Introduction to molecular biology

Nucleic acids: function and structure

Nucleic acids represent a prominent category of biomolecule present in living cells. The term incorporates both DNA and RNA. DNA represents the repository of genetic information (the genome) of most life forms. RNA replaces DNA as the repository of genetic information in some viruses. In most life forms, however, RNA plays a role in mediating the conversion of genetic information stored in specific DNA sequences (genes) into polypeptides.

There are three subcategories of RNA, each playing a different role in the conversion of gene sequences into the amino acid sequence of polypeptides.

- Messenger RNA (mRNA) carries the genetic coding information from the gene to the ribosome, where the polypeptide is actually synthesized.
- Ribosomal RNA (rRNA), along with a number of proteins, forms the ribosome itself,
- Transfer RNA (tRNA) functions as an adaptor molecule, transferring a specific amino acid to a growing polypeptide chain on the ribosomal site of polypeptide synthesis.

Therefore, nucleic acids, between them all, mediate the flow of genetic information via the processes of replication, transcription and translation as outlined in what has become known as the central dogma of molecular biology (Figure below).

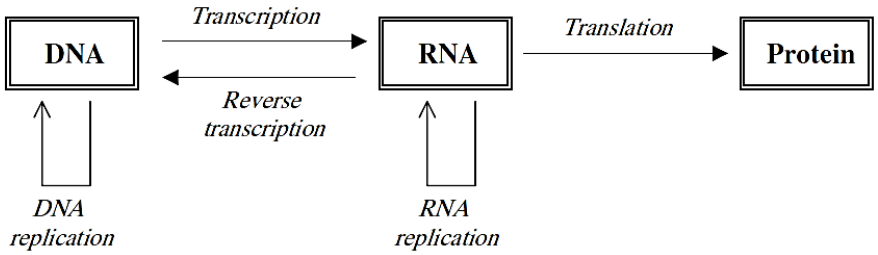


Figure Schematic representation of the so-called central dogma of molecular biology. DNA replication is essential to the transmission of genetic information from one generation to the next in most life forms (i.e. in living forms whose genomes are DNA based). RNA replication is essential to the transmission of genetic information in the context of a small number of viruses whose genomes are RNA based. Transcription describes the copying of selected DNA sequences into RNA, and translation describes the conversion of the genetic information inherent in mRNA into a polypeptide of defined amino acid sequence. The process of reverse transcription is a central feature of certain viruses (retroviruses) containing an RNA-based genome which, as part of their life cycle, infect eukaryotic cells and convert their RNA-based genomes into a DNA-based one.

Structurally, nucleic acids are polymers in which the basic recurring monomer is a **nucleotide** (i.e. nucleic acids are polynucleotides). Nucleotides themselves consist of three components:

A **phosphate group**, a **pentose** (five-carbon sugar) and a nitrogenous-containing cyclic structure known as a **base** (Figure below).

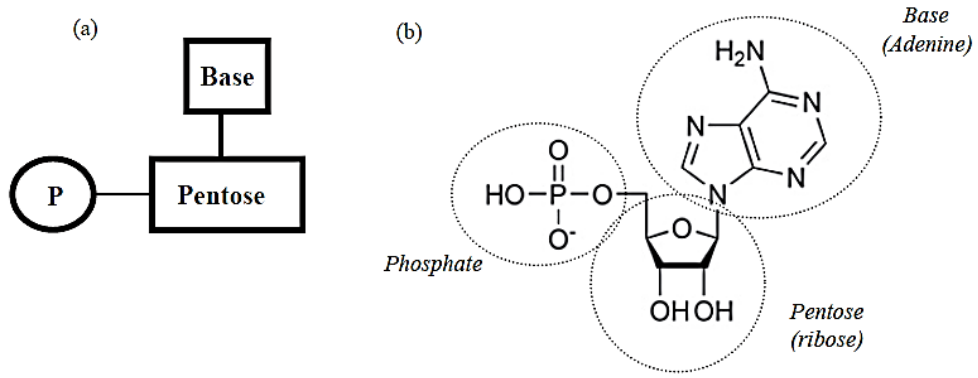


Figure:(a) The basic structure of a nucleotide. (b) The actual chemical structure of one representative nucleotide (adenylate, i.e. adenosine 5′-monophosphate)

The nucleotide sugar associated with RNA is ribose, whereas that found in DNA is deoxyribose. In total, five different bases are found in nucleic acids. They are categorized as either purines (adenine and guanine, or A and G, found in both RNA and DNA) or pyrimidines (cytosine, thymine and uracil, or C, T and U). Cytosine is found in both RNA and DNA, whereas thymine is unique to DNA and uracil is unique to RNA.

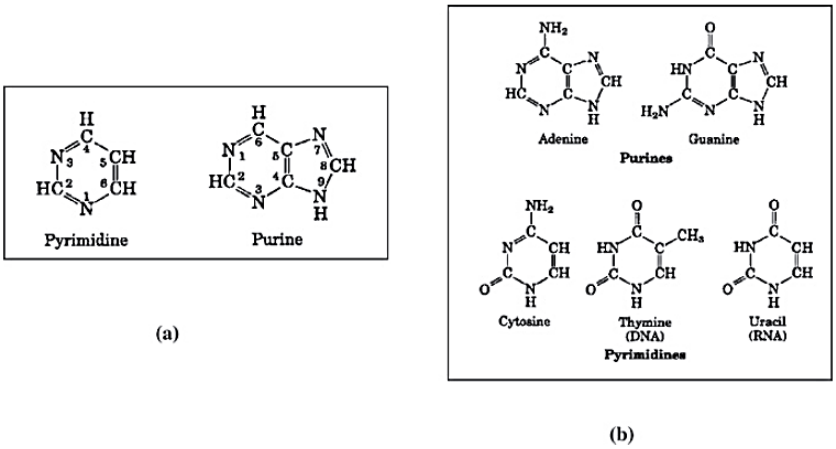


Figure : The five bases found in nucleic acids may be categorized as either pyrimidines or purines.

The DNA or RNA polymer consists of a chain of nucleotides of specific base sequence, linked via phosphodiester bonds. RNA is a single-stranded polynucleotide. DNA is a double stranded polynucleotide.

DNA molecules in a helix conformation are the predominant structures. Strands of DNA are composed of four specific building elements (shortly written as A, C, G, and T), the deoxyribonucleotides deoxyadenosine -triphosphate (dATP), deoxycytidine -triphosphate (dCTP), deoxyguanosine -triphosphate (dGTP), and deoxythymidine triphosphate (dTTP) linked by phosphodiester bonds.

The two strands in the DNA helix are held together through hydrogen bonds between the nucleotides in the various strands. The DNA strands in the helix are complementary in their

nucleotide composition: an A in one strand is always facing a T in the other one, while a C is always facing a G. Moreover, the strands in double-stranded DNA run antiparallel: the 5-P end of the one strand faces the 3-OH end of the complementary strand and the other way round.

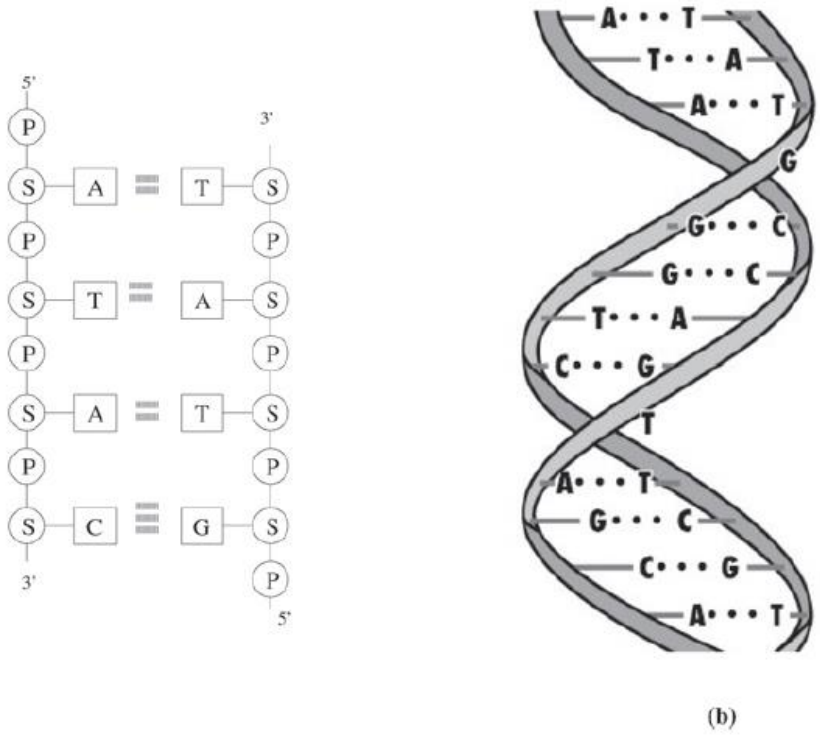


Figure :(a) DNA structure. The two complementary polynucleotide strands in DNA are antiparallel to each other in orientation (one runs 5′→3′, the other 3′→5′; The two strands are held together by hydrogen bonds between opposite complementary bases, as well as hydrophobic interactions between stacked bases (b) The double polynucleotide chain adopts a double helical structure

DNA Replication

During cell division the genetic information in a parental cell is transferred to the daughter cells by DNA replication. Essential in the very complex DNA replication process is the action of DNA polymerases.

During replication each DNA strand is copied into a complementary strand that runs antiparallel. The topological constraint for replication due to the double helix structure of the DNA is solved by unwinding of the helix, catalyzed by the enzyme helicase. In a set of biochemical events deoxyribonucleotide monomers are added one by one to the end of a growing DNA strand in a 5'to 3'direction.

DNA Replication in Prokaryotes

The prokaryotic chromosome is a circular molecule with a less extensive coiling structure than eukaryotic chromosomes. The eukaryotic chromosome is linear and highly coiled around proteins.

Replication of DNA cannot start at a random position on the strand, but only on a specific point of the beginning of replication, which in a prokaryotic DNA is called **ORI** (origin of replication). Bacteria *E. coli* contains only one DNA molecule and only one ORI, which has the size of 245 bp. An enzyme called **helicase** unwinds the DNA by breaking the hydrogen bonds between the nitrogenous base pairs. As the DNA opens up, Y-shaped structures called **replication forks** are formed. Two replication forks are formed at the origin of replication and these get extended bi-directionally as replication proceeds.

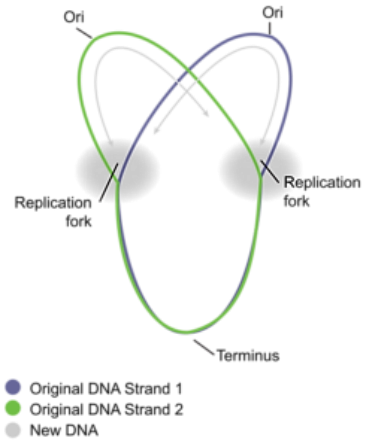


Figure: DNA replication in prokaryotes, which have one circular chromosome

After the separation of the DNA strands by helicases, to each strand of DNA immediately bind **Single-strand binding proteins**. Their task is not only stabilize the denatured part of the DNA (maintain the DNA in a single stranded form) but also prevent formation of loops called "hairpin.

RNA primase, synthesizes an RNA **primer** that is about five to ten nucleotides long and complementary to the DNA. Primers serve as the anchoring sites for of the DNA polymerase. In leading strand (with orientation 3´_ 5´) primer are applied at ORI. In case of the lagging strand (with 5´_ 3´ orientation) primase apply primers in relatively equal distances.

The most important enzymes in DNA replication are **DNA polymerases**. They ensure the prolongation (growth) of the DNA strand by assigning of new deoxyribonucleotides.

In prokaryotic cells 5 types of DNA polymerase were so far identified. The most know are DNA polymerase I, II and III.

DNA polymerase I is formed by one polypeptide chain with the molecular weight of 103 000 Da. Inside the cell of *E. coli* there are around 400 molecules of this enzyme.

The enzyme is responsible for repairing damaged DNA segments, and participates in the assembling of the lagging chain during DNA replication.

DNA polymerase II is also formed by one polypeptide chain with a molecular weight of 88 000 Da. Its *in vivo* function is not exactly defined, but it participated during the repair of DNA.

DNA polymerase III catalyzes the synthesis of the DNA strands.
The enzyme has a complicated quartiery structure, it consists of many monomers (900 000 Da).

After the adding of a primer to the beginning of the leading strand, the primase disconnects. DNA Polymerase III starts the **continual** adding of the particular deoxyribonucleotides to the 3´ end nucleotide according to the rules of complementarity

The antiparallel strand is replicated **discontinually**, therefore by parts. The primase creates the first RNA primer that connects to lagging strand, then primase detaches and moves forward, where it forms second RNA primer. Polymerase III connects to the second primer and synthesizes the complementary DNA strand from to the proceeding RNA primer. This process repeats, forming complementary strand to lagging strand, in which DNA and RNA regions alter. The fragment of DNA in this hybrid we call the **Okazaki fragment**.

RNA primers are then removed from the chain by DNA polymerase I, which at the same time fills the gap (after the primer) by adding it up with deoxyribonucleotides. Polymerase I has a sufficient amount of time, and therefore it works much slower then polymerase III. Alternating fragments are formed, which were made both by DNA Polymerase III and I. This short DNA fragments are finally connected together by **DNA ligase,** which forms phosphodiester bonds between the individual fragments of the DNA strand.

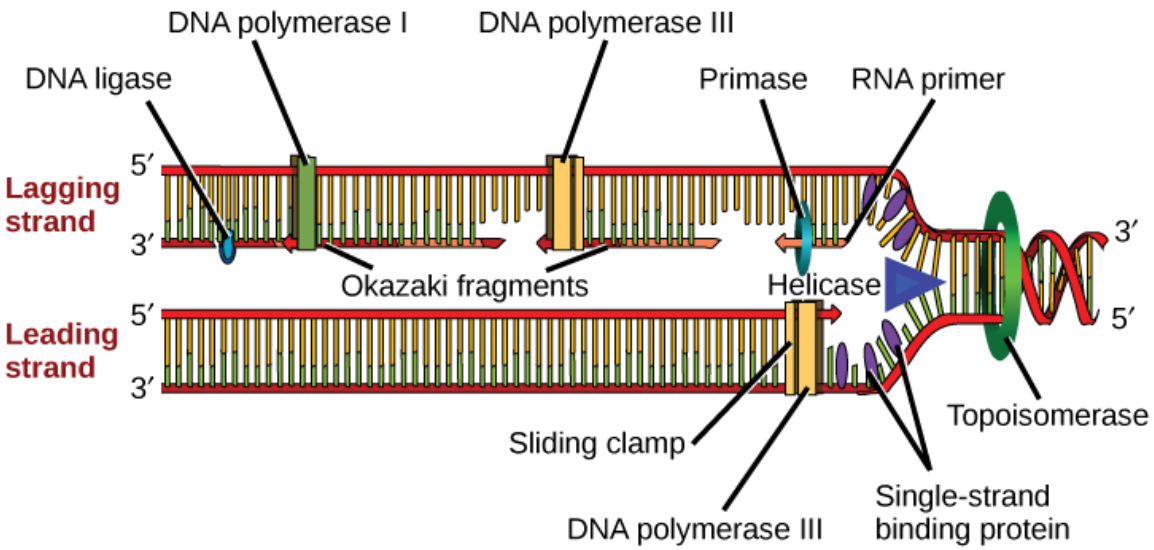


Figure: DNA replication in prokaryotes

Replication of DNA in eukaryotic cells

The essential steps of replication are the same as in prokaryotes. Starting replication is more complex in eukaryotes. At the origin of replication, a pre-replication complex is made with other initiator proteins. Other proteins are then recruited to start the replication process. The overall process is the same, although differently named enzymes fulfill the same function. The first important difference is in the polymerases. So far, 15 kinds of polymerases were isolated from the eukaryotic cells. Amongst the most important are:

• **DNA polymerase** δ catalyzes the synthesis of leading strand and finishes the synthesis of lagging strand;

• **DNA polymerase α** (DNA primase) catalyzes the synthesis of Okazaki fragments;

• **DNA polymerase β** catalyzes the synthesis of short fragments during DNA reparation;

• **DNA polymerase** γ catalyzes the formation of mitochondrial DNA in the mitochondria.

Here are the important differences between prokaryotic and eukaryotic replication:

Prokaryotic replication	**Eukaryotic replication**
semiconservative replication	semiconservative replication
single origin replication	multiple origins of replication
primer synthesized by primase	primer synthesized by subunits of DNA polymerase α
processing enzyme: DNA polymerase III	processing enzymes: DNA polymerases α and δ
removal of primer: DNA polymerase I	removal of primer: DNA polymerase β
DNA free in cytoplasm as nucleoid	chromatin structure, chromosomes, histones
circular DNA	linear DNA: problem of replication of chromosome ends → telomerase

Eukaryotic genomes are much more complex and larger in size than prokaryotic genomes. This means that there must be multiple origins of replication on the eukaryotic chromosome in order for the entire DNA to be replicated in a timely manner; humans can have up to 100,000 origins of replication.

DNA Packaging

Eukaryotic DNA is wound around proteins known as histones to form structures called **nucleosomes.** The DNA must be made accessible in order for DNA replication to proceed. The chromatin (the complex between DNA and proteins) may undergo some chemical modifications, so that the DNA may be able to slide off the histones or otherwise be accessible to the enzymes of the DNA replication machinery. Prokaryotes do not package their DNA by wrapping it around histones.

Telomere Replication

Unlike prokaryotic chromosomes, eukaryotic chromosomes are linear. The enzyme DNA pol can add nucleotides only in the 5′ to 3′ direction. In the leading strand, synthesis continues until the end of the chromosome is reached. On the lagging strand, DNA is synthesized in short stretches, each of which is initiated by a separate primer. When the replication fork reaches the end of the linear chromosome, there is no place for a primer to be made for the DNA fragment to be copied at the end of the chromosome. These ends thus remain unpaired, and over time these ends may get progressively shorter as cells continue to divide. The ends of the linear chromosomes are known as **telomeres**, which have repetitive sequences that do not code for a particular gene. These telomeres protect the genes that are located on the chromosome from getting deleted as cells continue to divide. In humans, a six base pair sequence, TTAGGG, is repeated 100 to 1000 times. The discovery of the enzyme telomerase (Figure below) helped in the understanding of how chromosome ends are maintained. The telomerase enzyme contains a catalytic part and a built-in RNA template. It attaches to the end of the chromosome, and complementary bases to the RNA template are added on the 3′ end of the DNA strand. Once the 3′ end of the lagging strand

template is sufficiently elongated, Telomerase then shifts, primase and DNA polymerase synthesize the rest of the complementary strand. Thus, the ends of the chromosomes are replicated.

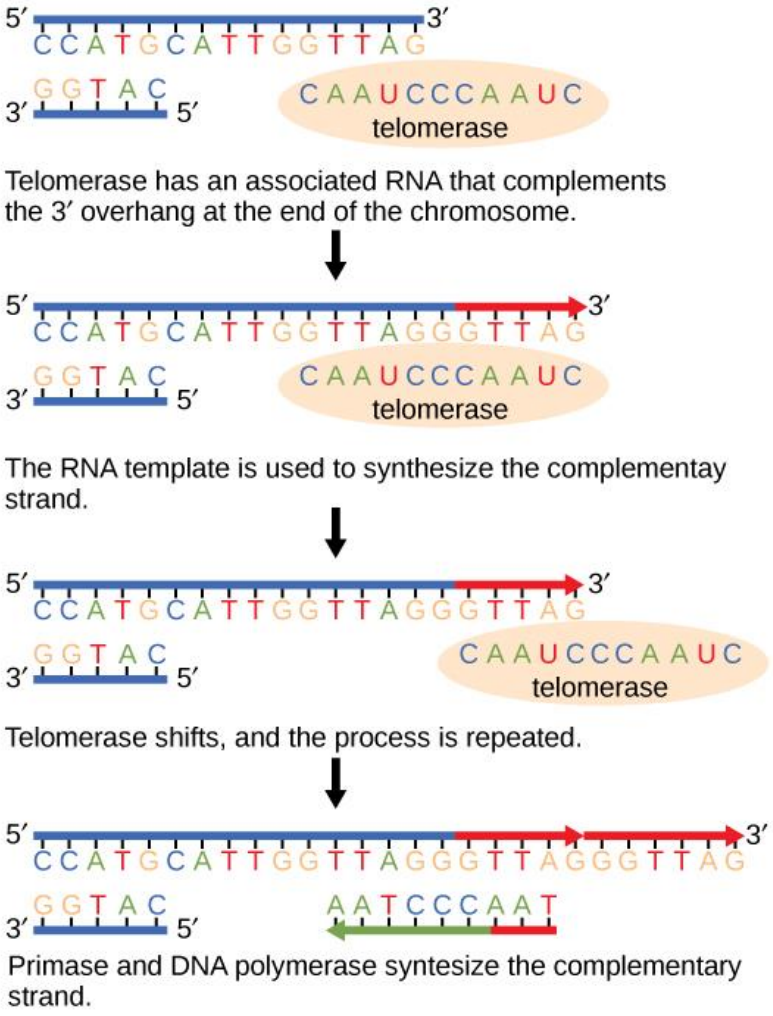


Figure: The ends of linear chromosomes are maintained by the action of the telomerase enzyme.

Gene

A gene is a basic functional and structural unit of genetic information. In meaning of molecular genetics gene is a segment of DNA codes for a protein and contain regulatory and coding parts. According to the genetic information that they carry, we can divide genes into three groups: structural genes, regulatory genes, and genes for the RNA molecules, except mRNA.

A Structural Gene is a part of the DNA chain, which codes for the primary structure of the proteins. The size of the genes is expressed in the number of base pairs (bp) in the DNA, it contains.

The structural gene consists of two parts: regulatory and coding part. The regulatory part is called the promoter. It contains important sequences (parts of base sequences, the so called boxes), for example TATA box, CAAT box etc. Their task is to bind regulatory proteins, the so called transcription factors. These proteins must bind on the mentioned sequences of the nitrogenous bases in the right order for transcription (syntheses of mRNA) to begin, which is carried out by RNA polymerase.

The prokaryotic promoter contains two important functional parts. It is the sequence in the area of the nucleotide in the -35 location (before the start codon), which is called the **GACA box** and has the following primary structure:

5´– T T **G A C A** T – 3´

The second important sequence is in the area of the nucleotide -10, which is called the **TATA box** (Pribnow box) and has a sequence:

5´– **T A T A** A T – 3´

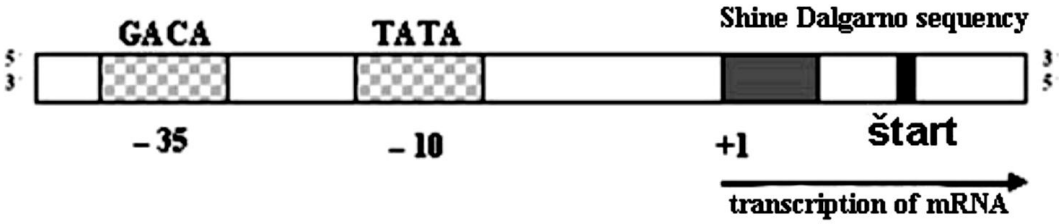


Figure: Prokaryotic promoter

The Eukaryotic promoter contains in the -34 to -26 area a **TATA box** (Hognes), to which the **TFIID** transcription factor binds.

5´– **T A T A** A A A – 3´

In the area between -75 and -80 is another important box called a **CAAT box**, to which the NF1 transcription factor binds, strengthening the promoter.

5´– G G C **C A A T** C T – 3´

The third box is called the **GC,** it is in the area - 90 and is bonded by the transcription factor SP1, which also strengthens the promoter.

5´– G G **G C** G G – 3´

In **the coding areas** the eukaryotic genes often contain noncoding parts (**introns**) in between the coding parts (**exons**). The coding part of the gene starts with the untranslated region (**UTR**) on the 5´ end, which serves as a connection of the mRNA to a small ribosomal subunit. Immediately after it follows the first exon, which starts with the so called **start triplet** (ATG). The first exon is followed by the first intron. Then proceeds another exon etc. This area therefore contains alternating exons and introns. The last exon ends with the so called **stop triplet** (TAA, TAG or TGA). Then untranslated region (**UTR**) on the 3´ end.

The regulatory gene is a part of the DNA which codes for the primary structure of the regulatory protein, which function is usually the induction or repression of the other genes expression.

The RNA coding genes are responsible for the primary structure of the ribosomal and transfer RNA and other types of smaller molecules of RNA.

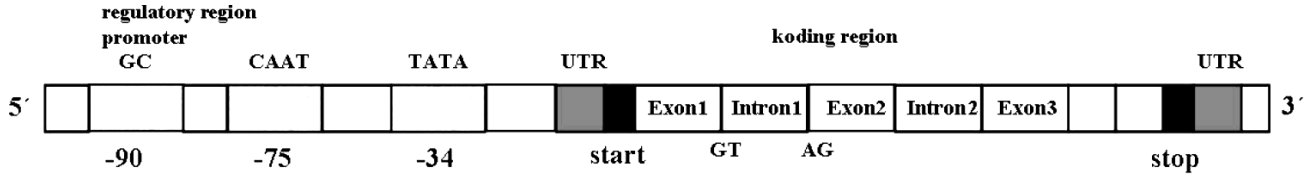


Figure: Structure of the Eukaryotic gene

In prokaryotes, genes of related function are often clustered together in **operons**, which are usually under the control of a single promoter/regulatory region. An example is the well-known ‘lac operon’. Transcribed operon mRNA thus usually contains coding sequence information for several polypeptides, and such mRNA is termed **polycistronic**. Although common in prokaryotes, the presence of polycistronic operons is infrequent in lower eukaryotes and essentially absent from higher eukaryotes, where virtually all protein-encoding genes are transcribed separately.

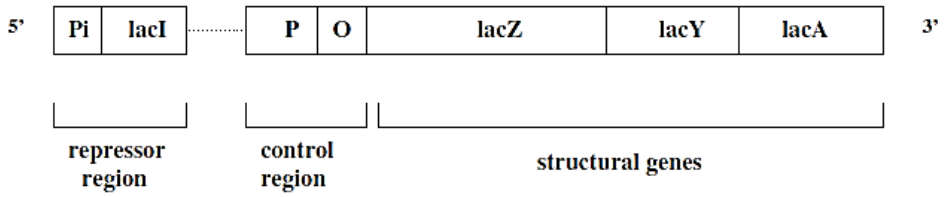


Figure: The *lac* operon houses three structural genes: *lacZ*, *lacY* and *lacA*. These code for three enzymes required for lactose metabolism (β-galactosidase, galactose permease and a transacetylase). Immediately upstream of these structural genes is a control region that houses a promoter (P) and operator (O) sequence. The operator represents a binding site for a 'repressor' protein that is in turn coded for by a repressor gene (*Pi*) found nearby but upstream of the *lac* operon. The repressor gene is in turn controlled by its own promoter. In the absence of the sugar lactose (or, more accurately, an isomer of lactose called 1,6-allolactose, which acts as an inducer) the repressor gene product is bound to the *lac* operator site, preventing transcription of the *lac* operon. In the presence of lactose (and hence the inducer), the inducer binds the repressor and the inducer–repressor complex disassociates from the operator, allowing transcription to go ahead.
mRNA is produced, but the operon also houses translational start and stop sites that allow for independent ribosomal production of the three gene products

Gene expression

Gene expression is the process by which the genetic code - the nucleotide sequence - of a gene is used to direct protein synthesis and produce the structures of the cell. Genes that code for amino acid sequences are known as 'structural genes'.

The process of gene expression involves two main stages:

Transcription: the production of messenger RNA (mRNA) by the enzyme RNA polymerase, and the processing of the resulting mRNA molecule.

Translation: the use of mRNA to direct protein synthesis, and the subsequent post-translational processing of the protein molecule.

Transcription in Prokaryotes

The principal enzyme responsible for RNA synthesis is RNA polymerase, which catalyzes the polymerization of ribonucleoside 5′-triphosphates (NTPs) as directed by a DNA template. The synthesis of RNA is similar to that of DNA, and like DNA polymerase, RNA polymerase catalyzes the growth of RNA chains always in the 5′ to 3′ direction. Unlike DNA polymerase, however, RNA polymerase does not require a preformed primer to initiate the synthesis of RNA. Instead, transcription initiates *de novo* at specific sites at the beginning of genes. The initiation process is particularly important because this is the primary step at which transcription is regulated.

E. coli RNA polymerase, like DNA polymerase, is a complex enzyme made up of multiple polypeptide chains. The intact enzyme consists of four different types of subunits, called α, β, β′, and σ (Figure.1). The σ subunit is relatively weakly bound and can be separated from the other subunits, yielding a core polymerase consisting of two α, one β, and one β′ subunits. The core polymerase is fully capable of catalyzing the polymerization of NTPs into RNA, indicating that σ is not required for the basic catalytic activity of the enzyme. However, the core polymerase does not bind specifically to the DNA sequences that signal the normal initiation of transcription; therefore, the σ subunit is required to identify the correct sites for transcription initiation. The selection of these sites is a critical element of transcription because synthesis of a functional RNA must start at the beginning of a gene.

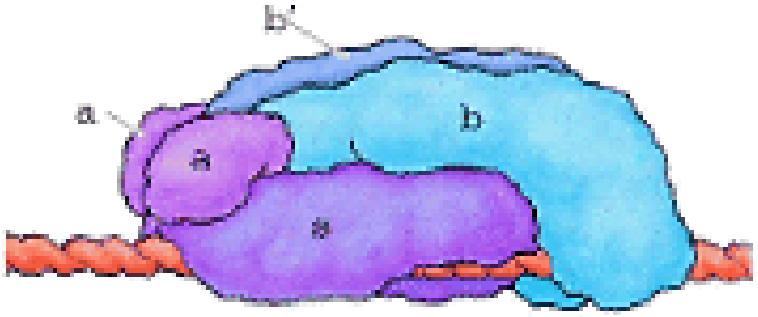


Figure 1: E. *coli* RNA polymerase. The complete enzyme consists of five subunits: two α, one β, one β′, and one σ. The σ subunit is relatively weakly bound and can be dissociated from the other four subunits, which constitute the core polymerase.

The DNA sequence to which RNA polymerase binds to initiate transcription of a gene is called the promoter. The DNA sequences involved in promoter function were first identified by comparisons of the nucleotide sequences of a series of different genes isolated from *E. coli*. These comparisons revealed that the region upstream of the transcription initiation site contains two sets of sequences that are similar in a variety of genes. These common sequences encompass six nucleotides each, and are located approximately 10 and 35 base pairs upstream of the transcription start site (Figure 2). They are called the -10 and -35 elements, denoting their position relative to the transcription initiation site, which is defined as the +1 position.

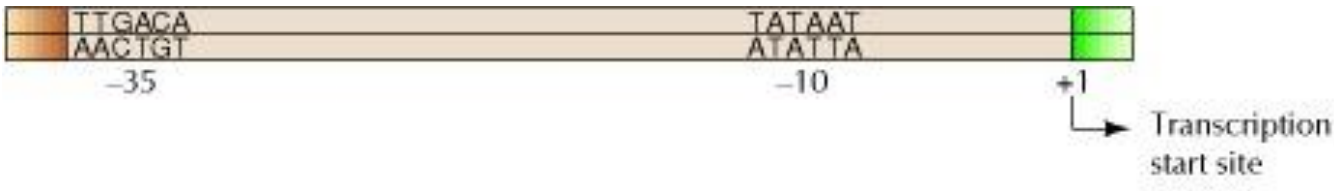


Figure 2 Sequences of *E. coli* promoter. *E. coli* promoters are characterized by two sets of sequences located 10 and 35 base pairs upstream of the transcription start site (+1).

The σ subunit binds specifically to sequences in both the -35 and -10 promoter regions, substantiating the importance of these sequences in promoter function.

In the absence of σ, RNA polymerase binds nonspecifically to DNA with low affinity. The role of σ is to direct the polymerase to promoters by binding specifically to both the -35 and -10 sequences, leading to the initiation of transcription at the beginning of a gene (Figure 3). The initial binding between the polymerase and a promoter is referred to as **a closed-promoter complex** because the DNA is not unwound. The polymerase then unwinds approximately 15 bases of DNA around the initiation site to form an open-promoter complex in which single-stranded DNA is available as a template for transcription. Transcription is initiated by the joining of two free NTPs. After addition of about the first 10 nucleotides, σ is released from the polymerase, which then leaves the promoter and moves along the template DNA to continue elongation of the growing RNA chain. As it travels, the polymerase unwinds the template DNA ahead of it and rewinds the DNA behind it, maintaining an unwound region of about 17 base pairs in the region of transcription.

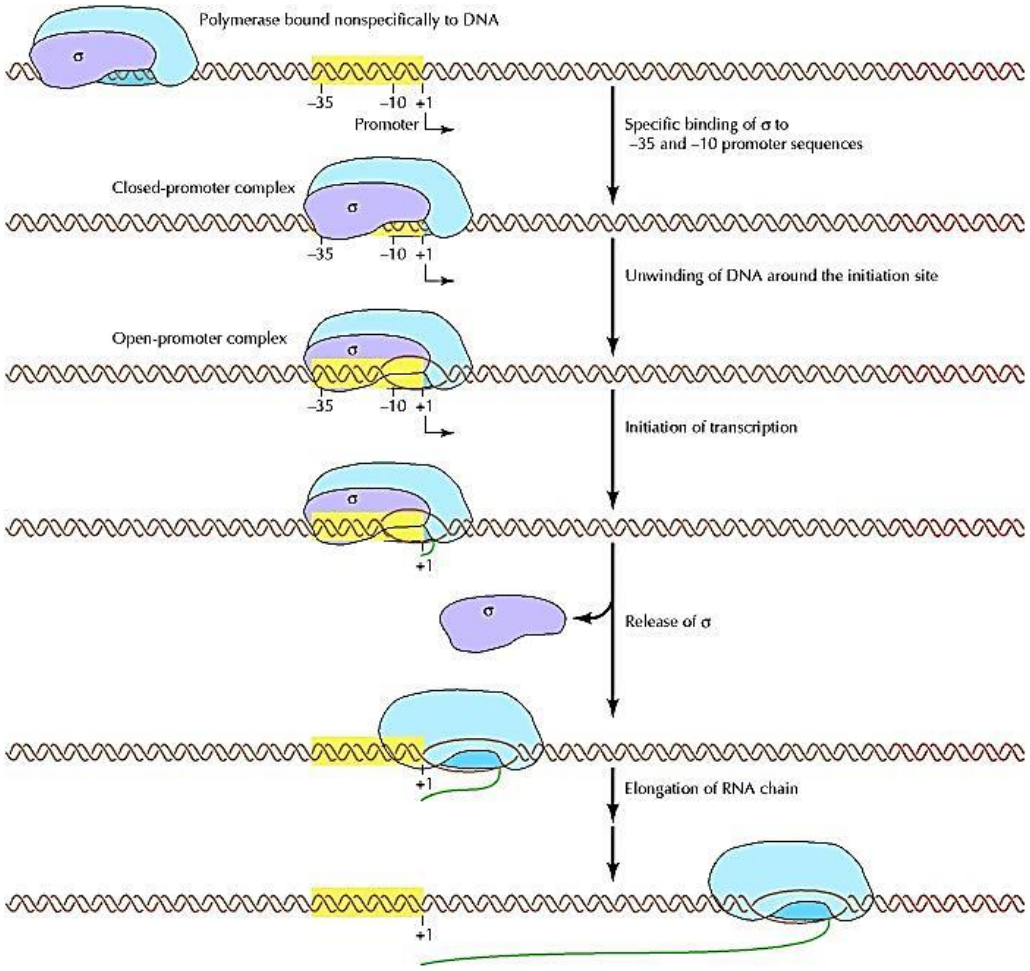


Figure 3: Transcription by *E. coli* RNA polymerase

RNA synthesis continues until the polymerase encounters a termination signal, at which point transcription stops, the RNA is released from the polymerase, and the enzyme dissociates from its DNA template. The simplest and most common type of termination signal in *E. coli* consists of a symmetrical inverted repeat of a GC-rich sequence followed by four or more A residues (Figure 4). Transcription of the GC-rich inverted repeat results in the formation of a segment of RNA that can form a stable stem-loop structure by complementary

base pairing. The formation of such a self-complementary structure in the RNA disrupts its association with the DNA template and terminates transcription. Because hydrogen bonding between A and U is weaker than that between **G and C,** the presence of A residues downstream of the inverted repeat sequences is thought to facilitate the dissociation of the RNA from its template. Other type of transcription termination called **Rho-dependent termination**, a protein factor called "Rho" is responsible for disrupting the complex involving the template strand, RNA polymerase and RNA molecule.

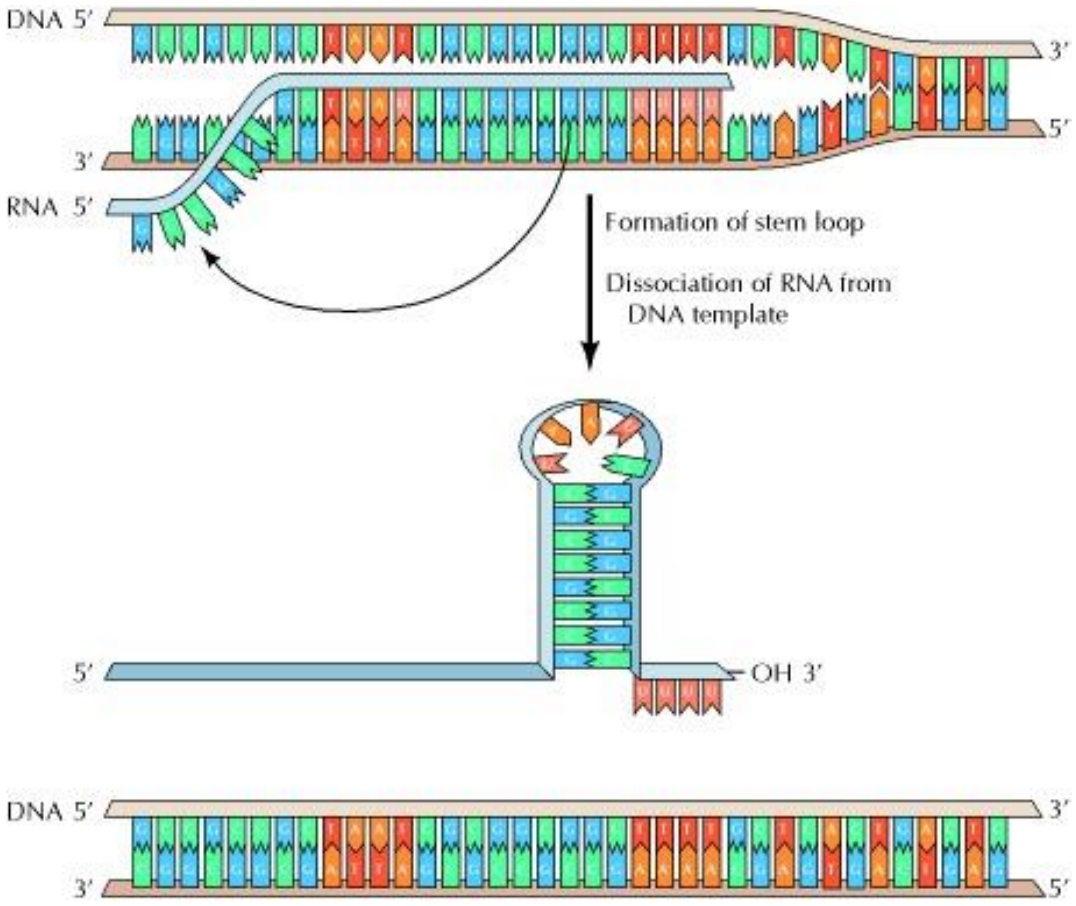


Figure 4.Transcription termination (stem-loop structure).

The termination of transcription is signaled by a GC-rich inverted repeat followed by four A residues. The inverted repeat forms a stable stem-loop structure in the RNA, causing the RNA to dissociate from the DNA template.

Eukaryotic transcription

The synthesis of RNA in eukaryotic cells is carried out by three kinds of RNA polymerases: **RNA polymerase I**, **II**, and **III** (Table below).

Enzyme	Location	Function
RNA polymerase I	Nucleolus	Transcribes genes for rRNA
RNA polymerase II	Nucleus	Transcribes structural genes and genes for certain small RNAs
RNA polymerase III	Nucleus	Transcribes genes for tRNA, 5S-rRNA and certain small RNAs

Table: Overview of eukaryotic RNA polymerases

Eukaryotic polymerases do not recognize directly their core promoter sequences. In eukaryotes, a collection of proteins called **transcription factors** mediate the binding of RNA polymerase and the initiation of transcription.

Only after attachment of certain transcription factors to the promoter, the RNA polymerase binds to it. The complete assembly of transcription factors and RNA polymerase bind-to the promoter, called **transcription initiation complex**.

The RNA Pol II is associated with six general transcription factors, designated as TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH.

TFIID consists of TBP (TATA-box binding protein) and TAFs (TBP associated factors). The role of TBP is to bind the core promoter. TAFs may assist TBP in this process. The transcription factor which catalyzes DNA melting is TFIIH. However, before TFIIH can unwind DNA, the RNA Pol II and at least five general transcription factors (have to form a **pre-initiation complex (PIC).** After pre-initiation complex [PIC] is assembled at the promoter, TFIIH can use its helicase activity to unwind DNA. Then, RNA Pol II uses ribonucleoside triphosphates (rNTPs) to synthesize a RNA transcript**.**

Eukaryotic genes also contain regulatory sequences enhancers(**Enhancers** are short nucleotide sequences in genomic DNA that have been found to influence the rate of transcription of particular target genes. enhancers may either increase or decrease transcription of their target genes. An enhancer that primarily suppresses transcription of a gene is often called a **silencer**).

When RNA polymerase II reaches a "termination sequence" (TTATTT on the DNA template and AAUAAA on the primary transcript (The polyadenylation signal), the end of transcription is signaled. This sequence is recognized by a protein complex with an endonuclease activity. It cleaves the chain in the distance of 10 to 30 nucleotides from polyadenylation sequence and the newly formed hnRNA is released from the transcription complex.

Post-transcription modification of RNA

The newly synthesized molecule of RNA is called the **primary transcript,** and must be modified. This process is called post-transcription modification or **RNA maturation**.

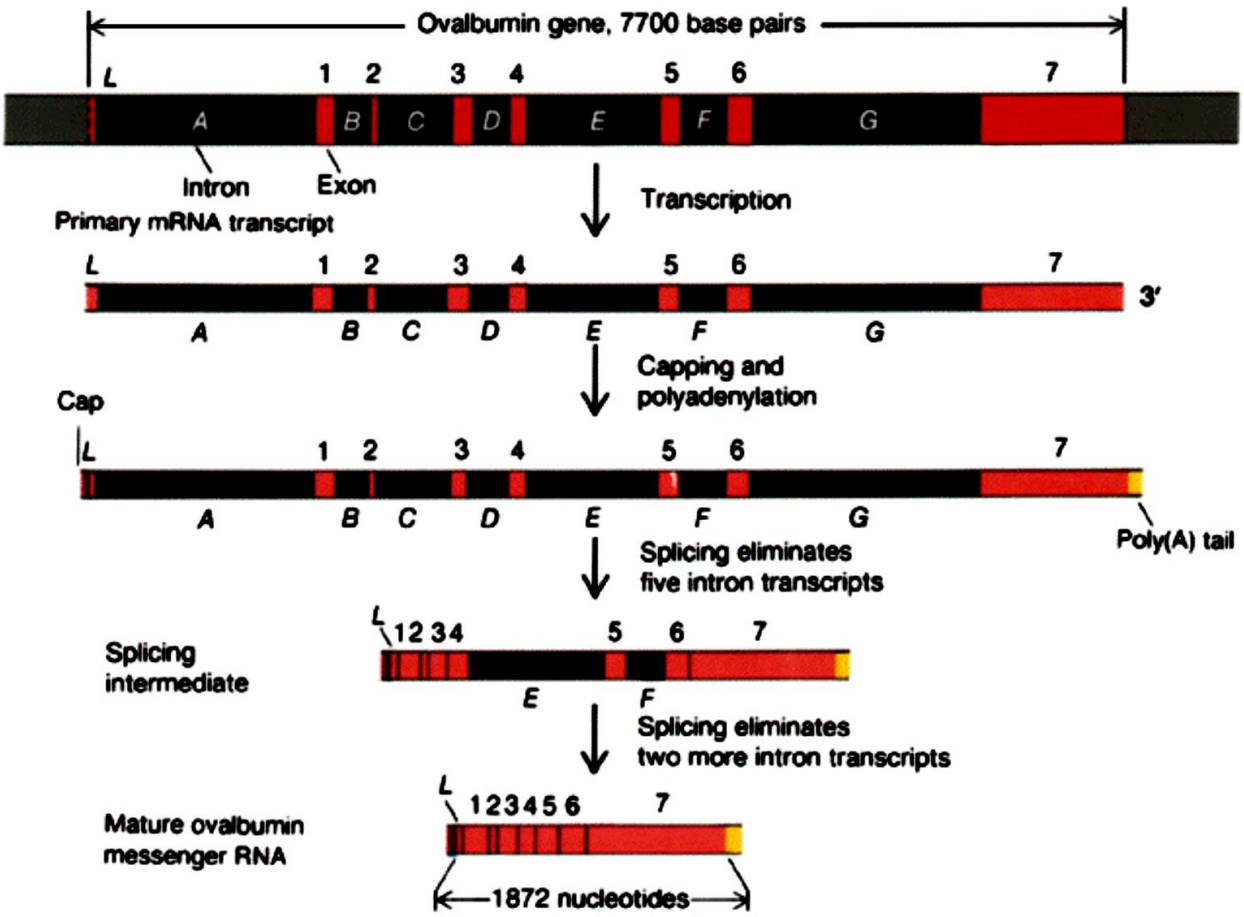
The first step is the binding of a 7-methylguanosine on to the 5´ beginning of the hnRNA chain with a non-typical 5´ – 5´ triphosphate bond, the so called **cap** is formed.

All mRNAs in eukaryotic cells have a cap. Its purpose is to protect the 5´ end of mRNA against the influence of exonucleases and to help during the binding of the mRNA to the small subunit of ribosome.

Formation of a **polyA tail** on the 3´ end of the mRNA. It is a sequence made up of 100 to 250 remnants of adenine assembled by polyA- polymerase. The polyA-sequence protects the mRNA from the 3´ end against exonucleases, during its transport to the ribosomes.

Splicing, during which RNA copy of non-coding gene sequences –**introns** – are cut out from the precursor molecule. In eukaryotic cells, splicing is catalyzed by large enzyme complexes (**spliceosomes**). They are able (in collaboration with small RNAs) to recognize an intron, cut it out, and connect together the coding sequences – **exons**.

After these modifications a functional mRNA molecule is created, and (after association with regulatory proteins) is transferred from the nucleus to the cytoplasm, where it binds to small subunit of ribosome.


Ovalbumin gene, 7700 base pairs
L
1 2 3 4 5 6 7
A B C D E F G
Intron
Exon
Primary mRNA transcript
Transcription
3'
Capping and polyadenylation
Cap
Poly(A) tail
Splicing eliminates five intron transcripts
Splicing intermediate
1 2 3 4 5 6 7
E F
Splicing eliminates two more intron transcripts
Mature ovalbumin messenger RNA
1 2 3 4 5 6 7
1872 nucleotides