Republic of Iraq Ministry of Higher Education and Scientific Research
University of Baghdad
College of Science
Department of Biology

Molecular Biology
2020-2021

ماجستير
الفصل الدراسي الثاني

مدرسي المادة

ا.د. خولة حوري زغير ا.م.د. انتصار جبار صاحب

Basic biology of nucleic acids

DNA

DNA, or deoxyribonucleic acid, has been described as the very "stuff of life." The central importance of DNA is that it is the genetic material—the substance that carries the information determining the properties of the organism. As we will discuss further, DNA is the master plan or program that specifies all the proteins that will be synthesized in an organism. As scientists you will need to have a good understanding of the structure and physical properties of DNA.

The discovery of DNA structure

DNA was first observed by a German biochemist named **Frederich Miescher** in 1869. But for many years, researchers did not realize the importance of this molecule. It was not until 1953 that **James Watson, Francis Crick, Maurice Wilkins and Rosalind Franklin** figured out the structure of DNA - a double helix - which they realized could carry biological information.

In the late 1940 the Austrian chemist **Erwin Chargaff'** state that DNA from any cell of any organisms should have a 1:1 ratio (base Pair Rule) of pyrimidine and purine bases and, more specifically, that the amount of guanine should be equal to cytosine and the amount of adenine should be equal to thymine. This pattern is found in both strands of the DNA.

In 1951 **Rosalind Franklin** used a technique called X-ray crystallography to find out the 3D shape of molecules. She applied this technique to different samples. Early in her career she worked on carbon and coal. Later she started working on biological subjects. She made major contributions to the discovery of the shape of DNA. After her work on this molecule, she also

gave new insights into the first virus that was ever discovered: The Tobacco Mosaic Virus. She thought the virus might be hollow and only consist of one strand of RNA. Although no proof existed at that time, she turned out to be right. Unfortunately, this was not confirmed until after her death.

The Hershey-Chase experiments were a series of experiments conducted in 1952 by **Alfred Hershey** and **Martha Chase** that helped to confirm that DNA is genetic material. While DNA had been known to biologists since 1869, many scientists still assumed at the time that proteins carried the information for inheritance because DNA appeared to be an inert molecule, and, since it is located in the nucleus, its role was considered to be phosphorus storage. In their experiments, Hershey and Chase showed that when bacteriophages, which are composed of DNA and protein, infect bacteria, their DNA enters the host bacterial cell, but most of their protein does not. Hershey and Chase and subsequent discoveries all served to prove that DNA is the hereditary material.

Hershey shared the 1969 Nobel Prize in Physiology or Medicine with Max Delbrück and Salvador Luria for their "discoveries concerning the genetic structure of viruses."

In 1953, Cambridge University scientists James D. Watson and Francis H.C. Crick announce that they have determined the double-helix structure of DNA, the molecule containing human genes. Watson, Crick and Wilkins were awarded the Nobel Prize in Medicine in 1962 "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material." Franklin was not included in the award, although her work was integral to the research.

I. DNA Structure

A. Nucleotides. DNA is a polymer of similar repeating units called **nucleotides**. The nucleotide is made up of three chemical entities: a five-carbon sugar called 2'-deoxyribose, a ring-shaped nitrogen-containing base, and a phosphate group (nucleosides do not have phosphate groups attached).

- ❖ Some things to notice about **the sugar**:

1. Ribose is one kind of five-carbon sugar. Sugars can circularize by eliminating water molecule and forming a bond between hydroxyl groups.
2. The carbons in the sugar are given numbers in a standard nomenclature. It is important that you know the structure and the numbering system of the sugar, because these numbers are used to distinguish important sites in the nucleotide and the DNA strand. (Numbers are indicated as "prime" to distinguish them from carbons in the bases, which are also numbered starting with 1.)

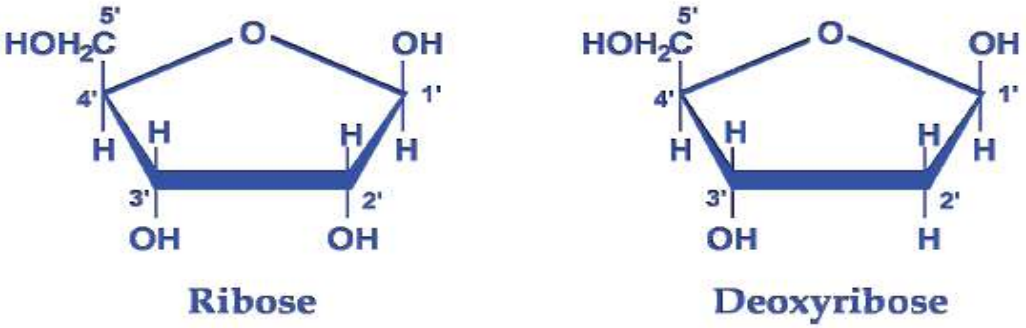


Fig (1) Difference between sugars of RNA (ribose) and DNA (deoxyribose).

3. The 2' position of deoxyribose is distinguished from ribose in that it lacks a hydroxyl (-OH) group on the 2' carbon. DNA is made up of nucleotides that lack the -OH group at this position. (In contrast, the sugar moiety of RNA contains an -OH at that position. This difference is responsible for important chemical differences between DNA and RNA.)
4. The -OH groups on the 5' and 3' carbons are important reactive groups through which nucleotides become joined. A nucleotide includes at least one phosphate group attached at the 5' position.

- Some things to notice about **the bases**:

1. There are four bases that are found in DNA: adenine (A), cytosine (C), guanine (G), and thymine (T).
2. Adenine and guanine are purines, they are two-ring structures. Thymine and cytosine are pyrimidines, they are single-ring structures.
3. The addition of a nitrogenous base to a ribose sugar results in the formation of a nucleoside.

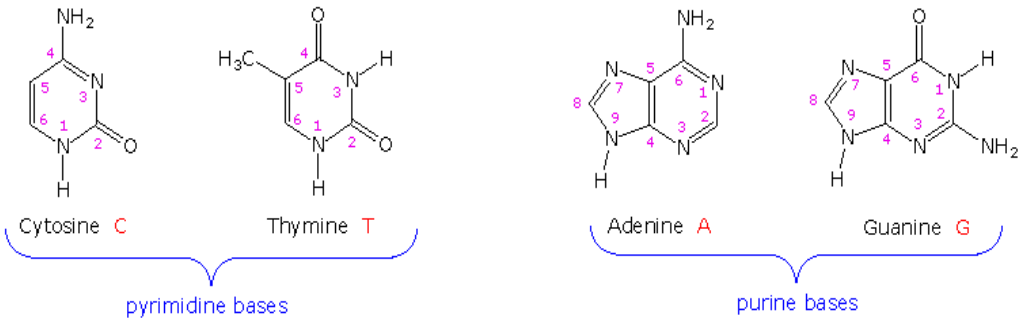


Fig (2) The four bases of DNA

❖ Some things to notice about **the phosphate**:

1. In a nucleotide the phosphate group is attached to the nucleoside via an ester linkage to the **5'** carbon.
2. Nucleotides can contain one or more phosphate groups. Thus, they can be designated as;
   i) Nucleoside monophosphates,
   ii) Nucleoside diphosphates, and
   iii) Nucleoside triphosphates.
3. The oxygen atoms on the phosphates are negatively charged at physiological pH. Thus, DNA carries a large net negative charge. This is an important physical property to remember.

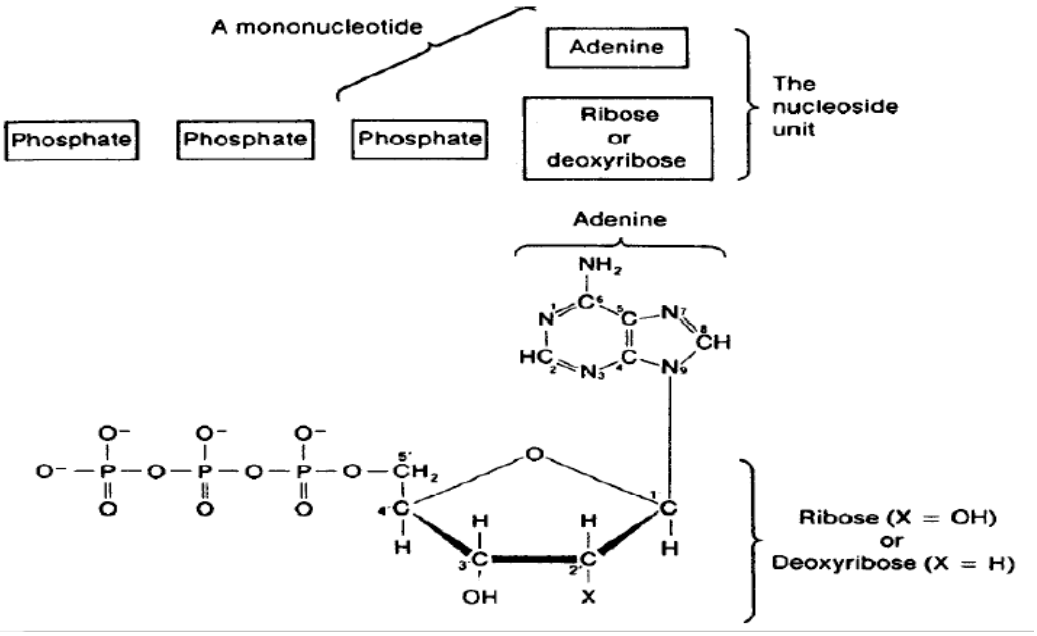


Fig (3) Nucleotide structure

B. The Polynucleotide Chain. To form the polynucleotide chain, nucleotide triphosphates are covalently linked together. In this reaction, the oxygen of the 3' hydroxyl group on the chain "attacks" the phosphate of a nucleotide triphosphate eliminating H2O and releasing the two outermost phosphate residues. The resulting bond between the sugar residues is called a phosphodiester bond. Notice that the DNA chain is synthesized in a 5' to 3' direction.

Bond formation:

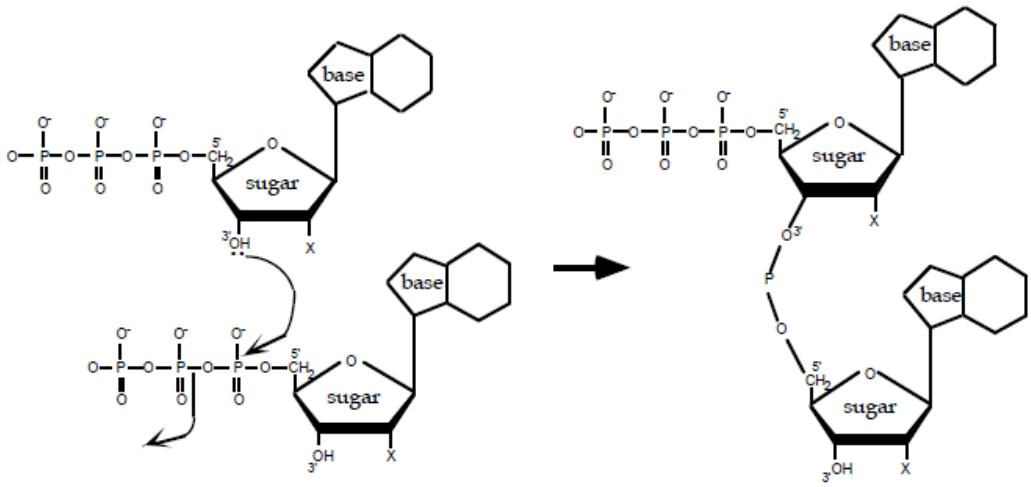


Fig (4) Formation of the phosphodiester backbone bond

An important thing to remember is that **there is an asymmetry to the DNA chain**. The 5' end is marked by a free phosphate and the 3'end is marked by a free hydroxyl group.

The example to the right is written as follows:

5' ATGC 3'

It should not be written as CGTA

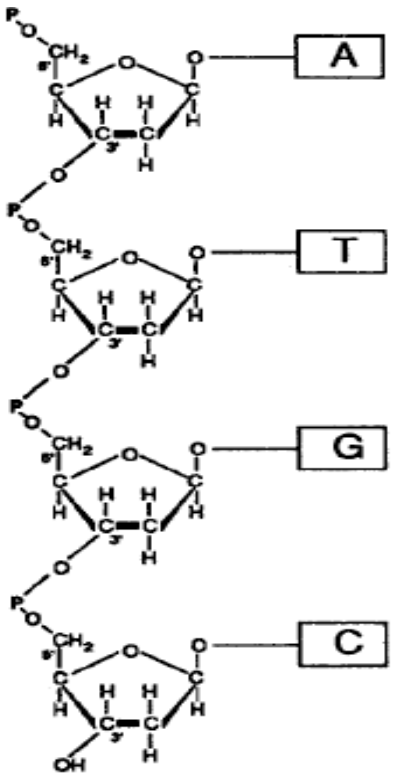


Fig (5) DNA chain

C. Structure of the Double Helix. DNA is "double-stranded." In other words, DNA is made up of two polynucleotide chains. These two strands are held together by hydrogen bonds that form between the two strands. Remember that only certain bases can pair well with other bases (one purine with one pyrimidine). **Standard Watson-Crick base pairing rules dictate that an A always pairs with T and G always pairs with C.** Note that A-T pairs (two hydrogen bonds) are held together less strongly that G-C pairs (three hydrogen bonds).

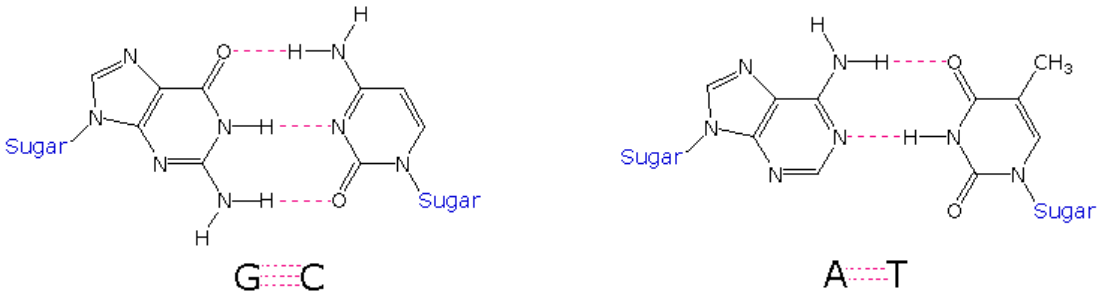
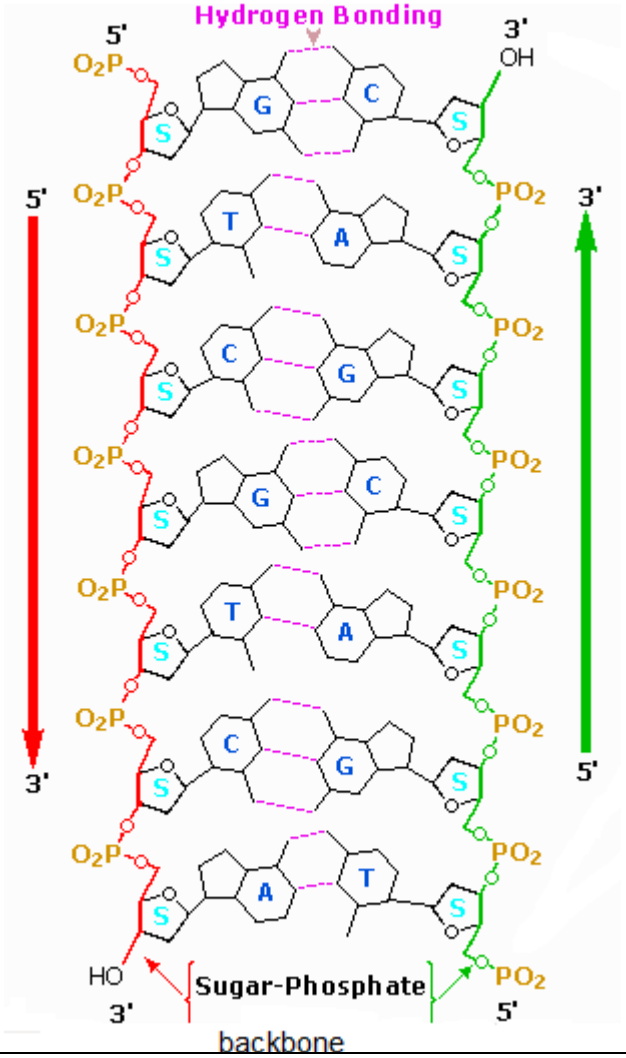


Fig (6) Formation of standard Watson-Crick base pairs in B-form DNA

Although DNA is a double helix, it is often depicted as a ladder. The base pairs make up the rungs of the ladder and the alternating sugar and phosphate groups make up the limbs of the latter. Unlike a ladder, however, DNA strands are arranged in an **antiparallel** manner. The 5' end of one strand is paired with the 3' end of the other strand. And, of course, the two strands of a DNA **(e.g., A pairs with T; G with C).** There are several different ways to depict the sequence to the right

5' GTCGTCA 3' 5' TGACGAC3'



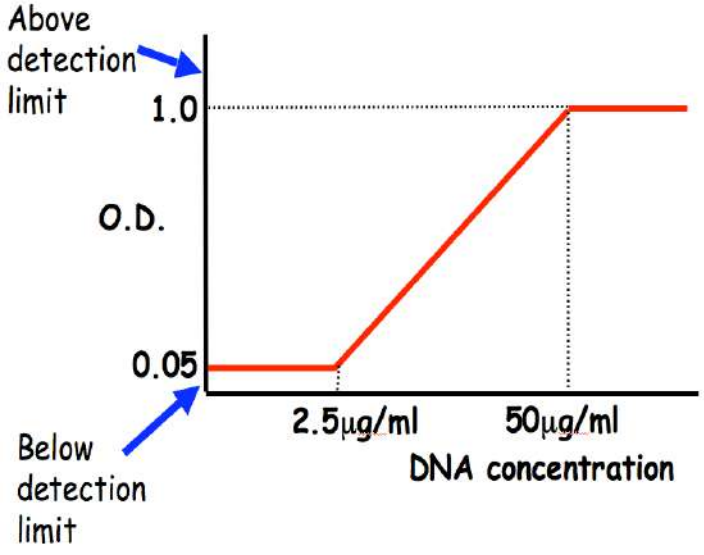
3' CAGCAGT 5' 3' ACTGCTG 5'

D. Some Useful Chemical Properties of DNA:

1. **DNA is negatively charged**. As noted above, the phosphate backbone confers a net negative charge on DNA, a property important for electrophoretic analysis.
2. **DNA can be denatured and renatured.** By raising the temperature or using chemical denaturants, it is possible to disrupt the hydrogen bonds between bases on different strands and "melt" the double helix into two polynucleotide strands. This process is called **denaturation**. The process is Reversible if conditions are favorable, the DNA strands can reanneal. **Renaturation** of two strands occurs efficiently only when the sequences of base pairs are complementary, i.e., when A-T and G-C pairs can form. Association of complementary single-stranded sequences is also called **nucleic acid hybridization**.
3. **DNA is soluble in water.** DNA is quite soluble in water and is generally kept as a buffered solution. Routinely, a buffer containing a small amount of the buffer Tris (to control pH) and the chelating agent EDTA (to trap cations which are cofactors for enzymes that can attack DNA) is used to solublize nucleic acids. **DNA is insoluble in ethanol.**

4. DNA absorbs ultraviolet light.

The purine and pyrimidine bases absorb light strongly in the ultraviolet end of the spectrum, with a maximum absorbance at 260 nm. Over a range of concentrations, the amount of light absorbed is proportional to the amount of DNA in solution. For double stranded DNA, an A_{260} reading of 1.0 corresponds to a concentration of 0.05 mg/ml.



Thus, by measuring the absorbance concentration, one can calculate the concentration of DNA in the solution.

Fig (8A): A plot of the Absorbance (OD) vs the concentration of DNA. Higher DNA concentrations absorb more light. The spectrophometers in the lab cannot accurately detect samples below 0.05 or above 1.0 OD.

5. DNA can be stained and quantified with ethidium bromide.

Ethidium bromide (EtBr) is a flat, ring-shaped molecule that can insert between base pairs of the double helix. This insertion, or intercalation, is useful for visualization of DNA because EtBr fluoresces when exposed to ultraviolet light. Thus, DNA, normally invisible, can be seen and photographed by staining with EtBr.

Additionally, because the amount of fluorescence is proportional to the total mass of DNA in a solution, the quantity of DNA can be estimated by comparing the fluorescent yield of the sample with that of a series of standards.

The intercalation of ethidium bromide into a DNA molecule is shown in the figure at the right. Note that the ethidium bromide increases the spacing of successive base pairs, distorts the regular sugar-phosphate backbone, and decreases the pitch of the helix.

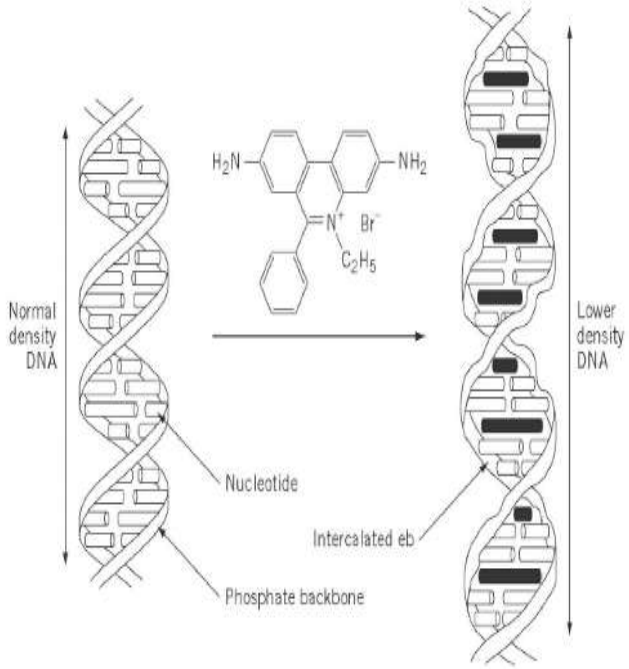


Fig (8b) Intercalation of EtBr into DNA

Variations in DNA structure:

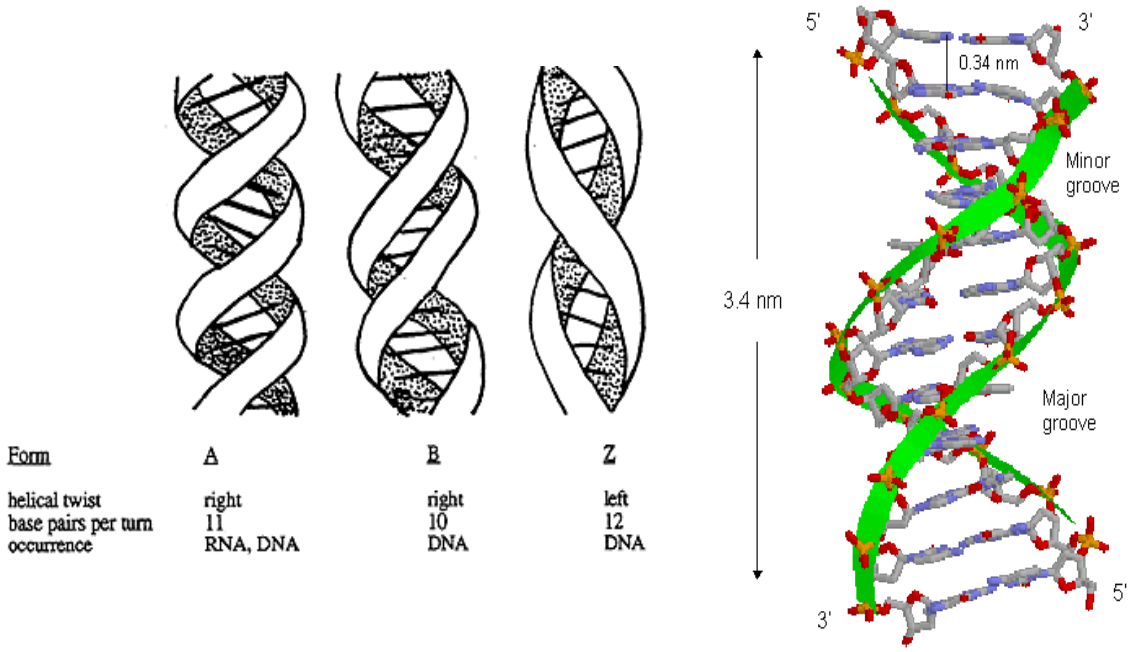
The structural model first proposed by Watson and Crick corresponds to the **B form** of DNA, which is typically prepared under conditions of **high humidity**. Under conditions of **partial dehydration**, however, DNA assumes an **A form**, in which the base pairs are no longer at right angles to the helical axis, but are tilted 20^o from the horizontal. The A form is observed when DNA is extracted

with solvents such as ethanol; the significance of the A form under physiological conditions is not currently apparent.

The **Z form**, named for its "zigzag" conformation, represents the most radical departure from the B form. DNA segments composed of alternating purine and pyrimidine bases, especially CGCGCG, are most likely to adopt the Z configuration. The table below lists those characteristics which serve to differentiate A, B and Z forms of DNA.

DNA Forms

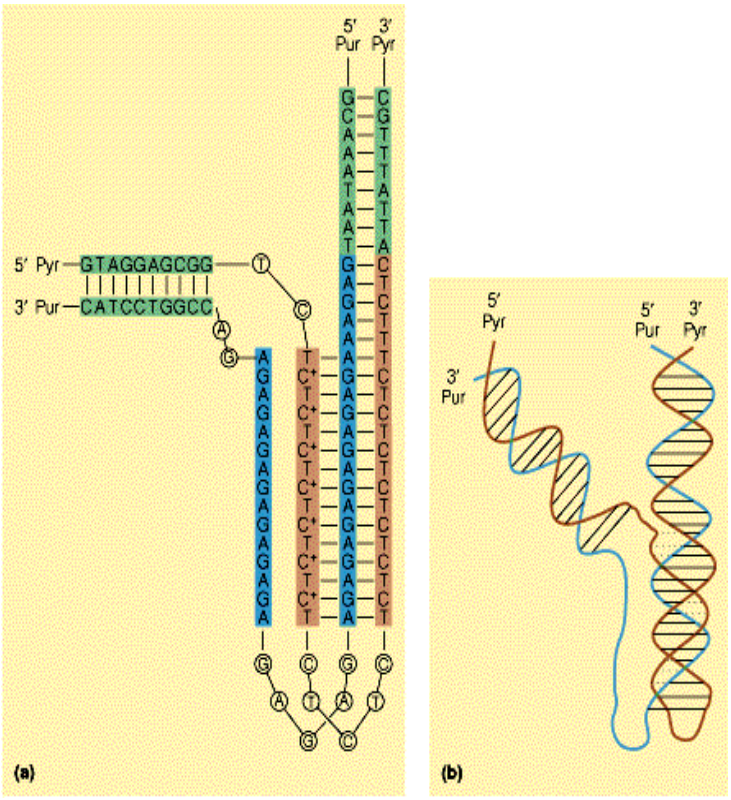
Characteristic	A-DNA	B-DNA	Z-DNA
diameter (D)	2.6 nm	2.0 nm	1.8 nm
bp/turn	11	10.4	12
degrees rotation/bp	+32.7	+34.6	-30.0
axial distance/turn	2.8 nm	3.4 nm	4.5 nm
axial distance between bp	0.25 nm	0.33 nm	0.38 nm

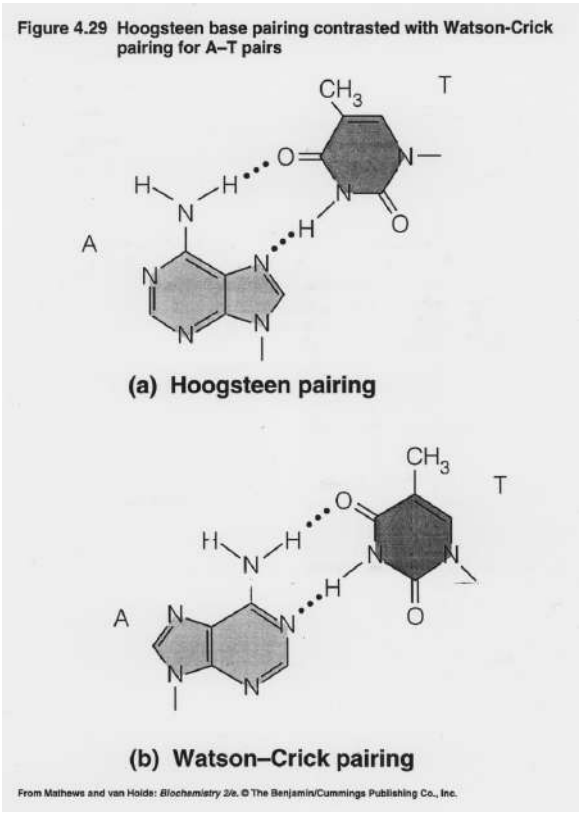


H-DNA: Under physiological conditions, e.g., low pH, a DNA segment consisting of a long polypurine strand base-paired to a polypyrimidine strand can form a triple helix. The formation of this triple helix, termed H-DNA, it was described as H form, because it was stabilized by hydrogen ions (it was a

protonated structure). This H form consisted of an intramolecular triple helix formed by the pyrimidine strand and half of purine strand, the other half of purine strand being single stranded.

Its formation depends on the phenomenon of nonconventional base-pairing (Hoogsteen base-pairing), which occurs without disruption of Watson-Crick base pairs. H-DNA formation has been postulated to have a role, not yet clarified, in gene expression.





H- DNA **Hoogsteen Base-pairing**

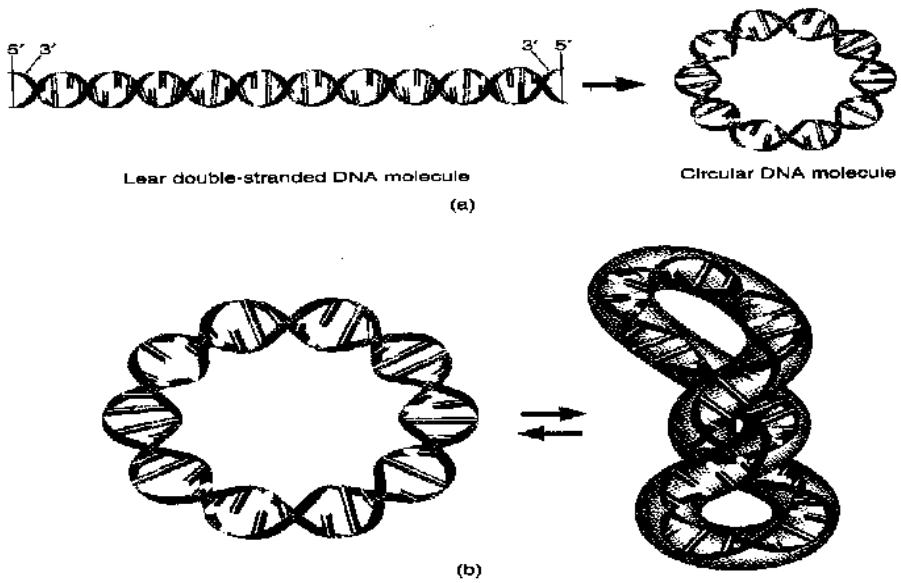
Supercoiled DNA:

A circular DNA molecule having the characteristics of B-DNA, i.e.,10.4 bp/turn, is termed "relaxed". The properties of such closed helical circles can be described by several quantitative parameters. The number of times each strand crosses the other is called the linking number (L). In a circular DNA molecule, L is equal to the number of base pairs per helical turn. The number of helical turns that occur in such a molecule is defined as twist (T).

Example, if there are 260 bp in a relaxed, circular DNA molecule, then L = 26 (i.e. 260/10) and T = 26.

A DNA molecule's linking number can change only if a covalent bond in the backbone of one or both strands is broken and subsequently resealed. Twist, however, can vary freely, if the relaxed circular DNA molecule is held and twisted a few times, it will adopt the shape shown in the figure below (panel b). When this twisted molecule is laid back on the flat surface, it will rotate so as to eliminate the twist. However, if this molecule is cut before it is twisted (e.g., by a topoisomerase), this "operation" changes both the linking number, as well as another parameter referred to as the writhing number (W). The writhing, or supercoiling, number is defined as the number of superhelical turns that occurs in a molecule's three-dimensional conformation.

The circular DNA molecule can be twisted either to the right or to the left before its strands are resealed. If it is twisted to the left, the L number decreases, and the molecules conformation is underwound, (negatively supercoiled), (Note that negative supercoiling, which causes a decrease in a DNA molecule's T value, promotes the formation of Z-DNA).

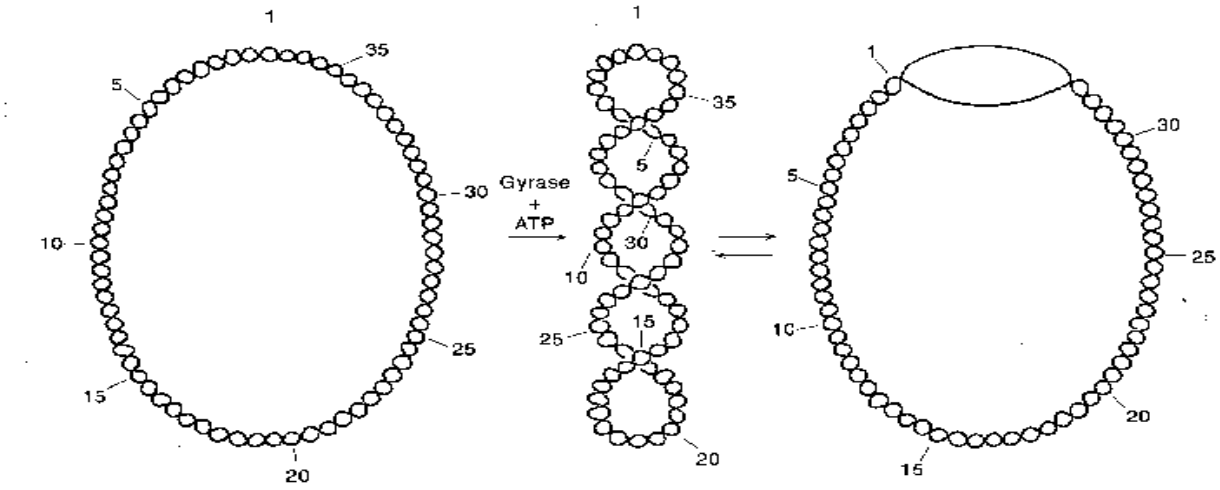


Linear, Circular and Supercoiled DNA

If, instead, the molecule is twisted to the right, it becomes overwound (the L number is higher), or positively supercoiled. All naturally occurring

superhelical DNA molecules are initially underwound, i.e., form negative superhelices. Also, the degree of twisting (superhelical density) is about the same for all molecules, namely, one negative twist per 200 base pairs or, 0.05 twist per helical turn [1 twist / (20 turns / twist) = 0.05 twist].

In addition to forming a supercoil, a strained circular negatively supertwisted DNA molecule could remain flat. If it does so, however, the T value of the molecule decreases. In this case, the strain introduced by the negative supercoiling process is relieved by a partial unwinding of the double helix (figure below). In cells, the formation of a supercoil requires less energy than the disruption of base pairs. However, cells apparently underwind DNA because it facilitates strand separation (e.g., during DNA replication and transcription), a process that is regulated by specific enzymes. DNA negative supercoiling also explains the propensity of certain DNA sequences to form cruciform and H-DNA.



		(1) Relaxed	(2) Strained: supertwisted	(3) Strained: disrupted base pairs
Base pairs	(bp):	360	360	360
Linking number	(*L*):	36	32	32
Twist	(*T*):	36	36	32
Supercoiling	(*S*):	0	−4	0

Lec. 2 Chromosomes

In **eukaryotes**, chromosomes consist of a **single molecule of DNA** associated with:

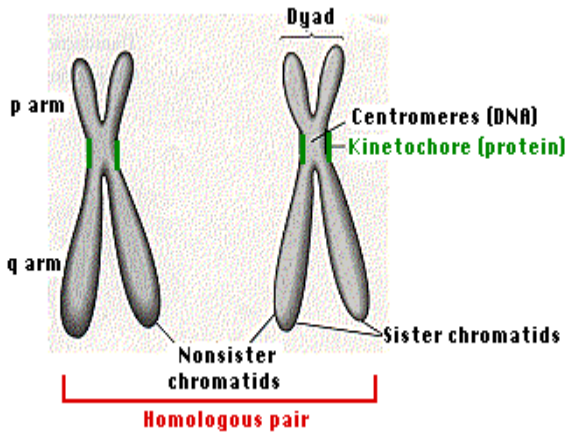
- Many copies of 5 kinds of **histones**. Histones are proteins rich in **lysine** and **arginine** residues and thus positively-charged. For this reason they bind tightly to the negatively-charged phosphates in DNA.
- A small number of copies of many different kinds of **non-histone proteins**. Most of these are **transcription factors** that regulate which parts of the DNA will be transcribed into RNA.

Examples of Chromosome Organization

I. Single stranded circular - Bacterial Xm, as well as ϕX174 phage.
II. Double stranded circular – various organisms.
III. Double stranded linear (i.e. phage T4).

Structure

- For most of the life of the cell, chromosomes are too elongated and tenuous to be seen under a microscope.
- Before a cell gets ready to divide by mitosis, each chromosome is duplicated (during **S phase** of the cell cycle).
- As mitosis begins, the duplicated chromosomes condense into short (~ 5 μm) structures which can be stained and easily observed under the light microscope. These duplicated chromosomes are called **dyads**.



- When first seen, the duplicates are held together at their **centromeres**. In humans, the centromere contains 1–10 million base pairs of DNA. Most of this is repetitive DNA: short sequences (e.g., 171 bp) repeated over and over in tandem arrays.
- While they are still attached, it is common to call the duplicated chromosomes **sister chromatids**, but this should not obscure the fact that each is a bona fide chromosome with a full complement of genes.
- The **kinetochore** is a complex of >80 different proteins that forms at each centromere and serves as the attachment point for the spindle fibers that will separate the sister chromatids as mitosis proceeds into anaphase.
- The shorter of the two arms extending from the centromere is called the **p arm**; the longer is the **q arm**.
- Staining with the trypsin-giemsa method reveals a series of alternating light and dark bands called **G bands**.
- G bands are numbered and provide "addresses" for the assignment of gene loci.

Chromosome Numbers

All animals have a characteristic number of chromosomes in their body cells called the **diploid** (or **2n**) number.

These occur as **homologous pairs**, one member of each pair having been acquired from the gamete of one of the two parents of the individual whose cells are being examined.

The gametes contain the **haploid** number (**n**) of chromosomes.

Diploid numbers of some commonly studied organisms

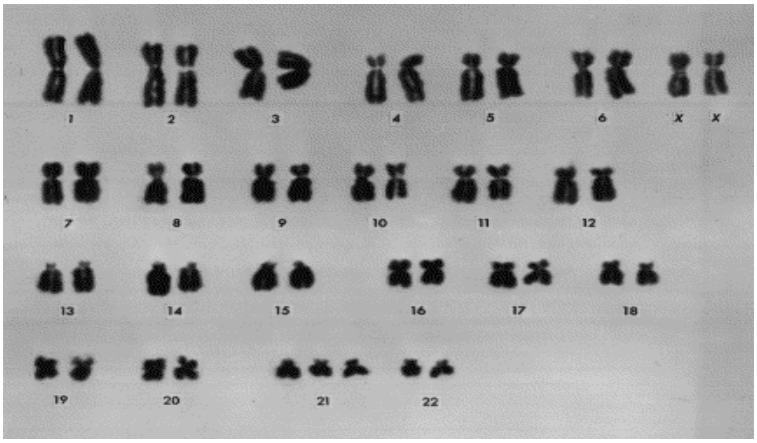
Homo sapiens (human)	46
Mus musculus (house mouse)	40
Drosophila melanogaster (fruit fly)	8
Caenorhabditis elegans (microscopic roundworm)	12
Saccharomyces cerevisiae (budding yeast)	32
Arabidopsis thaliana (plant in the mustard family)	10
Xenopus laevis (South African clawed frog)	36
Canis familiaris (domestic dog)	78
Gallus gallus (chicken)	78
Zea mays (corn or maize)	20
*Muntiacus reevesi (*the Chinese muntjac, a deer)	23
Muntiacus muntjac (its Indian cousin)	6
Myrmecia pilosula (an ant)	2
Parascaris equorum (parasitic roundworm)	2
Cambarus clarkii (a crayfish)	200
Equisetum arvense (field horsetail, a plant)	216

Karyotype:

The complete set of chromosomes in the cells of an organism is its **karyotype**. It is most often studied when the cell is at metaphase of mitosis and all the chromosomes are present as dyads. The karyotype of the human female contains:

Karyotype	Male	Female
Autosomes	22 pairs of autosomes	22 pairs of autosomes
Sex chromosome	one X chromosome one Y chromosome	1 pair of X chromosomes

(A gene on the Y chromosome designated SRY is the master switch for making a male.)



Above is a human karyotype (of which sex?). It differs from a normal human karyotype in having an extra #21 dyad. As a result, this individual suffered from a developmental disorder called **Down syndrome**. The inheritance of an extra chromosome is called **trisomy**, in this case trisomy 21. It is an example of **aneuploidy**.

Translocations

Karyotype analysis can also reveal translocations between chromosomes. A number of these are associated with cancers, for example

- The Philadelphia chromosome formed by a translocation between chromosomes 9 and 22 and a cause of Chronic **Myelogenous Leukemia** (CML).
- A translocation between chromosomes 8 and 14 that causes **Burkitt's lymphoma**.
- A translocation between chromosomes 18 and 14 that causes **B-cell leukemia**.

Human Chromosomes

I. Contour length of all DNA in a single cell: 2 meters! (*E.coli* length = 1.7mm)
II. DNA length in adult human body: 200,000,000,000 km (Earth circumference= 40,000 km)
III. Therefore genetic info is **phenomenally well packed**!

Other site for DNA in eukaryotic cells: **mitochondria** mDNA (mDNA codes for tRNA and rRNA's).

♣ Chromosome Physical Information:

- ½ weight is DNA
- ½ weight is protein: histone (H1, H2A, H2B, H3, H4).

  histone proteins is rich in arginine and lysine
- Non histone protein

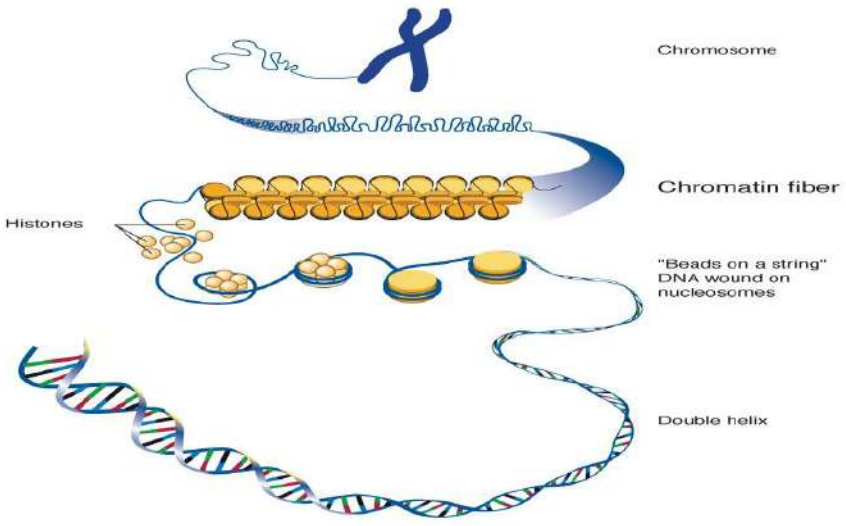
♣ Structure of Chromosomes

- "Chromatin" = chromosomal material.
- "Histone Complex" called "Nucleosome".
- ~200bp of DNA is wrapped around each histone complex.
- Histones (small basic proteins) package and order the DNA complex

Eukaryotic Chromosome has 3 Area types:

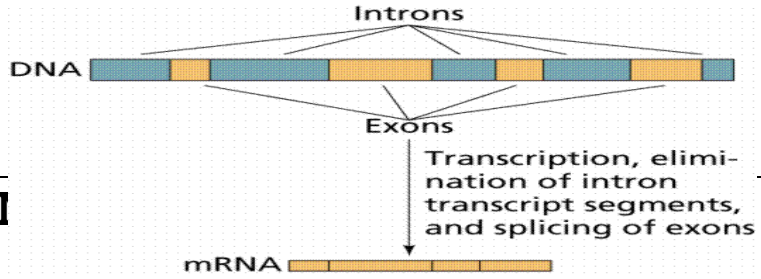
- Highly repetitive (10%), Moderately repetitive (20%), Unique segments (70%)

This is the overall relationship between chromosomal material and relaxed DNA

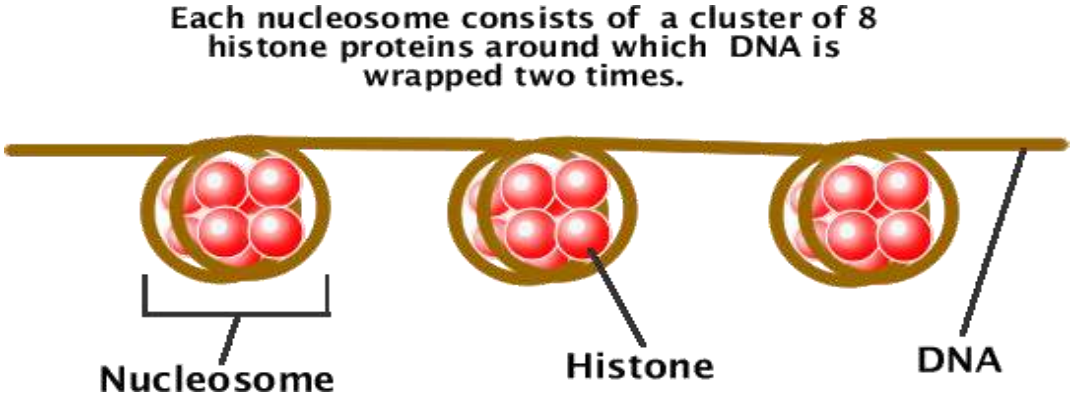


♣ Eukaryotic Gene Features

1. Have some "strange" features
2. Not all nucleotides code for amino acids
3. There are non-translated sequences called **Introns** or Intervening sequences.
4. Coding sequences are called **Exons** (amount of Exons is Variable).



A **nucleosome** is the basic unit of DNA packaging in eukaryotes, consisting of a segment of DNA wound in sequence around four histone protein cores. This structure is often compared to thread wrapped around a spool.



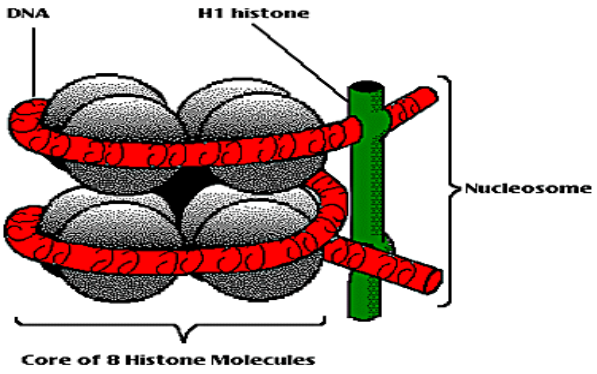
- Nucleosomes form the fundamental repeating units of eukaryotic chromatin, which is used to pack the large eukaryotic genomes into the nucleus while still ensuring appropriate access to it (in mammalian cells approximately 2 m of linear DNA have to be packed into a nucleus of roughly 10 µm diameter).
- Nucleosomes are folded through a series of successively higher order structures to eventually form a chromosome; this both compacts DNA and creates an added layer of regulatory control, which ensures correct gene expression.
- Nucleosomes are thought to carry epigenetically inherited information in the form of covalent modifications of their core histones.

- The nucleosome hypothesis was proposed by **Don** and **Ada Olins** in 1974 and **Roger Kornberg.**

The nucleosome core particle consists of approximately **147 base pairs** of DNA wrapped in **1.67 left-handed superhelical turns** around a **histone octamer** consisting of **2 copies** each of the core histones **H2A, H2B, H3,** and **H4**.

Core particles are connected by stretches of "linker DNA", which can be up to about **80 bp long**. Technically, a nucleosome is defined as the core particle plus one of these linker regions; however the word is often synonymous with the core particle. Genome-wide nucleosome positioning maps are now available for many model organisms including mouse liver and brain.

Linker histones such as **H1** and its isoforms are involved in chromatin compaction and sit at the base of the nucleosome near the DNA entry and exit binding to the linker region of the DNA. Non-condensed nucleosomes without the linker histone resemble "beads on a string of DNA" under an electron microscope.



Nucleosome

In contrast to most eukaryotic cells, mature sperm cells largely use protamines to package their genomic DNA, most likely to achieve an even higher packaging ratio. Histone equivalents and a simplified chromatin structure have also been found in Archea, suggesting that **eukaryotes are not the only organisms that use nucleosomes.**

Centromere

- Mitosis is the process by which a eukaryotic cell divides to produce two daughter cells that each contains the same number of chromosomes as the parent cell.
- The overall process of mitosis fails if the parent cell's chromosomes don't reach their correct destinations. One structure that plays a critical role in ensuring that this does not occur is the centromere.
- The centromere was first described by German biologist Walter Flemming in the 1880s as the "primary constriction" of the chromosome. Scientists now appreciate that the centromere is a region of specialized chromatin found within each constricted chromosome that provides the foundation for kinetochore assembly and serves as a site for sister chromatid attachment. Errors in centromere or kinetochore function can lead to aberrant division and chromosomal instability, both of which are often observed in cancerous cells.
- Eukaryotic chromosomes exist in four major types based on the position of the centromere. On the basis of the location of the centromere, chromosomes are classified into four types: metacentric, submetacentric, acrocentric, and telocentric.

- The centromere is easily visualized as the most constricted region of a condensed mitotic chromosome.
- Although the word "centromere" is derived from the Greek words *centro* ("central") and *mere* ("part"), centromeres are not always found in the center of chromosomes.
- In fact, only so-called metacentric chromosomes have centromeres at their middle; in other chromosomes, centromeres are located at a variety of positions that are characteristic for each particular chromosome.
- The position of the centromere, therefore, provides a useful landmark for dividing chromosomes into karyotype groups and for establishing a

standardized nomenclature for mapping the positions of genes on chromosomes.

- With a few exceptions, eukaryotic chromosomes have a single centromere that ensures their accurate segregation during mitosis. Chromosomes that lack centromeres segregate randomly during mitosis and are eventually lost from cells.

- Chromosomes with multiple centromeres are subject to fragmentation if the centromeres become attached to opposite spindle poles by way of their kinetochores (highly complex multi-protein structure, spindle fibers attach to the centromere via the kinetochore).

- There are two types of centromeres:

I. "**Point centromeres**" bind to specific proteins that recognise particular DNA sequences with high efficiency. Any piece of DNA with the **point centromere** DNA sequence on it will typically form a centomere if present in the appropriate species. The best characterised point centromeres are those of the budding yeast, ***Saccharomyces cerevisiae*****.**

II. "**Regional centromeres**" is the term coined to describe most centromeres, which typically form on regions of preferred DNA sequence, but which can form on other DNA sequences as well. The signal for formation of a regional centromere appears to be epigenetic. Most organisms, ranging from the fission yeast ***Schizosaccharomyces pombe*** to humans, have regional centromeres.

The centromere is visible as the primary constriction on mammalian chromosomes. In humans this site is marked by the presence of a particular type of tandemly repeated simple DNA sequence, termed the alpha-satellite.

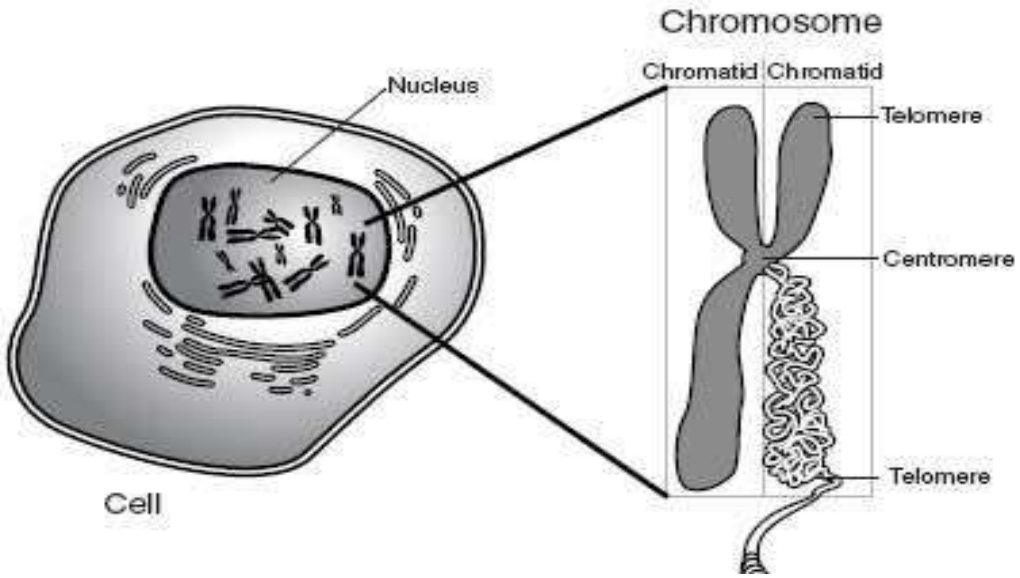
- Alpha-satellite DNA is a 170-bp sequence repeated over many millions of base pairs of DNA. In the mouse a different, but related, simple sequence (the minor satellite) occupies the equivalent position.

- In humans, variant chromosomes have shown that the presence of alpha-satellite DNA is neither necessary nor sufficient for centromere activity.
- Heterochromatin is visible at most mammalian centromeres and the histones in these regions are underacetylated.
- The study of proteins located at the centromere is essential to our understanding of centromere function and chromosome segregation. Several mammalian centromere-specific proteins have been identified.

Telomeres

Telomeres are the specialized structures at the ends of the linear DNA molecules. Telomeres 'hide' the ends of the chromosome from the mechanisms within the cell that monitor DNA damage. They are also needed to overcome the problem of end replication. The end-replication problem arises because all known DNA polymerases add nucleotides to a free 3' OH, i.e. they work only in the 5' to 3' direction.

Replicative DNA synthesis is primed from an RNA primer that is subsequently removed. At the extreme 3' end of a linear DNA strand being copied by lagging strand synthesis, removal of the RNA primer and ligation of Okazaki fragments leaves a gap at the end of the new strand. Because this is at the end of the DNA molecule/ chromosome there is no DNA template beyond this from which to prime synthesis of DNA across this gap. Hence, without a mechanism to counteract this, the ends of linear DNA molecules would get progressively shorter through subsequent rounds of conventional DNA replication.



Telomeric DNA sequences;

Most eukaryotes overcome the end-replication problem with an enzyme called **telomerase** (a ribonucleoprotein) that uses its own RNA template to add on simple repeats to the 3' ends of chromosomes, elongating them. Conventional DNA polymerases can use this extended DNA strand as a template on which to synthesize the complementary strand. The telomere sequence added is very similar in a wide variety of eukaryotes. In mammals it is **TTAGGG**. Large stretches of telomere-like repeats are found at the ends of chromosomes.

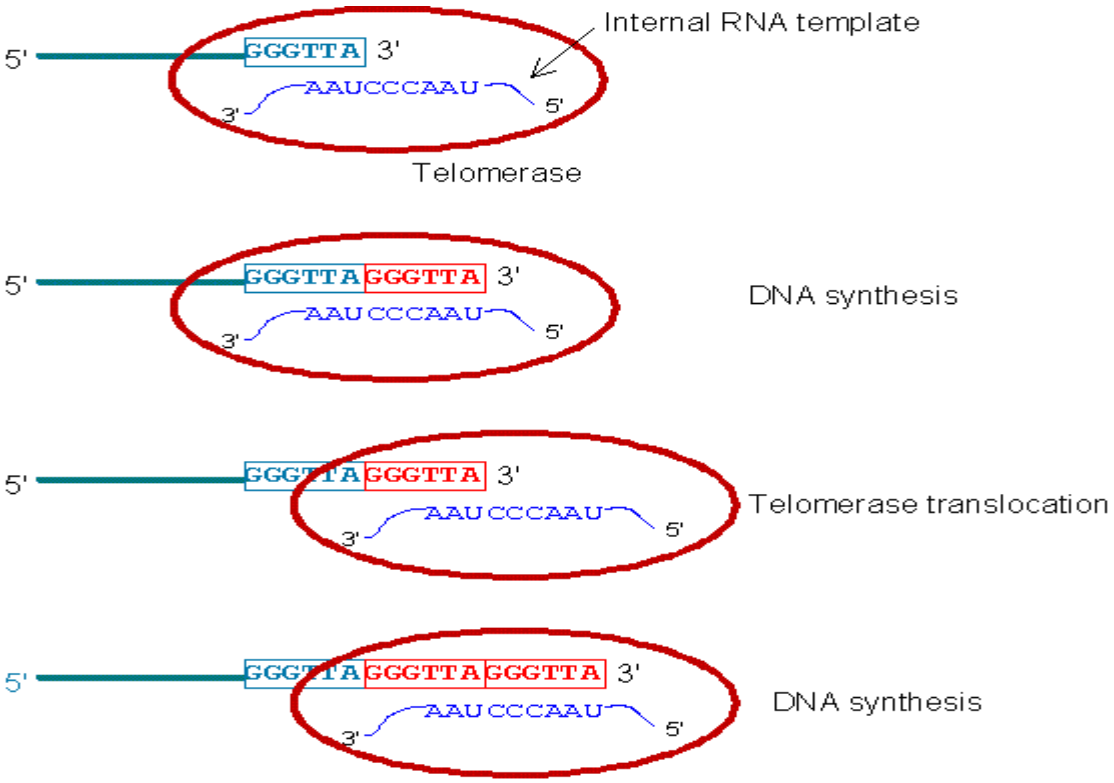
The telomeres of *Drosophila melanogaster* are unusual in that mobile repetitive sequences (non-long terminal repeat (LTR) retroposons) are found at the ends of the chromosomes rather than tandem repeats of TTAGGG.

Species	Repeat Sequence
Arabidopsis	TTTAGGG
Human	TTAGGG
Oxytricha	TTTTGGGG
Slime Mold	TAGGG
Tetrahymena	TTGGGG
Trypanosome	TAGGG
Yeast	$(TG)_{1-3}TG_{2-3}$

Telomerase activity:

Telomerase activity is highest in germ-line cells. Somatic cells have little telomerase and they have shorter telomeres than cells of the germ-line. It is estimated that 50–200 bp of DNA is lost from the end of the chromosome per cell division in the absence of telomerase activity.

Telomere length also decreases with age and this has led to speculation that telomere shortening may play some role in ageing. Indeed, in the human syndrome of premature ageing telomeres are excessively short.



Mammalian telomerase is a multi-subunit ribonucleoprotein complex of 41000 kDa. Two protein subunits have been identified:

p80 (TP1) binds telomerase RNA and **p123 (TERT1)** is similar to the budding yeast protein EST2p.

Mutation of EST2 in yeast results in short telomeres. Both EST2p and TERT1 have strong homology to reverse transcriptase, especially those of non-LTR retroposons. This has led to speculation that telomeres may have evolved from a cellular parasite, analogous to the presence of retroposons at *Drosophila* chromosome ends.

Experiment:

The gene encoding the RNA component of telomerase has been deleted in mice. These mice have shortened telomeres, losing _5 kb per generation, but they do not age prematurely, and in the first few generations they are fertile. By the fourth generation the mice begin to show chromosomal anomalies such as aneuploidy and chromosome fusions, and by the sixth generation they are unable to complete spermatogenesis successfully and the proliferative capacity of organs such as the testis, bone marrow and spleen is compromised. These experiments have demonstrated the role of telomeres and telomerase in both the maintenance of genome integrity and in the viability of organs with high rates of cell division.

Telomere capping

The ends of broken DNA molecules are 'sticky' and tend to fuse together. One job of the telomere is to 'camouflage' the natural ends of the linear chromosomal DNA molecules from the cellular mechanisms that monitor for DNA damage and also to stop the ends of real chromosomes fusing together. These functions are achieved by DNA binding proteins such as **taz 1** (in yeast) and **TRF1 and 2** (in mammals). These proteins bind to telomeres and so sequester away the linear DNA ends and they regulate telomere length through feedback to the telomerase enzyme. Over expression of TRF1 leads to telomere shortening, whereas abrogation of TRF1 function leads to elongated telomeres. When TRF2 function is perturbed in cells end-to-end fusions of chromosomes are seen.

Lec.3 *Dr.Khawla Alrawi*

The Genome

- The genome is the entire genetic material within an organism.
- An organism's genome contains DNA molecules or RNA (some viruses) and other organelles within a cell, such as mitochondria or chloroplasts.

- A genome includes all the coding regions of DNA that form discrete **GENES**, as well as all the noncoding stretches of DNA between genes.
- The sequence, structure, and chemical modifications of DNA not only provide the instructions needed to express the information held within the genome but also provide the genome with the capability to replicate, repair, package, and otherwise maintain itself.
- The human genome contains approximately 25,000 genes within its 3,000,000,000 base pairs of DNA, which form the 46 chromosomes found in a human cell. In contrast, *Nanoarchaeum equitans*, a parasitic prokaryote in the domain Archaea, has one of the smallest known genomes, consisting of 552 genes and 490,885 base pairs of DNA.

Genomics:

It is the study of the structure, function, and inheritance of genomes. It is useful for identifying genes, determining gene function, and understanding the evolution of organisms.

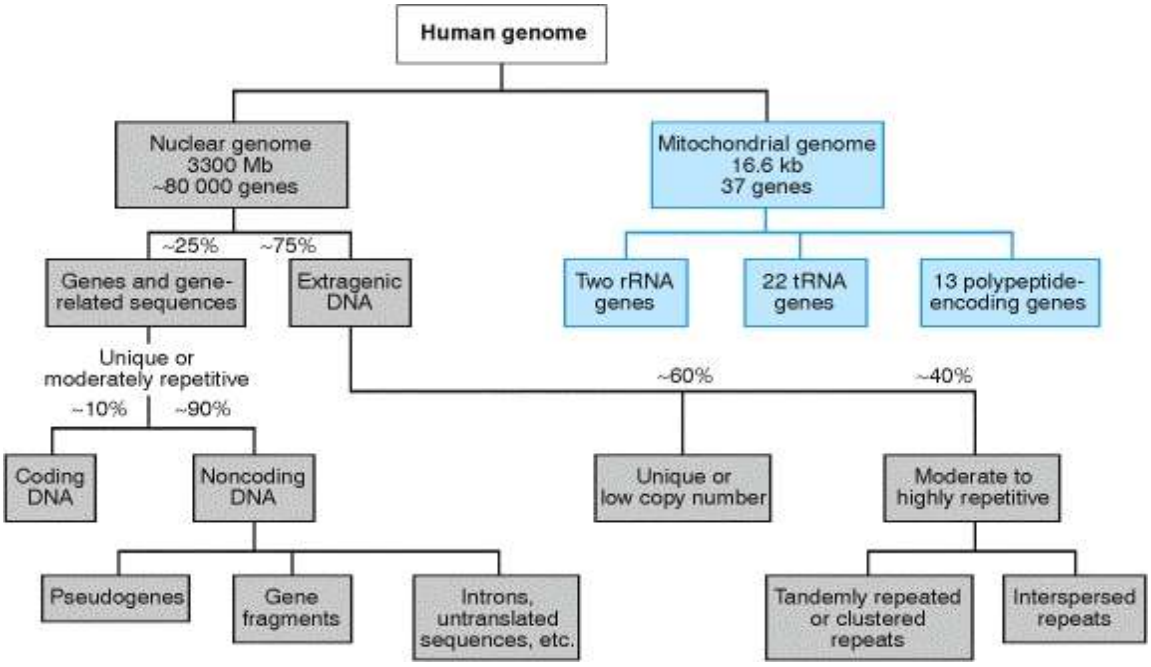
Human genome

The human genome is the total amount of genetic information in the chromosomes. This information is located as DNA sequences within the 23 chromosome pairs in cell nuclei and in a small DNA molecule found within individual mitochondria. The human genome is made up of about 35,000 genes. They include both protein-coding DNA genes and noncoding DNA.

Haploid human genomes (contained in egg and sperm cells) consist of three billion DNA base pairs, while diploid genomes (found in somatic cells) have twice the DNA content.

The haploid human genome contains approximately 20,000 protein-coding genes, significantly fewer than had been anticipated. Protein-coding sequences account for only a very small fraction of the genome (approximately 1.5%), and the rest is associated with non-coding RNA

molecules, regulatory DNA sequences, LINEs, SINEs, introns, and sequences for which as yet no function has been elucidated.



Human Genome Project (HGP)

International scientific research projects designed to:

- Study and identify all of the genes in the human genome,
- Determine the base-pair sequences in human DNA,
- Store this information in computer databases.

HGP began in the United States in 1990 and was completed in 2003. In 2012, thousands of human genomes have been completely sequenced, and many more have been mapped at lower levels of resolution.

The resulting data are used worldwide in biomedical science, anthropology, forensics and other branches of science.

There is a widely held expectation that genomic studies will lead to advances in the diagnosis and treatment of diseases, and to new insights in many fields of biology, including human evolution.

Genome size varies in different organisms:

Coding and noncoding DNA:

The content of the human genome is commonly divided into coding sequences (that can be transcribed into mRNA and translated into proteins) which occupy only a small fraction of the genome (<2%), and noncoding DNA sequences (not used to encode proteins, they occupy 98% of the genome).

I. Coding sequences (protein-coding genes)

These sequences ultimately lead to the production of all human proteins, although several biological processes (e.g. DNA rearrangements and alternative pre-mRNA splicing) can lead to the production of many more unique proteins than the number of protein-coding genes.

The complete modular protein-coding capacity of the genome is contained within the exons, and consists of DNA sequences encoded by exons that can be translated into proteins.

Protein-coding genes are distributed unevenly across the chromosomes, with an especially high gene density within chromosomes 19, 11, and 1. Each chromosome contains various gene-rich and gene-poor regions, which may be correlated with chromosome bands and GC-content. The significance of these nonrandom patterns of gene density is not well understood. The size of protein-coding genes within the human genome shows enormous variability.

II. Noncoding DNA (ncDNA)

Noncoding DNA is defined as all of the DNA sequences within a genome that are not found within protein-coding exons, and so are never

represented within the amino acid sequence of expressed proteins. By this definition, more than 98% of the human genomes are composed of ncDNA.

Numerous classes of noncoding DNA have been identified, including genes for **noncoding RNA** (e.g. tRNA and rRNA), **pseudogenes, introns, untranslated regions of mRNA, regulatory DNA sequences, repetitive DNA sequences, and sequences related to mobile genetic elements.**

Excluding protein-coding sequences, introns, and regulatory regions, much of the non-coding DNA is composed of:

- Many DNA sequences that do not play a role in gene expression have important biological functions. About 5% of the genome contains sequences of noncoding DNA that are **highly conserved**. Many of these sequences regulate the structure of chromosomes by limiting the regions of heterochromatin formation and regulating structural features of the chromosomes, such as the **telomeres** and **centromeres.**
- Other noncoding regions serve as origins of DNA replication.
- Finally several regions are transcribed into functional noncoding RNA that regulate the expression of protein-coding genes, mRNA translation and stability , chromatin structure (including histone modifications), DNA methylation , DNA recombination , and cross-regulate other noncoding RNAs .

1. Pseudogenes

Pseudogenes are inactive copies of protein-coding genes, often generated by gene duplication, that have become nonfunctional through the accumulation of inactivating mutations.

For example, the olfactory receptor gene family is one of the best-documented examples of pseudogenes in the human genome. More than 60% of the genes in this family are non-functional pseudogenes in humans. By comparison, only 20% of genes in the mouse olfactory receptor gene family are pseudogenes. Research suggests that this is a

species-specific characteristic, as the most closely related primates all have proportionally fewer pseudogenes. This genetic discovery helps to explain the less acute sense of smell in humans relative to other mammals.

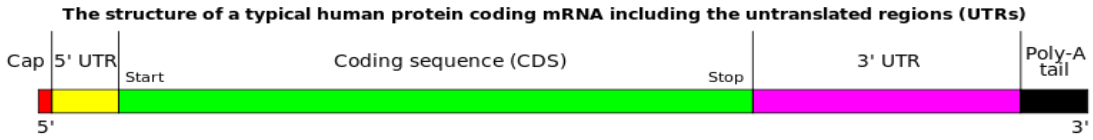
Types of pseudogene. (*A*) Mutation of existing genes gives rise to unitary pseudogenes. (*B*) Duplicated pseudogenes are produced following mutation of copied genes. (*C*) Reverse transcription of mRNA into cDNA followed by retrotransposition into genomic DNA leads to the generation of processed pseudogenes. Filled boxes represent exons; open boxes represent introns; X represents a crippling mutation that ablates protein-coding potential.

2. Genes for noncoding RNA (ncRNA)

Noncoding RNA molecules play many essential roles in cells, especially in the many reactions of protein synthesis and RNA processing. The human genome contains genes encoding 18,400 ncRNAs, including tRNA, rRNA, microRNA, and other non-coding RNA genes.

3. Introns and untranslated regions of mRNA

The initial transcripts of protein coding genes usually contain extensive noncoding sequences, in the form of introns, 5'-untranslated regions (5'-UTR), and 3'-untranslated regions (3'-UTR). Within most protein-coding genes of the human genome, the length of intron sequences is 10- 100 times the length of exon sequences.



Regulatory DNA sequences

The human genome has many different regulatory sequences which are crucial to controlling gene expression. Some types of non-coding DNA are genetic "switches" that do not encode proteins, but do regulate when genes are expressed (called enhancers).

4. Repetitive DNA sequences

About 8% of the human genome consists of repetitive DNA sequences, termed tandem DNA arrays or tandem repeats. The repeated sequences may be of variable lengths, from two nucleotides to tens of nucleotides. These sequences are highly variable, even among closely related individuals, and so are used for genealogical DNA testing and forensic DNA analysis.

Repeated sequences of fewer than ten nucleotides (e.g. the dinucleotide repeat $(AC)_n$ are termed **microsatellite** sequences. Among the microsatellite sequences, trinucleotide repeats are of particular importance, as sometimes occur within coding regions of genes for proteins and may lead to genetic disorders. For example, **Huntington's disease** results from an expansion of the trinucleotide repeat $(CAG)_n$ with in the *Huntingtin* gene on human chromosome 4.

Telomeres (the ends of linear chromosomes) end with a microsatellite hexanucleotide repeat of the sequence $(TTAGGG)_n$.

Tandem repeats of longer sequences (arrays of repeated sequences 10–60 nucleotides long) are termed **minisatellites**.

5. Mobile genetic elements (transposons)

Transposable genetic elements, DNA sequences that can replicate and insert copies of themselves at other locations within a host genome, are an abundant component in the human genome. The most abundant transposon lineage, ***Alu***, has about 50,000 active copies, while another lineage, ***LINE-1***(Long interspersed nuclear element)(6000-8000bp), has about 100 active copies per genome. Sometimes they called "**jumping genes**".

Transposons have played a major role in sculpting the human genome.

Mobile elements within the human genome can be classified into **LTR Long terminal repeat retrotransposons** (8.3% of total genome), **SINEs** (Short Interspersed Elements)(100-300bp) (13.1% of total genome)

including *Alu* elements, **LINEs** (20.4% of total genome), and **Class II DNA transposons** (2.9% of total genome).

Measuring human genetic variation

Most studies of human genetic variation have focused on single-nucleotide polymorphisms (SNPs), which are substitutions in individual bases along a chromosome. Most analysts estimate that SNPs occur 1 in 1000 base pairs, on average, in the euchromatic human genome, although they do not occur at a uniform density.

The genomic loci and length of certain types of small repetitive sequences are highly variable from person to person, which is the basis of DNA fingerprinting and DNA paternity testing technologies.

Epigenome

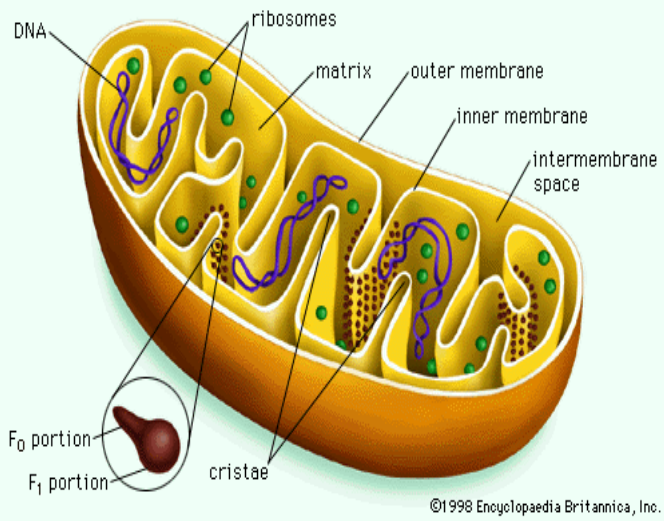
Epigenetics describes a variety of features of the human genome that transcend its primary DNA sequence, such as chromatin packaging, histone modifications and DNA methylation, which are important in regulating gene expression, genome replication and other cellular processes.

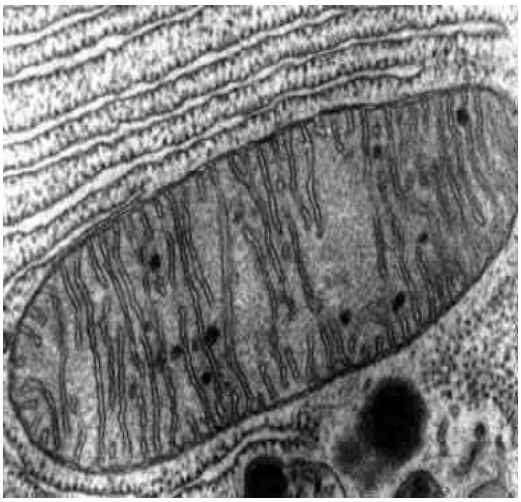
Epigenetic markers strengthen and weaken transcription of certain genes but do not affect the actual sequence of DNA nucleotides. DNA methylation is a major form of epigenetic control over gene expression and one of the most highly studied topics in epigenetics. During development, the human DNA methylation profile experiences dramatic changes. In early germ line cells, the genome has very low methylation levels. These low levels generally describe active genes. As development progresses, parental imprinting tags lead to increased methylation activity.

Mitochondrial Genome

Introduction:

- Membrane-bound organelle (eukaryotic only)
- All eukaryotic cells either have mitochondria or have nuclear genes that seem to have been derived from mitochondria.
- Each cell contains hundreds to thousands of mitochondria.
- Site of Krebs cycle and oxidative phosphorylation (the electron transport chain, or respiratory chain).
- Two membranes: outer and inner.
- Folds of the inner membrane, where most of oxidative phosphorylation occurs, are called cristae.
- Inside inner membrane = matrix
- Between membranes = intermembrane space
- Mitochondrial DNA is inside the inner membrane.





Endosymbiont Hypothesis

- Endosymbiont hypothesis: originally proposed in 1883 by Andreas Schimper, but extended by Lynn Margulis in the 1980s.
- Mitochondrial ribosomal RNA genes and other genes show that the original organism was in the alpha-proteobacterial family (similar to nitrogen-fixing bacteria)

- **Evidence:**

– Mitochondria have their own DNA (circular)

– The inner membrane is more similar to prokaryotic membranes than to eukaryotic. By the hypothesis, the inner membrane was the original prokaryotic membrane and the outer membrane was from the primitive eukaryote that swallowed it.

– Mitochondria make their own ribosomes, which are of the prokaryotic 70S type, not the eukaryotic 80S type.

– Mitochondria are sensitive to many bacterial inhibitors that don't affect the rest of the eukaryotic cell, such as streptomycin, chloramphenicol, rifampicin.

– Mitochondrial protein synthesis starts with N-formyl methionine, as in the bacteria but unlike eukaryotes.

- Most of the original bacterial genes have migrated into the nucleus.
- Eukaryotes that lack mitochondria generally have some mitochondrial genes in their nucleus, evidence that their ancestors had mitochondria that were lost during evolution.

Mitochondrial Function

- **Krebs cycle:**

– Pyruvate, the product of glycolysis, is produced in the cytoplasm.

– It is transported into the mitochondrial matrix. There it is converted into acetyl CoA.

– Fatty acids, from the breakdown of lipids, are also transported into the matrix and converted to acetyl CoA.

– The Krebs cycle then converts acetyl CoA into carbon dioxide and high energy electrons. The high energy electrons are carried by NADH and $FADH_2$.

- **Electron Transport:**

– The high energy electrons are removed from NADH and $FADH_2$, and passed through three protein complexes embedded in the inner membrane.

– Each complex uses some of the electrons' energy to pump H^+ ions out of the matrix into the intermembrane space.

- The final protein complex gives the electrons to oxygen, converting it to water.
- The H^+ ions come back into the matrix, down the concentration gradient, through a fourth complex, ATP synthase (also called ATPase), which uses their energy to generate ATP from ADP and inorganic phosphate.

• In brown fat, the synthesis of ATP is uncoupled from the flow of H^+ ions back into the matrix. The H^+ ions flow through a protein called thermogenin, and not through the ATPase. The energy is converted into heat: the primary way we keep warm in cold weather.

Genome Structure

• The mitochondrial genome is a circle, 16.6 kb of DNA. A typical bacterial genome is 2-4 Mbp.

• The two strands are notably different in base composition, leading to one strand being "heavy" (the H strand) and the other light (the L strand).

• Both strands encode genes, although more are on the H strand.

• A short region (1121 bp), the D loop (D = "displacement"), is a DNA triple helix: there are 2 overlapping copies of the H strand there.

• The D loop is also the site where most of replication and transcription is controlled.

• Genes are tightly packed, with almost no non-coding DNA outside of the D loop. In one case, two genes overlap: they share 43 bp, using different reading frames. Human mitochondrial genes contain no introns, although introns are found in the mitochondria of other groups (plants, for instance).

Mitochondrial Genes

• Genes: total of 37. 22 tRNAs, 2 rRNAs, 13 polypeptides.

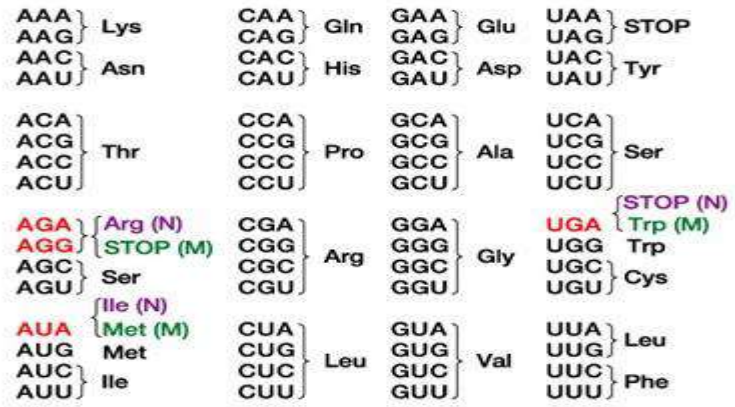
• tRNA: only 60 of the 64 codons codes for amino acids. 8 tRNAs cover all 4 3^{rd} base positions with the same amino acid, and the remaining 14 tRNAs each cover two 3^{rd} base positions (purines or pyrimidines). Thus, all 60 codons are covered.

• rRNA: 16S and 23S which are standard sizes for bacterial rRNAs. Bacterial ribosomes don't use 5S or 5.8S rRNAs.

- Polypeptides: all are components of the electron transport chain. Other components are encoded in the nucleus and transported to the mitochondria after translation.

Genetic Code

- The mitochondrial genetic code has drifted from the universal code: there are so few polypeptides that changes in the code are tolerated.
- Human mitochondrial code is different from other groups such as plants or fungi.
- Uses 2 of the 3 universal stop codons, but also uses 2 other codons as stop codons. Also, UGA codes for tryptophan in the mitochondrial, while it is a stop codon in the universal code. AUA gives methionine in the mitochondria instead of isoleucine.



Replicatio Transcription

- **<u>Replication</u>** starts with the H strand.
- The origin of replication for the H strand is in the D loop, and it is initiated by an RNA primer generated from the L strand transcript.
- After the new H strand is about 2/3 complete, the L strand origin of replication is uncovered. The L strand origin is on the old H strand; it is "uncovered" when the old H strand is displaced by the DNA polymerase synthesizing the new H strand.

- The L strand origin folds into a stem-loop structure, which acts as a primer, and replication of the L strand begins.
- Replication can be said to be bidirectional by asynchronous, unlike replication of nuclear DNA, which proceeds in both directions simultaneously.
- **Transcription.**
- Both strands are transcribed.
- The D loop contains one promoter for each strand, and the entire strand is transcribed.
- The RNA is then cut into individual RNAs for each gene.
- Protein-coding genes are given poly-A tails, and rRNA and tRNA molecules are modified as necessary.

Mitochondrial Genetics

- Maternal inheritance: Inherited through the mother (egg) only. Allows tracing female line back in time.
- A few sperm mitochondria enter the egg, but they are degraded and lost.

Mutation rate in mtDNA is very high: 10 times the nuclear rate. mtDNA is associated with the inner membrane, the site of oxidative phosphorylation. Large amounts of "reactive oxygen species" (peroxide and superoxide) are present, and they are quite mutagenic. The D loop has an especially high rate of mutation. Part of the effects of aging has been attributed to the gradual loss of mitochondria due to accumulated mutations in individual cells.

Heteroplasmy

- Sometimes an individual has more than one kind of mitochondria. This is called heteroplasmy. Since mitochondria are divided randomly during cell division, different cells get different proportions of the two types.
- If one mitochondrial type is mutant and the other is normal, severity of symptoms will vary in different tissues depending on the proportions of the two types.

During oogenesis (egg formation), random segregation of the two types can lead to some offspring inheriting a mitochondrial disease while other offspring are normal

Genetic Diseases

- In general: malfunctions of respiratory chain, so affects high metabolism tissues the most: nervous system, muscles, kidney, and liver.
- **Leber's hereditary optic neuropathy (LHON):**

LHON causes a painless loss of central vision between 12 and 30 years of age. Both eyes are affected at the same time. Males will not pass the gene to any of their children, but females with the mutation will pass it to all of their children, regardless of whether they are sons or daughters.

– Progressive loss of vision, from central to peripheral, usually beginning in 20's.

– Eyes can be affected several months apart, or simultaneously.

– About 85% are male (no good reason why).

– Recurrence risk for siblings around 20% (heteroplasmy); many spontaneous cases.

– Due to death of optic nerve fibers.

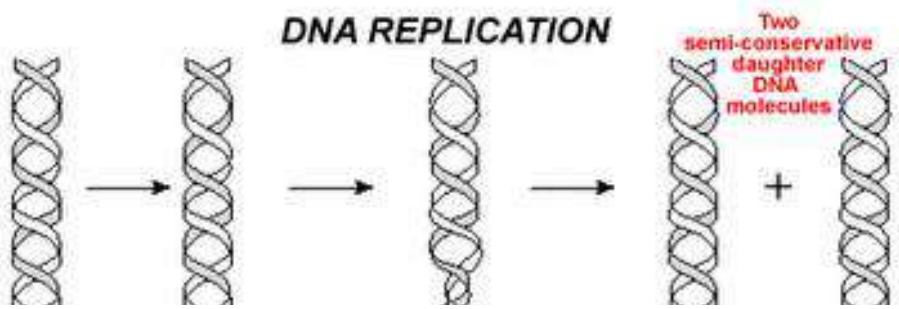
– Most due to change in conserved Arg to His in NADH dehydrogenase, but 18 total mutations known, all missense in respiratory chain.

- **Myoclonic epilepsy and ragged red fiber disease (MERRF):** CNS symptoms: epilepsy, deafness, dementia. Skeletal and heart muscles abnormal, mitochondria appear abnormal. Multiple enzyme defects in respiratory chain. Lots of variation in inheritance of disease. Most due to A --> G in lysine tRNA (mutation A8344G). Easy to assay for; CviJI restriction site is altered. Good correlation between % mutant mitochondria and disease severity.
- **Kearns-Sayre syndrome**: paralysis of the muscles, retinal degeneration, cardiac muscle problems, seizures, many other symptoms irregularly. Due to large deletions of mtDNA. Heteroplasmy necessary for survival. Mostly spontaneous--rarely passed to offspring. Many variants.

Lect 5

DNA replication

Is a biological process that occurs in all living organisms and copies their DNA; it is the basis for biological inheritance. The process starts when one double-stranded DNA molecule produces two identical copies of the molecule.



Repetitive DNA	Over 50% of genome	Very little
Transcription	The great bulk of genes are transcribed individually	Co-transcription of multiple genes from both the heavy and light strands
Introns	Found in most genes	Absent
% of coding DNA	~ 1.5%	~ 93%
Codon usage	Slightly different see slide	
Recombination	At least once for each pair of homologs at meiosis	No evidence for this occurring naturally
Inheritance	Mendelian for sequence on X and autosomes; paternal for sequence on Y	Exclusively maternal

DNA replication can also be performed *in vitro*. DNA polymerases, isolated from cells, and artificial DNA primers are used to initiate DNA synthesis at known sequences in a template molecule. The polymerase chain reaction (PCR), a common laboratory technique, employs such artificial synthesis in a cyclic manner to amplify a specific target DNA fragment from a pool of DNA.

DNA Replication models

The process of **DNA Replication** was hiding many secrets. One of the most important was how the two daughter strands are created. As we have noticed in previous studies, DNA molecule is a complex of two chains! In order the hereditary phenomenon to be explained, these strands should be accurately copied and transmitted from the parental cell to the daughter ones. These are three possible models that describe the accurate creation of the daughter chains.

1) **Semiconservative Replication** According to this model, DNA replication would create two molecules. Each of them would be a complex of an old parental and a daughter strand.

2) Conservative Replication According to this model, the DNA Replication process would create a brand-new DNA double helix made of two daughter strands while the parental chains would stay together.

3) Dispersive Replication According to this model the Replication Process would create two DNA double-chains, each of them with parts of both parent and daughter molecules.

The correct model is the first. Semiconservative DNA replication was proved by the experiment of **Meselson - Stahl**.

DNA Replication Enzymes

DNA replication is a very complex process that requires many proteins to act together. These proteins known as replication proteins are clustered together in the cell and that unit of the cell can be called the 'Replication Factory'. Here the replication, results in two DNA molecules.

The replication proteins each have a specific function in the production of new DNA strand:

- **Helicase**: The six proteins arranged in a ring shape, Unwind a portion of the DNA double Helix into single strand. Helicases are often utilized to separate strands of a DNA double helix or a self-annealed RNA molecule using the energy from ATP hydrolysis, a process characterized by the breaking of hydrogen bonds between annealed nucleotide bases. They move incrementally along one nucleic acid strand of the duplex with a directionality and processivity specific to each particular enzyme

- **Single-stranded DNA-binding protein (SSB**). The tetramers, that cover the single-stranded DNA. This prevents the DNA strands from re-annealing and forming the double stranded molecule.

- **Topoisomerases** are enzymes that regulate the overwinding or underwinding of DNA. They are separated into two types according to the number of strands cut in one round of action:

  ***Type I topoisomerase** cuts one strand of a DNA double helix, relaxation occurs, and then the cut strand is reannealed. Cutting one strand allows the part of the molecule on one side of the cut to rotate around the uncut strand, thereby reducing stress from too much or too little twist in the helix. Such stress is introduced when the DNA strand is "supercoiled".

  ***Type II topoisomerase** cuts both strands of one DNA double helix, passes another unbroken DNA helix through it, and then reanneals the cut strands. It is utilizing ATP hydrolysis. There are two subclasses: type IIA and type IIB topoisomerases, which share similar structure and mechanisms. Examples of type IIA topoisomerases include eukaryotic topo II, E. coli gyrase, and E. coli topo IV. Examples of type IIB topoisomerase include topo VI.

- **RNA Primase**: It is catalyzes the synthesis of a short RNA segment called a primer complementary to a ssDNA template. Primase is importance in DNA replication because no known DNA polymerases can initiate the synthesis of a DNA strand without an initial RNA or DNA primer (for temporary DNA elongation).

In bacteria, primase binds to the DNA helicase forming a complex called the **primosome**. Primase is activated by DNA helicase where it then synthesizes a short RNA primer approximately 11 ±1 nucleotides long, to which new nucleotides can be added by DNA polymerase.

- **DNA polymerase** "reads" an intact DNA strand as a template and uses it to synthesize the new strand. This process forms a new DNA strand complementary to the template strand and identical to the template's original partner strand. DNA polymerases use **magnesium ions** as cofactors. Human DNA polymerases are 900-1000 amino acids

1. **Prokaryotic DNA polymerase**

There are 5 known **Prokaryotic DNA polymerases**:

- **Pol I**: implicated in DNA repair; has 5'->3' polymerase activity, and both 3'->5' exonuclease activity (proofreading) and 5'->3' exonuclease activity (RNA primer removal).
- **Pol II**: involved in repairing damaged DNA; has 3'->5' exonuclease activity (proofreading).
- **Pol III**: the main polymerase in bacteria (responsible for 5'->3' elongation); has 3'->5' exonuclease activity (proofreading).
- **Pol IV**: Y-family DNA polymerase, Involved in SOS repair and translesion repair.
- **Pol V**: Y-family DNA polymerase; participates in by passing DNA damage, involved in SOS repair and translesion repair.

2. **Eukaryotic DNA polymerase**

There are at least 15 **Eukaryotic DNA polymerases** the main five are:

- **Pol α**: Replication of nuclear DNA, proofreading 3'→5' exonuclease activity.
- **Pol β**: Implicated in repairing DNA, in base excision repair and gap-filling synthesis.
- **Pol γ**: Replicates and repairs mitochondrial DNA and has proofreading 3'→5' exonuclease activity.
- **Pol δ**: Highly processive and has proofreading 3'→5' exonuclease activity. Thought to be the main polymerase involved in lagging strand synthesis.

- **Pol ε**: Also, highly processive and has proofreading 3'→5' exonuclease activity (Highly related to pol δ). It has 5'→3' exonuclease activity also and can remove primers on the lagging strand during the replication.

- **RNAse H and DNA polymerase I (exonuclease)**: recognizes the RNA polymers that are bound to the DNA template and removes the primers by RNA hydrolysis

- **DNA Ligase**: close nicks in the phosphodiester backbone of DNA. Biologically, DNA ligases are essential for the joining of Okazaki fragments during replication, and for completing short-patch DNA synthesis occurring in DNA repair process. There are two classes of DNA ligases. The first uses **NAD+** as a cofactor and only found in bacteria. The second uses **ATP** as a cofactor and found in eukaryotes, viruses and bacteriophages. Human DNA ligase is > 100KD.
- **Nucleases**: Remove wrong nucleotides from the daughter strand.

DNA polymerases:

DNA polymerases are a family of enzymes that carry out all forms of DNA replication. However, a DNA polymerase can only **extend an existing DNA** strand paired with a template strand; it cannot begin the synthesis of a new strand. To begin synthesis, a short fragment of DNA or RNA, called a primer, must be created and paired with the template DNA strand.

DNA polymerase then synthesizes a new strand of DNA by extending the 3' end of an existing nucleotide chain, adding new nucleotides matched to the template strand one at a time via the creation of **phosphodiester bonds**.

The energy for this process of DNA polymerization comes from **two of the three total phosphates** attached to each unincorporated base. (Free bases with their attached phosphate groups are called nucleoside triphosphates.) When a nucleotide is being added to a growing DNA strand, two of the phosphates are removed and the energy produced creates a phosphodiester bond that attaches the remaining phosphate to the growing chain. The energetics of this process also helps explain the directionality of synthesis.

If DNA were synthesized in the 3' to 5' direction, the energy for the process would come from the 5' end of the growing strand rather than from free nucleotides.

In general, DNA polymerases are extremely **accurate,** making less than **one mistake for every 10^7** nucleotides added. Even so, some DNA polymerases also have **proofreading** ability; they can remove nucleotides from the end of a strand in order to correct mismatched bases.

Note: If the 5' nucleotide needs to be removed during proofreading, the triphosphate end is lost. Hence, the energy source that usually provides energy to add a new nucleotide is also lost.

DNA replication process:

DNA Replication proceeds in three enzymatically catalyzed and coordinated steps: initiation, elongation and termination.

1. Initiation:

❖ This process is initiated at particular points in the DNA, known as "**origins**", which are targeted by proteins that separate the two strands and initiate DNA synthesis. Origins contain DNA sequences recognized by replication initiator proteins (e.g., **DnaA** in *E. coli'* and the **Origin Recognition Complex** in yeast). These initiators recruit other proteins to separate the strands and initiate replication forks.

Origins tend to be "**AT-rich**" (rich in adenine and thymine bases) to assist this process, because A-T base pairs have two hydrogen bonds (rather than the three formed in a C-G pair)—in general, strands rich in these nucleotides are easier to separate since less energy is required to break relatively fewer hydrogen bonds.

Only One Replication Origin in *E. coli*

All known DNA replication systems require a free 3' OH group before synthesis can be initiated (**DNA is read in 3' to 5' direction** whereas a new strand is synthesized in the 5' to 3' direction).

Four distinct mechanisms for synthesis have been described.

1. All cellular life forms and many DNA **viruses, phages and plasmids** use a primase to synthesize a short RNA primer with a free 3′ OH group which is subsequently elongated by a DNA polymerase.

2. The retroelements (including **retroviruses**) employ a transfer RNA that primes DNA replication by providing a free 3′ OH that is used for elongation by the reverse transcriptase.

3. In the **adenoviruses** and the **φ29** family of bacteriophages, the 3' OH group is provided by the side chain of an amino acid of the genome attached protein (the terminal protein) to which nucleotides are added by the DNA polymerase to form a new strand.

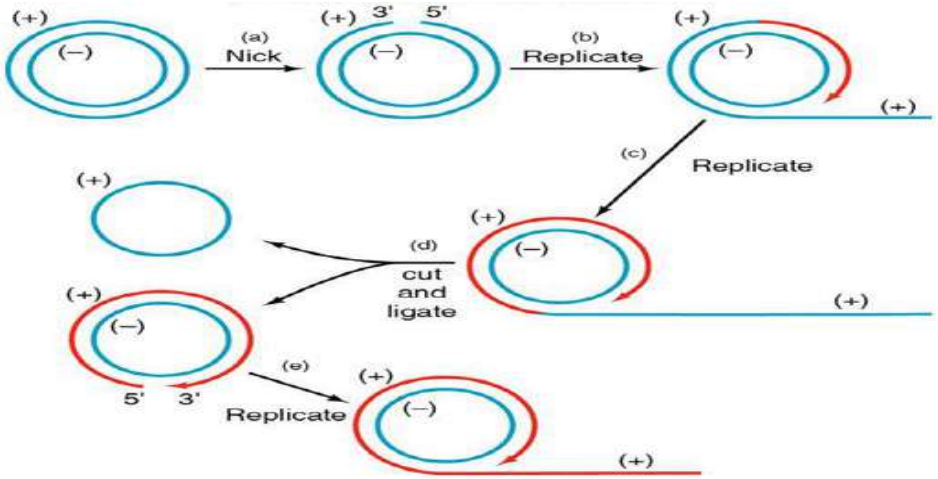
4. In the single stranded DNA **viruses - circoviruses, geminiviruses, parvoviruses** and also the many **phages and plasmids** that use the rolling circle replication mechanism, the endonuclease creates a nick the genome strand (single stranded viruses) or one of the DNA strands (plasmids). The 5′ end of the nicked strand is transferred to a tyrosine residue on the nuclease and the free 3′ OH group is then used by the DNA polymerase for new strand synthesis.

Rolling circle replication

Rolling circle replication describes a process of unidirectional nucleic acid replication that can rapidly synthesize multiple copies of **circular molecules of DNA or RNA**, such as **plasmids**, the genomes of **bacteriophages**, and the circular RNA genome of **viroids**.

This replication is initiated by an initiator protein encoded by the plasmid or bacteriophage DNA, which nicks one strand of the double-stranded, circular DNA molecule at a site called the **double-strand origin**, or DSO. The initiator protein remains bound to the 5' phosphate end of the nicked strand, and the free 3' hydroxyl end is released to serve as a primer for DNA synthesis by DNA polymerase III. Using the unnicked strand as a template, replication proceeds around the circular DNA molecule, displacing the nicked strand as single-stranded DNA. Displacement of the nicked strand is

carried out by a host-encoded helicase called **PcrA (plasmid copy reduced)** in the presence of the plasmid replication initiation protein



Rolling circle replication

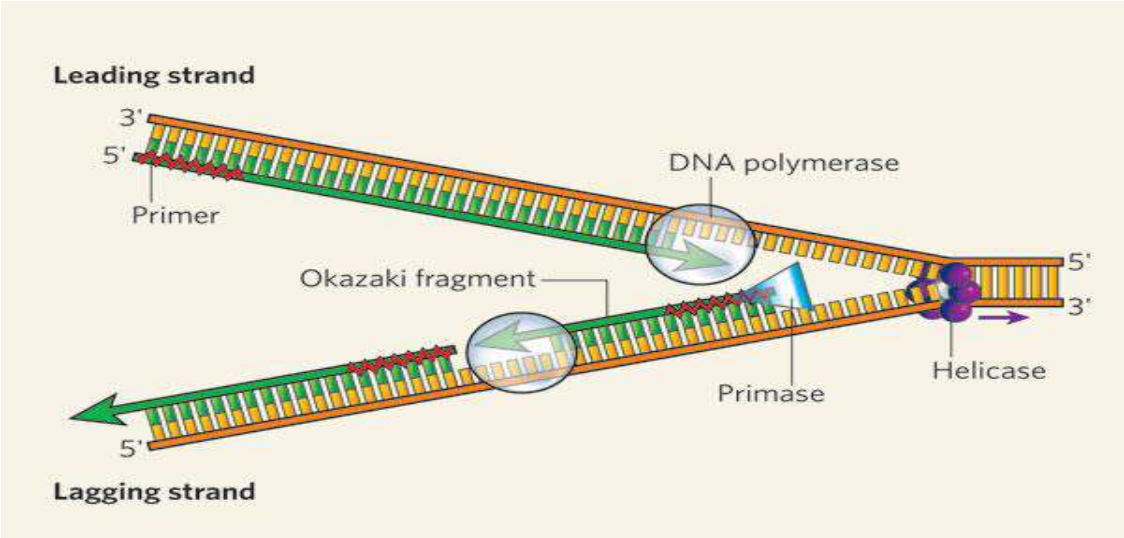
- The best known of these mechanisms is that used by the cellular organisms. In these once the two strands are separated, RNA primers are created on the template strands. To be more specific, the **leading strand** receives one RNA primer per active origin of replication while the **lagging strand** receives several; these several fragments of RNA primers found on the lagging strand of DNA are called **Okazaki fragments**.

  DNA polymerase extends the leading strand in one **continuous motion** and the lagging strand in a **discontinuous motion** (due to the Okazaki fragments). **RNase** removes the RNA fragments used to initiate replication by DNA polymerase, and another DNA polymerase enters to fill the gaps. When this is complete, a single nick on the leading strand and several nicks on the lagging strand can be found. **Ligase** works to fill these nicks in, thus completing the newly replicated DNA molecule.

- As DNA synthesis continues, the original DNA strands continue to unwind on each side of the bubble, forming a replication fork with two prongs.
- In bacteria, which have a single origin of replication on their circular chromosome, this process eventually creates a "theta structure" (resembling the Greek letter theta: θ). In contrast, eukaryotes have longer linear chromosomes and initiate replication at multiple origins.

2. Elongation:

The replication fork is a structure that forms within the nucleus during DNA replication. It is created by helicases, which break the hydrogen bonds holding the two DNA strands together. The resulting structure has two branching each one made up of a single strand of DNA. These two strands serve as the template for the leading and lagging strands, which will be created as DNA polymerase matches complementary nucleotides to the templates; the templates may be properly referred to as the leading strand template and the lagging strand templates.



Leading strand

The leading strand is the template strand of the DNA double helix so that the replication fork moves along it in the 3' to 5' direction. This allows the newly synthesized strand complementary to the original strand to be synthesized 5' to 3' in the same direction as the movement of the replication fork.

On the leading strand, a polymerase "reads" the DNA and adds nucleotides to it continuously. This polymerase is **DNA polymerase III** (DNA Pol III) in prokaryotes and presumably **Pol ε** in yeasts. In human cells the leading and lagging strands are synthesized by **Pol α** and **Pol δ** within the nucleus and **Pol γ** in the mitochondria. Pol ε can substitute for Pol δ in special circumstances.

Lagging strand

The lagging strand is the strand of the template DNA double helix that is oriented so that the replication fork moves along it in a **5' to 3' manner.** Because of its orientation, opposite to the working orientation of DNA polymerase III, which moves on a template in a 3' to 5' manner, replication of the lagging strand is more complicated than that of the leading strand.

On the lagging strand, **primase** "reads" the DNA and adds RNA to it in short, separated segments. In eukaryotes, primase is intrinsic to **Pol α.**

DNA polymerase III or **Pol δ** lengthens the primed segments forming okazaki fragments. Primer removal in eukaryotes is also performed by Pol δ.

In prokaryotes, **DNA polymerase I** "reads" the fragments, removes the RNA using its flap endonuclease domain (RNA primers are removed by 5'-3' exonuclease activity of polymerase I , and replaces the RNA nucleotides with DNA nucleotides this is necessary because RNA and DNA use slightly different kinds of nucleotides). DNA ligase joins the fragments together.

Dynamics at the replication fork

- As **helicase** unwinds DNA at the replication fork, the DNA ahead is forced to rotate. This process results in a build-up of twists in the DNA ahead. This build-up would form a resistance that would eventually halt the progress of the replication fork. DNA **Gyrase** is an enzyme that temporarily breaks the strands of DNA, relieving the tension caused by unwinding the two strands of the DNA helix; DNA Gyrase achieves this by adding negative supercoils to the DNA helix.
- Bare single-stranded DNA tends to fold back on itself and form secondary structures (hair pins); these structures can interfere with the movement of DNA polymerase. To prevent this, **single-strand binding (SSB) proteins** bind to the DNA until a second strand is synthesized, preventing secondary structure formation.

Regulation in Eukaryotes

Within eukaryotes, DNA replication is controlled within the context of the cell cycle. As the cell grows and divides, it progresses through stages in the cell cycle; DNA replication occurs during the **S phase** (synthesis phase). The progress of the eukaryotic cell through the cycle is controlled by cell cycle checkpoints which controlled through complex interactions between various proteins, including **cyclins** and **cyclin-dependent kinases.**

Regulation in prokaryote:

Most bacteria do not go through a well-defined cell cycle but instead continuously copy their DNA; during rapid growth, this can result in the concurrent occurrences of multiple rounds of replication. In *E. coli*, DNA replication is regulated through several mechanisms, including: the **hemi-methylation** and **sequestering of the origin sequence**, the **ratio of ATP to ADP**, and the levels of **protein DnaA** (required for initiation of replication). All these controls the process of initiator proteins binding to the origin sequences.

3. Termination:

- **Eukaryote:**

Eukaryotes initiate DNA replication at multiple points in the chromosome, so replication forks meet and terminate at many points in the chromosome; these are not known to be regulated in any particular way. Because eukaryotes have linear chromosomes, DNA replication is unable to reach the very end of the chromosomes, but ends at the **telomere region** of repetitive DNA close to the end.

This shortens the telomere of the daughter DNA strand. This is a normal process in somatic cells. As a result, cells can only divide a certain number of times before the DNA loss prevents further division. (This is known as the Hayflick limit.) Within the germ cell line, which passes DNA to the next generation, **telomerase** extends the repetitive sequences of the telomere region to prevent degradation. Telomerase can become mistakenly active in somatic cells, sometimes leading to **cancer** formation.

Additionally, to aid termination, the progress of the DNA replication fork must stop or be blocked. Essentially, there are two methods that organisms do this:

- First: it is to have a termination site sequence in the DNA.
- Second: it is to have a protein which binds to this sequence to physically stop DNA replication proceeding. This is named the DNA replication terminus site-binding protein or in other words, **Ter protein**.

❖ **Prokaryote:**

Because bacteria have circular chromosomes, termination of replication occurs when the two replication forks meet each other on the opposite end of the parental chromosome. *E coli* regulate this process through the use of termination sequences that, when bound by the **Tus protein**, enable only one direction of replication fork to pass through. As a result, the replication forks are constrained to always meet within the termination region of the chromosome.

Speed of Replication

◈ Eukaryotes

The average human chromosome contains 150 x 10^6 nucleotide pairs which are copied at about **50 base pairs per second.** The process would take a **month** (rather than the hour it actually does) but for the fact that there are many places on the eukaryotic chromosome where replication can begin. Replication begins at some replication origins earlier in S phase than at others, but the process is completed for all by the end of S phase. As replication nears completion, "bubbles" of newly replicated DNA meet and fuse, finally forming two new molecules.

Multiple Origins in Eukaryotes

Origins initiate replication at different times.

◈ Prokaryote

The single molecule of DNA that is the *E. coli* genome contains 4.7 x 10^6 nucleotide pairs. DNA replication begins at a single, fixed location in this molecule, the **replication origin**, proceeds at about **500 nucleotides per second**, and thus is done in no more than **40 minutes**. The process includes a "proof-reading" function, the job is done with only about one incorrect nucleotide for every 10^9 nucleotides inserted. In other words, more often than not, the *E. coli* genome (4.7 x 10^6) is copied without error!

Proofreading:

Proofreading is the error-correcting processes during DNA replication.

- In bacteria, all three DNA polymerases (I, II, and III) have the ability to proofread, using 3'----> 5' exonuclease activity. When an incorrect base pair is recognized, DNA polymerase reverses its direction by one base pair of DNAs and excises the mismatched base. Following base excision, the polymerase can re-insert the correct base and replication can continue.
- In eukaryotes only the polymerases that deal with the elongation (γ, δ and ε) have proofreading ability (3'--->5' exonuclease activity).

Paternity Testing

A paternity test is established to determine the biological father of a child likewise a maternity test is to prove the birth mother of a child. For obvious reasons paternity tests are much more common than maternity tests.

In a DNA paternity test, DNA samples from two possible fathers and the mother are compared with the offspring's DNA. In this procedure, the samples are digested with a type of enzyme that cuts DNA at specific sequences. The digested DNA is loaded onto a gel and separated according to size, by gel electrophoresis.

Every band of the offspring's DNA must match a band in at least one of its parents'. Let's first consider the offspring matches with the mother. Three bands match the mother. The three remaining bands must be shared

with the father. Only one of the possible fathers will share the three bands with the offspring. Therefore, this must be the biological father.

Lec. 6

Polymerase chain reaction

The **polymerase chain reaction** (**PCR**) is a scientific technique in molecular biology to amplify a single or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.

The method relies on:

- Thermal cycling, consisting of cycles of repeated heating and cooling of the reaction for DNA melting and enzymatic replication of the DNA.
- Primers (short DNA fragments) containing sequences complementary to the target region along with a DNA polymerase are key components to enable selective and repeated amplification.

As PCR progresses, the DNA generated is itself used as a template for replication, setting in motion a chain reaction in which the DNA template is exponentially amplified. PCR can be extensively modified to perform a wide array of genetic manipulations.

Primer is a strand of nucleic acid that serves as a starting point for DNA synthesis. They are required because the enzymes that catalyze replication, DNA

polymerases, can only add new nucleotides to an existing strand of DNA. The polymerase starts replication at the 3'-end of the primer, and copies the complementary strand.

In most cases of natural DNA replication, the primer for DNA synthesis and replication is a short strand of RNA .

Many of the laboratory techniques of biochemistry and molecular biology that involve DNA polymerase, such as DNA sequencing and the polymerase chain reaction (PCR), require DNA primers. These primers are usually **short**, chemically **synthesized oligonucleotides**, the length of primers is usually not more than 30 (usually 18–24) nucleotides. They are **hybridized** to a target DNA, which is then copied by the polymerase.

Almost all PCR applications employ a heat-stable **DNA polymerase**, such as Taq polymerase, an enzyme originally isolated from the bacterium ***Thermus aquaticus***. This DNA polymerase enzymatically assembles a new DNA strand from DNA building blocks, the nucleotides, by using single-stranded DNA as a template and DNA primers, which are required for initiation of DNA synthesis.

The vast majority of PCR methods use thermal cycling, alternately heating and cooling the PCR sample to a defined series of temperature steps. These thermal cycling steps are necessary first to physically separate the two strands in a DNA double helix at a high temperature in a process called **DNA melting.** At a lower temperature, each strand is then used as the template in DNA synthesis by the DNA polymerase to selectively amplify the target DNA.

Developed in 1983 by **Kary Mullis**. In 1993, Mullis was awarded the Nobel Prize in Chemistry for his work on PCR. PCR is now a common and often indispensable technique used in medical and biological research labs for a variety of applications. These include:

- DNA cloning for sequencing.
- DNA-based phylogeny.
- Functional analysis of genes.
- The diagnosis of hereditary diseases.

- The identification of genetic fingerprints (used in forensic sciences and paternity testing).
- The detection and diagnosis of infectious diseases.

I. History of polymerase chain reaction

A paper in the Journal of Molecular Biology by Kleppe and co-workers (1971) first described a method using an enzymatic assay to replicate a short DNA template with primers *in vitro*. However, this early manifestation of the basic PCR principle did not receive much attention, and the invention of the polymerase chain reaction in 1983 is generally credited to Kary Mullis.

At the core of the PCR method is the use of a suitable DNA polymerase able to withstand the high temperatures of **>90 °C (194 °F)** required for separation of the two DNA strands in the DNA double helix after each replication cycle. The DNA polymerases initially employed for *in vitro* experiments expect PCR were unable to withstand these high temperatures. So the early procedures for DNA replication were very inefficient, time consuming, and required large amounts of DNA polymerase and continual handling throughout the process.

The discovery in 1976 of Taq polymerase — a DNA polymerase purified from the thermophilic bacterium, *Thermus aquaticus*, which naturally lives in hot (50 to 80 °C) (122 to 176 °F)) environments such as hot springs — flatten the way for dramatic improvements of the PCR method. The DNA polymerase isolated from *T. aquaticus* is stable at high temperatures remaining active even after DNA denaturation, thus elemiate the need to add new DNA polymerase after each cycle. This allowed an automated thermocycler-based process for DNA amplification.

When Mullis developed the PCR in 1983, he was working in Emeryville, California for Cetus Corporation, one of the first biotechnology companies. There, he was responsible for synthesizing short chains of DNA. Mullis has written that he think of PCR while travelling along the Pacific Coast Highway one night in his car. He was playing in his mind with a new way of analyzing changes (mutations) in DNA when he realized that he had instead invented a method of amplifying any DNA region through repeated cycles of duplication driven by DNA polymerase. In *Scientific American*, Mullis summarized the procedure: "Beginning with a single molecule of the genetic material DNA, the PCR can generate 100 billion similar

molecules in an afternoon. The reaction is easy to performe. It requires no more than a test tube, a few simple reagents, and a source of heat.

PCR principles:

PCR is used to amplify a specific region of a DNA strand (the DNA target). Most PCR methods typically amplify DNA fragments of up to ~10 kilo base pairs (kb), although some techniques allow for amplification of fragments up to 40 kb in size.

A basic PCR set up requires several components and reagents. These components include:

- **DNA template** that contains the DNA region (target) to be amplified.
- **Two primers** that are complementary to the 3' ends of each of the sense and anti-sense strand of the DNA target.
- **Taq polymerase** or another DNA polymerase with a temperature optimum at around 70 °C.
- **Deoxynucleotide triphosphates** (dNTPs), the building blocks from which the DNA polymerase synthesizes a new DNA strand.
- **Buffer solution**, providing a suitable chemical environment for optimum activity and stability of the DNA polymerase.
- **Divalent cations**, magnesium or manganese ions; generally Mg^{2+} is used, but Mn^{2+} can be utilized for PCR-mediated DNA mutagenesis, as higher Mn^{2+} concentration increases the error rate during DNA synthesis.
- **Monovalent cation** potassium ions.

The PCR is commonly carried out in a reaction volume of 10–200 µl in small reaction tubes (0.2–0.5 ml volumes) in a **thermal cycler**. The thermal cycler heats and cools the reaction tubes to achieve the temperatures required at each step of the reaction. Many modern thermal cyclers make use of the Peltier effect which permits both heating and cooling of the block holding the PCR tubes simply by reversing the electric current. Thin-walled reaction tubes permit favorable thermal conductivity to allow for rapid thermal equilibration. Most thermal cyclers have heated lids to prevent condensation at the top of the reaction tube. Older thermocyclers lacking a heated lid require a layer of oil on the top of the reaction mixture or a ball of wax inside the tube.

II. PCR Procedure:

Typically, PCR consists of a series of 20-40 repeated temperature changes, called cycles, with each cycle commonly consisting of three discrete temperature steps.

The cycling is often preceded by a single temperature step (called *hold*) at a high temperature (>90°C), then followed by one hold at the end for final product extension or brief storage.

The temperatures used and the length of time they are applied in each cycle depend on a variety of parameters. These include the enzyme used for DNA synthesis, the concentration of divalent ions and dNTPs in the reaction, and the melting temperature (Tm) of the primers.

Initialization step: This step consists of heating the reaction to a temperature of 94–96 °C (or 98 °C if extremely thermostable polymerases are used), which is held for 1–9 minutes. It is only required for DNA polymerases that require heat activation by hot-start PCR.

- **Denaturation step:** This step is the first regular cycling event and consists of heating the reaction to 94–98 °C for 20–30 seconds. It causes DNA melting of the DNA template by disrupting the hydrogen bonds between complementary bases, yielding single-stranded DNA molecules.

- **Annealing step**: The reaction temperature is lowered to 50–65 °C for 20–40 seconds allowing annealing of the primers to the single-stranded DNA template. Typically, the annealing temperature is about 3-5 degrees Celsius below the **Tm** of the primers used. Stable DNA-DNA hydrogen bonds are only formed when the primer sequence very closely matches the template sequence. The polymerase binds to the primer-template hybrid and begins DNA synthesis.

  *****{Melting temperature can be calculated using the **Wallace** formula:

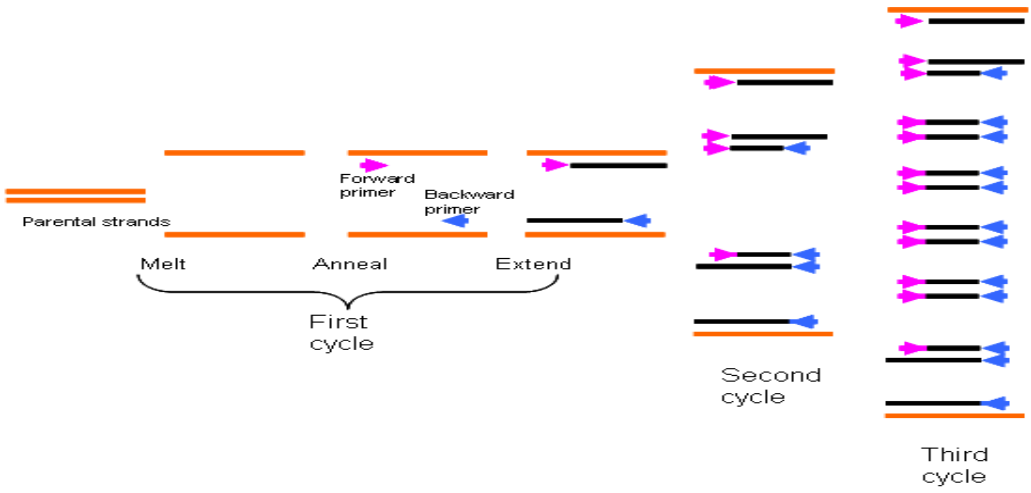
  $$T_m = 4\,(C+G) + 2\,(A+T)$$

- **Extension/elongation step:** The temperature at this step depends on the DNA polymerase used; Taq polymerase has its optimum activity temperature at 75–80 °C, and commonly a temperature of 72 °C is used with this enzyme. At this step the DNA polymerase synthesizes a new DNA strand complementary to the

DNA template strand by adding dNTPs that are complementary to the template in 5' to 3' direction, condensing the 5'-phosphate group of the dNTPs with the 3'-hydroxyl group at the end of the nascent (extending) DNA strand. The extension time depends both on the **DNA polymerase used** and on the **length of the DNA fragment** to be amplified. At its optimum temperature, the DNA polymerase will polymerize a **thousand bases per minute**. Under optimum conditions, if there are no limitations due to limiting substrates or reagents, at each extension step, the amount of DNA target is doubled, leading to exponential amplification of the specific DNA fragment.

Final elongation: This single step is occasionally performed at a temperature of 70–74 °C for 5–15 minutes after the last PCR cycle to ensure that any remaining single-stranded DNA is fully extended.

Final hold: This step at 4–15 °C for an indefinite time may be employed for short-term storage of the reaction.



To check whether the PCR generated the anticipated DNA fragment (also sometimes referred to as the amplimer or amplicon), agarose gel electrophoresis is employed for size separation of the PCR products. The size(s) of PCR products is determined by comparison with a DNA ladder (a molecular weight marker), which contains DNA fragments of known size; run on the gel alongside the PCR products (Fig. 2).

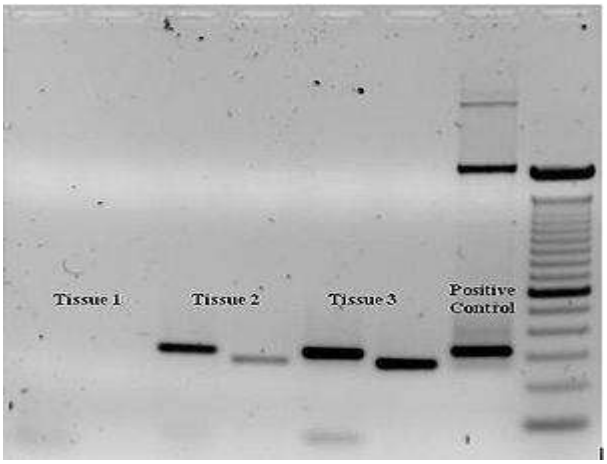


Figure 2: Ethidium bromide-stained PCR products after gel electrophoresis. Two sets of primers were used to amplify a target sequence from three different tissue samples. No amplification is present in sample #1; DNA bands in sample #2 and #3 indicate successful amplification of the target sequence. The gel also shows a positive control, and a DNA ladder containing DNA fragments of defined length for sizing the bands in the experimental PCRs.

III. PCR stages:

The PCR process can be divided into three stages:

First stage:

Exponential amplification: At every cycle, the amount of product is doubled (assuming 100% reaction efficiency). The reaction is very sensitive: only minute quantities of DNA need to be present.

Second stage:

Leveling off stage: The reaction slows as the DNA polymerase loses activity and as consumption of reagents such as dNTPs and primers causes them to become limiting.

Third stage:

Plateau: No more products accumulate due to exhaustion of reagents and enzyme.

IV. Polymerase chain reaction optimization

The polymerase chain reaction (PCR) is a commonly used molecular biology tool for amplifying DNA, and various techniques for **PCR optimization** have been developed by molecular biologists to improve PCR performance and minimize failure.

◈ Contamination and PCR

The PCR method is extremely sensitive, requiring only a few DNA molecules in a single reaction for amplification. Therefore, adequate measures to avoid contamination from any DNA present in the lab environment (bacteria, viruses, or human sources) are required. Because products from previous PCR amplifications are a common source of contamination, many molecular biology labs have implemented procedures that involve dividing the lab into **separate areas**. One lab area is dedicated to **preparation and handling of pre-PCR reagents** and the **setup of the PCR reaction**, and another area to **post-PCR processing**, such as gel electrophoresis or PCR product purification. For the setup of PCR reactions, many standard operating procedures involve using pipettes with **filter tips** and wearing fresh laboratory **gloves**, and in some cases a **laminar flow cabinet with UV lamp** as a work station formation is routinely assessed with a (**negative) control** PCR reaction. This control reaction is set up in the same way as the experimental PCRs, but without template DNA added, and is performed alongside the experimental PCRs.

◈ Hairpins

Secondary structures in the DNA can result in folding or knotting of DNA template or primers, leading to decreased product yield or failure of the reaction. Hairpins, which consist of internal folds caused by base-pairing between nucleotides in inverted repeats within single-stranded DNA, are common secondary structures and may result in failed PCRs.

Typically, primer design that includes a check for potential secondary structures in the primers, or addition of DMSO or glycerol to the PCR to minimize secondary structures in the DNA template are used in the optimization of PCRs that have a history of failure due to suspected DNA hairpins.

Polymerase errors

Taq polymerase lacks a 3' to 5' exonuclease activity. Thus, Taq has no error-proof-reading activity, which consists of excision of any newly miss-incorporated nucleotide base from the nascent (=extending) DNA strand that does not match with its opposite base in the complementary DNA strand. The lack in 3' to 5' proofreading of the Taq enzyme results in a high error rate (mutations per nucleotide per cycle) of approximately 1 in 10,000 bases, which affects the fidelity of the PCR, especially if errors occur early in the PCR with low amounts of starting material, causing accumulation of a large proportion of amplified DNA with incorrect sequence in the final product.

Several "high-fidelity" DNA polymerases, having engineered 3' to 5' exonuclease activity, have become available that permit more accurate amplification for use in PCRs for sequencing or cloning of products. Examples of polymerases with 3' to 5' exonuclease activity include:

- KOD DNA polymerase, a recombinant form of *Thermococcus kodakaraensis*
- KOD1; Vent, which is extracted from *Thermococcus litoralis*;
- Pfu DNA polymerase, which is extracted from *Pyrococcus furiosus*;
- Pwo, which is extracted from *Pyrococcus woesii*.

Magnesium concentration

Magnesium (Mg) is required as a co-factor for thermostable DNA polymerase. Taq polymerase is Mg-dependent enzyme and determining the optimum concentration to use is critical to the success of the PCR reaction. Some of the components of the reaction mixture such as **template concentration**, **dNTPs** and the presence of **chelating agents (EDTA**) or **proteins** can reduce the amount of free (Mg) present thus reducing the activity of the enzyme. Primers which bind to incorrect template sites are stabilized in the presence of excessive (Mg) concentrations and so results in decreased specificity of the reaction. Excessive (Mg) concentrations also stabilize double stranded DNA and prevent complete denaturation of the DNA during PCR reducing the product yield. Inadequate thawing of $MgCl_2$ may result in the formation of concentration gradients within the magnesium chloride solution

supplied with the DNA polymerase and also contribute too many failed experiments.

Size and other limitations

PCR works readily with a DNA template of up to two to three thousand base pairs in length. However, above this size, product yields often decrease, as with increasing length stochastic effects such as premature termination by the polymerase begin to affect the efficiency of the PCR. It is possible to amplify larger pieces of up to 50,000 base pairs with a slower heating cycle and special polymerases. These are polymerases fused to a processivity -enhancing DNA-binding protein, enhancing adherence of the polymerase to the DNA.

Non-specific priming

Non-specific binding of primers frequently occurs and can be due to repeat sequences in the DNA template, non-specific binding between primer and template, and incomplete primer binding, leaving the 5' end of the primer unattached to the template. Non-specific binding is also often increased when degenerate primers are used in the PCR. Manipulation of annealing temperature and magnesium ion (which stabilize DNA and RNA interactions) concentrations can increase specificity. Non-specific priming during reaction preparation at lower temperatures can be prevented by using "hot-start" polymerase enzymes whose active site is blocked by an antibody or chemical that only dislodges once the reaction is heated to 95˚C during the denaturation step of the first cycle.

Other methods to increase specificity include Nested PCR and Touchdown PCR. The earliest steps of a touchdown PCR cycle have high annealing temperatures. The annealing temperature is decreased in increments for every subsequent set of cycles. The primer will anneal at the highest temperature which is least-permissive of nonspecific binding that it is able to tolerate.

Primer dimers

Annealing of the 3' end of one primer to itself or the second primer may cause primer extension, resulting in the formation of so-called primer dimers, visible as low-molecular-weight bands on PCR gels. Primer dimer formation often competes with formation of the DNA fragment of interest, and may be avoided using primers

that are designed such that they lack complementarities—especially at the 3' ends—to itself or the other primer used in the reaction. If primer design is constraint by other factors and if primer-dimers do occur, methods to limit their formation may include optimization of the $MgCl_2$ concentration or increasing the annealing temperature in the PCR.

Deoxynucleotides concentrations:

Deoxynucleotides (dNTPs) may bind Mg^{2+} ions and thus affect the concentration of free magnesium ions in the reaction. In addition, excessive amounts of dNTPs can increase the error rate of DNA polymerase and even inhibit the reaction. An imbalance in the proportion of the four dNTPs can result in misincorporation into the newly formed DNA strand and contribute to a decrease in the fidelity of DNA polymerase.

V. Polymerase chain reaction inhibitors

PCR inhibitors usually affect PCR through interaction with DNA or interference with the DNA polymerase. Inhibitors can escape removal during the DNA purification procedure by binding directly to single or double-stranded DNA. Alternatively, by reducing the availability of cofactors (such as Mg^{2+}) or otherwise interfering with their interaction with the DNA polymerase, PCR is inhibited. In a multiplex PCR reaction, it is possible for the different sequences to suffer from different inhibition effects to different extents, leading to disparity in their relative amplifications.

Mechanisms of inhibition

A multitude of mechanisms can cause a PCR to fail, such as binding of an inhibitor to the active site of a DNA polymerase. Other reasons include sequestration of essentials co-factors like Mg^{2+}, e.g. by chelators like EDTA used in TE buffer. Many PCR also fail due to difficult template sequences, for example GC-rich, that cause the polymerisation process to be inefficient or even fail. In many cases, several factors only decrease the efficiency of the PCR but together lead the PCR product falling below the detection threshold.

- **Inhibitors from the purification process**

  - Alcohols like **ethanol** and **2-propanol** (isopropanol) from DNA template precipitation.
  - Organic solvents like **phenol** from phenol/chloroform purification.
  - Salts like **KCl**, **NaCl** from precipitation.
  - Detergents like **SDS**, from membrane lysis.

- **Inhibitors from the source tissue**

  - Components of blood: heme, hemoglobin, lactoferrin, immunoglobin G (IgG).
  - Liver, digestive tract, feces: bile salts, polysaccharides.
  - Connective tissue, skin: collagen, melanin.
  - Urine: urea.

Lec. 7

Quantifying extent of inhibition

In order to try to assess the extent of inhibition that occurs in a reaction, a control can be performed by adding a known amount of a template to the investigated reaction mixture (based on the sample under analysis). By comparing the amplification of this template in the mixture to the amplification observed in a separate experiment in which the same template is used in the absence of inhibitors, the extent of inhibition in the investigated reaction mixture can be concluded. Of course, if any part of the inhibition occurring in the sample-derived reaction mixture is sequence-specific, then this method will yield an under estimate of the inhibition as it applies to the investigate sequence(s).

Preventing PCR inhibition

1. Sample collection

The method of sample acquisition can be refined to avoid unnecessary collection of inhibitors. For example, in forensics, swab-transfer of blood on fabric or saliva on food, may prevent or reduce contamination with inhibitors present in the fabric or food.

2. DNA purification

Techniques exist and kits are commercially available to enable extraction of DNA to the exclusion of some inhibitors.

3. PCR reaction components

As well as methods for the **removal** of inhibitors from samples before PCR, some DNA polymerases offer varying **resistance** to different inhibitors and increasing the concentration of the chosen DNA polymerase also confers some resistance to polymerase-targeted inhibitors.

For PCR based on blood samples, the addition of bovine serum albumin reduces the effect of some inhibitors on PCR.

VI. Application of PCR

Selective DNA isolation

PCR allows **isolation of DNA fragments** from genomic DNA by selective **amplification of a specific region of DNA**. This use of PCR enhance many methods, such as generating hybridization probes for Southern or northern hybridization and DNA cloning, which require larger amounts of DNA, representing a specific DNA region. PCR supplies these techniques with high amounts of pure DNA, enabling analysis of DNA samples even from very small amounts of starting material.

Other applications of PCR include **DNA sequencing** to determine unknown PCR-amplified sequences in which one of the amplification primers may be used in Sanger sequencing, isolation of a DNA sequence to expedite recombinant DNA

technologies involving the insertion of a DNA sequence into a plasmid or the genetic material of another organism.

PCR may also be used for **genetic fingerprinting**; a forensic technique used to identify a person or organism by comparing experimental DNAs through different PCR-based methods.

Some PCR 'fingerprints' methods have high discriminative power and can be used to identify genetic relationships between individuals, such as parent-child or between siblings (brothers and sisters from the same parents) and are used in paternity testing (Fig. 3).

Amplification and quantification of DNA

Because PCR amplifies the regions of DNA that it targets, PCR can be used to analyze extremely small amounts of sample. This is often critical for forensic analysis, when only a trace amount of DNA is available as evidence. PCR may also be used in the analysis of ancient DNA that is tens of thousands of years old. These PCR-based techniques have been successfully used on animals, such as a forty-thousand-year-old mammoth, and also on human DNA, in applications ranging from the analysis of Egyptian mummies to the identification of a Russian tsar.

Quantitative PCR methods allow the estimation of the amount of a given sequence present in a sample—a technique often applied to quantitatively determine levels of gene expression. Real-time PCR is an established tool for DNA quantification that measures the accumulation of DNA product after each round of PCR amplification.

PCR in diagnosis of diseases

PCR permits **early diagnosis of malignant diseases** such as leukemia and lymphomas, which is currently the highest developed in cancer research and is already being used routinely. PCR assays can be performed directly on genomic

DNA samples to **detect translocation-specific malignant cells** at a sensitivity which is at least 10,000 fold higher than other methods.

PCR also permits identification of **non-cultivatable or slow-growing microorganisms** such as mycobacteria, anaerobic bacteria, or viruses from tissue culture assays and animal models. The basis for PCR diagnostic applications in microbiology is the detection of infectious agents and the **discrimination of non-pathogenic from pathogenic strains** by virtue of specific genes.

Viral DNA can likewise be detected by PCR. The primers used need to be specific to the targeted sequences in the DNA of a virus, and the PCR can be used for **diagnostic analyses or DNA sequencing of the viral genome.** The high sensitivity of PCR permits virus detection soon after infection and even before the onset of disease. Such early detection may give physicians a significant lead in treatment. The amount of virus ("viral load") in a patient can also be quantified by PCR-based DNA quantitation techniques.

Lec. 8 *Dr. Khawla Hori*

Transcription

Transcription is the process of creating a complementary RNA copy of a sequence of DNA. Both RNA and DNA are nucleic acids, which use base pairs of nucleotides as a complementary language that can be converted back and forth from DNA to RNA by the action of the correct enzymes.

During transcription, a DNA sequence is read by an RNA polymerase, which produces a complementary, anti-parallel RNA strand. As opposed to DNA

replication, transcription results in an RNA complement that includes **uracil (U)** in all instances where **thymine (T)** would have occurred in a DNA complement. Also unlike DNA replication where DNA is synthesized, transcription **does not involve an RNA primer** to initiate RNA synthesis.

Transcription is explained easily in 4 or 5 steps, each moving like a wave along the DNA:

1. RNA polymerase moves the transcription bubble, a stretch of unpaired nucleotides, by breaking the hydrogen bonds between complementary nucleotides.
2. RNA polymerase adds matching RNA nucleotides that are paired with complementary DNA bases.
3. RNA sugar-phosphate backbone forms with assistance from RNA polymerase.
4. Hydrogen bonds of the untwisted RNA + DNA helix break, freeing the newly synthesized RNA strand.
5. If the cell has a nucleus, the RNA is further processed (addition of a 3' poly-A tail and a 5' cap) and exits through to the cytoplasm through the nuclear pore complex.

- Transcription is the first step leading to **gene expression**. The stretch of DNA transcribed into an RNA molecule is called a **transcription unit** and **encodes at least one gene**. If the gene transcribed encodes a protein, the result of transcription is messenger RNA (mRNA), which will then be used to create that protein via the process of translation. Alternatively, the transcribed gene may encode for either non-coding RNA genes (such as microRNA, lincRNA, etc.) or ribosomal RNA (rRNA) or transfer RNA (tRNA), other components of the protein-assembly process, or other ribozymes.
- A DNA transcription unit encoding for a protein contains not only the sequence that will eventually be directly translated into the protein (**coding sequence**) but also **regulatory sequences** that direct and regulate the synthesis of that protein. The regulatory sequence before (upstream from) the coding sequence is called the **five prime un-translated region (5'UTR)**, and the sequence following (downstream from) the coding sequence is called the **three prime un-translated region (3'UTR).**

- Transcription has some **proofreading mechanisms**, but they are fewer and less effective than the controls for copying DNA; therefore, transcription has a **lower copying fidelity** than DNA replication.
- As in DNA replication, DNA is read from 3' → 5' during transcription. Meanwhile, the complementary RNA is created from the 5' → 3' direction. This means its 5' end is created first in base pairing. Although DNA is arranged as two anti-parallel strands in a double helix, only one of the two DNA strands, called the **template (none coding) strand**, is used for transcription. This is because RNA is only single-stranded, as opposed to double-stranded DNA. The other DNA strand is called the **lagging (coding) strand**, because its sequence is the same as the newly created RNA transcript (except for the substitution of uracil for thymine). The use of only the 3' → 5' strand eliminates the need for the Okazaki fragments seen in DNA replication.

RNA Polymerases

A. Prokaryotes RNA polymerase consists of:

1. **Core enzyme** = β **β'** α_2 polypeptide sub units has catalytic activity but cannot recognize start site of transcription
   - ~500,000 Daltons
   - requires **Mg2+** for activity
   - **β'** binds **2 Zn** atoms
2. **Sigma factor (σ) =** binds RNA pol. to promoter site in DNA.

Holoenzyme = core enzyme + sigma factor, carries out four functions:
(i) Template binding.
(ii) RNA chain initiation.
(iii) Chain elongation.
(iv) Chain termination.

Prokaryotes are generally in the ~106 bp size range. Eukaryotes are more in the ~109 bp size range. Larger genome means it requires more specificity. Also the diversity of function – organelles, different cell type, and so on.

B. Eukaryotic RNA polymerases.

I. There are 3 **Nuclear** RNA polymerases

a. RNA **polymerase** I- transcribes rRNA genes.

b. RNA **polymerase** II - transcribes mRNA genes.

c. RNA **polymerase** III - transcribes tRNA, 5S rRNA, and other small RNA genes.

II. Have 10-17 different subunits, large multi-subunit complexes are functionally similar to E. coli RNA polymerase.

III. Cannot bind to their respective promoters alone but requires transcription factor for promoter specific recruitment.

VI. Organelle specific RNA polymerases more prokaryotic-like:

1. Chloroplast
2. Mitochondria.

RNA polymerase II

1. Core subunits - have sequence similarity to the core subunits of E. coli core RNA polymerase or subunits of other eukaryotic RNA polymerases.
2. Shared or common subunits (same subunits) found in RNA Polymerase III and II or in RNA polymerase I and RNA Polymerase II.
3. Unique subunits - no similar homolog found anywhere else.
4. C-terminal domain (CTD) of the largest subunit.
5. Cycling of phosphorylated and dephosphorylated forms of RNA Pol. II associated with different stages in transcription.
6. CTD is also required to recruit proteins for capping of 5'-end of mRNA, as well as for splicing and polyadenylation of the 3'end of mRNA.

Transcription is divided into **5** stages: ***pre-initiation, initiation, promoter clearance, elongation*** **and** ***termination*****.**

Pre-initiation

In eukaryotes, RNA polymerase, and therefore the initiation of transcription, requires the presence of a **core promoter sequence** in the DNA.

Promoters are regions of DNA that promote transcription and, in eukaryotes, are found at **-30, -75, and -90 base pairs** upstream from the transcription start site (**TSS**). **Core promoters** are sequences within the promoter that are essential for transcription initiation. RNA polymerase is able to bind to core promoters in the presence of various specific transcription factors.

The most characterized type of core promoter in eukaryotes is a short DNA sequence known as a **TATA box,** found 25-30 base pairs upstream from the TSS. The TATA box, as a core promoter, is the binding site for a transcription factor known as TATA-binding protein (**TBP**), which is itself a subunit of another transcription factor, called Transcription Factor II D (**TFIID**).

After TFIID binds to the TATA box via the TBP, five more transcription factors and RNA polymerase combine around the TATA box in a series of stages to form a pre-initiation complex.

One transcription factor, **Transcription factor II H**, has two components with **helicase** activity and so is involved in the separating of opposing strands of double-stranded DNA to form the initial transcription bubble. However, only a low, or basal, rate of transcription is driven by the pre-initiation complex alone. Other proteins known as activators and repressors, along with any associated coactivators or corepressors, are responsible for modulating transcription rate.

Thus, pre-initiation complex contains:

1. Core Promoter Sequence.
2. Transcription Factors.
3. RNA Polymerase.
4. Activators and Repressors.

The transcription pre-initiation in **archaea** is, in essence, homologous to that of eukaryotes, but is much less complex. The archaeal pre-initiation complex assembles at a TATA-box binding site; however, in archaea, this complex is composed of only **RNA polymerase II, TBP, and TFB** (the archaeal homologue of eukaryotic transcription TFIIB).

- **General transcription factors**

❖ Initiation

In bacteria, transcription begins with the binding of RNA polymerase to the promoter in DNA. RNA polymerase is a core enzyme consisting of five subunits: **2 α subunits, 1 β subunit, 1 β' subunit, and 1 σ subunit**. At the start of initiation, the core enzyme is associated with a **sigma factor** that aids in finding the appropriate -35 and -10 base pairs downstream of promoter sequences. When the sigma factor and RNA polymerase combine, they form a **holoenzyme.**

Transcription initiation is more complex in eukaryotes. Eukaryotic RNA polymerase does not directly recognize the core promoter sequences. Instead, a collection of proteins called transcription factors mediate the binding of RNA polymerase and the initiation of transcription. Only after certain transcription factors are attached to the promoter does the RNA polymerase bind to it. The completed assembly of transcription factors and RNA polymerase bind to the promoter, forming a transcription initiation complex. Transcription in the archaea domain is similar to transcription in eukaryotes.

➢ Promoter clearance

After the first bond is synthesized, the RNA polymerase must clear the promoter. During this time there is a tendency to release the RNA transcript and produce truncated transcripts. This is called **abortive initiation** and is common for both eukaryotes and prokaryotes. Abortive initiation continues to occur until the **σ factor** rearranges, resulting in the transcription elongation complex (which gives a 35 bp moving footprint). The σ factor is released before 80 nucleotides of mRNA are synthesized. Once the transcript reaches approximately 23 nucleotides, it no longer slips, and elongation can occur. This, like most of the remainder of transcription, is an energy-dependent process, consuming adenosine triphosphate (ATP).

Elongation

One strand of the DNA, the **template (noncoding) strand**, is used as a template for RNA synthesis. As transcription proceeds, RNA polymerase traverses the template strand and uses base pairing complementarities with the DNA template to create an RNA copy. Although RNA polymerase traverses the template strand from 3' → 5', the coding (non-template) strand and newly-formed RNA can also be used as reference points, so transcription can be described as occurring 5' → 3'. This produces an RNA molecule from 5' → 3', an exact copy of the coding strand (except that thymine's are replaced with uracil's, and the nucleotides are composed of a ribose (5-carbon) sugar where DNA has deoxyribose in its sugar-phosphate backbone).

Unlike DNA replication, mRNA transcription can involve multiple RNA polymerases on a single DNA template and multiple rounds of transcription (amplification of particular mRNA), so many mRNA molecules can be rapidly produced from a single copy of a gene.

Elongation also involves a proofreading mechanism that can replace incorrectly incorporated bases. In eukaryotes, this may correspond with short pauses during transcription that allow appropriate RNA editing factors to bind. These pauses may be intrinsic to the RNA polymerase or due to chromatin structure.

Termination

- Bacteria use two different strategies for transcription termination.

1. Rho-independent transcription.

2. Rho-dependent transcription.

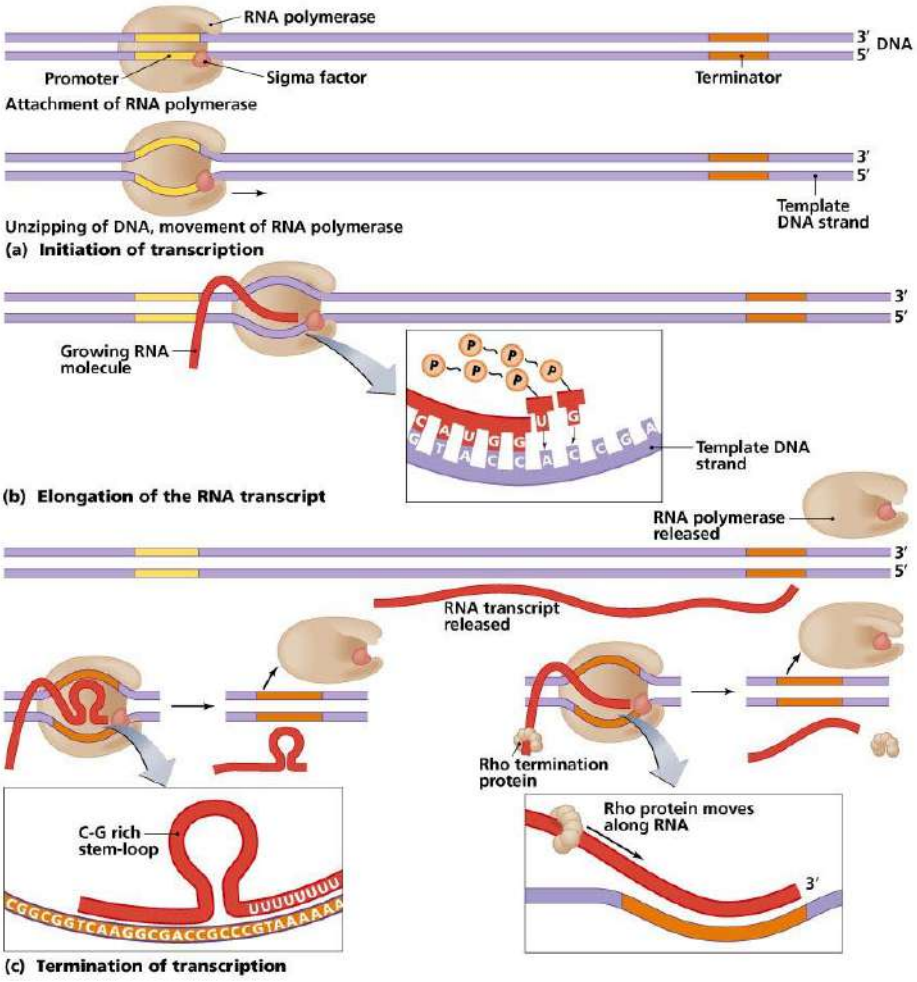
In **Rho-independent** transcription termination, also called intrinsic termination, RNA transcription stops when the newly synthesized RNA molecule forms a **G-C-rich hairpin loop** followed by a run of **Us**.

When the hairpin forms, the mechanical stress breaks the weak RNA (**U**)-DNA (**A**) bonds, now filling the DNA-RNA hybrid. This pulls the poly-U transcript

out of the active site of the RNA polymerase, in effect, terminating transcription.

In the "**Rho-dependent**" type of termination, a protein factor called "Rho" destabilizes the interaction between the template and the mRNA, thus releasing the newly synthesized mRNA from the elongation complex.

Transcription termination in eukaryotes is less understood but involves **cleavage** of the new transcript followed by template-independent addition of **As** at its new 3' end, in a process called **polyadenylation.**



❖ Post-transcriptional modification

Processing of mRNA differs greatly among eukaryotes, bacteria, and archaea. Non-eukaryotic mRNA is mature upon transcription and requires

no processing, except in rare cases. Eukaryotic pre-mRNA, however, requires extensive processing.

I. 5' cap addition

A **5' cap** (also termed an RNA cap, an RNA 7-methylguanosine cap, or an RNA m^7G cap) is a modified guanine nucleotide that has been added to the "front" or **5' end** of a eukaryotic mRNA shortly after the start of transcription. The 5' cap consists of a terminal **7-methylguanosine** residue that is linked through a **5'-5'-triphosphate bond** to the first transcribed nucleotide. Its presence is **critical for recognition by the ribosome and protection from RNases.**

II. Splicing

Splicing is the process by which pre-mRNA is modified to remove certain stretches of non-coding sequences called **introns**; the stretches that remain include protein-coding sequences and are called **exons.** Sometimes pre-mRNA messages may be spliced in several different ways, allowing a single gene to encode multiple proteins. This process is called **alternative splicing**. Splicing is usually performed by an RNA-protein complex called the **spliceosome.**

III. Polyadenylation

Polyadenylation is the covalent linkage of a polyadenylyl moiety to mRNA molecule. In eukaryotic organisms, with the exception of histones, all mRNA molecules are **polyadenylated at the 3' end**. The poly(A) tail and the protein bound to it aid in **protecting mRNA from degradation by exonucleases.** Polyadenylation is also important for **transcription termination, export of the mRNA from the nucleus, and translation.** mRNA can also be polyadenylated in prokaryotic organisms, where poly(A) tails act to facilitate, rather than impede, exonucleolytic degradation.

Polyadenylation occurs during and immediately after transcription of DNA into RNA. After the mRNA has been cleaved, around **250 adenosine residues are added to the free 3' end at the cleavage site**. This reaction is catalyzed by **polyadenylate polymerase**.

Measuring and detecting transcription

Transcription can be measured and detected in a variety of ways:

- Nuclear Run-on assay: measures the relative abundance of newly formed transcripts
- RNase protection assay and ChIP-Chip of RNAP: detect active transcription sites
- RT-PCR: measures the absolute abundance of total or nuclear RNA levels.
- DNA microarrays: measures the relative abundance of the global total or nuclear RNA levels.
- In situ hybridization: detects the presence of a transcript
- MS2 tagging: by incorporating RNA stem loops, such as MS2, into a gene; these become incorporated into newly synthesized RNA. The stem loops can then be detected using a fusion of GFP and the MS2 coat protein, which has a high affinity, sequence-specific interaction with the MS2 stem loops. The recruitment of GFP to the site of transcription is visualized as a single fluorescent spot. This remarkable new approach has revealed that transcription occurs in discontinuous bursts, or pulses. With the notable exception of in situ techniques, most other methods provide cell population averages, and are not capable of detecting this fundamental property of genes.
- Northern blot: the traditional method, and until the advent of RNA-Seq, the most quantitative
- RNA-Seq: applies next-generation sequencing techniques to sequence whole transcriptomes, which allows the measurement of relative abundance of RNA, as well as the detection of additional variations such as fusion genes, post-translational edits and novel splice sites.

Transcription factories

Active transcription units are clustered in the nucleus, in discrete sites called transcription factories or euchromatin. Such sites can be visualized by allowing engaged polymerases to extend their transcripts in tagged precursors (Br-UTP or Br-U) and immuno-labeling the tagged nascent RNA. Transcription factories

can also be localized using fluorescence in situ hybridization or marked by antibodies directed against polymerases. There are ~10,000 factories in the nucleoplasm of a HeLa cell, among which are ~8,000 polymerase II factories and ~2,000 polymerase III factories. Each polymerase II factory contains ~8 polymerases. As most active transcription units are associated with only one polymerase, each factory usually contains ~8 different transcription units. These units might be associated through promoters and/or enhancers, with loops forming a 'cloud' around the factor.

- Some viruses (such as HIV, the cause of AIDS), have the ability to transcribe RNA into DNA.
- HIV has an RNA genome that is duplicated into DNA. The resulting DNA can be merged with the DNA genome of the host cell.
- The main enzyme responsible for synthesis of DNA from an RNA template is called **reverse transcriptase**. In the case of HIV, reverse transcriptase is responsible for synthesizing a complementary DNA strand (cDNA) to the viral RNA genome.
- An associated enzyme, **ribonuclease H**, digests the RNA strand, and reverse transcriptase synthesizes a complementary strand of DNA to form a double helix DNA structure.
- This cDNA is integrated into the host cell's genome via another enzyme (**integrase**) causing the host cell to generate viral proteins that reassemble into new viral particles.
- In HIV, subsequent to this, the host cell undergoes programmed cell death, apoptosis of T cells. However, in other retroviruses, the host cell remains intact as the virus buds out of the cell.

Some eukaryotic cells contain an enzyme with reverse transcription activity called **telomerase**. Telomerase is a reverse transcriptase that lengthens the ends of linear chromosomes. Telomerase carries an RNA template from which it synthesizes DNA repeating sequence, or "junk" DNA. This repeated sequence of DNA is important because, every time a linear chromosome is duplicated, it is shortened in length. With "junk" DNA at the ends of chromosomes, the shortening eliminates some of the non-essential, repeated sequence rather than the protein-encoding DNA sequence farther away from the chromosome end.

Telomerase is often activated in **cancer cells** to enable cancer cells to duplicate their genomes indefinitely without losing important protein-coding DNA sequence. Activation of telomerase could be part of the process that allows cancer cells to become **immortal.**

Inhibitors

Transcription inhibitors can be used as antibiotics against, for example, pathogenic bacteria (**antibacterials**) and fungi (**antifungals**). An example of such an antibacterial is **rifampicin,** which inhibits prokaryotic DNA transcription into mRNA by inhibiting DNA-dependent RNA polymerase by binding its beta-subunit. **8-Hydroxyquinoline** is an antifungal transcription inhibitor.

Comparison of Transcription in Prokaryotes versus Eukaryotes

There are significant differences in the process of transcription in prokaryotes versus eukaryotes.

I. In prokaryotes, transcription occurs in the **cytoplasm**. Translation of the mRNA into proteins also occurs in the cytoplasm. In eukaryotes, transcription occurs in the cell's **nucleus**. mRNA then moves to the cytoplasm for translation.

II. DNA in prokaryotes is much more **accessible to RNA polymerase** than DNA in eukaryotes. Eukaryotic DNA is wrapped around proteins called histones to form structures called nucleosomes. **Eukaryotic DNA is packed** to form chromatin. While RNA polymerase interacts directly with prokaryotic DNA, other proteins mediate the interaction between RNA polymerase and DNA in eukaryotes.

III. mRNA produced as a result of transcription is **not modified** in prokaryotic cells. Eukaryotic cells modify mRNA by **RNA splicing, 5' end capping, and addition of a polyA tail.**

Lec 9

Protein Synthesis = Translation

Types of RNA:

RNA is all the same in that it is always:

- Composed of ribonucleotides,
- It contains uracil instead of thymine,
- Consists of a single strand (there are a few viral exceptions to this),
- It is synthesized using a gene (DNA) as a template (once again, there a few viral exceptions to this).

However, there are functionally three general categories of RNA. They have slight differences in structure that go along with their functions. Note also that within each of these categories, there are many versions, each with its own unique structure and properties.

Messenger RNA (mRNA):

mRNA has a central role in the production of proteins. mRNA carries the information found in a structural gene out of the nucleus (in the case of eukaryotic organisms) into the cytoplasm and to the ribosomes where other types of RNA (tRNA), the ribosomes themselves, and various enzymes cause appropriate amino acids to be lined up and polymerized into a protein. The mRNA molecule contains information that actually specifies the order of these amino acids. Since most proteins are large molecules consisting of several hundred or more amino acids, mRNAs also tend to be large molecules.

Since the mRNA contains a copy of the information found in the DNA, we often term it a "**transcript**" and the process of synthesizing mRNA as **transcription**. Transcription is no different than the RNA synthesis process. It is exactly the same process used in the synthesis of all functional types of RNA. **The only thing that makes one molecule mRNA and another tRNA is the function of the molecule** (which will

be reflected in its size, the nucleotide sequence, and the shape it assumes when it is completed – form and function again).

- The RNA polymerase reads a particular gene in the 3 to 5' direction
- The strand of RNA it synthesizes is made as the 5' side of the incoming NTP adds to the 3' end of the growing nucleic acid (5' to 3' addition).
- Thus, the resulting RNA molecules differ from each other according to differences in the DNA template used to direct their synthesis. In the drawing below, the gene template is shown at the bottom; the antiparallel strand is not shown:

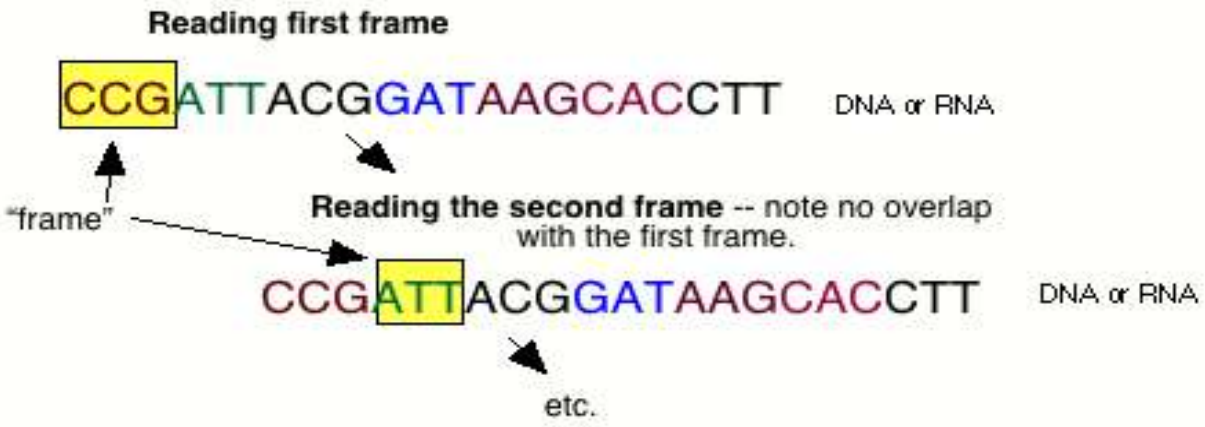
As you know, mRNA **contains information about the order of amino acids for a particular type of protein**. That is, the gene and then the mRNA molecule both contain a map of the **primary structure of a protein**. Recall that proteins are all made of amino acids of which there are 20 types.

In the 1950s one of the most important searches in molecular biology was to figure out how amino acids were specified to figure out the **basic arrangement of the nucleotide language.**

That is the **genetic code**:

- A certain **number of letters** would be joined together to stand for a certain "concept" – an **amino acid.** These sequences could be thought of as genetic code.
- They knew the structure of DNA and realized that, unlike human languages, different cods were not written with physical spaces between them. The reason for this expectation was that it was known that **in both DNA and RNA, sequential nitrogenous bases are the same distance apart**. Yet it still made sense that the codon would most likely **not overlap** each other anymore than they do in our written languages.
- The most logical conclusion was that a certain number of letters would be read as codon and then whatever it was that was doing the **reading would shift ahead an entire frame to the next codon**. Let's say the code were 3 letters long (which they turned out to be). Furthermore, let's assume that the **language consists of**

exactly 4 letters: A, C, G, and T (or U). The hypothesis as to how the information would be displayed and read is:



Now, why should the code be three long? The answer is that if they are any shorter it is not possible to specify all the information that we know genetic systems specify. And, if the codons are any longer, they will waste space, material, and energy. Here's the logic. Recall that since we are writing instructions to build proteins that we need 20 codons -- one for each amino acid.

Now, each nucleotide can be either A, C, G or T (U). How many unique codon can be produced using this 4 letter language if the words are always 3 letters long?

- If the codon was only **1 nucleotide** long, there would be **four possible codons** (A, C, G, or U).

- If they were **two nucleotides** long, there are a possible of **16 unique codons** (e.g., AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, and UU). Notice that the total number was equal to 4 length of the word which is $\mathbf{4^2 = 16}$ in this case.

- If the codons are actually **3 nucleotides** long, then there are a total of $\mathbf{4^3 = 64}$ **unique combinations.**

Notice that now we have **64 "codons" that we can use to code 20 amino acid**. Like most languages has more than one word for the same thing. **Many amino acids are specified by two or three different "codons".**

- 61 combinations specify amino acids (many amino acids **have more than one codon**). For example, **leucine,** have **six** different codons all of which will eventually produce a leucine (Leu) in the protein chain. There are also **six** for **serine** (Ser). In fact there are only two amino acids which have only **one** sequence of bases to code for them - **methionine** (Met) and **tryptophan** (Trp).

- 3 codons (**UAA, UAG, and UGA**) do not correspond to an amino acid and instead act as "**stop**" codons.

- The **triplet code** is "**degenerate**" because it has more than one codon for many amino acids. A better term is to call it **redundant** (excessive).

- The codon that marks the start of a protein chain is **AUG** the amino acid, methionine (Met). That ought to mean that every protein chain must start with methionine. That's not quite true because in some cases the methionine can get chopped off the chain after synthesis is complete.

- There are no "gaps" in RNA molecules between codons.

- The structural genes in eukaryotes have regions called "exons" and "introns". Both of these types of areas end up in the initial or "**primary RNA transcript**". The introns must all be excised before the primary transcript becomes a functional mRNA molecule.

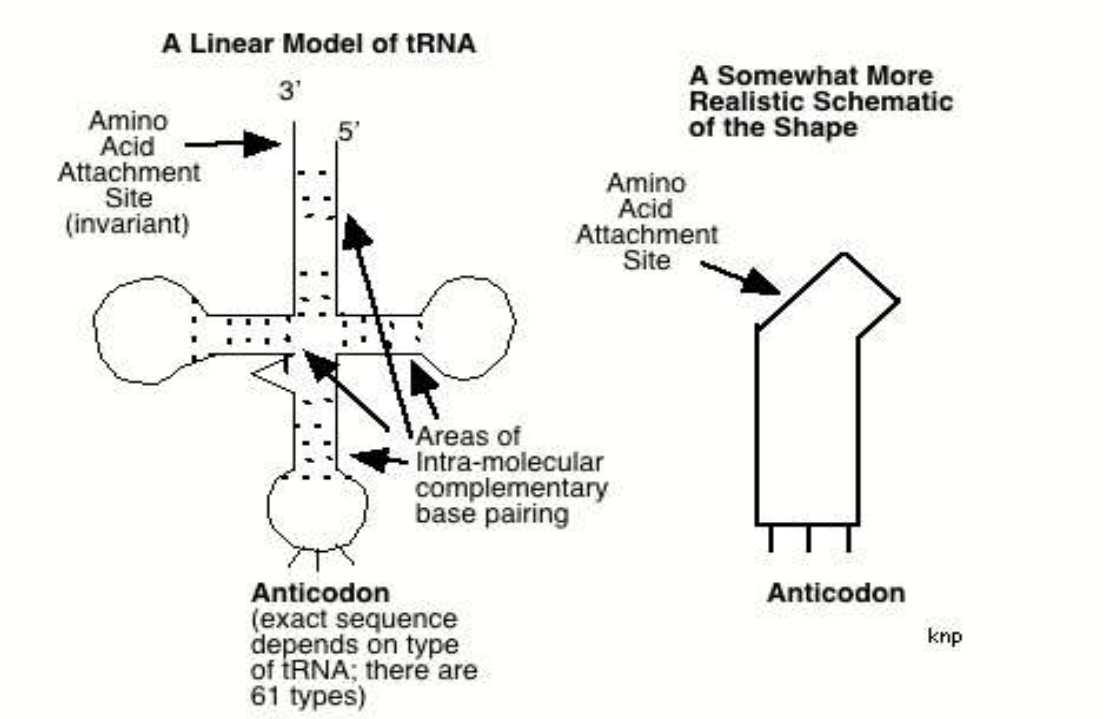
Transfer RNA (tRNA):

These are very short molecules that consist of only about 80 nucleotides. They are not straight, string-like molecules. As soon as they are synthesized, they tend to fold on themselves. This folding is caused by the fact that certain sequences that are some distance apart from each other are complementary. For example, the sequence:

This sort of thing is very common in RNA, especially in tRNA and ribosomal RNA.

In transfer RNA the result of this self-pairing is a molecule that has

(4) Paired areas and three loops. This general configuration, which is shared by all types of RNA molecules, is shown below.
Nevertheless, we will use the idea of an L structure because it greatly simplifies what we need to know about tRNA.



About this "L" shape, notice from the models above that the "top" is where the 3' and 5' ends is near together. This is **where an amino acid may be attached -- at the 3' end.** At the bottom of the tRNA molecule is a loop with something called an "**Anticodon**". This is a place where three nitrogenous bases stick more or less outwards from the molecule. This region is the part of the tRNA molecule that will interact with mRNA during protein synthesis.
Obviously, **a particular anticodon will only interact with its specific complementary mRNA codon .**
Moreover, the **anticodon is the central place for identifying the type of tRNA molecule**; in the "L" model, it is the bottom of the "L". So, the anticodon is used to identify:

- A particular type of tRNA.
- The type of amino acid that it carries.

Ribosomal RNA (rRNA):

These are **very large RNA molecules** with numerous folded and paired areas. They assume very complex shapes their shape determines their function. Most of the genes for the many types of rRNA are found near the part of the nucleus we call the **nucleolus.** These structures (there can be many in one nucleus) are most visible in cells that are **actively synthesizing proteins** and therefore that need large number of ribosomes. This leads us to ribosomes.

Ribosomes are small structures composed of a number of different types of structural **proteins, enzymes and rRNA**.

Functionally, they are the **sites of protein synthesis**. They are either found floating around more or less free in the cytoplasm or they are bound to the endoplasmic reticulum (rough endoplasmic reticulum or rough ER).

It is also worth noting once again that mitochondria produce their own ribosomes that although generally similar to those in the rest of the cell. **Mitochondrial ribosomes** are used to synthesize proteins that are coded for by the mitochondrial DNA.

All ribosomes are composed of two subunits called the **large** and **small** subunits:

- The small subunit has a binding site for mRNA.
- The large subunit has three regions called **A, P, and E** sites that are used to organize tRNA and the growing peptide chain.
- The large and small subunits normally only associate with each other during protein synthesis -- when it is complete, they dissociate.

Recycling: The "players" of protein synthesis (nucleotides, nucleic acids, proteins and amino acids) will have many lives. Keep in minds that not only are there many examples of each specific type of these chemicals but also that any individual molecule or structure will be used over and over again. Eventually, for some reason, each complex molecule (RNA or protein) will be broken down to its constituent parts. The most common fate of these parts (nucleotides or amino acids) will be recycled; it is far

less likely that they will be metabolized. This is exactly like the situation we saw with ATP, coenzymes, and enzymes.

An example; A molecule of a particular type of tRNA will be joined with the particular type of amino acid that it carries, and participate in protein synthesis where it loses its amino acid. It can then pick up another amino acid of the same type and once again participate in protein synthesis. This cycle will continue many times before this particular tRNA is broken down into NMP and these are then converted back to NTPs and used to synthesize new molecules of RNA.

In any cell that is capable of synthesizing proteins, there will already be plenty of tRNA of each type present, along with ribosome parts, all 20 types of amino acids, and plenty of NTPs for RNA synthesis. Moreover, at any time most of the tRNA molecules will already be carrying the type of amino acid for which they are specific.

Let's look at how amino acids come to be associated with a given type of tRNA:

Charge process of "tRNA molecules"

Production of Aminoacyl tRNAs:

When a given tRNA has the appropriate amino acid attached to it, it is called an **aminoacyl tRNA** or simply a "**charged tRNA**"(tRNA is carrying an amino acid).

This charging process is very important. For protein synthesis to work correctly, it is absolutely vital that the correct amino acid be attached to the correct tRNA – in other words, **the correct amino acid with the correct anticodon**. This vital task is performed by a class of more than 20 related enzymes called **aminoacyl-tRNA synthetases**.

Each species of aminoacyl-tRNA synthetase is specific for a particular type of amino acid and tRNA. Thus, the wrong amino acid cannot be put on a given type of tRNA.

Like any synthesis, the production of amino-acyl-tRNA requires an energy source. As usual, the energy source is ATP. Just as with nucleic acid synthesis, we will see that **2 ~P** are used to attach each amino acid. The result is that the process is very difficult to reverse and it requires lots of energy.
Here is an overview of the process of attaching one amino acid to its tRNA. Several steps are involved.

I. The first involves "activating the amino acid (AA):

AA + ATP + aminoacyl-tRNA synthetase -------->
AMP-AA..aminoacyl-tRNA synthetase + PPi

The pyrophosphate immediately is hydrolyzed (in the cytosol):

PPi ----------> 2 Pi – energy

This of course makes the previous step essentially irreversible.
When the stages above are complete, the appropriate tRNA attaches to the enzyme's active site next to the AMP-AA (already present from the previous step). The shape of the active site prevents anything but the correct amino acid and tRNA from coming together:

AMP-AA..aminoacyl-tRNA synthetase + tRNA ----->
AMP + aminoacyl-tRNA + aminoacyl-tRNA synthetase

The amino acid attaches to the 3' end of the tRNA at its carboxylic acid end. Thus, the amine end is left free.
There are a couple of additional things to realize about this reaction:

1. Notice how directly the enzyme (aminoacyl-tRNA synthetase) covalently **binds** to the AMP-aminoacyl. Actually, this sort of combination is common mechanism in catalyzed reactions

2. Notice that as usual the enzyme is not lost in the process -- by the end of the process, the **enzyme is ready to be used for another round**. On the other hand, the **substrates have definitely changed** -- ATP

was degraded to AMP and 2 Pi and a tRNA and amino acid have been joined.

Processes in the Nucleus – RNA synthesis and Processing

Primary Transcripts:

Under appropriate conditions RNA molecules are transcribed from genes that ultimately code for proteins. In eukaryotes these **primary transcripts** contain both introns and exons -- they contain a transcript of the entire gene.

Introns correspond to nucleotide sequences that do not code for anything. Introns need to be removed before protein synthesis.

mRNA Processing:

Before the primary transcript mRNA leaves the nucleus, it is intercepted by a series of enzymes and RNA molecules. There are three general processes that occur:

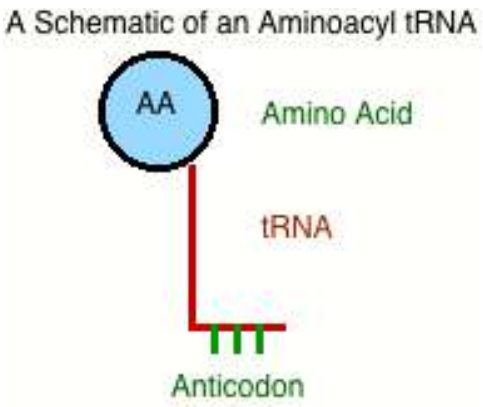
I. **5' Capping** (**attaching a modified version of the nucleotide guanine to the 5' end of the mRNA**.
II. **Poly A tails** (about 200 adenine)
III. **Intron removal and splicing**

Translation:

This is the process whereby the information contained in the mRNA transcript is used to direct the synthesis of proteins. The term "translation" refers to the fact that instructions in the form of a linear series of codons (linear sequence of nucleotides) are used to create a specific linear sequence of amino acids. Thus, information, written in the language of DNA and RNA (a sequence of nucleotides), becomes instead represented as a series of amino acids. The entire process occurs in the cytoplasm (or the matrix of a mitochondrion). The players in the translation process include:

- A fully processed **mRNA** molecule.
- **Ribosomes.**

• **Protein "helpers"** -- recognition, elongation, and termination factors. All of these have to with specific stages in the translation process.
• **aminoacyl-tRNA** ("charged" tRNA molecules).



The Steps of Translation:

Ribosomes consist of a small and large subunit; these subunits are not normally bound together. Before they can bind together, mRNA must bind to the small subunit. Then it joins with a large subunit to form a ribosome.

The ribosomal subunits of prokaryotes and eukaryotes are quite similar. The unit of measurement is the **Svedberg** unit, a measure of the rate of sedimentation in centrifugation rather than size, and this accounts for why fragment names do not add up (70S is made of 50S and 30S).

Prokaryotes have 70S ribosome, each consisting of:

- Small subunit (30S) has:

(**16S RNA subunit + 21 proteins**).

- large subunit(50S) is composed of:

(**5S RNA subunit, 23S RNA subunit and 31 proteins**).

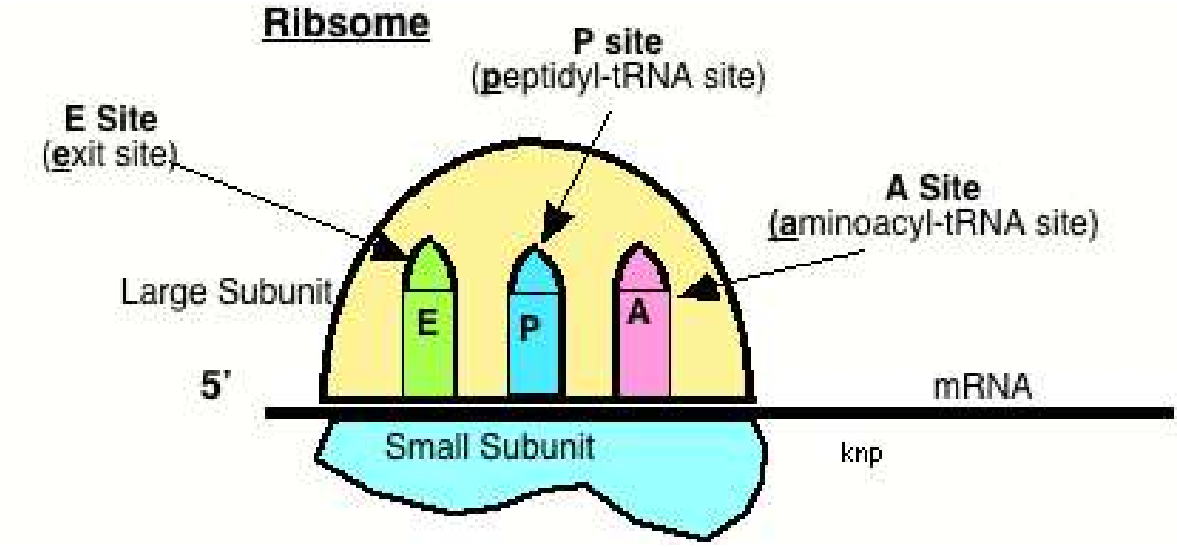
Eukaryotes have 80S ribosome, each consisting of:

- Small (40S) subunit composed of:

(**18S RNA and 33 proteins**).

- large subunit (60S) composed of:

(**5S RNA ,28S RNA, 5.8S subunits and 46 proteins)**.



In particular, note the three tRNA binding areas on the large subunit:

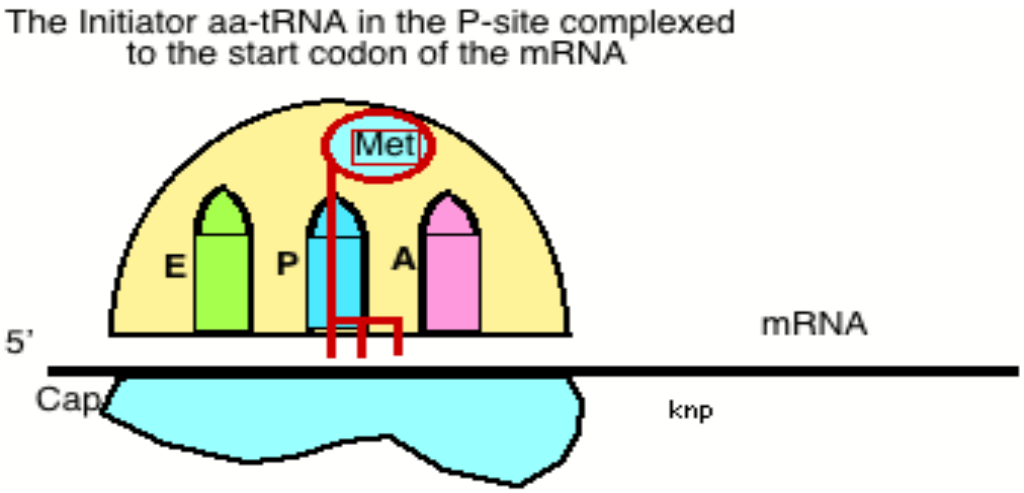
- The **E** or **exit site** is the last place where a tRNA molecule is bound to the complex before it is released so that it can go and pick up another amino acid.

- The **P** or **peptidyl-tRNA** site will normally hold a tRNA that contains the entire growing peptide. We will see that the tRNA molecule in this site changes with the reading of each codon.

- The **A** or **aminoacyl-tRNA** site is where each new incoming AA-tRNA (aminoacyl tRNA) first binds.

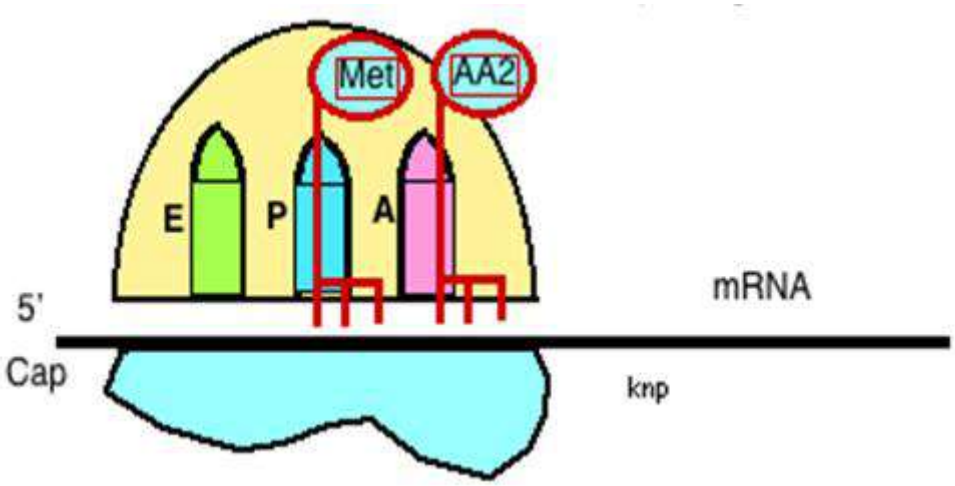
Step 1: Initiation:

The small subunit attaches to the 5' end of the mRNA (by identifying the cap). Using some energy carried in GTP and a series of proteins called **initiation factors** this complex join with the large subunit. The mRNA lines up with the tRNA sites in the large subunit such that a sequence of nucleotides that is located a short distance downstream from the cap lines up under the P site. This particular codon (the one now under the P site) is called the **start codon** (AUG). Thus, it will pair with an aa-tRNA with the anticodon UAC; this particular tRNA **carries the amino acid methionine**:



Step 2: Elongation:

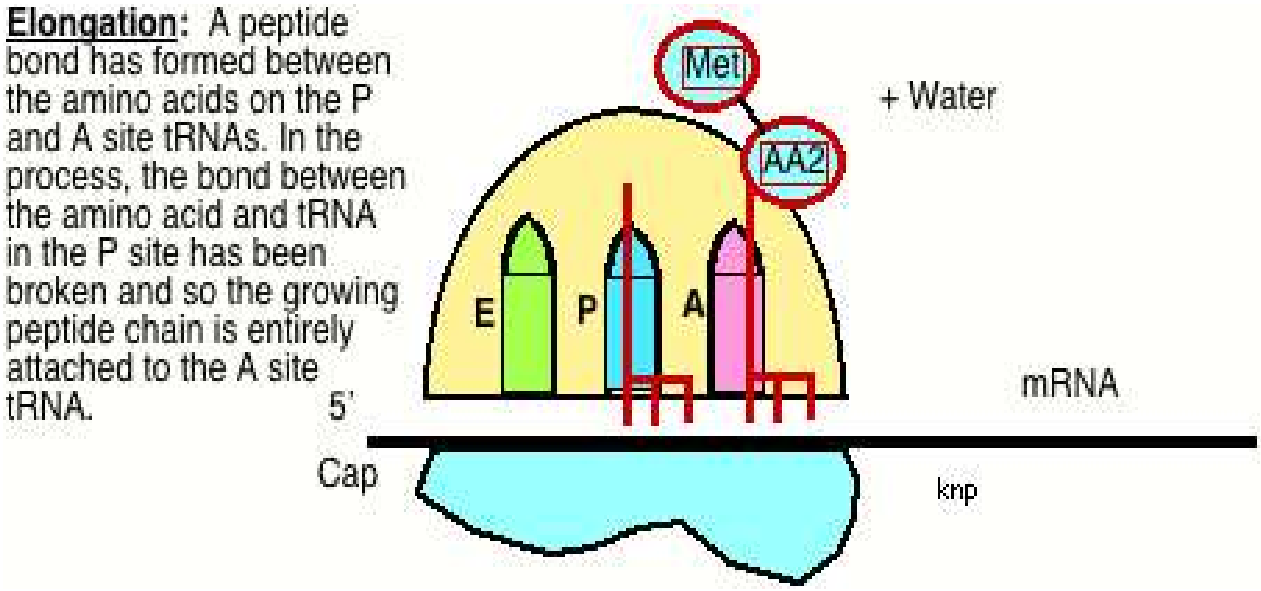
Now we will start building the polypeptide chain. aa-tRNAs are everywhere, having been synthesized in preparation for protein synthesis. At random they pop in and out of the **A site**. When one comes in that has the anticodon for the codon exposed under the A site, it will hydrogen bond. This is called **codon recognition**:



Step 3: Peptide Bond Formation:

the amino acid attached to the tRNA in the P-site is bonded to the amine (free) end of the amino acid in the A site. Since amino acids are bound to the tRNA via their carboxyl end, this means that the first amino acid has now been transferred to the tRNA in the A site:

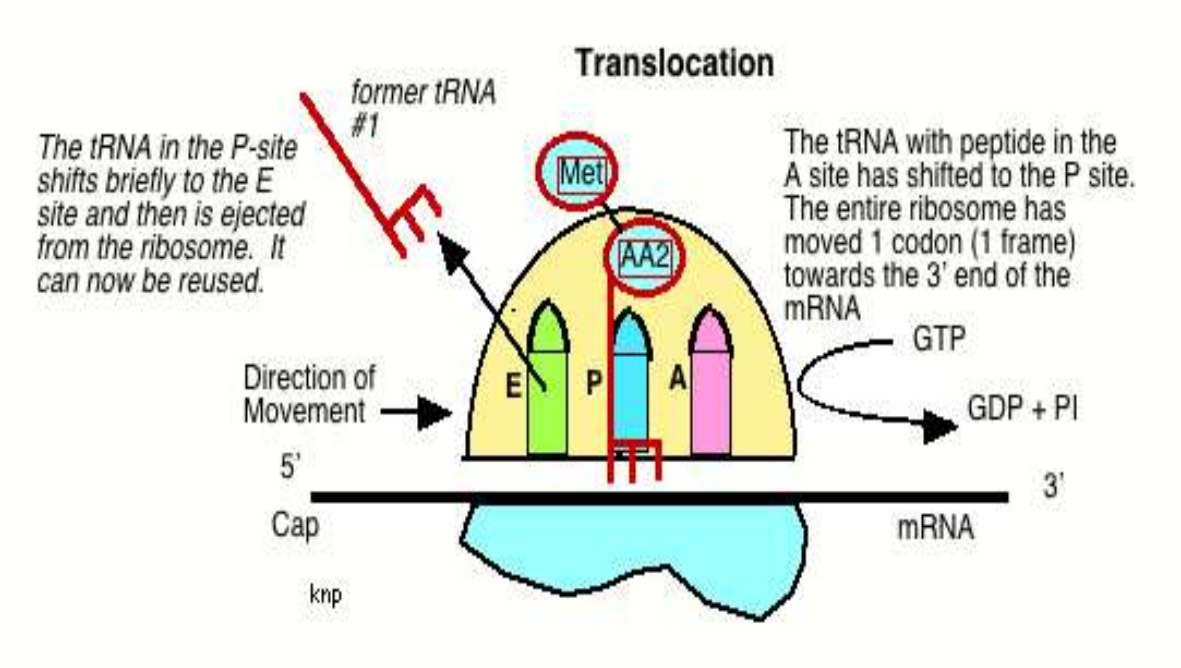
One item of interest here is that the formation of the peptide bond is catalyzed by a **ribozyme** found in the large subunit of the ribosome.



Step 4: Translocation:

In this step, t**he whole ribosome does the equivalent of walking one codon in the 5' to 3' direction**. The result is that:

- The **"amino acid-less" tRNA in the P site ends up in the E site** as a result of the entire ribosome moving one codon towards the 3' end of the mRNA. **This tRNA is ejected upon arrival in the E site;** it leaves the ribosome and is re-charged and reused.
- The **tRNA holding the growing chain ends up in the P site (from A)**, again as a result of the movement of the ribosome in the 3' direction along the mRNA. It remains bound to its codon on the mRNA.
- This process requires energy in the form of GTP.
- Notice again that **the translocation moves the reading exactly 1 codon or one reading frame.**



Step 5: Do Steps 1 to 4 Over and Over.

The process (codon recognition, elongation, and translocation) will continue over and over as the ribosome moves down the mRNA one codon at a time. Notice that this mechanism reads the information on the mRNA in a non-overlapping matter -- the reading codon (the exposed 3 nucleotide sequence) always moves exactly 3 nucleotides in each cycle. Thus, as we have mentioned before, codons are distinct.

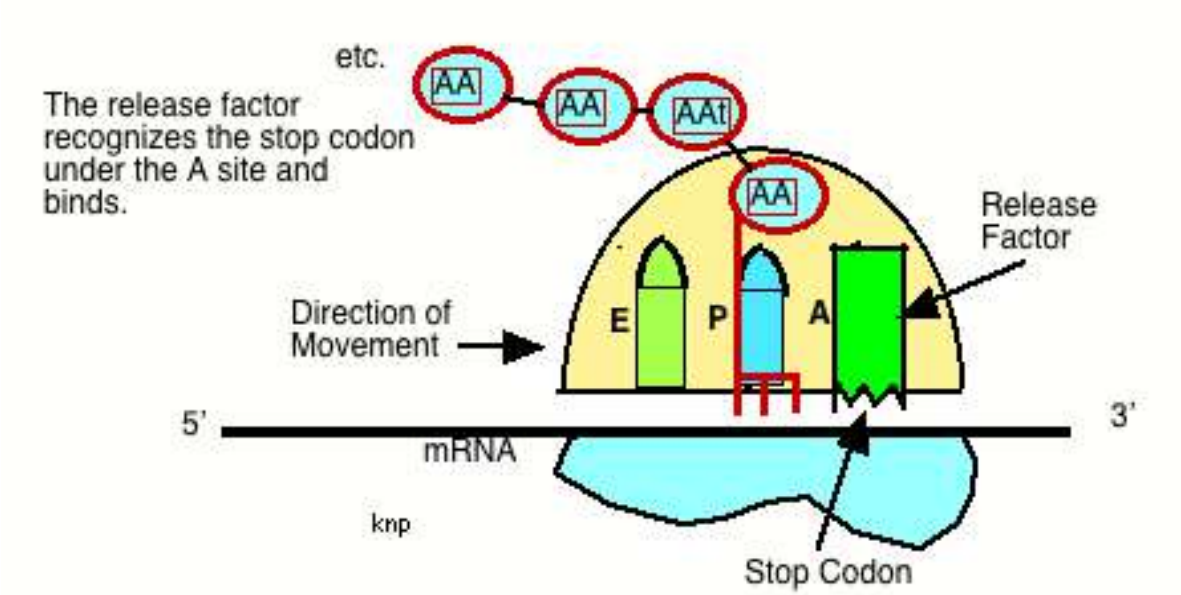
Note also the use of more ~P, this time for the translocation step. As usual, in animals this comes from oxidative phosphorylation.

Step 6: Termination:

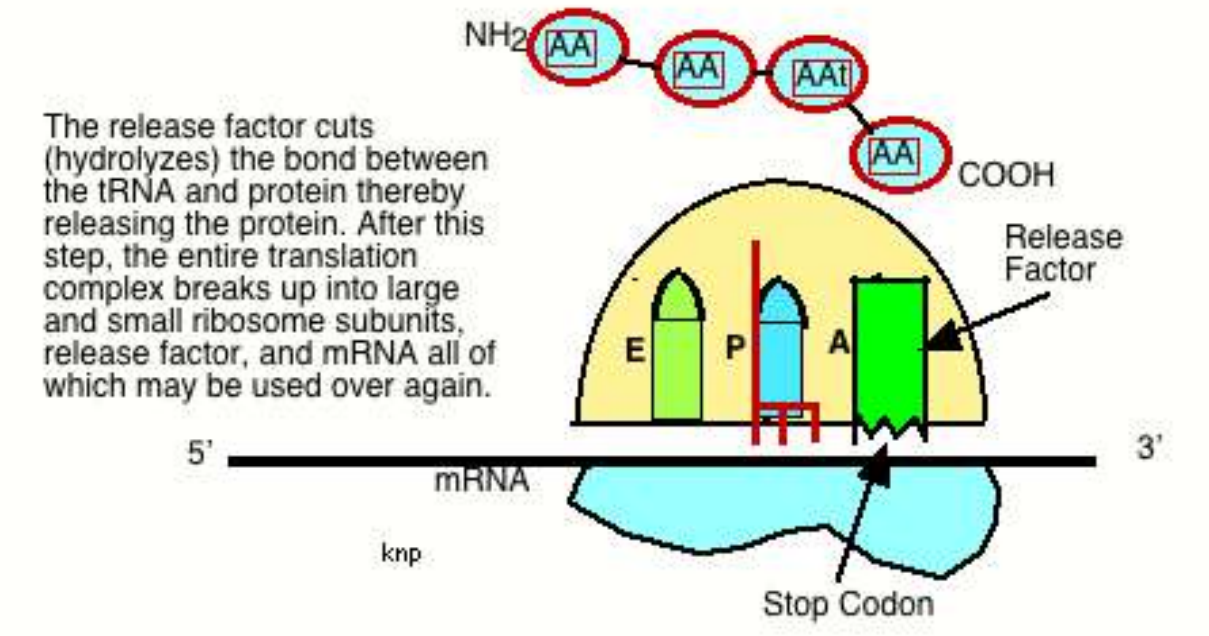
Eventually we get to near the end of the mRNA. Just before the end of the mRNA is a section called the trailer and at its start is what is called a **stop codon** (**UAA, UAG, and UGA**). They do not code for amino acids. When they are encountered the polypeptide chain cannot elongate further. What actually, happens is that:

- When a stop codon appears under the **A** site after the translocation step, **there is no tRNA that can attach**.
- However, a protein called a **release factor** binds to the mRNA at the stop codon.
- This factor will catalyze the hydrolysis of the bond between the

Polypeptide and the tRNA in the **P** site. As a result, the polypeptide is released.



- Once this happens, the entire ribosome breaks down into its small and large subunits, the mRNA is released so that it can potentially be translated again, and the release factor is also set free to be used by another ribosome (as usual, remember that there will be many release factor proteins milling about the cell:



Here is a more detailed view of the hydrolysis and its consequences;

Lec 10

Post-translational modification: the role of the "environment"

A gene directly specifies only the primary structure of a protein. The exact types of higher order structure that will develop depend very much on the sort of environment a protein finds itself in and the types of interactions a protein has with other compounds, especially other proteins.

In fact all of these things could be lumped together as the "environment". In this case, the environment is the chemical and physical conditions that a protein encounters in a cell over its lifetime (yes proteins eventually "die" – they are broken back down into amino acids). Let's look at these factors.

In all cells, there is an expected normal environment with respect to **pH, temperature, the concentrations of various ions,** and **osmolarity.** All of these factors clearly can affect the higher order structure of proteins. This expected environment is achieved partly as a result of regulation of these variables and partly as a matter of the conditions normally found in the external environment of the organism.

For instance, in animals that cannot regulate their body temperature, their cells will be at whatever temperature is found in the general environment. The animal will only be able to exist in environments where the range of temperatures is such that proteins fold and perform correctly.

Proteins assume their higher order structures in part during the translation process. As the protein is created and lengthens, various side groups begin to interact and cause various types of bending and folding. Amino acids with nonpolar side groups are twisted into the center of the protein away from water and from polar side groups. Obviously, anything that affects the polar properties of the cytoplasm will affect the folding process -- either more or less of any polar solute such as H+ or any other ions can obviously result in a protein with a different shape.

Beyond this there are often specific factors that work only on certain types or parts of proteins and that can have very significant effects.
Here are some examples of what are generally called **post-translational modifications**.

- **Disulfide bridges**:

One type of amino acid, cysteine contains sulfur in its side chain. This side chain can react with other cysteines -- the sulfur atoms in each side chain will bind to each other (C -- S -- S -- C) in what is called a disulfide bridge. Wherever these forms, a relatively strong connection is made between different areas of a protein. These covalent bonds tend to hold a molecule in particular shape with much more force than H bonds. It takes specific enzymes in most cases to catalyze the formation of disulfide bridges.

- **Change or modification of the primary sequence**:

An enzyme removes particular amino acids from end or the other or chemically modifies the side groups of particular types of amino acids. It should be obvious that such alterations may cause tremendous changes in the higher order structure.

In other cases, enzymes cut the "backbone" of the protein in a particular place and this induces important structural changes.
An important example is a functional condition for enzymes, where they are inactive and waiting to be activated. Such enzymes are usually called **Zymogens.**

Zymogens are enzymes that are produced so that they fold up into an inactive configuration. However, when needed, some factor acts on them to cause them to rapidly change shape into an active form.
Good examples are a number of **digestive enzymes** and **blood clotting proteins**. Obviously, you don't want digestive enzymes to be produced in their active state -- if you did, they might well digest the contents of the cell that makes them. Instead, you want to be able to turn them on only under the appropriate conditions. Likewise, with blood clotting, you want the proteins needed for clotting to be present in sufficient numbers but

you do not want them to do anything until you are bleeding. So specific substances are released when bleeding begins those interact with these clotting factors and cause them to become active.

DNA - MUTATIONS

Random changes to the genetic code. Copying errors when DNA replicates or is transcribed into RNA can cause changes in the sequence of bases which makes up the genetic code. Radiation and some chemicals can also cause changes.

- **Changes to individual bases**

There are several amino acids which are coded for by more than one base combination. For example, glycine (Gly) is coded for by GGT, GGC, GGA and GGG. It doesn't matter what the last base is - you would get glycine whatever base followed the initial GG.

I. That means that a mutation at the end of a codon like this wouldn't make any difference to the protein chain which would eventually form. These are known as **silent mutations**.
II. Alternatively, of course, you could well get a code for a different amino acid or even a stop codon. If a stop codon was produced in the middle of the gene, then the protein formed would be too short, and almost certainly wouldn't function properly.
III. If a different amino acid was produced, how much it mattered would depend on whereabouts it was in the protein chain.

    a. If it was near the active site of an enzyme, it might stop the enzyme from working entirely.
    b. If it was on the outside of an enzyme, and didn't affect the way the protein chain folded, it might not matter at all.

- **Inserting or deleting bases**

The situation is more dramatic if extra bases are inserted into the code, or some bases are deleted from the code. Then large chunks of the protein will consist of completely wrong amino acid residues.

We've looked so far at inserting or deleting *one* base. What if you do it for more than one?

The effect is the same unless you add or delete multiples of three bases - without changing any other codons. If you added an extra three bases between two existing codons, it represents an extra codon in the DNA, and so an extra amino acid residue in the protein chain. Does this matter?

It depends where it is in the chain (Is it important to the active site of an enzyme, for example?), and whether it affects the folding of the chain.

What if the three bases were inserted so that they broke up an existing codon?

The effect is again fairly limited. It will change one codon completely, and introduce an extra codon. That would give you one different amino acid and one extra amino acid in the chain. Again, how much that would affect the final protein depends on where it happens in the chain.

Deleting a whole codon again leaves most of the protein chain unchanged. Again, whether the function of the protein is affected depends on where the missing amino acid should have been and how critical it was to the way the protein folded.

• Some diseases caused by mutation

The following examples illustrate some of the changes we've looked at above and how they can result in disease.

♣ Cystic fibrosis

Cystic fibrosis is an inherited disease which affects the lungs and digestive system. It results from mutation in a gene responsible for making a protein which is involved in the transport of ions across cell boundaries.

The effect is to produce sticky mucus which clogs the lungs and can lead to serious infection. Similar sticky mucus also blocks the pancreas which provides enzymes for breaking down food. This gets in the way of the processes which convert the food into molecules which can be absorbed by the body. There are lots of different mutations which can cause this, but we'll just have a quick look at the one which accounts for about 70% of cystic fibrosis cases. The base sequence in the part of the gene affected ought to look like this:

...ATCATCTTTGGTGTTTCC...

Ile Ile Phe Gly Val Ser

The phenylalanine (Phe) is the amino acid which is missing from the final protein in many sufferers from cystic fibrosis. However, it isn't quite as simple as just losing the TTT codon.

Instead, the three bases lost are: That leaves the sequence:

The amino acid sequence is identical to before but without the phenylalanine. How did that happen when we didn't actually remove the whole of the phenylalanine codon?

If you look carefully, you will see that the codon for the second isoleucine (Ile) is different from before. It so happens that isoleucine is coded for by both ATC and ATT. Once the second T has joined the existing AT, all the rest of the base sequence is exactly what it was before.

♣ Sickle cell anemia

This is so called because red blood cells change their shape from the normal flexible doughnut shape to a much more rigid sickle shape - rather like a crescent moon.

It results from the change of a single base in a gene responsible for making one of the protein chains which makes up hemoglobin. The affected part of the gene should read:

What it actually reads in someone suffering from sickle cell anemia is:

The effect of this single change is to make the hemoglobin temporarily polymerise to make fibers after it has released the oxygen that it carries around the body. This changes the shape of the red blood cells so that they don't flow so easily - it makes them sticky, especially in small blood vessels. This can cause pain and lead to organ damage.

♣ Hemophilia

Sufferers from hemophilia lack a protein in the blood which allows it to clot. That means that if someone with hemophilia cuts themselves, the wound will just continue to bleed.

There are all sorts of mutations which cause hemophilia. One which is easy to understand is caused by changing a single base at the beginning of a codon for **argenine (CGA)** somewhere in the gene to give **UGA** which is a **stop codon.** All that will be produced is a useless fragment of the intended protein.

- ***Universal Fast Walking*** **(UFW):** for genome walking and genetic fingerprinting using a more specific 'two-sided' PCR than conventional 'one-sided' approaches (using only one gene-specific primer and one general primer - which can lead to artefactual 'noise') by virtue of a mechanism involving lariat structure formation. Streamlined derivatives of UFW are LaNe RAGE (lariat-dependent nested PCR for rapid amplification of genomic DNA ends), 5'RACE LaNe and 3'RACE LaNe.

- ***PAN-AC*****:** uses isothermal conditions for amplification, and may be used in living cells.