharge apotransferrin into the extracellular fluid. *The cycle takes about 16 minutes* : 4 minutes for the binding of transferrin, 5 minutes for transport to endosomes, and 7 minutes for the return of the iron carrier and the receptor to the cell surface. Toxins (diphtheria toxin, cholera toxin) as well as viruses (influenza virus) enter cells by receptor-mediated endocytosis.

Structure and assembly of a coated vesicle or clathrin



(*a*) **Electron micrograph of a metal-shadowed preparation of clathrin triskelions.** (*b*) **A typical coated vesicle containing a membrane vesicle about 40 nm in diameter surrounded by a fibrous network of 12 pentagons and 8 hexagons.** The fibrous coat is constructed of 36 clathrin triskelions.One clathrin triskelion is centered on each of the 36 vertices of the coat. Coated vesicles having other sizes and shapes are believed to be constructed similarly: each vesicle contains 12 pentagons but a variable number of hexagons. (*c*) **Detail of a clathrin triskelion.** Each of three clathrin heavy chains is bent into proximal arm and a distal arm. A clathrin light chain in attached to each heavy chain, most likely near the center (*d*) **An intermediate in the assembly of a coated vesicle, containing 10 of the final 36 triskelions, illustrates the packing of the clathrin triskelions.** Each of the 54 edges of a coated vesicle is constructed of two proximal and two distal arms intertwined. The 36 triskelions contain 36 × 3 = 108 proximal and 108 distal arms, and the coated vesicle has precisely 54 edges.

UNIT-IV

SYLLABUS

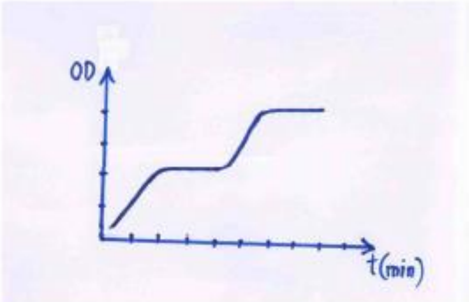
Regulation of gene expression in prokaryotes - Principles of gene regulation, negative and positive regulation, concept of operons, regulatory proteins, activators, repressors, DNA binding domains, regulation of lac operon and trp operon, induction of SOS response, synthesis of ribosomal proteins, regulation by genetic recombination, transcriptional regulation in λ bacteriophage.

REGULATION OF GENE EXPRESSION IN PROKARYOTES

Genes are expressed through transcription and translation, but what decide which gene, when, where and how it is expressed?

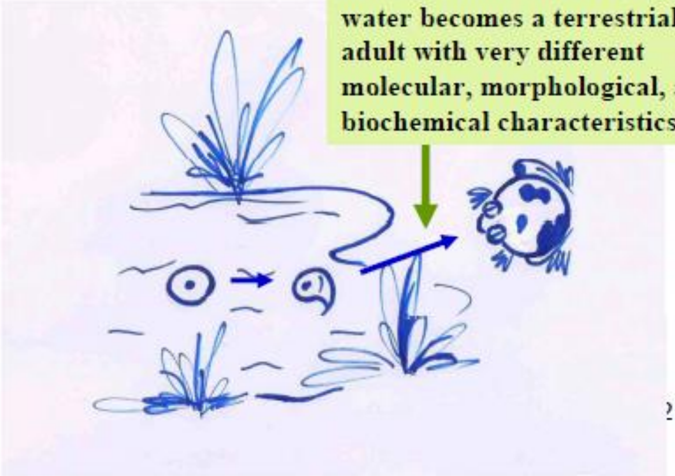
→ The expression of a gene (or a part of the genome) can be regulated in many ways depending on cell organization and needs of the organism

Examples concerning the regulation of gene expression in a bacterium and an animal



E. coli is grown in medium containing glucose and lactose. Cell density is measured according to culture time as OD (Optical density) value. Results are shown in the picture above

Metamorphosis : The transition period where a larvae living in water becomes a terrestrial adult with very different molecular, morphological, and biochemical characteristics



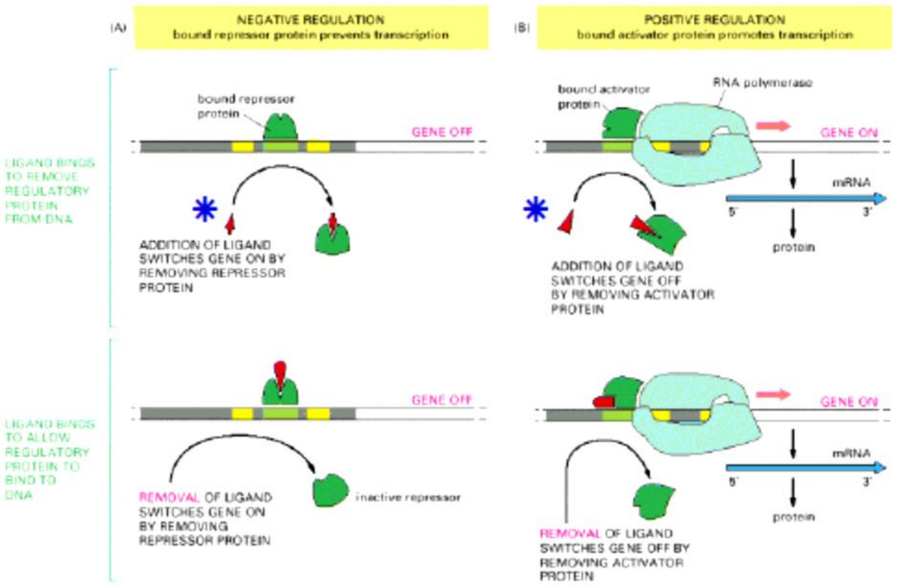
TRANSCRIPTIONAL CONTROL OF GENE EXPRESSION IN PROKARYOTES: The two well-studied main mechanisms of transcriptional control of gene expression are:

1. The operons : genes involved in a metabolic pathway are regrouped into a gene cluster controlled by common regulatory sequences and proteins. The expression of these genes are then rapid and synchronized. The operon model was developed by François Jacob and Jacques Monod (1961)

2. The cascades of gene expression : Under some environmental conditions, expression of a first set of genes can be "switch on", and one or more of the products of this first gene set will "switch on" a second gene set. This event could be repeated many times to mobilize wider gene sets to achieve a special metabolic pathway.

In all organisms, structural genes can be classified into two groups : 1. Constitutive genes, also called "housekeeping" genes : encoding RNA and proteins having basal vital functions such as rRNA, ribosomal proteins, proteins of cellular respiratory system, … These genes are mostly expressed continually and with a stable amount. 2. Inducible genes: encoding proteins necessary for the survival of the organism in changing environment. These genes must be rapidly "switch on" or "off" depending on the temporary needs of the organism for their products.

POSITIVE AND NEGATIVE CONTROL OF GENE EXPRESSION


(A)
NEGATIVE REGULATION
bound repressor protein prevents transcription
(B)
POSITIVE REGULATION
bound activator protein promotes transcription
LIGAND BINDS TO REMOVE REGULATORY PROTEIN FROM DNA
bound repressor protein
GENE OFF
ADDITION OF LIGAND SWITCHES GENE ON BY REMOVING REPRESSOR PROTEIN
bound activator protein
RNA polymerase
GENE ON
mRNA
5'
3'
protein
ADDITION OF LIGAND SWITCHES GENE OFF BY REMOVING ACTIVATOR PROTEIN
LIGAND BINDS TO ALLOW REGULATORY PROTEIN TO BIND TO DNA
GENE OFF
REMOVAL OF LIGAND SWITCHES GENE ON BY REMOVING REPRESSOR PROTEIN
inactive repressor
GENE ON
mRNA
5'
3'
protein
REMOVAL OF LIGAND SWITCHES GENE OFF BY REMOVING ACTIVATOR PROTEIN

Gene expression can be positively or negatively controlled. In positive control, binding of activator protein triggers the transcription whereas in negative control, binding of repressor protein inhibits the transcription.

 Ligands which bind to the activators to "switch on" gene expression in positive control are called inducers ; those binding to the repressors and "switching off" gene expression are called co-repressors. Inducers and co-repressors are known as effectors

THE OPERON

The purposes of the regulation of gene expression in prokaryotes are remarkably well served by the use of operons : (1) all genes of an operon are coordinateley expressed → the metabolic pathway controlled by this operon can be regulated very fast, (2) there is energy saving as the same set of regulatory sequences and proteins is used for all structural genes of the operon.

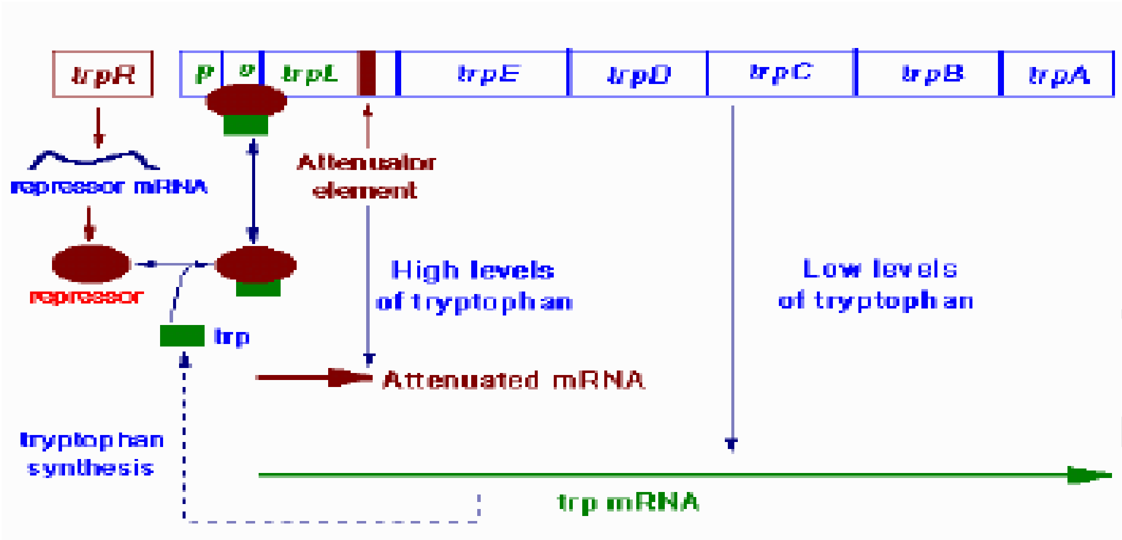
One of the most important challenges for prokaryotes is to adapt their metabolic processes to available environmental nutritive sources.Depending on the metabolic pathways, - In catabolic pathway (degradation of macromolecules into structural units), when a substrate to be degraded is present, the operon is "switched on". These operons are characterized as inducible - In anabolic pathway (synthesis of macromolecules from small ones), when a product needed by the cell is present, the corresponding operon is "switched off". These operons are considered as repressible operons.

 An operon is composed of : regulatory sequences (promoter, operator, other sequences,…), structural genes, regulatory gene (promoter + coding sequence of a regulatory protein)

genes, regulatory gene (promoter + coding sequence of a regulatory protein)

<u>Trp Operon</u>

The genes for the five enzymes in the Trp synthesis pathway are clustered on the same chromosome in what is called the Trp operon. The Trp operon has three components: Five Structural Genes: These genes contain the genetic code for the five enzymes in the Trp synthesis pathway. One Promoter: DNA segment where RNA polymerase binds and starts transcription. One Operator: DNA segment found between the promoter and structural genes. It determines if transcription will take place. If the operator is turned "on", transcription will occur. When nothing is bonded to the operator, the operon is "on". o RNA polymerase binds to the promoter and transcription is initiated.

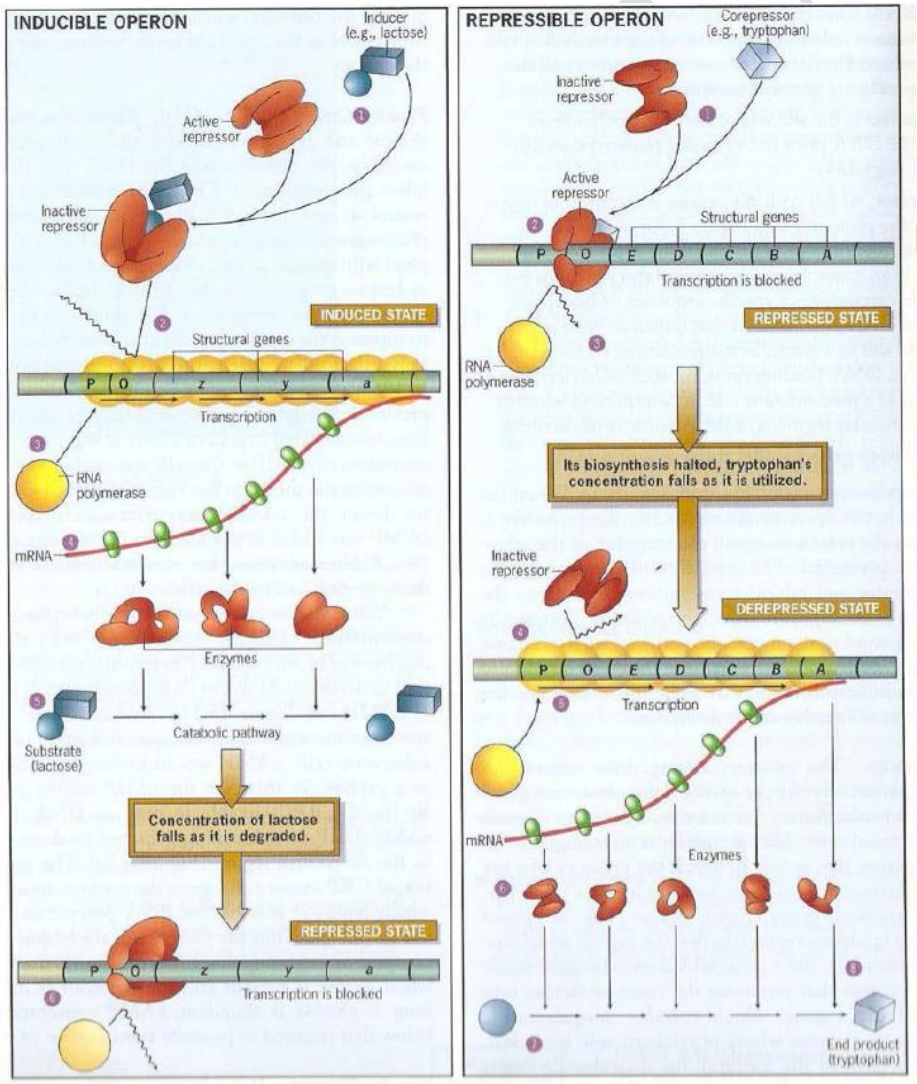


Structure of trp operon

The five structural genes are transcribed to one mRNA strand. The mRNA will then be translated into the enzymes that control the Trp synthesis pathway. The operon is turned "off" by a specific protein called the repressor. o The repressor is a product of the regulator gene which is found some distance from the operon. Transcription of the regulator produces mRNA which is translated into the repressor. The repressor is inactive in this form and cannot bind properly to the operator with this conformation. To become active and bind properly to the operator, a co-repressor must associate with the repressor. The co--repressor for this system is Trp. This makes sense because *E. coli* does not want to synthesize Trp if it is available from the environment. The more Trp available, the more that can associate with repressor molecules. An active repressor binds to the operator blocking the attachment of RNA polymerase to the promoter. Without RNA polymerase, transcription and translation of the structural genes can't occur and the enzymes needed for Trp synthesis are not made.

Repressible vs Inducible Systems

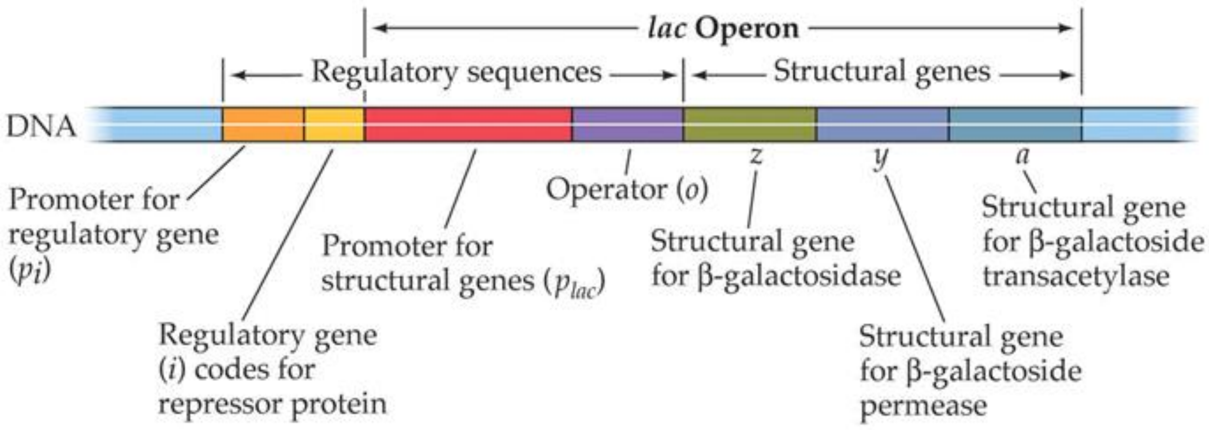
• The Trp pathway is anabolic as Trp is being synthesized. The Trp and other regulated anabolic pathways are usually repressible because the system can be repressed by an overabundance of the end product. The end product, Trp, in this case, decreases or stops the transcription of the enzymes necessary for its production. Regulated catabolic pathways, on the other hand, are usually inducible because the pathway is stimulated rather than inhibited by a specific molecule. An example of an inducible system is lactose metabolism.



Inducible and repressible operons

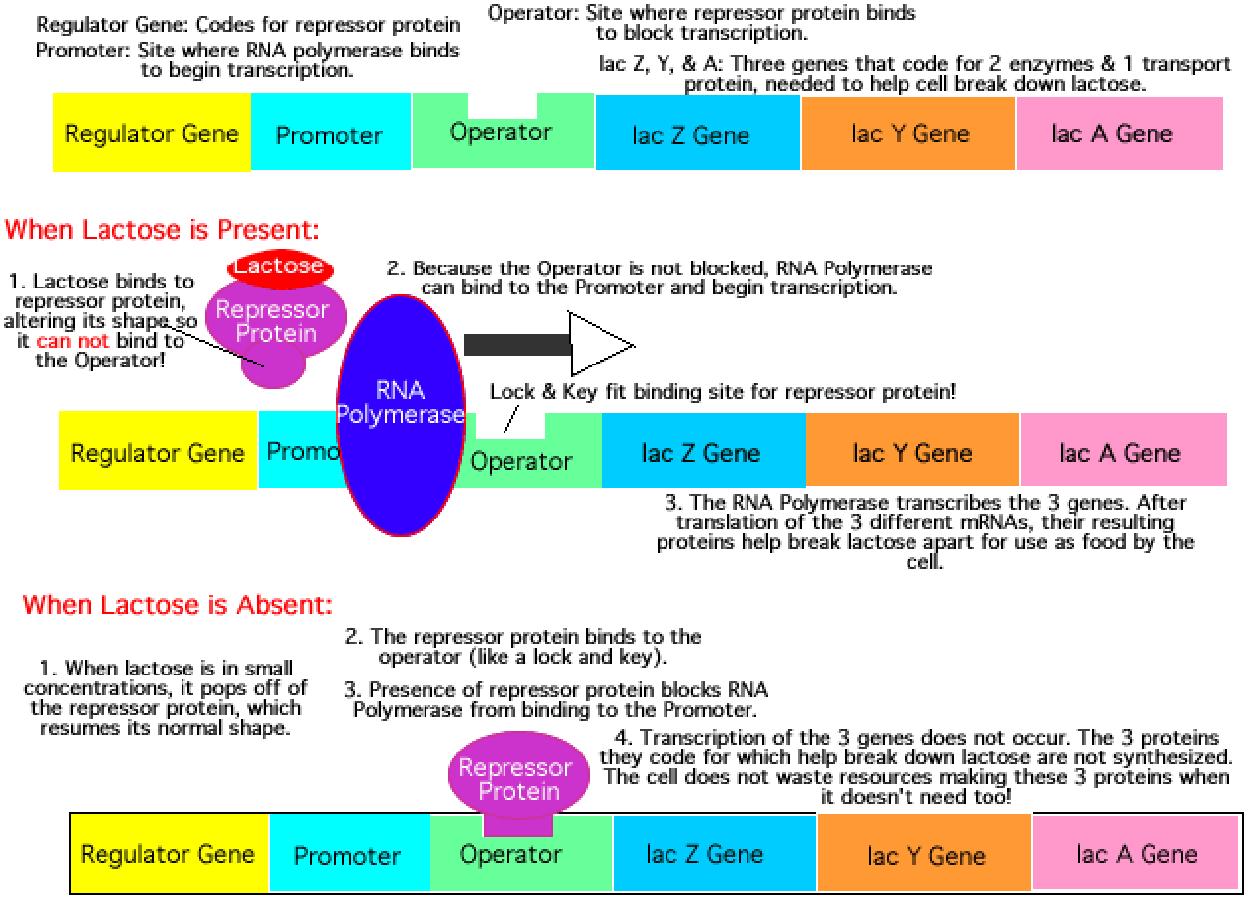
The lac Operon

The genes that code for the enzymes needed for lactose catabolism are clustered on the same chromosome in what is called the *lac* operon. The Lac operon has three components: Three Structural Genes: These contain the genetic code for the three enzymes in the lac catabolic pathway. One Promoter: DNA segment where RNA polymerase binds and starts transcription. One Operator: DNA segment found between the promoter and structural genes. It determines if transcription will take place. If the operator in turned "on", transcription will occur.


lac Operon
Regulatory sequences
Structural genes
DNA
z
y
a
Promoter for regulatory gene (p_i)
Regulatory gene (*i*) codes for repressor protein
Promoter for structural genes (p_{lac})
Operator (*o*)
Structural gene for β-galactosidase
Structural gene for β-galactoside permease
Structural gene for β-galactoside transacetylase

The *lac* operon

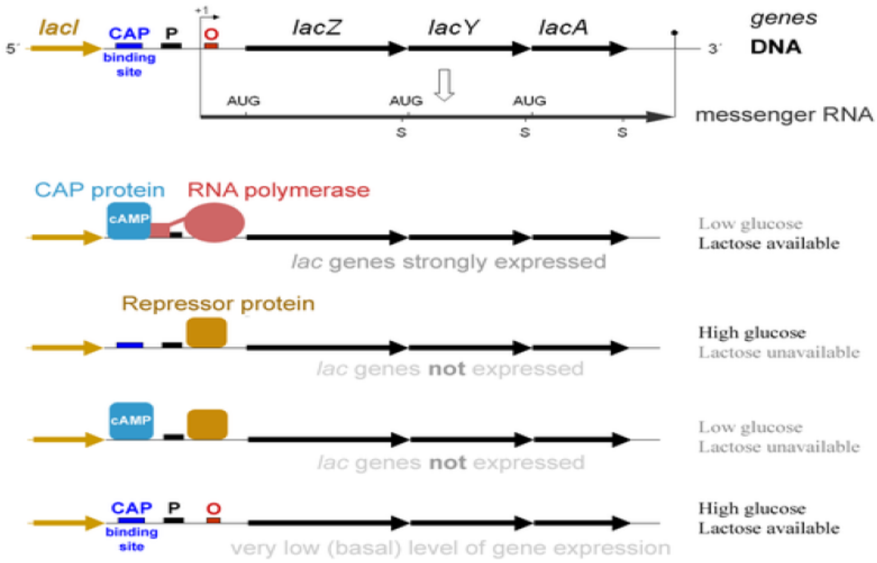
As in the Trp operon, the Lac operon is turned "off" by a specific protein called the repressor. The repressor is the product of the regulator gene which is found outside the operon. Transcription of the regulator produces mRNA which is translated into the repressor. But unlike the Trp operon, the repressor is active in this form and does not require a co-repressor. The active repressor binds to the operator blocking the advancement of RNA polymerase to the structural genes. Without RNA polymerase, transcription and translation of the genes can't occur and the enzymes needed for Lac metabolism are not made. What turns the Lac operon "on"? Lactose does! This makes sense because the cell only needs to make enzymes to catabolize lactose if lactose is present. When lactose enters the cell, allolactose, an isomer of lactose is formed. Allolactose binds to the repressor and alters its conformation so that it can't bind to the operator. RNA polymerase can now start transcription. The three structural genes are transcribed to one mRNA strand. The mRNA will then be translated into the enzymes that control lactose catabolism. In this sense, allolactose is an inducer.


Regulator Gene: Codes for repressor protein
Promoter: Site where RNA polymerase binds to begin transcription.
Operator: Site where repressor protein binds to block transcription.
lac Z, Y, & A: Three genes that code for 2 enzymes & 1 transport protein, needed to help cell break down lactose.
Regulator Gene
Promoter
Operator
lac Z Gene
lac Y Gene
lac A Gene
When Lactose is Present:
1. Lactose binds to repressor protein, altering its shape so it can not bind to the Operator!
Lactose
Repressor Protein
RNA Polymerase
2. Because the Operator is not blocked, RNA Polymerase can bind to the Promoter and begin transcription.
Lock & Key fit binding site for repressor protein!
Regulator Gene
Promo
Operator
lac Z Gene
lac Y Gene
lac A Gene
3. The RNA Polymerase transcribes the 3 genes. After translation of the 3 different mRNAs, their resulting proteins help break lactose apart for use as food by the cell.
When Lactose is Absent:
1. When lactose is in small concentrations, it pops off of the repressor protein, which resumes its normal shape.
2. The repressor protein binds to the operator (like a lock and key).
3. Presence of repressor protein blocks RNA Polymerase from binding to the Promoter.
4. Transcription of the 3 genes does not occur. The 3 proteins they code for which help break down lactose are not synthesized. The cell does not waste resources making these 3 proteins when it doesn't need too!
Repressor Protein
Regulator Gene
Promoter
Operator
lac Z Gene
lac Y Gene
lac A Gene

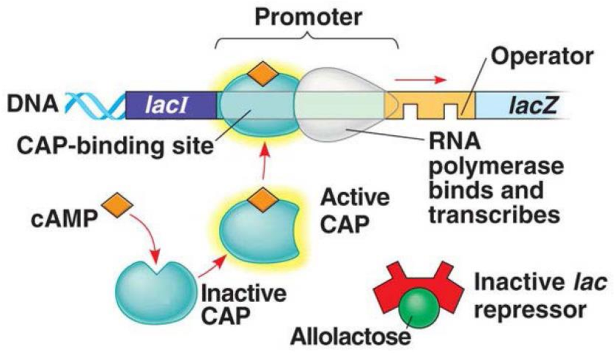
The *lac* operon: a model of gene regulation in prokaryotic cells

Negative vs Positive Control

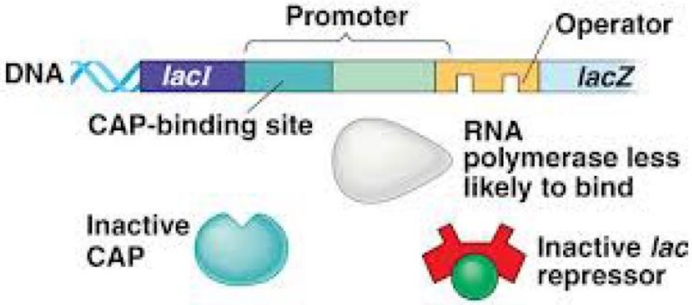
While the Trp operon is an example of repressible gene regulation and the Lac operon is an example of inducible gene regulation, both are examples of negative control of genes because both operons are shut "off" by an active repressor. Gene regulation would be positive; on the other hand, if an activator molecule turned the operon "on". The Lac operon is also an example of a positive control system and is turned on by the cAMP-CAP complex, as described below: *E. coli* can be described as a fussy eater. Its first choice at every meal is glucose because glucose supplies maximum energy for growth. Therefore, *E. coli* will only metabolize lactose if concentrations of glucose are low. For this to work, there must be a signal to tell the Lac operon that glucose is not available and to start transcribing the genes to metabolize lactose. This signal is a small molecule called cyclic AMP (cAMP). The amount of cAMP present in a cell is inversely proportional to the amount of glucose present. As a result, the absence of glucose results in an increase in cAMP in the cell. The following describes the situation where there is lactose but no glucose available to the cell: No glucose means high levels of cAMP. cAMP binds to a molecule known as CAP. CAP, when in association with cAMP, can bind to the promoter at the CAP binding site. Here, the cAMP-CAP complex stimulates transcription by helping RNA polymerase bind to the promoter. RNA polymerase has a weak affinity for the Lac promoter and will not bind without this help. Remember with lactose present so is allolactose. Allolactose binds to the repressor and prevents it from binding to the operator. Therefore, transcription and translation of the genes can occur. The following depicts what happens when glucose and lactose are both present for *E. coli* to metabolize: With glucose present, there is very little or no cAMP. It cannot bind to the CAP binding site. Without this complex, RNA polymerase cannot bind to the promoter and transcription cannot occur. Even though allolactose is present and blocks the action of the repressor, there is no transcription of the lac genes because glucose is present.



Overall structural elements of Lac Operon



Lactose present, glucose scarce (cAMP level high): abundant lac mRNA synthesized



Lactose present, glucose present (cAMO level low): little lac mRNA synthesized

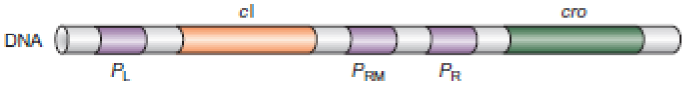
λ Bacteriophage _ is a virus that infects *E. coli*. Upon infection, the phage can propagate in either of two ways: **lytically** or **lysogenically,** as illustrated in Figure. Lytic growth requires replication of the phage DNA and synthesis of new coat proteins. These components combine to form new phage particles that are released by lysis of thehost cell. Lysogeny—the alternative propagation pathway—involves integration of the phage DNA into the bacterial chromosome where it is passively replicated at each cell division—just as though it were a legitimate part of the bacterial genome. A lysogen is extremely stable under normal circumstances, but the phage dormant within it—the **prophage**—can efficiently switch to lytic growth if the cell is exposed to agents that damage DNA (and thus threaten the host cell's continued existence). This switch from lysogenic to lytic growth is called **lysogenic induction.** The choice of developmental pathway depends on which of two alternative programs of gene expression is adopted in that cell. The program responsible for the lysogenic state can be maintained stably for many generations, but then, upon induction, switch over to the lytic program with great efficiency.

Alternative Patterns of Gene Expression Control Lytic and Lysogenic Growth

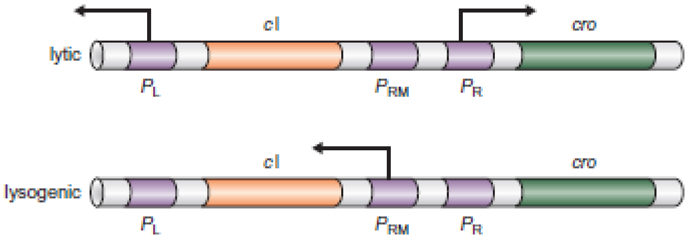
λ has a 50-kb genome and some 50 genes. Most of these encode coat proteins, proteins involved in DNA replication, recombination and lysis. The products of these genes are important in making new phage particles during the lytic cycle, but our concern here is restricted to the regulatory proteins, and where they act. Wecan therefore concentrate on just a few of them, and start by considering a very small area of the genome, shown in Figure. The depicted region contains two genes (*c*I and *cro*) and three promoters (*P*R, *P*L, and *P*RM). All the other phage genes (except one minor one) are outside this region and are transcribed directly from *P*R and *P*L (which stand for rightward and leftward promoter, respectively), or from other promoters whose activities are controlled by products of genes transcribed from *P*R and *P*L. *P*RM (promoter for repressor maintenance) transcribes only the *cI* gene. *P*R and *P*L are strong, constitutive promoters—that is, they have the elements required to bind RNA polymerase efficiently and direct transcription without help from an activator. *P*RM, in contrast, is a "weak" promoter and only directs efficient transcription when an activator is bound just upstream. Thus,*P*RM resembles the *lac* promoter. There are two arrangements of gene expression depicted in Figure: one renders growth lytic, the other lysogenic. Lytic growth proceeds when *P*L and *P*R remain switched on, while *P*RM is kept off. Lysogenic growth, in contrast, is a consequence of *P*L and *P*R being switched off, and *P*RM switched on. How are these promoters controlled?

Growth and Induction of _ Lysogen. Upon infection, _ can grow either lytically or lysogenically. A lysogen can be propogated stably for many generations, or it can be induced. Following induction, sets of the lytic genes are expressed sequentially, leading to the production of new phage particles.

Regulatory Proteins and Their Binding Sites - The *cI* gene encodes _ repressor, a protein of two domains joined by a flexible linker region 8). The N-terminal domain contains the DNA binding region (a helix-turn-helix domain, as we saw earlier). As with the majority of DNA binding proteins, _ repressor binds DNA as a dimer; the main dimerization contacts are made between the C-terminal domains. A single dimer recognizes a 17-bp DNA sequence, each monomer recognizing one half-site, again just as we saw in the *lac* system. Despite its name, _ repressor can both activate and repress transcription. When functioning as a repressor, it works in the same way as does Lac repressor—it binds to sites that overlap the promoter and excludes RNA polymerase. As an activator, _ repressor works like CAP, by recruitment. _ repressor's activating region is in the N-terminal domain of the protein. Its target on polymerase is a region of the _ subunit adjacent to the part of _ that recognizes the _35 region of the promoter



Promoters in the Right and Left Control Regions of Phage _.



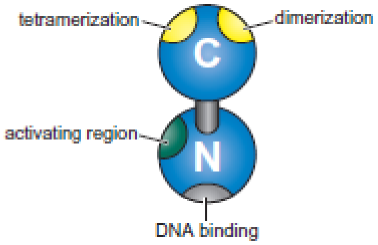
Transcription in the _

Control Regions in Lytic and Lysogenic Growth.

Cro (which stands for control of repressor and other things) only represses transcription, like Lac repressor. It is a single domain protein and again binds as a dimer to 17-bp DNA sequences. _ repressor and Cro can each bind to any one of six operators. These sites, which are shown in an expansion of our picture of the control region, are recognized with different affinities by each of the proteins. We will focus on the three operators on the right of the *c*I gene, but binding of repressor and Cro to the three operators on the left follows the same pattern. The three binding sites in the right operator are called *O*R1, *O*R2, and *O*R3; these sites are similar in sequence, but not identical, and each one—if isolated from the others and examined separately—can bind either a dimer of repressor or a dimer of Cro. The affinities of these various interactions, however, are not all the same. Thus, repressor binds *O*R1 tenfold better than it binds *O*R2. In other words, ten times more repressor—a tenfold higher concentration—is needed to bind *O*R2 than *O*R1. *O*R3 binds repressor with about the same affinity as does *O*R2. Cro, on the other hand, binds *O*R3 with highest affinity, and only binds *O*R2 and *O*R1 when present at tenfold higher concentration.

λ **Repressor Binds to Operator Sites Cooperatively -** repressor binds DNA cooperatively. This is critical to its function and occurs as follows. Consider repressor binding to sites in *O*R. In addition to providing the dimerization contacts, the C-terminal domain of _ repressor mediates interactions *between* dimers (the point of contact is the patch marked "tetramerization" in Figure). In this way, two dimers of repressor can bind cooperatively to adjacent sites on DNA. For example, repressor at *O*R1 helps repressor bind to the lower affinity site *O*R2 by cooperative binding. Repressor thus binds both sites simultaneously and does so at a concentration that would be sufficient to bind only *O*R1 were the two sites tested rately. (Recall that, without cooperativity, a tenfold higher concentration of repressor would be needed to bind *O*R2). *O*R3 is not bound: repressor bound cooperatively at *O*R1 and *O*R2 cannot simultaneously make contact with a third dimer at that adjacent site. We have already discussed the idea of cooperative binding and seen an example: activation of the *lac* genes by CAP. As in that case, cooperative binding of repressors is a simple consequence of their touching each other while simultaneously binding to sites on the same DNA molecule.

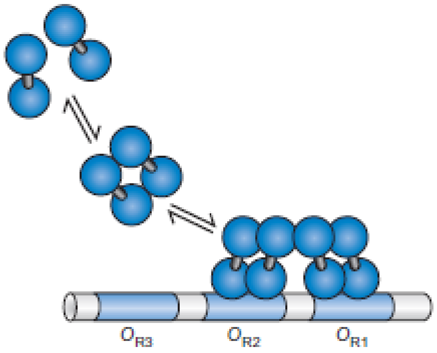
A more detailed discussion of the causes and effects of cooperative binding is given in Box on Concentration, Affinity, and Cooperative Binding. Cooperative binding of regulatory proteins is used to ensure that changes in the level of expression of a given gene can be dramatic even in response to small changes in the level of a signal that controls that gene. The lysogenic induction of _, discussed below, provides an excellent example of this sensitive aspect of control. In some systems, cooperative binding between activators is also the basis of signal integration



λ **Repressor.** N indicates the amino domain, C the carboxy domain. "Tetramerization" denotes the region where two dimers interact when binding cooperatively to sites on DNA. These patches mediate octamerization as well.

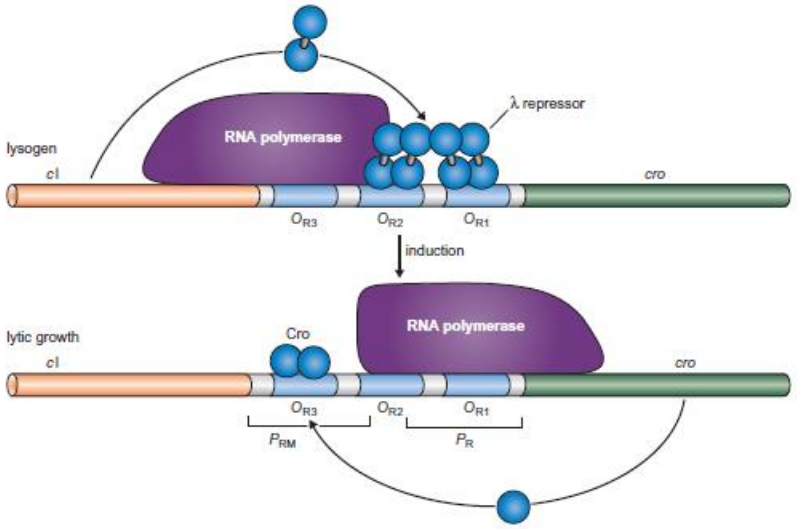
Repressor and Cro Bind in Different Patterns to Control Lytic

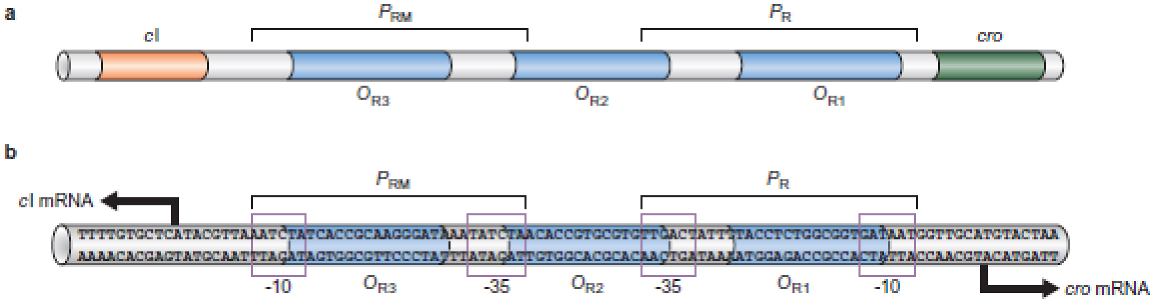
and Lysogenic Growth - How do repressor and Cro control the different patterns of gene expression associated with the different ways _ can grow?



Cooperative Binding of Repressor to DNA. The _ repressor monomers interact to form dimers, and those dimers interact to form tetramers. These interactions ensure that binding of repressor to DNA is cooperative. That cooperative binding is helped further by interactions between repressor tetramers at *O*R interacting with others at *O*L.

For lytic growth, a single Cro dimer is bound to *O*R3; this site overlaps *P*RM and so Cro represses that promoter (which would only work at a low level anyway in the absence of activator because the promoter is weak) neither repressor nor Cro is bound to *O*R1and *O*R2, *P*R binds RNA polymerase and directs transcription of lytic genes; *P*L does likewise. Recall that both *P*R and *P*L are strong promoters that need no activator. During lysogeny, *P*RM is on, while *P*R (and *P*L) are off. Repressorbound cooperatively at *O*R1 and *O*R2 blocks RNA polymerase binding at *P*R, repressing transcription from that promoter. But repressor bound at *O*R2 *activates* transcription from *P*RM.




a
cI
P_{RM}
P_R
cro
O_{R3}
O_{R2}
O_{R1}
b
cI mRNA
P_{RM}
P_R
TTTTGTGCTCATACGTTAAATCTATCACCGCAAGGGATAAATATCTAACACCGTGCGTGTTGACTATTTTACCTCTGGCGGTGATAATGGTTGCATGTACTAA
AAAACACGAGTATGCAATTTAGATAGTGGCGTTCCCTATTTATAGATTGTGGCACGCACAACTGATAAAATGGAGACCGCCACTATTACCAACGTACATGATT
-10
O_{R3}
-35
O_{R2}
-35
O_{R1}
-10
cro mRNA

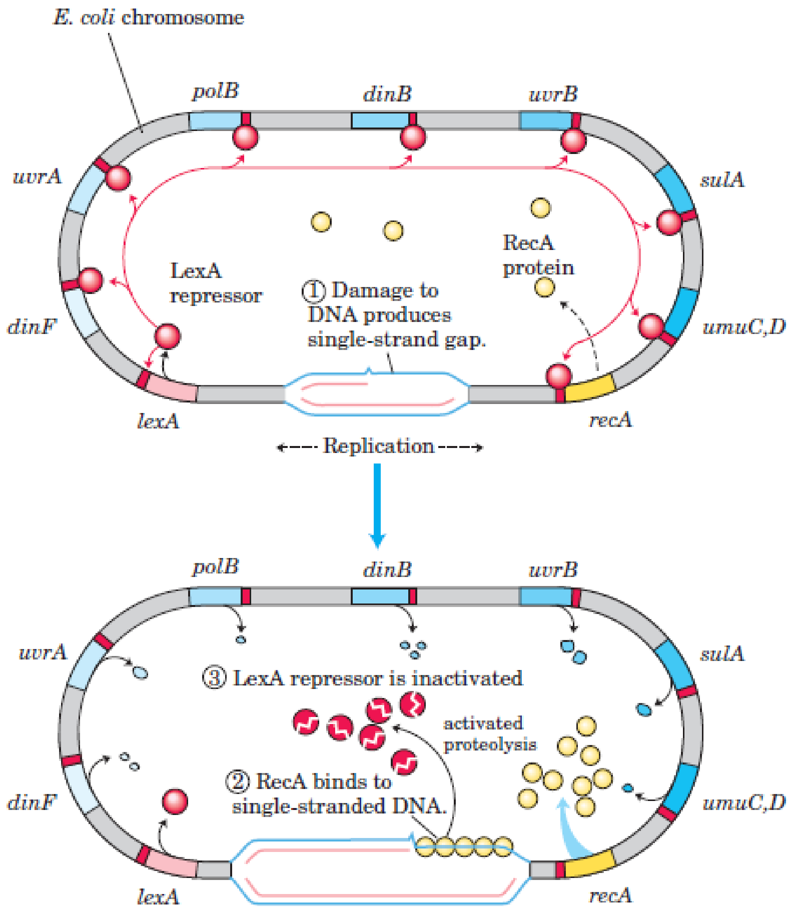
Relative Positions of Promoter and Operator Sites in *O*R. Note that *O*R2 overlaps the _35 region of *P*R by three base pairs, and that of *P*RM by two. This difference is enough for *P*R to be repressed and *P*RM activated by repressor bound at *O*R2.We return to the question of how the phage chooses between these alternative pathways shortly. But first we consider induction—how the lysogenic state outlined above switches to the alternative lytic one when the cell is threatened.

Lysogenic Induction Requires Proteolytic Cleavage of λ Repressor - *E. coli* senses and responds to DNA damage. It does this by activating the function of a protein called RecA. This enzyme is involved in recombination (which accounts for its name; but it has another function. That is, it stimulates the proteolytic autocleavage of cerain proteins. The primary substrate for this activity is a bacterial repressor protein called LexA that represses genes encoding DNA repair enzymes. Activated RecA stimulates autocleavage of LexA, releasing repression of those genes. This is called the SOS response.

If the cell is a lysogen, it is in the best interests of the prophage to escape under these threatening circumstances. To this end, _ repressor has evolved to resemble LexA, ensuring that _ repressor too undergoes autocleavage in response to activated RecA. The cleavage reaction removes the C-terminal domain of repressor, and so dimerization and cooperativity are immediately lost. As these functions are critical for repressor binding to *O*R1 and *O*R2 (at concentrations of repressor found in a lysogen), loss of cooperativity ensures that repressor dissociates from those sites (as well as from *O*L1 and *O*L2). Loss of repression triggers transcription from *P*R and *P*L leading to lytic growth. This switch from lysogenic to lytic growth is called **induction.** For induction to work efficiently, the level of repressor in a lysogen must be tightly regulated. If levels were to drop too low, the lysogen might spontaneously induce; if levels rose too high, appropriate induction would be inefficient. The reason for the latter is that more repressor would have to be inactivated (by RecA) for the concentration to drop enough to vacate *O*R1 and *O*R2. We have already seen how repressor ensures that its level never drops too low: it activates its own expression, an example of **positive autoregulation.** But how does it ensure levels never get too high? Repressor also regulates itself negatively.This **negative autoregulation** works as follows. As drawn, Figure shows *P*RM being activated by repressor (at *O*R2) to make more repressor. But if the concentration gets too high, repressor will bind to *O*R3 as well, and repress *P*RM (in a manner analogous to Cro binding *O*R3and repressing *P*RM during lytic growth). This prevents synthesis of new repressor until its concentration falls to a level at which it vacates *O*R3. It is interesting to note that the term “induction” is used to describe both the switch from lysogenic to lytic growth in _, and the switching on of the *lac* genes in response to lactose. This common usage stems from the fact that both phenomena were studied in parallel by Jacob and Monod. It is also worth noting that, just as lactose induces a conformational change in Lac repressor to relieve repression of the *lac* genes, so too the inducing signals of _ work by causing a structural

Induction of the SOS Response Requires Destruction of Repressor Proteins

Extensive DNA damage in the bacterial chromosome triggers the induction of many distantly located genes. This response, called the SOS response, provides another good example of coordinated gene regulation. Many of the induced genes are involved in DNA repair. The key regulatory proteins are the RecA protein and the LexA repressor. The LexA repressor (*M*r 22,700) inhibits transcription of all the SOS genes, and induction of the SOS response requires removal of LexA. This is not a simple dissociation from DNA in response to binding of a small molecule, as in the regulation of the *lac* operon described above. Instead, the LexA repressor is **SOS response in *E. coli.*** The LexA protein is the repressor in this system, which has an operator site (red) near each gene. Because the *recA* gene is not entirely repressed by the LexA repressor, the normal cell contains about 1,000 RecA monomers. 1 When DNA is extensively damaged (e.g., by UV light), DNA replication is halted and the number of single-strand gaps in the DNA increases. 2 RecA protein binds to this damaged, single-stranded DNA, activating the protein’s coprotease activity. 3 While bound toDNA, the RecA protein facilitates cleavage and inactivation of the LexA repressor. When the repressor is inactivated, the SOS genes, including*recA*, are induced; RecA levels increase 50- to100-fold.


E. coli chromosome
polB
dinB
uvrB
uvrA
sulA
RecA protein
LexA repressor
① Damage to DNA produces single-strand gap.
dinF
umuC,D
lexA
recA
Replication
polB
dinB
uvrB
uvrA
sulA
③ LexA repressor is inactivated
activated proteolysis
② RecA binds to single-stranded DNA.
dinF
umuC,D
lexA
recA

UNIT-IV

SYLLABUS

Regulation of gene expression in eukaryotes - Heterochromatin, euchromatin, chromatin remodeling, regulation of galactose metabolism in yeast, regulation by phosphorylation of nuclear transcription factors, regulatory RNAs, riboswitches, RNA interference, synthesis and function of miRNA molecules, phosphorylation of nuclear transcription factors.

Regulation of Gene Expression in Eukaryotes

Initiation of transcription is a crucial regulation point for both prokaryotic and eukaryotic gene expression. Although some of the same regulatory mechanisms are used in both systems, there is a fundamental difference in the regulation of transcription in eukaryotes and bacteria.

We can define a transcriptional ground state as the inherent activity of promoters and transcriptional machinery in vivo in the absence of regulatory sequences. In bacteria, RNA polymerase generally has access to every promoter and can bind and initiate transcription at some level of efficiency in the absence of activators or repressors; the transcriptional ground state is therefore nonrestrictive. In eukaryotes, however, strong promoters are generally inactive in vivo in the absence of regulatory proteins; that is, the transcriptional ground state is restrictive. This fundamental difference gives rise to at least four important features that distinguish the regulation of gene expression in eukaryotes from that in bacteria.

First, access to eukaryotic promoters is restricted by the structure of chromatin, and activation of transcription is associated with many changes in chromatin structure in the transcribed region. Second, although eukaryotic cells have both positive and negative regulatory mechanisms, positive mechanisms predominate in all systems characterized so far. Thus, given that the transcriptional ground state is restrictive, virtually every eukaryotic gene requires activation to be transcribed. Third, eukaryotic cells have larger, more complex multimeric regulatory proteins than do bacteria. Finally, transcription in the eukaryotic nucleus is separated from translation in the cytoplasm in both space and time.

The complexity of regulatory circuits in eukaryotic cells is extraordinary, as the following discussion shows. We conclude the section with an illustrated description of one of the most elaborate circuits: the regulatory cascade that controls development in fruit flies.

Transcriptionally Active Chromatin Is Structurally Distinct from Inactive Chromatin

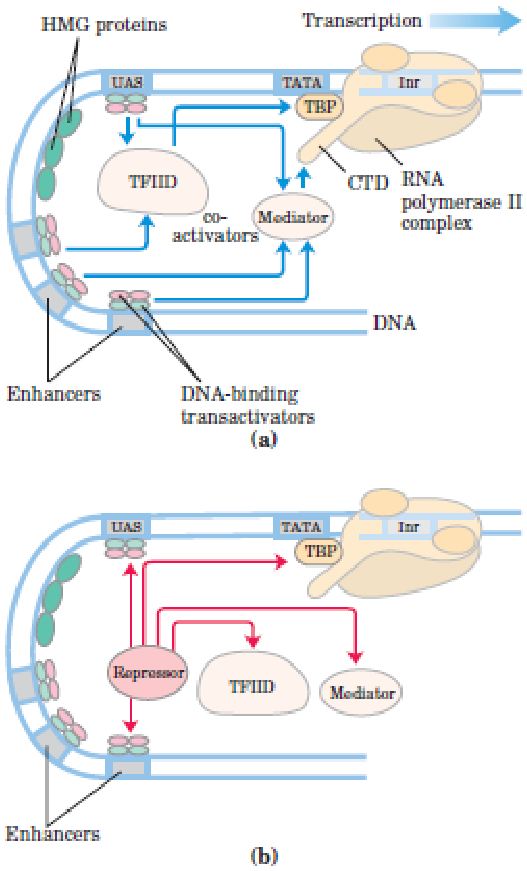
The effects of chromosome structure on gene regulation in eukaryotes have no clear parallel in prokaryotes. In the eukaryotic cell cycle, interphase chromosomes appear, at first viewing, to be dispersed and amorphous. Nevertheless, several forms of chromatin can be found along these chromosomes. About 10% of the chromatin in a typical eukaryotic cell is in a more condensed form than the rest of the chromatin. This form, **heterochromatin,** is transcriptionally inactive. Heterochromatin is generally associated with particular chromosome structures—the centromeres, for example. The remaining, less condensed chromatin is called **euchromatin.**

Transcription of a eukaryotic gene is strongly repressed when its DNA is condensed within heterochromatin. Some, but not all, of the euchromatin is transcriptionally active. Transcriptionally active chromosomal regions can be detected based on their increased sensitivity to nuclease-mediated degradation. Nucleases such as DNase I tend to cleave the DNA of carefully isolated chromatin into fragments of multiples of about 200 bp, reflecting the regular repeating structure of the nucleosome. In actively transcribed regions, the fragments produced by nuclease activity are smaller and more heterogeneous in size. These egions contain **hypersensitive sites,** sequences especially sensitive to DNase I, which consist of about 100 to 200 bp within the 1,000 bp flanking the 5’ ends of transcribed genes. In some genes, hypersensitive sites are found farther from the 5’ end, near the 3’ end, or even within the gene itself.

Many hypersensitive sites correspond to bindingsites for known regulatory proteins, and the relative absence of nucleosomes in these regions may allow the binding of these proteins. Nucleosomes are entirely absent in some regions that are very active in transcription, such as the rRNA genes.

Transcriptionally active chromatin tends to be deficient in histone H1, which binds to the linker DNA between nucleosome particles.

Histones within transcriptionally active chromatin and heterochromatin also differ in their patterns of covalent modification. The core histones of nucleosome particles (H2A, H2B, H3, H4; are modified by irreversible methylation of Lys residues, phosphorylation of Ser or Thr residues, acetylation (see below), or attachment of ubiquitin. Each of the core histones has two distinct structural domains. A central domain is involved in histone-histone interaction and the wrapping of DNA around the nucleosome. A second, lysine-rich amino-terminal domain is generally positioned near the exterior of the assembled nucleosome particle; the covalent modifications occur at specific residues concentrated in this amino-terminal domain. The patterns of modification have led some researchers to propose the existence of a histone code, in which modification patterns are recognized by enzymes that alter the structure of chromatin. Modifications associated with transcriptional activation would be recognized by enzymes that make the chromatin more accessible to the transcription machinery. 5-Methylation of cytosine residues of CpG sequences is common in eukaryotic DNA (p. 296), but DNA in transcriptionally active chromatin tends to be undermethylated. Furthermore, CpG sites in particular genes are more often undermethylated in cells from tissues where the genes are expressed than in those where the genes are not expressed. The overall pattern suggests that active chromatin is prepared for transcription by the removal of potential structural barriers.



Eukaryotic promoters and regulatory proteins. RNA polymerase II and its associated general transcription factors form a preinitiation complex at the TATA box and Inr site of the cognate promoters, a process facilitated by DNA-binding transactivators, acting through TFIID and/or mediator. **(a)** A composite promoter with typical sequence elements and protein complexes found in both yeast and higher eukaryotes. The carboxyl-terminal domain (CTD) of Pol II (see Fig. 26–9) is an important point of interaction with mediator and other protein complexes. Not shown are the protein complexes required for

histone acetylation and chromatin remodeling. For the DNA-binding transactivators, DNA-binding domains are shown in green, activation domains in pink. The interactions symbolized by blue arrows are discussed in the text. **(b)** A wide variety of eukaryotic transcriptional repressors function by a range of mechanisms. Some bind directly to DNA, displacing a protein complex required for activation; others interact with various parts of the transcription or activation complexes to prevent activation. Possible points of interaction are indicated with red arrows.

Chromatin Is Remodeled by Acetylation and Nucleosomal Displacements

The detailed mechanisms for transcription-associated structural changes in chromatin, called **chromatin remodeling,** are now coming to light, including identification of a variety of enzymes directly implicated in the process. These include enzymes that covalently modify the core histones of the nucleosome and others that use the chemical energy of ATP to remodel nucleosomes on the DNA.

The acetylation and deacetylation of histones figure prominently in the processes that activate chromatin for transcription. As noted above, the amino-terminal domains of the core histones are generally rich in Lys residues. Particular Lys residues are acetylated by histone acetyltransferases (HATs). Cytosolic (type B) HATs acetylate newly synthesized histones before the histones are imported into the nucleus. The subsequent assembly of the histones into chromatin is facilitated by additional proteins: CAF1 for H3 and H4, and NAP1 for H2A and H2B.

Where chromatin is being activated for transcription, the nucleosomal histones are further acetylated by nuclear (type A) HATs. The acetylation of multiple Lys residues in the amino-terminal domains of histones H3 and H4 can reduce the affinity of the entire nucleosome for DNA. Acetylation may also prevent or promote interactions with other proteins involved in transcription or its regulation. When transcription of a gene is no longer required, the acetylation of nucleosomes in that vicinity is reduced by the activity of histone deacetylases, as part of a general gene-silencing process that restores the chromatin to a transcriptionally inactive state. In addition to the removal of certain acetyl groups, new covalent modification of histones marks chromatin as transcriptionally inactive. As an example, the Lys residue at position 9 in histone H3 is often methylated in heterochromatin.

Chromatin remodeling also requires protein complexes that actively move or displace nucleosomes, hydrolyzing ATP in the process. The enzyme complex SWI/SNF found in all eukaryotic cells, contains 11 polypeptides (total *M*r 2 _ 106) that together create hypersensitive sites in the chromatin and stimulate the binding of transcription factors. SWI/SNF is not required for the transcription of every gene. NURF is another ATP-dependent enzyme complex that remodels chromatin in ways that complement and overlap the activity of SWI/SNF. These enzyme complexes play an important role in preparing a region of chromatin for active transcription.

Many Eukaryotic Promoters Are Positively Regulated

As already noted, eukaryotic RNA polymerases have little or no intrinsic affinity for their promoters; initiation of transcription is almost always dependent on the action of multiple activator proteins. One important reason for the apparent predominance of positive regulation seems obvious: the storage of DNA within chromatin effectively renders most promoters inaccessible, so genes are normally silent in the absence of other regulation. The structure of chromatin affects access to some promoters more than others, but repressors that bind to DNA so as to preclude access of RNA polymerase (negative regulation) would often be simply redundant. Other factors are at play in the use of positive regulation, and speculation generally centers around two: the large size of eukaryotic genomes and the greater efficiency of positive regulation.

First, nonspecific DNA binding of regulatory proteins becomes a more important problem in the much larger genomes of higher eukaryotes. And the chance that a single specific binding sequence will occur randomly at an inappropriate site also increases with genome size. Specificity for transcriptional activation can be improved if each of several positive-regulatory proteins must bind specific DNA sequences and then form a complex in order to become active. The average number of regulatory sites for a gene in a multicellular organism is probably at least five. The requirement for binding of several positive-regulatory proteins to specific DNA sequences vastly reduces the probability of the random occurrence of a functional juxtaposition of all the necessary binding sites. In principle, a similar strategy could be used by multiple negative-regulatory elements, but this brings us to the second reason for the use of positive regulation: it is simply more efficient. If the 30,000 to 35,000 genes in the human genome were negatively regulated, each cell would have to synthesize, at all times, this same number of different repressors (or many times this number if multiple regulatory elements were used at each promoter) in concentrations sufficient to permit specific binding to each “unwanted” gene. In positive regulation, most of the genes are normally inactive (that is, RNA polymerases do not bind to the promoters) and the cell synthesizes only the activator proteins needed to promote transcription of the subset of genes required in the cell at that time. These arguments notwithstanding, there are examples of negative regulation in eukaryotes, from yeast to humans, as we shall see.

DNA-Binding Transactivators and Coactivators Facilitate Assembly of the General Transcription Factors

To continue our exploration of the regulation of gene expression in eukaryotes, we return to the interactions between promoters and RNA polymerase II (Pol II), the enzyme responsible for the synthesis of eukaryotic mRNAs. Although most (but not all) Pol II promoters include the TATA box and Inr (initiator) sequences, with their standard spacing (see Fig. 26–8), they vary greatly in both the number and the location of additional sequences required for the regulation of transcription. These additional regulatory sequences are usually called **enhancers** in higher eukaryotes and **upstream activator sequences (UASs)** in yeast. A typical enhancer may be found hundreds or even thousands of base pairs upstream from the transcription start site, or may even be downstream, within the gene itself. When bound by the appropriate regulatory proteins, an enhancer increases transcription at nearby promoters regardless of its orientation in the DNA. The UASs of yeast function in a similar way, although generally they must be positioned upstream and within a few hundred base pairs of the transcription start site. An average Pol II promoter may be affected by a half-dozen regulatory sequences of this type, and even more complex promoters are quite common.

Successful binding of active RNA polymerase II holoenzyme at one of its promoters usually requires the action of other proteins (Fig. 28–27), of three types:
(1) **basal transcription factors** (see Fig. 26–9, Table 26–1), required at every Pol II promoter; (2) **DNAbinding transactivators,** which bind to enhancers or UASs and facilitate transcription; and (3) **coactivators.** The latter group act indirectly—not by binding to the DNA—and are required for essential communication between the DNA-binding transactivators and the complex composed of Pol II and the general transcription factors. Furthermore, a variety of repressor proteins can interfere with communication between the RNA polymerase and the DNA-binding transactivators, resulting in repression of transcription (Fig. 28–27b). Here we focus on the protein complexes shown in Figure 28–27 and on how they interact to activate transcription.

TATA-Binding Protein The first component to bind in the assembly of a preinitiation complex at the TATA box of a typical Pol II promoter is the **TATA-binding protein (TBP).** The complete complex includes the basal (or general) transcription factors TFIIB, TFIIE, TFIIF, TFIIH; Pol II; and perhaps TFIIA (not all of the factors are shown in Fig. 28–27). This minimal preinitiation complex, however, is often insufficient for the initiation of transcription and generally does not form at all if the promoter is obscured within chromatin. Positive regulation leading to transcription is imposed by the transactivators and coactivators.

DNA-Binding Transactivators The requirements for transactivators vary greatly from one promoter to another. A few transactivators are known to facilitate transcription at hundreds of promoters, whereas others are specific for a few promoters. Many transactivators are sensitive to the binding of signal molecules, providing the capacity to activate or deactivate transcription in response to a changing cellular environment. Some enhancers bound by DNA-binding transactivators are quite distant from the promoter's TATA box. How do the transactivators function at a distance? The answer in most cases seems to be that, as indicated earlier, the intervening DNA is looped so that the various protein complexes can interact directly. The looping is promoted by certain non-bind nonspecifically to DNA. These high mobility group (HMG) proteins (Fig. 28–27; "high mobility" refers to their electrophoretic mobility in polyacrylamide gels) play an important structural role in chromatin remodeling and transcriptional activation.

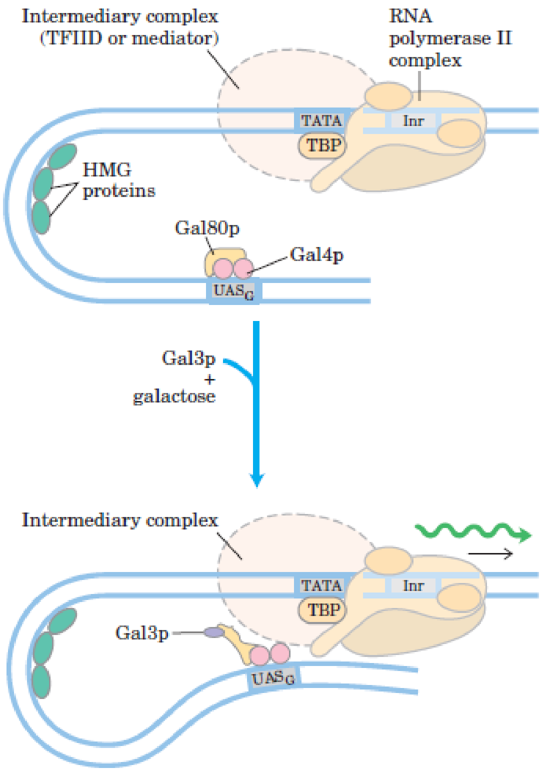
Coactivator Protein Complexes Most transcription requires the presence of additional protein complexes. Some major regulatory protein complexes that interact with Pol II have been defined both genetically and biochemically. These coactivator complexes act as intermediaries between the DNA-binding transactivators and the Pol II complex. The best-characterized coactivator is the transcription factor TFIID (Fig. 28–27). In eukaryotes, TFIID is a large complex that includes TBP and ten or more TBPassociated factors (TAFs). Some TAFs resemble histones and may play a role in displacing nucleosomes during the activation of transcription. Many DNA-binding transactivators aid in transcription initiation by interacting with one or more TAFs. The requirement for TAFs to initiate transcription can vary greatly from one gene to another. Some promoters require TFIID, some do not, and some require only subsets of the TFIID TAF subunits. Another important coactivator consists of 20 or more polypeptides in a protein complex called **mediator** (Fig. 28–27); the 20 core polypeptides are highly conserved from fungi to humans. Mediator binds tightly to the carboxyl-terminal domain (CTD) of the largest subunit of Pol II. The mediator complex is required for both basal and regulated transcription at promoters used by Pol II, and it also stimulates the phosphorylation of the CTD by TFIIH. Both mediator and TFIID are required at some promoters. As with TFIID, some DNAbinding transactivators interact with one or more components of the mediator complex. Coactivator complexes function at or near the promoter's TATA box.

Choreography of Transcriptional Activation We can now begin to piece together the sequence of transcriptional activation events at a typical Pol II promoter. First, crucial remodeling of the chromatin takes place in stages. Some DNA-binding transactivators have significant affinity for their binding sites even when the sites are within condensed chromatin. Binding of one transactivator may facilitate the binding of others, gradually displacing some nucleosomes. The bound transactivators can then interact directly with HATs or enzyme complexes such as SWI/SNF (or both), accelerating the remodeling of the surrounding chromatin. In this way a bound transactivator can draw in other components necessary for further chromatin remodeling to permit transcription of specific genes. The bound transactivators, generally acting through complexes such as TFIID or mediator (or both), stabilize the binding of Pol II and its associated transcription factors and greatly facilitate formation of the preinitiation transcription complex. Complexity in these regulatory circuits is the rule rather than the exception, with multiple DNA-bound transactivators promoting transcription.The script can change from one promoter to another, but most promoters seem to require a precisely ordered assembly of components to initiate transcription. The assembly process is not always fast. At some genes it may take minutes; at certain genes in higher eukaryotes the process can take days.

Reversible Transcriptional Activation Although rarer, some eukaryotic regulatory proteins that bind to Pol II promoters can act as repressors, inhibiting the formation of active preinitiation complexes (Fig. 28–27b). Some transactivators can adopt different conformations, enabling them to serve as transcriptional activators or repressors. For example, some steroid hormone receptors (described later) function in the nucleus as DNAbinding transactivators, stimulating the transcription of certain genes when a particular steroid hormone signal is present. When the hormone is absent, the receptor proteins revert to a repressor conformation, *preventing* the formation of preinitiation complexes. In some cases this repression involves interaction with histone deacetylases and other proteins that help restore the surrounding chromatin to its transcriptionally inactive state.

The Genes of Galactose Metabolism in Yeast Are Subject to Both Positive and Negative Regulation

Some of the general principles described above can be illustrated by one well-studied eukaryotic regulatory circuit (Fig. 28–28). The enzymes required for the importation and metabolism of galactose in yeast are encoded by genes scattered over several chromosomes (Table 28–3). Each of the *GAL* genes is transcribed separately, and yeast cells have no operons like those in bacteria. However, all the *GAL* genes have similar promoters and are regulated coordinately by a common set of proteins. The promoters for the *GAL* genes consist of the TATA box and Inr sequences, as well as an upstream activator sequence (UASG) recognized by a DNA-binding transcriptional activator known as Gal4 protein (Gal4p). Regulation of gene expression by galactose entails an interplay between Gal4p and two other proteins, Gal80p and Gal3p (Fig. 28–28). Gal80p forms a complex with Gal4p, preventing Gal4p from functioning as an activator of the *GAL* promoters. When galactose is present, it binds Gal3p, which then interacts with Gal80p, allowing Gal4p to function as an activator at the various *GAL* promoters. Other protein complexes also have a role in activating transcription of the *GAL* genes. These may include the SAGA complex for histone acetylation, the SWI/SNF complex for nucleosome remodeling, and the mediator complex. Figure 28–29 provides an idea of the complexity of protein interactions in the overall process of transcriptional activation in eukaryotic cells. Glucose is the preferred carbon source for yeast, as it is for bacteria. When glucose is present, most of the *GAL* genes are repressed—whether galactose is present or not. The *GAL* regulatory system described above is effectively overridden by a complex catabolite repression system that includes several proteins (not depicted in Fig. 28–29).



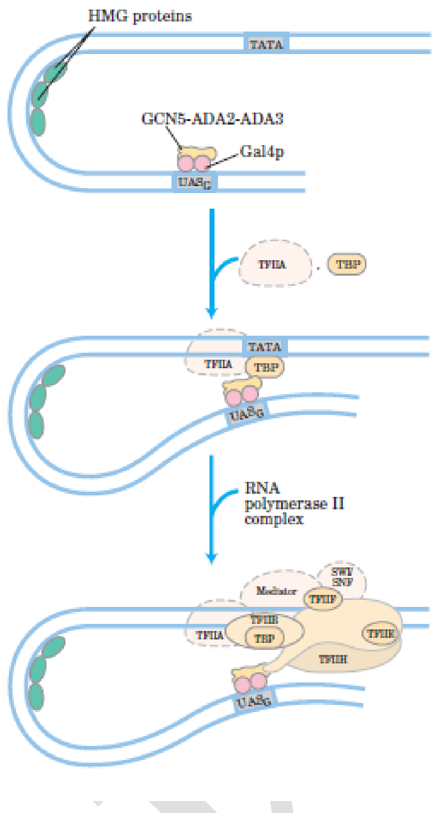
Regulation of transcription at genes of galactose metabolism in yeast. Galactose is imported into the cell and converted to galactose 6-phosphate by a pathway involving six enzymes whose genes are scattered over three chromosomes (see Table 28–3). Transcription of these genes is regulated by the combined actions of the proteins Gal4p, Gal80p, and Gal3p, with Gal4p playing the central role of DNA-binding transactivator. The Gal4p-Gal80p complex is inactive in gene activation. Binding of galactose to Gal3p and its interaction with Gal80p produce a conformational change in Gal80p that allows Gal4p to function in transcription activation.

DNA-Binding Transactivators Have a Modular Structure

DNA-binding transactivators typically have a distinct structural domain for specific DNA binding and one or more additional domains for transcriptional activation or for interaction with other regulatory proteins. Interaction of two regulatory proteins is often mediated by domains containing leucine zippers (Fig. 28–14) or helixloop- helix motifs (Fig. 28–15). We consider here three distinct types of structural domains used in activation by DNA-binding transactivators (Fig. 28–30a): Gal4p, Sp1, and CTF1.

Gal4p contains a zinc fingerlike structure in its DNA-binding domain, near the amino terminus; this domain has six Cys residues that coordinate two Zn2_. The protein functions as a homodimer (with dimerization mediated by interactions between two coiled coils) and binds to UASG, a palindromic DNA sequence about 17 bp long. Gal4p has a separate activation domain with many acidic amino acid residues. Experiments that substitute a variety of different peptide sequences for the **acidic activation domain** of Gal4p suggest that the acidic nature of this domain is critical to its function, although its precise amino acid sequence can vary considerably. Sp1 (*M*r 80,000) is a DNA-binding transactivator for a large number of genes in higher eukaryotes. Its DNA binding site, the GC box (consensus sequence GGGCGG), is usually quite near the TATA box. The DNA-binding domain of the Sp1 protein is near its carboxyl terminus and contains three zinc fingers. Two other domains in Sp1 function in activation, and are notable in that 25% of their amino acid residues are Gln. A wide variety of other activator proteins also have these **glutamine-rich domains.**

CCAAT-binding transcription factor 1 (CTF1) belongs to a family of DNA-binding transactivators that bind a sequence called the CCAAT site (its consensus sequence is TGGN6GCCAA, where N is any nucleotide). The DNA-binding domain of CTF1 contains many basic amino acid residues, and the binding region is probably arranged as an _ helix. This protein has neither a helixturn- helix nor a zinc finger motif; its DNA-binding mechanism is not yet clear. CTF1 has a **proline-rich activation domain,** with Pro accounting for more than 20% of the amino acid residues. The discrete activation and DNA-binding domains of regulatory proteins often act completely independently,as has been demonstrated in “domain-swapping”experiments. Genetic engineering techniques can join the proline-rich activation domain of CTF1 to the DNA-binding domain of Sp1 to create a protein that, like normal Sp1, binds to GC boxes on the DNA and activates transcription at a nearby promoter (as in Fig. 28–30b). The DNA-binding domain of Gal4p has similarly been replaced experimentally with the DNAbinding domain of the prokaryotic LexA repressor (of the SOS response; Fig. 28–22). This chimeric protein neither binds at UASG nor activates the yeast *GAL* genes (as would normal Gal4p) unless the UASG sequence in the DNA is replaced by the LexA recognition site.

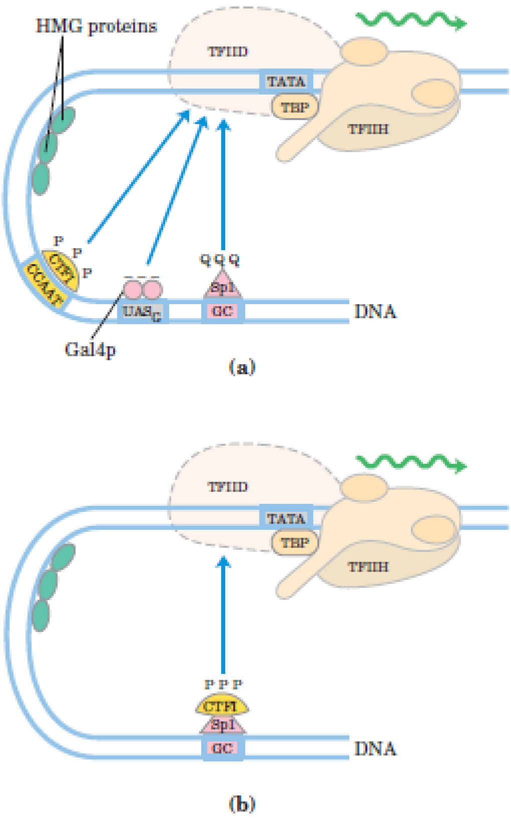


Protein complexes involved in transcription activation of a group of related eukaryotic genes. The *GAL* system illustrates the complexity of this process, but not all these protein complexes are yet known to affect *GAL* gene transcription. Note that many of the complexes (such as SWI/SNF, GCN5-ADA2-ADA3, and mediator) affect the transcription of many genes. The complexes assemble stepwise. First the DNA-binding transactivators bind, then the additional protein complexes needed to remodel the chromatin and allow transcription to begin.

Eukaryotic Gene Expression Can Be Regulated by Intercellular and Intracellular Signals

The effects of steroid hormones (and of thyroid and retinoid hormones, which have the same mode of action) provide additional well-studied examples of the modulation of eukaryotic regulatory proteins by direct interaction with molecular signals (see Fig. 12–40). Unlike other types of hormones, steroid hormones do not have to bind to plasma membrane receptors. Instead, they can interact with intracellular receptors that are themselves transcriptional transactivators. Steroid hormones too hydrophobic to dissolve readily in the blood (estrogen, progesterone, and cortisol, for example) travel on specific carrier proteins from their point of release to their target tissues. In the target tissue, the hormone passes through the plasma membrane by simple diffusion and binds to its specific receptor protein in the nucleus. The hormone-receptor complex acts by binding to highly specific DNA sequences called **hormone response elements (HREs),** thereby altering gene expression.

Hormone binding triggers changes in the conformation of the receptor proteins so that they become capable of interacting with additional transcription factors. The bound hormone-receptor complex can either enhance or suppress the expression of adjacent genes. The DNA sequences (HREs) to which hormonereceptor complexes bind are similar in length and arrangement, but differ in sequence, for the various steroid hormones. Each receptor has a consensus HRE sequence (Table 28–4) to which the hormone-receptor complex binds well, with each consensus consisting of two six-nucleotide sequences, either contiguous or separated by three nucleotides, in tandem or in a palindromic arrangement. The hormone receptors have a highly conserved DNA-binding domain with two zinc fingers. The hormone-receptor complex binds to the DNA as a dimer, with the zinc finger domains of each monomer recognizing one of the six-nucleotide sequences. The ability of a given hormone to act through the hormone-receptor complex to alter the expression of a specific gene depends on the exact sequence of the HRE, its position relative to the gene, and the number of HREs associated with the gene. Unlike the DNA-binding domain, the ligand-binding region of the receptor protein—always at the carboxyl terminus—is quite specific to the particular receptor. In the ligand-binding region, the glucocorticoid receptor is only 30% similar to the estrogen receptor and 17% similar to the thyroid hormone receptor. The size of the ligand- binding region varies dramatically; in the vitamin D receptor it has only 25 amino acid residues, whereas in the mineralocorticoid receptor it has 603 residues. Mutations that change one amino acid in these regions can result in loss of responsiveness to a specific hormone. Some humans unable to respond to cortisol, testosterone, vitamin D, or thyroxine have mutations of this type.



DNA-binding transactivators. (a) Typical DNA-binding transactivators such as CTF1, Gal4p, and Sp1 have a DNA-binding domain and an activation domain. The nature of the activation domain is indicated by symbols: ___, acidic; Q Q Q, glutamine-rich; P P P, proline-rich. Some or all of these proteins may activate transcription by interacting with intermediary complexes such as TFIID or mediator. Note that the binding sites illustrated here are not generally found together near a single gene. **(b)** A chimeric protein containing the DNA-binding domain of Sp1 and the activation domain of CTF1activates transcription if a GC box is present.

Regulation Can Result from Phosphorylation of Nuclear Transcription Factors

We noted in Chapter 12 that the effects of insulin on gene expression are mediated by a series of steps leading ultimately to the activation of a protein kinase in the nucleus that phosphorylates specific DNA-binding proteins and thereby alters their ability to act as transcription factors (see Fig. 12–6). This general mechanism mediates the effects of many nonsteroid hormones. For example, the _-adrenergic pathway that leads to elevated levels of cytosolic cAMP, which acts as a second messenger in eukaryotes as well as in prokaryotes (see Figs 12–12, 28–18), also affects the transcription of a set of genes, each of which is located near a specific DNA sequence called a cAMP response element (CRE).

The catalytic subunit of protein kinase A, released when cAMP levels rise (see Fig. 12–15), enters the nucleus and phosphorylates a nuclear protein, the CRE-binding protein (CREB). When phosphorylated, CREB binds to CREs near certain genes and acts as a transcription factor, turning on the expression of these genes.

Many Eukaryotic mRNAs Are Subject to Translational Repression

Regulation at the level of translation assumes a much more prominent role in eukaryotes than in bacteria and is observed in a range of cellular situations. In contrast to the tight coupling of transcription and translation in bacteria, the transcripts generated in a eukaryotic nucleus must be processed and transported to the cytoplasm before translation. This can impose a significant delay on the appearance of a protein. When a rapid increase in protein production is needed, a translationally repressed mRNA already in the cytoplasm can be activated for translation without delay. Translational regulation may play an especially important role in regulating certain very long eukaryotic genes (a few are measured in the millions of base pairs), for which transcription and mRNA processing can require many hours. Some genes are regulated at both the transcriptional and translational stages, with the latter playing a role in the finetuning of cellular protein levels. In some anucleate cells, such as reticulocytes (immature erythrocytes), transcriptional control is entirely unavailable and translational control of stored mRNAs becomes essential. As described below, translational controls can also have spatial significance during development, when the regulated translation of prepositioned mRNAs creates a local gradient of the protein product. Eukaryotes have at least three main mechanisms of translational regulation. *1.* Initiation factors are subject to phosphorylation by a number of protein kinases. The phosphorylated forms are often less active and cause a general depression of translation in the cell.

2. Some proteins bind directly to mRNA and act as translational repressors, many of them binding at specific sites in the 3_ untranslated region (3_UTR). So positioned, these proteins interact with other translation initiation factors bound to the mRNA or with the 40S ribosomal subunit to prevent translation initiation (Fig. 28–32; compare this with Fig. 27–22). *3.* Binding proteins, present in eukaryotes from yeast to mammals, disrupt the interaction between eIF4E and eIF4G (see Fig. 27–22). The mammalian versions are known as 4E-BPs (eIF4E binding proteins). When cell growth is slow, these proteins limit translation by binding to the site on eIF4E that normally interacts with eIF4G. When cell growth resumes or increases in response to growth factors or other stimuli, the binding proteins are inactivated by protein kinase– dependent phosphorylation. The variety of translational regulation mechanisms provides flexibility, allowing focused repression of a few mRNAs or global regulation of all cellular translation. Translational regulation has been particularly well studied in reticulocytes. One such mechanism in these cells involves eIF2, the initiation factor that binds to the initiator tRNA and conveys it to the ribosome; when Met-tRNA has bound to the P site, the factor eIF2B binds to eIF2, recycling it with the aid of GTP binding and hydrolysis. The maturation of reticulocytes includes destruction of the cell nucleus, leaving behind a plasma membrane packed with hemoglobin. Messenger RNAs deposited in the cytoplasm before the loss of the nucleus allow for the replacement of hemoglobin. When reticulocytes become deficient in iron or heme, the translation of globin mRNAs is repressed. A protein kinase called HCR (*h*emin-*c*ontrolled *r*epressor) is activated, catalyzing the phosphorylation of eIF2. In its phosphorylated form, eIF2 forms a stable complex with eIF2B that sequesters the eIF2, making it unavailable for participation in translation. In this way, the reticulocyte coordinates the synthesis of globin with the availability of heme. Many additional examples of translational regulation have been found in studies of the development of multicellular organisms, as discussed in more detail below.

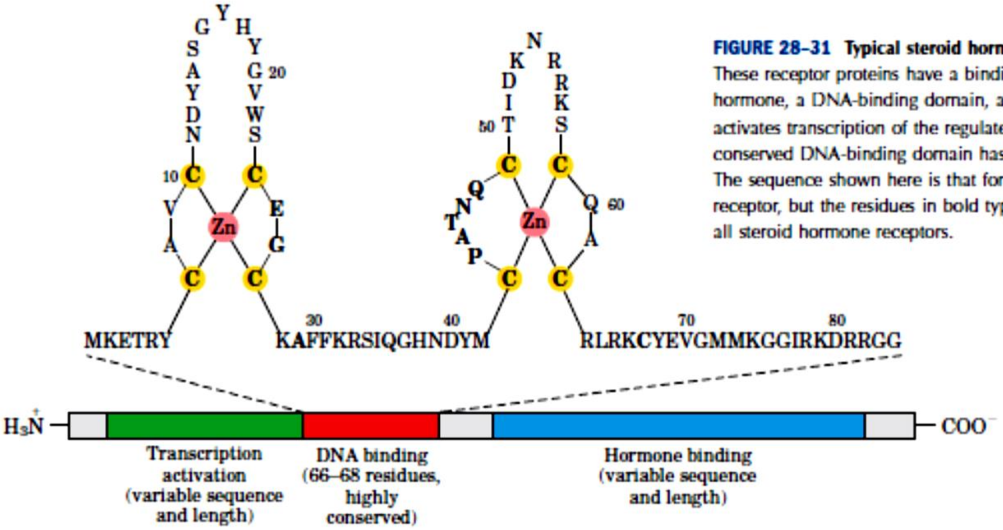
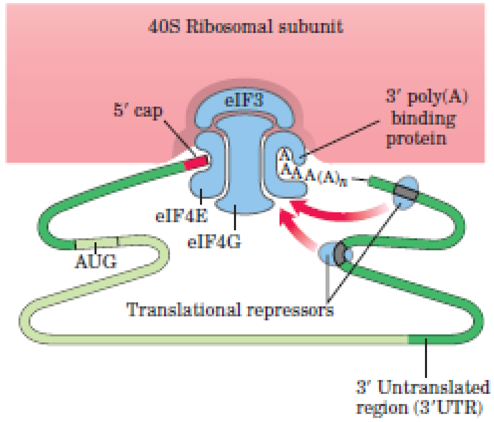

Zn
Zn
10
20
30
40
50
60
70
80
MKETRY
KAFFKRSIQGHNDYM
RLRKCYEVGMMKGGIRKDRRGG
$H_3\overset{+}{N}$
COO^-
Transcription activation (variable sequence and length)
DNA binding (66–68 residues, highly conserved)
Hormone binding (variable sequence and length)

FIGURE 28–31 Typical steroid hormone receptors. These receptor proteins have a binding site for the hormone, a DNA-binding domain, and a region that activates transcription of the regulated gene. The highly conserved DNA-binding domain has two zinc fingers. The sequence shown here is that for the estrogen receptor, but the residues in bold type are common to all steroid hormone receptors.

Posttranscriptional Gene Silencing Is Mediated by RNA Interference

In higher eukaryotes, including nematodes, fruit flies, plants, and mammals, a class of small RNAs has been discovered that mediates the silencing of particular genes. The RNAs function by interacting with mRNAs, often in the 3_UTR, resulting in either mRNA degradation or translation inhibition. In either case, the mRNA, and thus the gene that produces it, is silenced. This form of gene regulation controls developmental timing in at least some organisms. It is also used as a mechanism to protect against invading RNA viruses (particularly **1110** Chapter 28 Regulation of Gene Expression **FIGURE 28–32 Translational regulation of eukaryotic mRNA.** One of the most important mechanisms for translational regulation in eukaryotes involves the binding of translational repressors (RNA-binding

proteins) to specific sites in the 3_ untranslated region (3_UTR) of the mRNA. These proteins interact with eukaryotic initiation factors or with the ribosome (see Fig. 27–22) to prevent or slow translation. important in plants, which lack an immune system) and to control the activity of transposons. In addition, small RNA molecules may play a critical (but still undefined) role in the formation of heterochromatin. The small RNAs are sometimes called micro-RNAs (miRNAs). Many are present only transiently during development, and these are sometimes referred to as small temporal RNAs (stRNAs). Hundreds of different miRNAs have been identified in higher eukaryotes. They are transcribed as precursor RNAs about 70 nucleotides long, with internally complementary sequences that form hairpinlike structures (Fig. 28–33). The precursors are cleaved by endonucleases to form short duplexes about 20 to 25 nucleotides long. The best-characterized nuclease goes by the delightfully suggestive name Dicer; endonucleases in the Dicer family are widely distributed in higher eukaryotes. One strand of the processed miRNA is transferred to the target mRNA (or to a viral or transposon RNA), leading to inhibition of translation or degradation of the RNA (Fig. 28–33a). This gene regulation mechanism has an interesting and very useful practical side. If an investigator introduces into an organism a duplex RNA molecule corresponding in sequence to virtually any mRNA, the Dicer endonuclease cleaves the duplex into short segments, called small interfering RNAs (siRNAs). These bind to the mRNA and silence it (Fig. 28–33b). The process is known as **RNA interference** (**RNAi).** In plants, virtually any gene can be effectively shut down in this way. In nematodes, simply introducing the duplex RNA into the worm’s diet produces very effective suppression of the target gene. The technique has rapidly become an important tool in the ongoing efforts to study gene function, because it can disrupt gene function without creating a mutant organism. The procedure can be applied to humans as well. Laboratory-produced siRNAs have already been used to block HIV and poliovirus infections in cultured human cells for a week or so at a time. Although this work is in its infancy, the rapid progress makes RNA interference a field to watch for future medical advances.



Translational regulation of eukaryotic mRNA. One of the most important mechanisms for translational regulation in eukaryotes involves the binding of translational repressors (RNA-binding proteins) to specific sites in the 3_ untranslated region (3_UTR) of the mRNA. These proteins interact with eukaryotic initiation factors or with the ribosome to prevent or slow translation.

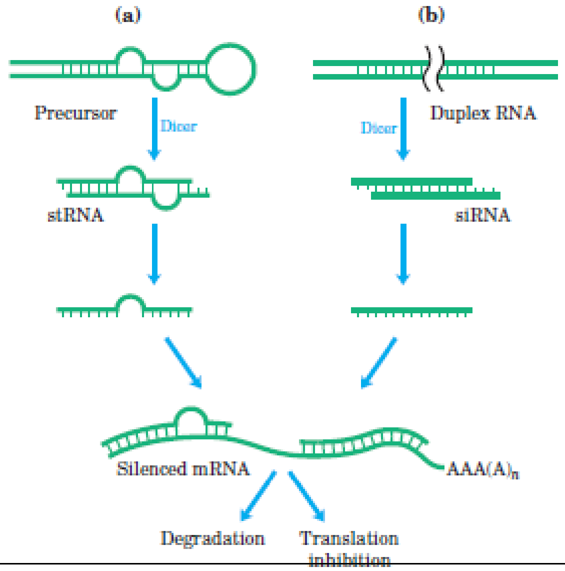
Development Is Controlled by Cascades of Regulatory Proteins

For sheer complexity and intricacy of coordination, the patterns of gene regulation that bring about development of a zygote into a multicellular animal or plant have no peer. Development requires transitions in morphology and protein composition that depend on tightly coordinated changes in expression of the genome. More genes are expressed during early development than in any other part of the life cycle. For example, in the sea urchin, an oocyte has about 18,500 *different* mRNAs, compared with about 6,000 different mRNAs in the cells of a typical differentiated tissue. The mRNAs in the oocyte give rise to a cascade of events that regulate the expression of many genes across both space and time. Several animals have emerged as important model systems for the study of development, because they are easy to maintain in a laboratory and have relatively short generation times. These include nematodes, fruit flies, zebra fish, mice, and the plant *Arabidopsis*. This discussion focuses on the development of fruit flies. Our understanding of the molecular events during development of

Drosophila melanogaster is particularly well advanced and can be used to illustrate patterns and principles of general significance.

The life cycle of the fruit fly includes complete metamorphosis during its progression from an embryo to an adult. Among the most important characteristics of the embryo are its **polarity** (the anterior and posterior parts of the animal are readily distinguished, as are its dorsal and ventral parts) and its **metamerism** (the embryo body is made up of serially repeating segments, each with characteristic features). During development, these segments become organized into a head, thorax, and abdomen. Each segment of the adult thorax has a different set of appendages. Development of this complex pattern is under genetic control, and a variety of pattern-regulating genes have been discovered that dramatically affect the organization of the body.

The *Drosophila* egg, along with 15 nurse cells, is surrounded by a layer of follicle cells. As the egg cell forms (before fertilization), mRNAs and proteins originating in the nurse and follicle cells are deposited in the egg cell, where some play a critical role in development. Once a fertilized egg is laid, its nucleus divides and the nuclear descendants continue to divide in synchrony every 6 to 10 min. Plasma membranes are not formed around the nuclei, which are distributed within the egg cytoplasm (or syncytium). Between the eighth and eleventh rounds of nuclear division, the nuclei migrate to the outer layer of the egg, forming a monolayer of nuclei surrounding the common yolk-rich cytoplasm; this is the syncytial blastoderm. After a few additional divisions, membrane invaginations surround the nuclei to create a layer of cells that form the cellular blastoderm. At this stage, the mitotic cycles in the various cells lose their synchrony. The developmental fate of the cells is determined by the mRNAs and proteins originally deposited in the egg by the nurse and follicle cells.



Gene silencing by RNA interference. (a) Small temporal RNAs (stRNAs) are generated by Dicer-mediated cleavage of longer precursors that fold to create duplex regions. The stRNAs then bind to mRNAs, leading to degradation of mRNA or inhibition of translation. **(b)** Double-stranded RNAs can be constructed and introduced into a cell. Dicer processes the duplex RNAs into small interfering RNAs (siRNAs), which interact with the target mRNA. Again, the mRNA is either degraded or its translation inhibited.

Proteins that, through changes in local concentration or activity, cause the surrounding tissue to take up a particular shape or structure are sometimes referred to as **morphogens;** they are the products of pattern regulating genes. As defined by Christiane Nüsslein- Volhard, Edward B. Lewis, and Eric F. Wieschaus, three major classes of pattern-regulating genes—maternal, segmentation, and homeotic genes—function in successive stages of development to specify the basic features of the *Drosophila* embryo's body. **Maternal genes** are expressed in the unfertilized egg, and the resulting **maternal mRNAs** remain dormant until fertilization. These provide most of the proteins needed in very early development, until the cellular blastoderm is formed. Some of the proteins encoded by maternal mRNAs direct the spatial organization of the developing embryo at early stages, establishing its polarity.

Segmentation genes, transcribed after fertilization, direct the formation of the proper number of body segments. At least three subclasses of segmentation genes act at successive stages: **gap genes** divide the developing embryo into several broad regions, and **pair-rule genes** together with **segment polarity genes** define 14 stripes that become the 14 segments of a normal embryo. **Homeotic genes** are expressed still later; they specify which organs and appendages will develop in particular body segments.

The many regulatory genes in these three classes direct the development of an adult fly, with a head, thorax, and abdomen, with the proper number of segments, and with the correct appendages on each segment. Although embryogenesis takes about a day to complete, all these genes are activated during the first four hours. Some mRNAs and proteins are present for only a few minutes at specific points during this period. Some of the genes code for transcription factors that affect the expression of other genes in a kind of developmental cascade. Regulation at the level of translation also occurs, and many of the regulatory genes encode translational repressors, most of which bind to the 3_UTR of the mRNA (Fig. 28–32). Because many mRNAs are deposited in the egg long before their translation is required, translational repression provides an especially important avenue for regulation in developmental pathways.

Maternal Genes Some maternal genes are expressed within the nurse and follicle cells, and some in the egg itself. Within the unfertilized *Drosophila* egg, the maternal gene products establish two axes—anterior-posterior and dorsal-ventral—and thus define which regions of the radially symmetric egg will develop into the head and abdomen and the top and bottom of the adult fly. A key event in very early development is establishment of mRNA and protein gradients along the body axes. Some maternal mRNAs have protein products that diffuse through the cytoplasm to create an asymmetric distribution in the egg. Different cells in the cellular blastoderm therefore inherit different amounts of these proteins, setting the cells on different developmental paths. The products of the maternal mRNAs include transcriptional activators or repressors as well as translational repressors, all regulating the expression of other pattern regulating genes. The resulting specific patterns and sequences of gene expression therefore differ between cell lineages, ultimately orchestrating the development of each adult structure. The anterior-posterior axis in *Drosophila* is defined at least in part by the products of the *bicoid* and *nanos* genes. The *bicoid* gene product is a major anterior morphogen, and the *nanos* gene product is a major posterior morphogen. The mRNA from the *bicoid* gene is synthesized by nurse cells and deposited in the unfertilized egg near its anterior pole. Nüsslein-Volhard found that this mRNA is translated soon after fertilization, and the Bicoid protein diffuses through the cell to create, by the seventh nuclear division, a concentration gradient radiating out from the anterior pole (Fig. 28–36a). The Bicoid protein is a transcription factor that activates the expression of a number of segmentation genes; the protein contains a homeodomain (p. 1090). Bicoid is also a translational repressor that inactivates certain mRNAs. The amounts of Bicoid protein in various parts of the embryo affect the subsequent expression of a number of other genes in a threshold dependent manner. Genes are transcriptionally activated or translationally repressed only where the Bicoid protein concentration exceeds the threshold. Changes in the shape of the Bicoid concentration gradient have dramatic effects on the body pattern. Lack of Bicoid protein results in development of an embryo with two abdomens but neither head nor thorax (Fig. 28–36b); however, embryos without Bicoid will develop normally if an adequate amount of *bicoid* mRNA is injected into the egg at the appropriate end. The *nanos* gene has an analogous role, but its mRNA is deposited at the posterior end of the egg and the anterior-posterior protein gradient peaks at the posterior pole. The Nanos protein is a translational repressor.

A broader look at the effects of maternal genes reveals the outline of a developmental circuit. In addition to the *bicoid* and *nanos* mRNAs, which are deposited in the egg asymmetrically, a number of other maternal mRNAs are deposited uniformly throughout the egg cytoplasm. Three of these mRNAs encode the Pumilio, Hunchback, and Caudal proteins, all affected by *nanos* and *bicoid* (Fig. 28–37). Caudal and Pumilio are involved in development of the posterior end of the fly. Caudal is a transcriptional activator with a homeodomain; Pumilio is a translational repressor. Hunchback protein plays an important role in the development of the anterior end and is also a transcriptional regulator

of a variety of genes, in some cases a positive regulator, in other cases negative. Bicoid suppresses translation of *caudal* in the anterior and also acts as a transcriptional activator of *hunchback* in the cellular blastoderm. Because *hunchback* is expressed both from maternal mRNAs and from genes in the developing egg,

it is considered both a maternal and a segmentation gene. The result of the activities of Bicoid is an increased concentration of Hunchback at the anterior end of the egg. The Nanos and Pumilio proteins act as translational repressors of *hunchback*, suppressing synthesis of its protein near the posterior end of the egg. Pumilio does not function in the absence of the Nanos protein, and the gradient of Nanos expression confines the activity of both proteins to the posterior region. Translational repression of the *hunchback* gene leads to degradation of *hunchback* mRNA near the posterior end. However, lack of Bicoid protein in the posterior leads to expression of *caudal*. In this way, the Hunchback and Caudal proteins become asymmetrically distributed in the egg.

Segmentation Genes Gap genes, pair-rule genes, and segment polarity genes, three subclasses of segmentation genes in *Drosophila*, are activated at successive stages of embryonic development. Expression of the gap genes is generally regulated by the products of one or more maternal genes. At least some of the gap genes encode transcription factors that affect the expression of other segmentation or (later) homeotic genes. One well-characterized segmentation gene is *fushi tarazu* (*ftz*), of the pair-rule subclass. When *ftz* is deleted, the embryo develops 7 segments instead of the normal 14, each segment twice the normal width. The Fushi-tarazu protein (Ftz) is a transcriptional activator with a homeodomain. The mRNAs and proteins derived from the normal *ftz* gene accumulate in a striking pattern of seven stripes that encircle the posterior twothirds of the embryo (Fig. 28–38). The stripes demarcate the positions of segments that develop later; these segments are eliminated if *ftz* function is lost. The Ftz protein and a few similar regulatory proteins directly or indirectly regulate the expression of vast numbers of genes in the continuing developmental cascade.

Homeotic Genes Loss of homeotic genes by mutation or deletion causes the appearance of a normal appendage or body structure at an inappropriate body position. An important example is the *ultrabithorax* (*ubx*) gene. When Ubx function is lost, the first abdominal segment develops incorrectly, having the structure of the third thoracic segment. Other known homeotic mutations cause the formation of an extra set of wings, or two legs at the position in the head where the antennae are normally found (Fig. 28–39). The homeotic genes often span long regions of DNA. The *ubx* gene, for example, is 77,000 bp long. More than 73,000 bp of this gene are in introns, one of which is more than 50,000 bp long. Transcription of the *ubx* gene takes nearly an hour. The delay this imposes on *ubx* gene expression is believed to be a timing mechanism involved in the temporal regulation of subsequent steps in development. The Ubx protein is yet another transcriptional activator with a homeodomain (Fig. 28–13). Many of the principles of development outlined above apply to eukaryotes from nematodes to humans. Some of the regulatory proteins themselves are conserved. For example, the products of the homeo box containing genes *HOX 1.1* in mouse and *antennapedia* in fruit fly differ in only one amino acid residue. Of course, although the molecular regulatory mechanisms may be similar, many of the ultimate events are not conserved developmental (humans do not have wings or antennae). The discovery of structural determinants with identifiable molecular functions is the first step in understanding the molecular events underlying development. As more genes and their protein products are discovered, the biochemical side of this vast puzzle will be elucidated in increasingly rich detail.

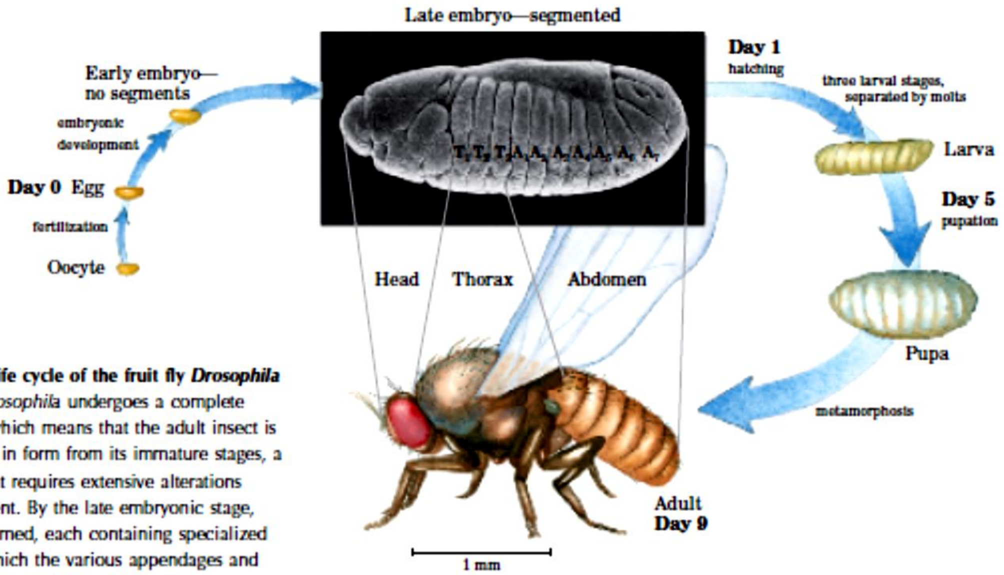


FIGURE 28–34 Life cycle of the fruit fly *Drosophila melanogaster*. *Drosophila* undergoes a complete metamorphosis, which means that the adult insect is radically different in form from its immature stages, a transformation that requires extensive alterations during development. By the late embryonic stage, segments have formed, each containing specialized structures from which the various appendages and other features of the adult fly will develop.