ENTRI

Unit VI

Cell Biology, Molecular Biology and Bioinformatics

Module 2. Molecular biology

Nucleic acids: DNA – structure and Conformations of DNA

Nucleic acids are the organic materials present in all organisms in the form of DNA or RNA. These nucleic acids are formed by the combination of nitrogenous bases, sugar molecules and phosphate groups that are linked by different bonds in a series of sequences. The DNA structure defines the basic genetic makeup of our body. In fact, it defines the genetic makeup of nearly all life on earth.

This is also true for viruses, as most of these entities have either RNA or DNA as their genetic material. For instance, some viruses may have RNA as their genetic material, while others have DNA as the genetic material. The Human Immunodeficiency Virus (HIV) contains RNA, which is then converted into DNA after attaching itself to the host cell.

Apart from being responsible for the inheritance of genetic information in all living beings, DNA also plays a crucial role in the production of proteins. Nuclear DNA is the DNA contained within the nucleus of every cell in a eukaryotic organism. It codes for the majority of the organism's genomes while the mitochondrial DNA and plastid DNA handles the rest.

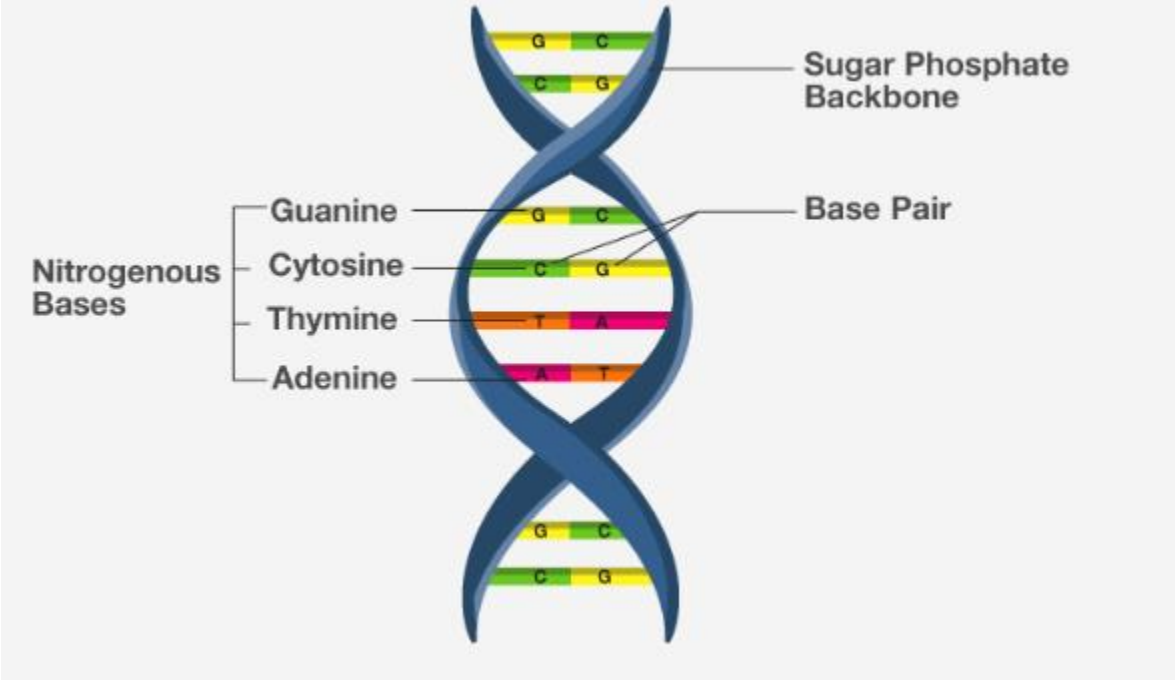
The DNA present in the mitochondria of the cell is termed mitochondrial DNA. It is inherited from the mother to the child. In humans, there are approximately 16,000 base pairs of mitochondrial DNA. Similarly, plastids have their own DNA, and they play an essential role in photosynthesis.

DNA Types

There are **three different DNA types:**

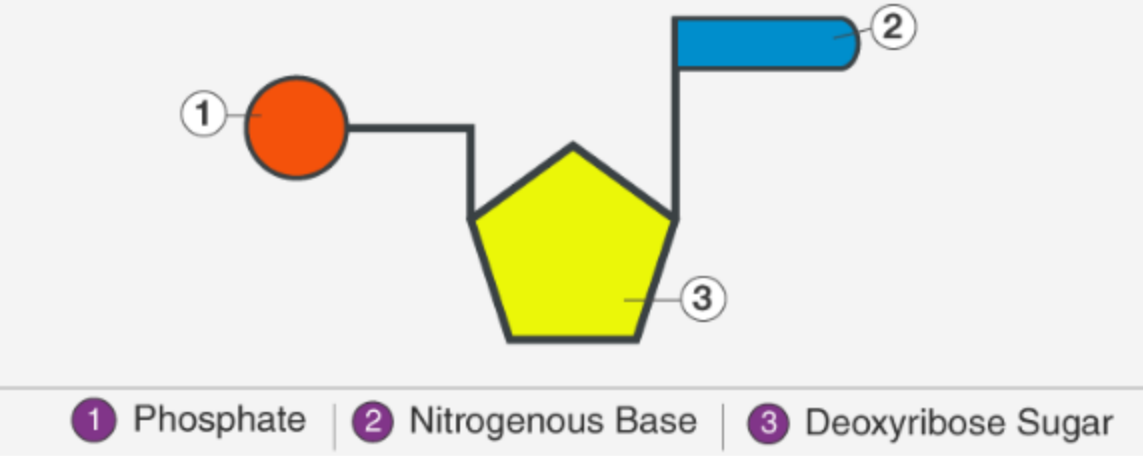
- **A-DNA:** It is a right-handed double helix similar to the B-DNA form. Dehydrated DNA takes an A form that protects the DNA during extreme conditions such as desiccation. Protein binding also removes the solvent from DNA, and the DNA takes an A form.
- **B-DNA:** This is the most common DNA conformation and is a right-handed helix. The majority of DNA has a B type conformation under normal physiological conditions.
- **Z-DNA:** Z-DNA is a left-handed DNA where the double helix winds to the left in a zig-zag pattern. It was discovered by Andres Wang and Alexander Rich. It is found ahead of the start site of a gene and hence, is believed to play some role in gene regulation.

DNA Structure

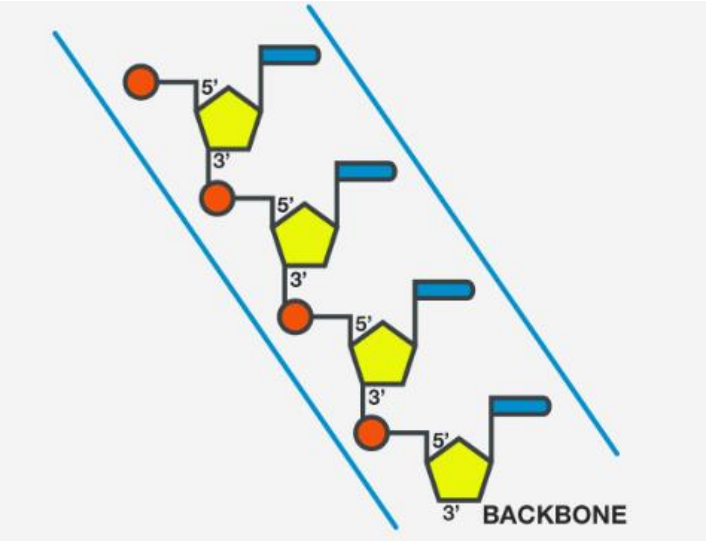


- The DNA structure can be thought of as a **twisted ladder.**
- This structure is described as a double-helix.
- It is a nucleic acid, and all nucleic acids are made up of nucleotides.

- The DNA molecule is composed of units called nucleotides, and each nucleotide is composed of three different components such as sugar, phosphate groups and nitrogen bases.
- The basic building blocks of DNA are nucleotides, which are composed of a sugar group, a phosphate group, and a nitrogen base.
- The sugar and phosphate groups link the nucleotides together to form each strand of DNA.
- **Adenine (A), Thymine (T), Guanine (G) and Cytosine (C)** are four types of nitrogen bases.
- These 4 Nitrogenous bases pair together in the following way: A with T, and C with G.
- These base pairs are essential for the DNA's double helix structure, which resembles a twisted ladder.
- The order of the nitrogenous bases determines the genetic code or the DNA's instructions.



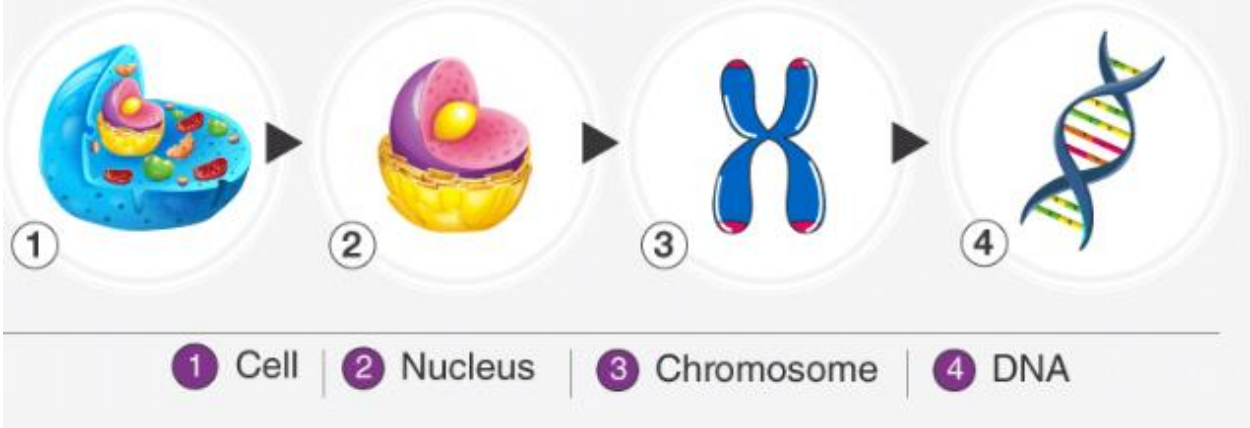
Among the three components of DNA structure, sugar is the one which forms the backbone of the DNA molecule. It is also called deoxyribose. The nitrogenous bases of the opposite strands form hydrogen bonds, forming a ladder-like structure.



The DNA molecule consists of 4 nitrogen bases, namely adenine (A), thymine (T), cytosine (C) and Guanine (G), which ultimately form the structure of a nucleotide. The A and G are purines, and the C and T are pyrimidines.

The two strands of DNA run in opposite directions. These strands are held together by the hydrogen bond that is present between the two complementary bases. The strands are helically twisted, where each strand forms a right-handed coil, and ten nucleotides make up a single turn.

The pitch of each helix is 3.4 nm. Hence, the distance between two consecutive base pairs (i.e., hydrogen-bonded bases of the opposite strands) is 0.34 nm.



The DNA coils up, forming chromosomes, and each chromosome has a single molecule of DNA in it. Overall, human beings have around twenty-three pairs of chromosomes in the nucleus of cells. DNA also plays an essential role in the process of cell division.

Conformations of DNA

Commonly occurring structural conformations of DNA are – A-DNA, B-DNA and Z-DNA.

The key difference between form B DNA and Z DNA is that the B-DNA is right-handed, while the Z-DNA is left-handed.

B DNA

- Commonly occurring DNA form in normal physiological conditions, this form of DNA is a right-handed double helix
- The two strands of this DNA run in two different directions
- They show an asymmetrical structure, with the alternate presence of major and minor grooves. It is a result of the glycosidic bonds of a base pair not being diametrically opposed to one another
- Between the adjacent deoxyribonucleotides, there is a distance of 0.34 nm and each turn comprises 10.5 base pairs of length 3.4 nm
- The helical width of B-DNA is 2 nm and its backbone comprises sugar phosphates associated continuously through phosphodiester bonds. The core comprises nitrogenous bases.

Z DNA

- Structurally differing, this form of DNA is a left-handed double helix
- The helical width of Z-DNA is 1.8 nm, making it the narrowest compared to the other DNA conformations
- Its distinguishing factor is its backbone appearing as though a zigzag
- Each turns comprises 12 base pairs, 4.56 nm long.

- Two adjacent deoxyribonucleotides are 0.37 nm apart with the presence of hydrogen bonds between two strands.

DNA replication in prokaryotes and eukaryotes

DNA Replication in Prokaryotes

Centring on the general principle of DNA replication, the prokaryotic DNA replication in prokaryotic cells takes place just before a cell divides in an organism and ensures both daughter cells receive an exact copy of the parent's genetic material. The process uses the semiconservative model of replication which results in a double-stranded DNA with one parental and one daughter strand.

The **Steps of Prokaryotic DNA Replication** are as follows:

- The DNA replication process is bi-directional begins at a spot on the DNA molecule called the origin of replication.
- At this spot, enzymes unwind the double helix structure of the DNA which makes its components accessible for replication.
- The helix is unwound by the helicase enzyme to form a pair of replication forks, and the unwound helix is stabilised by SSB proteins and DNA isomerases.
- Primase forms 10 base RNA primers which initiate the synthesis of the leading and the lagging strand.
- The leading continues to synthesise in the 5′ to 3′ direction by DNAP III (DNA Polymerase III)
- The lagging strand is also synthesised in the 5′ to 3′ direction but it is discontinued through the formation of Okazaki fragments.
- DNA polymerase I removes the 10 base RNA primers and replaces the gap with deoxynucleotides.

- Then DNA ligase seals the breaks between Okazaki fragments as well as around the primers to form continuous strands.
- The entire process of replication takes place in the cell cytoplasm.

DNA Replication in Eukaryotes

The eukaryotic DNA replication takes place in the cell nucleus and only occurs in the S phase at many chromosomal origins. Similar to prokaryotic DNA replication, eukaryotic cells also use the semi-conservative process of replication but there are multiple origins of replication.

The **Steps of the Eukaryotic DNA Replication** are as follow:

- The replication process starts in a chromosome at multiple origins, with one origin being at 30-300 kb of DNA depending on the tissue and species.
- A replication bubble of two forks forms at each origin. The DNA replicated under the control of a single origin is called a replicon. The synthesis proceeds until all bubbles merge together.
- The process starts with the unwinding of DNA with the help of enzymes, which makes its components accessible for replication.
- The unwound helix forms a pair of replication forks and is stabilised by DNA topoisomerases and SSB proteins.
- The RNA primers required for the process are made by DNA polymerases α which initiates the synthesis of the lagging strand and makes the first primer. It then extends it with a short region of DNA.
- The Okazaki fragments and the leading strand are synthesised by DNA polymerase δ.
- The leading strand is synthesised continuously whilst the lagging strand is synthesised discontinuously. Both strands are synthesised in the 5′to 3′ direction.
- At completion, DNA ligase seals the breaks around the primers and between the Okazaki fragments.

Although there are some similarities between DNA replication in prokaryotes and eukaryotes, the differences are many. Here we will discuss the differences between prokaryotes' and eukaryotes' DNA replication process.

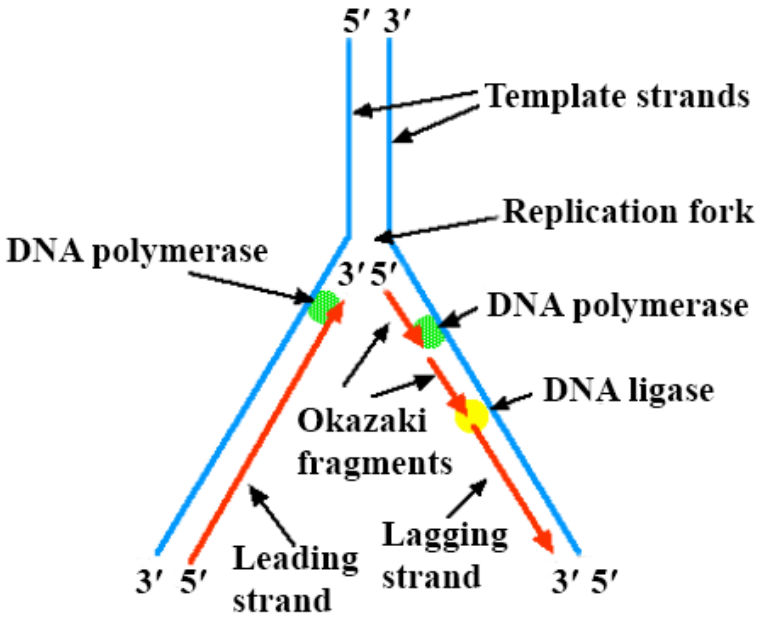
Replication machinery, mechanisms and repair

Replication machinery, mechanism

The **process of DNA duplication** is called DNA replication. Replication follows several steps that involve multiple proteins called replication enzymes and RNA. In eukaryotic cells, such as animal cells and plant cells, DNA replication occurs in the S phase of interphase during the cell cycle. The process of DNA replication is vital for cell growth, repair, and reproduction in organisms.

Preparation for Replication

Step 1: Replication Fork Formation



- Before DNA can be replicated, the double stranded molecule must be "unzipped" into two single strands.

- DNA has four bases called adenine (A), thymine (T), cytosine (C) and guanine (G) that form pairs between the two strands.
- Adenine only pairs with thymine and cytosine only binds with guanine.
- In order to unwind DNA, these interactions between base pairs must be broken.
- This is performed by an enzyme known as DNA helicase. DNA helicase disrupts the hydrogen bonding between base pairs to separate the strands into a Y shape known as the replication fork. This area will be the template for replication to begin.
- DNA is directional in both strands, signified by a 5' and 3' end.
- This notation signifies which side group is attached the DNA backbone. The 5' end has a phosphate (P) group attached, while the 3' end has a hydroxyl (OH) group attached.
- This directionality is important for replication as it only progresses in the 5' to 3' direction.
- However, the replication fork is bi-directional; one strand is oriented in the 3' to 5' direction (leading strand) while the other is oriented 5' to 3' (lagging strand).
- The two sides are therefore replicated with two different processes to accommodate the directional difference.

Replication Begins

Step 2: Primer Binding

- The leading strand is the simplest to replicate.
- Once the DNA strands have been separated, a short piece of RNA called a primer binds to the 3' end of the strand.
- The primer always binds as the starting point for replication. Primers are generated by the enzyme DNA primase.

DNA Elongation

Step 3: Elongation

- Enzymes known as DNA polymerases are responsible creating the new strand by a process called elongation.
- There are five different known types of DNA polymerases in bacteria and human cells. In bacteria such as E. coli, polymerase III is the main replication enzyme, while polymerase I, II, IV and V are responsible for error checking and repair.
- DNA polymerase III binds to the strand at the site of the primer and begins adding new base pairs complementary to the strand during replication.
- In eukaryotic cells, polymerases alpha, delta, and epsilon are the primary polymerases involved in DNA replication.
- Because replication proceeds in the 5' to 3' direction on the leading strand, the newly formed strand is continuous.
- The lagging strand begins replication by binding with multiple **primers**.
- Each primer is only several bases apart.
- DNA polymerase then adds pieces of DNA, called **Okazaki fragments**, to the strand between primers.
- This process of replication is discontinuous as the newly created fragments are disjointed.

Step 4: Termination

- Once both the continuous and discontinuous strands are formed, an enzyme called **exonuclease** removes all RNA primers from the original strands.
- These primers are then replaced with appropriate bases.
- Another exonuclease “proofreads” the newly formed DNA to check, remove and replace any errors.
- Another enzyme called **DNA ligase** joins Okazaki fragments together forming a single unified strand.

- The ends of the linear DNA present a problem as DNA polymerase can only add nucleotides in the 5′ to 3′ direction.
- The ends of the parent strands consist of repeated DNA sequences called **telomeres.**
- Telomeres act as protective caps at the end of chromosomes to prevent nearby chromosomes from fusing.
- A special type of DNA polymerase enzyme called **telomerase** catalyzes the synthesis of telomere sequences at the ends of the DNA.
- Once completed, the parent strand and its complementary DNA strand coils into the familiar double helix shape.
- In the end, replication produces two DNA molecules, each with one strand from the parent molecule and one new strand.

Enzyme involved in replication

DNA replication would not occur without enzymes that catalyze various steps in the process. Enzymes that participate in the eukaryotic DNA replication process include:

- **DNA helicase** - unwinds and separates double stranded DNA as it moves along the DNA. It forms the replication fork by breaking hydrogen bonds between nucleotide pairs in DNA.
- **DNA primase** - a type of RNA polymerase that generates RNA primers. Primers are short RNA molecules that act as templates for the starting point of DNA replication.
- **DNA polymerases** - synthesize new DNA molecules by adding nucleotides to leading and lagging DNA strands.
- **Topoisomerase or DNA Gyrase** - unwinds and rewinds DNA strands to prevent the DNA from becoming tangled or supercoiled.
- **Exonucleases** - group of enzymes that remove nucleotide bases from the end of a DNA chain.
- **DNA ligase** - joins DNA fragments together by forming phosphodiester bonds between nucleotides.

DNA damage and repair

Damage to cellular DNA is involved in mutagenesis and the development of cancer. The DNA in a human cell undergoes several thousand to a million damaging events per day, generated by both external (exogenous) and internal metabolic (endogenous) processes. Changes to the cellular genome can generate errors in the transcription of DNA and ensuing translation into proteins necessary for signaling and cellular function. Genomic mutations can also be carried over into daughter generations of cells if the mutation is not repaired prior to mitosis. Once cells lose their ability to effectively repair damaged DNA, there are three possible responses.

- The cell may become senescent, i.e., irreversibly dormant. In 2005, multiple laboratories reported that senescence could occur in cancer cells in vivo as well as in vitro, stopping mitosis and preventing the cell from evolving further.
- The cell may become apoptotic. Sufficient DNA damage may trigger an apoptotic signaling cascade, forcing the cell into programmed cell death.
- The cell may become malignant, i.e., develop immortal characteristics and begin uncontrolled division.

To compensate for the degree and types of DNA damage that occur, cells have developed multiple repair processes including mismatch, base excision, and nucleotide excision repair mechanisms, with little process redundancy. Cells may have evolved to proceed into apoptosis or senescence if overwhelming damage occurs rather than expend energy to effectively repair the damage. The rate at which a cell is able to make repairs is contingent on factors including cell type and cell age.

Source of DNA Damage

The four types of factors that cause DNA damage are:

1. **Hydrolysis**
2. **Deamination**
3. **Alkylation and**
4. **Oxidation**

1. Hydrolysis:

- DNA consists of long strands of sugar molecules called deoxyribose that are linked together by phosphate groups.
- Each sugar molecule carries one of the four natural DNA bases: adenine, guanine, cytosine, or thymine (A, G, C, or T).
- The chemical bond between a DNA base and its respective deoxyribose, although relatively stable, is nonetheless subject to chance cleavage by a water molecule in a process known as spontaneous hydrolysis.
- Loss of the "purine" bases (guanine and adenine) is referred to as depurination, whereas loss of the "pyrimidine" bases (cytosine and thymine) is called depyrimidination. In mammalian cells, it is estimated that depurination occurs at the rate of about 10,000 purine bases lost per cell generation.
- The rate of depyrimidination is considerably slower, resulting in the loss of about 500 pyrimidine bases per cell generation.
- The baseless sugars that result from these processes are commonly referred to as AP-sites (apurinic/apyrimidinic).
- They are potentially lethal to the cell, as they act to block the progress of DNA replication, but are efficiently repaired in a series of enzyme-catalyzed reactions collectively referred to as the base excision repair (BER) pathway.
- In fact, AP-sites are intentionally created during the course of BER.

2. Deamination:

- The bases that make up DNA are also vulnerable to modification of their chemical structure.
- One form of modification, called spontaneous deamination, is the loss of an amino group (-NH2). For example, cytosine (C), which is paired with guanine (G) in normal, double-stranded DNA, has an amino group attached to the fourth carbon (C4) of the base.
- When that amino group is lost, either through spontaneous, chemical, or enzymatic hydrolysis, a uracil (U) base is formed, and a normal C-G DNA base pair is changed to a pre-mutagenic U-G base pair (uracil is not a normal part of DNA).
- The U-G base pair is called pre-mutagenic because if it is not repaired before DNA replication, a mutation will result.
- During DNA replication, the DNA strands separate, and each strand is copied by a DNA polymerase protein complex.
- On one strand, the uracil (U) will pair with a new adenine (A), while on the other strand the guanine (G) will pair with a new cytosine(C).
- Thus, one DNA double-strand contains a normal C-G base pair, but the other double-strand has a mutant U-A base pair.
- This process is called mutation fixation, and the mutation of the G to an A is said to be fixed (meaning “fixed in place,” not “repaired”).
- In other words, the cell now accepts the new mutant base pair as normal.
- It is estimated that approximately 400 cytosine deamination events per genome occur every day.
- Clearly, it is very important for the cell to repair DNA damage before DNA replication commences, in order to avoid mutation fixation.
- One cause of normal human aging is the gradual accumulation over time of mutations in our cellular DNA.

3. Alkylation:

- Another type of base modification is alkylation.

- Alkylation occurs when a reactive mutagen transfers an alkyl group (typically a small hydrocarbon side chain such as a methyl or ethyl group, denoted as -CH3 and –C2H5, respectively) to a DNA base.
- The nitrogen atoms of the purine bases (N3 of adenine and N7 of guanine) and the oxygen atom of guanine (O6) are particularly susceptible to alkylation in the form of methylation.
- Methylation of DNA bases can occur through the action of exogenous (environmental) and endogenous (intracellular) agents. For example, exogenous chemicals such as dimethyl sulfate, used in many industrial processes and formed during the combustion of sulfur-containing fossil and N-methyl-N-nitrosoamine, a component of tobacco smoke, are powerful alkylating agents.
- These chemicals are known to greatly elevate mutation rates in cultured cells and cause cancer in rodents.
- Inside every cell is a small molecule known as S-adenosylmethionine or “SAM” SAM, which is required for normal cellular metabolism, is an endogenous methyl donor.
- The function of SAM is to provide an activated methyl group for virtually every normal biological methylation reaction.
- SAM helps to make important molecules such as adrenaline, a hormone secreted in times of stress; creatine, which provides energy for muscle contraction; and phosphatidylcholine, an important component of cell membranes.
- However, SAM can also methylate inappropriate targets, such as adenine and guanine. Such endogenous DNA-alkylation damage must be continually repaired; otherwise, mutation fixation can occur.

4. Oxidation:

- Oxidative damage to DNA bases occurs when an oxygen atom binds to a carbon atom in the DNA base.
- High-energy radiation, like X-rays and gamma radiation, causes exogenous oxidative DNA base damage by interacting with water molecules to create highly reactive oxygen species, which then attack DNA bases at susceptible carbon atoms.
- Oxidative base damage is also endogenously produced by reactive oxygen species released during normal respiration in mitochondria, the cell's "energy factories."
- Humans enjoy a long life span; thus, it would seem that healthy, DNA repair-proficient cells could correct most of the naturally occurring endogenous DNA damage.
- Unfortunately, when levels of endogenous DNA damage are high, which might occur as the result of an inactivating mutation in a DNA repair gene, or when we are exposed to harmful exogenous agents like radiation or dangerous chemicals, the cell's DNA repair systems become overwhelmed.
- Lack of DNA repair results in a high mutation rate, which in turn may lead to cell death, cancer, and other diseases.
- Also, if the level of DNA repair activity declines with age, then the mutational burden of the cell will increase as we grow older.

Methods for Repairing DNA Damages

(a) Direct Reversal of Base Damage:

- Spontaneous addition of a methyl group (CH3-) (an example of alkylation) to Cs followed by domination to a T is the most frequent cause of point mutations in humans.
- Fortunately, most of these changes are repaired by enzymes which are known as glycosylases that remove the mismatched T

restoring the correct C. The DNA backbone need not be broken for this.

- DNA used to get damaged by alkylation in cancer chemotherapy ("chemo") due to some of the drugs used also damage DNA by alkylation. Some of the methyl groups can be removed by a protein encoded by our MGMT gene. The removal of each methyl group requires another molecule of protein as the protein can only do it once.
- Each of the myriad types of chemical alterations to bases requires its own mechanism to correct. The cell needs are more general mechanisms capable of correcting all sorts of chemical damage with a limited toolbox. The mechanisms of excision repair this requirement.

b) Excision Repair:

- In this process the damaged base or bases are removed and then replaced with the correct ones in a localized burst of DNA synthesis.
- There are **three modes of excision repair**, each of which employs specialized sets of enzymes namely.
    - **Base Excision Repair (BER)**
    - **Nucleotide Excision Repair (NER)**
    - **Mismatch Repair (MMR)**

Base Excision Repair (BER)

The **steps** and by players of BER are:

(i) Removal of the damaged base by a DNA glycosylase.
(ii) Removal of its deoxyribose phosphate in the backbone which produces a gap.

(iii) Replacement with the correct nucleotide. This relies on DNA polymerase beta, one of at least 11 DNA polymerases encoded by our genes.

(iv) Ligation of the break in the strand. Two enzymes are known that can do this; both require ATP to provide the needed energy.

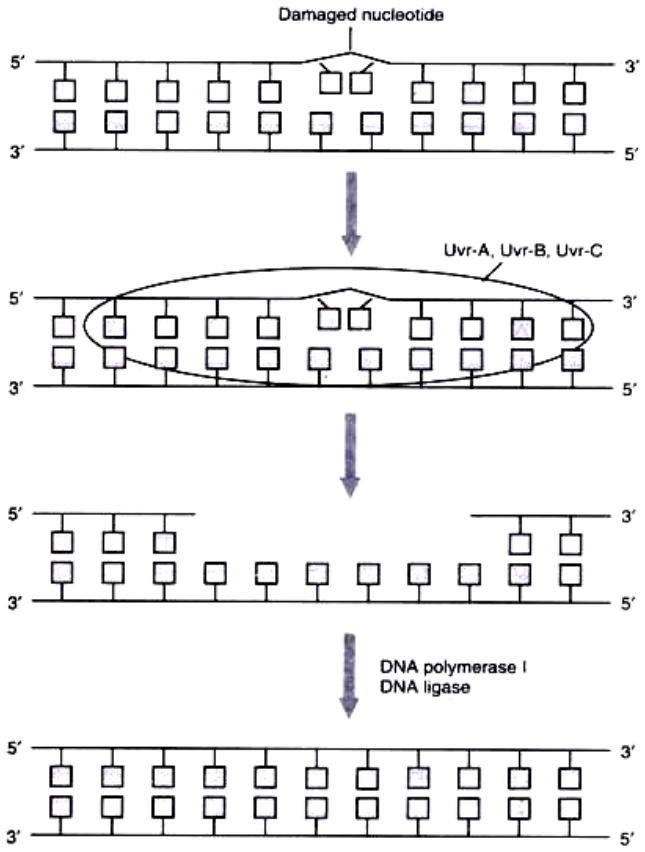


Fig. 4.5. Nucleotide Excision Repair (NER) steps

Nucleotide Excision Repair (NER)

Nucleotide Excision Repair (NER) differs from BER in several ways. It uses different enzymes. NER removes a large "patch" around the damage.

The **steps** and key players of NER are:

(i) The damage one or more protein factors recognize. These assemble at the location.
(ii) The DNA is unwound which produces a "bubble". The enzyme system that does this is Transcription Factor IIH, TFIIH, (which also functions in normal transcription).
(iii) Cuts are made on both the 3′ side and the 5′ side of the damaged area so the tract containing the damage can be removed.
(iv) A fresh burst of DNA synthesis – using the intact (opposite) strand as a template – fills in the correct nucleotides. The DNA polymerases responsible are designated polymerase delta and epsilon.
(v) A DNA ligase covalently inserts the fresh piece into the backbone.

Mismatch Repair (MMR)

Mismatch Repair (MMR) deals with correcting mismatches of the normal bases.

1. It can enlist the aid of enzymes involved in both base-excision repair (BER) and nucleotide-excision repair (NER) as well as using enzymes specialized for this function.
2. Recognition of a mismatch requires several different proteins including one encoded by MSH2.
3. Cutting the mismatch out also requires several proteins, including one encoded by MLH1.
4. The process of repairing starts with the protein MutS which binds to mismatched base pairs.
5. MutL is recruited to the complex and activates MutH which binds to GATC sequences.

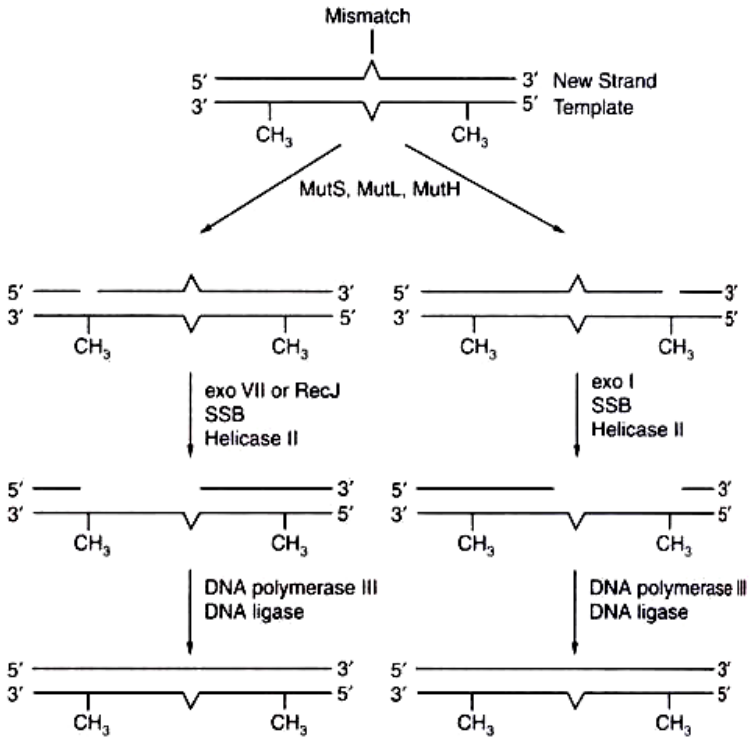


Fig. 4.6. Mismatch Repair (MMR) steps

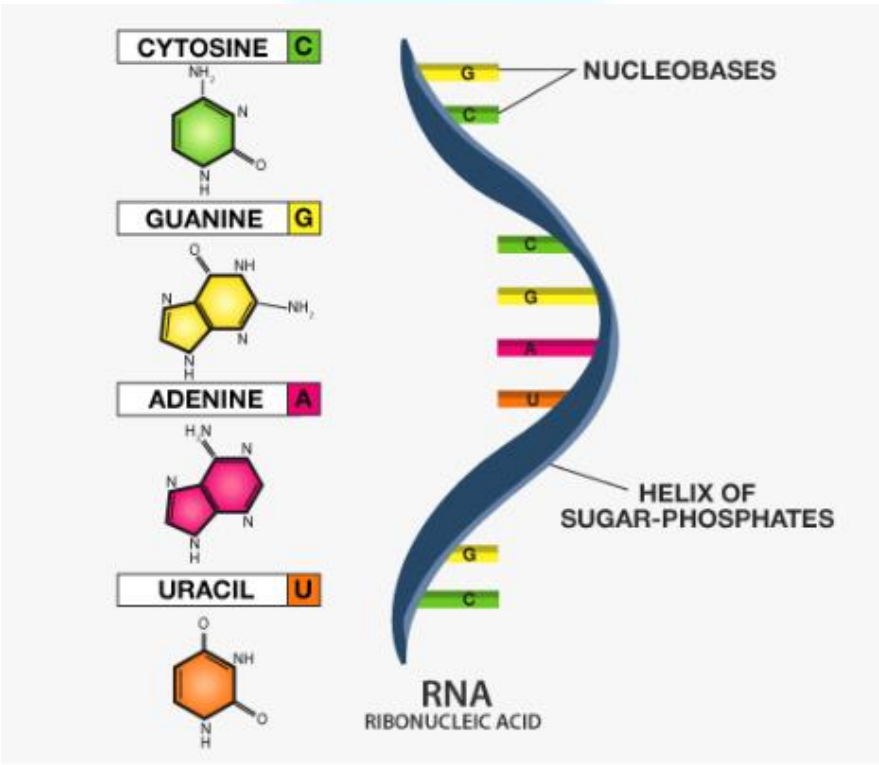
6. Activation of MutH cleaves the unmethylated strand at the GATC site.
7. Then, the segment from the cleavage site to the mismatch is removed by exonuclease (with assistance from helices II and SSB proteins).
8. If the cleavage occurs on the 3′ side of the mismatch, this step is carried out by exonuclease I.
9. It degrades a single strand only in the 3′ to 5′ direction.
10. If the cleavage occurs on the 5′ side of the mismatch, exonuclease VII or RecJ is used to degrade the single stranded DNA.
11. Mismatch repair is very expensive and inefficient as the distance between the GATC site and the mismatch could be as long as 1,000 base pairs.
12. Homologs of MutS and MutL have been found in yeast, mammals, and other eukaryotes. MSH1 to MSH5 are

homologous to MutS; MLH1, PMS1 and PMS2 are homologous to MutL. Colon cancer relate to mutations of MSH2, PMS1 and PMS2.

RNA - Types of RNA and functions

RNA or ribonucleic acid is a polymer of nucleotides which is made up of a **ribose sugar, a phosphate, and bases** such as adenine, guanine, cytosine, and uracil.

It is a polymeric molecule essential in various biological roles in coding, decoding, regulation, and expression of genes.



Like DNA, RNA is a long polymer consisting of nucleotides.

- RNA is a single-stranded helix.
- The strand has a 5′end (with a phosphate group) and a 3′end (with a hydroxyl group).
- It is composed of ribonucleotides.
- The ribonucleotides are linked together by 3′ –> 5′ phosphodiester bonds.

- The nitrogenous bases that compose the ribonucleotides include adenine, cytosine, uracil, and guanine.

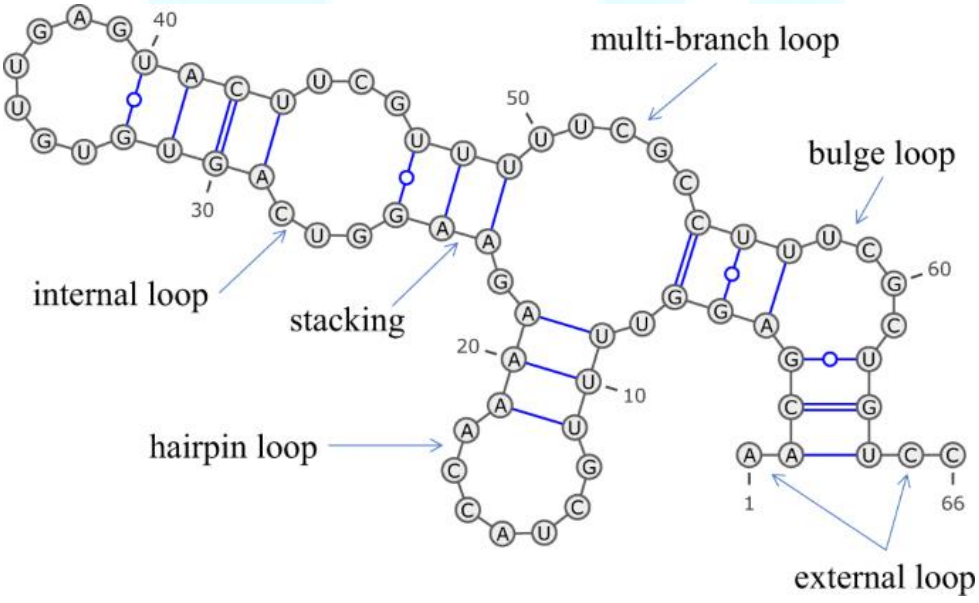
Thus, the difference in the structure of RNA from that of DNA include:

- The bases in RNA are a**denine (abbreviated A), guanine (G), uracil (U) andcytosine (C).**

Thus thymine in DNA is replaced by uracil in RNA, a different pyrimidine. However, like thymine, uracil can form base pairs with adenine.

- The sugar in RNA is ribose rather than deoxyribose as in DNA.
- The corresponding ribonucleosides are adenosine, guanosine, cytidine and uridine. The corresponding ribonucleotides are adenosine 5′-triphosphate (ATP), guanosine 5′-triphosphate (GTP), cytidine 5′-triphosphate (CTP) and uridine 5′-triphosphate (UTP).

RNA Secondary Structure

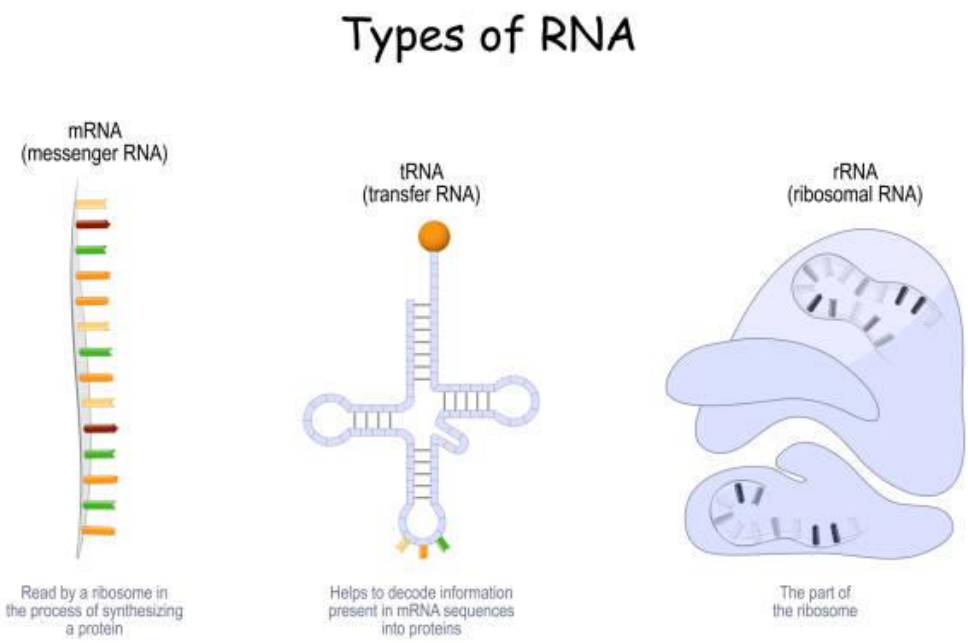


- Most RNA molecules are single-stranded but an RNA molecule may contain regions which can form complementary base pairing where the RNA strand loops back on itself.
- If so, the RNA will have some double-stranded regions.
- Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) exhibit substantial secondary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

Types of RNA

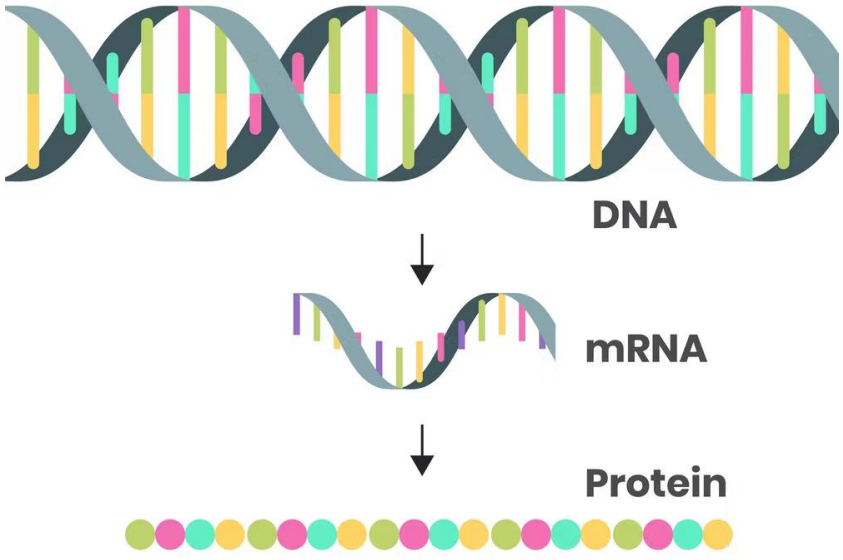
In both prokaryotes and eukaryotes, there are three main types of RNA –

- **mRNA (messenger)**
- **rRNA (ribosomal)**
- **tRNA (transfer)**



1. **Messenger RNA (mRNA)**
    - Accounts for about 5% of the total RNA in the cell.
    - Most heterogeneous of the 3 types of RNA in terms of both base sequence and size.

- It carries the genetic code copied from the DNA during transcription in the form of triplets of nucleotides called codons.
- As part of post-transcriptional processing in eukaryotes, the 5′ end of mRNA is capped with a guanosine triphosphate nucleotide, which helps in mRNA recognition during translation or protein synthesis.
- Similarly, the 3′ end of an mRNA has a poly A tail or multiple adenylate residues added to it, which prevent enzymatic degradation of mRNA. Both 5′ and 3′ end of an mRNA imparts stability to the mRNA.
- Function -
  - mRNA transcribes the genetic code from DNA into a form that can be read and used to make proteins. mRNA carries genetic information from the nucleus to the cytoplasm of a cell.



2. **Ribosomal RNA (rRNA)**
   - **Found in the ribosomes** and account for **80% of the total RNA** present in the cell.

- Ribosomes consist of two major components: the small ribosomal subunits, which read the RNA, and the large subunits, which join amino acids to form a polypeptide chain. Each subunit comprises one or more ribosomal RNA (rRNA) molecules and a variety of ribosomal proteins (r-protein or rProtein).
- Different rRNAs present in the ribosomes include small rRNAs and large rRNAs, which denote their presence in the small and large subunits of the ribosome.
- rRNAs combine with proteins in the cytoplasm to form ribosomes, which act as the site of protein synthesis and has the enzymes needed for the process.
- These complex structures travel along the mRNA molecule during translation and facilitate the assembly of amino acids to form a polypeptide chain. They bind to tRNAs and other molecules that are crucial for protein synthesis.
- Function
  - rRNA directs the translation of mRNA into proteins.

3. Transfer RNA (tRNA)

- tRNA is the smallest of the 3 types of RNA having about 75-95 nucleotides.
- tRNAs are an essential component of translation, where their main function is the transfer of amino acids during protein synthesis. Therefore they are called transfer RNAs.
- Each of the 20 amino acids has a specific tRNA that binds with it and transfers it to the growing polypeptide chain. tRNAs also act as adapters in the translation of the genetic sequence of mRNA into proteins. Therefore they are also called adapter molecules.

Structure of tRNA

tRNAs have a clover leaf structure which is stabilized by strong hydrogen bonds between the nucleotides. Apart from the usual 4 bases, they normally contain some unusual bases mostly formed by methylation of the usual bases, for example, methyl guanine and methylcytosine.

- Three structural loops are formed via hydrogen bonding.
- The 3′ end serves as the amino acid attachment site.
- The center loop encompasses the anticodon.
- The anticodon is a three-base nucleotide sequence that binds to the mRNA codon.
- This interaction between codon and anticodon specifies the next amino acid to be added during protein synthesis.

Function

Transfer RNA brings or transfers amino acids to the ribosome that correspond to each three-nucleotide codon of rRNA. The amino acids then can be joined together and processed to make polypeptides and proteins.

Genome organisation- Exons, introns, overlapping genes and transposons

Genome organisation

- Organisms have a vast array of ways in which their respective genomes are organized.
- A comparison of the genomic organization of six major model organisms shows size expansion with the increase of complexity of the organism.
- There is a more than 300-fold difference between the genome sizes of yeast and mammals, but only a modest 4- to 5-fold increase in overall gene number.

- However, the ratio of coding to noncoding and repetitive sequences is indicative of the complexity of the genome: The largely "open" genomes of unicellular fungi have relatively little noncoding DNA compared with the highly heterochromatic genomes of multicellular organisms.
- In particular, mammals have accumulated considerable repetitive elements and noncoding regions, which account for the majority of their DNA sequences (52% non-coding and 44% repetitive DNA).
- Only 1.2% of the mammalian genome thus encodes for protein function.
- This massive expansion of repetitive and noncoding sequences in multicellular organisms is most likely due to the incorporation of invasive elements, such as DNA transposons, retrotransposons, and other repetitive elements.
- The expansion of repetitive elements (such as Alu sequences) has even infiltrated the transcriptional units of the mammalian genome.
- This results in transcription units that are frequently much larger (30–200 kb), commonly containing multiple promoters and DNA repeats within untranslated introns.
- The vast expansion of the genome with noncoding and repetitive DNA in higher eukaryotes implies more extensive epigenetic silencing mechanisms.
- Studies of the genomic organization is thought to be the future of genomic medicine, which will provide the opportunity for personalized prognoses in clinics.

Exons and Introns

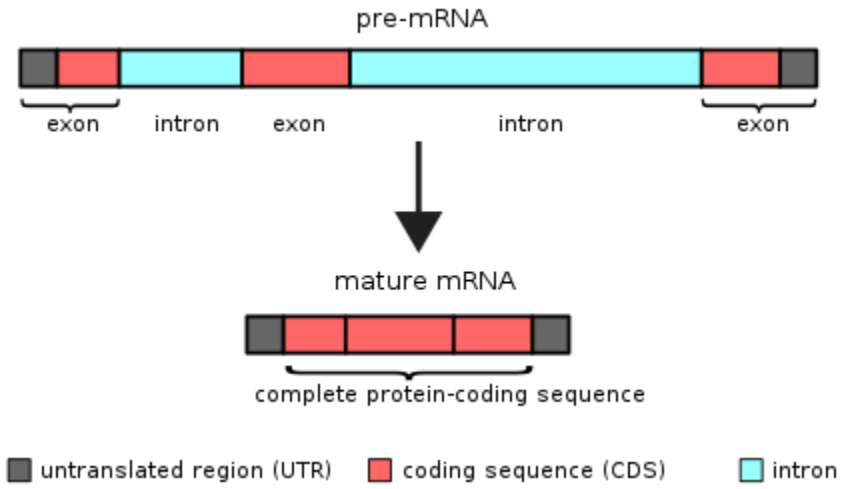
Any nucleotides sequence within a gene that is removed by RNA splicing of the final product during maturation is known as an intron. The final mature RNA produced by that gene is encoded by exons.

A gene containing coding regions unknown exons are interrupted by non-coding regions known as introns. The DNA region between exons and introns is encoded by exons proteins. In the coding region, only the eukaryotes contain introns.
In eukaryotes, both exons and introns are transcribed into the mRNA primary transcript.

Exons are termed nucleic acid sequences represented in the RNA molecule. **Introns** can be defined as the nucleotide sequences that are found in the genes that are removed by the process of RNA splicing. We can also say that exons are coding areas, whereas introns are non-coding areas. Both Roberts and Phillip Sharp discovered introns and exons respectively. The introns change their sequences frequently with time, whereas the exon sequences are highly conserved.

What are Introns?

- Introns play the role of intervening sequences between two exons found in eukaryotes. They do not directly code for proteins. They are removed before the mRNA forms proteins. Therefore, these introns undergo the process of splicing.
- Introns are the non-coding parts of the nucleotides and aren't highly conserved. Therefore, it's essential to get rid of introns to stop the formation of incorrect proteins.
- The word intron means 'In the Nucleus'. Thus, the Universal feature in introns is to remove the splicing within the nucleus by RNA.
-



- The sequence of nucleotides found in RNA and DNA that interrupts the sequence of the gene is known as an intron.
- Introns are found in the mRNA, primary transcript, and intergenic regions of the gene.
- Hence, mature RNA lacks introns on the other hand RNA splicing mechanism lacks in prokaryotes.

Function of Introns

- While introns were initially – and to an extent still are – considered **'junk DNA'**, it's been shown that introns likely play a crucial role in regulation and organic phenomenon.
- As introns cause a rise in gene length, this increases the likelihood of crossover and recombination between sister chromosomes.
- This increases genetic variation and may end in new gene variants through duplications, deletions, and exon shuffling.
- Introns also allow for alternative splicing.
- This allows one gene to encode multiple proteins because the exons are often assembled in multiple ways.
- The RNA polymerase makes a copy of the whole gene during transcription, both introns, and exons, into the initial mRNA transcript referred to as pre-mRNA or heterogeneous nuclear RNA (hrRNA).

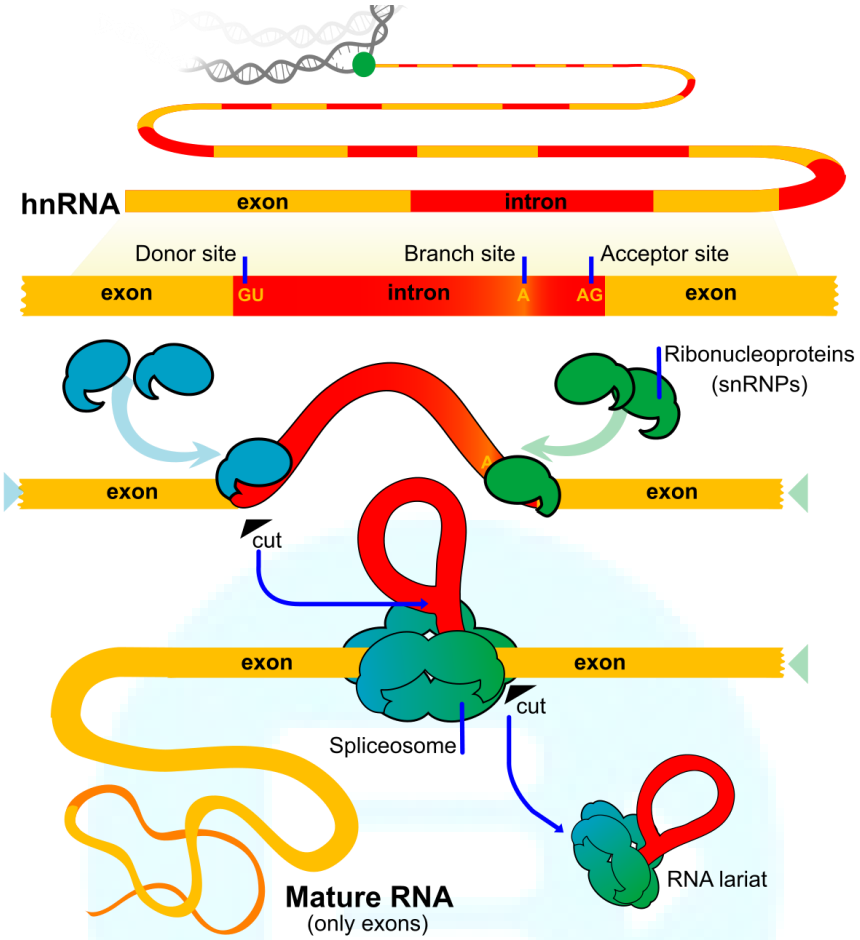
- As introns aren't transcribed, they need to then be removed before translation can occur.
- The excision of introns and therefore the connection of exons into a mature mRNA molecule occurs within the nucleus and is understood as splicing.
- Introns contain a variety of sequences that are involved in splicing including spliceosome recognition sites.
- These sites help the spliceosome to identify the boundary between the introns and exons.
- **Nucleolar ribonucleoproteins (snRNPs)** are recognised by small sites themselves. There are a variety of snRNPs involved in mRNA splicing which combine to create a spliceosome. The splicing takes place in three steps.

What are Exons?

- Exons are the coding sequences that code for the amino acid sequence of the protein. The exons are transcribed into mature mRNA after post-transcriptional modification. These are highly conserved sequences, i.e., they are not changing frequently with time.
- Splicing is the process of removing introns. The production of different combinations of amino acids is promoted by alternative splicing by combining different combinations of exons together.
- Therefore the Amino acid sequence of the polypeptide exons is responsible.

Structure and Function of Exons

- In protein-coding genes, the exons include both the protein-coding sequence and therefore the 5′- and 3′-untranslated regions (UTR).
-



- Often the primary exon includes both the 5′-UTR and therefore the first a part of the coding sequence, but exons containing only regions of 5′-UTR or (more rarely) 3′-UTR occur in some genes, i.e. the UTRs may contain introns. Some non-coding RNA transcripts even have exons and introns.
- Mature mRNAs originating from an equivalent gene needn't include equivalent exons, since different introns within the pre-mRNA are often removed by the method of another splicing.
- Exonization is the creation of a replacement exon, as a result of mutations in introns.

Overlapping genes

- **An overlapping gene (or OLG)** is a gene whose expressible nucleotide sequence partially overlaps with the expressible nucleotide sequence of another gene.
- In this way, a nucleotide sequence may make a contribution to the function of one or more gene products.
- Overlapping genes are present and a fundamental feature of both cellular and viral genomes.
- The current definition of an overlapping gene varies significantly between eukaryotes, prokaryotes, and viruses.
- In prokaryotes and viruses overlap must be between coding sequences but not mRNA transcripts, and is defined when these coding sequences share a nucleotide on either the same or opposite strands.
- In eukaryotes, gene overlap is almost always defined as **mRNA transcript overlap**.
- Specifically, a gene overlap in eukaryotes is defined when at least one nucleotide is shared between the boundaries of the primary mRNA transcripts of two or more genes, such that a DNA base mutation at any point of the overlapping region would affect the transcripts of all genes involved. This definition includes 5′ and 3′ untranslated regions (UTRs) along with introns.

Genes may overlap in a variety of ways and can be **classified by their positions** relative to each other.

- **Unidirectional or tandem overlap:** the 3' end of one gene overlaps with the 5' end of another gene on the same strand. This arrangement can be symbolized with the notation → → where arrows indicate the reading frame from start to end.
- **Convergent or end-on overlap:** the 3' ends of the two genes overlap on opposite strands. This can be written as → ←.

- **Divergent or tail-on overlap:** the 5' ends of the two genes overlap on opposite strands. This can be written as ← →.

Overlapping genes can also be **classified by phases**, which describe their relative reading frames:

- **In-phase overlap** occurs when the shared sequences use the same reading frame. This is also known as "phase 0". Unidirectional genes with phase 0 overlap are not considered distinct genes, but rather as alternative start sites of the same gene.
- **Out-of-phase overlaps** occurs when the shared sequences use different reading frames. This can occur in "phase 1" or "phase 2", depending on whether the reading frames are offset by 1 or 2 nucleotides. Because a codon is three nucleotides long, an offset of three nucleotides is an in-phase, phase 0 frame.

Transposons

- **transposon**, class of genetic elements that **can "jump"** to different locations within a genome. Although these elements are frequently called **"jumping genes**," they are always maintained in an integrated site in the genome. In addition, most transposons eventually become inactive and no longer move.
- Transposons were first discovered in corn (maize) during the 1940s and '50s by American scientist **Barbara McClintock**, whose work won her the Nobel Prize for Physiology or Medicine in 1983.
- Since McClintock's discovery, three basic types of transposons have been identified.
- These include class II transposons, miniature inverted-repeat transposable elements **(MITEs, or class III transposons)**, and retrotransposons **(class I transposons)**.

Class II transposons

- Class II elements are simply segments of DNA that move from one place to another via a **"cut and paste" mechanism**.
- Most, if not all, of these elements encode an enzyme called transposase, which acts to cleave the ends of the transposon, freeing it from its initial location in the genome.
- Transposase also cleaves target sites where the element is to be inserted. Once the transposon is ligated (bound) into its new position, gaps that are left in the DNA sequence are filled in through the synthesis of nucleotides.
- Class II transposons range in length from 1,000 to as many as 40,000 base pairs.

Miniature inverted-repeat transposable elements

- **MITEs** are characterized by their short lengths, generally about 400 to 600 base pairs, and by a stretch of about 15 base pairs that occurs at each end of each element in an inverted fashion (as mirror sequences).
- The mechanism by which these elements move about genomes is not well understood.
- Thousands of MITEs have been identified in the genomes of Oryza sativa (cultivated rice), Caenorhabditis elegans (a type of nematode), and other organisms.
- Unlike some types of transposons, MITEs do not appear to encode proteins, and most insertions of these elements occur in euchromatin, the form of chromosomal material that contains the majority of active genes.
- As a result, a genetic regulatory function of MITEs has been proposed, and this has received support from evidence that some microRNAs (miRNAs), which play a role in RNA interference (a form of gene regulation), are derived from MITEs.

Retrotransposons

- Retrotransposons represent a highly unique group of transposable elements and form large portions of the genomes of many eukaryotes (organisms with cells containing a clearly defined nucleus).
- Retrotransposons function by a **"copy and paste" mechanism**.
- Thus, they leave behind the original copy and generate a second copy that is inserted elsewhere in the genome.
- This process results in the insertion of repetitive sequences of DNA throughout the genome and is the mechanism responsible for the vast spread of transposable elements in many higher organisms.
- The first step in retrotransposition occurs when the transposable DNA is copied into RNA.
- The RNA segment then jumps to another location in the genome. However, in order to be inserted into the genome at the new site, the RNA must be copied back into DNA by an enzyme called reverse transcriptase.
- There are several different types of retrotransposons, including long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs).
- About 20 percent of the human genome is made up of LINEs.

Transposons and antibiotic resistance

- The simplest kinds of transposons merely contain a copy of the transposase with no additional genes. They behave as parasitic elements and usually have no known associated function that is advantageous to the host. More often, transposable elements have additional genes associated with them—for example, antibiotic resistance factors.
- Antibiotic resistance typically occurs when an infecting bacterium acquires a plasmid that carries a gene encoding resistance to one or more antibiotics.

- Typically, these resistance genes are carried on transposable elements that have moved into plasmids and are easily transferred from one organism to another.
- Once a bacterium picks up such a gene, it enjoys a great selective advantage because it can grow in the presence of the antibiotic.
- Indiscriminate use of antibiotics actually promotes the buildup of these drug-resistant plasmids and strains.

Genetic code- characteristic features

The information transferred or passed from the parent generation to the offspring is called **genetic code**. The process begins at the cellular level when the genes get split into the parental reproductive cells and then get united to form a hybrid set of genes during fertilization.

Apart from the union of the male and female reproductive cells in sexual reproduction, there are other means of reproducing young ones where genetic code is passed. For instance, in asexual reproduction, the splitting of the cells of one parent-cell to form daughter cells is also the process of passing genetic information from one generation to another.

This is an intricate process of replicating and transcribing genes. The biochemical process of copying a very complex set of nucleic acids or genetic material within the reproductive cells and their union with the other half of the reproductive cell results in the formation of genetic information. In a nutshell, the answer to what is genetic code is the set of information passed from a parental generation to the offspring in the form of genes to conduct various physiological functions.

Characteristics of Genetic Code

The intricate process of decoding genes and then recording to form a new set of genes in the offspring is fascinating. The entire biochemical process

comprises millions of chemical reactions occurring within a minute yet controlled environment of a cell.

To understand its features, first, we need to observe the scientific discovery and theories behind it. George Gamow, a renowned physicist proposed that the four genetic bases (guanine, adenine, thymine, and cytosine) are the primary constituents of the code for 20 existing amino acids. Hence, the combination of these bases will occur in the set of three resulting in 43 or 64 combinations. These combinations are called codons.

Renowned biochemist Har Gobind Khorana developed a brilliant method to synthesize artificial RNA by defining the combination of these bases and took a step ahead to decipher a genetic code. Marshall Nirenberg, another geneticist developed a cell-free environment to synthesize protein and to decipher this code. Severo Ochoa, a brilliant biochemist, discovered an enzyme that can synthesize RNA with proper sequence without using DNA templates.

Let us move forward to understand the salient features of genetic code now.

- Each codon thus has 3 bases producing 64 codons. Out of these 64 codons, only 61 are capable of producing amino acids. The rest 3 are used as stop codons during the process of translation.
- One codon is responsible to direct the reactions for producing an amino acid. Hence, the process is absolutely specific and focused.
- Some of the amino acids can have multiple codons for production. This is called the degeneracy of genetic code. For example, Valine (Val) has four different sequential codons for production. They are GUA, GUC, GUU, and GUG.
- Codons are responsible for the formation of mRNA. On the other hand, mRNA is responsible for the generation of genes.
- Another prime feature of the genetic code is that it is universal. It means that one codon will lead to the formation of one amino acid. For instance,

Phenylalanine (Phe) has the genetic code UUU. It is universal across all living beings. It means that the Phe of a bacterium will be similar to that of a human being.

- Sometimes, codons have dual functions too. For instance, AUG is the genetic code for Methionine (Met). It also acts as an initiator or start codon.
- The standard genetic code is universal virtually among all extant life forms.

Genetic Code - Properties

It is a must that the genetic codes properties are known by all the students who are in touch with their Biology, or who study it:

- They are mostly triplet coded
- They are unambiguous as well as universal in nature
- They have a degenerate code
- They contain start and stop codons
- They showcase polarity
- Their code is mostly non overlapping
- They are commaless, hence have no indication of an end or a beginning

Mutations of Genetic Codes

Not every individual is similar. In fact, it has been observed that a particular physiological trait goes missing. It happens when the genetic codes get rearranged and deleted during transcription and replication. The different segments of DNA get rearranged and deleted during the process resulting in mutations.

Genes are gained and lost in the process resulting in new physical traits of an organism. Let us consider an example. If Valine (Val) is replaced by Glutamine

(Gln) in a particular gene sequence, the individual will develop sickle cell anemia, a blood disorder.

Wrapping up, the genetic code of an organism depends on the genes received from the parental generation and the way it is produced from replication and transcription. We have learned how the bases form codons and how they direct the genesis of amino acids. These amino acids are units of protein that form the genetic material of an organism following a proper sequence. From the above-given information, students can easily prepare a genetic code PPT and check your understanding of the genetic code.

Deciphering genetic code

- The most important feature of the genetic code is that it is a triplet codon.
- Three consecutive nucleotides of a single strand of DNA contain the information for coding a specific amino acid.
- It is known as a triplet codon.
- Translation takes place in such a way that these nucleotide triplets are read in a successive non-overlapping fashion.
- The information is first transcribed into messenger RNA, which has a sequence of bases complementary to DNA from which it is copied.

DNA has four types of bases C, T, G, A while RNA has four complementary bases G, A, C, U. The four base language of DNA is translated into language of 20 amino acids. Deciphering or cracking of genetic code is the outcome of research of various scientists like Marshal Nirenberg, Steve Ochoa, Hargobind Khorana, Francis Crick, Mattaei and many others.

They discovered that the order in which the nucleotides are arranged – mRNA would determine the sequence of amino acids in polypeptides. Nirenberg and Mattaei gave the first experimental proof for the triplet codon. They used artificial mRNA raide of only uracil nucleotides (Poly U) in a cell free system. It

resulted in the synthesis of polypeptide chain made up of only one kind of amino acid, phenylalanine. It was concluded that codon for phenylalanine was uridylic acid basis (uracil), UUU.

Similarly poly C(CCC) codon represented amino acid proline and poly A(AAA) codon represents am: no acid chain of lysine.

Later Hargobind Khorana confirmed the genetic code to be triplet codon. Using synthetic mRNA have alternating polynucleotides in a cell free system, discovered the chain of alternating amino acid using alternating uracil (U) and guanine (G) triplets which showed the following results.

GUG	UGU	GUG	UGU	GUG
Valine	Cystine	Valine	Cystine	Valine

Similarly, alternating ACA and CAC triplets produced a chain of following amino acids.

ACA	CAC	ACA	CAC	ACA
Threonine	Histidine	Threonine	Histidine	Threonine

This also confirmed that each codon is a triplet. The cell free protein-synthesizing system was an extract of E. coli without walls. It contained ribosomes, tRNA, tRNA synthetase enzymes, ATP and radioactive amino acids. Use of artificial trinucleotide templates resulted in determination of base composition of all genetic codons.

Khorana Technique:

H. G. Khorana and associates synthesized DNA of known sequences in vitro conditions. From this DNA, the mRNA is transcribed. These mRNA molecules, which have known sequences of nucleotides are then directed to synthesize polypeptides.

These polypeptides are then sequentially degraded to know their amino acid sequences. The nucleotide sequence is then compared with amino acid sequence. This nucleotide sequence specifies codons of amino acids. This technique led to the determination of codons for all amino acids.

DNA Genetic Codon is Triplet:

If a genetic codon consisted of two consecutive bases, the number of codons would be 42 = 16. Since the number of amino acids is 20, this is insufficient. Therefore, three is the minimum number of bases needed to code for 20 amino acids 43 = 64. George Gamow in 1964 pointed out that the code would contain at least three consecutive bases.

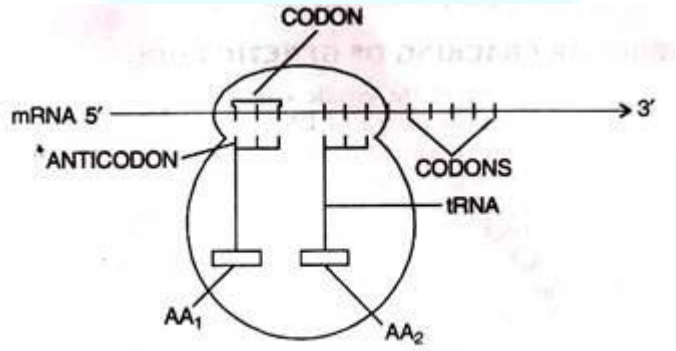


Fig. 11.1. Triplet codons on mRNA dictate the arrangement of amino acids of a polypeptide chain.

The length of the coding portion of a gene called reading frame depends upon the length of the message to be translated. For example, a sequence of 600 nucleotides will code for a polypeptide having a chain of 200 amino acids. Therefore, the length of mRNA depends upon the length of polypeptide it codes for.

Codons provide key to the translation of genetic information dictating the synthesis of specific proteins. The protein synthesis machinery reads triplet codons sequentially from one triplet codon to the next. Genetic code sequences explain how protein sequence information is stored in nucleic acid and how the information is translated into proteins.

Reading frame

- In molecular biology, a reading frame is a way of dividing the sequence of nucleotides in a nucleic acid (DNA or RNA) molecule into a set of consecutive, non-overlapping triplets.
- Where these triplets equate to amino acids or stop signals during translation, they are called codons.
- A single strand of a nucleic acid molecule has a phosphoryl end, called **the 5′-end**, and a **hydroxyl or 3′-end**.
- These define the 5′→3′ direction. There are three reading frames that can be read in this 5′→3′ direction, each beginning from a different nucleotide in a triplet.
- In a double stranded nucleic acid, an additional three reading frames may be read from the other, complementary strand in the 5′→3′ direction along this strand.
- As the two strands of a double-stranded nucleic acid molecule are antiparallel, the 5′→3′ direction on the second strand corresponds to the 3′→5′ direction along the first strand.
- In general, at the most, one reading frame in a given section of a nucleic acid, is biologically relevant (open reading frame).
- Some viral transcripts can be translated using multiple, overlapping reading frames.
- There is one known example of overlapping reading frames in mammalian mitochondrial DNA: coding portions of genes for 2 subunits of ATPase overlap.

Frameshift

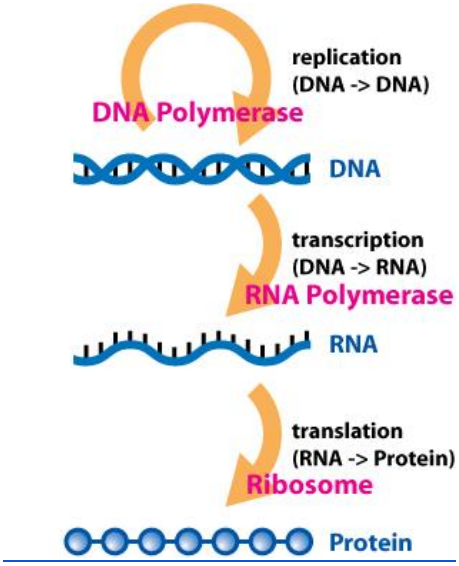
- A frameshift mutation (also called a **framing erro**r or a reading frame shift) is a genetic mutation caused by indels (insertions or deletions) of a number of nucleotides in a DNA sequence that is not divisible by three.

- Due to the triplet nature of gene expression by codons, the insertion or deletion can change the reading frame (the grouping of the codons), resulting in a completely different translation from the original.
- The earlier in the sequence the deletion or insertion occurs, the more altered the protein.
- A frameshift mutation is not the same as a single-nucleotide polymorphism in which a nucleotide is replaced, rather than inserted or deleted.
- A frameshift mutation will in general cause the reading of the codons after the mutation to code for different amino acids. The frameshift mutation will also alter the first stop codon ("UAA", "UGA" or "UAG") encountered in the sequence. The polypeptide being created could be abnormally short or abnormally long, and will most likely not be functional.
- Frameshift mutations are apparent in severe genetic diseases such as Tay–Sachs disease; they increase susceptibility to certain cancers and classes of familial hypercholesterolaemia; in 1997, a frameshift mutation was linked to resistance to infection by the HIV retrovirus.
- Frameshift mutations have been proposed as a source of biological novelty, as with the alleged creation of nylonase, however, this interpretation is controversial.
- A study by Negoro et al (2006) found that a frameshift mutation was unlikely to have been the cause and that rather a two amino acid substitution in the active site of an ancestral esterase resulted in nylonase.

Protein synthesis: Central dogma, Transcription, Transcription factors, Transcription activators and repressors, RNA polymerases, capping, elongation and termination.

ENTRI

Transcription



- It is one of the first processes in gene expression.
- The genetic information flows from DNA to protein and this flow of information takes place in a sequential process of transcription and translation.
- Only one strand of DNA is copied during the process of transcription known as the template strand and the RNA formed is called the mRNA.
- The main motive of transcription is to make a copy of RNA from the DNA sequence.
- The RNA transcript carries the information used to encode a protein.

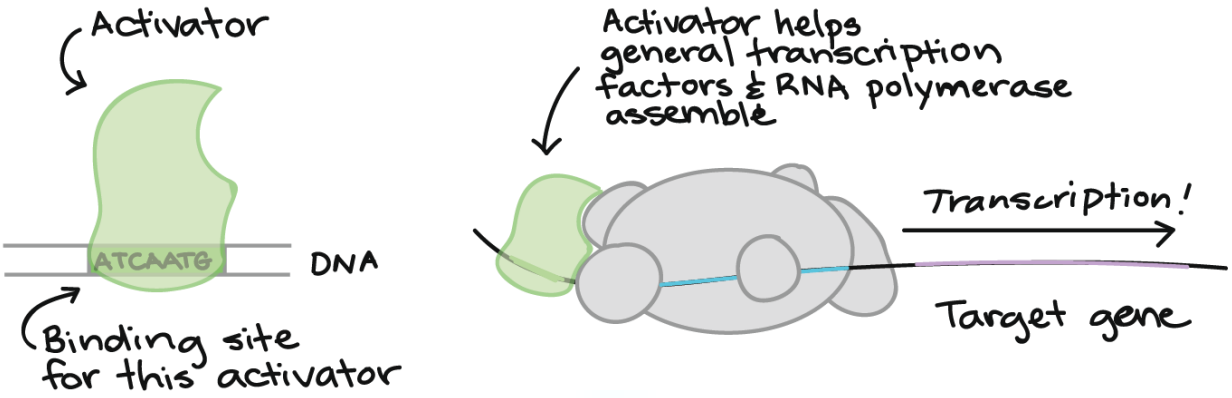
Transcription factors

- Transcription factors are proteins that help turn specific genes "on" or "off" by binding to nearby DNA.
- Transcription factors that are **activators** boost a gene's transcription. **Repressors** decrease transcription.

    - **Activators**

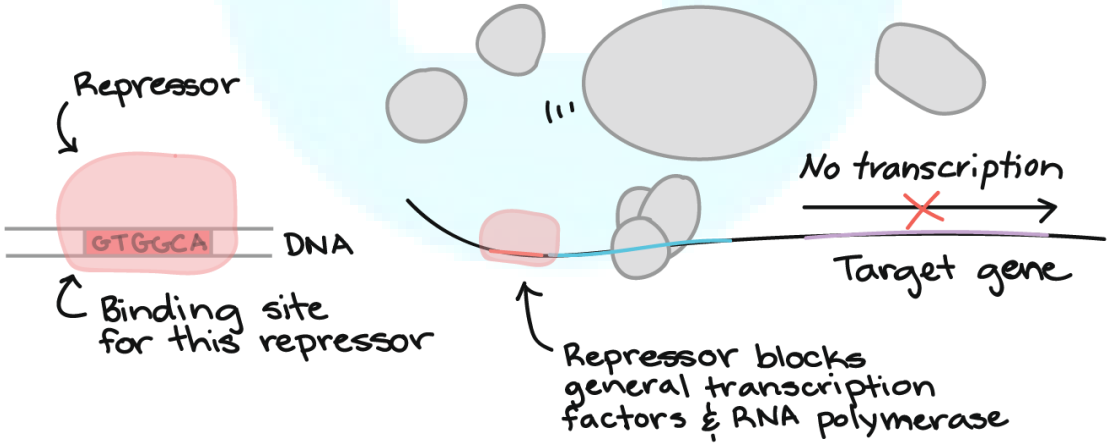
        Some transcription factors activate transcription. For instance, they may help the general transcription factors and/or

RNA polymerase bind to the promoter, as shown in the diagram below.



- **Repressors**

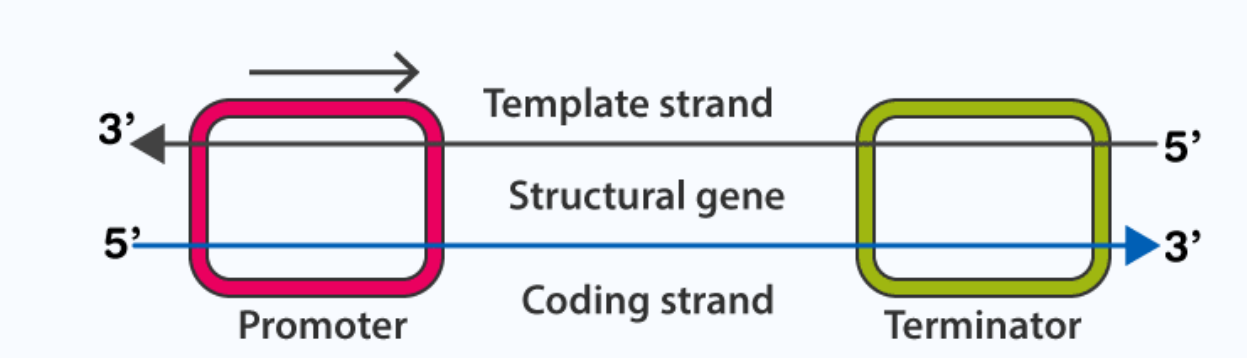
  Other transcription factors repress transcription. This repression can work in a variety of ways. As one example, a repressor may get in the way of the basal transcription factors or RNA polymerase, making it so they can't bind to the promoter or begin transcription.



RNA Polymerase

- The RNA polymerase is the main enzyme involved in transcription.
- It uses single-strand DNA to synthesize a complementary RNA strand.

- The DNA-dependent RNA polymerase binds to the promoter and catalyses the polymerization in the 5′ to 3′ direction on the template strand.
- Once it reaches the terminator sequence, the process terminates and the newly synthesised RNA strand is released.
- Transcription Unit is a stretch of a DNA transcribed into an RNA molecule.
- Its function is to encode at least one gene.
- Suppose if gene encodes protein than mRNA is produced by transcription.
- A protein encoded by the DNA transcription unit may comprise a coding sequence.
- Compared to DNA replication, transcription has a lower copying fidelity.



Stages of Transcription

Transcription proceeds in enzymatically catalysed steps i.e.

1. Initiation
2. Elongation
3. Termination

Initiation

- RNA polymerase attaches to the DNA molecule and moves along the DNA strand until it recognises a promoter sequence.
- These are known as the transcription start sites.

- The DNA double helix then unwinds and all the bases on each of the DNA strands are exposed.
- This acts as a template for a new mRNA strand.

Elongation

- Ribonucleotides are added to the template strand that enables the growth of mRNA growth.

Termination

- RNA polymerase encounters a terminator sequence and the transcription stops.
- RNA polymerase then releases the DNA template.

RNA Processing

The transcribed RNA is known as the pre-mRNA. It is processed further to convert it into mature RNA. RNA processing include:

1. Capping
2. Polyadenylation
3. Splicing

Capping

A methylated guanine cap is added to protect the mRNA. It involves:

- Addition of methylated guanine
- It occurs at 5′ end of mRNA transcript
- It protects the mRNA from degradation

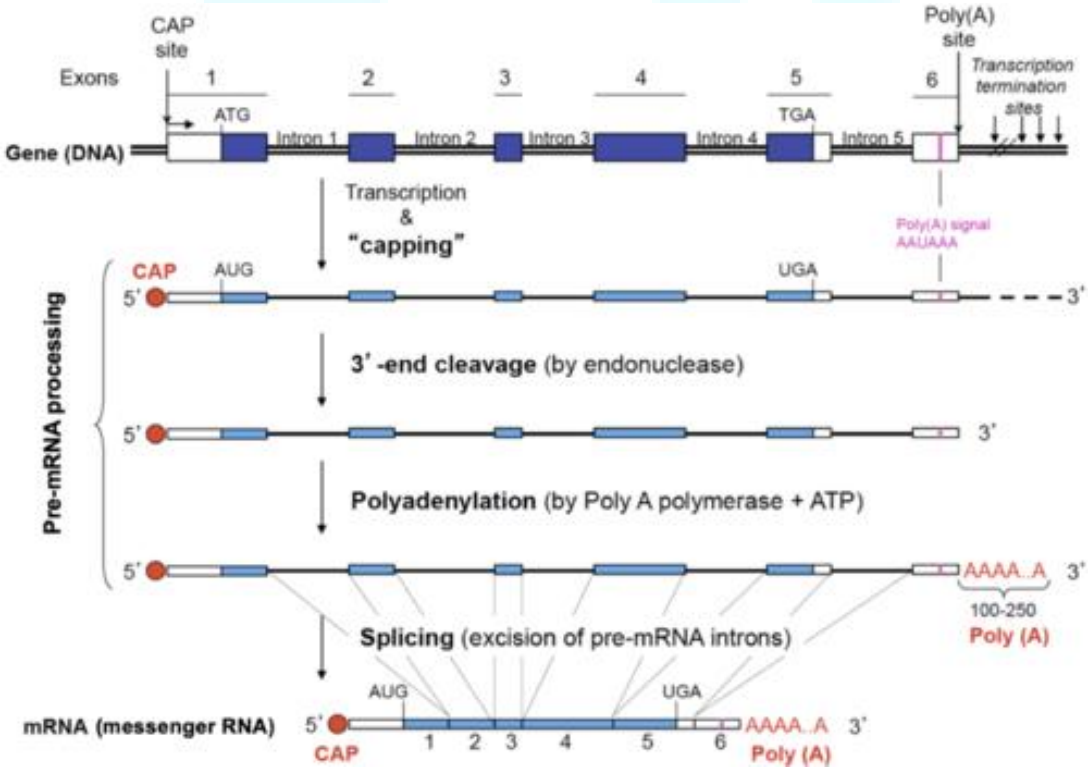
Polyadenylation

The poly-A tail also protects the mRNA from degradation. It involves:

- The endonucleases cleave the mRNA at a specific sequence.
- The enzyme polyA polymerase facilitates the addition of several adenine nucleotides.

Splicing

- The non-coding sequences, i.e., the introns are removed by spliceosome excision.
- The coding sequences or the exons join together by ligation.



Thus several proteins can be made from a single pre-mRNA. A mature mRNA is obtained at the end of transcription.

Post-transcriptional processing in eukaryotes

- Post-transcriptional modification or co-transcriptional modification is a set of biological processes common to most eukaryotic cells by which an RNA primary transcript is chemically altered following transcription from a gene to produce a mature, functional RNA molecule that can then leave the nucleus and perform any of a variety of different functions in the cell.
- There are many types of post-transcriptional modifications achieved through a diverse class of molecular mechanisms.
- One example is the conversion of precursor messenger RNA transcripts into mature messenger RNA that is subsequently capable of being translated into protein.
- This process includes three major steps that significantly modify the chemical structure of the RNA molecule: the addition of a **5' cap**, the addition of a **3' polyadenylated tail**, and RNA splicing. Such processing is vital for the correct translation of eukaryotic genomes because the initial precursor mRNA produced by transcription often contains both exons (coding sequences) and introns (non-coding sequences); splicing removes the introns and links the exons directly, while the cap and tail facilitate the transport of the mRNA to a ribosome and protect it from molecular degradation.
- Post-transcriptional modifications may also occur during the processing of other transcripts which ultimately become transfer RNA, ribosomal RNA, or any of the other types of RNA used by the cell.

5' processing

Capping

- Capping of the pre-mRNA involves the addition of 7-methylguanosine (m7G) to the 5' end.
- To achieve this, the terminal 5' phosphate requires removal, which is done with the aid of a phosphatase enzyme.
- The enzyme guanosyl transferase then catalyses the reaction, which produces the diphosphate 5' end.
- The diphosphate 5' end then attacks the alpha phosphorus atom of a GTP molecule in order to add the guanine residue in a 5'5' triphosphate link.
- The enzyme (guanine-N7-)-methyltransferase ("cap MTase") transfers a methyl group from S-adenosyl methionine to the guanine ring.
- This type of cap, with just the (m7G) in position is called a **cap 0 structure**.
- The ribose of the adjacent nucleotide may also be methylated to give a cap 1.
- Methylation of nucleotides downstream of the RNA molecule produce cap 2, cap 3 structures and so on.
- In these cases the methyl groups are added to the 2' OH groups of the ribose sugar.
- The cap protects the 5' end of the primary RNA transcript from attack by ribonucleases that have specificity to the 3'5' phosphodiester bonds.

3' processing

Cleavage and polyadenylation

- The pre-mRNA processing at the 3' end of the RNA molecule involves cleavage of its 3' end and then the addition of about 250 adenine residues to form a poly(A) tail.

- The cleavage and adenylation reactions occur primarily if a polyadenylation signal sequence (5'- AAUAAA-3') is located near the 3' end of the pre-mRNA molecule, which is followed by another sequence, which is usually (5'-CA-3') and is the site of cleavage.
- A GU-rich sequence is also usually present further downstream on the pre-mRNA molecule.
- More recently, it has been demonstrated that alternate signal sequences such as UGUA upstream off the cleavage site can also direct cleavage and polyadenylation in the absence of the AAUAAA signal.
- It is important to understand that these two signals are not mutually independent and often coexist.
- After the synthesis of the sequence elements, several multi-subunit proteins are transferred to the RNA molecule.
- The transfer of these sequence specific binding proteins cleavage and polyadenylation specificity factor (CPSF), Cleavage Factor I (CF I) and cleavage stimulation factor (CStF) occurs from RNA Polymerase II.
- The three factors bind to the sequence elements. The AAUAAA signal is directly bound by CPSF.
- For UGUA dependent processing sites, binding of the multi protein complex is done by Cleavage Factor I (CF I).
- The resultant protein complex formed contains additional cleavage factors and the enzyme Polyadenylate Polymerase (PAP).
- This complex cleaves the RNA between the polyadenylation sequence and the GU-rich sequence at the cleavage site marked by the (5'-CA-3') sequences.
- Poly(A) polymerase then adds about 200 adenine units to the new 3' end of the RNA molecule using ATP as a precursor.

- As the poly(A) tail is synthesized, it binds multiple copies of poly(A)-binding protein, which protects the 3'end from ribonuclease digestion by enzymes including the CCR4-Not complex.

Introns Splicing

- RNA splicing is the process by which introns, regions of RNA that do not code for proteins, are removed from the pre-mRNA and the remaining exons connected to re-form a single continuous molecule.
- Exons are sections of mRNA which become "expressed" or translated into a protein. They are the coding portions of a mRNA molecule.
- Although most RNA splicing occurs after the complete synthesis and end-capping of the pre-mRNA, transcripts with many exons can be spliced co-transcriptionally.
- The splicing reaction is catalyzed by a large protein complex called the **spliceosome** assembled from proteins and small nuclear RNA molecules that recognize splice sites in the pre-mRNA sequence.
- Many pre-mRNAs, including those encoding antibodies, can be spliced in multiple ways to produce different mature mRNAs that encode different protein sequences.
- This process is known as alternative splicing, and allows production of a large variety of proteins from a limited amount of DNA.

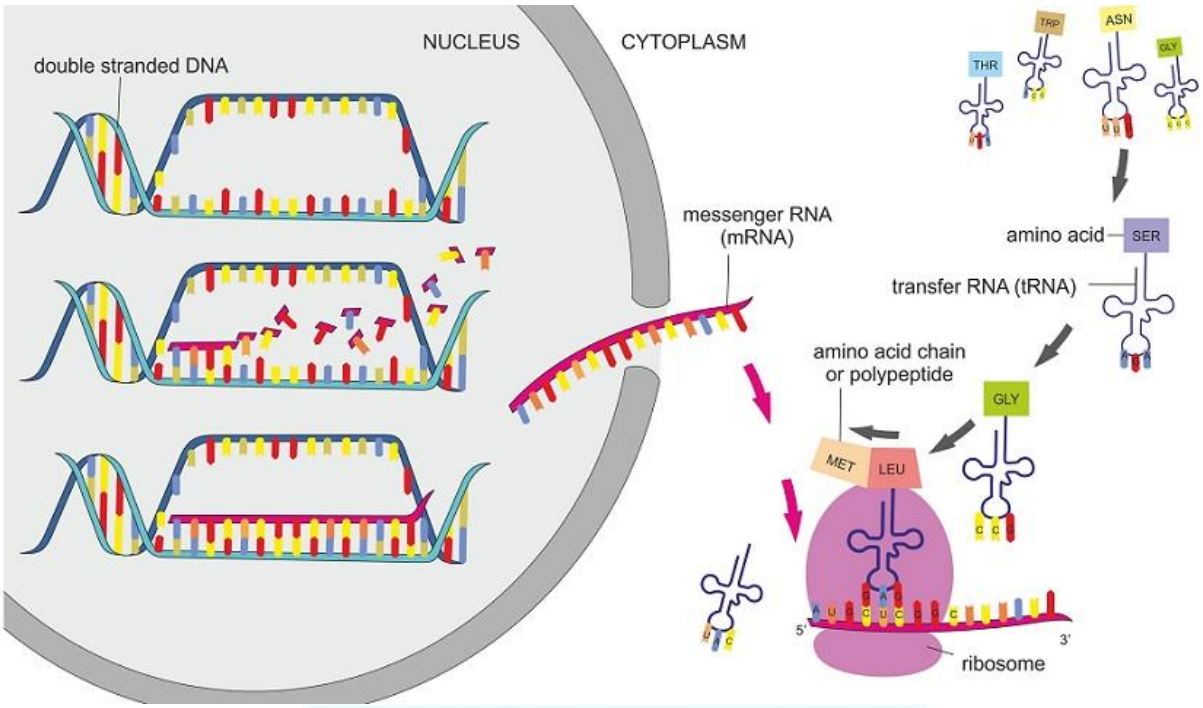
Histone mRNA processing

- Histones H2A, H2B, H3 and H4 form the core of a nucleosome and thus are called core histones.
- Processing of core histones is done differently because typical histone mRNA lacks several features of other eukaryotic mRNAs, such as poly(A) tail and introns.
- Thus, such mRNAs do not undergo splicing and their 3' processing is done independent of most cleavage and polyadenylation factors.

- Core histone mRNAs have a special stem-loop structure at 3-prime end that is recognized by a stem–loop binding protein and a downstream sequence, called **histone downstream element (HDE)** that recruits U7 snRNA.
- Cleavage and polyadenylation specificity factor 73 cuts mRNA between stem-loop and HDE.
- Histone variants, such as H2A.Z or H3.3, however, have introns and are processed as normal mRNAs including splicing and polyadenylation.

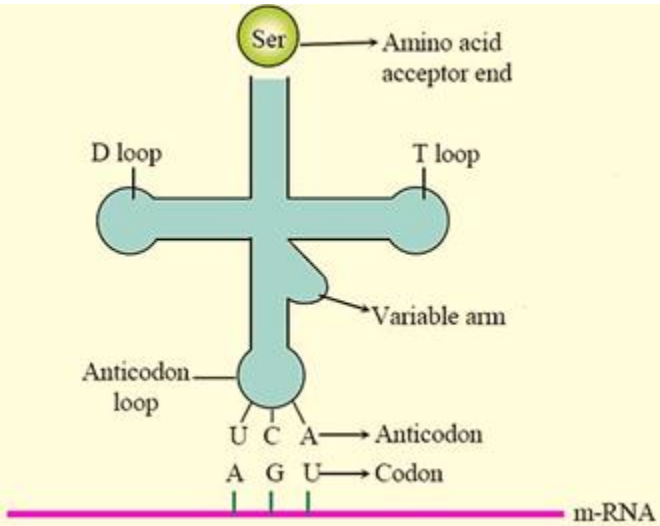
Translation: Mechanism, initiation complex, elongation and termination, Post-translational modifications of proteins.

Translation Process in Protein Synthesis



- The translation is a process of protein synthesis from mRNA with the help of ribosomes.
- Translational unit of mRNA from 5′ to 3` includes start codon, region coded polypeptide, a stop codon, and untranslated regions (UTRs) at 5`end & 3`end both for more efficiency of the process.
- The ribosome is the place where the whole machinery of translation is present.
- Each eukaryotic ribosome has two parts: a smaller 40S subunit and a larger 60S subunit.
- The smallest unit has a point for attachment of mRNA.
- Along with the largest subunit, it forms a P-site or peptidyl transfer (Donor site).
- There are binding sites for initiation factors, elongation factors, translocation, etc.

Structure and Role of tRNA in Protein Synthesis

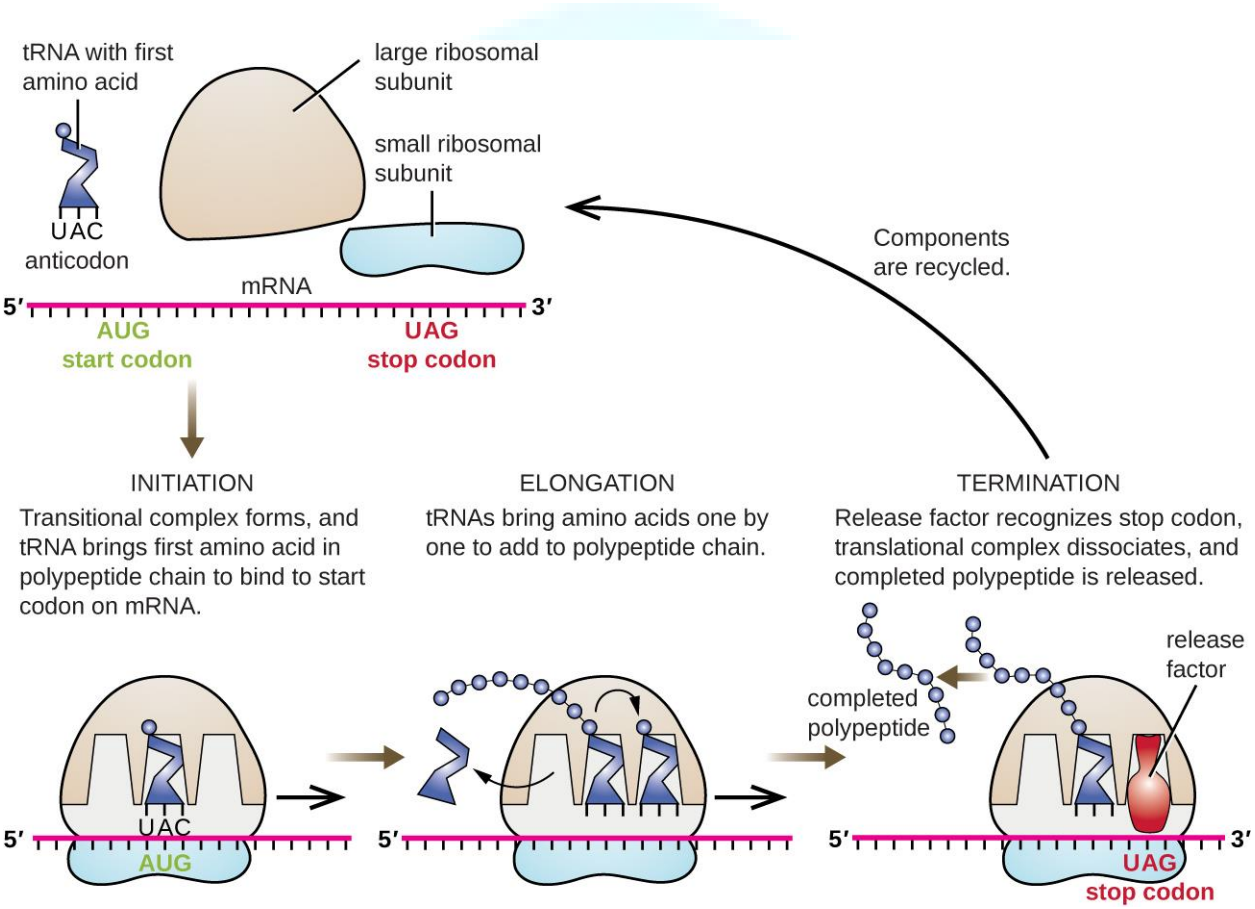


- The transfer RNA(tRNA) is a family of about 60 small sized ribonucleic acids that can recognize the codon of mRNA and exhibit a higher affinity for 21 activated amino acids which combine with them and carry them to the site of protein synthesis. tRNA molecules have been variously termed as soluble RNA or supernatant RNA or adapted RNA of the cell.
- Structurally, tRNA looks like a cloverleaf or inverted L shaped molecule which on one end has an amino acid receptor end and on the other end has an anticodon loop.
- The L shape results due to the modification in the nucleotides of tRNA such as pseudouridine, dihydrouridine(DHU), inosine and ribothymidine.
- The bent in the chain of each tRNA molecule contains a definite sequence of three nitrogenous bases that constitute the anticodon.
- It recognizes the codon on mRNA. The main constituents of tRNA are-
    - **Anticodon Loop:** It contains 7 bases out of which three bases form the anticodon loop and attaches to the codon of mRNA.
    - **DHU Loop:** This loop serves as the binding site for aminoacyl synthetase enzyme and it contains around 8-12 bases. The D arm contains the modified nucleotide called dihydrouridine.
    - **T Ψ C Loop:** This loop contains two modified nucleotides- pseudouridine and ribothymidine. This loop serves as the attachment site for ribosomes.

- **AA Binding Site:** This site serves as the binding site for amino acid. It contains a CCA- OH group.
- **Variable Loop:** It is generally present between the TΨC loop and anticodon loop.

The function of tRNA is specific in protein synthesis as they pick up specific amino acids from the amino acid pool and carry over the mRNA strand.

Protein Synthesis Steps Involved



The three stages of translation are-

- Initiation involves assembling ribosomes around mRNA and activating amino acid and delivering it to the transfer RNA.

- Elongation is the process in which the RNA strand gets longer by adding amino acids.
- The termination process only involves releasing a polypeptide chain.

Explanation of Steps of Translation

1.Initiation

- Initiation in prokaryotes requires large and small ribosome subunits, the mRNA, initiating transfer RNA, and 3 initiation factors (IFs).
- Amino acids are activated by binding with the enzyme called aminoacyl tRNA synthetase in presence of ATP forming an enzyme complex and P site.

Amino acid and ATP in the presence of aminoacyl transfer RNA synthetase Pi + AA-AMP-Enzyme complex

- Transfer of amino acid to tRNA -

AA-AMP-Enzyme complex + transfer RNA Amino Acid- tRNA + AMP + Enzyme.

- Two sites at ribosome are present that are called A-site and P-site where units of ribosome bind to the cap region of messenger RNA and comparatively smaller units bind to mRNA followed by binding of them with the larger subunits. It makes AUG lie on P-site and methionyl tRNA binds to P-site.

2.Elongation of the Polypeptide Chain

- At the 2nd codon, other aminoacyl transfer RNA complexes that are charged initiate binding at A-site.
- At P-site- peptide bond between the carboxyl molecule and the amino molecule is observed whereas at A-site bond between amino molecule

and amino acid is formed through the enzyme named as a peptidyl transferase.

- Sliding of ribosome over messenger RNA from one codon to its alternate codon in the direction of 5′ to 3`.
- A polypeptide chain is formed by the attachment of amino acids to one alternate to another in a chain formed by the peptide bond, and the attachment is based in accordance with the sequence of codons resulting in elongation of the protein chain.

3. Termination of Polypeptide

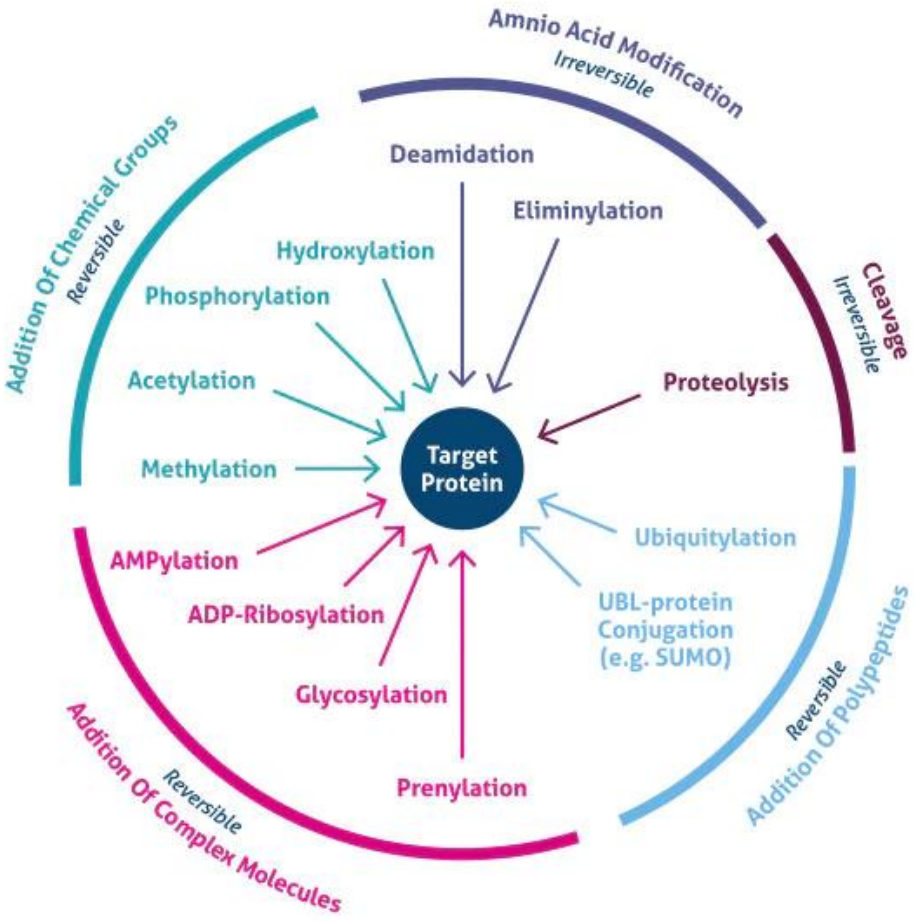
- Reaching the A-site of the ribosome at a termination codon which is present, not coding for any amino acid, no charged transfer RNA binds to the A-site of ribosome.
- A polypeptide is now not associated with the ribosome and dissociates and is catalyzed by a "release factor", a factor that releases 3 termination codons called UGA, UAG, and UAA.

Post-translational modification(PTM) of protein

- PTMs can happen at any step of the protein lifespan.
- Many proteins are modified shortly after translation is completed to mediate proper folding or to direct the nascent protein to distinct cellular locations (such as the nucleus or membrane).
- Other modifications occur after folding and localization are completed to activate or inactivate catalytic activity.
- Proteins are also covalently linked to tags that target a protein for degradation.
- They are modified through a combination of post-translational cleavage and the addition of functional groups through a step-wise mechanism of protein maturation or activation.

- PTMs occur at distinct amino acid side chains or peptide linkages and are most often mediated by enzymatic activity.
- Indeed, 5% of the proteome comprises enzymes that perform more than 200 types of PTMs (4).
- These enzymes include kinases, phosphatases, transferases, and ligases, which add or remove functional groups, proteins, lipids, or sugars to or from amino acid side chains, and proteases, which cleave peptide bonds to remove specific sequences or regulatory subunits.
- Many proteins can also modify themselves using autocatalytic domains, such as autokinase and autoprotolytic domains.
- PTMs can also be reversible based on the nature of the modification.
- As an example, phosphatases hydrolyze the phosphate group to remove it from the protein and reverse its biological activity.

Types of post-translational modifications (PTMs)



1. **Protein phosphorylation**

- Protein phosphorylation is the one of the most commonly occurring and most-studied post-translational modifications.
- It entails the phosphorylation of a specific amino acid residue through the addition of a phosphate group to a polar group R via a kinase, most commonly occurring at serine, tyrosine or threonine residues.
- The addition of the phosphate group results in a modification of the protein, whereby it transitions from being hydrophobic apolar to hydrophilic polar, enabling its interaction with other molecules – essentially "activating" it.
- A reversible post-translational modification, protein phosphorylation is important for cell regulation and the activation and deactivation of

enzymes and receptors, which can be implicated in disease processes such as cancer.

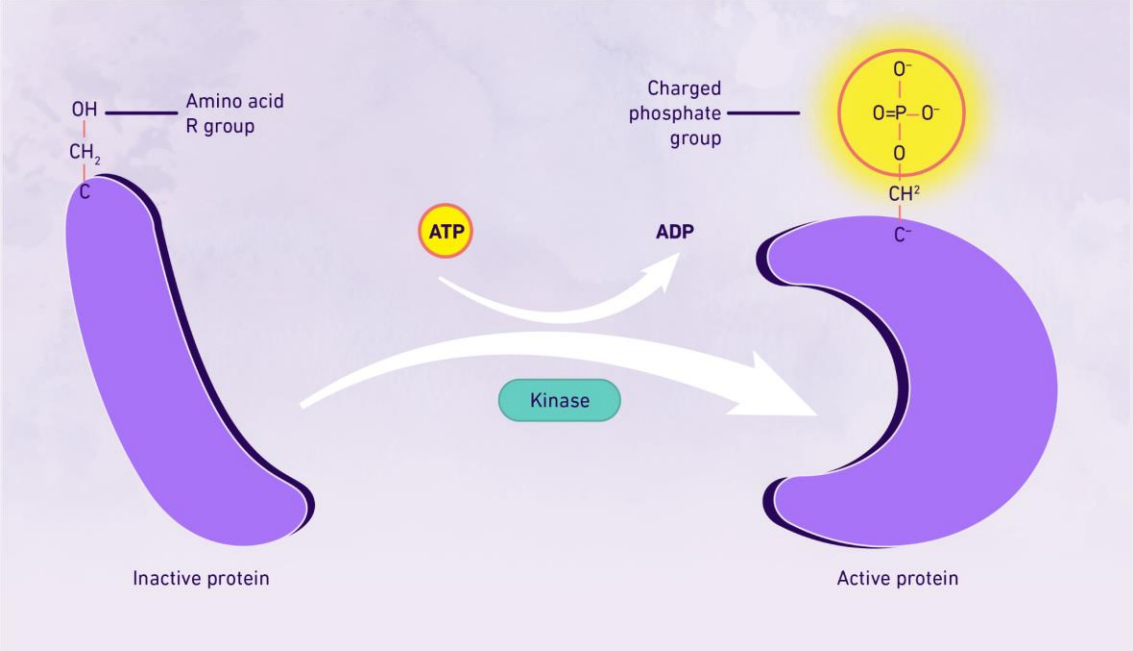


Fig: The key steps in protein phosphorylation

2. Protein glycosylation

- Protein glycosylation is recognized as one of the most “complicated" yet most commonly occurring post-translational modifications.
- It involves the covalent addition of a carbohydrate moiety to an amino acid, forming a glycoprotein.
- Glycosylation reactions are diverse and catalyzed by various different enzymes, which attach specific glycans to specific amino acids.
- Glycoproteins are estimated to make up ~50% of the proteome; however, the study of the glycoproteome is challenging because of the large number and diversity of glycoprotein isoforms.
- Glycosylation of eukaryotic proteins is usually categorized into two major types; N-linked, whereby a sugar molecule is attached to the amide nitrogen of asparagine, and O-linked, where a sugar molecule is attached to the oxygen atom of serine or threonine.

- There are an array of applications from glycoproteome research; many glycoproteins serve structural functions, whereas immunoglobulins are central to immunity and surface-presenting glycoproteins and glycolipids determine human blood group type.

3. Protein ubiquitination

- Ubiquitin is a small protein – approximately 8kDa in size – that can bind to a substrate protein in a process known as ubiquitination, a type of post-translation modification that serves to regulate a protein's function or mark it for degradation.

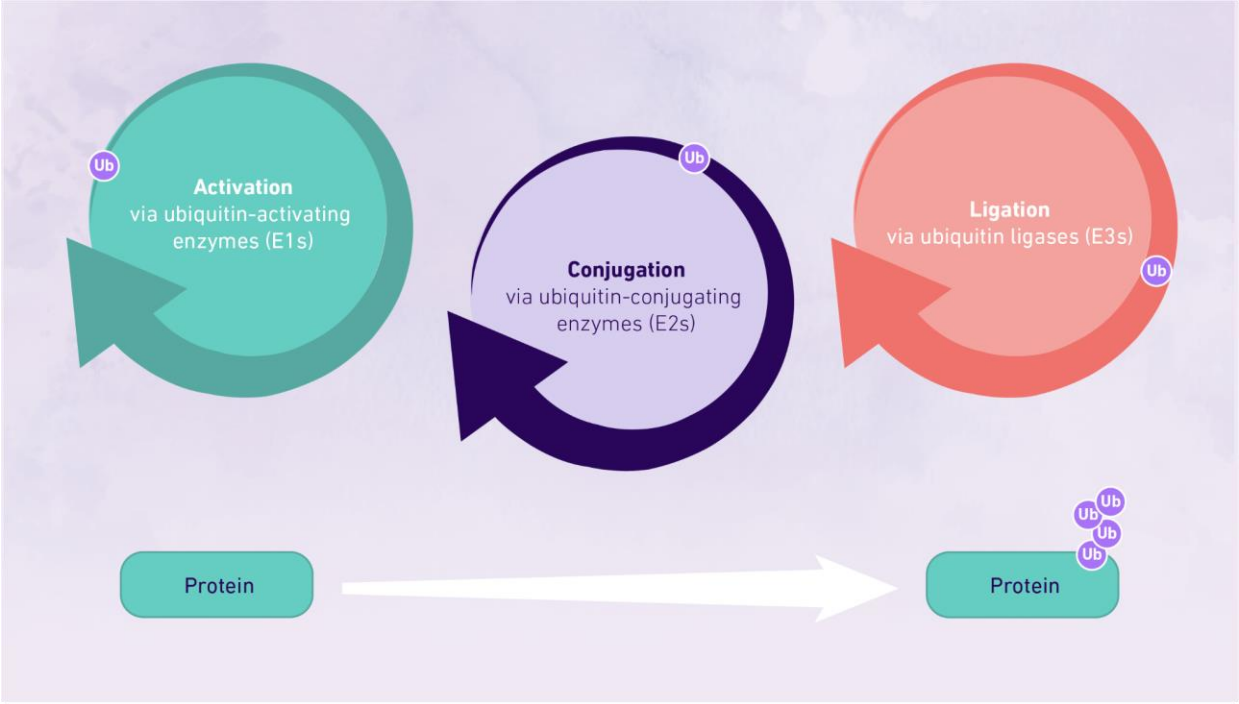


Fig: The key steps involved in protein ubiquitination

- Ubiquitination occurs in three sequential steps that are catalyzed by three groups of enzymes.
- This process generally culminates with an isopeptide bond forming between ubiquitin and the lysine residue of the protein substrate.
- Monoubiquitination refers to the addition of one ubiquitin molecule, whereas the addition of several ubiquitin proteins is known as polyubiquitination.

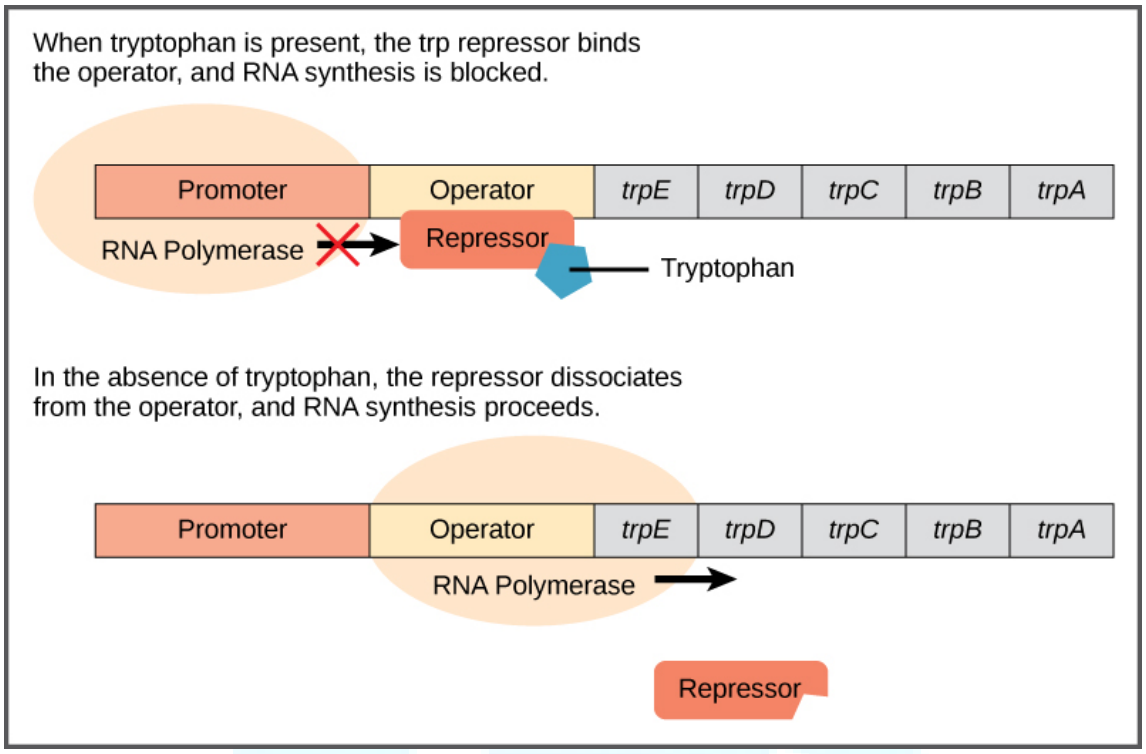
- Ubiquitination serves several functions, the most common being to flag proteins for degradation by the proteasome, but there are others including: immune and inflammatory response, organelle biogenesis and signaling roles in DNA repair.

Regulation of gene expression in bacteria. Operon model: lac operon, constitutive mutants, Catabolite repression.

- The DNA of prokaryotes is organized into a circular chromosome supercoiled in the nucleoid region of the cell cytoplasm.
- Proteins that are needed for a specific function, or that are involved in the same biochemical pathway, are encoded together in blocks called **operons.**
    - For example, all of the genes needed to use lactose as an energy source are coded next to each other in the lactose (or lac) operon.
- In prokaryotic cells, there are three types of regulatory molecules that can affect the expression of operons: repressors, activators, and inducers.
- **Repressors** are proteins that suppress transcription of a gene in response to an external stimulus, whereas **activators** are proteins that increase the transcription of a gene in response to an external stimulus.
- Finally, inducers are small molecules that either activate or repress transcription depending on the needs of the cell and the availability of substrate.

The trp Operon: A Repressor Operon

- Bacteria such as E. coli need amino acids to survive.
- Tryptophan is one such amino acid that E. coli can ingest from the environment.

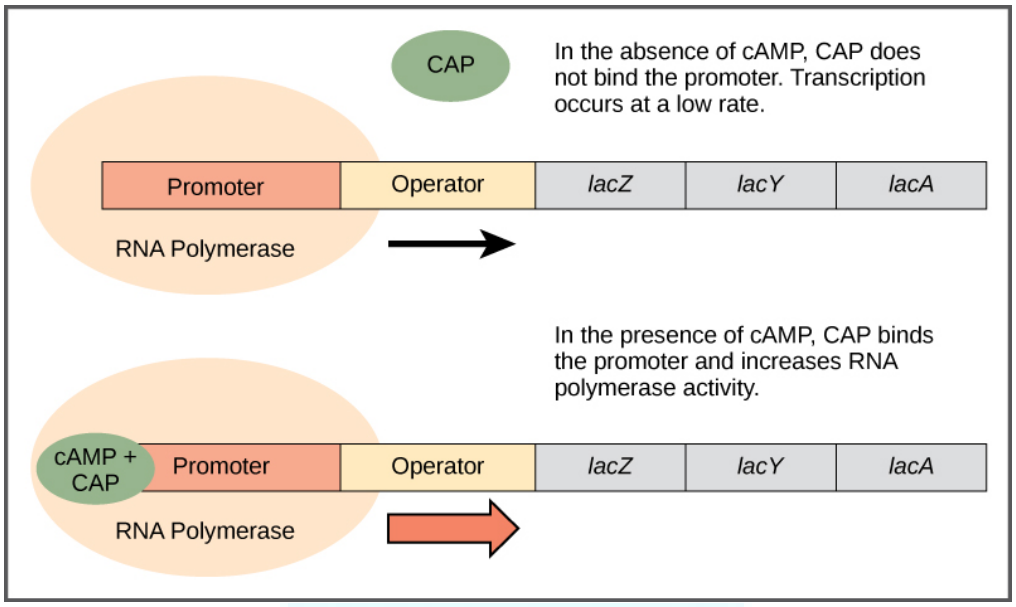


- E. coli can also synthesize tryptophan using enzymes that are encoded by five genes.
- These five genes are next to each other in what is called the **tryptophan (trp) operon.**
- If tryptophan is present in the environment, then E. coli does not need to synthesize it and the switch controlling the activation of the genes in the trp operon is switched off.
- However, when tryptophan availability is low, the switch controlling the operon is turned on, transcription is initiated, the genes are expressed, and tryptophan is synthesized.
- The five genes that are needed to synthesize tryptophan in E. coli are located next to each other in the trp operon.
- When tryptophan is plentiful, two tryptophan molecules bind the repressor protein at the operator sequence.
- This physically blocks the RNA polymerase from transcribing the tryptophan genes.

- When tryptophan is absent, the repressor protein does not bind to the operator and the genes are transcribed.
- A DNA sequence that codes for proteins is referred to as the coding region.
- The five coding regions for the tryptophan biosynthesis enzymes are arranged sequentially on the chromosome in the operon.
- Just before the coding region is the transcriptional start site.
- This is the region of DNA to which RNA polymerase binds to initiate transcription.
- The promoter sequence is upstream of the transcriptional start site; each operon has a sequence within or near the promoter to which proteins (activators or repressors) can bind and regulate transcription.
- A DNA sequence called the operator sequence is encoded between the promoter region and the first trp coding gene.
- This operator contains the DNA code to which the repressor protein can bind.
- When tryptophan is present in the cell, two tryptophan molecules bind to the trp repressor, which changes shape to bind to the trp operator.
- Binding of the tryptophan–repressor complex at the operator physically prevents the RNA polymerase from binding, and transcribing the downstream genes.
- When tryptophan is not present in the cell, the repressor by itself does not bind to the operator; therefore, the operon is active and tryptophan is synthesized.
- Because the repressor protein actively binds to the operator to keep the genes turned off, the trp operon is negatively regulated and the proteins that bind to the operator to silence trp expression are negative regulators.

Catabolite Activator Protein (CAP): An Activator Regulator

- Just as the trp operon is negatively regulated by tryptophan molecules, there are proteins that bind to the operator sequences that act as a positive regulator to turn genes on and activate them.
- For example, when glucose is scarce, E. coli bacteria can turn to other sugar sources for fuel.
- To do this, new genes to process these alternate genes must be transcribed.
- When glucose levels drop, cyclic AMP (cAMP) begins to accumulate in the cell.
- The cAMP molecule is a signaling molecule that is involved in glucose and energy metabolism in E. coli.
- When glucose levels decline in the cell, accumulating cAMP binds to the positive regulator catabolite activator protein (CAP), a protein that binds to the promoters of operons that control the processing of alternative sugars.
- When cAMP binds to CAP, the complex binds to the promoter region of the genes that are needed to use the alternate sugar sources.
- In these operons, a CAP binding site is located upstream of the RNA polymerase binding site in the promoter.
- This increases the binding ability of RNA polymerase to the promoter region and the transcription of the genes.
- When glucose levels fall, E. coli may use other sugars for fuel but must transcribe new genes to do so. As glucose supplies become limited, cAMP levels increase.
- This cAMP binds to the CAP protein, a positive regulator that binds to an operator region upstream of the genes required to use other sugar sources.



The lac Operon: An Inducer Operon

- The third type of gene regulation in prokaryotic cells occurs through inducible operons, which have proteins that bind to activate or repress transcription depending on the local environment and the needs of the cell.
- The lac operon is a typical inducible operon.
- As mentioned previously, E. coli is able to use other sugars as energy sources when glucose concentrations are low.
- To do so, the cAMP–CAP protein complex serves as a positive regulator to induce transcription.
- One such sugar source is lactose. The lac operon encodes the genes necessary to acquire and process the lactose from the local environment.
- CAP binds to the operator sequence upstream of the promoter that initiates transcription of the lac operon.
- However, for the lac operon to be activated, two conditions must be met.

- First, the level of glucose must be very low or non-existent. Second, lactose must be present.
- Only when glucose is absent and lactose is present will the lac operon be transcribed.
- This makes sense for the cell, because it would be energetically wasteful to create the proteins to process lactose if glucose was plentiful or lactose was not available.

Regulation of gene expression in eukaryotes

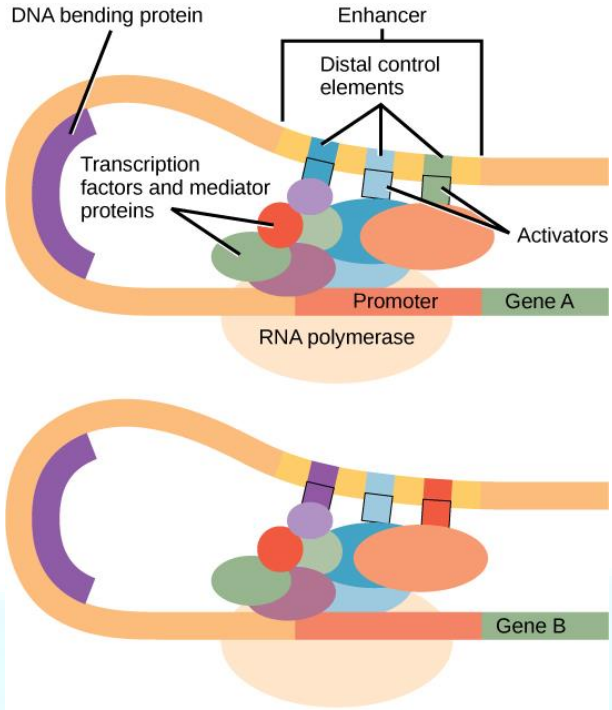
- Like prokaryotic cells, the transcription of genes in eukaryotes requires the actions of an RNA polymerase to bind to a sequence upstream of a gene to initiate transcription.
- However, unlike prokaryotic cells, the eukaryotic RNA polymerase requires other proteins, or transcription factors, to facilitate transcription initiation.
- Transcription factors are proteins that bind to the promoter sequence and other regulatory sequences to control the transcription of the target gene.
- RNA polymerase by itself cannot initiate transcription in eukaryotic cells.
- Transcription factors must bind to the promoter region first and recruit RNA polymerase to the site for transcription to be established.

The Promoter and the Transcription Machinery

- Genes are organized to make the control of gene expression easier.
- The promoter region is immediately upstream of the coding sequence.
- This region can be short (only a few nucleotides in length) or quite long (hundreds of nucleotides long).
- The longer the promoter, the more available space for proteins to bind.
- This also adds more control to the transcription process.

- The length of the promoter is gene-specific and can differ dramatically between genes.
- Consequently, the level of control of gene expression can also differ quite dramatically between genes.
- The purpose of the promoter is to bind transcription factors that control the initiation of transcription.
- Within the promoter region, just upstream of the transcriptional start site, resides the TATA box.
- This box is simply a repeat of thymine and adenine dinucleotides (literally, TATA repeats).
- RNA polymerase binds to the transcription initiation complex, allowing transcription to occur.
- To initiate transcription, a transcription factor (TFIID) is the first to bind to the TATA box.
- Binding of TFIID recruits other transcription factors, including TFIIB, TFIIE, TFIIF, and TFIIH to the TATA box.
- Once this complex is assembled, RNA polymerase can bind to its upstream sequence.
- When bound along with the transcription factors, RNA polymerase is phosphorylated.
- This releases part of the protein from the DNA to activate the transcription initiation complex and places RNA polymerase in the correct orientation to begin transcription; DNA-bending protein brings the enhancer, which can be quite a distance from the gene, in contact with transcription factors and mediator proteins.
- An enhancer is a DNA sequence that promotes transcription.
- Each enhancer is made up of short DNA sequences called **distal control elements.**
- Activators bound to the distal control elements interact with mediator proteins and transcription factors.

- Two different genes may have the same promoter but different distal control elements, enabling differential gene expression.



- In addition to the general transcription factors, other transcription factors can bind to the promoter to regulate gene transcription.
- These transcription factors bind to the promoters of a specific set of genes.
- They are not general transcription factors that bind to every promoter complex, but are recruited to a specific sequence on the promoter of a specific gene.
- There are hundreds of transcription factors in a cell that each bind specifically to a particular DNA sequence motif.
- When transcription factors bind to the promoter just upstream of the encoded gene, it is referred to as a **cis-acting element**, because it is on the same chromosome just next to the gene.
- The region that a particular transcription factor binds to is called the **transcription factor binding site**.

- Transcription factors respond to environmental stimuli that cause the proteins to find their binding sites and initiate transcription of the gene that is needed.

Enhancers and Transcription

- In some eukaryotic genes, there are regions that help increase or enhance transcription.
- These regions, called enhancers, are not necessarily close to the genes they enhance.
- They can be located upstream of a gene, within the coding region of the gene, downstream of a gene, or may be thousands of nucleotides away.
- Enhancer regions are binding sequences, or sites, for transcription factors.
- When a DNA-bending protein binds, the shape of the DNA changes.
- This shape change allows for the interaction of the activators bound to the enhancers with the transcription factors bound to the promoter region and the RNA polymerase.
- Whereas DNA is generally depicted as a straight line in two dimensions, it is actually a three-dimensional object.
- Therefore, a nucleotide sequence thousands of nucleotides away can fold over and interact with a specific promoter.

Turning Genes Off: Transcriptional Repressors

- Like prokaryotic cells, eukaryotic cells also have mechanisms to prevent transcription.
- Transcriptional repressors can bind to promoter or enhancer regions and block transcription.
- Like the transcriptional activators, repressors respond to external stimuli to prevent the binding of activating transcription factors.

Transcription factors

- In molecular biology, **a transcription factor (TF)** (or sequence-specific DNA-binding factor) is a protein that controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence.
- The **function of TFs** is to regulate—**turn on and off—genes** in order to make sure that they are expressed in the desired cells at the right time and in the right amount throughout the life of the cell and the organism.
- Groups of TFs function in a coordinated fashion to direct cell division, cell growth, and cell death throughout life; cell migration and organization (body plan) during embryonic development; and intermittently in response to signals from outside the cell, such as a hormone.
- There are up to 1600 TFs in the human genome. Transcription factors are members of the proteome as well as regulome.
- TFs work alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase (the enzyme that performs the transcription of genetic information from DNA to RNA) to specific genes.
- A defining feature of TFs is that they contain at least one DNA-binding domain (DBD), which attaches to a specific sequence of DNA adjacent to the genes that they regulate.
- TFs are grouped into classes based on their DBDs.
- Other proteins such as coactivators, chromatin remodelers, histone acetyltransferases, histone deacetylases, kinases, and methylases are also essential to gene regulation, but lack DNA-binding domains, and therefore are not TFs.
- TFs are of interest in medicine because TF mutations can cause specific diseases, and medications can be potentially targeted toward them.

Histone acetylation and deacetylation

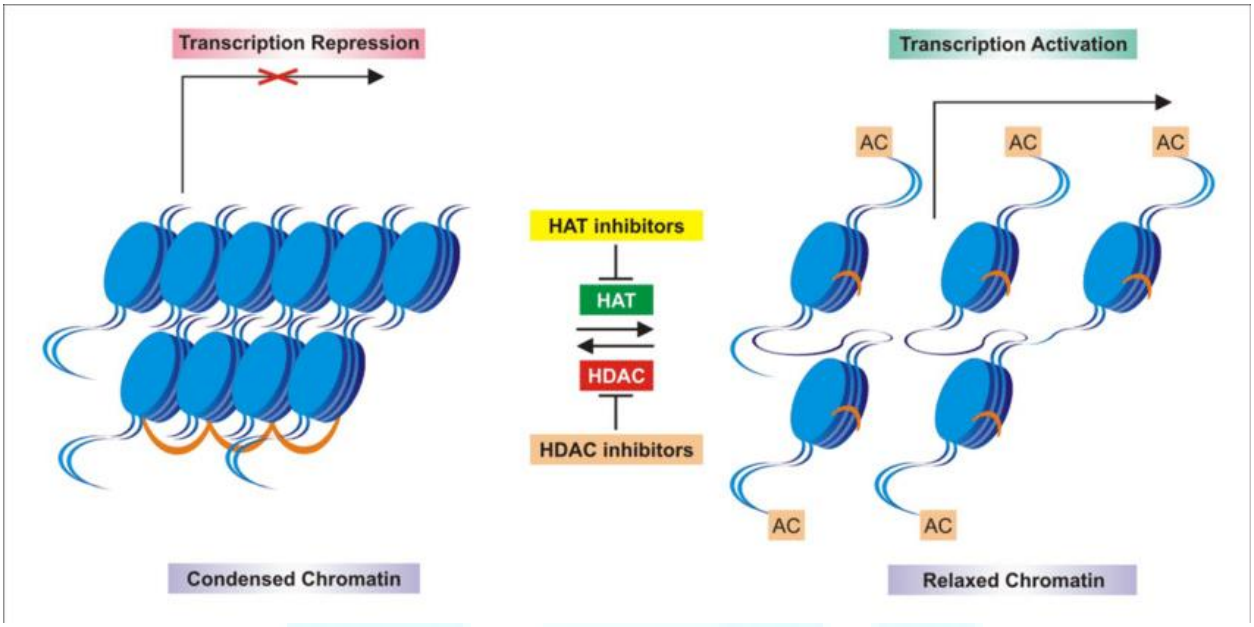
- **Histone acetylation and deacetylation** are the processes by which the lysine residues within the N-terminal tail protruding from the histone core of the nucleosome are acetylated and deacetylated as part of gene regulation.
- Histone acetylation and deacetylation are essential parts of gene regulation.
- These reactions are typically catalysed by enzymes with **"histone acetyltransferase" (HAT) or "histone deacetylase" (HDAC)** activity.
- **Acetylation** is the process where an acetyl functional group is transferred from one molecule (in this case, acetyl coenzyme A) to another.
- **Deacetylation** is simply the reverse reaction where an acetyl group is removed from a molecule.
- Acetylated histones, octameric proteins that organize chromatin into nucleosomes, the basic structural unit of the chromosomes and ultimately higher order structures, represent a type of epigenetic marker within chromatin.
- Acetylation removes the positive charge on the histones, thereby decreasing the interaction of the N termini of histones with the negatively charged phosphate groups of DNA.
- As a consequence, the condensed chromatin is transformed into a more relaxed structure that is associated with greater levels of gene transcription.
- This relaxation can be reversed by deacetylation catalyzed by HDAC activity.
- Relaxed, transcriptionally active DNA is referred to as euchromatin.
- More condensed (tightly packed) DNA is referred to as heterochromatin.
- Condensation can be brought about by processes including deacetylation and methylation.

Mechanism of action

- Nucleosomes are portions of double-stranded DNA (dsDNA) that are wrapped around protein complexes called histone cores.
- These histone cores are composed of 8 subunits, two each of **H2A, H2B, H3 and H4 histones.**
- This protein complex forms a cylindrical shape that dsDNA wraps around with approximately 147 base pairs.
- Nucleosomes are formed as a beginning step for DNA compaction that also contributes to structural support as well as serves functional roles.
- These functional roles are contributed by the tails of the histone subunits.
- The histone tails insert themselves in the minor grooves of the DNA and extend through the double helix, which leaves them open for modifications involved in transcriptional activation.
- **Acetylation** has been closely associated with increases in transcriptional activation while deacetylation has been linked with transcriptional deactivation. These reactions occur post-translation and are reversible.
- The mechanism for acetylation and deacetylation takes place on the NH_3^+ groups of lysine amino acid residues.
- These residues are located on the tails of histones that make up the nucleosome of packaged dsDNA.
- The process is aided by factors known as histone acetyltransferases (HATs).
- HAT molecules facilitate the transfer of an acetyl group from a molecule of **acetyl-coenzyme A (Acetyl-CoA)** to the NH_3^+ group on lysine.
- When a lysine is to be deacetylated, factors known as histone deacetylases (HDACs) catalyze the removal of the acetyl group with a molecule of H_2O.

- Acetylation has the effect of changing the overall charge of the histone tail from positive to neutral.

- Nucleosome formation is dependent on the positive charges of the H4 histones and the negative charge on the surface of H2A histone fold domains.
- Acetylation of the histone tails disrupts this association, leading to weaker binding of the nucleosomal components.



- By doing this, the DNA is more accessible and leads to more transcription factors being able to reach the DNA.
- Thus, acetylation of histones is known to increase the expression of genes through transcription activation.
- Deacetylation performed by HDAC molecules has the opposite effect.
- By deacetylating the histone tails, the DNA becomes more tightly wrapped around the histone cores, making it harder for transcription factors to bind to the DNA.
- This leads to decreased levels of gene expression and is known as gene silencing.

- Acetylated histones, the octomeric protein cores of nucleosomes, represent a type of epigenetic marker within chromatin.

- Studies have shown that one modification has the tendency to influence whether another modification will take place.
- Modifications of histones can not only cause secondary structural changes at their specific points, but can cause many structural changes in distant locations which inevitably affects function.
- As the chromosome is replicated, the modifications that exist on the parental chromosomes are handed down to daughter chromosomes.
- The modifications, as part of their function, can recruit enzymes for their particular function and can contribute to the continuation of modifications and their effects after replication has taken place.
- It has been shown that, even past one replication, expression of genes may still be affected many cell generations later.
- A study showed that, upon inhibition of HDAC enzymes by Trichostatin A, genes inserted next to centric heterochromatin showed increased expression.
- Many cell generations later, in the absence of the inhibitor, the increased gene expression was still expressed, showing modifications can be carried through many replication processes such as mitosis and meiosis.

Regulation at transcriptional and translational level

Transcriptional regulation

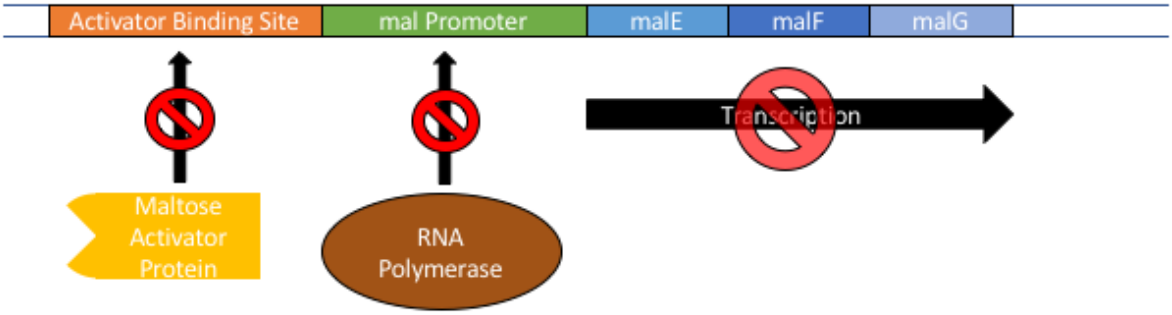
- In molecular biology and genetics, transcriptional regulation is the means by which a cell regulates the conversion of DNA to RNA (transcription), thereby orchestrating gene activity.
- A single gene can be regulated in a range of ways, from altering the number of copies of RNA that are transcribed, to the temporal control of when the gene is transcribed.
- This control allows the cell or organism to respond to a variety of intra- and extracellular signals and thus mount a response.

- **Some examples** of this include producing the mRNA that encode enzymes to adapt to a change in a food source, producing the gene products involved in cell cycle specific activities, and producing the gene products responsible for cellular differentiation in multicellular eukaryotes, as studied in evolutionary developmental biology.
- The regulation of transcription is a vital process in all living organisms.
- It is orchestrated by transcription factors and other proteins working in concert to finely tune the amount of RNA being produced through a variety of mechanisms.
- Bacteria and eukaryotes have very different strategies of accomplishing control over transcription, but some important features remain conserved between the two. Most importantly is the idea of combinatorial control, which is that any given gene is likely controlled by a specific combination of factors to control transcription.
- In a hypothetical example, the factors A and B might regulate a distinct set of genes from the combination of factors A and C.
- This combinatorial nature extends to complexes of far more than two proteins, and allows a very small subset (less than 10%) of the genome to control the transcriptional program of the entire cell.

In bacteria

- Much of the early understanding of transcription came from bacteria, although the extent and complexity of transcriptional regulation is greater in eukaryotes.
- Bacterial transcription is governed by **three main sequence elements**:
    - Promoters are elements of DNA that may bind RNA polymerase and other proteins for the successful initiation of transcription directly upstream of the gene.
    - Operators recognize repressor proteins that bind to a stretch of DNA and inhibit the transcription of the gene.

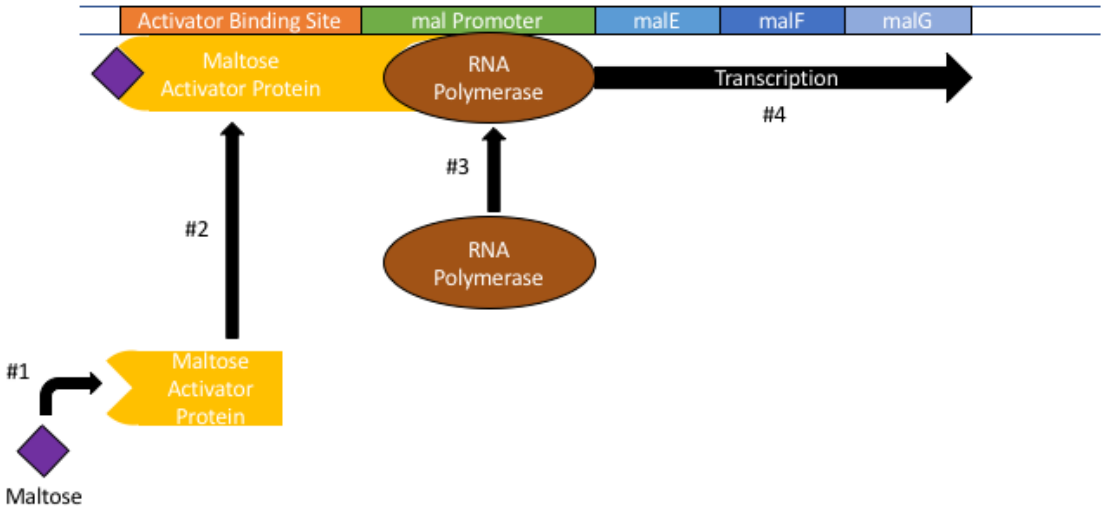
- Positive control elements that bind to DNA and incite higher levels of transcription.



- While these means of transcriptional regulation also exist in eukaryotes, the transcriptional landscape is significantly more complicated both by the number of proteins involved as well as by the presence of introns and the packaging of DNA into histones.
- The transcription of a basic bacterial gene is dependent on the strength of its promoter and the presence of activators or repressors.
- In the absence of other regulatory elements, a promoter's sequence-based affinity for RNA polymerases varies, which results in the production of different amounts of transcript.
- The variable affinity of RNA polymerase for different promoter sequences is related to regions of consensus sequence upstream of the transcription start site.
- The more nucleotides of a promoter that agree with the consensus sequence, the stronger the affinity of the promoter for RNA Polymerase likely is.
- In the absence of other regulatory elements, the default state of a bacterial transcript is to be in the “on” configuration, resulting in the production of some amount of transcript.
- This means that transcriptional regulation in the form of protein repressors and positive control elements can either increase or decrease transcription.

- **Repressors** often physically occupy the promoter location, occluding RNA polymerase from binding.
- Alternatively a repressor and polymerase may bind to the DNA at the same time with a physical interaction between the repressor preventing the opening of the DNA for access to the minus strand for transcription.
- This strategy of control is distinct from eukaryotic transcription, whose basal state is to be off and where co-factors required for transcription initiation are highly gene dependent.
- **Sigma factors** are specialized bacterial proteins that bind to RNA polymerases and orchestrate transcription initiation.
- Sigma factors act as mediators of sequence-specific transcription, such that a single sigma factor can be used for transcription of all housekeeping genes or a suite of genes the cell wishes to express in response to some external stimuli such as stress.

Figure 2: Maltose Present



- In addition to processes that regulate transcription at the stage of initiation, mRNA synthesis is also controlled by the rate of transcription elongation.

- RNA polymerase pauses occur frequently and are regulated by transcription factors, such as NusG and NusA, transcription-translation coupling, and mRNA secondary structure.

Translational regulation

- Translational regulation refers to the control of the levels of protein synthesized from its mRNA.
- This regulation is vastly important to the cellular response to stressors, growth cues, and differentiation.
- In comparison to transcriptional regulation, it results in much more immediate cellular adjustment through direct regulation of protein concentration.
- The corresponding mechanisms are primarily targeted on the control of ribosome recruitment on the initiation codon, but can also involve modulation of peptide elongation, termination of protein synthesis, or ribosome biogenesis.
- While these general concepts are widely conserved, some of the finer details in this sort of regulation have been proven to differ between prokaryotic and eukaryotic organisms.

In prokaryotes

- Initiation
    - Initiation of translation is regulated by the accessibility of ribosomes to the Shine-Dalgarno sequence.
    - This stretch of four to nine purine residues are located upstream the initiation codon and hybridize to a pyrimidine-rich sequence near the 3' end of the 16S RNA within the 30S bacterial ribosomal subunit.
    - Polymorphism in this particular sequence has both positive and negative effects on the efficiency of base-pairing and subsequent protein expression.

- Initiation is also regulated by proteins known as initiation factors which provide kinetic assistance to the binding between the initiation codon and tRNAfMet, which supplies the 3'-UAC-5' anticodon.
- **IF1** binds the 30S subunit first, instigating a conformational change[3] that allows for the additional binding of IF2 and IF3.
- **IF2** ensures that tRNAfMet remains in the correct position while IF3 proofreads initiation codon base-pairing to prevent non-canonical initiation at codons such as AUU and AUC.
- Generally, these initiation factors are expressed in equal proportion to ribosomes, however experiments using cold-shock conditions have shown to create stoichiometric imbalances between these translational machinery. In this case, two to three fold changes in expression of initiation factors coincide with increased favorability towards translation of specific cold-shock mRNAs.

- Elongation
  - Due to the fact that translation elongation is an irreversible process, there are few known mechanisms of its regulation.
  - However, it has been shown that translational efficiency is reduced via diminished tRNA pools, which are required for the elongation of polypeptides.
  - In fact, the richness of these tRNA pools are susceptible to change through cellular oxygen supply.

- Termination
  - The termination of translation requires coordination between release factor proteins, the mRNA sequence, and ribosomes. Once a termination codon is read, release factors RF-1, RF-2, and RF-3 contribute to the hydrolysis of the growing polypeptide, which terminates the chain.

- Bases downstream the stop codon affect the activity of these release factors.
- In fact, some bases proximal to the stop codon suppress the efficiency of translation termination by reducing the enzymatic activity of the release factors.
- For instance, the termination efficiency of a UAAU stop codon is near 80% while the efficiency of UGAC as a termination signal is only 7%.

In eukaryotes

- Initiation
  - When comparing initiation in eukaryotes to prokaryotes, perhaps one of the first noticeable differences is the use of a larger 80S ribosome.
  - Regulation of this process begins with the supply of methionine by a tRNA anticodon that basepairs AUG.
  - This base pairing comes about by the scanning mechanism that ensues once the small 40S ribosomal subunit binds the 5' untranslated region (UTR) of mRNA.
  - The usage of this scanning mechanism, in opposition to the Shine-Dalgarno sequence that was referenced in prokaryotes, is the ability to regulate translation through upstream RNA secondary structures.
  - This inhibition of initiation through complex RNA structures may be circumvented in some cases by way of internal ribosomal entry sites (IRESs) that localize pre-initiation complexes (PIC) to the start site.
  - In addition to this, the guidance of the PIC to the 5' UTR is coordinated by subunits of the PIC, known as **eukaryotic initiation factors (eIFs)**.

- When some of these proteins are down-regulated through stresses, translation initiation is reduced by inhibiting cap dependent initiation, the activation of translation by binding eIF4E to the 5' 7-methylguanylate cap.
- eIF2 is responsible for coordinating the interaction between the Met-tRNAiMet and the P-site of the ribosome.
- Regulation by phosphorylation of eIF2 is largely associated with the termination of translation initiation.
- Serine kinases, GCN2, PERK, PKR, and HRI are examples of detection mechanisms for differing cellular stresses that respond by slowing translation through eIF2 phosphorylation.

- Elongation
  - The hallmark difference of elongation in eukaryotes in comparison to prokaryotes is its separation from transcription.
  - While prokaryotes are able to undergo both cellular processes simultaneously, the spatial separation that is provided by the nuclear membrane prevents this coupling in eukaryotes.
  - **Eukaryotic elongation factor 2 (eEF2)** is a regulateable GTP-dependent translocase that moves nascent polypeptide chains from the A-site to the P-site in the ribosome.
  - Phosphorylation of threonine 56 is inhibitory to the binding of eEF2 to the ribosome.
  - Cellular stressors, such as anoxia have proven to induce translational inhibition through this biochemical interaction.

- Termination
  - Mechanistically, eukaryotic translation termination matches its prokaryotic counterpart.

- In this case, termination of the polypeptide chain is achieved through the hydrolytic action of a heterodimer consisting of release factors, **eRF1 and eRF3.**
- Translation termination is said to be leaky in some cases as noncoding-tRNAs may compete with release factors to bind stop codons.
- This is possible due to the matching of 2 out 3 bases within the stop codon by tRNAs that may occasionally outcompete release factor base pairing.
- An example of regulation at the level of termination is functional translational readthrough of the lactate dehydrogenase gene LDHB. This readthrough provides a peroxisomal targeting signal that localizes the distinct LDHBx to the peroxisome.

Antisense RNA strategies- siRNA, miRNA

- Antisense technology is a recent approach to specific modification or inhibition of gene expression in vitro or in vivo.
- It is a tool to study gene function and utilize it to manipulate the gene expression within cells to treat an endless number of diseases.
- The antisense approach utilizes **antisense agents** to alter the expression of viral genome inside the host cell or regulate the expression of specific genes that causes that particular disease.
- **Sense strand or sequence:** it is the coding strand within double-stranded DNA that carries the translatable code in the 5′ to 3′ direction. It is complementary to the template strand. The sense strand have sequences similar to that of mRNA.
- **Antisense strand:** it is the template strand of ds DNA, from which mRNA is transcribed. Thus, the antisense strand is complementary to mRNA.

- In simple term 'sense' refers to original sequence of DNA or RNA molecule and 'antisense' refers to Complementary copy.
- The antisense strand base pairs with its complementary mRNA strands (sense RNA) and thus prevents it from being translated into a protein.
- Therefore in antisense technology, the complementary nucleic acid sequence (antisense agents) is utilized to silence gene expression. The binding, or hybridization, of antisense nucleic acid sequences to a specific mRNA target will inhibit normal gene expression (ie. either interrupt transcription or translation) resulting in flow of message from DNA to protein.
- There are several mechanisms to interrupt the gene expression by antisense technology. Sometime the gene expression is completely inhibited known as knock-out and other time it is partially interrupted known as knock-down.
- **Some of the antisense agents are:**
    - **Antisense oligonucleotide:** like oligodeoxyribonucleotides (ODN) having less than 30 mucleotides or longer antisense RNA (a RNA) sequences
        - First generation
        - Second generation
        - Third generation
    - **Ribozymes**
    - **RNA interference (RNAi)**

1. **Antisense oligonucleotide:**

- Zamecnik and Stephenson first demonstrated the antisense effect of synthetic oligonucleotide
- Zamecnik and Stephenson identified a repeated sequence of 21 nucleotides that was crucial to viral integration with the help of nucleotide sequences from the 5' and 3' ends of the 35S RNA of Rous sarcoma virus (RSV).

- They synthesized a 13-mer oligonucleotide, d(AATGGTAAAATGG), complement to the portion of this viral sequence.
- Viral production got inhibited when synthetic oligonucleotide was introduced into cultured fibroblast cells. Thus, they concluded that oligonucleotide was inhibiting viral integration by hybridizing to the crucial sequences and blocking them. They introduced the term 'hybridon' to describe such oligonucleotides.
- At the same time, Tennant et al and Miller et al reported similar effects for synthetic oligonucleotides in other systems.
- **Criteria for successful oligonucleotide**
    - Specific target recognition by Watson-Crick pairing
    - Good structural Mimicry
    - Activation of RNaseH
    - Enhanced cellular uptake
    - Enhanced resistance to various nucleases: Synthetic oligonucleotides are foreign to the cells into which they are introduced and thus becomes prey for endogenous nucleases.
    - Synthetic oligonucleotides were protected from endogenous nuclease when they attained the persistence level in cell.
    - There are three possible sites on a nucleotide where protective modifications could be introduced.

Antisense Oligonucleotide modification: first, second and third generation

- The three possible sites for oligonucleotide modification are- at the position of Nitrogenous Base, Ribose sugar (2' OH group) and the Phosphate backbone.
- The main purpose of modification is to protect the antisense nucleotide from nuclease degradation when introduced inside cells.
- And at the same time it should be considered that the modification do not alter the inhibit hybridization ability of the antisense nucleotide.

First generation modification:

- The first generation antisense-motivated nucleotide modification is made by Eckstein and colleagues in the late 1960s by replacing one of the oxygen atom (non-bridging oxygen) of the phosphate backbone with a sulfur atom.
- This modified antisense agent is known as phosphorothioate.
- Phosphorothioate oligonucleotide is more nuclease resistant than the original oligonucleotide. The nuclease resistance was measured by an increased half-life for a phosphorothioated oligonucleotide upto ten hours in human serum as compared to that of one hour of an unmodified oligonucleotide having the same sequence.
- However, the phosphorothionated nucleotide displayed slight reduced hybridization ability and also a tendency to bind un-specifically to certain proteins in cell. High concentration of phosphrothionated nucleotide thus result in cytotoxicity.
- Matsukura and colleagues demonstrated that phosphorothioated oligonucleotides were effective hybridons against the HIV replication in the cultured cells.
- The first FDA-approved antisense drug is Vitravene from ISIS (Carlsbad, CA, USA.)

Second generation modification:

- The second generation modification focused on non-specific bind with certain proteins and cytotoxic effect of phosphorothioated nucleotides.
- In this modification, the antisense oligonucleotide undergoes alkyl modifications at the carbon no 2 (C2 position) of the ribose sugar.
- The two most important of these modifications are 2′-O-methyl and 2′-O-methoxy-ethyl at the C2 position.

- After alkylation at the C2 position of ribose sugar, the antisense oligonucleotides become resistant to nuclease degradation and shows low cytotoxicity effect.
- However, the antisense oligonucleotide with 2'-O-alkyl modification are unavailable for RNase H cleavage after hybridization with sense RNA.
- Since RNase H cleavage is the most desirable mechanism for antisense effect, A hybrid oligonucleotide is constructed containing both the desirable characteristics of nuclease resistance and RNase cleavage and it is known as gapmer antisense oligonucleotide.
- **Gapmer antisense oligonucleotide:**
    - This hybrid antisense oligonucleotide contains a central block of deoxynucleotides sufficient to induce RNase H cleavage flanked by blocks of 2'-O-methyl modified nuclease resistance ribonucleotide.

<u>Third generation modification:</u>

- Antisense oligonucleotide forms either DNA: DNA homo-duplex or DNA: RNA hetero-duplex depending upon the nature of oligonucleotide.
- The unmodified oligo-deoxynucleotides form such desired DNA: DNA or DNA:RNA duplexes.
- However variety of nucleic acid analogs have been developed having high affinity with target DNA or RNA and as a modification these analogs are utilized for antisense oligonucleotide construction.
- Some of these nucleic acid analogs are peptide nucleic acids (PNAs), 2'-fluoro N3-P5'-phosphoramidites, 1', 5'- anhydrohexitol nucleic acids (HNAs) and locked nucleic acids.

- Among these analogs locked nucleic acid (LNA) is very desirable as shows promising effect. The LNA is composed of nucleotides that is locked into a single conformation through a 2'-0', 4'-C methylene linkage in 1,2:5,6-di-O-isopropylene-α-allofuranose. LNA has increased the thermodynamic stability and enhanced nucleic acid recognition.

2. Ribozymes:

- Ribozymes are known as catalytic RNA, first described by Tom Cech (1982) from ribosomal RNA precursor from Tetrahymena thermophilia.
- Ribozymes acts as an enzymes to processes RNA precursors. And catalyze the modification or alteration of RNA or even DNA.

3. RNA Interference (RNAi):

- RNA interference (RNAi) was first described by Fire and colleagues in Caenorhabditis elegans.
- Several types of very short RNAs repress or silence the expression of genes and such silencing is known as RNA interference.
- RNA interference manifest in different ways; some time by inhibiting translation of mRNA and in other case by destruction of mRNA or silencing of promoter.

siRNA, miRNA

- Both siRNA and miRNA are proteomics tools used to study various aspects of gene expression. Proteomics is the study of proteins by which a cell's complete complement of proteins is examined at once. Technological advances have made such study possible.

- So are siRNA and miRNA similar or different? The jury is still somewhat out on that question, depending on whom you ask. Some sources feel that

siRNA and miRNA are the same things, while others indicate that they're separate entities entirely.

- The disagreement comes about because the two are both formed in the same manner.
- They emerge from longer RNA precursors. They're also both processed in the cytoplasm by an enzyme called **Dicer** before becoming part of the protein complex RISC.
- Enzymes are proteins that can improve the rate of reaction between biomolecules.

Difference between siRNA & miRNA

- The process of RNA interference (RNAi) can be moderated by either siRNA or miRNA, and there are subtle differences between the two.
- As mentioned, both are processed inside the cell by the enzyme Dicer and incorporated into the complex RISC.
- siRNA is considered exogenous double-stranded RNA that is taken up by cells. In other words, it enters through vectors, such as viruses.
- Vectors arise when geneticists use bits of DNA to clone a gene to produce a genetically modified organism (GMO).
- The DNA used in this process is called a vector.
- Although siRNA is thought to be exogenous double-stranded RNA, miRNA is single-stranded.
- It comes from endogenous noncoding RNA, meaning that it's made inside the cell.
- This RNA is found within the introns of larger RNA molecules.