

**Department of BSBE
Indian Institute Of Technology Guwahati**



Genes

Dr. Sanjukta Patra

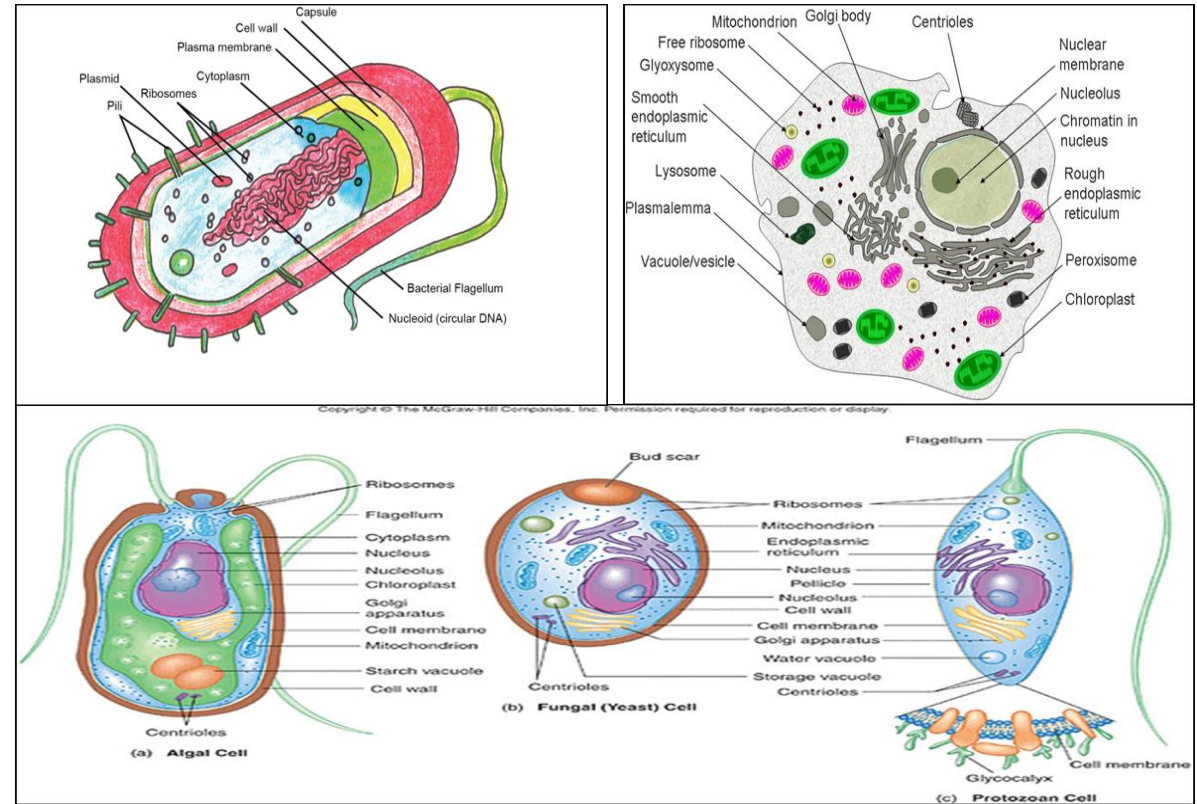
BT 207

Jan - May 2023

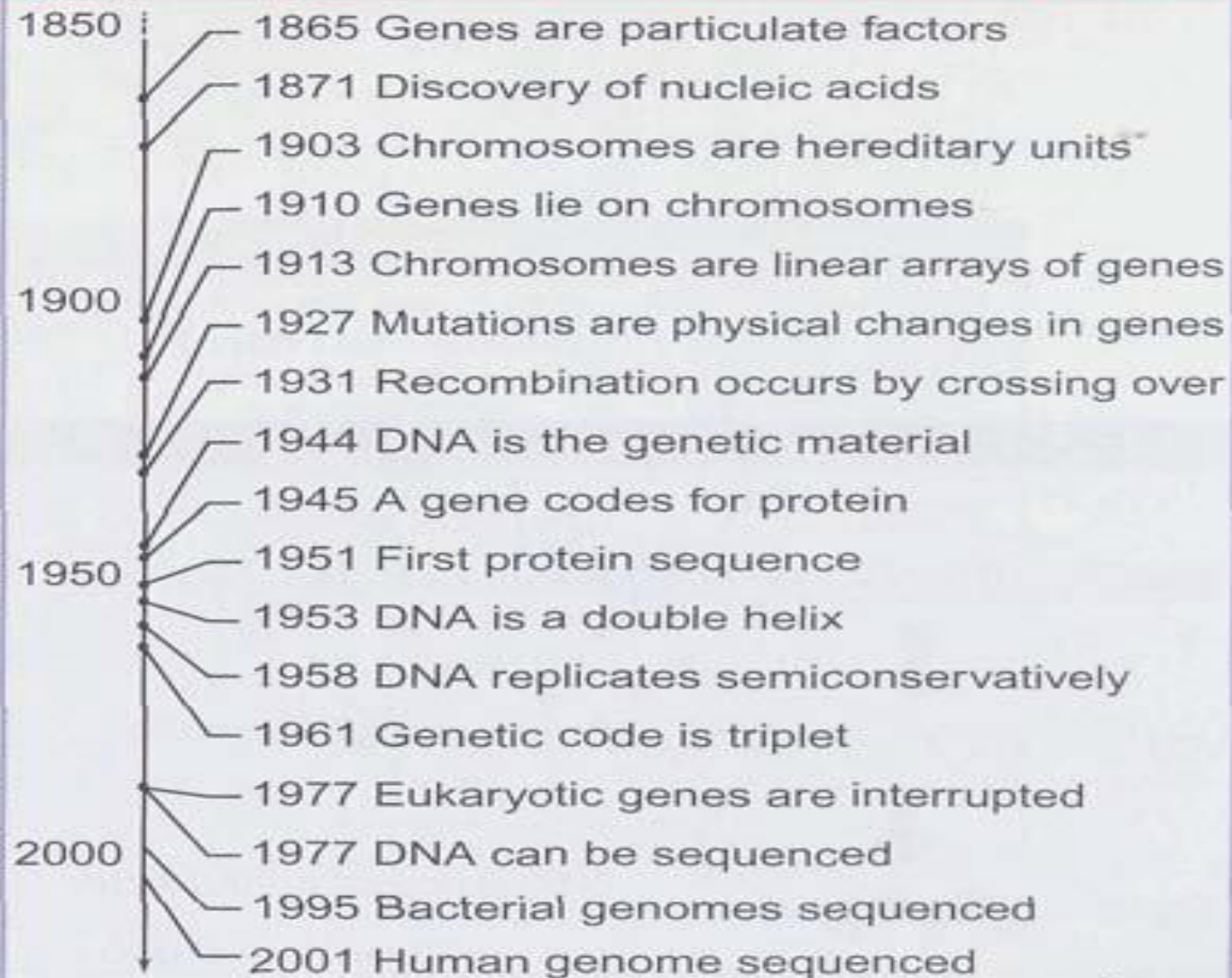
“A organism is what its genome is”

“Phenotype is an expression of the genotype”

- The genome contains the complete set of hereditary information for any organism.
- Physically the genome may be divided into a number of different nucleic acid molecules.
- Functionally it may be divided into genes.
- Each gene is a sequence within the nucleic acid that represents a single protein.
- Genomes for living organisms may contain as few as <500 genes (for a mycoplasma, a type of bacterium) to as many as >40,000 for Man.



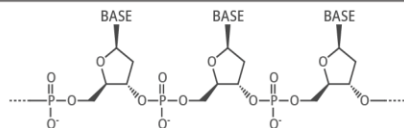
Major events in the genetics century



THE CHEMICAL STRUCTURE OF DNA

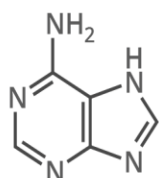
DNA (deoxyribonucleic acid) carries genetic information in all multicellular forms of life. It carries instructions for the creation of proteins, which carry out a wide range of roles in the body.

THE SUGAR PHOSPHATE 'BACKBONE'

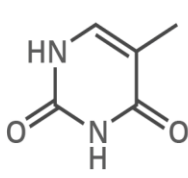


DNA is a polymer made up of units called nucleotides. The nucleotides are made of three different components: a sugar group, a phosphate group, and a base. There are four different bases: adenine, thymine, guanine & cytosine.

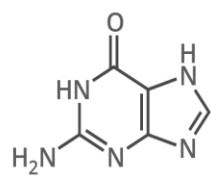
A ADENINE



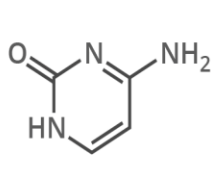
T THYMINE



G GUANINE

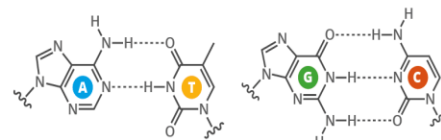


C CYTOSINE



WHAT HOLDS DNA STRANDS TOGETHER?

DNA strands are held together by hydrogen bonds between bases on adjacent strands. Adenine (A) always pairs with thymine (T), whilst guanine (G) always pairs with cytosine (C).



FROM DNA TO PROTEINS

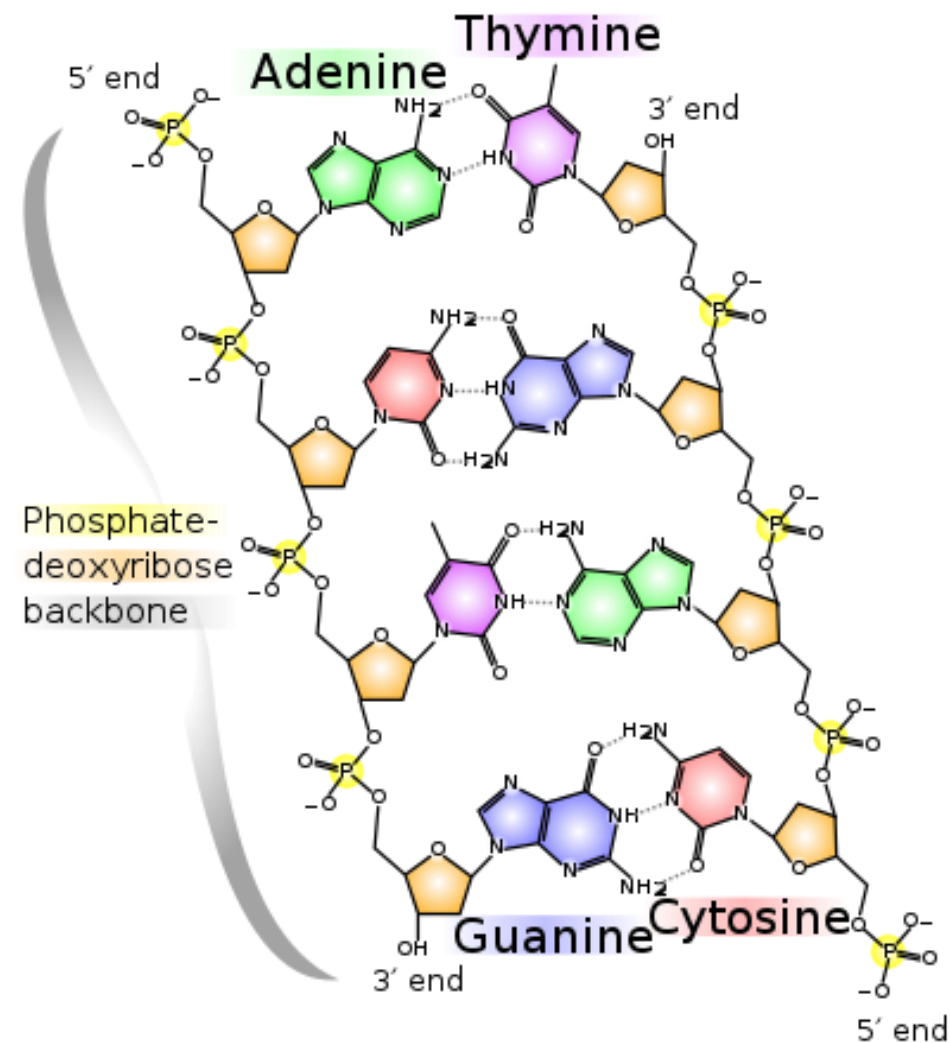


The bases along a single strand of DNA act as a code. The letters form three letter 'words', or codons, which code for different amino acids - the building blocks of proteins.

An enzyme, RNA polymerase, transcribes DNA into mRNA (messenger ribonucleic acid). It does this by splitting apart the two strands that form the double helix, then reading a strand and copying the sequence of nucleotides. The only difference between the RNA and the original DNA is that in the place of thymine (T), another base with a similar structure is used: uracil (U).

DNA SEQUENCE	T	T	C	T	G	A	A	C	C	G	T	T	A	
mRNA SEQUENCE	U	U	C	C	U	G	A	A	C	C	G	U	U	A
AMINO ACID	Phenylalanine		Leucine		Asparagine		Proline		Leucine					

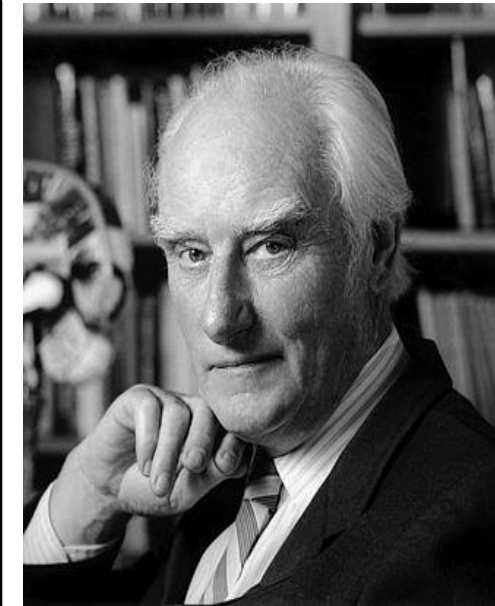
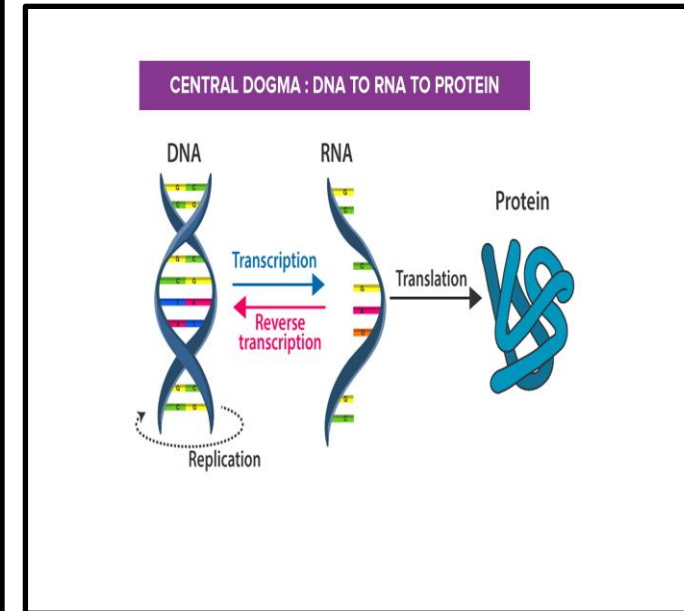
In multicellular organisms, the mRNA carries genetic code out of the nucleus, to the cell's cytoplasm. Here, protein synthesis takes place. 'Translation' is the process of converting turning the mRNA's 'code' into proteins. Molecules called ribosomes carry out this process, building up proteins from the amino acids coded for.



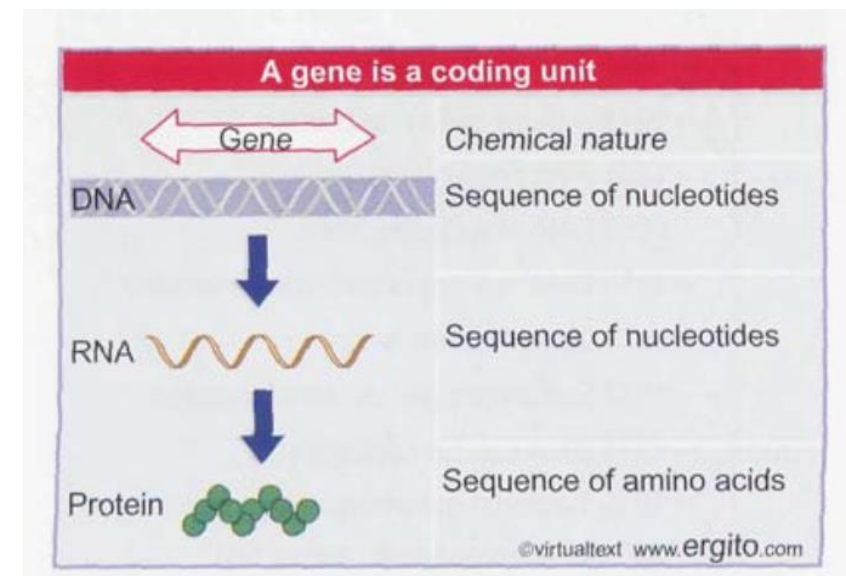
© COMPOUND INTEREST 2015 - WWW.COMPOUNDCHEM.COM | Twitter: @compoundchem | Facebook: www.facebook.com/compoundchem
This graphic is shared under a Creative Commons Attribution-NonCommercial-NoDerivatives licence.



- A gene is a sequence of DNA that produces another nucleic acid, RNA.
- The DNA has two strands of nucleic acid, and the RNA has only one strand. The sequence of the RNA is determined by the sequence of the DNA (in fact, it is identical to one of the DNA strands).
- The RNA is in turn used to direct production of a protein.
- *Thus a gene is a sequence of DNA that codes for an RNA; in protein-coding genes, the RNA in turn codes for a protein.*
- The basic behavior of the **gene was defined by Mendel** in his two laws, the gene was recognized as a "**particulate factor**" that passes unchanged from parent to progeny.
- A gene may exist in alternative forms **called alleles**.
- In diploid organisms, which have two sets of chromosomes, one copy of each chromosome is inherited from each parent.
- **Each chromosome consists of a linear array of genes with each gene residing at a particular location on the chromosome called a genetic locus.**
- The key to understanding the organization of genes into chromosomes was the discovery of **genetic linkage**.



Francis Crick



Genes and Enzymes

Early genetic studies - identification and chromosomal localization of genes that control readily observable characteristics.

The first insight into the **relationship between genes and enzymes came in 1909**, when it was realized that the inherited human disease **phenylketonuria** results from a genetic defect in metabolism of the amino acid phenylalanine.

This defect was hypothesized to result from a deficiency in the enzyme needed to catalyze the relevant metabolic reaction, leading to the general suggestion that genes specify the synthesis of enzymes.

Clearer evidence linking genes with the synthesis of enzymes came from experiments of **George Beadle and Edward Tatum, performed in 1941 with the fungus Neurospora crassa**.

In the laboratory, Neurospora can be grown on minimal or rich media similar consisting of only of salts, glucose, and biotin; rich media are supplemented with amino acids, vitamins, purines and pyrimidines.

Beadle and Tatum isolated mutants of Neurospora that grew normally on rich media but could not grow on minimal media. Each mutant was found to require a specific nutritional supplement, such as a particular amino acid, for growth.

Furthermore, the requirement for a specific nutritional supplement correlated with the failure of the mutant to synthesize that particular compound. Thus each mutation resulted in a deficiency in a specific metabolic pathway. **Since such metabolic pathways were known to be governed by enzymes, the conclusion from these experiments was that each gene specified the structure E' of a single enzyme-the one gene-one enzyme hypothesis.**

Many enzymes are now known to consist of multiple polypeptides, so the currently accepted statement of this hypothesis is that **each gene specifies the structure of a single polypeptide chain.**

DNA is the genetic material of bacteria

What causes transformation of bacteria?

Bacterial transformation provided the first proof that DNA is the genetic material.

Griffith - 1928

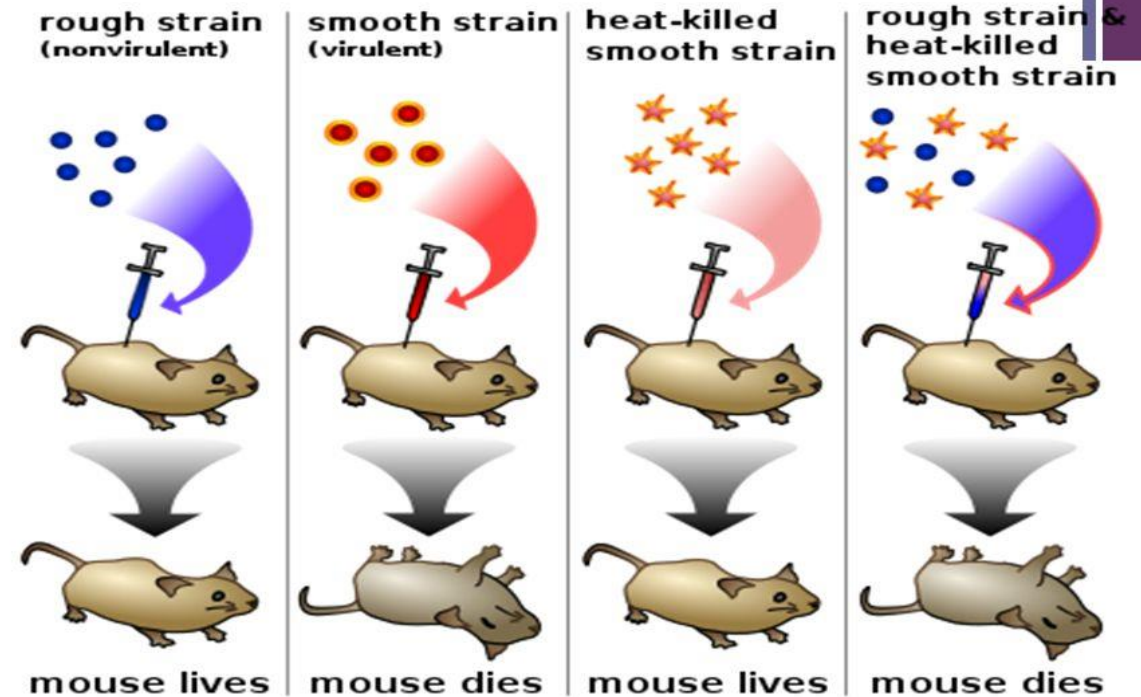
Genetic properties can be transferred from one bacterial strain to another by extracting DNA from the first strain and adding it to the second strain.

The bacterium *Pneumococcus* kills mice by causing pneumonia.

The virulence of the bacterium is determined by its capsular polysaccharide.

This is a component of the surface that allows the bacterium to escape destruction by the host. Several types (I, II, III) of *Pneumococcus* have different capsular polysaccharides. They have a smooth (S) appearance

Griffith's Experiment



Neither heat-killed S-type nor live R-type bacteria can kill mice, but simultaneous infection of them both can kill mice just as effectively as the live S-type.

DNA is the genetic material of bacteria

In 1928 it was observed (Griffith et al) that mice inoculated with nonencapsulated (R) bacteria plus heat-killed encapsulated (S) bacteria developed pneumonia and died. Importantly, the bacteria that were then isolated from these mice were of the S type.

Subsequent experiments showed that a cell-free extract of S bacteria was similarly capable of converting (or transforming) R bacteria to the S state.

Thus a substance in the S extract (called the transforming principle) was responsible for inducing the genetic transformation of R to S bacteria.

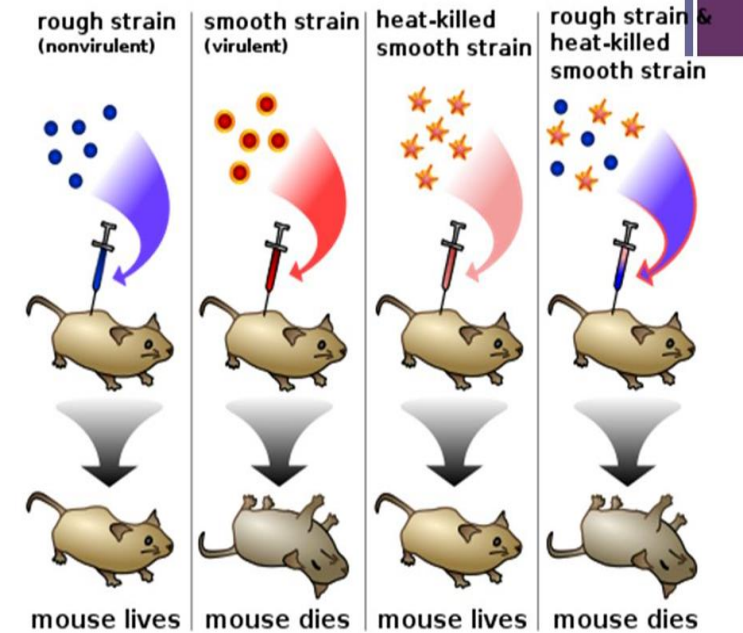
Identification of DNA as the Genetic Material

Oswald Avery, Colin Macleod, and Maclyn McCarty established in 1944 that the transforming principle (genetic material) is DNA were derived from studies of the bacterium that causes pneumonia (*Pneumococcus*).

Virulent strains of *Pneumococcus* are surrounded by a polysaccharide capsule that protects the bacteria from attack by the immune system of the host giving bacterial colonies a smooth appearance in culture, encapsulated strains are denoted S.

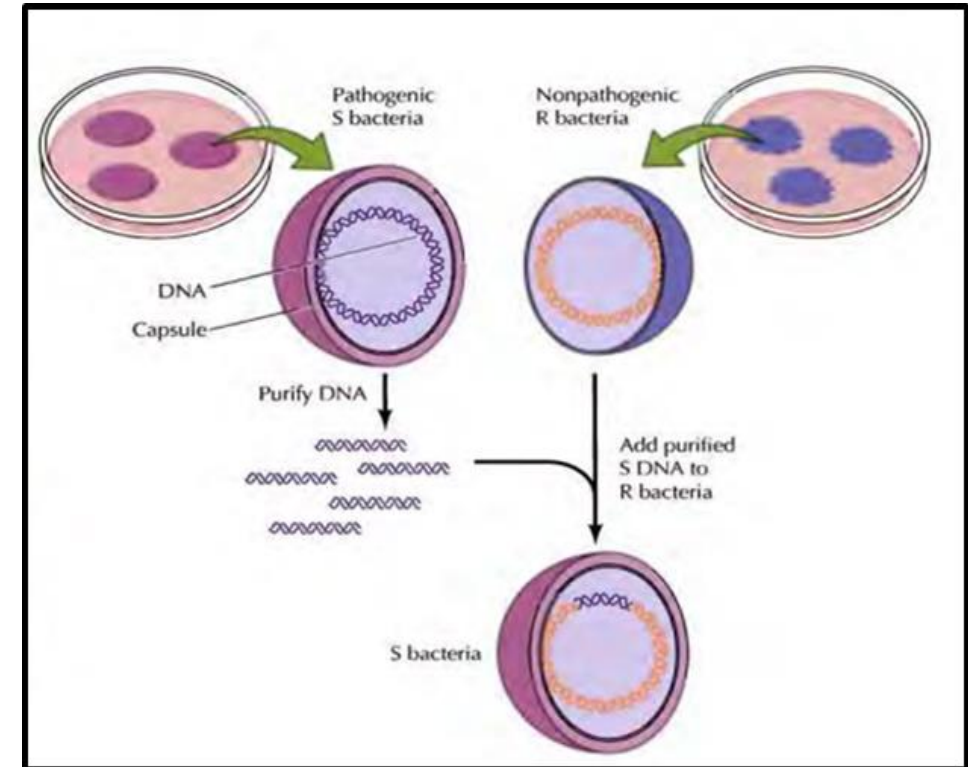
Mutant strains that have lost the ability to make a capsule (denoted R) form rough-edged colonies in culture and are no longer lethal when inoculated into mice.

Griffith's Experiment



Is DNA the genetic material of bacteria?

- In 1944 Oswald Avery, Colin MacLeod, and Maclyn McCarty established that the transforming principle was DNA,
- (i) By purifying it from bacterial extracts and by demonstrating that the activity of the **transforming principle is abolished by enzymatic digestion of DNA but not by digestion of proteins.**
- Although these studies did not immediately lead to the acceptance of DNA as the genetic material, they were extended within a few years **by experiments with bacterial viruses.**
- (ii) In particular, when a **bacterial virus infects a cell, the viral DNA rather than the viral protein must enter the cell in order for the virus to replicate.**
- Moreover, the parental viral DNA (but not the protein) is transmitted to progeny virus particles.
- The concurrence of these results with continuing studies of the activity of DNA in bacterial transformation led to acceptance of the idea that DNA is the genetic material.



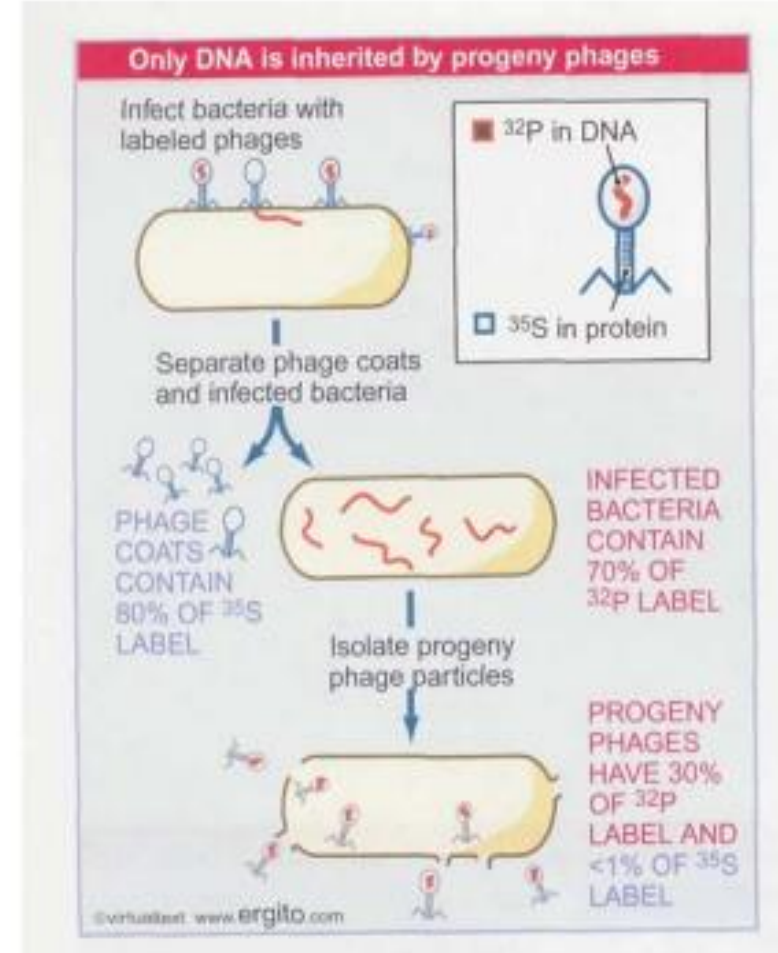
Transfer of genetic information by DNA DNA is extracted from a pathogenic strain of *Pneumococcus*, which is surrounded by a capsule and forms smooth colonies (S). Addition of the purified S DNA to a culture of nonpathogenic, nonencapsulated bacteria (R for "rough" colonies) results in the formation of S colonies. The purified DNA therefore contains the genetic information responsible for transformation of R to S bacteria.

DNA is the genetic material of viruses

Hershey and Chase in 1952

Phage infection proved that DNA is the genetic material of viruses.

When the DNA and protein components of bacteriophages are labeled with different radioactive isotopes, only the DNA is transmitted to the progeny phages produced by infecting bacteria.



The genetic material of Phage T2 is DNA

- Bacteria were infected with T2 phages that had been radioactively labelled either in their DNA component (with ^{32}P) or in their protein component (with ^{35}S).
- The infected bacteria were agitated in a blender, and two fractions were separated by centrifugation.
- One contained the empty phage coats that were released from the surface of the bacteria. The other fraction consisted of the infected bacteria themselves.
- Most of the ^{32}P label was present in the infected bacteria. The progeny phage particles produced by the infection contained ~30 % of the original ^{32}P label.
- The progeny received very little less than 1%—of the protein contained in the original phage population. The phage coats consist of protein and therefore carried the ^{35}S radioactive label. This experiment therefore showed directly that only the DNA of the parent phages enters the bacteria and then becomes part of the progeny phages, exactly the pattern of inheritance expected of genetic material.
- A phage (virus) reproduces by commandeering the machinery of an infected host cell to manufacture more copies of itself.
- Conclusion that the genetic material is DNA, of a cell or virus.

DNA is the genetic material of animal cells

When DNA is added to population of single eukaryotic cells growing in culture, the nucleic acid enters the cells, and in some of them results in the production of new proteins .

When a purified DNA is used, its incorporation leads to the production of a particular protein .

Figure depicts one of the standard systems.

The DNA that is introduced into the recipient cell becomes part of its genetic material, and is inherited in the same way as any other part.

Its expression confers a new trait upon the cells (synthesis of thymidine kinase in the example of the figure).

At first, these experiments were successful only with individual cells adapted to grow in a culture medium.

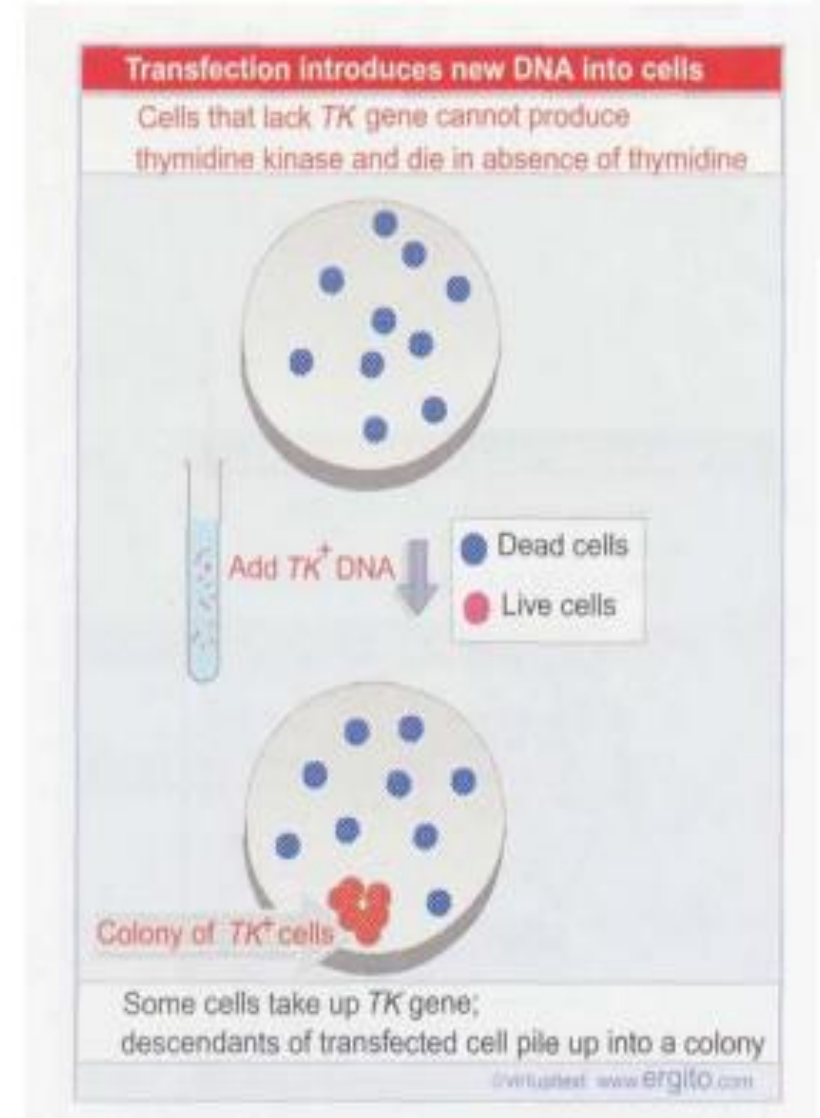
Since then, however, DNA has been introduced into mouse eggs by microinjection; and it may become a stable part of the genetic material of the mouse.

Such experiments show directly not only that **DNA is the genetic material in eukaryotes, but also that it can be transferred between different species and yet remain functional.**

The genetic material of all known organisms and many viruses is DNA .

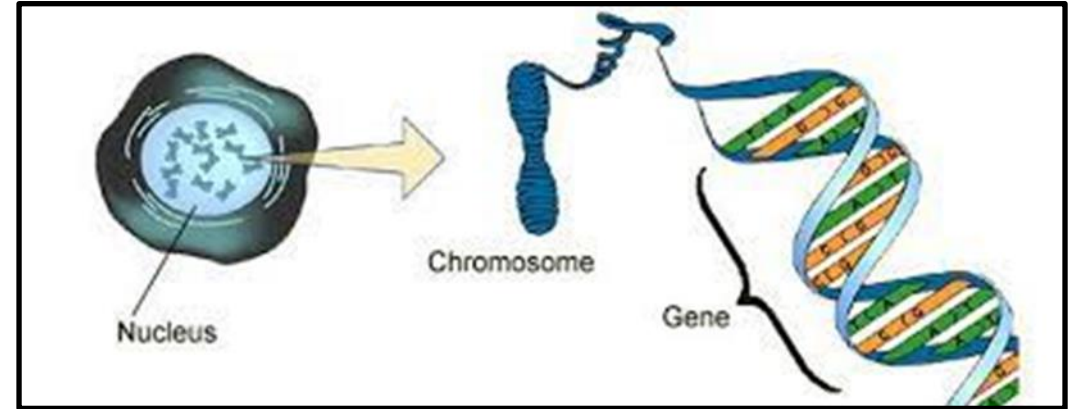
However, some viruses use an alternative type of nucleic acid , ribonucleic acid (RNA), as the genetic material.

The general principle of the nature of the genetic material, is that it is always nucleic acid; in fact, it is DNA except in the RNA viruses.



Concept of gene

- The classical principles of genetics were deduced by Gregor Mendel in 1865 based on breeding experiments with peas.
- He assumed that each trait is determined by a pair of inherited 'factors' which are now called gene.
- In 1909 Wilhelm Johannsen coined the term 'GENE'




Features of a Gene

Several genes are located on each chromosome.

- The genes are arranged in a single linear order like beads on a string.
- Each gene occupies specific position called locus.
- If the position of gene changes, character changes.
- Genes can be transmitted from parent to off springs.
- Genes may exist in several alternate form called alleles.
- Genes can combine together or can be replicated during a cell division.
- Genes may undergo for sudden changes in position and composition called mutation.
- Genes are capable of self-duplication producing their own exact copies

Why do we need to know the gene?


A gene is a specific sequence of DNA containing genetic information required to make a specific protein.



Types of Gene based on organism

Prokaryotic gene (which is seen in prokaryotes, example : Bacteria, Cyanobacteria)

Eukaryotic gene (Which is seen in higher organisms such as Plants, Animals)



Types of gene based on activity

House-keeping genes (genes which are always active)

Specific genes. (Those genes which are getting active only during some special condition)

Types of gene based on behavior

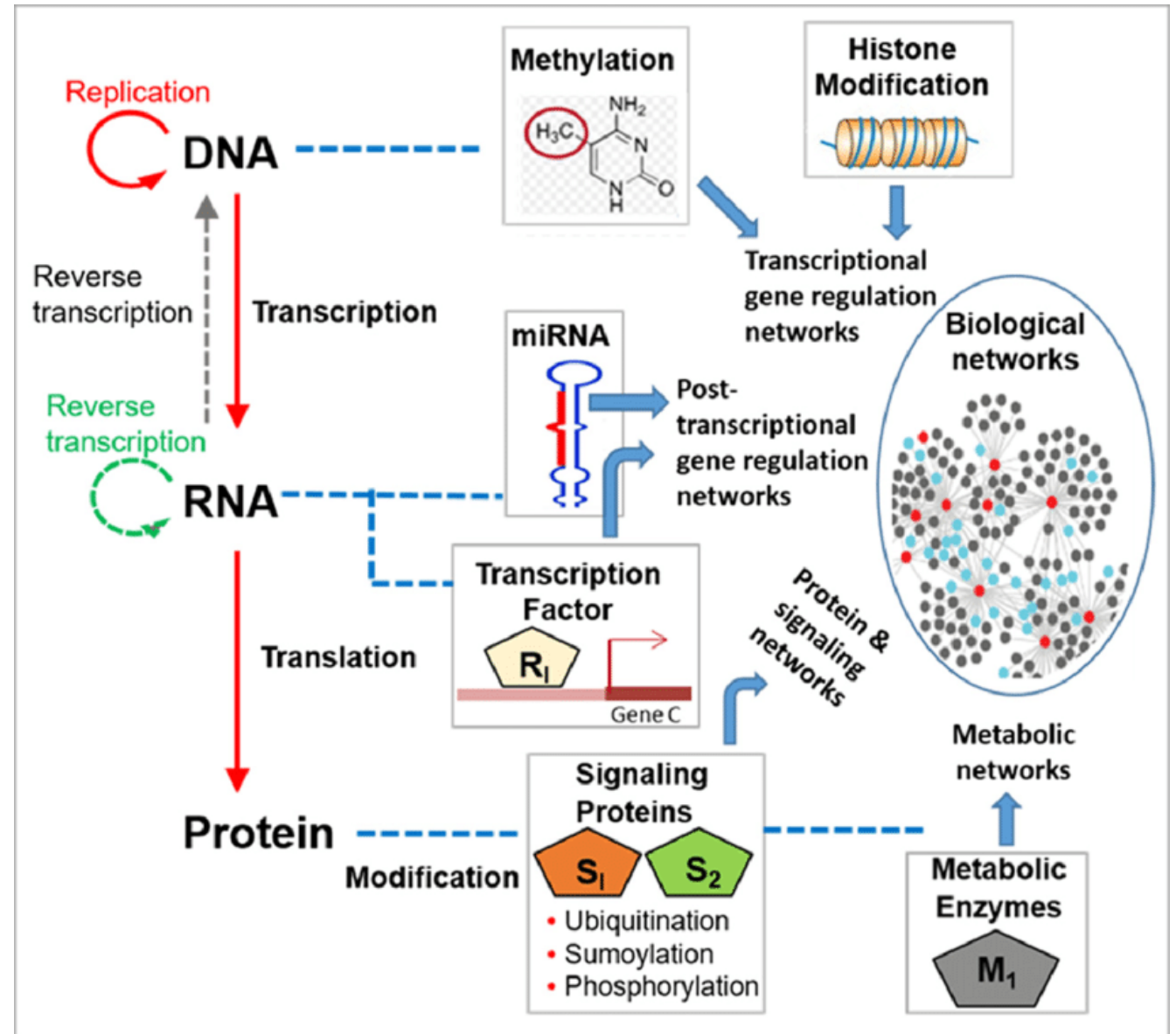
- Basic genes: These are the fundamental genes that bring about expression of particular character.
 - Lethal genes: These bring about the death their possessor.
 - Multiple gene: When two or more pairs of independent genes act together to produce a single phenotypic trait.
 - Cumulative gene: Some genes have additive effects on the action of other genes. These are called cumulative genes.
 - Pleiotropic genes: The genes which produce changes in more than one character is called pleiotropic gene.
 - Modifying gene: The gene which cannot produce a character by itself but interacts with other to produce a modified effect is called modifier gene.
 - Inhibitory gene: The gene which suppresses or inhibits the expression of another gene is called inhibitory gene
- Introduction and History of gene.

Codons

- Discovered by **Sydney Brenner and Francis Crick** in 1961.
- In every triplet of nucleotides, each codon codes for one amino acid in a protein.

Central dogma

- Proposed in 1958 by Francis Crick.
- He postulated that all possible information transferred, are not viable.



Uninterrupted DNA – For cloning

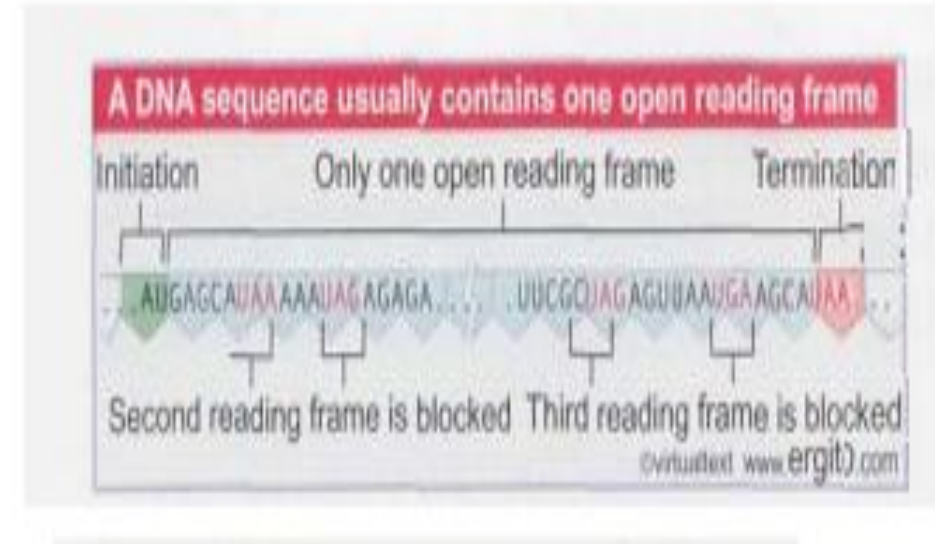
Why do we need to know the genetic code?

The genetic code is triplet

1. The relationship between a sequence of DNA and the sequence of the corresponding protein is called the **genetic code**.
2. The genetic code is read in triplet nucleotides called **codons** - which are **nonoverlapping** and are **read from a fixed starting point**.
3. A gene includes a series of codons that is **read sequentially** from a starting point at one end to a **termination point** at the other end.
4. Written in the conventional **5'—>3' direction**, the nucleotide sequence of the DNA strand that codes for protein corresponds to the amino acid sequence of the protein written in the direction from **N-terminus to C-terminus**.
5. The genetic code is read in nonoverlapping triplets from a fixed starting point: ' **Nonoverlapping implies that each codon consists of three nucleotides and that successive codons are represented by successive trinucleotides**.
6. The use of a **fixed starting point means that assembly of a protein must start at one end and work to the other**, so that **different parts of the coding sequence cannot be read independently**.

Every sequence has three possible reading frames

- A **reading frame** is a way of dividing the sequence of nucleotides in a nucleic acid (DNA or RNA) molecule into a set of consecutive, non-overlapping triplets.
- Where these **triplets equate to amino acids or stop signals during translation.**
- A single strand of a nucleic acid molecule has a phosphoryl end, called the 5'-end, and a hydroxyl or 3'-end. These define the 5'→3' direction.
- **There are three reading frames that can be read in this 5'→3' direction, each beginning from a different nucleotide in a triplet.**
- **Usually only one reading frame is translated** and the other two are blocked by frequent termination signals.
- **An open reading frame starts with AUG and continues in triplets to a termination codon.**
- Blocked reading frames may be interrupted frequently by termination codons.



5' AGGTGACACCGCAAGCCTTATATTAGC 3'

Possible reading frames:

AGG·TGA·CAC·CGC·AAG·CCT·TAT·ATT·AGC

A·GGT·GAC·ACC·GCA·AGC·CTT·ATA·TTA·GC

AG·GTG·ACA·CCG·CAA·GCC·TTA·TAT·TAG·C

		Second nucleotide					
		U	C	A	G		
First nucleotide	U	UUU Phe	UCU	UAU Tyr	UGU Cys	Third nucleotide	U
		UUC	UCC Ser	UAC	UGC		C
		UUA Leu	UCA	UAA STOP	UGA STOP		A
		UUG	UCG	UAG STOP	UGG Trp		G
	C	CUU	CCU	CAU His	CGU		U
		CUC Leu	CCC Pro	CAC	CGC Arg		C
		CUA	CCA	CAA Gln	CGA		A
		CUG	CCG	CAG	CGG		G
	A	AUU Ile	ACU	AAU Asn	AGU Ser		U
		AUC	ACC Thr	AAC	AGC		C
		AUA	ACA	AAA Lys	AGA Arg		A
		AUG Met	ACG	AAG	AGG		G
	G	GUU	GCU	GAU Asp	GGU		U
		GUC Val	GCC Ala	GAC	GGC Gly		C
		GUA	GCA	GAA Glu	GGA		A
		GUG	GCG	GAG	GGG		G

Effect of Mutations:

- Mutations that insert or delete individual bases cause a shift in the triplet sets after the site of mutation.
- Combinations of mutations that together insert or delete 3 bases (or multiples of three) insert or delete amino acids but do not change the reading of the triplets beyond the last site of mutation.

1. Point mutation - Can it effect?

- The nature of the code predicts that two types of mutations will have different effects.
- If a particular sequence is read sequentially,

such as UUU AAA GGG CCC (codons)

a a 1 aa2 aa3 aa4 (amino acids) then a point mutation will affect only one amino acid.

For example, because only the second codon has been changed, the substitution of an A by some other base (X) causes aa2 to be replaced by aa5:

UUU **AAX** GGG CCC

aa1 aa5 aa3 aa4

2. Insertions and deletions:

But a mutation that inserts or deletes a single base will change the triplet sets for the entire subsequent sequence.

A change of this sort is called a frameshift.

An insertion might take the following form:

UUU AAX GGG CCC - treated as sequence for introducing A

UUU AAX AGG GCC C

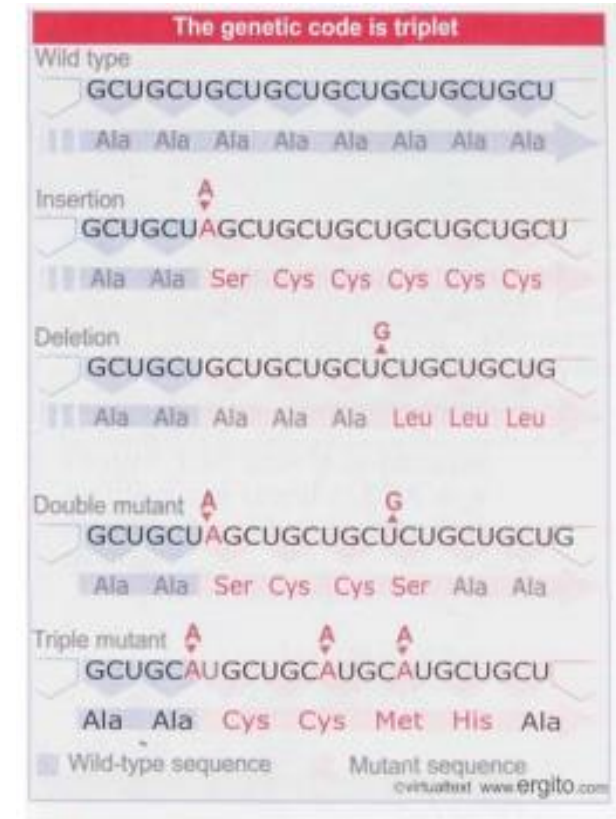
aa1 aa5 aa6 aa7

Because the new sequence of triplets is completely different from the old one, **the entire amino acid sequence of the protein is altered beyond the site of mutation**. So the function of the protein is likely to be lost completely.

Frameshift mutations are **induced by the acridines, compounds that bind to DNA** and distort the structure of the double helix, causing additional bases to be incorporated or omitted during replication.

Each mutagenic event sponsored by an acridine results in the addition or removal of a single base pair.

GET YOUR CLONED GENE SEQUENCED



Frameshift mutation

- The structure and/or enzymatic activity of each protein follows from its primary sequence of amino acids. By determining the sequence of amino acids in each protein, the gene is able to carry all the information needed to specify an active polypeptide chain.
- In addition to sequences that code for proteins, DNA also contains certain sequences whose function is to be recognized by regulator molecules, usually proteins. Here the function of the DNA is determined by its sequence directly, not via any intermediary code.
- Both types of regions, genes expressed as proteins and sequences recognized as such, constitute genetic information.
- In any given region, only one of the two strands of DNA codes for protein, so we write the genetic code as a sequence of bases (rather than base pairs).
- The identification of a lengthy open reading frame is taken to be prima facie evidence that the sequence is translated into protein in that frame.
- An open reading frame (ORF) for which no protein product has been identified is sometimes called an unidentified reading frame (URF)

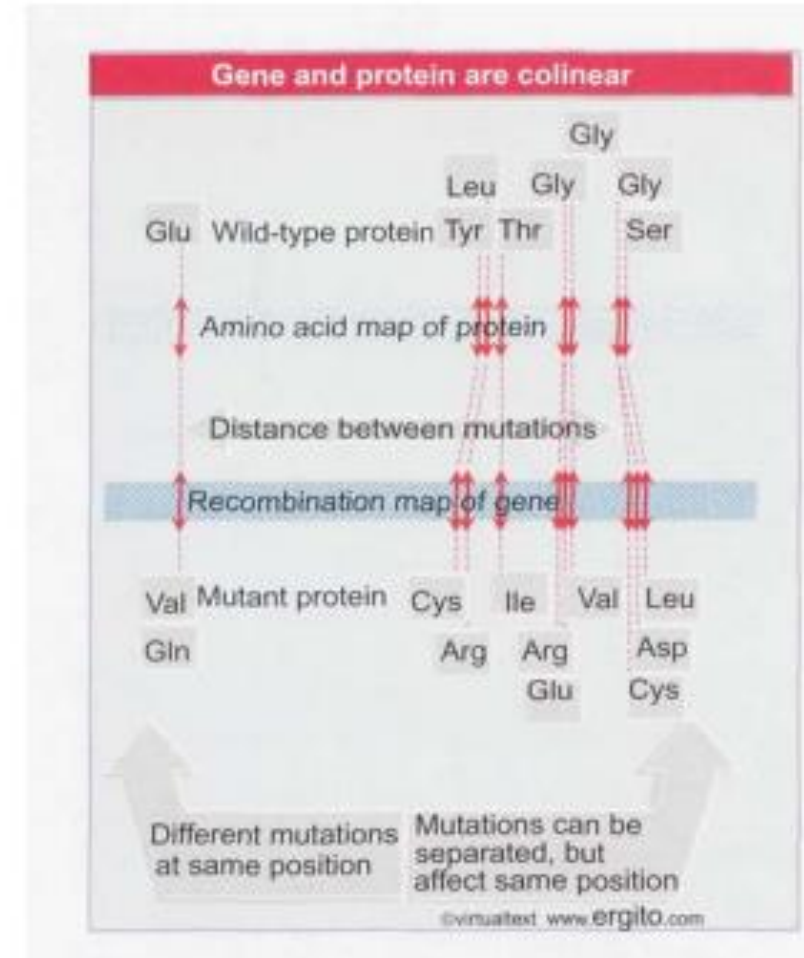
Prokaryotic genes are co linear with their proteins

By comparing the nucleotide sequence of a gene with the amino acid sequence of a protein, we can determine directly whether the gene and the protein are colinear: whether the sequence of nucleotides in the gene corresponds exactly with the sequence of amino acids in the protein.

In bacteria and their viruses, there is an exact equivalence.

Each gene contains a continuous stretch of DNA whose length is directly related to the number of amino acids in the protein that it represents.

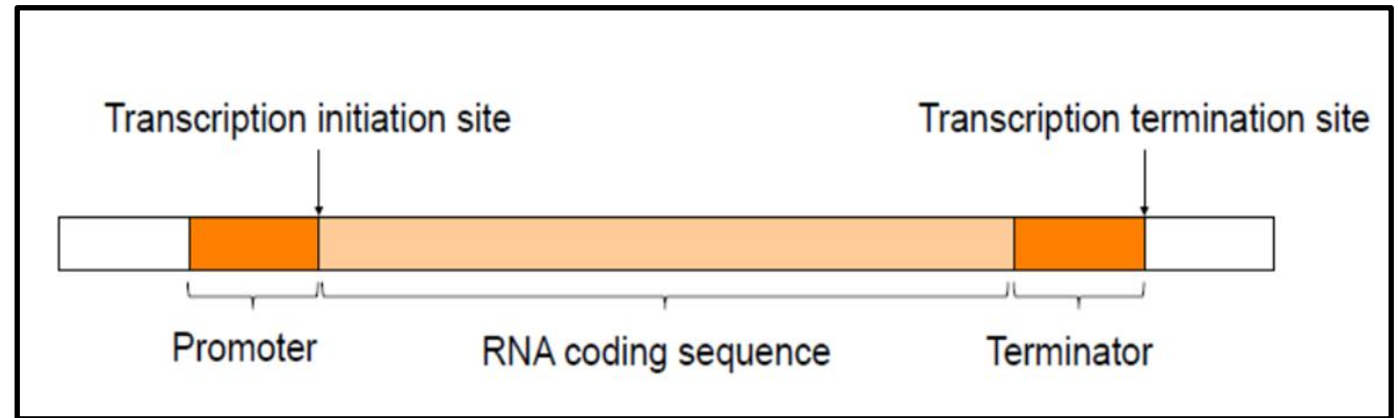
A gene of $3N$ bp is required to code for a protein of N amino acids, according to the genetic code.



A prokaryotic gene consists of a continuous length of $3N$ nucleotides that codes for N amino acids. The gene, mRNA, and protein are all colinear.

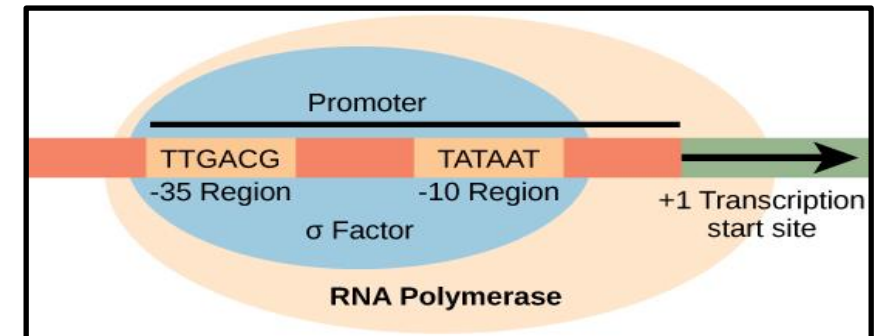
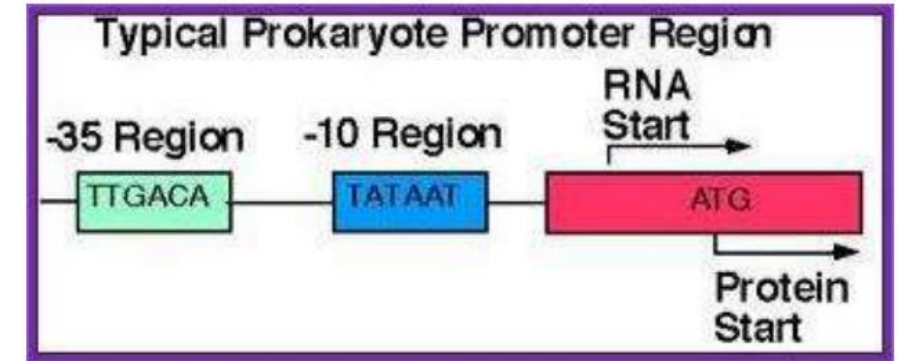
Components of prokaryotic gene

1. Promoter region.
 2. RNA coding sequence
 3. Terminator region
- Prokaryotic gene is continuous and with no introns
 - The region 5' of the promoter sequence is called upstream sequence and the region 3' of the terminator sequence is called downstream sequence.



Promoter region

- This is situated on upstream of the sequence that codes for RNA.
- This is the site that **interact RNA polymerase** before RNA synthesis (Transcription).
- Promoter region **provides the location and direction** to initiate transcription.
- At -10 there is a sequence TATAAT or PRIBNOW BOX.
- At -35 another consensus sequence TTGACA.
- These two are the most important promoter elements recognized by transcription factors.



The σ subunit of prokaryotic RNA polymerase recognizes consensus sequences found in the promoter region upstream of the transcription start sight. The σ subunit dissociates from the polymerase after transcription has been initiated.

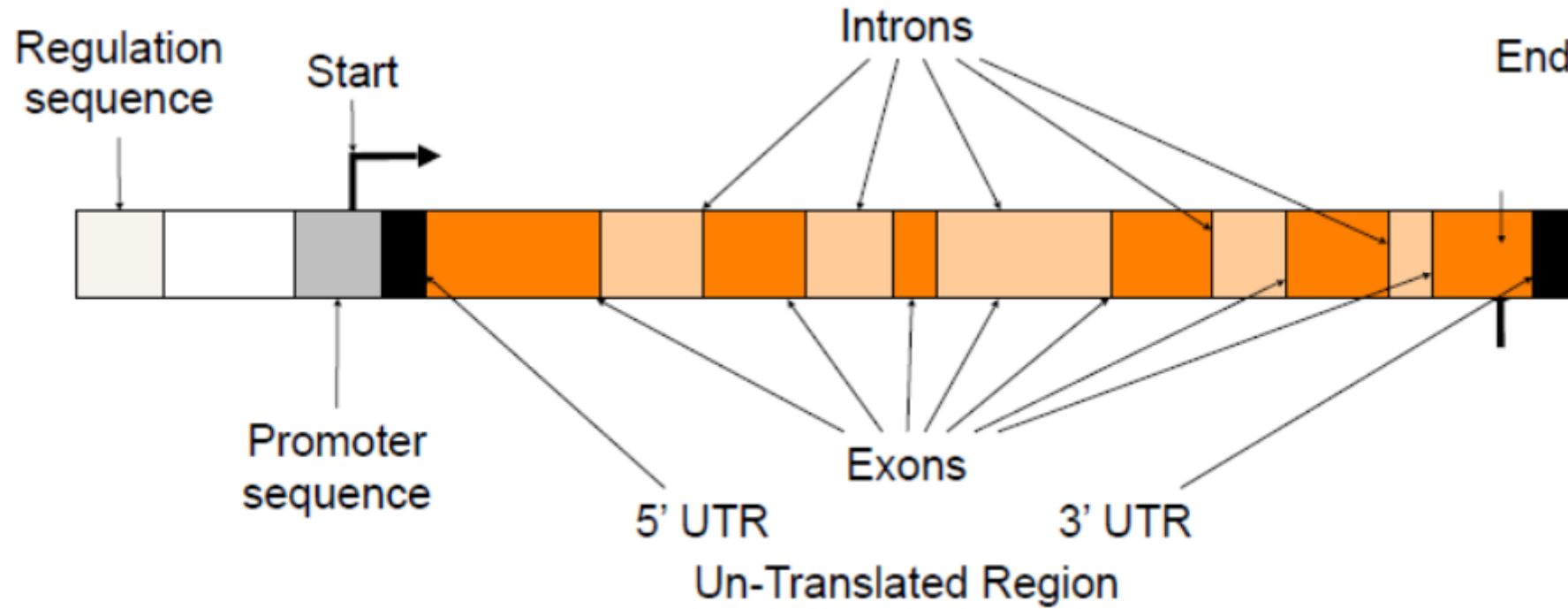
RNA coding sequence:

- The DNA sequence that will become copied into an RNA molecule (RNA transcript).
- Starts with an initiator codon and ends with termination codon.
- No introns (uninterrupted).
- Collinear to its mRNA.
- Any nucleotide present on the left is denoted by (-)symbol and the region is called upstream element. e.g. -10,-20,-35 etc.
- Any sequence to the right of the start is downstream elements and numbered as +10,+35 etc.).

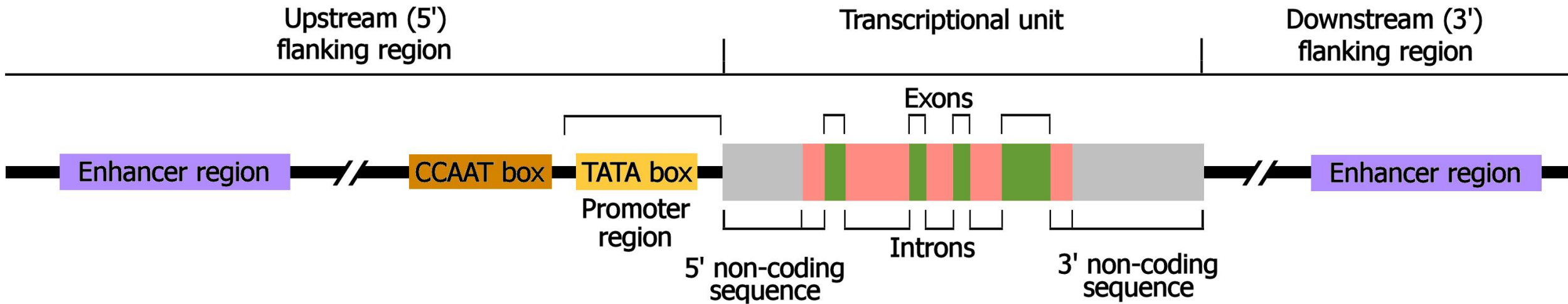
Terminator region:

- The region that signal the RNA polymerase to stop transcription from DNA template.
- Transcription termination occur through **Rho dependent or Rho independent manner.**

Eukaryotic gene structure



Eukaryotic gene structure



Eukaryotic gene structure

- They are composed of following regions :
- Exons
- Introns
- Promoter sequences
- Terminator sequences
- Upstream sequences
- Downstream sequences
- Enhancers and silencers(upstream or downstream)
- Signals (Upstream sequence signal for addition of cap. Downstream sequences signal for addition of poly A tail.)

Exons

- Coding sequence, transcribed and translated.
- Coding for amino acids in the polypeptide chain.
- Vary in number, sequence and length. A gene starts and ends with exons.(5' to 3').

Introns

- Coding sequences are separated by noncoding sequences called introns.
- They are removed when the primary transcript is processed to give the mature RNA.
- Introns were 1st discovered in 1977 independently by Phillip Sharp and Richard Roberts.
- Introns don't specify the synthesis of proteins but have other important cellular activities.
- Many introns encodes RNA's that are major regulators of gene expression.
- Contain regulatory sequences that control transcription and mRNA processing.
- Introns allow exons to be joined in different combinations(**alternative splicing**), resulting in the synthesis of different proteins from the same gene.
- Important role in evolution by facilitating recombination between exons of different genes(**exon shuffling**).

Promoters

- A promoter is a regulatory region of DNA located upstream controlling gene expression.
- 1. Core promoter – transcription start site(-34) Binding site for RNA polymerase and it is a general transcription factor binding sites.
- 2. Proximal promoter-contain primary regulatory element.
- These together are responsible for binding of RNA polymerase II which is responsible for transcription.

Upstream (5'end)

- 5'UTR serve several functions including mRNA transport and initiation of translation.
- Signal for addition of cap(7 methyl guanine) to the 5'end of the mRNA.
- The cap facilitates the initiation of translation.
- Stabilization of mRNA.

Downstream (3'end)

- 3'UTR serves to add mRNA
- stability and attachment site for poly-A-tail.
- The translation termination codon TAA.
- AATAA sequence signal for addition of poly A tail.

- Terminator- Recognized by RNA polymerase as a signal to stop transcription
- Enhancer -Enhances the transcription of a gene upto few thousand bp upstream.
- Silencers -Reduce or shut down the expression of a near by gene.

Know the gene

- It is a prerequisite for detailed functional annotation of genes and genomes.
- It can detect location of ORFs(Open Reading Frames), structures of introns and exons.
- It describes all the genes computationally with near 100% accuracy.
- It can reduce the amount of experimental verification work required.

Gene prediction is easier in prokaryotic genomes.

Smaller genomes, high gene density, very few repetitive sequence, more sequenced genomes.

Start codon is ATG.

Ribosomal binding site/Shine Dalgarno sequence.

Eukaryotic gene prediction

- Genomes are much larger than prokaryotes(10Mbp to 670 Gbp).
- Low gene density.
- Space between genes very large and rich in repetitive sequences & transposable elements.
- Splitting of genes by intervening non-coding sequences (introns) and joining of coding sequences(exons).
- Splice junctions follow GT-AG rule.
- An intron at the 5' splice junction has a consensus motif GTAAGT and that at 3' end NCAG. exon 1
exon 2
- Genes have a high density of CG dinucleotides near the transcription start site. This region is CpG island. It helps to identify the transcription initiation site of an eukaryotic gene.
- Some post-transcriptional modification occur with the transcript to become mature mRNA viz. Capping, Splicing and Polyadenylation. Acceptor Site Donor Site GT AG
- CAPPING: Occurs at the 5' end of the transcript. It involves methylation at the initial residue of the RNA.
- SPLICING: Process of removal of introns and joining of exons. It involves a large RNA-protein complex called spliceosome.
- POLYADENYLATION: Addition of a stretch of As (~250) at the 3' end of the RNA. The process is accomplished by **poly-A polymerase**.

Gene Prediction Methods

There are mainly two classes of methods for computational gene prediction.

1. Sequence similarity searches.
2. The other is gene structure and signal-based searches, which is also referred to as *ab initio* gene finding.

Sequence similarity searches

- Sequence similarity search is a conceptually simple approach that is based on finding similarity in gene sequences between ESTs (expressed sequence tags), proteins, or other genomes to the input genome.

This approach is based on the assumption that functional regions (exons) are more conserved evolutionarily than nonfunctional regions (intergenic or intronic regions).

Once there is similarity between a certain genomic region and an EST, DNA, or protein, the similarity information can be used to infer gene structure or function of that region.

EST-based sequence similarity usually has **drawbacks in that ESTs only correspond to small portions of the gene sequence**, which means that it is often difficult to predict the complete gene structure of a given region.

Local alignment and global alignment are two methods based on similarity searches.

The most common **local alignment tool is the BLAST family of programs**, which detects sequence similarity to known genes, proteins, or ESTs.

Two more types of software, PROCRUSTES and GeneWise, use global alignment of a homologous protein to translated ORFs in a genomic sequence for gene prediction.

A new heuristic method based on pairwise genome comparison has been implemented in the software called CSTfinder. The biggest limitation to this type of approaches is that only about half of the genes being discovered have significant homology to genes in the databases.

***Ab initio* gene prediction methods** – This methods for the computational identification of genes is to use **gene structure as a template to detect genes**, which is also called *ab initio* prediction.

Ab initio gene predictions rely on two types of sequence information: **signal sensors and content sensors**.

Signal sensors refer to **short sequence motifs**, such as **splice sites**, **branch points**, **polypyrimidine tracts**, **start codons** and **stop codons**.

Exon detection must rely on the content sensors, which refer to the patterns of codon usage that are unique to a species, and allow coding sequences to be distinguished from the surrounding non-coding sequences by statistical detection algorithms.

- Many algorithms are applied for modeling gene structure, such as Dynamic Programming, linear discriminant analysis, Linguist methods, Hidden Markov Model and Neural Network. Based on these models, a great number of *ab initio* gene prediction programs have been developed.

Table - ab initio gene prediction programs

<i>Ab initio</i> Gene Prediction Programs (Possibly with Homology Integration)				
Program	Organism	Algorithm [*]	Website	Homology
GeneID	Vertebrates, plants	DP	http://www1.imim.es/geneid.html	
FGENESH	Human, mouse, Drosophila, rice	HMM	http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind	
GeneParser	Vertebrates	NN	http://beagle.colorado.edu/~eesnyder/GeneParser.html	EST
Genie	Drosophila, human, other	GHMM	http://www.fruitfly.org/seq_tools/genie.html	protein
GenLang	Vertebrates, Drosophila, dicots	Grammar rule	http://www.cbil.upenn.edu/genlang/genlang_home.html	
GENSCAN	Vertebrates, Arabidopsis, maize	GHMM	http://genes.mit.edu/GENSCAN.html	
GlimmerM	Small eukaryotes, Arabidopsis, rice	IMM	http://www.tigr.org/tdb/glimmerm/glmr_form.html	
GRAIL	Human, mouse, Arabidopsis, Drosophila	NN, DP	http://compbio.ornl.gov/Grail-bin/EmptyGrailForm	EST, cDNA
HMMgene	Vertebrates, <i>C. elegans</i>	CHMM	http://www.cbs.dtu.dk/services/HMMgene/	
AUGUSTUS	Human, Arabidopsis	IMM, WWAM	http://augustus.gobics.de/	
MZEF	Human, mouse, Arabidopsis, Fission yeast	Quadratic discriminant analysis	http://rulai.cshl.org/tools/genefinder/	
[*] DP, dynamic programming; NN, neural network; MM, Markov model; HMM, Hidden Markov model; CHMM, class HMM; GHMM, generalized HMM; IMM, interpolated MM.				

Among which the programs GeneParser, Genie and GRAIL combine similarity searches.

Gene finding programs

- GeneMark.hmm (microbial genomes)
- Glimmer (UNIX program from TIGR). Computation involves two steps viz. model building & gene prediction.
- FGENESB (bacterial sequences). It uses Vertibi algorithm & linear discriminant analysis(LDA).
- RBSfinder- Searches from ribosomal binding site or shine dalgarno sequence for prediction of translation initiation site.



Questions