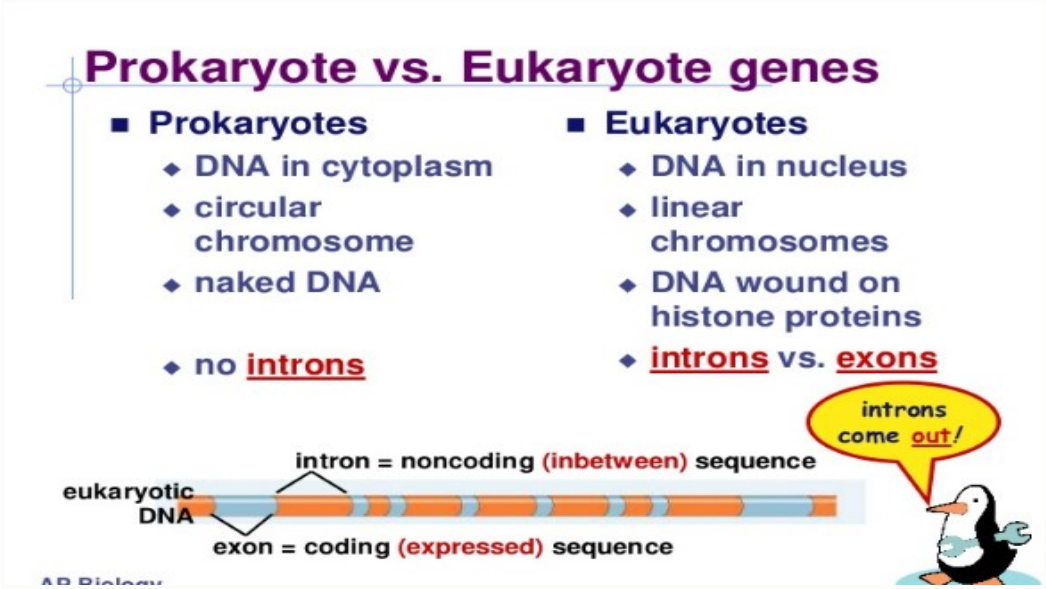
ModuleV

General introduction to Gene expression in prokaryotes and eukaryotes- Prokaryotic Genomes – Gene structure, GC content, Gene Density, Eukaryotic Genomes- Gene structure, GC content, Gene Density, Gene Expression, Transposition, Gene prediction approaches.

Genes



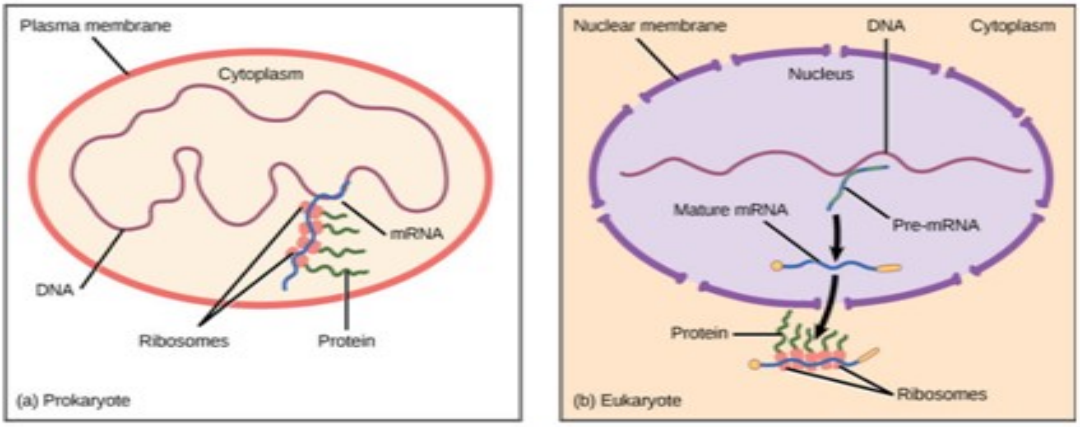


Figure 1. Prokaryotic transcription and translation occur simultaneously in the cytoplasm, and regulation occurs at the transcriptional level. Eukaryotic gene expression is regulated during transcription and RNA processing, which take place in the nucleus, and during protein translation, which takes place in the cytoplasm. Further regulation may occur through post-translational modifications of proteins.

Table 1. Differences in the Regulation of Gene Expression of Prokaryotic and Eukaryotic Organisms

Prokaryotic organisms	Eukaryotic organisms
Lack nucleus	Contain nucleus
DNA is found in the cytoplasm	DNA is confined to the nuclear compartment
RNA transcription and protein formation occur almost simultaneously	RNA transcription occurs prior to protein formation, and it takes place in the nucleus. Translation protein occurs in the cytoplasm.
Gene expression is regulated primarily at the transcriptional level	Gene expression is regulated at many levels (epigenetic, transcriptional, nuclear shuttling, post-transcriptional, translational, and post-translational)

Gene expression

Gene expression is the process by which the instructions in our DNA are converted into a functional product, such as a protein.

- The information stored in our DNA is converted into instructions for making proteins or other molecules, it is called gene expression.
- Gene expression is a tightly regulated process that allows a cell to respond to its changing environment.
- It acts as both an on/off switch to control when proteins are made and also a volume control that increases or decreases the amount of proteins made.
- The following are the steps involved in gene expression

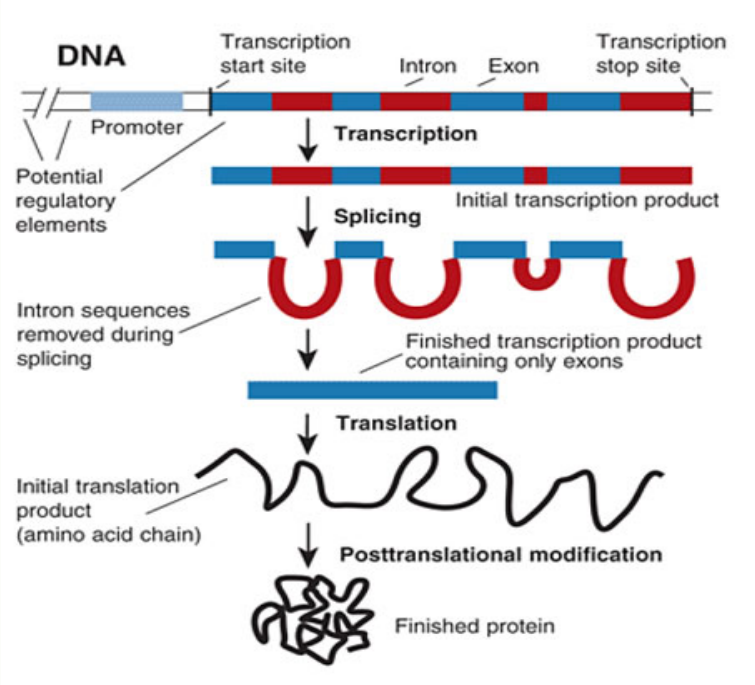
Transcription

- The first step is transcription.Here the DNA is converted into pre mRNA
- This is carried out by an enzyme called RNA polymerase which uses available bases from the nucleus of the cell to form the mRNA.

- The first (primary) transcript from DNA is called a pre-mRNA and contains both introns and exons. Introns and exons are parts of genes. Exons code for proteins, whereas introns do not
- RNA is a chemical similar in structure and properties to DNA, but it only has a single strand of bases and instead of the base thymine (T), RNA has a base called Uracil(U)

RNA Splicing

pre-mRNA contains both introns and exons . Introns and exons are parts of genes. Exons code for proteins, whereas introns do not .Pre-mRNA requires splicing (removal) of introns to produce the final mRNA molecule containing only exons(can be converted to proteins)



An illustration showing the process of transcription. Image credit: Genome Research Limited

Translation

- Translation occurs after the messenger RNA (mRNA) has carried the transcribed 'message' from the DNA to protein-making factories in the cell, called ribosomes.
- The message carried by the mRNA is read by a carrier molecule called transfer RNA (tRNA).
- The mRNA is read three letters (a codon) at a time.
- Each codon specifies a particular amino acid. For example, the three bases 'GGU' code for an amino acid called glycine.
- As there are only 20 amino acids but 64 potential combinations of codon, more than one codon can code for the same amino acid. For example, the codons 'GGU' and 'GGC' both code for glycine.
- Each amino acid is attached specifically to its own tRNA molecule.
- When the mRNA sequence is read, each tRNA molecule delivers its amino acid to the ribosome and binds temporarily to the corresponding codon on the mRNA molecule.
- Once the tRNA is bound, it releases its amino acid and the adjacent amino acids all join together into a long chain called a polypeptide.
- This process continues until a protein is formed.
- Proteins carry out most of the active functions of a cell.

Post-translational modification

Post-translational modification (**PTM**) refers to modification of proteins following protein biosynthesis. Proteins are synthesized by ribosomes translating mRNA into polypeptide chains, which may then undergo PTM to form the mature protein product. PTMs are important components in cell signaling, as for example when prohormones are converted to hormones.They can extend the chemical repertoire of the 20 standard amino acids by modifying an existing functional group or introducing a new one such as phosphate. Phosphorylation is a very common mechanism for regulating the activity of enzymes and is the most common post-translational modification.

Procaryotic genome

Prokaryotic organisms are single-celled organisms that lack a cell nucleus, and their DNA therefore floats freely in the cell cytoplasm. To synthesize a protein, the processes of transcription and translation occur almost simultaneously. When the resulting protein is no longer needed, transcription stops. As a result, the primary method to control what type of

protein and how much of each protein is expressed in a prokaryotic cell is the regulation of DNA transcription. All of the subsequent steps occur automatically. When more protein is required, more transcription occurs. Therefore, in prokaryotic cells, the control of gene expression is mostly at the transcriptional level.

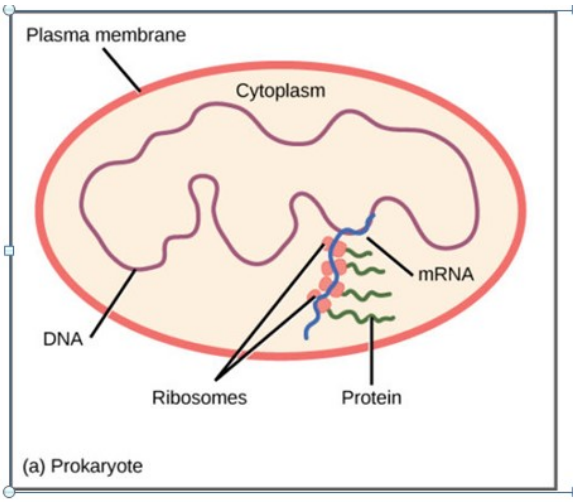
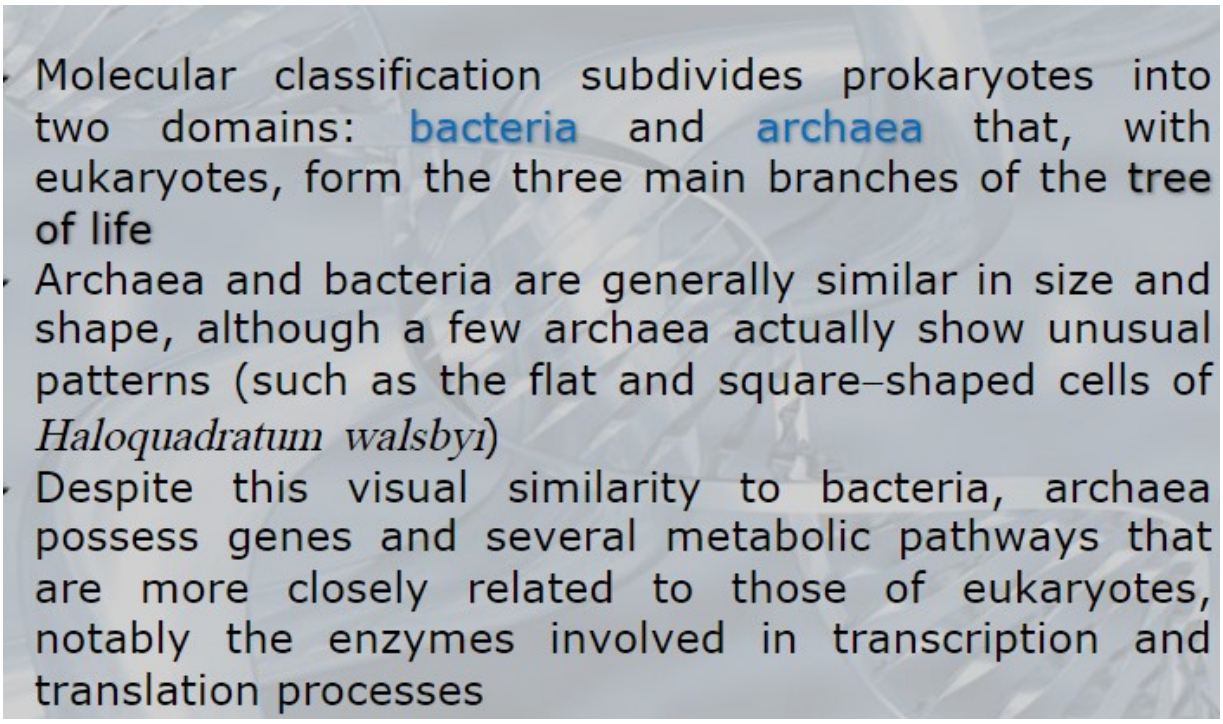
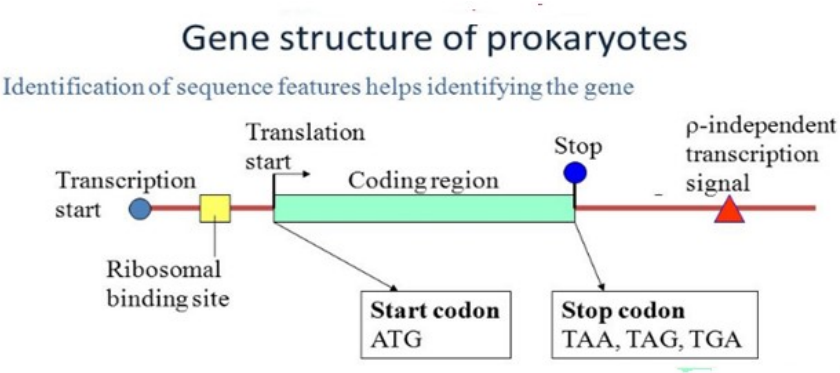


Figure 1. Prokaryotic transcription and translation occur simultaneously in the cytoplasm, and regulation occurs at the transcriptional level



Prokaryotic gene structure



-Before transcription start there is a region called promoter region ,where tha RNA polymerase starts transcrition
-Coding region also known as Open reading frame

Based on activity prokaryotic genes are classified as

1. House keeping genes-Housekeeping genes are involved in basic cell maintenance and, therefore, are expected to maintain constant expression levels in all cells and conditions.
2. Specific genes- Genes that are turned on in each cell and give specific properties to that cell

Structural features

- Simple gene structure
- Small genomes(.5-10 million bp)
- Prokaryotic coding region is collinear to its mRNA, which is collinear to polypeptide chain
- the coding region of structural genes is not split, but rRNA genes have spacers with in them.

The Structure contains the following

Promoter elements

The upstream elements from the start of the coding region include promoter elements with initiate RNA polymerase to start transcription

Transcription start site

The **transcription start site** is the location where transcription starts at the 5'-end of a gene sequence.

- Nearly 50 to 100 ntds upstream of the start codon, is the first nucleotide at which transcription initiates, it means, it is at this site the first nucleotide is incorporated into the transcribed RNA. The site is called transcriptional initiation site or START.

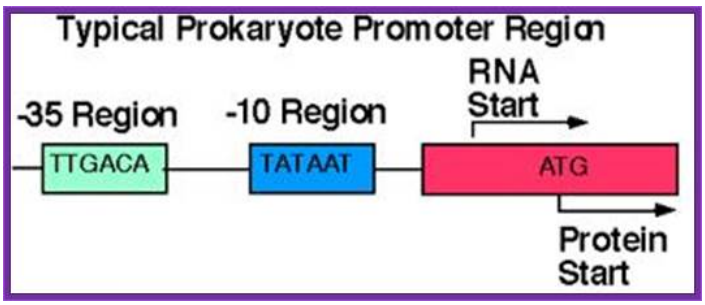
Ribosome binding site

A **ribosome binding site**, or **ribosomal binding site** (**RBS**), is a sequence of nucleotides upstream of the start codon of an mRNA transcript that is responsible for the recruitment of a ribosome during the initiation of protein translation.

Translation start site

The start codon is often preceded by a 5' untranslated region (5' UTR). In prokaryotes this includes the ribosome binding site.

- Nearly 10 nucleotides upstream of the start, there is a sequence called TATAAT or Pribnow box. Any nucleotide present on the left of the start is denoted by (-) symbol and the region is called upstream element. The numbers are written as -10, -20, -35 etc.
- The start site is the first ntds and symbolized by +1, any sequence to the right of the start is called down stream elements and numbered as +10, +35 and so on.



- At –35 there is another consensus sequence TTGACA. These two sequences are the most important promoter elements, for if there is any change in their sequence and position, transcriptional initiation suffers.

- The meaning of a promoter essentially is a distinct sequence module recognized by transcription factors that recruits RNA polymerase (as a holozyme) and bind to the sequence tightly and initiate transcription by unwinding the helically coiled DNA into transcriptional bubble.
- The said sequences not only facilitate the binding of TFs and enzyme and also provide sequence information for the site at which the enzyme to initiate transcription. If any one of the consensus sequences is deleted or changed drastically, the enzyme won't bind, even if it binds, it initiates transcription at different positions.

- Any nucleotide present on the left of the start is denoted by (-) symbol and the region is called upstream element. The numbers are written as -10, -20, -35 etc.

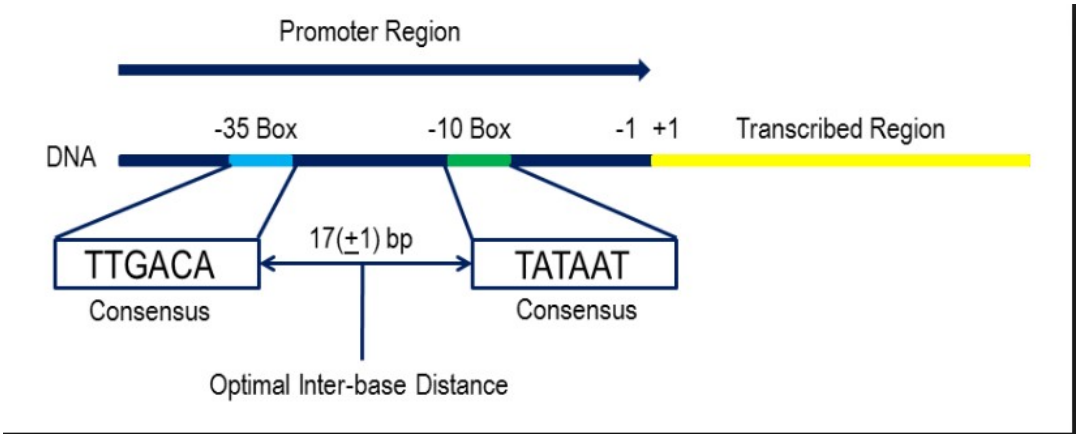
- The start site is the first ntds and symbolized by +1, any sequence to the right of the start is called down stream elements and numbered as +10, +35 and so on.

Coding region

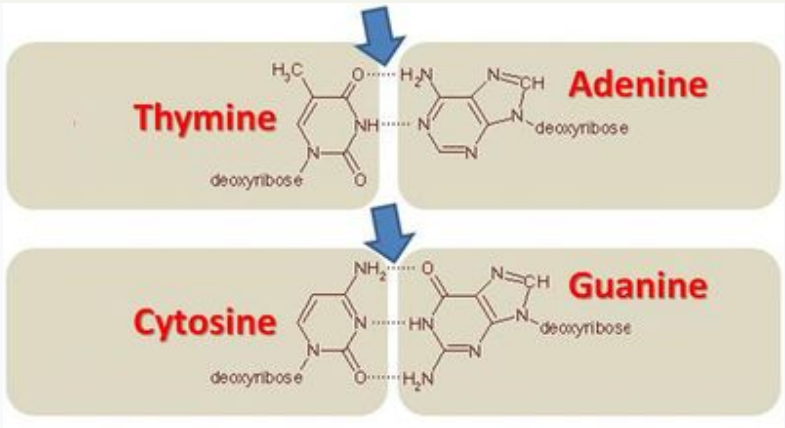
Coding region start with an initiator codon and ends with an end codon.This contain no introns. The **coding region** of a gene, also known as the **CDS** (from *coding sequence*), is that portion of a gene's DNA or RNA that codes for protein. The region usually begins at the 5' endby a start codon(ATG) and ends at the 3' end with a stop codon.(TAA,TAG,TGA)

Terminal region

Region of termination of transcription



GC content

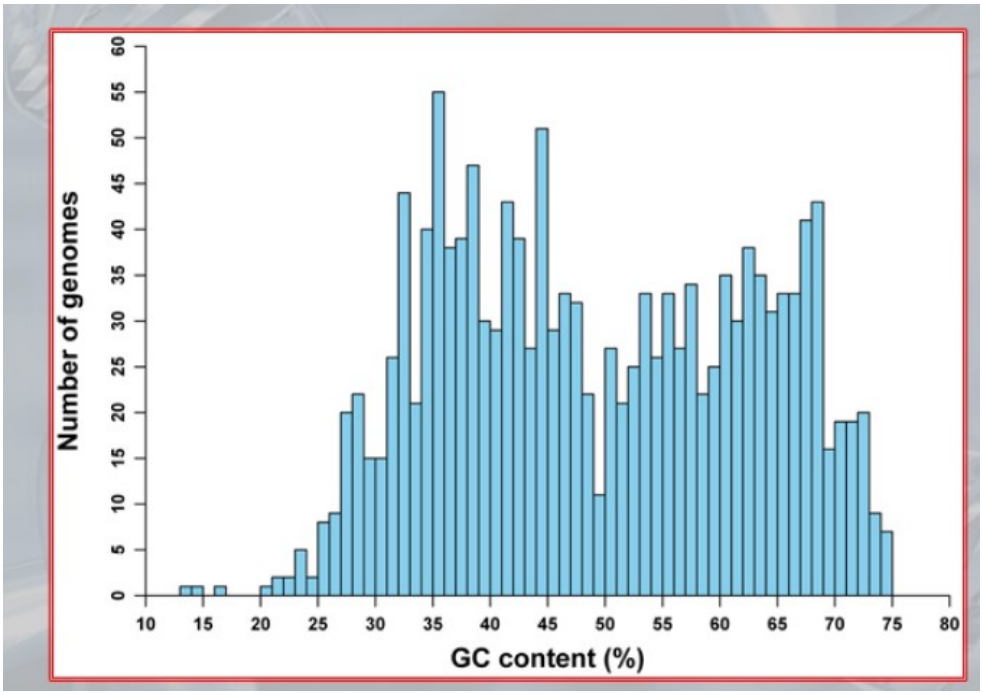


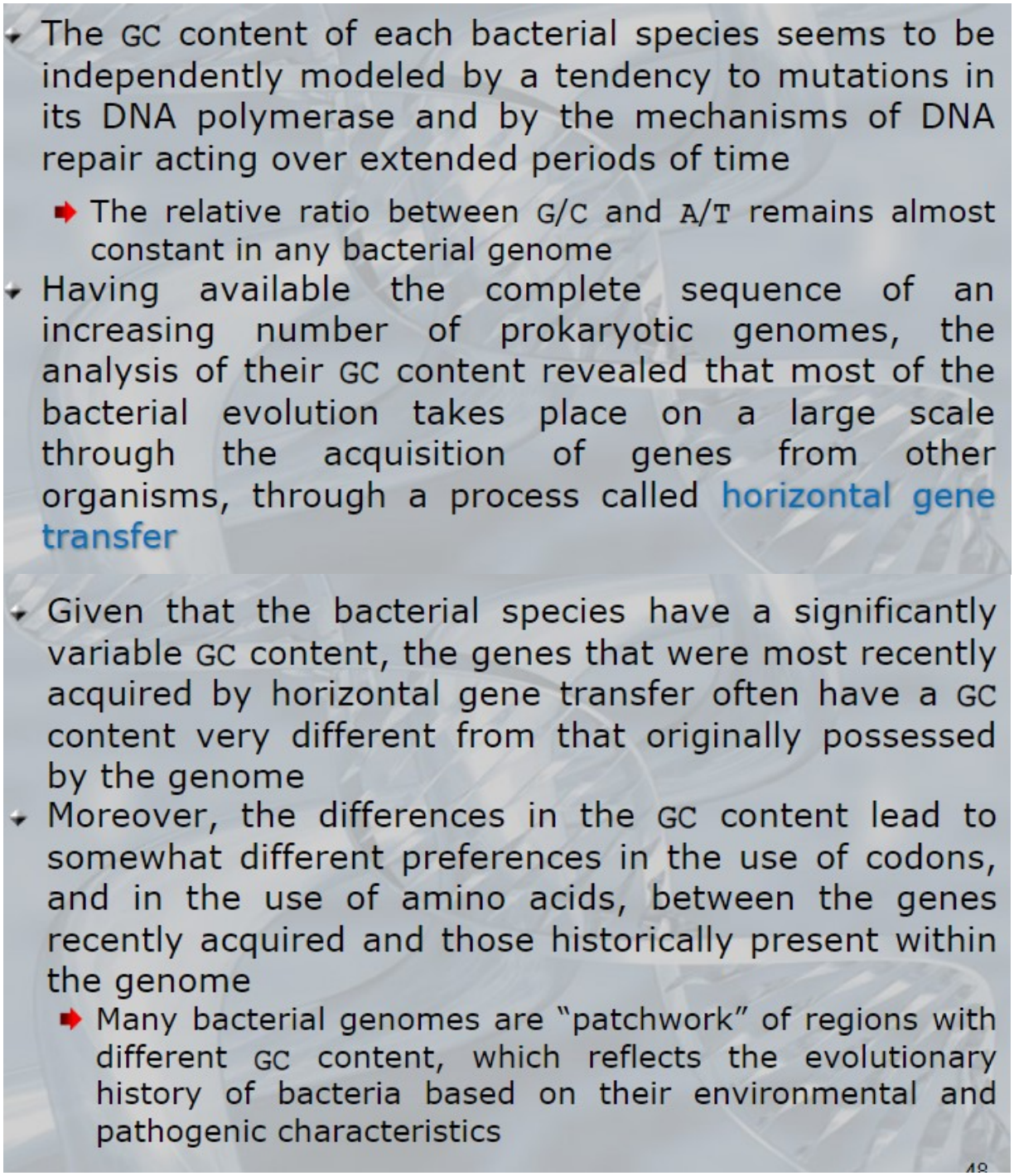
Nucleotide bonds showing AT and GC pairs. Arrows point to the hydrogen bonds.

GC-content

In molecular biology and genetics, **GC-content** (or **guanine-cytosine content**) is the percentage of nitrogenous bases on a DNA or RNA molecule that are either guanine or cytosine (from a possibility of four different ones, also including adenine and thymine in DNA and adenine and uracil in RNA).[1] This may refer to a certain fragment of DNA or RNA, or that of the whole genome. When it refers to a fragment of the genetic material, it may denote the GC-content of section of a gene (domain), single gene, group of genes (or gene clusters), or even a non-coding region.

- The coupling rules between bases require that, in a double stranded DNA, each G corresponds to a complementary C, but the only physical constraint with regard to the fraction of nucleotides G/C as opposed to that of A/T is that they sum up to 100%
- The abundance of nucleotides G and C with respect to A and T has long been recognized as a distinctive attribute of bacterial genomes
- The measurement of the GC content in prokaryotic genomes is very variable, ranging from 25% to 75%
- It was also noted that the base composition is not uniform along the genome





Procaryotic gene density

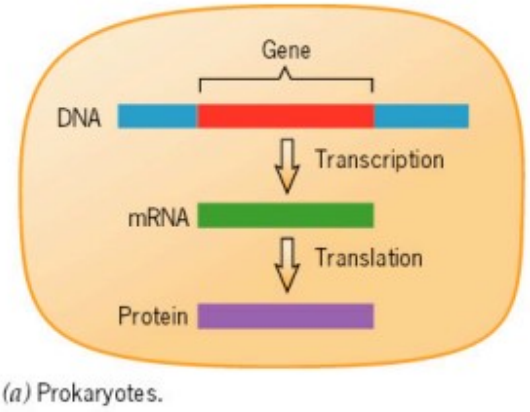
In genetics, the **gene density** of an organism's genome is the ratio of the number of genes per number of base pairs, usually written in terms of a million base pairs, or *megabase*(Mb). The human genome has a gene density of 12-15 genes/Mb Seemingly simple organisms, such as bacteria and amoebas, have a much higher gene density than humans. Bacterial DNA has a

gene density on the order of 500-1000 genes/Mb. This is due several factors, including that the fact that bacterial DNA has no introns. There are also fewer codons in bacterial genes.

Procaryotic Gene expression

Gene expression is the process by which the instructions in our DNA are converted into a functional product, such as a protein.

- The information stored in our DNA is converted into instructions for making proteins or other molecules, it is called gene expression.
- Gene expression is a tightly regulated process that allows a cell to respond to its changing environment.
- It acts as both an on/off switch to control when proteins are made and also a volume control that increases or decreases the amount of proteins made.
- The following are the steps involved in gene expression\



(a) Prokaryotes.

Transcription

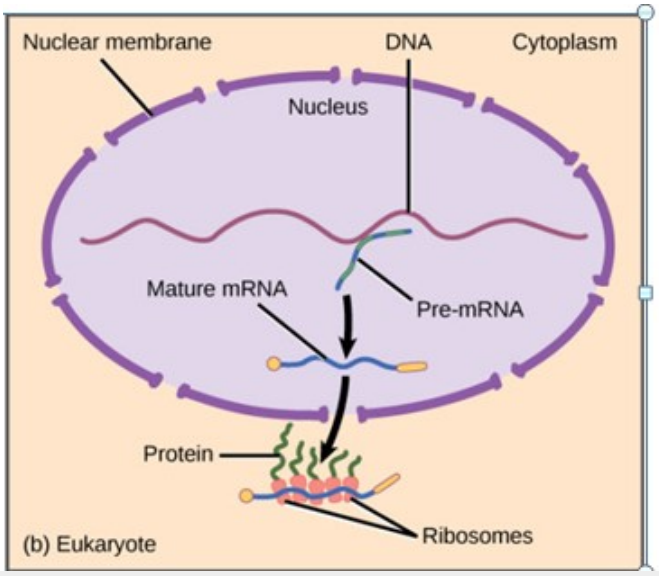
- The first step is transcription.Here the DNA is converted into mRNA
- This is carried out by an enzyme called RNA polymerase
- The transcript from DNA is called a mRNA and contains only exons. Exons code for proteins.
- RNA is a chemical similar in structure and properties to DNA, but it only has a single strand of base and instead of the base thymine (T), RNA has a base called Uracil(U)

Translation

- Translation occurs after the messenger RNA (mRNA) has carried the transcribed 'message' from the DNA to protein-making factories in the cell, called ribosomes.
- The message carried by the mRNA is read by a carrier molecule called transfer RNA (tRNA).
- The mRNA is read three letters (a codon) at a time.
- Each codon specifies a particular amino acid. For example, the three bases 'GGU' code for an amino acid called glycine.
- As there are only 20 amino acids but 64 potential combinations of codon, more than one codon can code for the same amino acid. For example, the codons 'GGU' and 'GGC' both code for glycine.
- Each amino acid is attached specifically to its own tRNA molecule.
- When the mRNA sequence is read, each tRNA molecule delivers its amino acid to the ribosome and binds temporarily to the corresponding codon on the mRNA molecule.
- Once the tRNA is bound, it releases its amino acid and the adjacent amino acids all join together into a long chain called a polypeptide.
- This process continues until a protein is formed.
- Proteins carry out most of the active functions of a cell.

Eucaryotic Genomes

Eukaryotic cells, in contrast, have intracellular organelles that add to their complexity. In eukaryotic cells, the DNA is contained inside the cell's nucleus and there it is transcribed into RNA. The newly synthesized RNA is then transported out of the nucleus into the cytoplasm, where ribosomes translate the RNA into protein. The processes of transcription and translation are physically separated by the nuclear membrane; transcription occurs only within the nucleus, and translation occurs only outside the nucleus in the cytoplasm. The regulation of gene expression can occur at all stages of the process (Figure 1). Regulation may occur when the DNA is uncoiled and loosened from nucleosomes to bind transcription factors (**epigenetic** level), when the RNA is transcribed (transcriptional level), when the RNA is processed and exported to the cytoplasm after it is transcribed (**post-transcriptional** level), when the RNA is translated into protein (translational level), or after the protein has been made (**post-translational** level



Eukaryotic gene expression is regulated during transcription and RNA processing, which take place in the nucleus, and during protein translation, which takes place in the cytoplasm. Further regulation may occur through post-translational modifications of proteins.

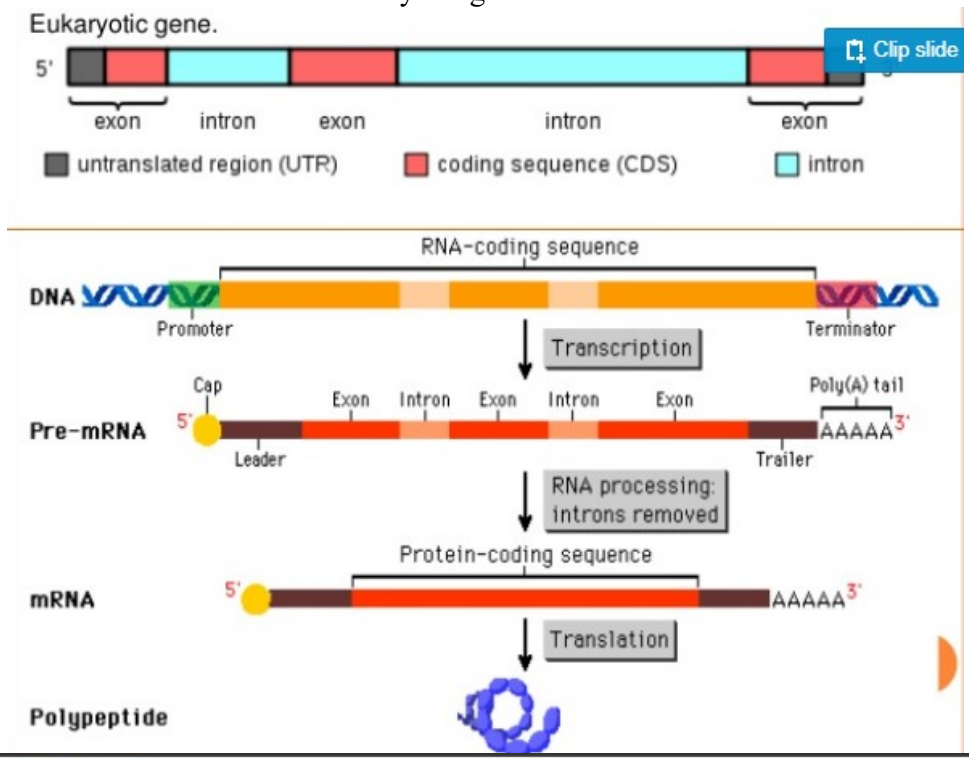
The differences in the regulation of gene expression between prokaryotes and eukaryotes are summarized in Table 1. The regulation of gene expression is discussed in detail in subsequent modules.

Table 1. Differences in the Regulation of Gene Expression of Prokaryotic and Eukaryotic Organisms

Prokaryotic organisms	Eukaryotic organisms
Lack nucleus	Contain nucleus
DNA is found in the cytoplasm	DNA is confined to the nuclear compartment
RNA transcription and protein formation occur almost simultaneously	RNA transcription occurs prior to protein formation, and it takes place in the nucleus. Translation of RNA to protein occurs in the cytoplasm.
Gene expression is regulated primarily at the transcriptional level	Gene expression is regulated at many levels (epigenetic, transcriptional, nuclear shuttling, post-transcriptional, translational, and post-translational)

Eucaryotic gene expression
Refer page 1(same):

Eukaryotic gene structure



❑EXONS –coding sequence, transcribed and translated. Coding for amino acids in the polypeptide chain.
Vary in number ,sequence and length. A gene starts an ends with exons.(5' to 3').
Some exon includes untranslated(UTR)region.

❑INTRONS- coding sequences are separated by non-coding sequences called introns.
Any nucleotide sequence that are removed when the primary transcript is processed to give the mature RNA are called introns.
All introns share the base sequence GT in the 5'end and AG in the 3'end.
Introns were 1st discovered in 1977 independently by Phillip Sharp and Richard Roberts.

❑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.
General transcription factor binding sites.
2. Proximal promoter-contain primary regulatory element
Apprx. -250,specific transcription factor binding sites.
❑TATA box or hogness box (-30 to -80)and CAAT(upstream TATA) are two distinct sequences.

These together are responsible for binding of RNA polymerase II which is responsible for transcription

❑UPSTREAM(5'END)- 5'UTR serve sevral functions including mRNA transport and initiation of translation.

Signal for addition of cap(7 methyl guanisine) 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

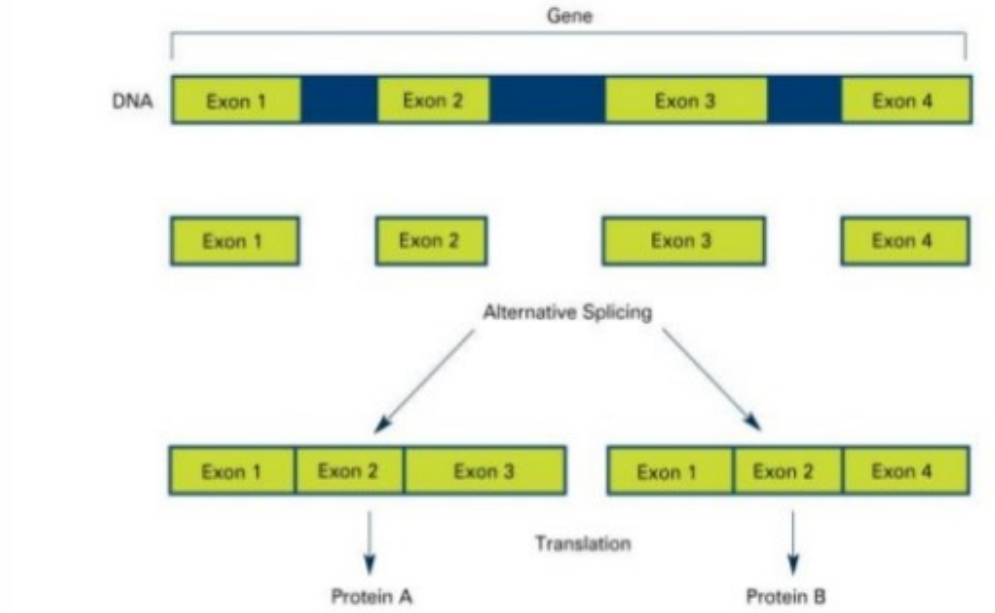
❑ENHANCER-enhances the transcription of a gene. Upto few thousand bp upstream.

❑SILENCERS-reduce or shut off the expression of a near by gene.

Significance of introns

Clip slide

✓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 trancription 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).



INTRON SIGNIFICANCE: ALTERNATIVE SPLICING

GC CONTENT

GC content plays an important in gene recognition algorithm because of two reasons
1.Eukaryotic ORF are much harder to recognize
2.Large scale variation of GC content within eukaryotic genomes underlies usefl correlation between genes and upsteam promoter sequence ,codon choices,gene legth and gene density

Transposition

The final method of changing the DNA in a genome that we will consider is **transposition**, which is the movement of DNA from one location to another. Segments of DNA with this ability to move are called **transposable elements**. Transposable elements were formerly thought to be found only in a few species, but now they are recognized as components of the genomes of virtually all species.They have significant influence on evolution

Transposable element is only one of several types found in Nature. Transposable elements can be divided into two major classes based on method of transposition: Transposons and retrotransposons .

Transposons and retrotransposons .

They are genetic components of DNA, and there are major differences between them. The percentage presence of these genetic materials varies across species, and their functions determine the fates of the organism with mutations and other phenotypically important changes. Transposons and retrotransposons are genes or collections of certain genes located in the DNA strands, and alterations of their locations have been the main causes for these consequences

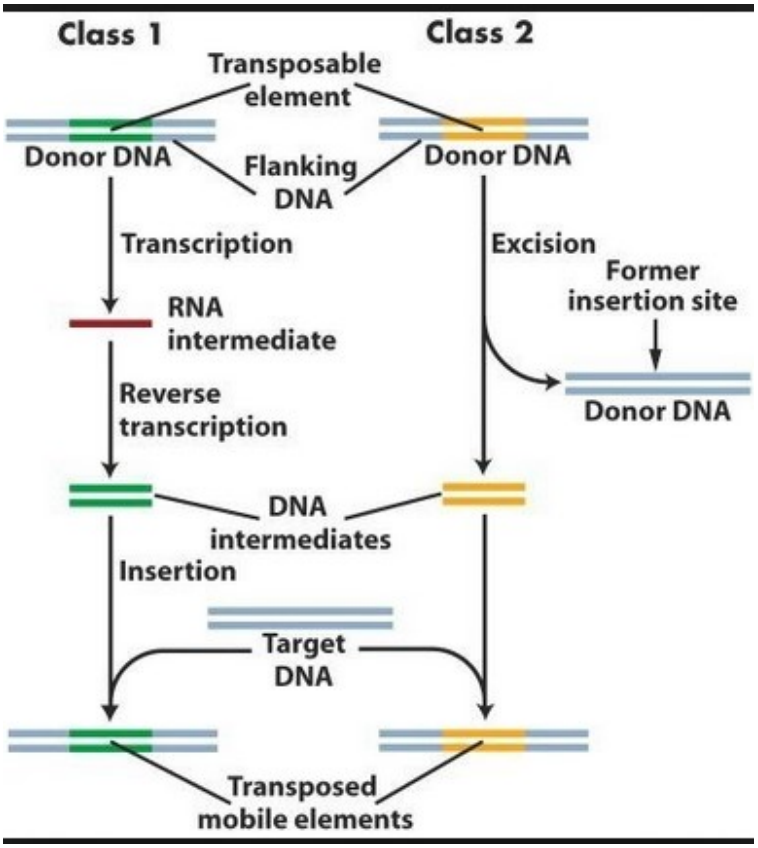
Retrotransposons (class 1)

- Use reverse transposase to make RNA intermediate for transposition.
- these move through the genome through the mechanism of copy and paste.
- The mechanism of the mobility of retrotransposons involves few major steps such as copying of the gene segment of the DNA strand into RNA, transfer of the copy of RNA to the target site, transcription of the RNA sequence back to DNA using reverse transcriptase, and insertion of the gene into the new location of DNA strand of genome
- Use reverse transcriptase for transposition.
- Found in viruses.

Transposons (class 2)

- Transposons are interesting fragments or segments of DNA with the ability to change the location of the DNA strand in the form of cut and paste mechanism.
- Because of this mobile nature of the transposons, these are known as jumping genes
- The processes of cutting and pasting of mobile DNA segments are regulated by the enzyme transposase. The enzyme binds to the both ends of the transposon and cuts the phosphodiester bonds of the DNA strand, isolate the transposon, move it to the target site, and bind in the new location.
- DNA fragments transpose directly from DNA segment to DNA segment
- Producing a DNA copy that transposes (replicative transposition) Or, cut/paste into a new locus (conservative transposition).
- Found in eukaryotes and prokaryotes.

The transpososome provides a scaffold to support the transposition reactions, changing its conformation to accommodate the different steps in transposition.



Gene prediction

In computational biology, **gene prediction** or **gene finding** refers to the process of identifying the regions of genomic DNA that encode genes. This includes protein-coding genes as well as RNA genes, but may also include prediction of other functional elements such as regulatory regions. Gene finding is one of the first and most important steps in understanding the genome of a species once it has been sequenced

Gene prediction methods

1. **Laboratory based methods**
2. **Feature based methods**
3. **Homology based methods**
4. **Statitical and HMM based methods**

Laboratory based methods

1 Blotting methods

Blotting is a technique used for detection of nucleic acids and proteins .The procedure involved following steps

1. first step preparing is cell free extract re containing biomolecule of interest
2. resolving the mixture by gel electrophoresis

3. transferring the unresolved mixture onto a membrane support such as nitrocellulose paper
4. incubating the paper with the detection system that specifically hybridize to the molecule of interest

Some important blotting methods are

Southern blotting

- Southern blotting is a technique for detecting specific **DNA** fragments in a complex mixture. The technique was invented in mid-1970s
- Southern blotting is an example of RFLP (restriction fragment length polymorphism). It was developed by Edward M. Southern (1975). Southern blotting is a hybridization technique for identification of particular size of **DNA** from the mixture of other similar molecules**.** This technique is based on the principle of separation of DNA fragments by gel electrophoresis and identified by labelled probe hybridization.
- Basically, the DNA fragments are separated on the basis of size and charge during electrophoresis. Separated DNA fragments after transferring on nylon membrane, the desired DNA is detected using specific DNA probe that is complementary to the desired DNA.
- A hybridization probe is a short (100-500bp), single stranded DNA. The probes are labeled with a marker so that they can be detected after hybridization.

Northern Blotting

- Northern blotting is used for detecting **RNA** fragments, instead of DNA fragments. The technique is called "Northern" simply because it is similar to "Southern", not because it was invented by a person named "Northern".
- In the Southern blotting, DNA fragments are denatured with alkaline solution. In the Northern blotting, RNA fragments are treated with formaldehyde to ensure linear conformation.

Western Blotting

- Western blotting is used to detect a particular **protein** in a mixture.
- The probe used is therefore not DNA or RNA, but antibodies.
- The technique is also called "immunoblotting"

Zoo bolts

- A **zoo blot** or **garden blot** is a type of Southern blot that demonstrates the similarity between specific, usually protein-coding, DNA sequences of different species.

- A zoo blot compares animal species while a garden blot compares plant species. The purpose of the zoo blot is to detect the conservation of the gene(s) of interest throughout the evolution of different species.[

2. Primer extension

- Primer extension is used to measure the amount of a given RNA and to map the 5' end of that RNA.

3.S1 nuclease mapping

- Is a single strand- specific endonuclease.
- Is useful for identifying 5' end of DNA and RNA
- S1 nuclease is useful for the removal of unpaired gegions following hybridization
- S1 mapping removes single stranded regions of DNA and RNA from double stranded versions
- It digests both single-stranded DNA & RNA.
- Is thermostable & resistant to denaturants/high salt concentration.
- It works best in an acidic P^H

4.Exon trapping

- **Exon trapping** is a molecular biology technique to identify potential exons in a fragment of eukaryote DNA of unknown intron-exon structure

5.*Reverse transcription polymerase chain reaction (RT-PCR)*

- **Reverse transcription polymerase chain reaction** (**RT-PCR**), a variant of polymerase chain reaction (PCR), is a technique commonly used detect RNA transcript of any gene

6.In situ hybridization

- Is a versatile method for the localization of specific mRNA in cells or tissues

Feaure based methods

Featurebaseed approaches were based on patern recognition and treats DNA fragments as sequences

Genefinding by ORF prediction

The region of the nucleotide sequences from the start codon (ATG) to the stop codon is called the Open Reading frame. Gene finding in organism specially prokaryotes starts form searching for an open reading frames (ORF). An ORF is a sequence of DNA that starts with start codon

"ATG" (not always) and ends with any of the three termination codons (TAA, TAG, TGA). Depending on the starting point, there are six possible ways (three on forward strand and three on complementary strand) of translating any nucleotide sequence into amino acid sequence according to the genetic code .These are called reading frames.While eukaryotic gene finding is altogether a different task as the eukaryotic genes are not continuous and interrupted by intervening noncoding sequences called 'introns'. Moreover organization of genetic information in eukaryotes and prokaryotes is different .

Following are some ORF prediction tools

GRAIL

- GRAIL (gene recognition and analysis Internet link) is an integrated artificial intelligent system
- This system uses a combination of a multi-sensor/neural network, expert system, and parallel search tools to recognize and interpret genes in DNA sequences.
- A simple electronic mail (E-mail) interface makes the system accessible through Internet.
- The strength of the system in recognizing and interpreting genes in DNA sequences and the simple E-mail interface have already attracted more than 150 users.

GRAIL II

- Helps in analyzing protein coding regions,poly(A) sites ans promoters
- Enables to construct gene models
- Predict encoded protein sequences
- Provised database searching capabilities
- Use neural network approach for evalution

Find patterns

- Used for scanning ORF frames

Frames

- Show ORF for the 6 traslation frames of a DNA sequence

Mac Vector 6.5

- **MacVector** is a commercial sequence analysis application for Apple Macintosh computers running Mac OS X.
- It is intended to be used by molecular biologists to help analyze, design, research and document their experiments in the laboratory.
- Does ORF detection based on the designation of sequence ends as start and stop codons

Sequencher

- Used for conig assembly,restriction enzyme mapping,heterozygote detection,cDNA to Genomic DNA large gap alignment,motif and SNP analysis

  Following are the major problem faced by the above tools

  Small proteins:cannot able to predict small protein

  Small exons: Exons smaller than about 3o neucleotide cannot be reliabily predict by the above methods

Homology based methods

Features in species being compared that are similar because they are ancestrally related based on genome pattern the homologous gene finding can be identified as:
1.Genome pattern based on number of nucleotides
2. genome size based on nucleotide base pairs
3.percentage of similarity between two similar species

Procrustes

- Is a hology based program
- It accept a genomic DNA sequence as and one or more protein sequences
- The proteins are assumed similar to the protein encoded in the genomic DNA fragment.
- It then find a chain of exon withbest bit to the target

TBLASTX

- Is the tool used for similarity search in the database
- if a protein sequence matches get its DNA sequence and align with your known sequence

statistical and HMM approches

HMM gene prediction

-use simplified grammer rules such as start codon,end codon ,legth is divisible by 3 and no stop codon in the reading frame.

- GeneMark: use HMM for gene identification
- HMM Gene:
  - For predicting anonymous genes
  - Apply prediction on whole genes so the predicted exons splice correctly
  - Also predict splice sites ,start and stop codons
- Glimmer

  - Gene Locator and Interpolated Morkov Modeler
  - Is the Predictor of genes in bacterial DNA

- Veil

  - Stands for Viterbi Exon Intron Locator
  - Used For Eucaryotic gene prediction

- GenScan

  - Used to predict complete gene structure
  - Used to identify inrons,exons,promoter elements,poly A signals etc..

- Genie

  - Use Hmm and Neural networks for prediction

- Signal Scan

  - Used for finding protein biding sites in DNA sequence

- GenLang

  - Is the syntactic pattern recognizing system for gene finding

- BCM GeneFinder

  - Predicting Exons and Introns and construct gene model

- Gene parser

  - Use likelyhood score for gene prediction

- Gene ID

  - Use position weight array for gene prediction