Institutional Development Plan (IDP),SKUAST Jammu

Strengthening Institutional Capacities for Delivering Competent Skilled Professionals

LECTURES DELIVERED IN REMEDIAL CLASSES OF SCHOOL F BIOTECHNOLOGY

Complied by:
Dr. Manmohan Sharma, Coordinator, SBT

Organized By : NAHEP-IDP, SKUAST- Jammu

CHAPTER 1

REGULATION OF GENE EXPRESSION

I. Introduction

A. No operon structures in eukaryotes

B. Regulation of gene expression is frequently tissue specific. This tissue specific gene expression is fundamental to the function of a particular cell or tissue

C. How does an organism express a subset of genes in one cell type and another subset in another cell type? Activation and repression.

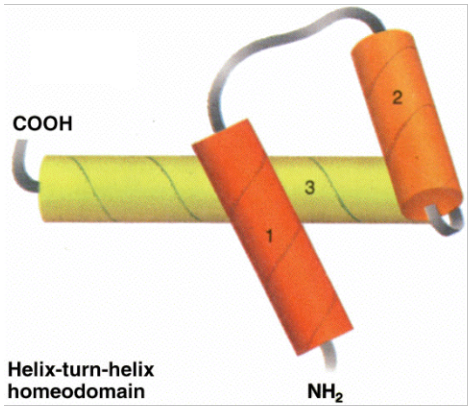
D. Multiple levels of regulation (transcriptional initiation, mRNA processing, mRNA stability, gene redundancy, gene amplification)

II. Transcriptional regulation

A. As with prokaryotes, transcriptional regulation is accomplished using cis-acting DNA sequences and trans-acting factors

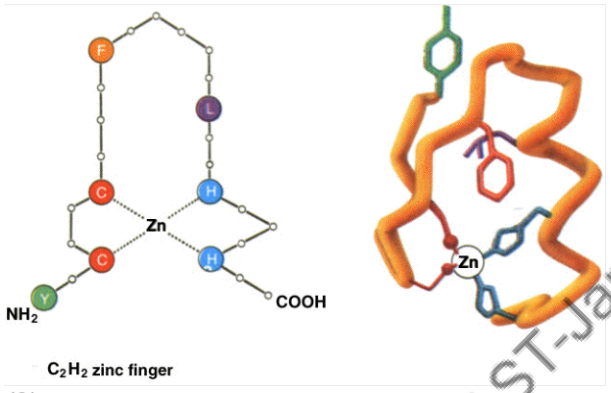
1. Cis-acting sequences
    a) promoters (see Chapter 13 notes)
    b) enhancers (see Chapter 13 notes)
2. Trans-acting proteins (generally have two domains – one to interact witha specific cis-acting DNA sequence and one to activate transcription)
    a) DNA binding motifs (also called homeodomains)
        (1) Helix-turn-helix (*Drosophila* developmental regulators; prokaryotic regulators)

Helices 1 and 2 make contact with other proteins and helix 3 contacts the DNA.



(2) Zinc fingers (Many steroid receptors and transcription factors for mRNAs)

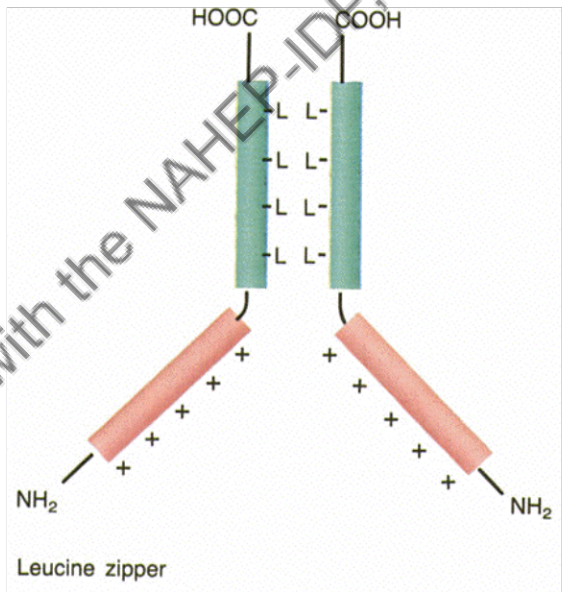
Histidine and cysteine bind zinc forming a finger-like structure that can bind DNA



C_2H_2 zinc finger

(3) Leucine zippers (many proto-oncogenes)

Dimer is formed between leucine rich regions of two monomers and is required for DNA binding; the DNA binding region is a positively charged region

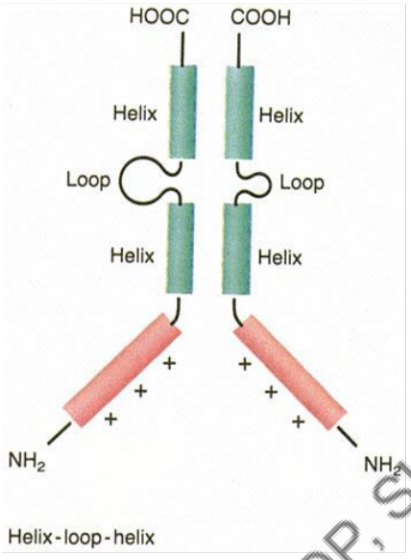


Leucine zipper

(From: AN INTRODUCTION TO GENETIC ANALYSIS 6/E BY Griffiths, Miller, Suzuki, Leontin, Gelbart © 1996 by W. H. Freeman and Company. Used with permission.)

(4) Helix-loop-helix

Dimer is formed between helix-loop-helix of two monomers and is required for DNA binding; the DNA binding region is a positively charged region



b) Transcriptional activation by trans-acting factors

(1) Stabilize RNA polymerase binding

(2) Unwind the DNA

(3) Attract other factors

(4) Formation of a DNA loop that places previously distantly bound activators in proximity to each other.

B. Regulation of the transcriptional regulator's activity (Example - steroid receptors and hormones)

Eukaryotic cells are capable of activating gene expression in response to hormones secreted by other cells. Steroid hormones enter the cell by diffusing through the membrane. Once inside the cell, they bind to a specific receptor. The receptor complexed with the hormone is now capable of activating transcription of genes that contain a hormone-responsive element (HRE). In some ways the mechanism of action of steroid hormone regulation is similar to the mechanism of action of the arabinose operon regulation in prokaryotes. Specifically, a small external effector molecule (arabinose or the hormone) binds to a cytoplasmic transcription factor (AraC or steroid receptor) which then binds to a specific DNA element (araO or HRE) and activates transcription.

C. DNA methylation to regulate transcription

1. The cytosine in the sequence CG is frequently methylated in many eukaryotes.

2. There is a correlation between decreased methylation of CG sequences and increased transcription.

a) The inactivated X chromosome is over-methylated except for the few genes that are transcribed.

b) Methylation patterns are tissue specific and are heritable for all cells in the tissue

c) Addition of the cytosine analogue 5'-azacytodine which can not be methylated activates previously unexpressed genes

D. Histone acetylation to regulate transcription

1. Histones are proteins that form a unit upon which the DNA is coiled in chromosomes.

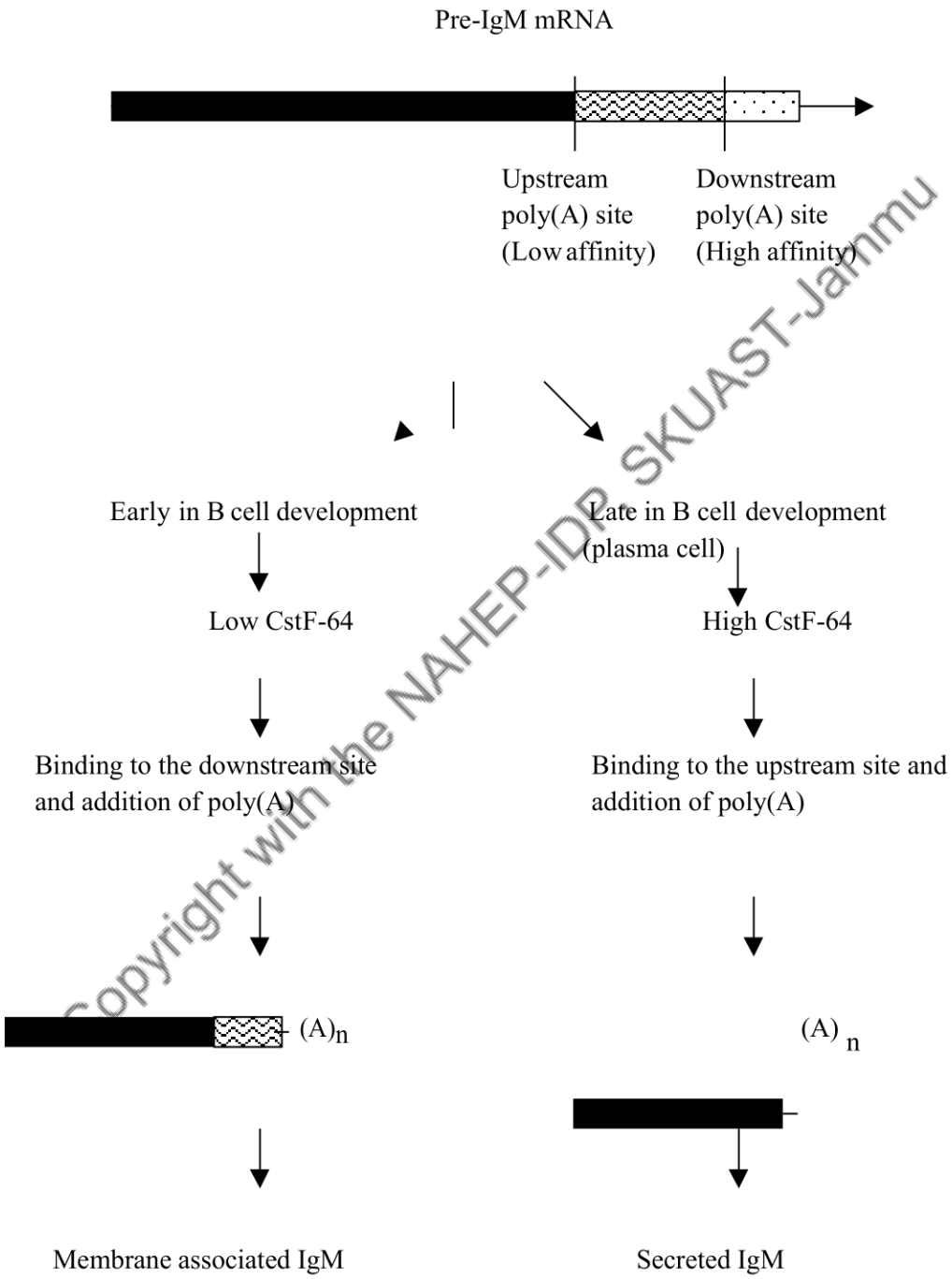
2. Acetylation of histones may loosen the DNA around the histone to allow the transcriptional apparatus access.

III. Post transcriptional regulation

A. Differential processing

1. Poly(A) site selection (example - IgM mRNA)

B cells produce proteins called antibodies. The IgM antibody is composed of two types of protein chains called heavy and light. The pre-mRNA encoding the heavy chain has two possible sites for addition of the poly(A) tail. Early in development of the B cell, the poly(A) tail is added to the downstream site producing an IgM molecule that is associated with the cell membrane. Later when the B cell has differentiated into a plasma cell, the poly(A) tail is added to the upstream site producing a IgM molecule that is secreted. Recent work suggests that poly(A) site selection is regulated by the concentration of a subunit of the enzyme cleavage stimulation factor (CstF-64). In early stage B cells, CstF-64 accumulation is repressed and poly(A) selection is directed to the downstream site. Artificially overexpressing CstF-64 in early B cells results in use of the upstream splice site and production of secreted IgM. The current model is that the CstF-64 has a higher affinity for the downstream site and this is why it is preferentially used in early B cells when the concentration of CstF-64 is low.

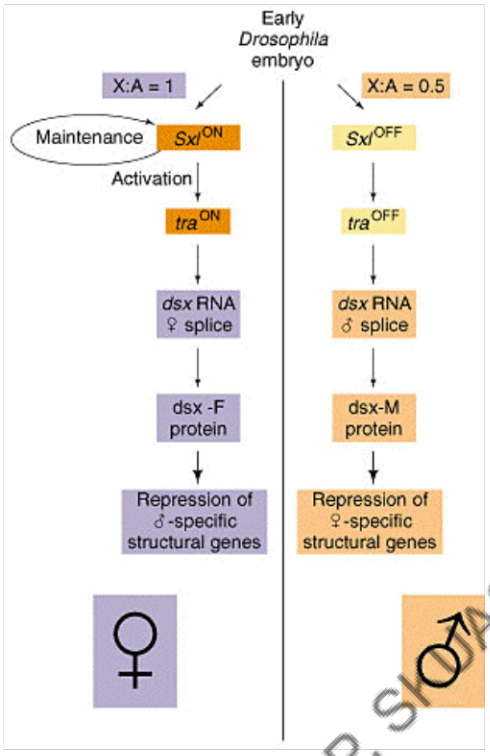


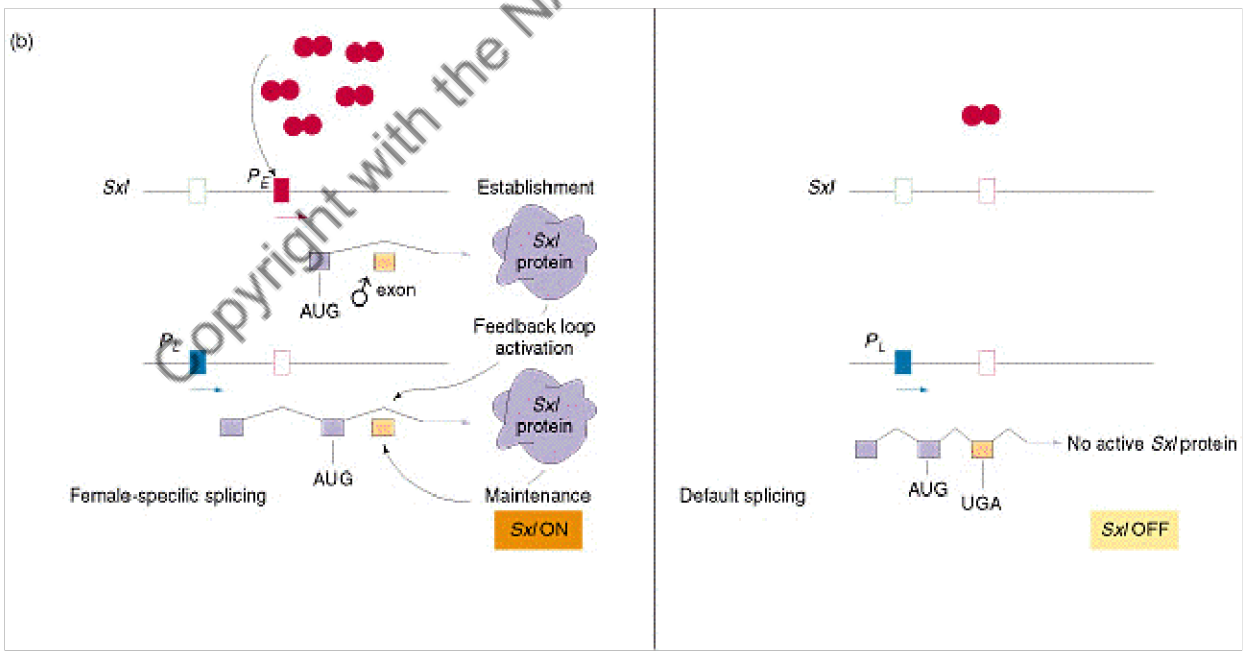
2. Splice site determination (example sex determination in *Drosophila*) In *Drosophila*, differential splicing of one mRNA transcript (*sxl*) initiates a cascade that eventually determines the sex characteristics of the fly. A transcription factor that activates a promoter of the *sxl* gene early in development is encoded on the X chromosome of flies. This factor functions as a homodimer. Another factor, encoded on an autosome, can interact with the X encoded factor and inactivate it by formation of nonfunctional heterodimers. Early in development in XX flies, there is sufficient X factor made relative to the autosomal inhibitor so that a low level of *sxl* is transcribed. In contrast, in XY flies, since the ratio of X to autosomal chromosome is only 0.5, there is less X encoded factor relative to the autosomal inactivator and so not enough functional X factor to activate *sxl* transcription.

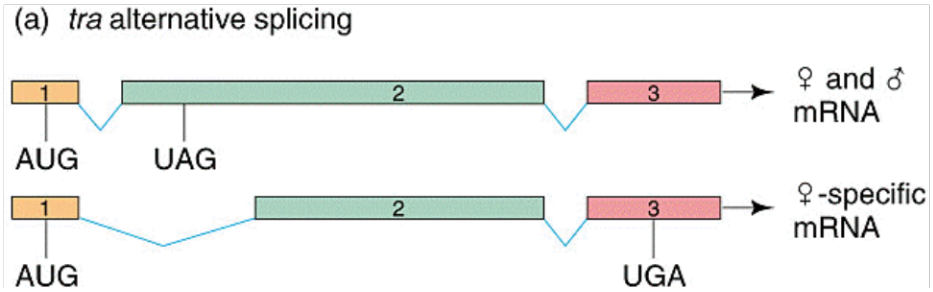
Later in development, the X encoded factor is not produced. *sxl* is now transcribed from a different promoter producing a longer pre mRNA.
For this pre mRNA to form mRNA encoding a functional Sxl protein, an exon containing a stop codon must be spliced out. Sxl represses splicing at the site that would leave the stop codon in the mRNA. Since there is Sxl is the female cells, the correct splicing of the *sxl* pre mRNA transcript will occur and more Sxl will be made which will catalyze more splicing of *sxl* pre mRNA. This is a positive autoregulatory loop. In contrast in the male cells, there is no Sxl and the exon containing a stop codon is not spliced out of the *sxl* pre mRNA. Thus, no Sxl is made.

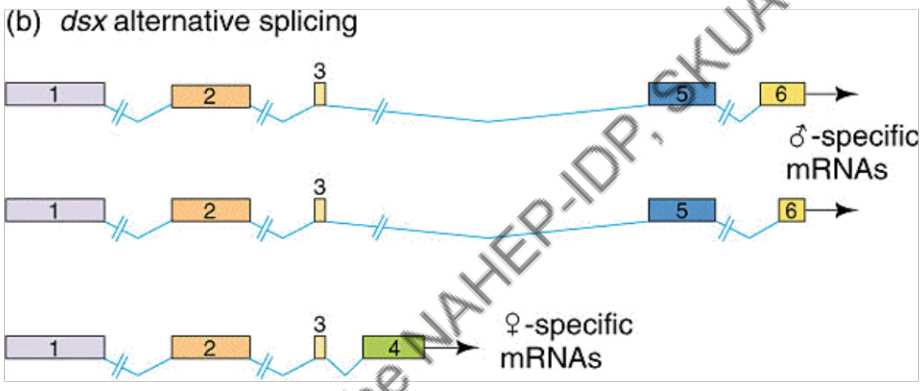
Sxl goes on to catalyze the proper splicing of another mRNA encoding Tra using a similar mechanism as for regulation of its own splicing. In the absence of Sxl (in males), splicing of the *tra* mRNA results in the incorporation of a premature stop codon and no full length Tra is produced.

Tra goes on to catalyze the splicing of another mRNA encoding Dsx to a "female" specific conformation. The Dsx-F protein represses transcription of genes that encode male traits. In the absence of Tra (in males), splicing of the *dsx* mRNA results in the production of a mRNA encoding a "male" specific Dsx. The Dsx-M protein represses transcription of genes that encode female traits.





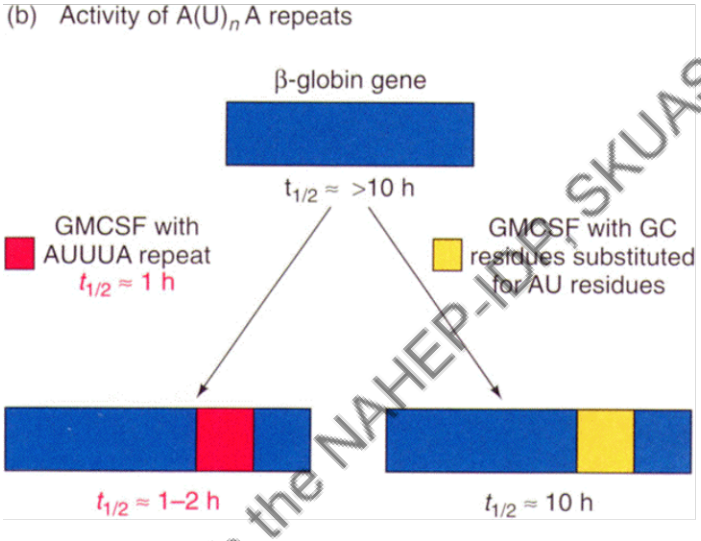




Figures 23-16, 23-17 and 23-18
(From: AN INTRODUCTION TO GENETIC ANALYSIS 6/E BY Griffiths, Miller, Suzuki, Leontin, Gelbart 1996 by W. H. Freeman and Company. Used with permission.)

B. mRNA stability

1. Gene expression can be regulated by altering the stability of the mRNA.
2. mRNAs with short or no poly(A) tails are rapidly degraded.
3. Specific sequences in the mRNA that may affect stability. For example, a sequence of AUUUA repeats at 3' end reduces the mRNA stability. Insertion of this sequence from a gene encoding an unstable mRNA (GMCSF gene for granulocyte-monocyte stimulating factor) into a gene that encodes a stable mRNA (□-globin) decreases the stability of the □-globin mRNA as measured by mRNA half life.



C. Protein degradation

IV. Gene redundancy

For some genes whose products are needed in high amounts, there are multiple copies of the genes in the chromosome.

A. Hundreds of copies of rRNA genes in *Xenopus* (frogs).

1. Nucleolar organizer (region of chromosome that physically associates with the nucleolus) has 450 copies of the 18S and 28S rRNA genes.
2. 20,000 copies of the 5S rRNA genes (not associated with the NO)

B. Several hundred copies of the histone genes in sea urchin chromosomes

CHAPTER 2

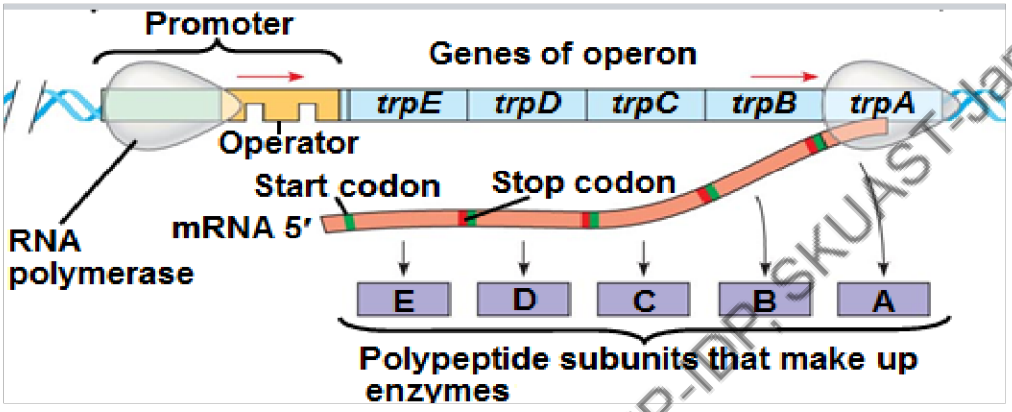
OPERON

Operon- A group of prokaryotic genes with a related function that are often grouped and transcribed together. In addition, the operon has only one promoter region for the entire operon. An operon is composed of the following:

Structural genes- genes that are related and used in a biochemical pathway.

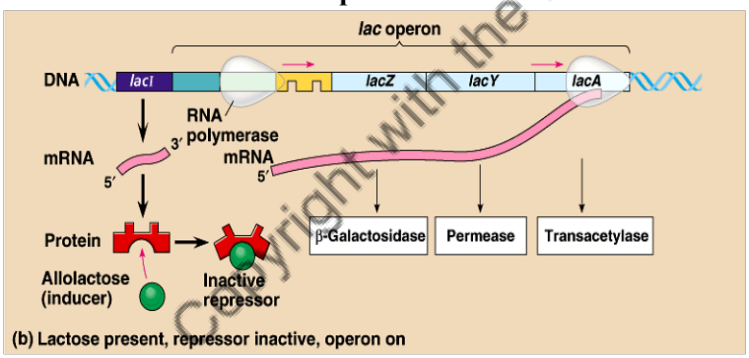
Promoter-The nucleotide sequence that can bind with RNA polymerase to start transcription. This sequence also contains the operator region.

Operator-The nucleotide sequence that can bind with repressor protein to inhibit transcription.

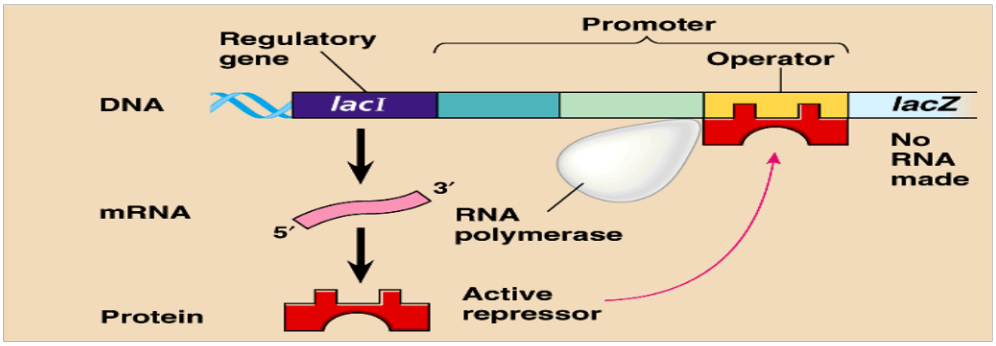


Regulator gene- This gene produces a protein called a repressor that can inhibit the transcription of an operon by attaching to the operator.

Lactose and inducible lac operon

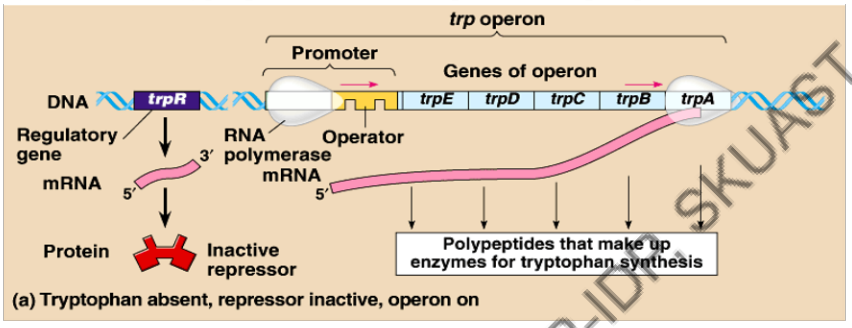


Absence of Lactose and *lac* operon

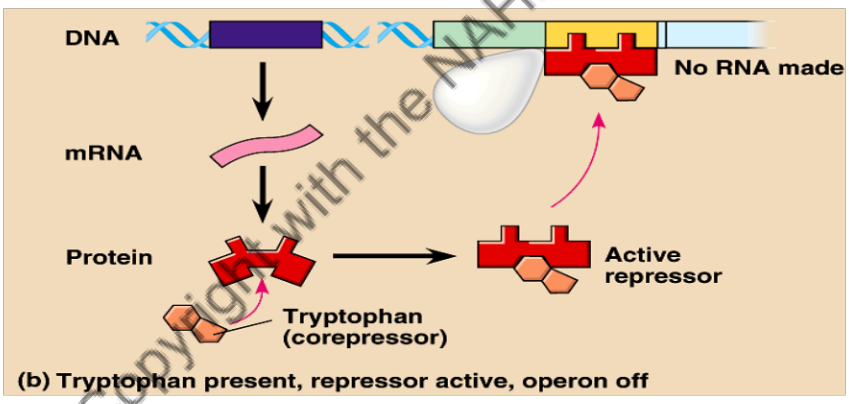


If no lactose or allolactose is present, the repressor protein is active, binding to the operator site. This prohibits the RNA polymerase from transcribing the operon.

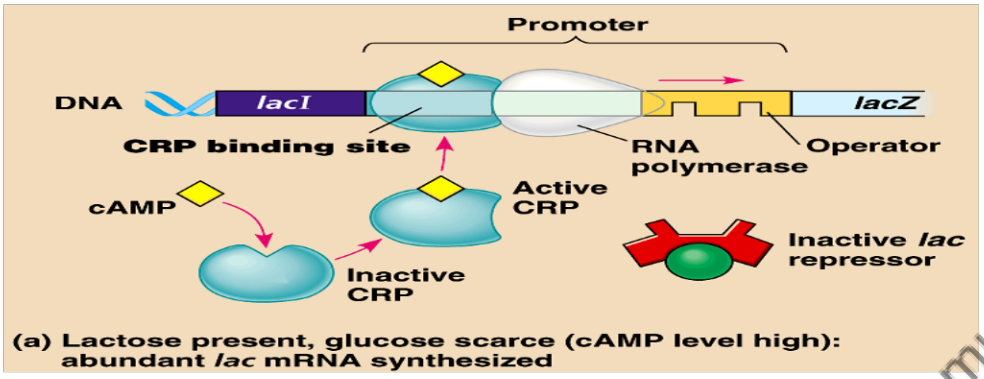
Synthesis of Tryptophan and the Repressible *trp* Operon



Tryptophan Present and the Repressible *trp* Operon



POSITIVE GENE REGULATION



CHAPTER 3

TRANSLATION

Translation is the process by which ribosomes convert the information carried by messenger RNA(mRNA) to the synthesis of proteins. It can also be defined as the process in which sequence of nucleotides in mRNA is translated into the sequence of amino acids. It can also be defined as the translation of the language available in the form of mRNA into the language of proteins.

mRNA (translation) → Proteins

Translation involves the transport of amino acids from the intercellular pool to the ribosomes where they are assembled into proteins elsewhere in the cytoplasm. Transfer of amino acids to the ribosome surface is accomplished by mRNA.

Requirement of protein synthesis:

Various molecules are required for the process of protein synthesis. They are:

1)D.N.A - D.N.A is a double helical prime molecule that determines the kind of protein needed to be synthesized. The protein synthesis is initiated, guided and regulated by DNA molecule.

2)Messenger R.N.A(mRNA)- mRNA is a single-stranded molecule that carries information from D.N.A to the cytoplasm for protein synthesis. The information stored in the form of a base sequence of mRNA is complementary to the base sequence present on template D.N.A.

3)Transfer R.N.A(tRNA)-tRNA helps in protein synthesis by picking up activated amino acids from the amino acid pool and transporting them to the ribosomes where it recognizes a specific triplet codon of mRNA. Each amino acid is carried by a specific tRNA as the lowermost segment of tRNA has three base sequences anticodon loop which are complementary to the triplet codons of mRNA.

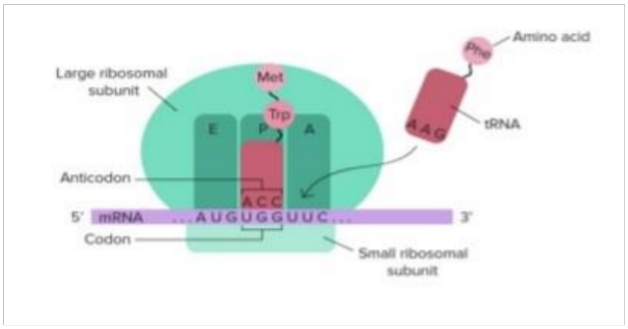
4)Ribosomes-These are the sites of protein synthesis and are found in the cytoplasm, They contain a number of enzymes responsible for the formation of the polypeptide chain. Each ribosome has two subunits- a larger subunit and a smaller subunit.

Larger subunit has two sites:

I) Aminoacyl site (A site) or acceptor site

II) Peptide site (P site) or donor site

5)Amino acids-These are the building blocks of a polypeptide chain or protein. There are 20 types of



amino acids which occur in cytoplasm forming an amino acid pool. These amino acids are assembled in polypeptide chain to form a protein.

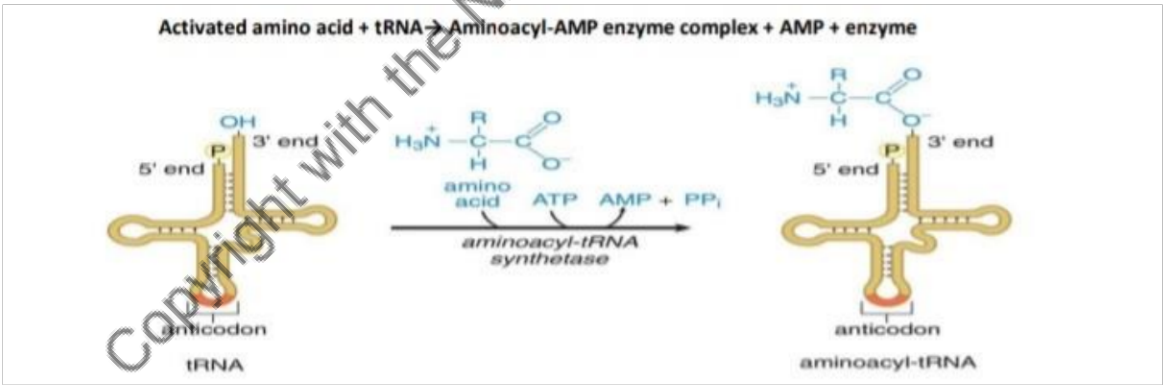
6)Enzymes-A number of enzymes are responsible for the process of transcription. Aminoacyl-tRNA synthetase is one of them.

The process of translation is much more complex than that of transcription. It involves the following steps:

1)Binding of mRNA to ribosomes: During transcription, DNA molecule synthesizes three types of RNAs inside the nucleus. Then, these RNAs migrate into the cytoplasm through the nuclear pore. Out of these RNAs, mRNA carries the genetic information and it is joined to the ribosomal subunits by the initiation codon 'AUG' found on its 5'end. This union forms mRNA ribosomal complex. [Many ribosomes lined up on a chain is known as poly-ribosomes.

2)Activation of amino acid: Amino acids are found in the amino acid pool in the cytoplasm in an inactive form. To form polypeptide chain, the amino acids must be activated before they are joined to the tRNA. The enzyme aminoacyl synthetase activates the amino acid in the presence of ATP and Mg . Amino acid + Aminoacyl-Synthetase + ATP→ Aminoacyl-AMP enzyme complex(Activated amino acid) + Ppi

3)Attachment of activated amino acid with tRNA: The activated amino acids are joined to the 3' end of the tRNA and form amino-acyl-tRNA complex. Activated amino acid + tRNA→ Aminoacyl-AMP enzyme complex + AMP + enzyme. There are more than 20 different enzymes and 20 tRNA molecules in the cell. So a specific amino acid attaches to a specific aminoacyl-tRNA molecule to form chained tRNA. This chain of tRNA serves as an adaptor molecule for decoding the information to mRNA till it reaches the last codon. As one ribosome moves along mRNA, the initiating part of mRNA becomes free. In this site, new ribosomes get lined up to form polyribosome.



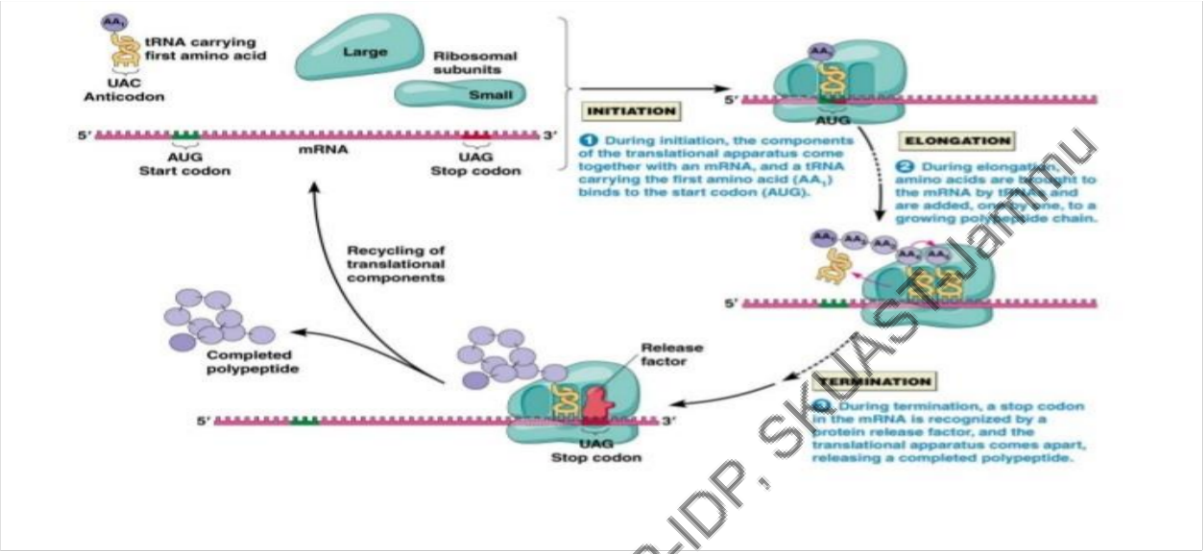
4)Initiation of polypeptide chain: Each mRNA molecule has initiation codon AUGm which signals the beginning of polypeptide chain. In this process,mRNA first binds to the subunits of ribosomes. The AUG codon lies near 'P' peptidyl site of the larger subunit of the ribosome. This codon codes for amino acid methionine. This means, activated methionine bearing tRNA has anticodon UAC. The second codon on mRNA leads close to 'A' site of the ribosome. Then, the 2 aminoacyl-tRNA complex with anticodon bonds with the 2 codon of mRNA and occupies the 'A'-site of the ribosome.

5)Elongation of polypeptide chain: The elongation begins with the formation of the peptide bond (-CO-NH-) between the amino acids present in 'P' and 'A' sites of the ribosomes. This is catalyzed by enzyme peptide synthetase. It causes the transfer of amino acid from 'A' site to 'P' site and formation of amino acid chain on 'A' site and releases the tRNA from P-site.During the elongation of the polypeptide chain,

ribosomes move along mRNA till it reaches the last codon. As one ribosome moves along mRNA, the initiating point of mRNA becomes free. In this site, new ribosome gets lined up to form polyribosomes.

6) Termination of polypeptide chain: When the ribosome reaches the end of mRNA strand (3' end) the synthesis of the polypeptide chain is completed. It is signaled by the termination codon UAA, UGA, and UAG.

During this process:



CHAPTER 4

GENETIC CODE

The genetic code is the code which body uses to convert the instructions contained in DNA. It is typically discussed using the "codons" found in mRNA, as mRNA is the messenger that carries information from the DNA to the site of protein synthesis. Therefore, Genetic code is the set of rules used by living cells to translate information encoded within genetic material (DNA or mRNA sequence) into proteins. Translation is accomplished by the ribosome, which links amino acids in an order specified by messenger RNA (mRNA), using transfer RNA (tRNA) molecules to carry amino acids and to read the mRNA three nucleotides at a time.

General features of genetic code:

1. Linear: Genetic code is always written in linear form using ribonucleotide bases that compose mRNA molecules as letters. DNA is a linear polynucleotide chain and a protein is a linear polypeptide chain. The sequence of amino acids in a polypeptide chain corresponds to the sequence of nucleotide bases in the gene (DNA) that codes for it. Change in a specific codon in DNA produces a change of amino acid in the corresponding position in the polypeptide. The gene and the polypeptide it codes for are said to be co-linear.

2. Triplet: The Crick, Brenner experiment first demonstrated that codons consist of three DNA bases. Marshall Nirenberg and Heinrich J. Matthaei was the first to reveal the nature of a codon in 1961. Each word within the mRNA contains three ribonucleotide letters. Each group of three ribonucleotide called a codon specifies one amino acid and hence the code is triplet. The code defines how sequences of nucleotide triplets, called codons, specify which amino acid will be added next during protein synthesis.

3. Universal: The same code is used throughout all the life forms being universal in nature. The universal genetic code is a common language for almost all organisms to translate nucleotide sequences of deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) to amino acid sequences of proteins eg UUU codes phenylalanine in every organisms..

4. Degenerate: Degeneracy is the redundancy of the genetic code which means given amino acid can be specified by more than one triplet codon. Ex. codons GAA and GAG both specify glutamic acid (redundancy). The codons encoding one amino acid may differ in any of their three positions.

5. Nonoverlapping and commaless: The genetic code is composed of nucleotide triplets. In other words, three nucleotides in mRNA (a codon) specify one amino acid in a protein. The code is non-overlapping. This means that successive triplets are read in order. Each nucleotide is part of only one triplet codon. A non-overlapping code means that a base in a mRNA is not used for different codons and once read for a amino acid will not participate for another amino acid. In Figure it has been shown nine bases code for not more than three amino acids. A comma less code means that no nucleotide or comma (or punctuation) is present in between two codons.

The most common start codon is AUG, which is read as methionine or, in bacteria, as formylmethionine. Alternative start codons depending on the organism include "GUG" or "UUG"; these codons normally represent valine and leucine, respectively, but as start codons they are translated as methionine or formylmethionine. The three stop codons have names: UAG is amber, UGA is opal (sometimes also called umber), and UAA is ochre. Stop codons are also called "termination" or "nonsense" codons as they encode no amino acid. They signal release of the nascent polypeptide from the ribosome because no cognate tRNA has anticodons complementary to these stop signals, allowing a release factor to bind to the ribosome instead.

6. Unambiguous: There is no ambiguity in the genetic code. A given codon always codes for a particular amino acid, wherever it is present. Each codon specifies one amino acid only. For instance UAU codes for only Tyrosine and none other amino acid.

Translation is the communication of the meaning of a source-language text by means of an equivalent target-language text. The English language draws a terminological distinction (which does not exist in every language) between *translating* (a written text) and *interpreting* (oral or signed communication between users of different languages); under this distinction, translation can begin only after the appearance of writing within a language community.

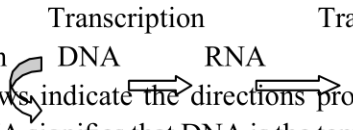
A translator always risks inadvertently introducing source-language words, grammar, or syntax into the target-language rendering. On the other hand, such "spill-overs" have sometimes imported useful source-language calques and loanwords that have enriched target languages. Translators, including early translators of sacred texts, have helped shape the very languages into which they have translated.

Because of the laboriousness of the translation process, since the 1940s efforts have been made, with varying degrees of success, to automate translation or to mechanically aid the human translator. More recently, the rise of the Internet has fostered a world-wide market for translation services and has facilitated "language localisation".

CHAPTER 5

THE CENTRAL DOGMA

By the fall of 1953, the working hypothesis was adopted that chromosomal DNA functions as the template for RNA molecules, which subsequently move to the cytoplasm, where they determine the arrangement of amino acids within proteins. In 1956 Francis Crick referred to this pathway for the flow of genetic information as the central dogma:



Here the arrows indicate the directions proposed for the transfer of genetic information. The arrow encircling DNA signifies that DNA is the template for its self-replication. The arrow between DNA and RNA indicates that RNA synthesis (called transcription) is directed by a DNA template. Correspondingly, the synthesis of proteins (called translation) is directed by an RNA template. Most importantly, the last two arrows were presented as unidirectional; that is, RNA sequences are never determined by protein templates nor was DNA then imagined ever to be made on RNA templates. The idea that proteins never serve as templates for RNA has stood the test of time. RNA templates sometimes do act as templates for DNA chains of complementary sequence. Such reversals of the normal flow of information are very rare events compared with the enormous number of RNA molecules made on DNA templates. Thus, the Central dogma as originally proclaimed more than 50 years ago still remains essentially valid.

The Adaptor Hypothesis of Crick

At first it seemed simplest to believe that the RNA templates forprotein synthesis were folded upto create cavities on their outer surfaces specific for the 20 different amino acids. The cavities would be so shaped that only one given amino acid would fit,and in this way RNA would provide the information to order amino acids during protein synthesis.By 1955,however,Crick became disenchanted with this conventional wisdom, arguing that it would never work. In the first place, the specific chemical groups on the four bases of RNA (A, U, G, and C) should mostly interact with water- soluble groups. Yet, the specific side groups of many amino acids (e.g., leucine, valine, and phenylalanine) strongly prefer interactions with water-insoluble (hydrophobic) groups. In the second place, even if somehow RNA could be folded so as to display some hydrophobic surfaces, it seemed at the time unlikely that an RNA template would be used to discriminate accurately between chemically very similar amino acids like glycine and alanine or valine and isoleucine, both pairs differing only by the presence of single methyl (CH_3) groups. Crick thus proposed that prior to incorporation into proteins,amino acids are first attached to specific adaptor molecules,which in turn possess unique surfaces that can bind specifically to bases on the RNA templates.

Discovery of Messenger RNA (mRNA)

Cells infected with phage T4 provided the ideal system to find the true template.Following infection by this virus, cells stop synthesizing E.coli RNA; the only RNA synthesized is transcribed off the T4 DNA. Most strikingly,not only does T4 RNA have a base composition very similar to T4 DNA, but it does not bind to the ribosomal proteins that normally associate with rRNA to form ribosome Instead, after first

attaching to previously existing ribosomes, T4 RNA moves across their surface to bring its bases into positions where they can bind to the appropriate tRNA–amino acid precursors for protein synthesis (Fig. 2-15). In so acting, T4 RNA orders the amino acids and is thus the long-sought- for RNA template for protein synthesis. Because it carries the information from DNA to the ribosomal sites of protein synthesis, it is called messenger RNA(mRNA).The observation of T4 RNA binding to E.coli ribosomes, first made in the spring of 1960,was soon followed with evidence for a separate messenger class of RNA within uninfected E. coli cells, thereby definitively ruling out a template role for any rRNA. Instead, the rRNA components of ribosomes, together with some 50 different ribosomal proteins that bind to them, serve as the factories for protein synthesis, functioning to bring the tRNA–amino acid precursors into positions where they can read off the information provided by the mRNA templates. Only a few percent of total cellular RNA is mRNA. This RNA shows the expected large variations in length and nucleotide composition required to encode the many different proteins found in a given cell. Hence, it is easy to understand why mRNA was first overlooked. Because only a small segment of mRNA is attached at a given moment to a ribosome, a single mRNA molecule can simultaneously be read by several ribosomes. Most ribosomes are found as parts of polyribosomes (groups of ribosomes translating the same mRNA), which can include more than 50 members (Fig.2-16).

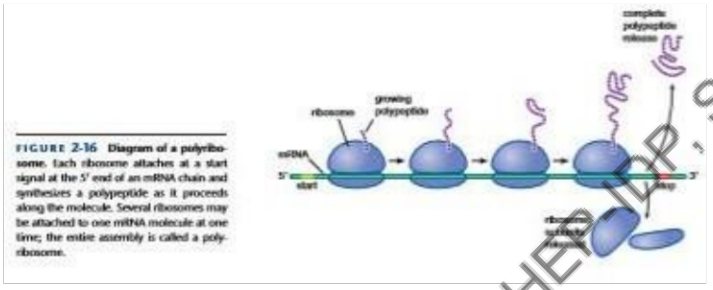


FIGURE 2-16 Diagram of a polyribosome. Each ribosome attaches at a start signal at the 5′ end of an mRNA chain and synthesizes a polypeptide as it proceeds along the molecule. Several ribosomes may be attached to one mRNA molecule at one time; the entire assembly is called a polyribosome.

CHAPTER 6

TRANSPOSABLE ELEMENTS (TES)

INTRODUCTION

Transposable Elements (TES) are defined as DNA sequences that are able to move from one location to another in the genome. TEs have been identified in all organisms, prokaryotic and eukaryotic, and can occupy a high proportion of a species genome. The mobilization of TEs is termed transposition or retro transposition, depending on the nature of the intermediate used for mobilization.

TYPES OF TRANSPOSABLE ELEMENTS:

There are three types of transposable elements described in prokaryotes.

1. Insertion Sequences (IS elements)
2. Composite transposons
3. Tn3 elements

1. IS elements:

a) IS elements are relatively small transposable elements that range in size from 760 to lessthan 2,500 base pairs (bp). They can insert at many different sites in bacterial and viral chromosomes and plasmids and episomes, and they contain genes whose products are involved in promoting and regulating transposition. One of the gene is a *transposase* that functions in excision of the element from a chromosome, plasmid, or episome.

b) IS elements typically generate unstable mutants that revert to wild-type at a detectable frequency. For that reason, IS elements originally were called "mutable" genes.

c) All IS elements contain *inverted terminal repeats* that range in size (length) from 9 to 40 base pairs. At the site of integration there invariably is a target site duplication of from 2-13 base pairs.

2. Composite transposons (denoted by symbol *Tn)*:

a) *Tn* elements stem from two IS elements that insert near one other. The regions (sequences) between the two elements can be "mobilized" by the joint action of the two IS elements. This is of significance in that many *Tn* elements possess genes that confer resistance to antibiotics between the two IS elements.

b) *Tn* transposition is regulated by a "repressor" that appears to exist to keep the elements somewhat quiescent.

(c) *Tn3* elements are simply large transposable elements that are not generated by flanking IS elements (as in *Tn* elements). They are generally ~5,000 bp, have ~386 bp inverted repeats at both ends, and carry antibiotic resistant genes.

B. Transposable elements in eukaryotes:

They are of two types: those that have DNA as their genetic material, and those that have RNA as their genetic material.

1. DNA transposable elements: These are exemplified by the P elements in *Drosophila*.

a) P elements were discovered when it was found that certain strains of *Drosophila* exhibited an assortment of aberrant phenotypes, including elevated mutation (and reversion), chromosome breakage, and sterility. This phenomenon was termed “hybrid

dysgenesis" and turned out to be situations where transposable P elements had been induced to "jump." The phenomenon was termed "hybrid dysgenesis" because normally (within populations) the P elements are quiescent and do not "jump." When "hybrids" were made between individuals from different geographic populations, the elements "moved" and promoted the dysgenic phenotypes.

b) P elements vary in size (the largest are nearly 3,000 base pairs in length). Complete (intact) P elements possess a gene for a transposase. The number of P elements per individual varies from a few to up to 50.

c) P elements characteristically have a 31 bp inverted repeat at both ends and an 8 bp target-site duplication.

2. Retrotransposons are transposable elements that have RNA as their genetic material. There are two types: retrovirus-like elements and retroposons.

a) Retrovirus-like elements:

(i) The basic structure of retrovirus-like elements is a central coding region of two genes flanked by long terminal repeats (LTRs) that are oriented in the *same* direction and bounded by short inverted repeats. The LTRs play a role in integration of the element into the host genome. The two genes are homologous to two genes in retroviruses and encode a structural protein of the virus capsule and a reverse transcriptase/integrase enzyme.

(ii) Minimal (active) retroviruses carry a third gene that codes for a protein of the virus envelope. Active retroviruses are capable of exiting cells and infecting other cells.

(iii) Transposition involves transcription (RNA synthesis) of the DNA sequence integrated in the chromosome, reverse transcription of the RNA, synthesis of a double-stranded DNA from the RNA, and insertion into a new chromosomal location.

b) Retroposons:

(i) These are elements that move through an RNA intermediary but do *not* possess direct or inverted repeats at their termini. They possess instead a string of A=T base pairs at one end (of the DNA), and presumably represent a copy from reverse transcription of the poly-A tail of the mature RNA transcript.

(ii) Several retroposons are known in *Drosophila*. They appear to occur non- randomly at the ends of chromosomes and to function in replicating telomeres.

(iii) Some LINE sequences in mammals are retroposons, and the LlNE-1 retroposon is the only transposable element thus far documented in humans.

CHAPTER 7

EPIGENETIC CONTROL OF GENE EXPRESSION

1. Introduction

Plants being sessile organisms are being constantly challenged by various biotic and abiotic stresses. In order to adapt themselves to the changing environments they need constant changes at molecular level. These efficient and effective controls are provided by epigenetic regulations which improve the survivability of plants by increasing their tolerance toward stress. It is now evident that heritable phenotypic variation does not need to be based on DNA sequence polymorphism. These epigenetic regulations involve different chemical modifications at molecular level that influence gene expression. Epigenetic as defined by Conrad Waddington, is "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence".

Epigenetic refers mainly to the changes that do not relate to the changes in DNA sequence but to chemical modification that can be inherited from one generation to the next. Three types of epigenetic regulatory mechanisms viz. DNA methylation, histone modification and RNA interference (RNAi) are exploited by plants in order to survive adverse conditions.

DNA methylation is a chemical modification, catalyzed by cytosine methyltransferases which involves addition of a methyl group in a DNA sequence onto the cytosine residue in a sequence specific manner, primarily within CpG dinucleotide. The added methyl group provides platform for attachment of various protein complexes that modifies the histone scaffolds resulting in altered gene expression.

In eukaryotic nuclei DNA is organized in the form of nucleosome where it is wrapped around by Histone proteins. Histone comprise a family of highly conserved globular proteins whose N-terminal tails are exposed on the surface of the nucleosome octamer for chemical modifications. Histone modifications include acetylation, methylation, ubiquitination and phosphorylation of histone proteins. Acetylation and phosphorylation are mostly associated with induced gene expression while on the other hand modifications like biotinylation represses gene expression. Such modifications not only impinge on DNA accessibility, but also on the recruitment of specific proteins involved in several processes, including transcription, DNA replication and repair. Histone proteins are not only modified, but can also be replaced by histone variants with different physical properties, or released, in order to allow gene expression

2. Different types of epigenetic modifications

DNA methylation modification

DNA methylation arises as a result of addition of a methyl group to the nitrogenous base in the DNA strand in a sequence specific manner. DNA methylation occurs at the fifth carbon position of a cytosine ring. Methylation of cytosine leads to the generation of 5-methyl cytosine. On the basis of the target sequence, methylation is classified either as symmetrical or asymmetrical methylation. CG and CHG methylation are termed as symmetrical and CHH methylation as asymmetrical. Plants methylate only some genes and this methylation is usually restricted to CGs located within the gene body while Transposable Element sequences tend to be methylated at most of their CG, CHG, and CHH sites.

Thus DNA methylation results into following (i) methylcytosines in the gene body play an important role in regulating the gene expression and (ii) methylcytosines in repetitive sequences (transposable elements), are thought to prevent repetitive sequences from compromising normal genome function. Increased methylation of genomic DNA down regulates gene expression. Down regulated gene expression enable the plants to conserve energy for the sake of biotic or abiotic stress. In contrast, the reduction in methylation of resistance-related genes favours chromatin activation and the expression of novel genes, which provides long-term or permanent resistance for stress.

Histone modifications in plants

In addition to DNA methylation, histone N-terminal tail modifications constitute an attractive area in epigenetics. Plants contain several histone variants and enzymes that posttranslationally modify histones and influence gene regulation. Application of chromatin immunoprecipitation followed by deep sequencing has given insight into the genome-wide distribution of histone variants and histones bearing different posttranslational modifications Histone proteins are wrapped around DNA and forms a highly compact structure called nucleosome. Nucleosomes are composed of histone octamers that comprise two copies each of H2A, H2B, H3, and H4. A total of 147 base pair of DNA sequence is wrapped around the histone core. The N termini of histone proteins called N terminal tails undergo various chemical modifications like methylation or acetylation. Such histone modifications are associated with either gene repression or gene activation. In plants methylation and deacetylation of H3K9 and H3K27 results into gene repression whereas acetylation and methylation of H3K4 and H3K36 is associated with gene activation and thus induces gene expression. These covalent modifications in response to various environmental stresses regulates the transcription of wrapped DNA sequence by altering the packaging structure which either activates the DNA for the transcription or makes the structure more condensed so that transcription machinery is unable to reach it

Histone acetylation/deacetylation:

Addition of acetyl group to the N terminal Lysine of histones results into transcriptional activation of a DNA sequence. Acetylation of N terminal lysine causes reduction in the net positive charge of histone and as a result the electrostatic force of attraction between the negatively charged DNA and positively charged histone reduces which leads to the loosening of chromatin and transcriptional activation of DNA. The addition of acetyl group to Lysine is catalyzed by histone acetyltransferases (HATs).

Histone methylation: Arginine and Lysine amino acids in histone proteins undergo methylation. Different arginine and lysine residues in different histones undergo different types of methylation (R3 of H2A, R3, K20 of H4 and K4, K9, K27, K36, R2, and R17 of H3 etc.) and these residues can be mono, di or tri methylated. Methylation affects the hydrophobicity of the histone and hence may change histone DNA interactions or may create binding site for various proteins which restricts the binding of transcription machinery and prevents transcription. Histone lysine methyltransferases (HKMT) and protein arginine methyltransferases (PRMT) catalyze methylation of lysine and arginine residues respectively

miRNA directed DNA methylationRNA directed DNA methylation (RdDM) is *de novo* cytosine methylation primarily within the region of RNA-DNA sequence identity. Although this pathway can methylate all sequence contexts, but it specifically methylates CHH sequences. The reason for this is that symmetrical methylation is maintained by maintenance methylation each time the DNA is replicated whereas the CHH methylation at many silenced loci is dependent on RNA-guided *de novo* methylation.The 24-nt siRNAs are generated by DNA dependent RNA polymerase Pol IV enzyme, in association with RNA-dependent RNA polymerase 2 (RDR2), and processed by dicer-like 3 (DCL3).One strand of the resulting 24-nt dsRNA fragments is loaded onto argonaute 4 (AGO4) leading to generation of a silencing effector

complex. DNA methylation at sites having sequence homology to the siRNA is dependent on, Pol V, which is a DNA dependent RNA polymerase that transcribes non-coding RNAs. Transcription of Pol V is facilitated by a chromatin remodeling protein which is defective in RNA-directed DNA methylation 1 (DRD1). KOW domain transcription factor1 (KTF1) which is an adaptor protein, mediates binding of AGO4 and AGO4-bound siRNAs onto the transcripts generated by Pol V forming a silencing effector. This effector acts as signal for DRM2 to introduce methylation at target sites. Development of stress tolerant crop has successfully been achieved by the use of RNAi technology.

Summary

A course remedial cum jrf claases on bioinformatics was organized in the month of March, 2022 under the IDP project 'Strengthning institutional capacities for delivering competent skilled professionals at SKUAST- Jammu sponsored under NAHEP-ICAR

A total of 13 students participated in the course. As bioinformatics becomes increasingly central to research in the molecular life sciences, the need to train the life science students and others working in the area of lifescience to make the most of bioinformatics resources is growing. Here, I tried to deliver and talk about the bioinformaics (introduction; basic to advance) and its key applications in the area of Lifescience and medical science. The course content was in the form on power point presentation and handouts. In total 6 lectures were delivered on the different topics of Bioinformatics in the form of power point presentations.

The details of lectures are

1	Web resources for Bioinformatics
2	Biological Databases
3	Introcuction to BLAST
4	Multiple Sequence Alignment and Tools
5	Phylogenetic Analysis
6	Algorithm (Bioinformatics)

1. **Web resources for Bioinformatics:** The vast amount of information generated has made computational analysis critical and has increased demand for skilled bioinformaticians. There are thousands of bioinformatics and genomics resources that are free and publicly accessible. However, trying to find the right resource and to learn how to use the complex features and functions can be difficult. Under this topic, I focussed on the ways that we can quickly find and effectively learn how to use resources. It included a tour of examples, resources, organized by categories such as Algorithms and Analysis tools, expression resources, genome browsers, Literature and text mining resources.

Bioinformatics is a new and emerging branch of Biotechnology that has come up very recently. It mainly involves the use of software to utilize information from vast biological database that is developed by experienced Biotechnologists. Gene sequencing is a part of Bioinformatics wherein a lot of data related to biotechnology is processed. This brings biotechnology within the ambit of information technology, and hence the label, Bioinformatics. In fact Bioinformatics is a new discipline that involves molecular biology and computer science. Presently genomic research, sequencing of human genome and advances in disease related issues have required and helped in developing this at fast level. So we can say that in Bioinformatics, computers are required to store, retrieve, analyze or predict the composition or the structure of biomolecules. It is the fascinating hybrid of computer science and biology

The National Center for Biotechnology Information (NCBI 2001) defines Bioinformatics as: "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline...There are three important sub-disciplines within Bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information." Bioinformatics is a particularly international subject, with a notably high degree of information sharing among researchers in different countries. It is also known as computational biology e.g., USC Computational Biology, NCSA Computational Biology.

USE OF INTERNET IN BIOINFORMATICS

When we talk about sources of biological information and computers for providing it, we can not ignore the role and impact of information superhighway i.e., Internet. Internet is the most potential tool of this information age and it is serving as a platform for Bioinformatics tool. It provides the opportunity to search that information, which was available only by reaching to the information centre.

Areas of Services The Internet provides various facilities for Bioinformatics, such as;

- Bioinformatics research
- Courses
- Resources
- Biological databases
- Construction tools
- Software resources
- WWW search tools
- Advanced topics in Bioinformatics
- Scientific databases
- Electronic journals

2. **Biological Databases:** Biological databases emerged as a response to the huge data generated by low-cost DNA sequencing technologies. One of the first databases to emerge was GenBank, which is a collection of all available protein and DNA sequences. It is maintained by the National Institutes of Health (NIH) and the National Center for Biotechnology Information (NCBI). GenBank paved the way for the Human

Genome Project (HGP). The HGP allowed complete sequencing and reading of the genetic blueprint. The data stored in biological databases is organized for optimal analysis and consists of two types: raw and curated (or annotated). Biological databases are complex, heterogeneous, dynamic, and yet inconsistent.

Bioinformatics is characterized by an abundance of data stored in very large databases. Local databases with capacities measured in the tens of terabytes are common. As such, fluency in data warehousing, data dictionaries, database design, archiving, and knowledge management techniques are mandatory for the design and maintenance of these systems. Most of the large biology databases are based on traditional relational databases architectures; whereas others, especially systems dealing with images and other multimedia, are based on object-oriented designs. In recent years, biological databases have greatly developed, and became a part of the biologist's everyday toolbox. There are several reasons to search databases, for instance:

When obtaining a new DNA sequence, one needs to know whether it has already been deposited in the databanks fully or partially, or whether they contain any homologous sequences(sequences which are descended from a common ancestor).

- Some of the databases contain annotation which has already been added to a specific sequence. Finding annotation for the searched sequence or its homologous sequences can facilitate its research.
- Find similar non-coding DNA stretches in the database. For instance repeat elements or regulatory sequences.
- Other uses for specific purpose, like locating false priming sites for a set of PCR oligonucleotides.
- Search for homologous proteins - proteins similar in their sequence and therefore also in their presumed folding or structure or function.

Primary sequence databases: In the early 1980's, several primary database projects evolved in different parts of the world. There are two main classes of databases: DNA (nucleotide) databases and protein databases. The primary sequence databases have grown tremendously over the years.

DNA (nucleotide) Databases		**Protein Databases**	
EMBL	UK	Swiss-Prot	Swiss
GenBank	US	PIR	US
DDBJ	Japan	MIPS	Germany
		TrEMBL	Swiss
		GenPept NRL 3D	US
			US

Today they suffer from several problems, unpredicted in early years (when their sizes were much smaller):

- Databases are regulated by users rather than by a central body (except for SwissProt).
- Only the owner of the data can change it.
- Sequences are not up to date.
- Large degree of redundancy in databases and between databases.
- Lack of standard for fields or annotation.

DNA Databases (Nucleotide Sequences):
The growth rate of DNA databases is much higher than that of the protein databases. This is because most of the DNA is not coding for proteins and because DNA sequencing is the most prominent source of database entries. The large DNA databases are:Genbank (US), EMBL (Europe - UK), DDBJ (Japan). These databases are quite similar regarding their contents and are updating one another periodically. This was is a result of the International Nucleotide Sequence Database Collaboration.

EMBL:
EMBL is a DNA sequence database from European Bioinformatics Institute (EBI). EMBL includes sequences from direct submissions, from genome sequencing projects, scientific literature and patent applications. EMBL supports several retrieval tools: SRS for text based retrieval and Blast and FastA for sequence based retrieval.

GenBank:
GenBank is a DNA sequence database from National Center Biotechnology Information (NCBI). It incorporates sequences from publicly available sources (direct submission and large-scale sequencing). DDBJ (DNA Data Bank of Japan): DNA Data Bank of Japan began DNA data bank activities in earnest in 1986 at the National Institute of Genetics (NIG). DDBJ has been functioning as the international nucleotide sequence database in collaboration with EBI/EMBL and NCBI/GenBank.

Protein Databases (Amino Acid Sequence)
PIR - (International Protein Sequence Database):
PIR - The Protein Sequence Database was developed in the early 1960's. It is located at the National Biomedical Research Foundation (NBRF). Since 1988 it has been maintained by PIR-International

PIR is split into four distinct sections that differ in quality of the data and the level of annotation: PIR1 - fully classified and annotated entries.
PIR2 - preliminary entries, not thoroughly reviewed.
PIR3 - unverified entries, not reviewed.
PIR4 - conceptual translations.

Swiss-Prot:
Swiss-Prot was established in 1986. It is maintained collaboratively by SIB (Swiss Institute of Bioinformatics) and EBI/EMBL. Provides high-level annotations, including description of protein function, structure of protein domains, post-translational, modifications, variants, etc. It aims to be minimally redundant. Swiss-Prot is linked to many other resources, including other sequence databases.

TrEMBL:
Translated EMBL: Translated EMBL was created in 1996 as a computer annotated supplement to Swiss-Prot. It contains translations of all coding sequences in the EMBL nucleotide sequence database. SP-TrEMBL contains entries that will be incorporated into Swiss-Prot REM TrEMBL contains entries that are not destined to be included in Swiss-Prot, (for example, T-cell receptors, patented sequences). The entries in REM-TrEMBL have no accession number.

GenPept:
GenPept is a supplement to the GenBank nucleotide sequence database. Its entries are translation of coding regions in GenBank entries. They contain minimal annotation, primarily extracted from the corresponding GenBank entries. For the complete annotations, one must refer to the GenBank entry or entries referenced by the accession number(s) in the GenPept entry.

NRL 3D:
NRL 3D is produced and maintained by PIR. It contains sequences extracted from the Protein DataBank (PDB). The entries include secondary structure, active site, binding site and modified site annotations, details of experimental method, resolution, Rfactor, etc. NRL 3D makes the sequence data in the PDB available for both text based and sequencebased searching. It also provides cross-reference information for use with the other PIR Protein Sequence Databases.

Summary of protein sequence databases
PIR(1-4) - comprehensive, poor quality of annotation (even in PIR1).
Swiss-Prot - poor sequence coverage, highly structured, excellent annotation.
GenPept- most comprehensive, poor quality of annotation.
NRL 3D - least comprehensive but is directly relating to structural information. When searching for a protein sequence, it is recommended to search all databases.

Secondary Databases: There are various databases containing secondary structure information. Each has its advantages and disadvantages, so it is advisable to try more than one database when searching. This section will show some popular databases.

Prosite:
The Prosite database is based on SwissPort and thus is very well annotated, but small. Characterization of protein families is done by the single most conserved motif observed in a multiple sequence alignment of known homologous. These conserved motifs usually relate to biological functions such as active sites or binding sites. The search in Prosite does not require an exact match in structure. Prosite enables searches using complex patterns. It is possible to search textually using regular expressions for names of known proteins, etc. It is also possible to scan a protein sequence using prosite for structural pattern matches. The database is well cross-linked to SwissProt and TrEMBL.

FingerPrints:
Unlike Prosite, FingerPrints has an improved diagnostic reliability which is achieved by using more than one conserved structural motif to characterize a protein family. With FingerPrints, many motifs are encoded using ungapped and unweighed local alignments. The input to FingerPrints is a small multiple alignment, which has some conserved motifs. These motifs are searched for in the database, and only sequences that match all the motifs are considered for further analysis. With the new alignment, the database is searched for more sequences until no further complete fingerprint matches can be identified. These final aligned motifs constitute the refined fingerprint that is entered into the database.

Blocks:
Blocks uses multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. Block Searcher ,Get Blocks and Block Maker are aids to detection and verification of protein

sequence homology. They compare a protein or DNA sequence to a database of protein blocks, retrieve blocks, and create new blocks, respectively.

Profiles:

The Profiles databases use the notion of profiles to achieve a good detection of distant sequence relationships. A profile is a scoring table with multiple alignment information for the whole sequences, not just for conserved regions.

Pfam:

Pfam uses a different method for its database. High quality seed alignments are used to create Hidden Markov Models to which sequences are aligned. Pfam has to classes of alignments, according to their credibility. Pfam-a – Non-edited seed alignments which are deemed to be accurate.. Pfam-b – Alignments derived by automatic clustering of the SwissPort database. These alignments are, of course, less reliable.

3. **Introcuction to BLAST:** BLAST which is a sequence similarity search program is an excellent starting point for teaching bioinformatics to students and it has the potential to enhance a student's grasp of biomedical, biochemical, and biogeochemical concepts. Under this segment the students were introduced with the underlying concepts of the BLAST algorithm, the scores and statistics of the alignments; with illustrations using the NCBI BLAST. This segment also emphasizes the need for students to be familiarized with the basic concepts and programs of bioinformatics which is a necessity in biological sciences now-a-days because of the recent advances in high-throughput techniques for data generation and analysis.

BLAST

One of the most widely used bioinformatics software Blast was developed in 1990 and since then have been available to everyone at NCBI site. This software can be accessed by any one and can be modified according to ones need. Blast is the software in which input data of a sequence to be compared is in Fasta format and output data can be obtained in plain text, HTML or XML. Blast works on the principle of searching for localized similarities between the two sequences and after short listing the similar sequences it searches for neighborhood similarities. The software searches for high number of similar local regions and gives the result after a threshold value is reached. This process differs from earlier software in which entire sequence was searched and compared which took a lot of time. Blast is used for many purposes like DNA mapping, comparing two identical genes in different species, creating phylogenetic tree.

FASTA

Fasta program was written in 1985 for comparing protein sequences only but was later modified to conduct searches on DNA also. Fasta software uses the principle of finding the similarity between the two sequences statistically. This software matches one sequence of DNA or protein with the other by local sequence alignment method. It searches for local region for similarity and not the best match between two sequences. Since this software compares localized similarities at times it can come up with a mismatch. In a sequence Fasta takes a small part known as k-tuples where tuple can be from 1 to 6 and matches with k-tuples of other sequence and once a threshold value of matching is reached it comes up with the result. It is a program that is used to shortlist prospects of matching sequence from a large number for full comparison as it is very fast.

BLAST stands for Basic Local Alignment Search Tool. It is a local alignment algorithm-based tool that is used for aligning multiple sequences and to find similarity or dissimilarity among various species. In this article, we will explain different kinds of BLAST tools and how does BLAST algorithm works.
BLAST is a heuristic method which means that it is a dynamic programming algorithm that is faster, efficient but relatively less sensitive.

For BLAST(ing) any sequence, there is a query sequence and a target sequence/database. The query sequence is the sequence for which we want to find out the similarity and the target sequence is a sequence/database against which the query sequence is aligned. Blast returns the output in the form of hit tables that are arranged in decreasing order of matched accession number along with their titles, query coverage, sequence identity, score, and an e-value in separate columns. The reliability of the matched sequences is assessed by e-value.

BLAST has different programs to align sequences of nucleotides, proteins, etc. It consists of other multiple BLAST programs, but the basic kinds of BLAST are as follows:

blastn

It is a type of blast where the query sequence is a nucleotide and the target sequence is also a nucleotide, i.e., it is a nucleotide against a nucleotide.

blastp

Blastp is a protein-to-protein blast where the query sequence is a protein and the target sequence is also a protein.

blastx

In this type of blast, the query sequence is a nucleotide sequence and the target is a protein sequence/database. First, the nucleotide sequence is converted into its protein sequence in three reading frames, then it is searched against the protein.

tblastn

In tblastn, the query is a protein and the target is a nucleotide sequence/database. Here, the protein sequence is searched against a nucleotide database which is translated to its corresponding proteins. The translation occurs in all reading frames, but the reading frame is only for the conventional 5' to 3' site in the databases, therefore, only 3 reading frames are compared.

tblastx

It is a type of blast in which the nucleotide sequence is against the nucleotide database but at the protein level. In other words, the nucleotide query sequence and the target sequences are both translated into their corresponding protein sequences and then aligned together. Both the query and the target are translated in all 6 reading frames.

How does Blast work?

Blast is a greedy algorithm that was developed by Altschul et al. It is similar to FASTA but more efficient. As FASTA uses a ktup parameter, similarly BLAST also uses a window size for proteins and nucleotides. Both assume that good alignments contain short stretches of exact matches. BLAST is an improvisation over FASTA in the sense that it is faster, more sensitive, more statistically significant, and easy to use. There is a threshold in blast known as 'Minimal Score denoted as '*S*'. It means that whatever the match is between the query and the database is it must have a value equal to or greater than *S*.
BLAST performs the alignment in 3 basic steps:

- First, blast applies the word search in which it removes the higher complex regions and then looks for short stretches of a fixed length of the query sequence.
- Secondly, blast identifies the exact word matches from the database. Those words which have scored equal to or greater than the threshold (S) are taken for alignment. These obtained alignments are called "Hits".
- Lastly, the blast extends the alignment in both directions as an ungapped alignment that stops at the maximum score and inserts a gap.

Blast vs Fasta

Blast and Fasta are two software that are used to compare biological sequences of DNA, amino acids, proteins and nucleotides of different species and look for the similarities. These algorithms were written keeping speed in mind because as the data bank of the sequences swelled once DNA was isolated in the laboratory by the scientists in mid 1980s there raised a need to compare and find identical genes for further research at high speed. Blast is an acronym for Basic Local Alignment Search Tool and uses localized approach in comparing the two sequences. Fasta is a software known as Fast A where A stands for All because it works with the alphabet like Fast A for DNA sequencing and Fast P for protein. Both Blast and Fasta are very fast in comparing any genome database and are therefore very viable monetarily as well as in saving time.

In brief:

Blast vs Fasta

• Blast is much faster than Fasta.
• Blast is much more accurate than Fasta.
• For closely matched sequences Blast is very accurate and for dissimilar sequence Fasta is better software.
• Blast can be modified according to the need but Fasta cannot be modified.
• Blast has to use Fasta input format to get the output data.
• Blast is much more versatile and widely used than Fasta.

4. **Multiple sequence alignment (MSA)**: It is a tool used to identify the evolutionary relationships and common patterns between genes. Precisely it refers to the sequence alignment of three or more biological sequences, usually DNA, RNA or protein. Alignments are generated and analysed with computational algorithms. Dynamic and heuristic approaches are used in most MSA algorithms. One of the objectives of MSA is to detect structural or functional similarities between proteins in the comparison of another protein sequence. MSAs require advanced approaches rather than parallel alignment because the computational complexity is greater. A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a

common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. Visual depictions of the alignment as in the image at right illustrate mutation events such as point mutations (single amino acid or nucleotide changes) that appear as differing characters in a single alignment column, and insertion or deletion mutations (indels or gaps) that appear as hyphens in one or more of the sequences in the alignment. Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides.

Multiple sequence alignment also refers to the process of aligning such a sequence set. Because three or more sequences of biologically relevant length can be difficult and are almost always time-consuming to align by hand, computational algorithms are used to produce and analyze the alignments. MSAs require more sophisticated methodologies than pairwise alignment because they are more computationally complex. Most multiple sequence alignment programs use heuristic methods rather than global optimization because identifying the optimal alignment between more than a few sequences of moderate length is prohibitively computationally expensive.

ClustalW is a tool to align three or more sequences together in a computationally efficient manner

Why is ClustalW useful:
Aligning multiple sequences highlights areas of similarity which may be associated with specific features that have been more highly conserved than other regions. These regions in turn can help classify sequences.

Multiple sequence alignment is also an important step for phylogenetic analysis, which aims to model the substitutions that have occurred over evolution and derive the evolutionary relationships between sequences

What inputs does ClustalW accept?

The program accepts nucleic acid or protein sequences, in the following multiple sequence formats:

- NBRF/PIR
- EMBL/UniProt
- Pearson (FASTA)
- GDE
- ALN/ClustalW
- GCG/MSF
- RSF

The sequences can either be pasted into the web form or uploaded to the web form in a file. It is very important that each of the sequences has a unique name. If they do not, the program will fail. There must be no empty lines, white spaces or control characters between sequences or at the top of the file. This will also cause the program to fail.

What do the consensus symbols mean in the alignment?

An * (asterisk) indicates positions which have a single, fully conserved residue.
A : (colon) indicates conservation between groups of strongly similar properties
A . (period) indicates conservation between groups of weakly similar properties

Applicability:

1. Very useful in the development of PCR primers and hybridization probes
2. Great for producing annotated, publication quality, graphics and illustrations
3. Invaluable in structure/function studies through homology inference
4. Essential for building "profiles" for remote homology searching and required for molecular evolutionary phylogenetic inference programs such as those from PAUP, PHYLIP, RAXML (phylogenetic Analysis Software)

5. **Phylogenetic Analysis:** It is the study of the evolutionary development of a species or a group of organisms or a particular characteristic of an organism. In phylogenetic analysis, branching diagrams are made to represent the evolutionary history or relationship between different species, organisms, or characteristics of an organism (genes, proteins, organs, etc.) that are developed from a common ancestor. Phylogenetic analysis is important for gathering information on biological diversity, genetic classifications, as well as learning developmental events that occur during evolution.With advancements in genetic sequencing techniques, phylogenetic analysis now involves the sequence of a gene to understand the evolutionary relationships among species. DNA being the hereditary material can now be sequenced easily, rapidly, and cost-effectively, and the data obtained from genetic sequencing is very informative and specific.Also, morphological estimates can be used to infer evolutionary developments, especially in cases where genetic material is not available (fossils).

Phylogenetic Analysis

Phylogenetic Analysis is the study of evolutionary relationships. The evolutionary history inferred from phylogenetic analysis is usually depicted as branching tree like diagram that represents an estimated pedigree of the inherited relationships among the molecules (gene tree), organisms or both. It is sometimes called cladistic because the word "clade", a set of descendants from a single ancestor is derived from the Greek word for branch. However, cladistic is a particular method of hypothesizing about evolutionary relationship.

Phylogenetic sequence data usually consist of multiple sequence alignments. The purpose of phylogeny is to reconstruct the history of life and explain the present diversity of living creatures. This can be represented as genealogic tree (the tree of life). The underlying principle of phlogeny is to group living creatures according to their level of similarity. Phylogenetics is a special kind of phylogeny that relies on the comparison of equivalent genes coming from several species and finding out who is the closest relative of whom in the family. If necessary we can also apply phylogenetic methods to various genes of gene family to reconstruct the history of the gene family by the same means.

****Note these trees make sense only if you believe in evolution****

In the context of Bioinformatics analysis, there are three major reasons why we may want to use phylogenetics

1. **Determining the closest relatives of the organism that we are interested in:** For instance, if we are studying a new bacterium, we can sequence its ribosomal RNA (rRNA) and place it on a phylogenetic tree computed with all known ribosoman RNAs (rRNAs). This can give us a fairly good idea of who this bacterium really is.

2. **Discovering the function of Gene:** If we are studying a gene, we can use phylogenetic tree to make sure that the gene we are interested in is orthologous to another well characterised gene in another species.

3. **Retracing the origin of gene:** Most genes within a genome travel together through evolutionary time. However from time to time, individual genes may jump from one species to other one. Phylogenetic trees are great to reveal such events which are called horizontal (or lateral) transfer.

There are three main families of Methods:
• Parsimony
• Distance Method
• Maximum likelihood Methods

Methods directly based on sequences :

1. Maximum Parsimony : find a phylogenetic tree that explains the data, with as few evolutionary changes as possible.

2. Maximum likelihood : find a tree that maximizes the probability of the genetic data given the tree.

Methods indirectly based on sequences :

1. Distance based methods (Neighbour Joining (NJ) : find a tree such that branch lengths of paths between sequences (species) fit à matrix of pairwise distances between sequences.

Maximum Parsimony

Maximum parsimony is a useful approach to creating phylogenetic trees. By evaluating different possibilities for the evolutionary relationships among a set of organisms and selecting the most parsimonious, or simplest, option, we can maximize the likelihood that the hypothesis we select is true.

We can define maximum parsimony as the state in which our phylogenetic tree includes the fewest possible number of evolutionary steps between the organisms it features. Therefore, out of all the possible ways in which a group of organisms might be evolutionarily related, the simplest explanation is said to have maximum parsimony.

Of all possible phylogenetic trees for our group of organisms, we would therefore choose the tree with the least number of branches and intermediate common ancestors as the most likely to be true.

Why is the tree with the fewest branches most likely to be correct? To answer this question, consider what each branch on a phylogenetic tree represents. Every time a phylogenetic tree branches out, this signifies that the lineage it represents has diverged down two separate paths. The point at which these paths separate is called a *node,* and represents the last common ancestor shared between the two new lineages.

A tree with fewer branches indicates that fewer of these divergences occurred throughout the evolutionary history of the tree's organisms. This is parsimony in phylogeny: the simplest possible evolutionary path that could reasonably explain the relationships between the organisms on the tree.

Conversely, a tree that has more branches represents an evolutionary path that is more complicated. A hypothesis like this suggests that there were many divergence events, and many separate lineages, which arose over the course of the evolution of the creatures on this tree.

Since this is not the simplest possible tree, the hypothesis it represents does not abide by the principle of maximum parsimony. Rather, the tree that we could create with the fewest possible number of branches is the most parsimonious, and therefore the most likely to be true.

It's important to remember that there are limitations to this approach. Sometimes, as in parallel or convergent evolution, the actual evolutionary path is not the most parsimonious. However, seeking maximum parsimony is overall an effective way to analyze phylogenetic trees.

Maximum parsimony predicts the evolutionary tree or trees that minimize the number of steps required to generate the observed variation in the sequences from common ancestral sequences. For this reason, the method is also sometimes referred to as the minimum evolution method. A multiple sequence alignment (msa) is required to predict which sequence positions are likely to correspond. These positions will appear in vertical columns in the msa. For each aligned position, phylogenetic trees that require the smallest number of evolutionary changes to produce the observed sequence changes from ancestral sequences are identified. This analysis is continued for every position in the sequence alignment. Finally, those trees that produce the smallest number of changes overall for all sequence positions are identified. This method is best suited for sequences that are quite similar and is limited to small numbers of sequences.

Maximum Likelihood

Maximum Likelihood is a method for the inference of phylogeny. It evaluates a hypothesis about evolutionary history in terms of the probability that the proposed model and the hypothesized history would give rise to the observed data set. The supposition is that a history with a higher probability of reaching the observed state is preferred to a history with a lower probability. The method searches for the tree with the highest probability or likelihood.

Programs

The Maximum Likelihood method of inference is available for both nucleic acid and protein data. The following programs are available from the web:

- DNAML (DNA data only. By Joe Felsenstein in the Phylip package)
- FastDNAML (DNA data only. A faster algorithm applied by Garry Olsen applied to Joe Felsenstein's program DNAML)
- ProtML (DNA and protein. By Adachi and Hasegawa)
- Puzzle (DNA and protein. By Strimmer and von Haeseler). This program is much faster than PROTML

Advantages and disadvantages of maximum likelihood methods:

There are some supposed advantages of maximum likelihood methods over other methods.

They have often lower variance than other methods (ie. it is frequently the estimation method least affected by sampling error)

They tend to be robust to many violations of the assumptions in the evolutionary model

Even with very short sequences they tend to outperform alternative methods such as parsimony or distance methods.

The method is statistically well founded

They evaluate different tree topologies

They use all the sequence information

Disadvantages

Maximum likelihood is very CPU intensive and thus extremely slow

The result is dependent on the model of evolution used

6. **Algorithms (Bioinformatics):** It introduces readers to the art of algorithms in a practical manner which is linked with biological theory and interpretation including key concepts and algorithms, the development of the field historically, its applications and relevant ethical considerations. Topics covered are retrieval of information from biological databases, pairwise and multiple sequence alignment, phylogenetic trees, score matrices, sequence search in databases with BLAST, statistical evaluation of alignment scores, and measures of classification performance.

Alignment methods

There are various alignment methods used within multiple sequences to maximize scores and correctness of alignments. Each is usually based on a certain heuristic with an insight into the evolutionary process.

Most try to replicate evolution to get the most realistic alignment possible to best predict relations between sequences.

Dynamic programming

A direct method for producing an MSA uses the dynamic programming technique to identify the globally optimal alignment solution. For proteins, this method usually involves two sets of parameters: a gap penalty and a substitution matrix assigning scores or probabilities to the alignment of each possible pair of amino acids based on the similarity of the amino acids' chemical properties and the evolutionary probability of the mutation. For nucleotide sequences, a similar gap penalty is used, but a much simpler substitution matrix, wherein only identical matches and mismatches are considered, is typical. The scores in the substitution matrix may be either all positive or a mix of positive and negative in the case of a global alignment, but must be both positive and negative, in the case of a local alignment.

For *n* individual sequences, the naive method requires constructing the *n*-dimensional equivalent of the matrix formed in standard pairwise sequence alignment. The search space thus increases exponentially with increasing *n* and is also strongly dependent on sequence length. Expressed with the big O notation commonly used to measure computational complexity, a naïve MSA takes $O(Length^{Nseqs})$ time to produce. To find the global optimum for *n* sequences this way has been shown to be an NP-complete problem. In 1989, based on Carrillo-Lipman Algorithm, Altschul introduced a practical method that uses pairwise alignments to constrain the n-dimensional search space. In this approach pairwise dynamic programming alignments are performed on each pair of sequences in the query set, and only the space near the n-dimensional intersection of these alignments is searched for the n-way alignment. The MSA program optimizes the sum of all of the pairs of characters at each position in the alignment (the so-called *sum of pair* score) and has been implemented in a software program for constructing multiple sequence alignments. In 2019, Hosseininasab and van Hoeve showed that by using decision diagrams, MSA may be modeled in polynomial space complexity.

Progressive alignment construction

The most widely used approach to multiple sequence alignments uses a heuristic search known as progressive technique (also known as the hierarchical or tree method) developed by Da-Fei Feng and Doolittle in 1987. Progressive alignment builds up a final MSA by combining pairwise alignments beginning with the most similar pair and progressing to the most distantly related. All progressive alignment methods require two stages: a first stage in which the relationships between the sequences are represented as a tree, called a guide tree, and a second step in which the MSA is built by adding the sequences sequentially to the growing MSA according to the guide tree. The initial guide tree is determined by an efficient clustering method such as neighbor-joining or UPGMA, and may use distances based on the number of identical two-letter sub-sequences (as in FASTA rather than a dynamic programming alignment).

Progressive alignments are not guaranteed to be globally optimal. The primary problem is that when errors are made at any stage in growing the MSA, these errors are then propagated through to the final result. Performance is also particularly bad when all of the sequences in the set are rather distantly related. Most modern progressive methods modify their scoring function with a secondary weighting function that assigns scaling factors to individual members of the query set in a nonlinear fashion based on their phylogenetic

distance from their nearest neighbors. This corrects for non-random selection of the sequences given to the alignment program.

Progressive alignment methods are efficient enough to implement on a large scale for many (100s to 1000s) sequences. Progressive alignment services are commonly available on publicly accessible web servers so users need not locally install the applications of interest. The most popular progressive alignment method has been the Clustal family, especially the weighted variant ClustalW to which access is provided by a large number of web portals including Genome Net, EBI, and EMBNet. Different portals or implementations can vary in user interface and make different parameters accessible to the user. ClustalW is used extensively for phylogenetic tree construction, in spite of the author's explicit warnings that unedited alignments should not be used in such studies and as input for protein structure prediction by homology modeling. Current version of Clustal family is ClustalW2. EMBL-EBI announced that CLustalW2 will be expired in August 2015. They recommend Clustal Omega which performs based on seeded guide trees and HMM profile-profile techniques for protein alignments. They offer different MSA tools for progressive DNA alignments. One of them is MAFFT (Multiple Alignment using Fast Fourier Transform).

Another common progressive alignment method called T-Coffee is slower than Clustal and its derivatives but generally produces more accurate alignments for distantly related sequence sets. T-Coffee calculates pairwise alignments by combining the direct alignment of the pair with indirect alignments that aligns each sequence of the pair to a third sequence. It uses the output from Clustal as well as another local alignment program LALIGN, which finds multiple regions of local alignment between two sequences. The resulting alignment and phylogenetic tree are used as a guide to produce new and more accurate weighting factors.

Because progressive methods are heuristics that are not guaranteed to converge to a global optimum, alignment quality can be difficult to evaluate and their true biological significance can be obscure. A semi-progressive method that improves alignment quality and does not use a lossy heuristic while still running in polynomial time has been implemented in the program PSAlign.

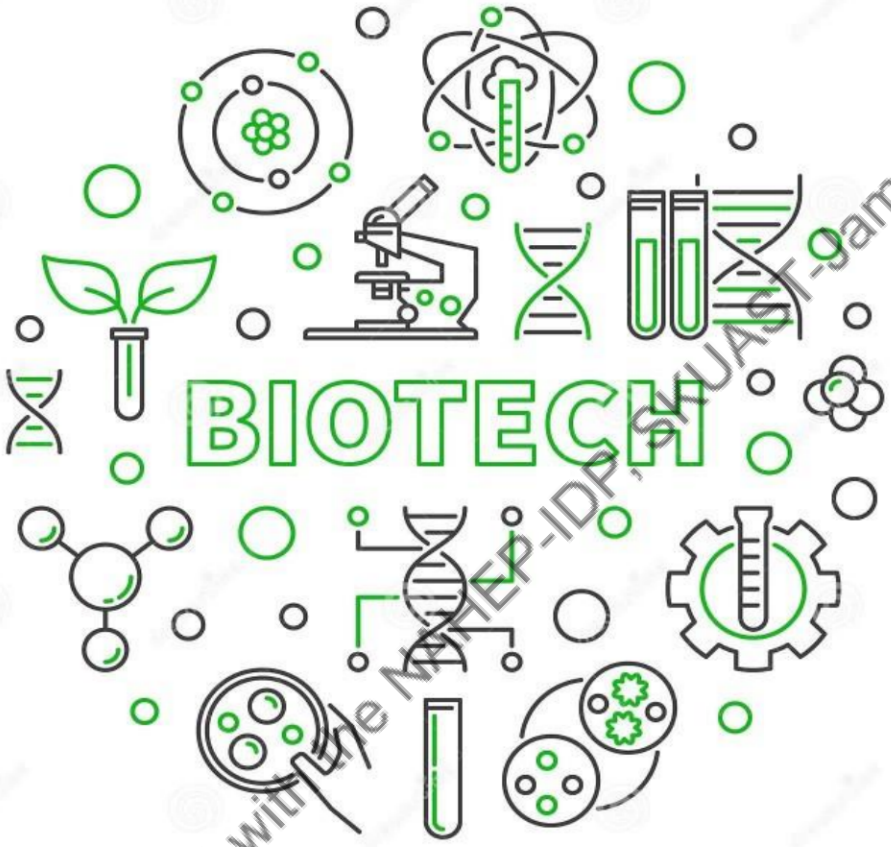
Iterative methods

A set of methods to produce MSAs while reducing the errors inherent in progressive methods are classified as "iterative" because they work similarly to progressive methods but repeatedly realign the initial sequences as well as adding new sequences to the growing MSA. One reason progressive methods are so strongly dependent on a high-quality initial alignment is the fact that these alignments are always incorporated into the final result - that is, once a sequence has been aligned into the MSA, its alignment is not considered further. This approximation improves efficiency at the cost of accuracy. By contrast, iterative methods can return to previously calculated pairwise alignments or sub-MSAs incorporating subsets of the query sequence as a means of optimizing a general objective function such as finding a high-quality alignment score.

A variety of subtly different iteration methods have been implemented and made available in software packages; reviews and comparisons have been useful but generally refrain from choosing a "best" technique. The software package PRRN/PRRP uses a hill-climbing algorithm to optimize its MSA alignment score and iteratively corrects both alignment weights and locally divergent or "gappy" regions of the growing MSA. PRRP performs best when refining an alignment previously constructed by a faster method

Another iterative program, DIALIGN, takes an unusual approach of focusing narrowly on local alignments between sub-segments or sequence motifs without introducing a gap penalty. The alignment of individual motifs is then achieved with a matrix representation similar to a dot-matrix plot in a pairwise alignment. An alternative method that uses fast local alignments as anchor points or "seeds" for a slower global-alignment procedure is implemented in the CHAOS/DIALIGN suite.

A third popular iteration-based method called MUSCLE (multiple sequence alignment by log-expectation) improves on progressive methods with a more accurate distance measure to assess the relatedness of two sequences. The distance measure is updated between iteration stages (although, in its original form, MUSCLE contained only 2-3 iterations depending on whether refinement was enabled).



Designed by : Magandeep Kaur

Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu (SKUAST-Jammu)

www.skuast.org