# Unit 2 - Session 1

## SLO 1: Structure & function of carbohydrates

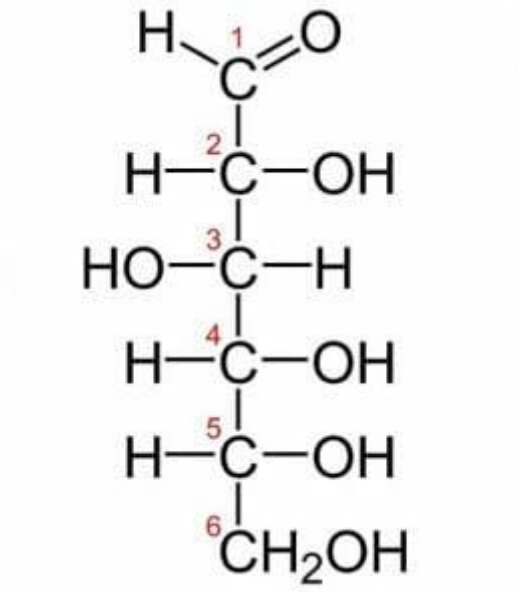
- Carbohydrates are biological molecules made of carbon, hydrogen, and oxygen in a ratio of approximately one carbon atom to one water molecule.
- Carbohydrates are, in fact, an essential part of our diet; grains, fruits, and vegetables are all natural sources of carbohydrates.
- Carbohydrates provide energy to the body, particularly through glucose, a simple sugar.
- Carbohydrates are classified into three subtypes:

1. Monosaccharides,
2. Disaccharides, and
3. Oligosaccharides
4. Polysaccharides.

**Monosaccharides: Monosaccharides** are otherwise known as simple sugar (that cannot be hydrolyzed into simpler carbohydrates).

**Ex. Aldose sugar (Glucose) and Keto sugar (Fructose)**

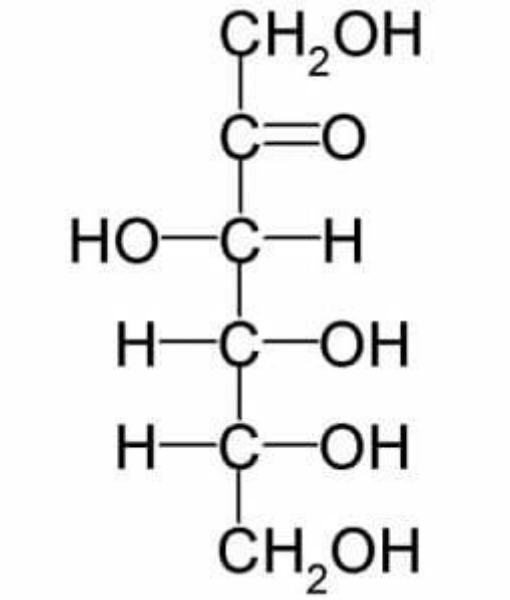
- All monosaccharides have the same general formula of $(CH_2O)_n$, which designates a central carbon molecule bonded to two hydrogens and one oxygen.
- The oxygen will also bond to a hydrogen, creating a hydroxyl group. Because carbon can form 4 bonds, several of these carbon molecules can bond together.
- One of the carbons in the chain will form a double bond with an oxygen, which is called a carbonyl group. If this carbonyl occurs at the end of the chain, the monosaccharide is in the *aldose* family.
- If the carboxyl group is in the middle of the chain, the monosaccharide is in the *ketose* family.




.

## Fructose

Although almost identical to glucose, fructose is a slightly different molecule. The formula ($(CH_2O)_6$) is the same, but the structure is much different. Below is an image of fructose:




Notice that instead of the carbonyl group being at the end of the molecule, as in glucose, it is the second carbon down. This makes fructose a ketose, instead of an aldose. Like glucose, fructose still has 6 carbons, each with a hydroxyl group attached. However, because the double bonded oxygen in fructose exists in a different place, a slightly different shaped ring is formed. In nature, this makes a big difference in how the sugar is processed. Most reactions in cells are catalyzed by specific enzymes. Different shaped monosaccharides each need a specific enzyme to be broken down.

Fructose, because it is a monosaccharide, can be combined with other monosaccharides to form oligosaccharides. A very common disaccharide made by plants is sucrose. Sucrose is one fructose molecule connected to a glucose molecule through a glycosidic bond.

**Disaccharides** : formed when two monosaccharides undergo a dehydration reaction (a reaction in which the removal of a water molecule occurs).

**Ex. Lactose, Sucrose**

**Oligosaccharides : Oligosaccharides** (*oligo* means in Greek – few) are condensation products of two to ten monosaccharides units joined by characteristics linkages called glycosidic bonds.

Ex. maltotriose (three glucose molecule).

**Polysaccharides Polysaccharides** (also known as glycanes or poly sides) are condensation products may have hundreds or thousands of monosaccharides units.

Ex. Storage (starch, glycogen, inulin) and structural (cellulose, pectin, chitin) form of polysaccharides.

- **Starch** is the stored form of sugars in plants and is made up of amylose and amylopectin (both polymers of glucose).
- Plants are able to synthesize glucose, and the excess glucose is stored as starch in different plant parts, including roots and seeds.
- **Glycogen** is the storage form of glucose in humans and other vertebrates, and is made up of monomers of glucose.
- Glycogen is the animal equivalent of starch and is a highly branched molecule usually stored in liver and muscle cells. Whenever glucose levels decrease, glycogen is broken down to release glucose.
- **Cellulose** is one of the most abundant natural biopolymers. The cell walls of plants are mostly made of cellulose, which provides structural support to the cell.
- Cellulose is made up of glucose monomers that are linked by bonds between particular carbon atoms in the glucose molecule.
- Cellulose gives rigidity and high tensile strength—which is so important to plant cells
- Cellulases can break down cellulose into glucose monomers that can be used as an energy source by the animal.
- Carbohydrates serve other functions in different animals. Arthropods, such as insects, spiders, and crabs, have an outer skeleton, called the exoskeleton, which protects their internal body parts. This exoskeleton is made of the biological macromolecule **chitin**, which is a nitrogenous carbohydrate. It is made of repeating units of a modified sugar containing nitrogen.

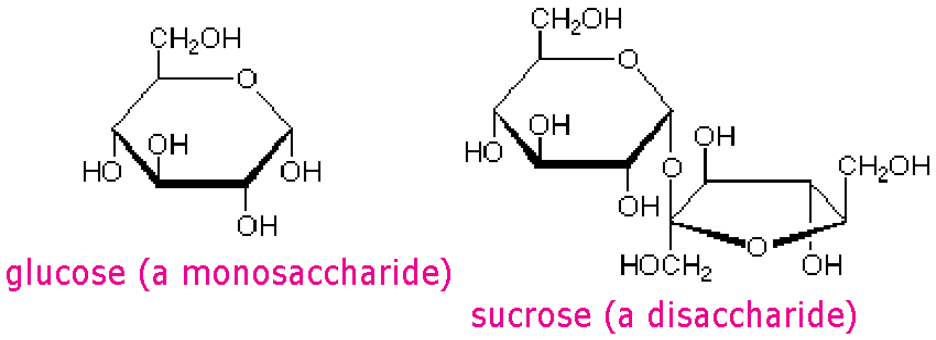



Fig 1: Glucose & Sucrose

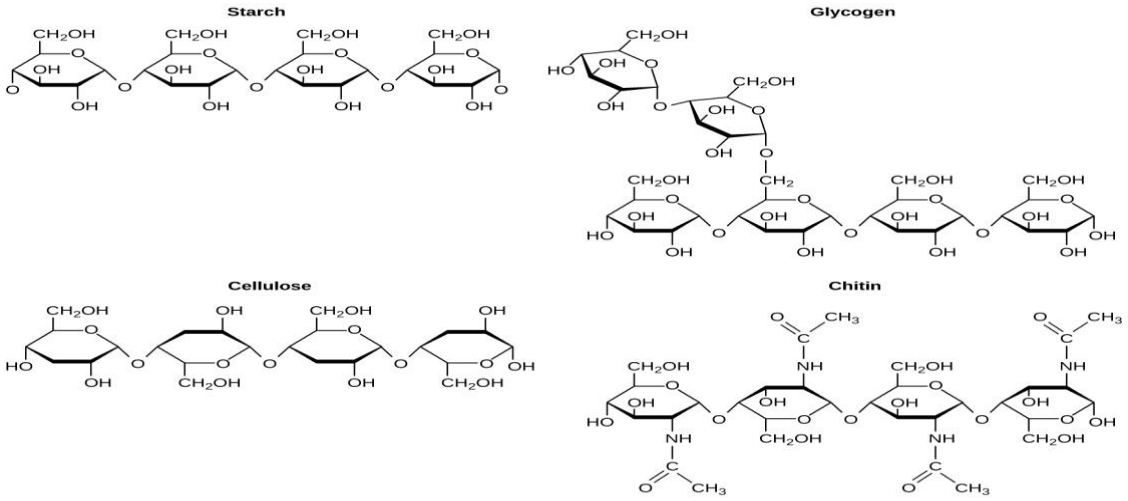



**Fig 2: Polysaccharides**

## Unit 2 - Session 2

## SLO 2: lipids

- The lipids form a heterogeneous class of compounds related to fatty acids and include fats, oils, waxes, steroids, and related compounds.
- Phospholipids and sterols are major structural elements of biological membranes.
- Other lipids although present in small quantities, plays a crucial role as enzyme cofactors, electron carriers, light-absorbing pigments, hydrophobic anchors, emulsifying agents, hormones and intercellular messengers

**Classification of lipids**

1) **Simple lipids or Homolipids :**

- These are esters of fatty acids.
- The Fats and oils are stored form of energy in many organisms are derivatives of fatty acids.
- Fats are stored in adipose tissue, where they also serve as a thermal insulator in subcutaneous tissues and around nearby organs.

2) **<u>Triacylglycerols</u>**

These are are fatty acid esters of glycerol.

It is mainly defined as the simplest lipids. These are composed of three fatty acids are linked in ester bond with single glycerol.

3) **Waxes** are esters of fatty acids. They can serve as a major storage form of metabolic fuels. Waxes are helped to protect hair and skin.

**<u>Fatty acids</u>**

- Fatty acids may be saturated or unsaturated.

- **Saturated fatty acids** : if there are only single bonds between neighboring carbons in the hydrocarbon chain, the fatty acid is saturated
- When the hydrocarbon chain contains a double bond, the fatty acid is an **unsaturated fatty acid**.

**Phospholipids**

- **Phospholipids** are the major constituent of the plasma membrane.
- A phospholipid has both hydrophobic and hydrophilic regions. The fatty acid chains are hydrophobic and exclude themselves from water, whereas the phosphate is hydrophilic and interacts with water.
- Cells are surrounded by a membrane, which has a bilayer of phospholipids. The fatty acids of phospholipids face inside, away from water, whereas the phosphate group can face either the outside environment or the inside of the cell, which are both aqueous.

**Cholesterol**

- Cholesterol is a steroid. Cholesterol is mainly synthesized in the liver and is the precursor of many steroid hormones, such as testosterone and estradiol. It is also the precursor of vitamins E and K
- . Cholesterol is the precursor of bile salts, which help in the breakdown of fats and their subsequent absorption by cells. Although cholesterol is often spoken of in negative terms, it is necessary for the proper functioning of the body. It is a key component of the plasma membranes of animal cells.
- Waxes are made up of a hydrocarbon chain with an alcohol (–OH) group and a fatty acid. Examples of animal waxes include beeswax and lanolin. Plants also have waxes, such as the coating on their leaves, that helps prevent them from drying out.

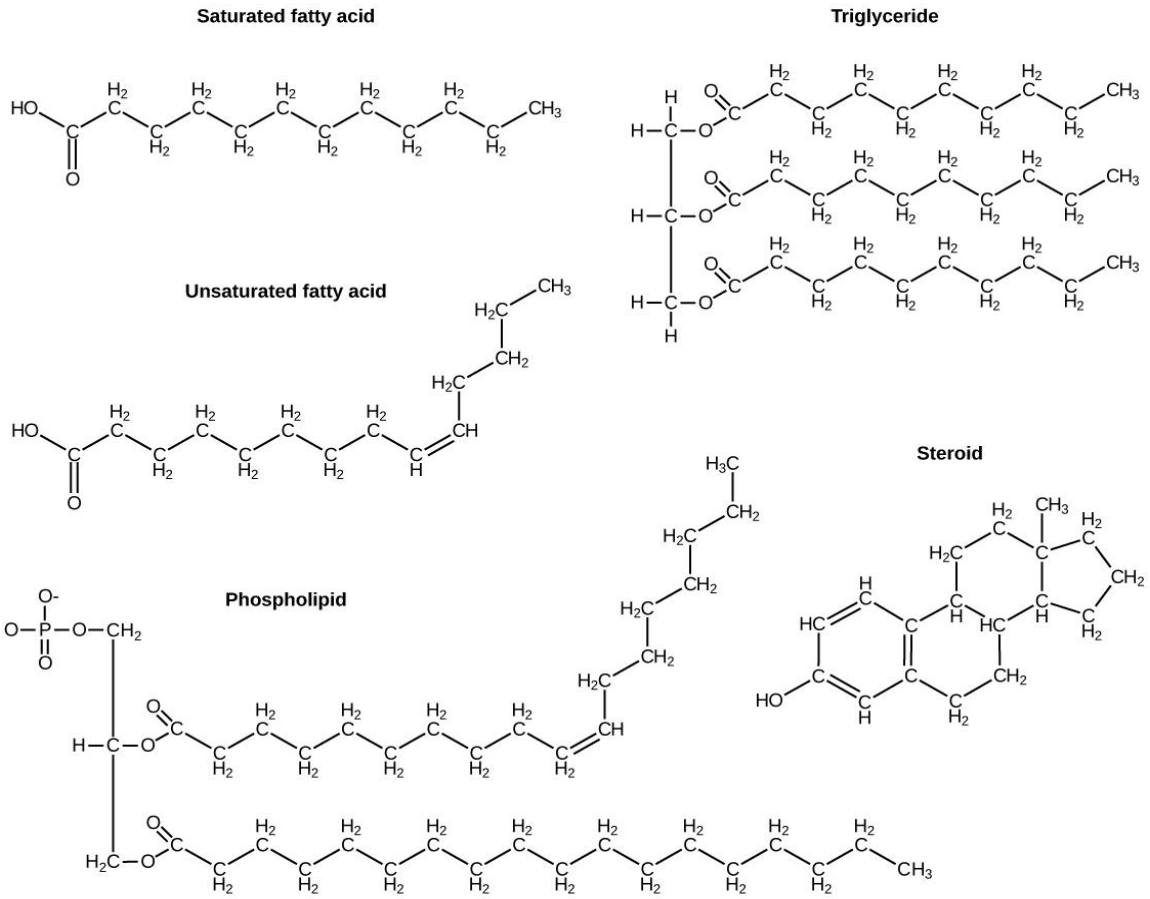
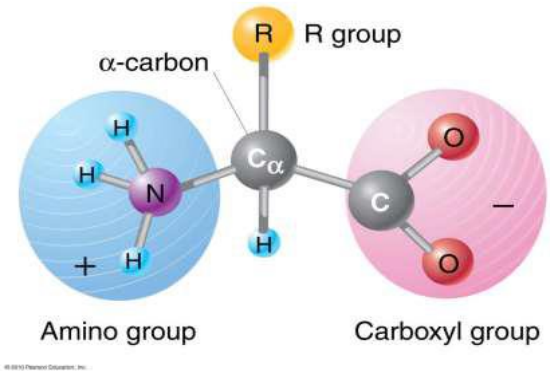



**Fig 3: lipids**

## Unit 2 - Session 3

## SLO 3: Proteins

- Proteins are the most abundant organic molecules of the living system. Proteins are made up of amino acids linked by peptide bond.
- Each amino acid has 4 different groups attached to α- Carbon (which is C atom next to COOH).
- These 4 groups are: Amino group, COOH group, Hydrogen atom and side chain (R).



| Amino Acid | Three Letter Code | One Letter Code |
|---|---|---|
| Alanine | Ala | A |
| Arginine | Arg | R |
| Aspartic Acid | Asp | D |
| Asparagine | Asn | N |
| Cysteine | Cys | C |
| Glutamic Acid | Glu | E |
| Glutamine | Gln | Q |
| Glycine | Gly | G |
| Histidine | His | H |
| Isoleucine | Ile | I |
| Leucine | Leu | L |
| Lysine | Lys | K |
| Methionine | Met | M |
| Phenylalanine | Phe | F |
| Proline | Pro | P |
| Serine | Ser | S |
| Threonine | Thr | T |
| Tryptophan | Trp | W |
| Tyrosine | Tyr | Y |
| Valine | Val | V |

## PROTEIN STRUCTURE

- Primary
- Secondary
- Tertiary &
- Quaternary structure

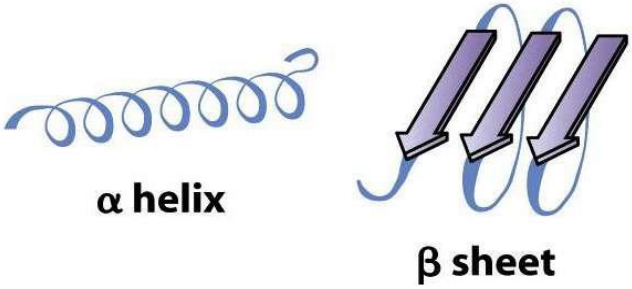
## PRIMARY STRUCTURE

- The primary structure of protein refers to the sequence of amino acids present in the polypeptide chain.
- Primary structure of protein starts from the amino terminal (N) end and ends in the carboxyl terminal (C) end.
- Amino acids are covalently linked by peptide bonds or covalent bonds.

## Secondary structure

- Pauling & Corey studied the secondary structures and proposed 2 conformations

1. **α helix**
2. **β sheets**




## ALPHA HELIX

- Right handed spiral structure.
- Stabilized by H bonding

- Amino acids per turn – 3.6
- Alpha helical segments are found in many globular proteins like Myoglobin, Troponin C.
- phi = -60 degrees, psi = -45 degrees falls within the fully allowed regions of the Ramachandran plot.

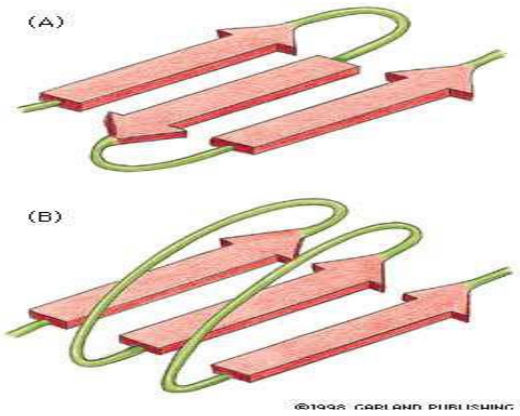
**BETA PLEATED SHEET**

- Formed when 2 or more polypeptides line up side by side.
- Individual polypeptide – beta strand.
- They are stabilized by hydrogen bond between N- H and carbonyl groups of adjacent chains.

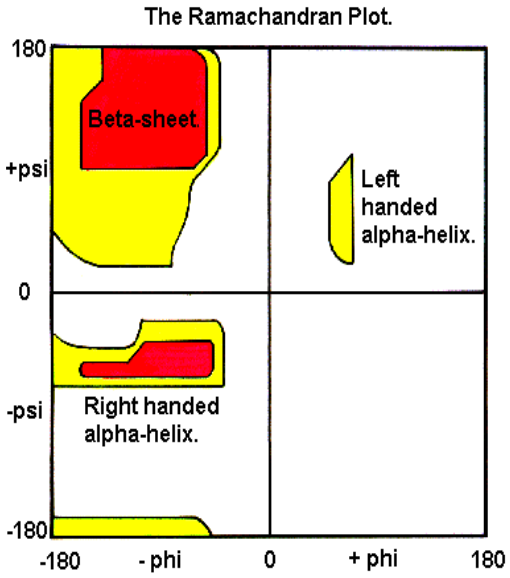
Beta sheets come in two varieties

1. Antiparallel beta sheet – neighbouring hydrogen bonded polypeptide chains run in opposite direction.
2. Parallel beta sheet - hydrogen bonded chains extend in the same direction.

- The connection between two antiparallel strands may be just a small loop




- The conformation of the backbone can therefore be described by the torsion angles (also called dihedral angles or rotational angles).
- These angles are both defined as 180° when the polypeptide chain is in full conformation.
- The rotations around the Cα – N bond (measured as Φ) and the Cα –C bond (measured as Ψ).
- The aminoacids with larger side chains will show less number of allowed regions within the ramachandran plot.
- Ramachandran plot – to visualize the backbone of amino acid residues.




- Loops and turns connect α helices and β strands.
- In addition to α helices and β strands, a folded polypeptide chain contains two other types of secondary structure called loops and turns. θ

**TERTIARY STRUCTURE**

- It is based on various types of interactions between the side-chains of the peptide chain.
- The tertiary structure defines the specific overall 3-D shape of the protein.
- The disulfide bonds between cysteine residue helps to maintain the protein's tertiary structure.
- Interactions stabilizing tertiary structure :

1. Disulfide bonds
2. Hydrophobic interactions
3. Hydrogen bonds
4. Ionic interactions
5. Vander Waals force
6. Salt bridges

**QUATERNARY PROTEIN**

The quaternary protein involves the clustering of several individual peptide or protein chains into a final specific shape.

. Two kinds of quaternary structures:

1) Homodimer : association between identical polypeptide chains.
2) Heterodimer : interactions between subunits of very different structures

**MYOGLOBIN**

- Myoglobin is closely related to hemoglobin,
- It is found in abundance in the skeletal muscle of vertebrates, and is responsible for the characteristic red color of muscle tissue.
- Myoglobin is a small, monomeric protein which serves as an intracellular oxygen storage site.

**HAEMOGLOBIN**

- Haemoglobin is a globular protein with 4 polypeptide chains bonded together.
- The four polypeptide chain consists of two alpha and two beta chains.
- There are 4 haem groups each contain iron.
- Each haem group can carry one molecule of oxygen.

**Types of Proteins**

**Fibrous proteins**

Fibrous proteins are a secondary structure, insoluble in water, physically tough, long parallel polypeptide chains cross-linked at intervals forming long fibers or sheets. These are mainly of animal origin.

Functions

Fibrous proteins perform structural functions or protective role in cells and organisms examples, collagen (tendons, bone, connective tissue), myosin (in muscle), silk (spiders webs), keratin (hair, nails, feathers).

**Globular proteins**

Globular proteins are more complex than fibrous proteins, biological Globular proteins are mostly tertiary structure, polypeptide chains tightly folded to form spherical shape, easily soluble in nature.

Functions

Enzymes, antibodies, blood transport proteins, nutrient storage proteins, and some hormone

**Conjugated proteins**

Conjugated proteins are complex globular proteins and were tightly bound. These are the proteins are linked with non-protein part is called prosthetic group. Further classified based on the nature of the prosthetic group.

They are metalloproteins, glycoproteins, chromoproteins, lipoproteins, phosphoproteins, and nucleoproteins.

## Unit 2 - Session 3

## SLO 3: Enzymes

- Enzymes can be defined as biological polymers that catalyze biochemical reactions.
- The initial stage of metabolic process depends upon the enzymes, which react with a molecule and is called the **substrate.**
- Enzymes convert the substrates into other distinct molecules, which are known as **products.**

- The macromolecular components of all enzymes consist of protein, except in the class of RNA catalysts called ribozymes. The word ribozyme is derived from the ribonucleic acid enzyme.
- Enzymes are said to possess an active site. The active site is a part of the molecule that has a definite shape and the functional group for the binding of reactant molecules.

According to the International Union of Biochemists (I U B), enzymes are divided into six functional classes and are classified based on the type of reaction in which they are used to catalyze.

**The six categories of enzymes are**

1. **Hydrolases,**
2. **Oxidoreductases,**
3. **Lyases,**
4. **Transferases,**
5. **Ligases and**
6. **Isomerases.**

**The six categories of enzymes are**

**Oxidoreductases**

These catalyze oxidation and reduction reactions, e.g. pyruvate dehydrogenase, catalysing the oxidation of pyruvate to acetyl coenzyme A.

**Transferases**

These catalyze transferring of the chemical group from one to another compound. An example is a transaminase, which transfers an amino group from one molecule to another.

**Hydrolases**

They catalyze the hydrolysis of a bond. For example, the enzyme pepsin hydrolyzes peptide bonds in proteins.

**Lyases**

These catalyze the breakage of bonds without catalysis, e.g. aldolase (an enzyme in glycolysis) catalyzes the splitting of fructose-1, 6-bisphosphate to glyceraldehyde-3-phosphate and dihydroxyacetone phosphate.

**Isomerases**

They catalyze the formation of an isomer of a compound. Example: phosphoglucomutase catalyzes the conversion of glucose-1-phosphate to glucose-6-phosphate (phosphate group is transferred from one to another position in the same compound) in glycogenolysis (glycogen is converted to glucose for energy to be released quickly).

**Ligases**

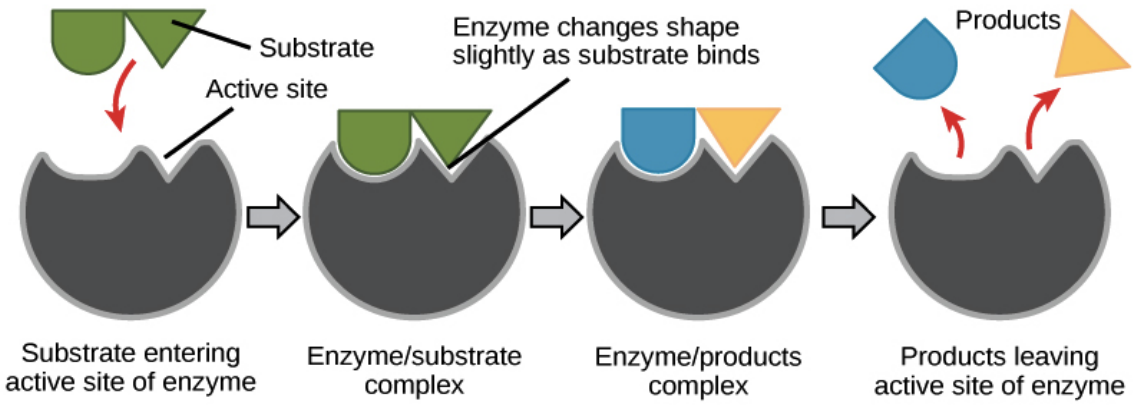
Ligases catalyze the association of two molecules. For example, DNA ligase catalyzes the joining of two fragments of DNA by forming a phosphodiester bond.

## Enzyme-Substrate Interactions

- The substrate which has an opposite charge of the enzyme fits into these spaces, just like a key fits into a lock. This substrate binding site is called the active site of an enzyme (E).
- The favourable model of enzyme-substrate interaction is called the **induced-fit model.**
- This model states that the interaction between substrate and enzyme is weak, and these weak interactions induce conformational changes rapidly and strengthen binding and bring catalytic sites close enough to substrate bonds.

### Covalent Catalysis

- The substrate is oriented to active place on the enzymes in such a manner that a covalent intermediate develops between the enzyme and the substrate, in catalysis that occurs by covalent mechanisms.
- The best example of this involves proteolysis by serine proteases that have both digestive enzymes and various enzymes of the blood clotting cascade.
- These proteases possess an active site serine whose R group hydroxyl generates a covalent bond with a carbonyl carbon of a peptide bond and results in the hydrolysis of the peptide bond.




### Factors affecting enzyme activity

### Temperature and pH

- Enzymes require an optimum temperature and pH for their action.
- The temperature or pH at which a compound shows its maximum activity is called optimum temperature or optimum pH, respectively.
- Enzymes are protein compounds, a temperature or pH more than optimum may alter the molecular structure of the enzymes.
- Generally, an optimum pH for enzymes is considered to be ranging between 5 and 7.
- Human enzymes have a optimum temperature range of 35°- 40°C

**Concentration and Type of Substrate**

- Enzymes have a saturation point, i.e., once all the enzymes added are occupied by the substrate molecules, its activity will be ceased.
- When the reaction begins, the velocity of enzyme action keeps on increasing on further addition of substrate.
- However, at a saturation point where substrate molecules are more in number than the free enzyme, the velocity remains the same.
- The type of substrate is another factor that affects the enzyme action. The chemicals that bind to the active site of the enzyme can inhibit the activity of the enzyme and such substrate is called **an inhibitor**.
- **Competitive inhibitors** are chemicals that compete with the specific substrate of the enzyme for the active site. They structurally resemble the specific substrate of the enzyme and bind to the enzyme and inhibit the enzymatic activity. This concept is used for treating bacterial infectious diseases.

**Functions of Enzymes**

1. Enzymes help in signal transduction. The most common enzyme used in the process includes protein kinase that catalyzes the phosphorylation of proteins.
2. They break down large molecules into smaller substances that can be easily absorbed by the body.
3. They help in generating energy in the body. ATP synthase is the enzyme involved in the synthesis of energy.
4. Enzymes are responsible for the movement of ions across the plasma membrane.
5. Enzymes perform a number of biochemical reactions, including oxidation, reduction, hydrolysis, etc. to eliminate the non-nutritive substances from the body.
6. They function to reorganize the internal structure of the cell to regulate cellular activities.

## Unit 2 - Session 3

## SLO 3: Hormones

Hormones are chemicals synthesized and produced by the specialized glands to control and regulate the activity of certain cells and organs. These specialized glands are known as endocrine glands.

**Types of Hormones**

- Peptide Hormones
- Steroid Hormones

**Peptide Hormones**

- Peptide hormones are composed of **amino acids** and are soluble in water.
- Peptide hormones are unable to pass through the cell membrane as it contains a phospholipid bilayer that stops any fat-insoluble molecules from diffusing into the cell.
- Insulin is an important peptide hormone produced by the pancreas.

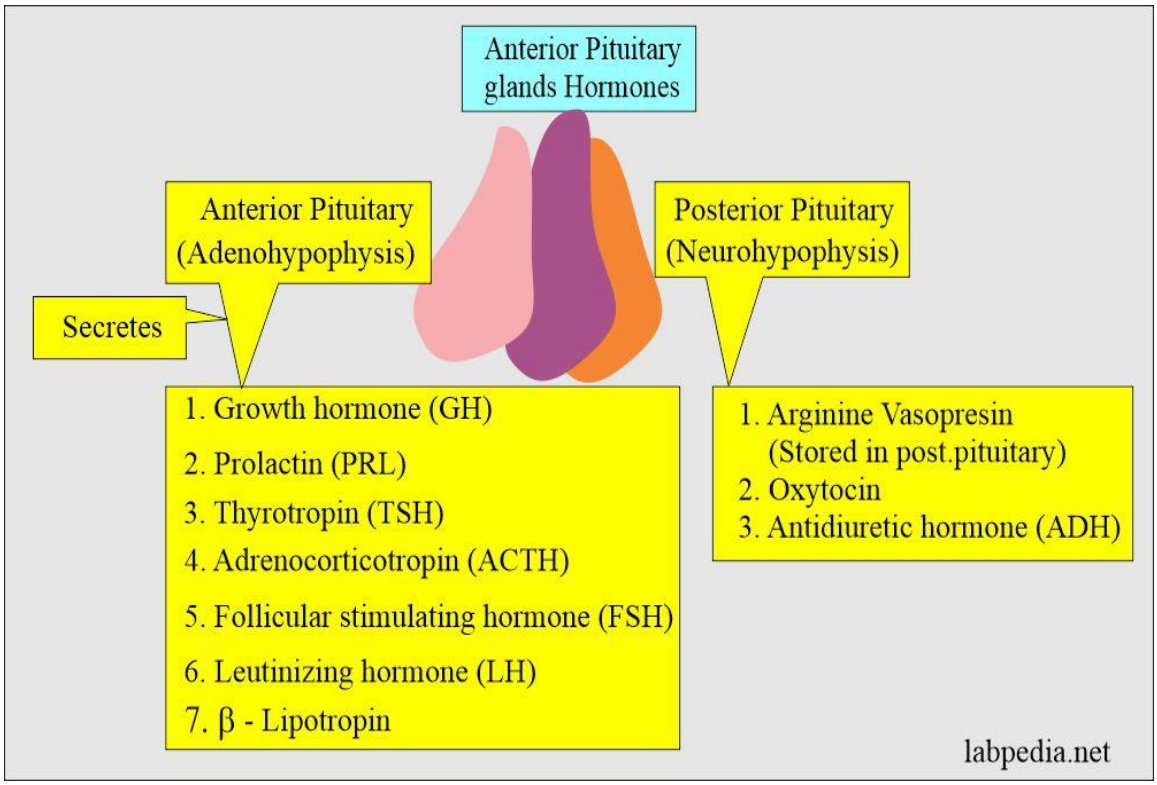
**Steroid Hormones**

- Steroid hormones are fat-soluble and are able to pass through a cell membrane.
- Sex hormones such as testosterone, estrogen and progesterone are examples of steroid hormone

**Endocrine Glands and the Hormones Secreted**

Hormones are released by the endocrine glands and are ductless

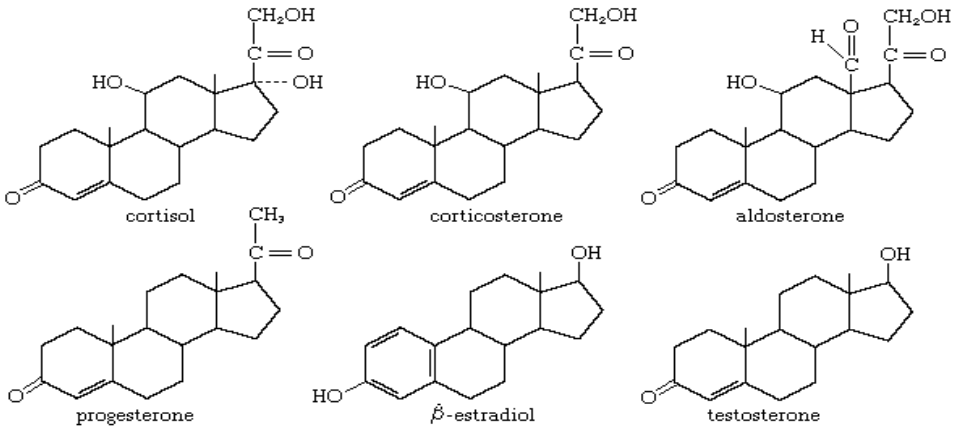
1. **Hypothalamus:** It controls the body temperature, regulates emotions, hunger, thirst, sleep, moods and allow the production of hormones.
2. **Pineal**: Pineal is also known as the thalamus. It produces serotonin derivatives of melatonin, which affects sleep patterns.
3. **Parathyroid**: This gland helps in controlling the amount of calcium present in the body.
4. **Thymus**: It helps in the production of T-cells, functioning of the adaptive immune system and maturity of the thymus.
5. **Thyroid:** It produces hormones that affect the heart rate and how calories are burnt.
6. **Adrenal**: This gland produces the hormones that control the sex drive, cortisol and stress hormone.
7. **Pituitary**: It is also termed as the "master control gland". This is because the pituitary gland helps in controlling other glands. Moreover, it develops the hormones that trigger growth and development.



8. **Pancreas**: This gland is involved in the production of insulin hormones, which plays a crucial role in maintaining blood sugar levels.
9. **Testes:** In men, the testes secrete the male sex hormone, testosterone. It also produces sperm.
10. **Ovaries**: In the female reproductive system, the ovaries release estrogen, progesterone, testosterone and other female sex hormones.

**Important Hormones**

1. **Cortisol** – It has been named as the "stress hormone" as it helps the body in responding to stress. This is done by increasing the heart rate, elevating blood sugar levels etc.
2. **Estrogen**-This is the main sex hormone present in women which bring about puberty, prepares the uterus and body for pregnancy and even regulates the menstrual cycle. Estrogen level changes during menopause because of which women experience many uncomfortable symptoms.
3. **Melatonin** – It primarily controls the circadian rhythm or sleep cycles.
4. **Progesterone** – It is a female sex hormone also responsible for menstrual cycle, pregnancy and embryogenesis.
5. **Testosterone** – This is the most important sex hormone synthesized in men, which cause puberty, muscle mass growth, and strengthen the bones and muscles, increase bone density and controls facial hair growth.




**Functions of hormones**

- Food metabolism.
- Growth and development.
- Controlling thirst and hunger.
- Maintaining body temperature.
- Regulating mood and cognitive functions.
- Initiating and maintaining sexual development and reproduction.

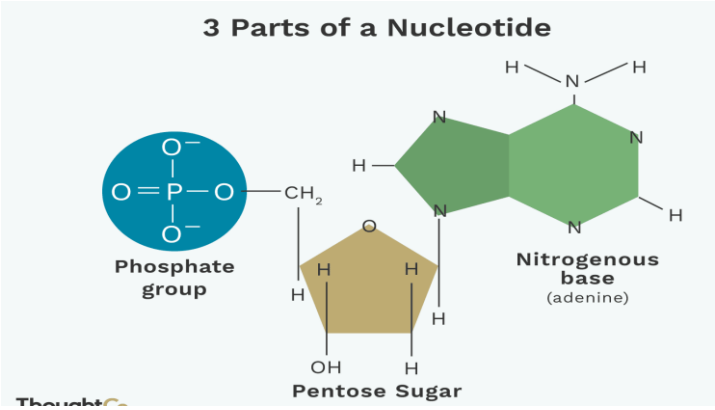
# Unit 2 - Session 4

## SLO 4: DNA

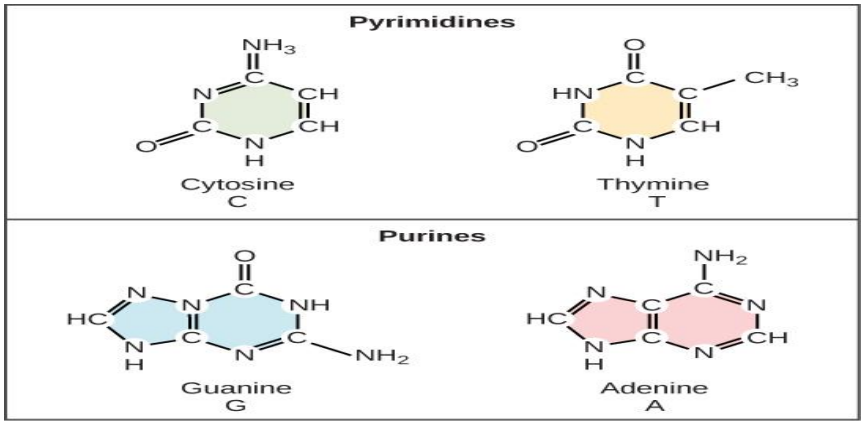
- DNA (Deoxyribonucleic acid) is a molecule that is responsible for carrying and transmitting the hereditary materials or the genetic instructions from parents to offsprings and for the production of proteins.
- DNA is found in both animal and plant cell in chromosomes, mitochondria, and chloroplasts.

**DNA structure**

- The double helix structure of a DNA molecule was discovered by James Watson, Francis Crick & Rosalin Franklin.
- The DNA structure is a long, double helix that resembles a ladder, which is twisted at both the ends.
- The DNA molecule is composed of basic materials called **nucleotides** and each **nucleotide** is composed of three different components such as
- **Sugar,**
- **Phosphate groups and**
- **Nitrogen bases.**




- The sugar and phosphate groups link the **nucleotides** together to form each strand of DNA.
- The purine and pyrimidine bases face the inside of the helix, with guanine always opposite cytosine and adenine always opposite thymine.




- Adenine (A), Thymine (T), Guanine (G) and Cytosine (C) are four types of nitrogen bases.

- Adenine and Thymine are the complementary pairs, Guanine and Cytocine are the complementary pairs.
- Erwin Chargaff, a Biochemist, found that the number of nitrogenous bases is present in equal amounts in the DNA. The amount of adenine (A) is equal to the amount of thymine (T). Whereas, the amount of guanosine (G) is equal to the amount of Cytosine (C). Adenine is linked with Thymine by a double bond and Guanosine is linked with Cytosine by a triple bond.

**Three different forms of DNA**

1) **A DNA**
2) **B DNA**
3) **Z DNA**

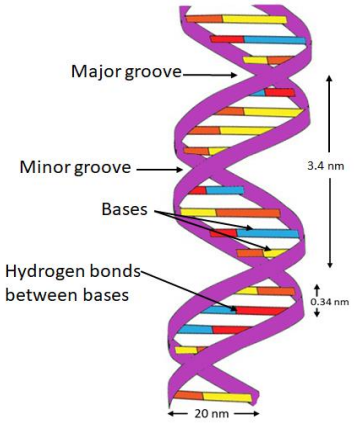
**B-DNA**

- Described by James D. Watson & Francis Crick. Most commonly forms of DNA.
- Right handed double helix. DNA molecule consists of 2 helical polynucleotide chains coiled around common axis.

2 helices are wound in such a way so as to produce 2 interchain spacing or groove:

- Major/wide groove (width 12A°,depth 8.5A°)
- Minor/narrow groove (width 6A°,depth 7.5A°)

- These grooves provide surface with which proteins, chemicals, drugs can interact.
- 2 chains run in opposite direction, they are antiparallel, the plane of bases are perpendicular to helix axis.
- Base pair per turn is 10.4.
- Rise per base pair is 3.4A°.




**DNA replication**

- DNA replicates by separating into two single strands, each of which serves as a template for a new strand.
- The new strands are copied by the same principle of hydrogen-bond pairing between bases that exist in the double helix.

- Two new double-stranded molecules of DNA are produced, each containing one of the original strands and one new strand.
- This "semiconservative" replication is the key to the stable inheritance of genetic traits.

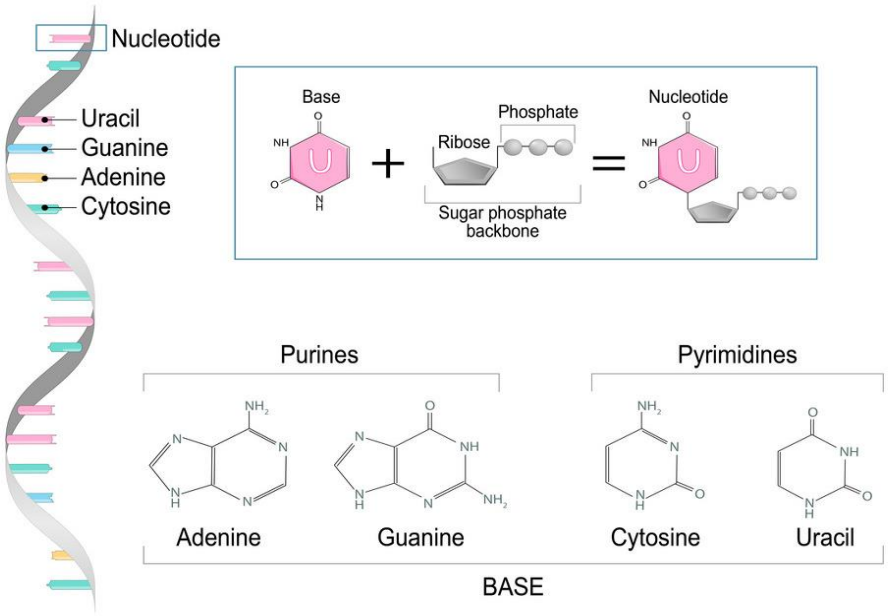
**DNA Functions**:

- DNA is the genetic material which carries all the hereditary information.
- Replication process: Transferring the genetic information from one cell to its daughters and from one generation to the next and equal distribution of DNA during the cell division
- Mutations: The changes which occur in the DNA sequences
- Transcription
- Cellular Metabolism
- DNA Fingerprinting
- Gene Therapy

## Unit 2 - Session 4

## SLO 4: RNA

- RNA is typically single stranded and is made of ribonucleotide that are linked by phosphodiester bonds
- A ribonucleotide in the RNA chain contains ribose (the pentose sugar), one of the four nitrogenous bases (A, U, G, and C), and a phosphate group.
- The RNA-specific pyrimidine uracil forms a complementary base pair with adenine and is used instead of the thymine (DNA)




- Even though RNA is single stranded, most types of RNA molecules show extensive intramolecular base pairing between complementary sequences within the RNA strand, creating a predictable three-dimensional structure essential for their function

- Although RNA is not used for long-term genetic information in cells, many viruses do use RNA as their genetic material.

**There are three main types of RNA, all involved in protein synthesis.**

1) **Messenger RNA (mRNA)** :

   serves as the intermediary between DNA and the synthesis of protein products during translation.

2) **Ribosomal RNA (rRNA) :**

   is a type of stable RNA that is a major constituent of ribosomes. It ensures the proper alignment of the mRNA and the ribosomes during protein synthesis and catalyzes the formation of the peptide bonds between two aligned amino acids during protein synthesis.

3) **Transfer RNA (tRNA)** :

   is a small type of stable RNA that carries an amino acid to the corresponding site of protein synthesis in the ribosome**.** It is the base pairing between the tRNA and mRNA that allows for the correct amino acid to be inserted in the polypeptide chain being synthesized.

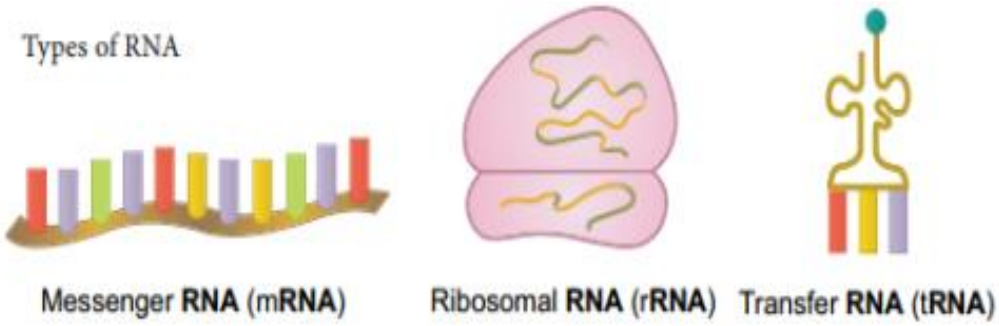



Fig : Types of RNA & Function

| Type | Abbreviation | Function(s) |
|---|---|---|
| Messenger RNA | mRNA | Transfers genetic information from genes to ribosomes to synthesize proteins. |
| Heterogeneous nuclear RNA | hnRNA | Serves as precursor for mRNA and other RNAs |
| Transfer RNA | tRNA | Transfers amino acid to mRNA for protein synthesis. |
| Ribosomal RNA | rRNA | Provides structural framework for ribosomes |
| Small nuclear RNA | snRNA | Involved in mRNA processing |
| Small nucleolar RNA | snoRNA | Plays a key role in processing of rRNA molecules |
| Small cytoplasmic RNA | scRNA | Involved in selection of proteins for export. |
| Transfer messenger RNA | tmRNA | Mostly present in Bacteria. Adds short peptide tags to proteins to facilitate the degradation of incorrectly synthesized proteins. |

## Unit 2 - Session 5

## SLO 5: The HUMAN GENOME PROJECT (HGP)

- The Human Genome Project (HGP) was an international scientific research project that aimed to determine the complete sequence of nucleotide base pairs (3.2 billion bp) that make up human DNA and all the genes.
- The **$3-billion project** was formally launched in 1990 by the US Department of Energy **(DOE)** and the National Institute of Health. (**NIH)**
- **James Watson** headed the NIH Genome Program.
- The Human Genome Project was a **13-year-long**, publicly funded project initiated in 1990 with the objective of determining the DNA sequence of the entire euchromatic human genome
- The 1st gene to be mapped was BRCA1 - gene for breast cancer
- HGP was declared complete in April 14, 2003 (final sequencing mapping of human genome 99.9% accuracy)
- Francis Collins succeeded James Watson in 1993 as the overall Project Head and the Director of the NIH and was in power until the completion of HGP in 2003.

**Goals of HGP:**

- To identify and map all the **20,000-25,000 genes (approx)** in the human DNA
- To determine the sequences of the **3 billion chemical base pairs of the human DNA.**

- To store these informations in **databases.**
- To discover more **efficient technologies for data analysis.**

**Timeline of HGP**

- **1997 –** E.coli genome sequence completed
- 1998 - Genome of the roundworm *Caenorhabditis elegans* was sequenced. SNP (single nucleotide polymorphism) sequencing was initiated
- 1998 - sequencing of human chromosome 22 was completed and was published in the journal "The Nature."
- 2000 - working draft of human genome completed
- 2001 – working draft of human genome sequence was published in "The Nature" & "Science".
- **2002 – working draft of mouse genome sequence was** completed & published.

**2003 - Finished version of human genome sequence** was completed

**METHODOLOGY:**

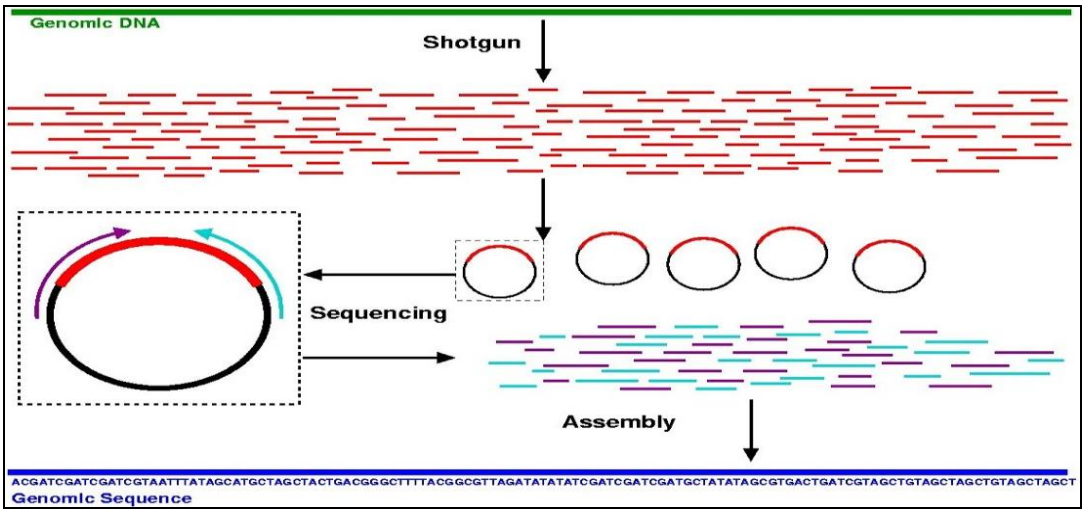
- DNA isolation
- PCR (polymerase chain reaction)
- RFLP
- Cloning
- Sequencing:

➢ Shot gun Sequencing

➢ Sanger Sequencing

**SHOT GUN SEQUENCING**

- ➢ Shotgun sequencing is a laboratory technique for determining the DNA sequence of an organism's genome.
- ➢ The method involves breaking the genome into a collection of small DNA fragments that are sequenced individually.
- ➢ The computer program looks for overlaps in the DNA sequences and uses them to place the individual fragments in their correct order to reconstitute the genome.
- ➢ The initial random fragmenting and reading of the DNA gave this approach the name "shotgun sequencing".




**SANGER SEQUENCING**

- Invented by Frederick Sanger in 1977
- Nobel prize - 1980
- It is also termed as **chain termination & dideoxy method** sequencing

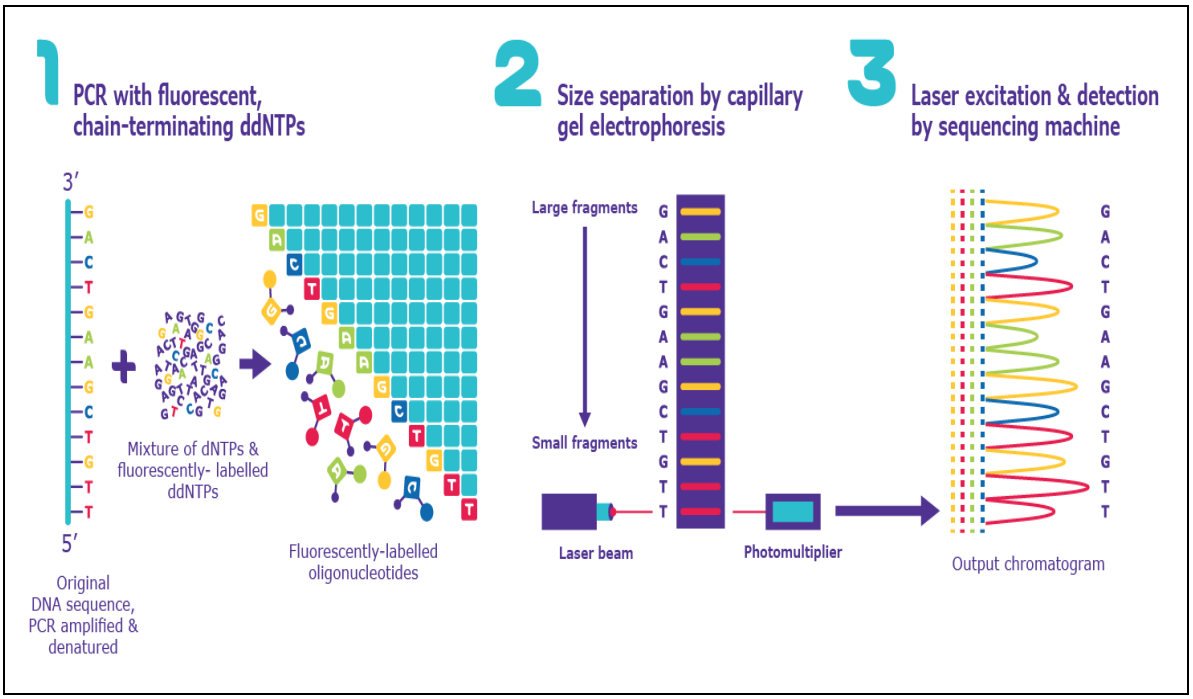
**Requirements :**

**DNA sequencing is performed in four separate tubes, each containing**

- **Single stranded DNA to be sequenced**
- **DNA polymerase**
- **Primers**
- **The four dNTPs (dATP, dCTP, dTTP and dGTP)**
- **Small amount of one of the four ddNTPs (ddATP or ddCTP or ddTTP or ddGTP)**

**Steps of Dideoxy Sequencing :**

- A primer is annealed to a single-stranded section of DNA
- DNA- primer mixture is put into 4 separate tubes with DNA polymerase and a solution of dNTPs at a concentration of 100 times lower than the dNTP concentration.
- DNA Polymerase uses dNTPs (deoxynucleotide triphosphates) to extend the DNA
- ddNTPs (dideoxy nucleotide triphosphate) are put together randomly, resulting in different lengths of fragments
- Fragments that are from each of the reactions are denatured and separated by size using gel electrophoresis
- The gel is used to visually detect the DNA fragments. The fragments are to be read from bottom to top and this represents the complementary sequence of the original strand of DNA.




## Unit 2 - Session 5

## SLO 5: Genomics

**Genome**

Genome is the entire set of genetic material of an organism (it can be either DNA or RNA).

**Genomics**

- Genomics is the study of the structure, function and inheritance of the genome of the organism.
- **Genomics** is a discipline in genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble and analyze the function and structure of genomes.

**<u>It comprises of:</u>**

- **Structural genomics** refers to the initial phase of genome analysis, which includes the construction of genetic and physical maps of a genome, identification of genes, annotation of gene features, and comparison of genome structures.
- **Functional genomics** is the study of how genes and intergenic regions of the genome contribute to different biological processes.. Functional genomics focuses on the dynamic expression of gene products in a specific context, for example, at a specific developmental stage or during disease.
- **Comparative genomics** is the comparison of gene number, gene location, and gene content from these genomes. The comparison helps to reveal the extent of conservation among genomes, which will provide insights into the mechanism of genome evolution and gene transfer among genomes.
- **Epigenomics**

- **Metagenomics**
- **Pharmacogenomics**
- **Mutation Genomics.**

**Methods in Genomics**

Genome mapping

- Genome mapping is a process of identifying relative locations of genes, mutations or traits on a chromosome.
- It involves assigning/locating of a specific gene to particular region of a chromosome and determining the location of and relative distances between genes on the chromosome.
- **Linkage maps** show the arrangement of genes and genetic markers along the chromosomes as calculated by the frequency with which they are inherited together.
- **Physical maps** represent chromosomes and provide physical distances between chromosomal landmarks ideally measured in nucleotide bases.

**Genome sequence Assembly**

- Initial DNA sequencing reactions generate short sequence reads from DNA clones.
- The average length of the reads is about 500 bases. To assemble a whole genome sequence, these short fragments are joined to form larger fragments after removing overlaps. These longer, merged sequences are termed contigs, which are usually 5,000 to 10,000 bases long.
- A number of overlapping contigs can be further merged to form scaffolds (30,000–50,000 bases, also called supercontigs), which are uni-directionally oriented along a physical map of a chromosome.
- Overlapping scaffolds are then connected to create the final highest resolution map of the genome.
- Correct identification of overlaps and assembly of the sequence reads into contigs need computational tools.

**Gene Annotation**

- Before the assembled sequence is deposited into a database, it has to be analyzed for useful biological features. The genome annotation process provides comments for the features.
- This involves two steps:

1. gene prediction and
2. functional assignment

**Applications**

• Gene discovery and diagnosis of rare monogenic disorder.

• Gene therapy, Gene Editing, Pharmacogentics and Targeted Therapy

• Analysis of Gene Expression Profile

• Data base of model organism

• Identify and Comparison of new nucleic acid sequencing.

## • Unit 2 - Session 6

## • SLO 6: Sequence databases

### SEQUENCE DATABASES

A database is a computerized archive used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria.

Biological databases can be divided into three categories:

1) Primary database
2) Secondary database
3) Specialized database

#### PRIMARY DATABASE

There are three major public sequence databases that store raw nucleic acid sequence data produced and submitted by researchers worldwide.

Major databases and retrival systems are:

1) **GENBANK**
2) **ENA (European Nucleotide Archive)**
3) **DDBJ (DNA DATA BANK OF JAPAN)**

These three public databases closely collaborate and exchange new data daily. They together constitute the International Nucleotide Sequence Database Collaboration (INSD)

#### GENBANK

- primary nucleotide sequence database in NCBI (National Centre for Biotechnology Information) (USA)
- www.ncbi.nlm.nih.gov/genbank

#### ENA (European Nucleotide Archive)

- primary nucleotide sequence database in Europe
- www.ebi.ac.uk/embl/index.html

#### DDBJ (DNA Data Bank of Japan)

- primary nucleotide sequence database in Japan
- www.ddbj.nig.ac.jp

#### PDB (Protein data Bank)

- Three-dimensional structures of biological macromolecules are stored in PDB.

- This database archives atomic coordinates of macromolecules (both proteins and nucleic acids) determined by x-ray crystallography and NMR.
- It uses a flat file format to represent protein name, authors, experimental details, secondary structure, cofactors, and atomic coordinates.
- The web interface of PDB also provides viewing tools for simple image manipulation.

**SECONDARY DATABASES**

- Secondary databases contain computationally processed sequence information derived from the primary databases.

1) **SWISS-PROT** (**www.ebi.ac.uk/swissprot/access.html)** - which provides detailed sequence annotation of proteins which includes structure, function, domain structure, catalytic sites, cofactor binding, posttranslational modification, metabolic pathway information, disease association.
2) The sequence data are mainly derived from TrEMBL, a database of translated nucleic acid sequences stored in the EMBL database.
3) **PFAM AND BLOCKS DATABASES** - secondary databases containing aligned protein sequence information as well as derived motifs and patterns, which can be used for classification of protein families and inference of protein functions.
4) **DALI DATABASE** - compares protein structures in 3D. It is vital for protein structure classification and threading analysis to identify distant evolutionary relationships among proteins.

**SPECIALIZED DATABASES**

- Specialized databases normally serve a specific research community.
- The sequences in these databases may overlap with a primary database, but may also have new data submitted directly by authors.
- Many genome databases that are taxonomic specific fall, unique organizations and additional annotations within this category.
- Examples of specialized databases include

1) **Flybase -** (A database of the *Drosophila* genome) http://flybase.bio.indiana.edu/
2) **Microarray Gene Expression Database** - (DNA microarray data and analysis tools) www.ebi.ac.uk/microarray
3) **GenBank EST database** (European Bioinformatics Institute)
4) **WormBase**
5) **AceDB,**
6) **TAIR**

**Constraints in the biological database**

1) **Redundancy** includes repeated submission of identical or overlapping sequences by the same or different authors.
2) **Erroneous annotations** - the same gene sequence is found under different names resulting in multiple entries and confusion about the data.

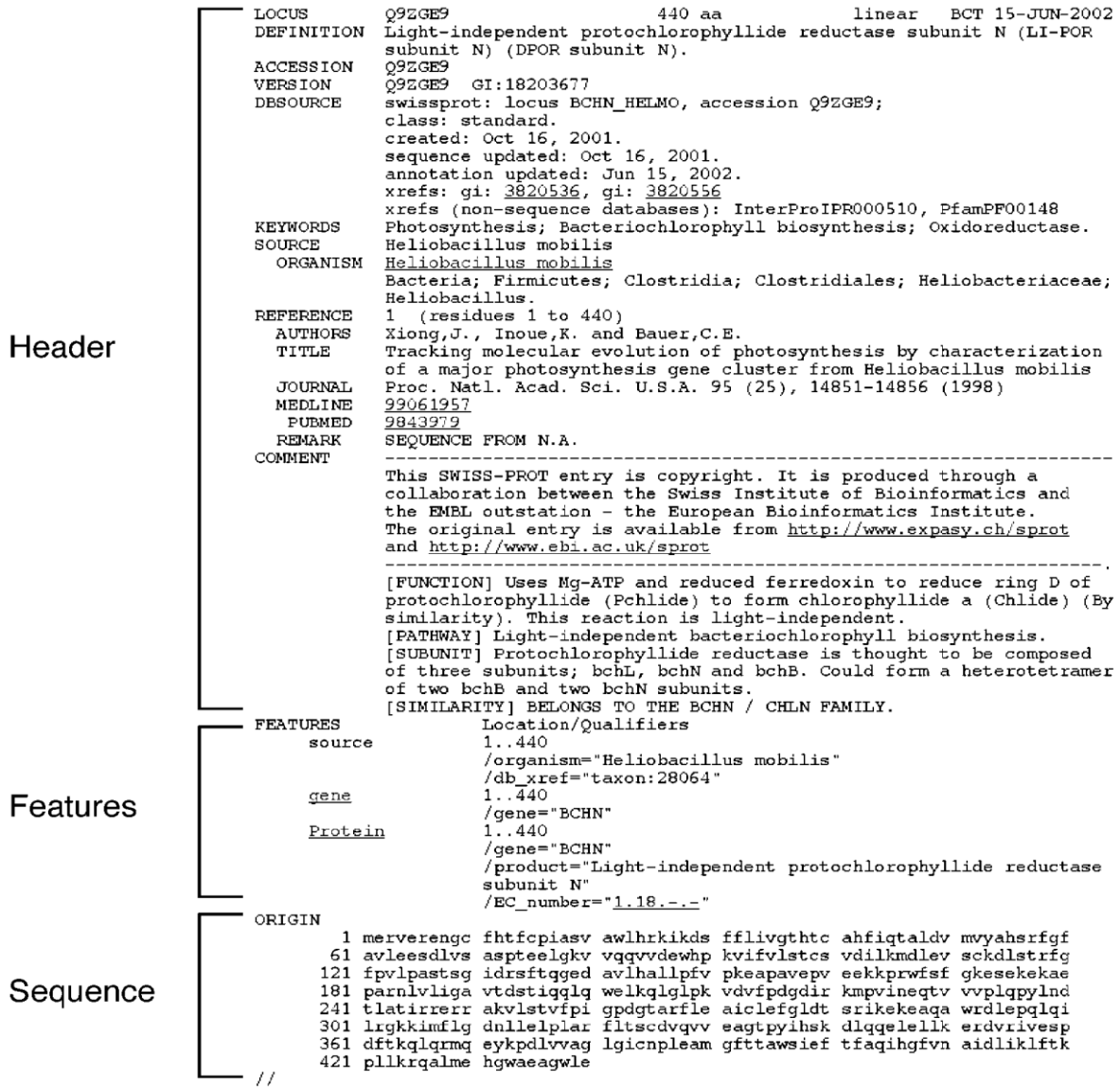
**Information retrivel system from different biological output sources**

**GenBank**

- GenBank is the most complete collection of annotated nucleic acid sequence data for almost every organism.
- The content includes genomic **DNA, mRNA, cDNA, ESTs, high throughput raw sequence data, and sequence polymorphisms.**
- There are two ways to search for sequences in GenBank.
  - **Text-based keywords** - similar to a **PubMed** search.
  - **Molecular sequences** - search by sequence similarity using **BLAST**

**GenBank Sequence Format**

- GenBank is a relational database and the search output for sequence files is produced as flat files for easy reading.
- The resulting flat files contain three sections :
  **Header, Features, and Sequence entry** (Fig).
- Each field has an unique identifier for easy indexing by computer software.
- The **Header** section describes the origin of the sequence, identification of the organism, and unique identifiers associated with the record.
- The top line of the Header section is the **Locus**, which contains a unique database identifier for a sequence location in the database (not a chromosome locus).
- The **identifier** is followed by sequence length and molecule type (e.g., DNA or RNA).
- This is followed by a **three-letter code** for GenBank divisions. There are 17 divisions in total, whichwere set up simply based on convenience of data storage for example, PLN for plant, fungal, and algal sequences; PRI for primate sequences; MAM for nonprimate mammalian sequence
- "**DEFINITION,**" provides the summary information for the sequence record including the name of the sequence, the name and taxonomy of the source organism if known, and whether the sequence is complete or partial.
- **ACCESSION NUMBER** for the sequence, which is a unique number assigned to a piece of DNA when it was first submitted to GenBank and is permanently associated with that sequence.




- The next line in the Header section is the "**ORGANISM**" field, which includes the source of the organism, taxonomic classification, scientific name of the species and tissue type.
- "**REFERENCE**" field, which provides the publication citation, author and title information of the published work.
- The "**JOURNAL"** field includes the citation information as well as the date of sequence submission.
- The "**FEATURES**" section includes annotation information about the gene and gene product,
- . The "Source" field provides the length of the sequence, the scientific name of the organism, and the taxonomy identification number.
- The "gene" field is the information about the nucleotide coding sequence and its name.
- The third section of the flat file is the sequence itself starting with the label "ORIGIN."

**ENTREZ**

- The **NCBI** developed and maintains Entrez, a biological database retrieval system.

- It is a gateway that allows text-based searches for a wide variety of data, including annotated genetic sequence information, structural information, as well as citations and abstracts, full papers, and taxonomic data.
- The key feature of Entrez is its ability to integrate information, which comes from cross-referencing between NCBI databases.
- One of the databases accessible from Entrez is a biomedical literature database known as **PubMed**, which contains abstracts and in some cases the full text articles from nearly 4,000 journals.
- Another unique database accessible from Entrez is **Online Mendelian Inheritance in Man** (OMIM),which is a non-sequence-based database of human disease genes and
  human genetic disorders
- One option is "Limits," which helps to restrict the search to a subset of a particular database. It can also be set to restrict a search to a particular database (e.g., the field for author or publication date) or a particular type of data (e.g., chloroplast DNA/RNA).
- The search can also be limited to a particular search field (e.g., **gene name or accession number)**.
- The "History" option provides a record of the previous searches so that the user can review, revise, or

```
>gi|18203677|sp|Q9ZGE9|BCHN
MERVERENGCFHTFCPIASVAWLHRKIKDSFFLIVGTHTCAHFIQTALDVMVYAHSRFGFAVLEESDLVS
ASPTEELGKVVQQVVDEWHPKVIFVLSTCSVDILKMDLEVSCKDLSTRFGFPVLPASTSGIDRSFTQGED
AVLHALLPFVPKEAPAVEPVEEKKPRWFSFGKESEKEKAEPARNLVLIGAVTDSTIQQLQWELKQLGLPK
VDVFPDGDIRKMPVINEQTVVVPLQPYLNDTLATIRRERRAKVLSTVFPIGPDGTARFLEAICLEFGLDT
SRIKEKEAQAWRDLEPQLQILRGKKIMFLGDNLLELPLARFLTSCDVQVVEAGTPYIHSKDLQQELELLK
ERDVRIVESPDFTKQLQRMQEYKPDLVVAGLGICNPLEAMGFTTAWSIEFTFAQIHGFVNAIDLIKLFTK
PLLKRQALMEHGWAEAGWLE
```

  combine the results of earlier searches.

**Alternative Sequence Formats**

***FASTA***

- FASTA is one of the simplest and the most popular sequence formats because it contains plain sequence information that is readable
- It has a single definition line that begins with a right angle bracket (>) followed by a sequence name
- The plain sequence in standard one-letter symbols starts in the second line.
- Each line of sequence data is limited to sixty to eighty characters in width.
- The drawback of this format is that much annotation information is lost

# Unit 2 - Session 6

## SLO 6: BLAST tool

- Basic Local Alignment Search Tool (BLAST) is an algorithm and program that finds the region of local similarity between the sequences by the comparison of the primary biological sequence information of amino acids found in proteins or the nucleotides of DNA or RNA sequences.
- It is a search program for sequence similarity that can quickly search a sequence database to match with the query sequences.
- It can be used to observe the functional and evolutionary relationship between sequences as well as help identify members of a gene family.

**Four types of BLAST**

1. **BLASTn (Nucleotide BLAST)**

- This tool helps to compare one or more nucleotide sequences to reference sequences or a database of nucleotide sequences.
- It is used during the determination of the evolutionary relationship among different organisms.
- Enter into the NCBI website and search for the Nucleotide BLAST option. After this, add the accession number of the reference sequences and the query sequences. Then, for comparing sequences, find the box called align two or more sequences under the query sequence box. Finally, click on the BLAST options leaving other settings to their default options.

2. **BLASTx (translated nucleotide sequence searched against protein sequence)**

- BLASTx compares a nucleotide query sequence that is translated into protein sequences against the protein sequence database.
- BLASTx is particularly important when the reading frame of the query sequence is unknown, or it contains errors that may lead to frameshift or other coding errors because it translates the query sequence in all six reading frames and provides a combined statistical significance for hits to different frames.
- With a newly determined sequence, BLASTx is often the first analysis that is performed.

3. **tBLASTn (protein sequence searched against translated nucleotide sequences)**

- A query protein sequence is compared against the six-frame translation of a database of nucleotide sequences in this type of BLAST search tool.

- Homologous protein-coding regions in unannotated nucleotide sequences like expressed sequence tags (ESTs) and draft genome records (HTG), located in the BLAST database est and htgs, respectively, can be found by tBLASTn.
- A short single-read cDNA sequence is called ESTs, which consists of the largest pool of the sequenced data for many organisms and also consists of proportions of transcripts from many uncharacterized genes.

4. **BLASTp**

- One or more protein sequences are compared to subject protein sequences or a database of protein sequences by this type of BLAST search.
- It is used for the identification of protein sequences.

**BLAST score**

- Once a similar sequence has been found for the query sequence in the database through BLAST, then it becomes essential to have the idea of whether the alignment is good or whether it shows the possible biological relationships or not. So BLAST uses statistical theory to produce a bit score for each alignment pair.
- The indication of the good alignment is given by the bit score, which shows the higher the scores, the better the alignments.
- Generally, this score is calculated by taking into consideration the alignment of the similar or identical residues and the gaps introduced while aligning the sequences.
- It uses the "substitution matrix" for the alignment of any possible residues.
- For most of the BLAST programs, the BLOSUM62 matrix is the default with the exception of BLASTn and MegaBLAST as these are the programs that perform nucleotide-nucleotide comparisons and do not use protein-specific matrices.

**BLAST E-Value**

- E-value is the statistical theory used in the BLAST for the alignment of each pair of sequences and provides the idea of whether the alignment is good or not and whether the two sequences match with it or not.
- The number of expected hits of similar quality (score) that could be found just by chance is the BLAST E-value and the E-value of 10 means that up to 10 hits can be expected to be found by chance.
- The E-value provides the information about the likelihood that a given sequence match is purely by chance and is used as the first quality filter for the BLAST search result.
- The lower the E-value the better the match which means if E is less than 1e-50, then there is high confidence that the database match is a result of homologous relationships.

- If the value of E is between 0.01 and 10 then the match is considered to be non-significant but may have a weak homology relationship.
- Similarly, if the value of E is greater than 10, then the sequence under consideration is either unrelated or if related then has an extremely distant relationship.

**Application**

- **DNA mapping**: BLAST helps in the identification and mapping of the gene between the known and unknown species.
- **Domains location:** BLAST helps to identify and locate the domains in the protein sequences of interest.
- **Comparison:** while comparing the sequences between the two different or similar species, BLAST is used and thus it helps to identify the similar genes present or the functions between the species.
- **Identification of species:** BLAST can also be used for the identification of the species by the comparison of the sequences of the DNA between the different organisms.
- **Establishing phylogeny:** after the alignment of the sequences using BLAST, one can identify the result and observe the phylogenetic relationships between the known and unknown species.