BTY 102 BIOMOLECULES

UNIT 1

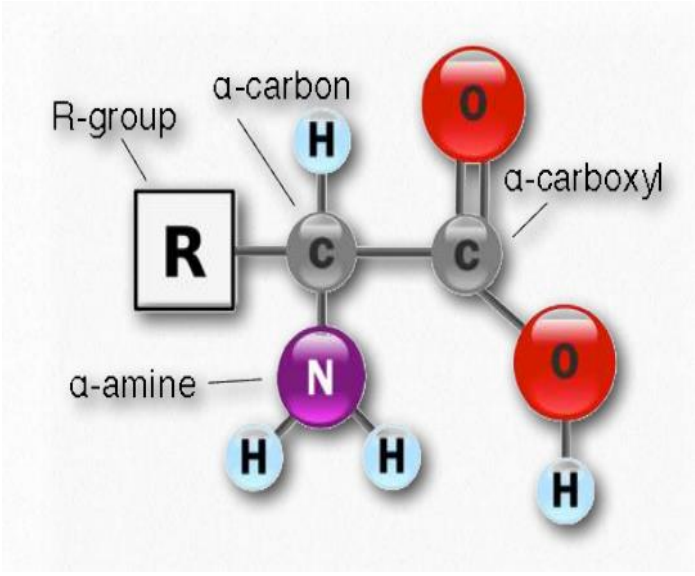
Classification of Amino acids:

All of the proteins on earth are made up of the same 20 amino acids. Linked together in long chains called polypeptides, amino acids are the building blocks for the vast assortment of proteins found in all living cells.

It is one of the more striking generalizations of biochemistry …that the twenty amino acids and the four bases, are, with minor reservations, the same throughout Nature." – Francis Crick

All amino acids have the same basic structure, shown in Figure 2.1. At the center of each amino acid is a carbon called the α carbon and attached to it are four groups – a hydrogen, a carboxylic acid group, an amine group, and an R-group, sometimes referred to as a variable group or side chain. The α carbon, carboxylic acid, and amino groups are common to all amino acids, so the R-group is the only variable feature. With the exception of glycine, which has an R-group consisting of a hydrogen atom, all of the amino acids in proteins have four different groups attached to them and consequently can exist in two mirror isomeric forms.

The designations used in organic chemistry are not generally applied to amino acid nomenclature, but a similar system uses L and D to describe these enantiomers. Nature has not distributed the stereoisomers of amino acids equally. Instead, with only very minor exceptions, every amino acid found in cells and in proteins is in the L configuration.



There are 22 amino acids that are found in proteins and of these, only 20 are specified by the universal genetic code. The others, selenocysteine and pyrrolysine use tRNAs that are able to base pair with stop codons in the mRNA during translation. When this happens, these unusual amino acids can be incorporated into proteins. Enzymes containing selenocysteine, for example, include glutathione peroxidases, tetraiodothyronine 5′ deiodinases, thioredoxin reductases, formate dehydrogenases, glycine reductases, and selenophosphate synthetase. Pyrrolysine-containing proteins are much rarer and are mostly confined to archaea.

Essential and non-essential

Nutritionists divide amino acids into two groups – essential amino acids and non-essential amino acids. Essential amino acids must be included in our diet because our cells can't synthesize them. What is essential varies considerably from one organism to another and even differ in humans, depending on whether they are adults or children. Table 2.1 shows essential and non-essential amino acids in humans.

Some amino acids that are normally nonessential, may need to be obtained from the diet in certain cases. Individuals who do not synthesize sufficient amounts of arginine, cysteine, glutamine, proline, selenocysteine, serine, and tyrosine, due to illness, for example, may need dietary supplements containing these amino acids.

Essential	Non-Essential
Histidine	Alanine
Isoleucine	Arginine
Leucine	Asparagine
Lysine	Aspartic acid
Methionine	Cysteine
Phenylalanine	Glutamic acid
Threonine	Glutamine
Tryptophan	Glycine
Valine	Proline
	Selenocysteine
	Serine
	Tyrosine

Table 2.1 - Essential and non-essential amino acids

Non-protein amino acids

There are also amino acids found in cells that are not incorporated into proteins. Common examples include ornithine and citrulline. Both of these compounds are intermediates in the urea cycle, an important metabolic pathway.

R-group chemistry

Non-Polar	Carboxyl	Amine	Aromatic	Hydroxyl	Other
Alanine	Aspartic Acid	Arginine	Phenylalanine	Serine	Asparagine
Glycine	Glutamic Acid	Histidine	Tryptophan	Threonine	Cysteine
Isoleucine		Lysine	Tyrosine	Tyrosine	Glutamine
Leucine					Selenocysteine
Methionine					Pyrrolysine
Proline					
Valine					

Amino acids can be classified based on the chemistry of their R-groups. It is useful to classify amino acids in this way because it is these side chains that give each amino acid its characteristic properties. Thus, amino acids with (chemically) similar side groups can be expected to function in similar ways, for example, during protein folding. The specific category divisions may vary, but all systems are attempts to organize and understand the relationship between an amino acid's structure and its properties or behavior as part of a larger system.

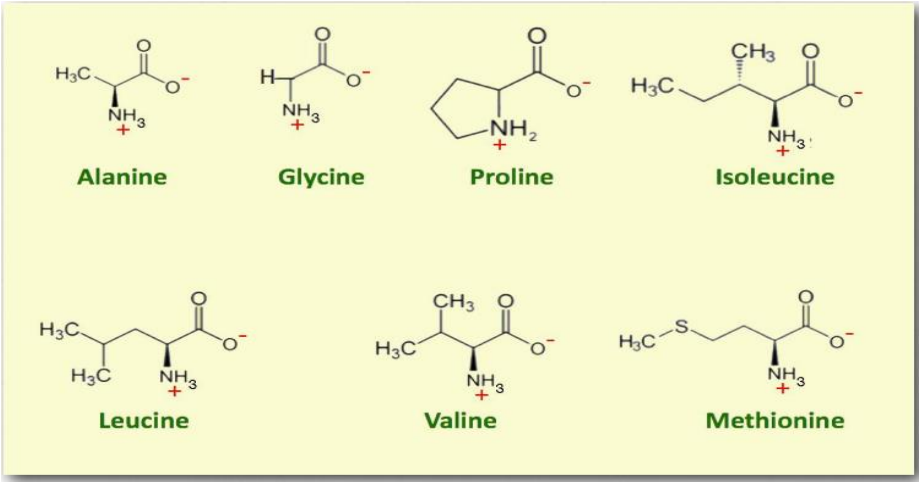
Non-polar amino acids

- Amino acids in this group include:
- Alanine (Ala/A)
- Glycine (Gly/G)
- Isoleucine (Ile/I)
- Leucine (Leu/L)
- Methionine (Met/M)
- Valine (Val/V)

The amino acids in this group have nonpolar, hydrophobic R groups. When incorporated into globular proteins they tend to pack inward among other hydrophobic groups. In proteins that embed themselves into or through membranes, these amino acids can orient themselves toward hydrophobic portions of the inside of the membrane.

The small R groups here are more readily packed into tight formations. Proline is exceptional in that it has an R group that folds back and covalently bonds to the backbone of the amino acid, creating a more rigid element in a protein chain that reduces free movement of the polypeptide chain. Additionally, proline can undergo hydroxylation reactions, stabilizing the protein structure. This occurs in collagen with the aid of ascorbic acid (Vitamin C). One symptom of the vitamin C deficiency syndrome 'scurvy' is the reduced quality of collagen in tissues, including the skin and gums. This can lead to the deterioration and loss of teeth.

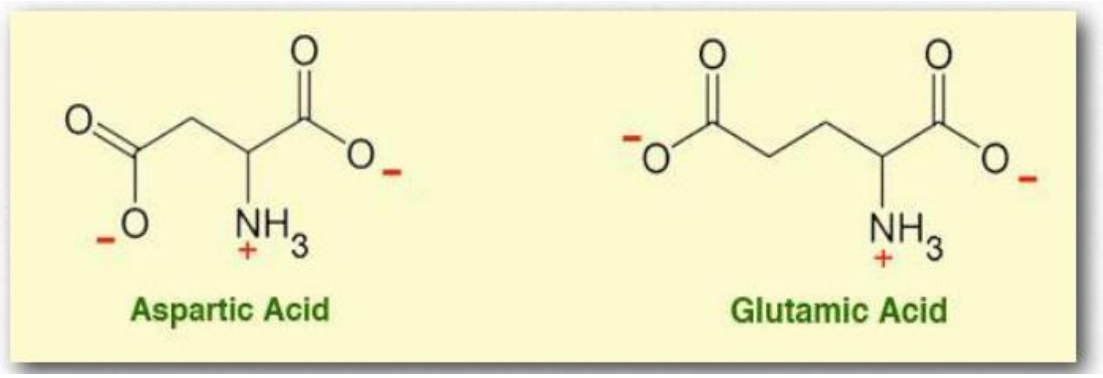
Amino acid	Short	Abbrev.	Side chain	Hydro-phobic	pKa	Polar	pH	Small	Tiny	Aromatic or Aliphatic	van der Waals volume
Alanine	A	Ala	$-CH_3$	X	-	-	-	X	X	-	67
Cysteine	C	Cys	$-CH_2SH$	-	8.18	-	acidic	X	X	-	86
Aspartic acid	D	Asp	$-CH_2COOH$	-	3.90	X	acidic	X	-	-	91
Glutamic acid	E	Glu	$-CH_2CH_2COOH$	-	4.07	X	acidic	-	-	-	109
Phenylalanine	F	Phe	$-CH_2C_6H_5$	X	-	-	-	-	-	Aromatic	135
Glycine	G	Gly	-H	X	-	-	-	X	X	-	48
Histidine	H	His	$-CH_2-C_3H_3N_2$	-	6.04	X	weak basic	-	-	Aromatic	118
Isoleucine	I	Ile	$-CH(CH_3)CH_2CH_3$	X	-	-	-	-	-	Aliphatic	124
Lysine	K	Lys	$-(CH_2)_4NH_2$	-	10.54	X	basic	-	-	-	135
Leucine	L	Leu	$-CH_2CH(CH_3)_2$	X	-	-	-	-	-	Aliphatic	124
Methionine	M	Met	$-CH_2CH_2SCH_3$	X	-	-	-	-	-	-	124
Asparagine	N	Asn	$-CH_2CONH_2$	-	-	X	-	X	-	-	96
Pyrrolysine	O	Pyl	$-(CH_2)_4NHCOC_4H_5NCH_3$	-	-	X	weak basic	-	-	-	
Proline	P	Pro	$-CH_2CH_2CH_2-$	X	-	-	-	X	-	-	90
Glutamine	Q	Gln	$-CH_2CH_2CONH_2$	-	-	X	weak basic	-	-	-	114
Arginine	R	Arg	$-(CH_2)_3NH-C(NH)NH_2$	-	12.48	X	strongly basic	-	-	-	148
Serine	S	Ser	$-CH_2OH$	-	5.68	X	weak acidic	X	X	-	73
Threonine	T	Thr	$-CH(OH)CH_3$	-	5.53	X	weak acidic	X	-	-	93
Selenocysteine	U	Sec	$-CH_2SeH$	-	5.73	-	acidic	X	X	-	
Valine	V	Val	$-CH(CH_3)_2$	X	-	-	-	X	-	Aliphatic	105
Tryptophan	W	Trp	$-CH_2C_8H_6N$	-	5.885	X	weak basic	-	-	Aromatic	163
Tyrosine	Y	Tyr	$-CH_2-C_6H_4OH$	-	10.46	X	weak acidic	-	-	Aromatic	141



Acidic Amino Acids (Carboxylic acid side chains)

Amino acids in this group include:

- Aspartic acid (Asp/D)
- Glutamic acid (Glu/E)



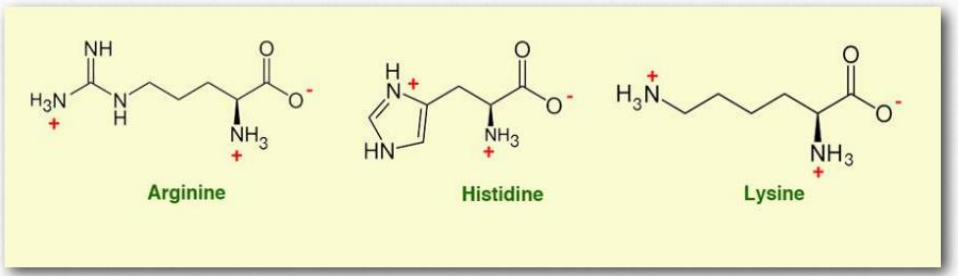
These amino acids each contain a carboxylic acid group as part of the variable group. At physiological pH, these groups exist primarily in their deprotonated state. It is easy to be confused if they are drawn in this state, because their names include "acid" while the structure shows no ionizable proton and the charge on the R group is negative.

In addition to its role as a building block in proteins, glutamic acid (with the deprotonated form named "glutamate") is a neurotransmitter. It also is recognized by a receptor in our mouths, contributing to a taste sensation described as "umami." Many foods contain appreciable amounts of glutamate that are recognized by our taste receptors, and encourage us to eat these substances. Those foods frequently contain protein that has broken down to some degree: cooked meats, fermented sauces like Worcestershire or soy, tahini, broths, and yeast extracts.

Basic amino acids (Nitrogen-containing side chains)

Included in this group of amino acids are:

- Arginine (Arg/R)
- Histidine (His/H)
- Lysine (Lys/K)



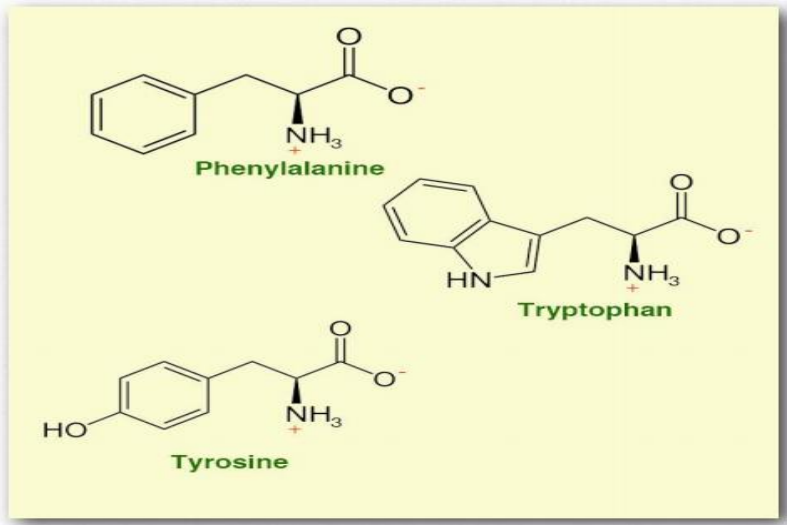
The variable group in each of these amino acids contains nitrogen, which imparts to the group the ability to exist in protonated and deprotonated states. They are frequently called basic, but also are often drawn in their protonated state which is more prevalent at physiological pH.

Arginine (Arg/R) is interesting due to the fact it is an essential dietary amino acid for premature infants, who cannot synthesize it. In addition, surgical trauma, sepsis, and burns increase demand for arginine and proper healing can require dietary intake.

Histidine contains a nitrogen-containing imidazole functional group that has a pKa of 6. This means it can pick up or donate hydrogen ions in response to small changes in pH. In proteins, histidine frequently has important roles participating directly in reactions involving hydrogen ion transfer.

The R group on lysine is frequently chemically modified in order for it to make unusual linkages to other chemical groups or to take part in specific chemical reactions. Lysine is often added to animal feed because it is a limiting amino acid and is necessary for optimizing growth of animals raised for consumption.

Aromatic amino acids



Amino acids with aromatic side chains include:

- Phenylalanine (Phe/ F)
- Tryptophan (Trp/W)
- Tyrosine (Tyr/Y)

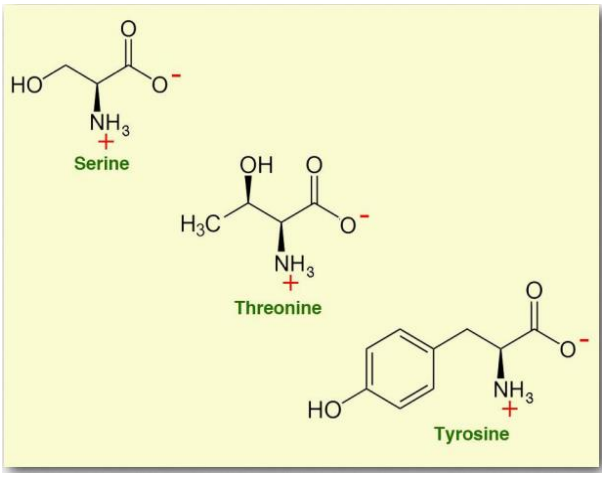
These amino acids are included in protein structures but also serve as precursors in some important biochemical pathways, leading to the production of hormones such as L-Dopa and serotonin.

Hydroxyl amino acids

This group includes

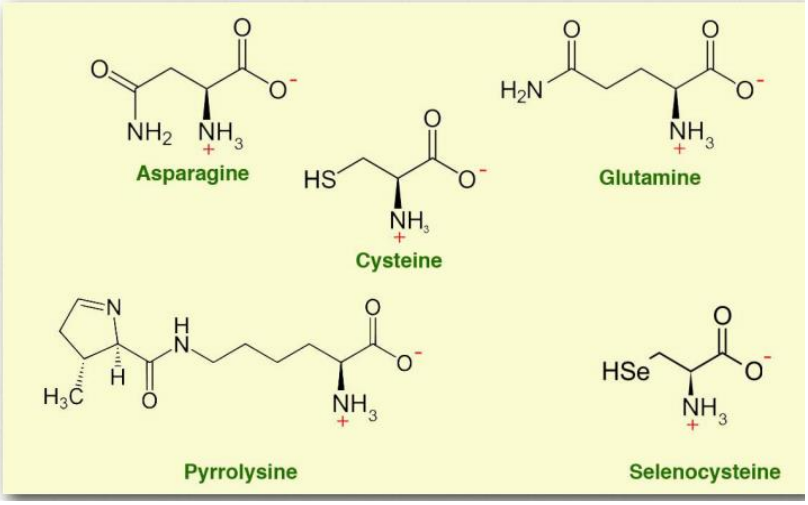
- Threonine (Thr/T)
- Serine (Ser/S)
- Tyrosine (already discussed as an aromatic amino acid)

The amino acids in this group contain alcohol groups, which can engage in hydrogen-bonding interactions. As part of protein molecules they are hydrophilic and can be oriented outward in watery environments. The alcohol group is subject to chemical reactions or modifications, for instance when carbohydrate groups are covalently linked to proteins.



Other amino acids

- Asparagine (Asn/N) is a polar amino acid. The amide on the functional group is not basic.
- Cysteine (Cys/C)
- Glutamine (Gln/Q)

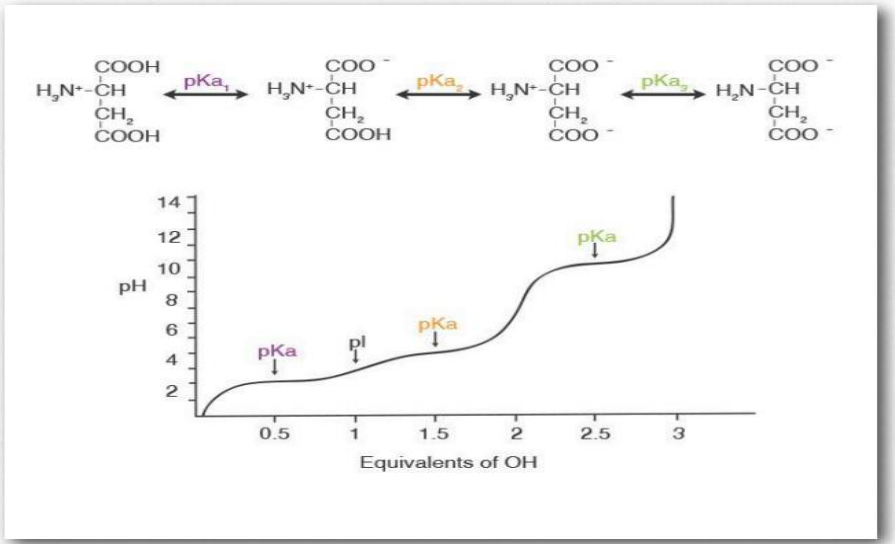


Cysteine, which contains a thiol. Thiols can react with one another via oxidation, forming disulfide links containing two covalently-linked sulfur atoms. Variable groups on methionine in protein chains can undergo such reactions, covalently tying the chains to one another with a short tether. Such disulfide links or bridges restrict the mobility of protein chains and contribute to more defined structures.

- Selenocysteine (Sec/U) is a component of selenoproteins found in all kingdoms of life. Twenty five human proteins contain selenocysteine. It is a component in several enzymes, including glutathione peroxidases and thioredoxin reductases. It is not coded for by the standard genetic code.
- Pyrrolysine (Pyl/O) is a twenty second amino acid, but is rarely found in proteins. Like selenocysteine, it is not coded for in the genetic code and must be incorporated by unusual means.

Ionizing groups

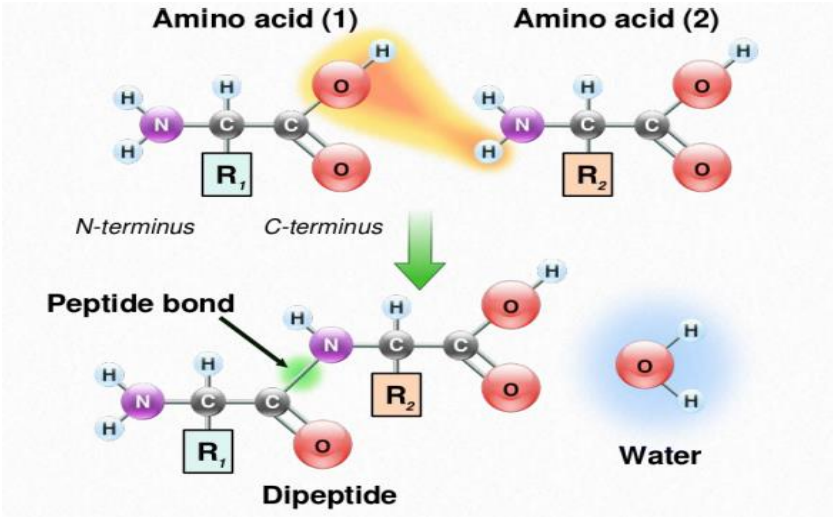
Some, but not all amino acids have R-groups that can ionize. The charge of a protein then arises from the charges of the amine group, the carboxyl group, and the sum of the charges of the ionized R-groups. Titration/ionization of aspartic acid is depicted in Figure 2.10. Ionization (or deionization) within a protein's structure can have significant effect on the overall conformation of the protein and, since structure is related to function, a major impact on the activity of a protein.



Building Polypeptides

Although amino acids serve other functions in cells, their most important role is as constituents of proteins. Proteins, as we noted earlier, are polymers of amino acids.

Amino acids are linked to each other by peptide bonds, in which the carboxyl group of one amino acid is joined to the amino group of the next, with the loss of a molecule of water. Additional amino acids are added in the same way, by formation of peptide bonds between the free carboxyl on the end of the growing chain and the amino group of the next amino acid in the sequence. A chain made up of just a few amino acids linked together is called an oligopeptide (oligo=few) while a typical protein, which is made up of many amino acids is called a polypeptide (poly=many). The end of the peptide that has a free amino group is called the N-terminus (for NH2), while the end with the free carboxyl is termed the C-terminus (for carboxyl).



As we've noted before, function is dependent on structure, and the string of amino acids must fold into a specific 3-D shape, or conformation, in order to make a functional protein. The folding of polypeptides into their functional forms is the topic of the next section.

Classification of Proteins:

Proteins fold into stable three-dimensional shapes, or conformations, that are determined by their amino acid sequence. The complete structure of a protein can be described at four different levels of complexity: primary, secondary, tertiary, and quaternary structure.

As a multitude of protein structures are rapidly being determined by X-ray crystallography and by nuclear magnetic resonance (NMR), it is becoming clear that the number of unique folds is far less than the total number of protein structures. Not only do functionally related proteins generally have similar tertiary structures (see below), but even proteins with very different functions are often found to share the same tertiary folds. As a consequence, structural conservation at the tertiary level is perhaps more profound than it is at the primary. The identification of the fold of a protein has therefore become an invaluable tool since it can potentially provide a direct extrapolation to function, and may allow one to map functionally important regions in the amino acid sequence.

Several groups have already attempted to classify protein structures into fold families and superfamiles without focusing on function (Orengo et al., 1993; Murzin et al., 1995). The scope of this unit is not to enumerate all the existing folds and tertiary structures determined to date, but rather to provide a comprehensive overview of some commonly observed protein fold families and commonly observed structural motifs which have functional significance. Likewise, the PDB-entry tables given in this unit provide some examples of various folds, but are not comprehensive lists. The unit is organized into sections based on both structural and functional relations.

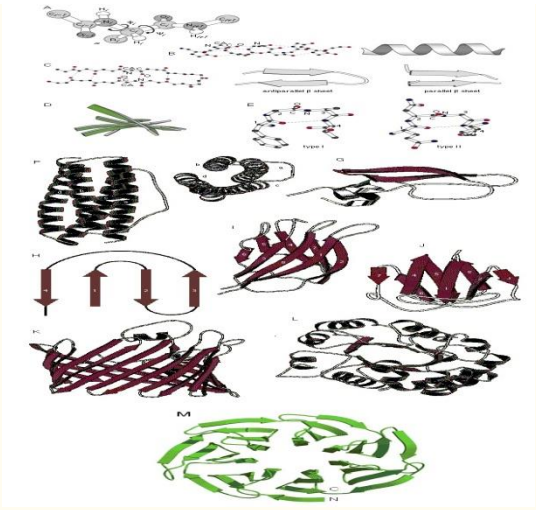
Primary Structure

Primary structure is defined as the linear amino acid sequence of a protein's polypeptide chain. In fact, the term **protein sequence** is often used interchangeably with primary structure. In 1973, Chris Anfinsen demonstrated that the primary amino sequence of a protein uniquely determines the higher orders of structure for a protein and is thus of fundamental importance (Anfinsen, 1973). It is noteworthy, however, that changes in the local biological environment of a protein molecule can sometimes perturb its

three-dimensional structure. For example, interactions with ligands, substrates, or other proteins can bring about controlled conformational changes producing potentially profound effects. Furthermore, a few proteins have been found to have intrinsically unstructured regions (Wright and Dyson, 1999; Tompa, 2002). Hence, although structural uniqueness associated with a protein sequence is a powerful principle, it can sometimes depend on the local environment and is not rigidly followed in every case.

Secondary Structure

Secondary structure is defined as the local spatial conformation of the polypeptide backbone excluding the side chains. Regular secondary structures (also referred to as **secondary structure elements**) common to many proteins include α-helices, β-sheets, and turns (see below). They can vary widely in length, from as few as three to five residues in short helices and sheets, to over fifty residues in some coiled-coil helices (see Frequently Observed Secondary Structure Assemblies or Structural Motifs). Such structures are generally defined both by characteristic main chain φ and ψ dihedral angles (i.e., the torsion angles between backbone atoms C_{i-1}–N_i–Ciα–C_i and N_i–Ciα–C_i–N_{i+1}, respectively; Fig. Fig.1A)1A) and by regular main chain hydrogen bonding patterns. When discussing the structure of a protein, the term **topology** is often used to refer to the connectivity of secondary structure elements. For example, a portion of a protein containing a β-strand connected to an α-helix and then another β-strand is said to have a βαβ topology. The complete secondary structure of a polypeptide chain is often represented in two dimensions by a topology diagram (e.g., Fig. Fig.1H)1H) which shows both the connectivity and the relative orientation of neighboring secondary structure elements. Such diagrams are particularly useful in classifying β-sheets.



α-Helices

An α-helix is formed when a protein backbone adopts a right-handed helical conformation with 3.6 residues per turn and a set of hydrogen bonds formed between the main chain carbonyl (CO) of the i^{th} residue and the main chain NH of the $(i+4)^{th}$ residue (Fig. (Fig.1B).1B). Occasionally, hydrogen bonds are observed between residues i and i+3, resulting in a 3_{10} helix. Compared to the more common α-helices, 3_{10} helices are much shorter, usually comprising only one to three turns. The optimal backbone φ and ψ dihedral angles for a right-handed α-helix are −57° and −47°, while for a 3_{10} helix they and −49° and −26°, respectively.

β-Sheets

β-sheets (also referred to as β-pleated sheets) make up another major secondary structural element in proteins. A β-sheet consists of at least two β-strands, each approaching an extended backbone conformation with dihedral angles confined to the region where the φ torsion angle is between −60° and −180°, and the ψ torsion angle is between 30° and 180°. All main-chain CO-to-NH hydrogen bonds lie between adjacent strands . A parallel β-sheet is formed when all the sheet-forming strands run parallel to each other and in the same direction from N to C termini, whereas an antiparallel β-sheet is formed when the strands are still parallel to each other, but run in opposite directions a β-sheet formed from both parallel and antiparallel strands in referred to as a mixed β-sheet. A characteristic feature of β-sheets is the right handed twist which is visible when the sheet is viewed edge-on

Turns and Loops

Loops and turns connect helices and β-sheets in protein structures. Most turns and loops assume irregular secondary structures, but important exceptions are the type I and II reverse turns (also referred to as β-turns or β-bends), which are tight turns connecting adjacent, antiparallel β-strands. Shown in Figure Figure1E,1E, type I and II turns each have a hydrogen bond between the CO of the first residue in a turn and the NH of the fourth residue (also the last residue) in the turn. The difference between the type I and II turn is a 180° flip of the peptide plane in the middle of the turn. The type I turn is more frequently observed than the type II. The terms loops and coils generally refer to secondary structure elements that display less regular hydrogen bonding patterns than those observed in α-helices, β-sheets, and reverse turns.

Tertiary Structure

Tertiary structure refers to the three-dimensional arrangement of all the atoms that constitute a protein molecule. It relates the precise spatial coordination of secondary structure elements and the location of all functional groups of a single polypeptide chain.

Domains, Folds, and Motifs

The tertiary structure of a protein can often be divided into domains—i.e., distinct compact folding units usually comprising 100 to 200 residues. Small proteins may contain a single domain, whereas larger proteins often contain multiple domains. A fold refers to a characteristic spatial assembly of secondary structure elements into a domain-like structure that is common to many different proteins. Particular folds are often, though not always, related to a certain function (e.g., nucleotide-binding folds). A structural motif is similar to a fold, but is generally smaller, tending to form the building blocks of folds. Some structural motifs (e.g., β-barrels) are observed in a vast array of unrelated proteins with many variations, while others, especially those related to a unique biochemical function (e.g., zinc fingers; see Zinc-Containing DNA-Binding Motifs), are highly conserved, generally occurring in protein domains of similar function. Highly conserved structural motifs often display a characteristic signature at their amino acid sequence level, the so-called **sequence motif**. Very often the term **motif** is used alone to refer to either a **structural motif** or a **sequence motif**. This can lead to confusion since structural motifs do not always have a unique amino acid signature. It is important to note that the terms **domain**, **fold**, and **motif** are often used interchangeably with blurred distinctions. This happens especially for larger, highly conserved structural motifs which may be as equally well known as a particular type of domain or fold.

Quaternary Structure

The structure of many proteins, especially those >100 kDa in mass, are oligomers consisting of more than one polypeptide chain. The precise spatial arrangement of the subunits within a protein is referred to as the quaternary structure.

UNIT – II

Classification Of Carbohydrates:

- Carbohydrates are defined as biomolecules containing a group of naturally occurring carbonyl compounds (aldehydes or ketones) and several hydroxyl groups.
- It consists of carbon (C), hydrogen (H), and oxygen (O) atoms, usually with a hydrogen-oxygen atom ratio of 2:1 (as in water). It's represented with the empirical formula Cm(H2O)n (where m and n may or may not be different) or (CH2O)n.
- But some compounds do not follow this precise stoichiometric definition, such as uronic acids. And there are others that, despite having groups similar to carbohydrates, are not classified as one of them, e.g., formaldehyde and acetic acid.

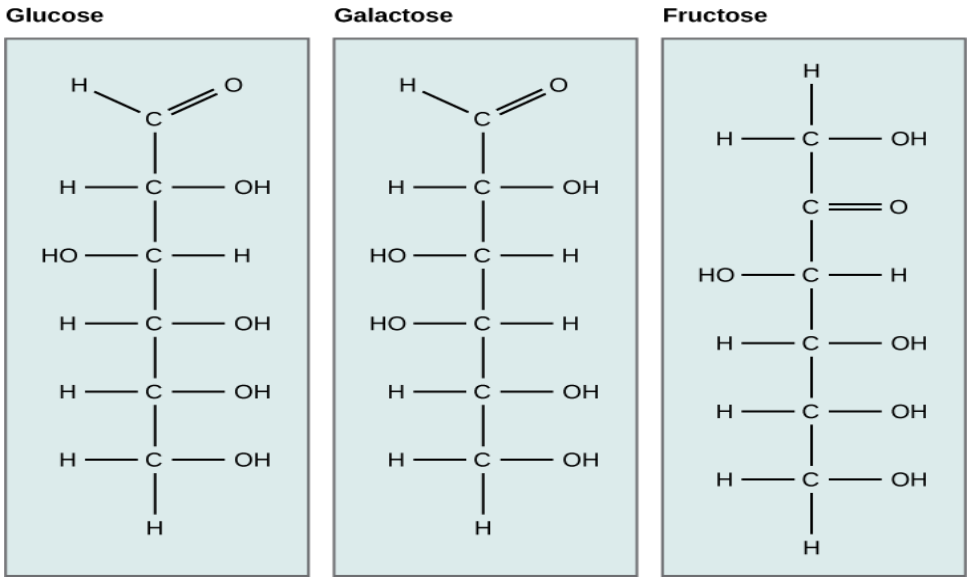
Classification of Carbohydrates

Carbohydrates are divided into four major groups based on the degree of polymerization: monosaccharides, disaccharides, oligosaccharides, and polysaccharides. Given below is a brief account of the structure and functions of carbohydrate groups.

➢ Monosaccharides

- Monosaccharides are the simplest carbohydrates and cannot be hydrolyzed into other smaller carbohydrates. The "mono" in monosaccharides means one, which shows the presence of only one sugar unit.
- They are the building blocks of disaccharides and polysaccharides. For this reason, they are also known as simple sugars. These simple sugars are colorless, crystalline solids that are soluble in water and insoluble in a nonpolar solvent.
- The general formula representing monosaccharide structure is $Cn(H_2O)n$ or CnH_2nOn. Dihydroxyacetone and D- and L-glyceraldehydes are the smallest monosaccharides – here, n=3.
- The monosaccharides containing the aldehyde group (the functional group with the structure, R-CHO) are known as aldolases and the one containing ketone groups is called ketoses (the functional group with the structure RC(=O)R′). Some examples of monosaccharides are glucose, fructose, erythrulose, and ribulose.
- D-glucose is the most common, widely distributed, and abundant carbohydrate. It's commonly known as dextrose and it's an aldehyde containing six carbon atoms, called aldohexose. It's present in both, open-chain and cyclic structures.

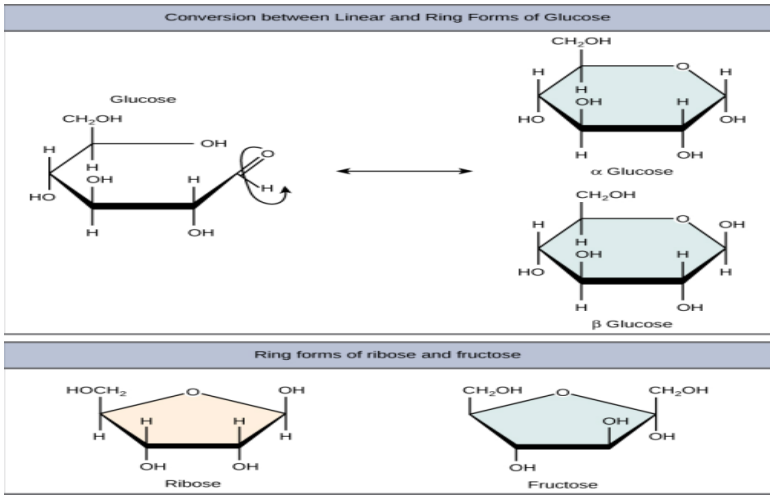
- Most monosaccharide names end with the suffix -ose. And based on the number of carbons, which typically ranges from three to seven, they may be known as trioses (three carbons), tetroses (four carbons), pentoses (five carbons), hexoses (six carbons), and heptoses (seven carbons).
- Although glucose, galactose, and fructose all have the chemical formula of $C_6H_{12}O_6$, they differ at the structural and chemical levels because of the different arrangement of functional groups around their asymmetric carbon.



.Structure of Monosaccharides

Monosaccharides are either present as linear chains or ring-shaped molecules. In a ring form, glucose's hydroxyl group (-OH) can have two different arrangements around the anomeric carbon (carbon-1 that becomes asymmetric in the process of ring formation).

If the hydroxyl group is below carbon number 1 in the sugar, it is said to be in the alpha (α) position, and if it is above the plane, it is said to be in the beta (β) position.

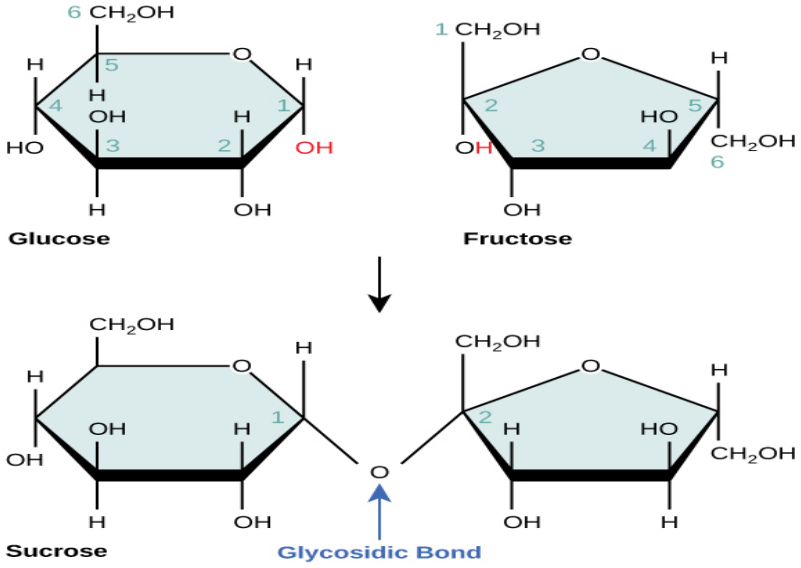


Functions of Monosaccharides

- Glucose ($C_6H_{12}O_6$) is an important source of energy in humans and plants. Plants synthesize glucose using carbon dioxide and water, which in turn is used for their energy requirements. They store the excess glucose as starch which humans and herbivores consume.
- The presence of galactose is in milk sugar (lactose), and fructose in fruits and honey makes these foods sweet.
- Ribose is a structural element of nucleic acids and some coenzymes.
- Mannose is a constituent of mucoproteins and glycoproteins required for the proper functioning of the body.

2. Disaccharides

- Disaccharides consist of two sugar units. When subjected to a dehydration reaction (condensation reaction or dehydration synthesis), they release two monosaccharide units.
- In this process, the hydroxyl group of one monosaccharide combines with the hydrogen of another monosaccharide through a covalent bond, releasing a molecule of water. The covalent bond formed between the two sugar molecules is known as a **glycosidic bond**.
- The glycosidic bond or glycosidic linkage can be alpha or beta type. The alpha bond is formed when the OH group on the carbon-1 of the first glucose is below the ring plane, and a beta bond is formed when the OH group on the carbon-1 is above the ring plane.



Some examples of disaccharides are lactose, maltose, and sucrose. Sucrose is the most abundant disaccharide of all and is composed of one D-glucose molecule and one D-fructose molecule. The systematic name for sucrose is O-α-D-glucopyranosyl-(1→2)-D-fructofuranoside.

Lactose occurs naturally in mammalian milk and is composed of one D-galactose molecule and one D-glucose molecule. The systematic name for lactose is O-β-D-galactopyranosyl-(1→4)-D-glucopyranose.

Disaccharides can be classified into two groups based on their ability to undergo oxidation-reduction reactions.

- **Reducing sugar:** A disaccharide in which the reducing sugar has a free hemiacetal unit serving as a reducing aldehyde group. Examples include maltose and cellobiose.
- **Non-reducing Sugar:** Disaccharides that do not have a free hemiacetal because they bond through an acetal linkage between their anomeric centers. Examples are sucrose and trehalose.

Some other examples of disaccharides include lactulose, chitobiose, kojibiose, nigerose, isomaltose, sophorose, laminaribiose, gentiobiose, turanose, maltulose, trehalose, palatinose, gentiobiulose, mannobiose, melibiose, melibiulose, rutinose, rutinulose, and xylobiose.

A list of disaccharides with their monomer units is given below:

Disaccharide	**Monomer Units**
Sucrose	Glucose and Fructose
Lactose	Galactose and Glucose
Maltose	Glucose and Glucose (alpha-1,4 linkage)

Disaccharide	Monomer Units
Trehalose	Glucose and Glucose (alpha-1, alpha-1 linkage)
Cellobiose	Glucose and Glucose (beta-1,4 linkage)
Gentiobiose	Glucose and Glucose (beta-1,6 linkage)

Functions of Disaccharides

- Sucrose is a product of photosynthesis, which functions as a major source of carbon and energy in plants.
- Lactose is a major source of energy in animals.
- Maltose is an important intermediate in starch and glycogen digestion.
- Trehalose is an essential energy source for insects.
- Cellobiose is essential in carbohydrate metabolism.
- Gentiobiose is a constituent of plant glycosides and some polysaccharides.

3. Oligosaccharides

Oligosaccharides are compounds that yield 3 to 10 molecules of the same or different monosaccharides on hydrolysis. All the monosaccharides are joined through glycosidic linkage. And based on the number of monosaccharides attached, the oligosaccharides are classified as trisaccharides, tetrasaccharides, pentasaccharides, and so on.

The general formula of trisaccharides is $Cn(H_2O)n-2$, and that of tetrasaccharides is $Cn(H_2O)n-3$, and so on. The oligosaccharides are normally present as glycans. They are linked to either lipids or amino acid side chains in proteins by N- or O-glycosidic bonds known as glycolipids or glycoproteins.

The glycosidic bonds are formed in the process of glycosylation, in which a carbohydrate is covalently attached to an organic molecule, creating structures such as glycoproteins and glycolipids.

- **N-Linked Oligosaccharides:** It involves the attachment of oligosaccharides to asparagine via a beta linkage to the amine nitrogen of the side chain. In eukaryotes, this process occurs at the membrane of the endoplasmic reticulum. Whereas in prokaryotes, it occurs at the plasma membrane.
- **O-Linked Oligosaccharides:** It involves the attachment of oligosaccharides to threonine or serine on the hydroxyl group of the side chain. It occurs in the Golgi apparatus, where monosaccharide units are added to a complete polypeptide chain.

Functions of Oligosaccharides

- Glycoproteins are carbohydrates attached to proteins involved in critical functions such as antigenicity, solubility, and resistance to proteases. Glycoproteins are relevant as cell-surface receptors, cell-adhesion molecules, immunoglobulins, and tumor antigens.
- Glycolipids are carbohydrates attached to lipids that are important for cell recognition and modulate membrane proteins that act as receptors.
- Cells produce specific carbohydrate-binding proteins known as lectins, which mediate cell adhesion with oligosaccharides.
- Oligosaccharides are a component of fiber from plant tissues.

4. Polysaccharides

Polysaccharides are a chain of more than 10 carbohydrates joined together through glycosidic bond formation. They are ubiquitous and mainly involved in the structural or storage functions of organisms. They are also known as glycans.

These compounds' physical and biological properties depend on the components & the architecture of their binding or reacting molecules and their interaction with the enzymatic machinery.

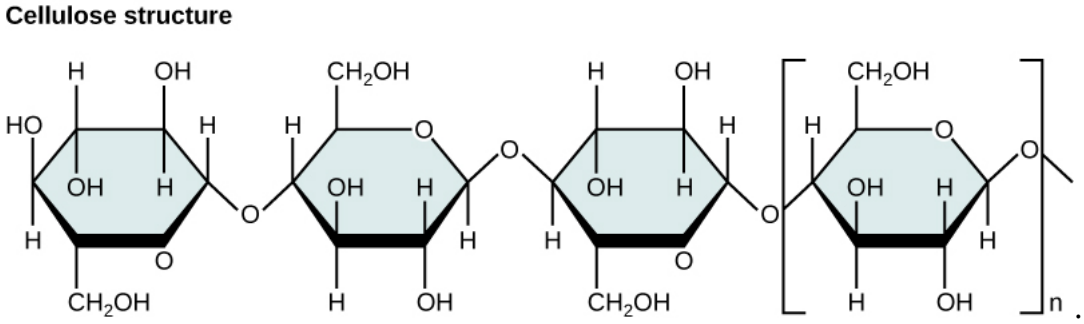
Polysaccharides are classified based on their functions, the type of monosaccharide units they contain, or their origin.

Based on the type of monosaccharides involved in the formation of polysaccharide structures, they are classified into two groups: homopolysaccharides and heteropolysaccharides.

Homopolysaccharides:

They are composed of repeating units of only one type of monomer. A few examples of homopolysaccharides include cellulose, chitin, starches (amylose and amylopectin), glycogen, and xylans. And based on their functional roles, these compounds are classified into structural polysaccharides and storage polysaccharides.

- Cellulose is a linear, unbranched polymer of glucose units joined by beta 1-4 linkages. It's one of the most abundant organic compounds in the biosphere.



- Chitin is a linear, long-chain polymer of N-acetyl-D-glucosamine (a derivative of glucose) residues/units, joined by beta 1-4 glycosidic linkages. It's the second most abundant natural biopolymer after cellulose.
- Starch is made of repeating units of D-glucose that are joined together by alpha-linkages. It's one of the most abundant polysaccharides found in plants and is composed of a mixture of amylose (15-20%) and amylopectin (80-85%).

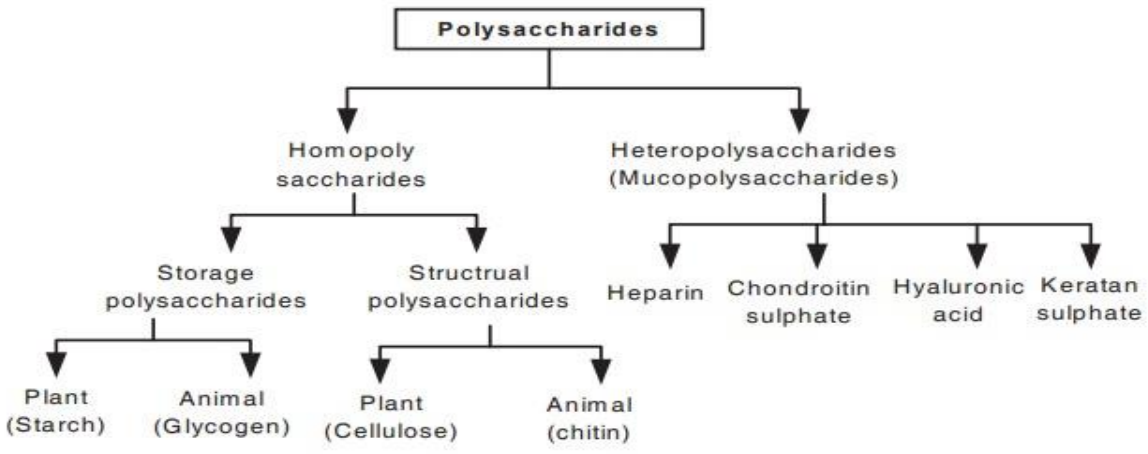
Heteropolysaccharides:

They are composed of two or more repeating units of different types of monomers. Examples include glycosaminoglycans, agarose, and peptidoglycans. In natural systems, they are linked to proteins, lipids, and peptides.

- Glycosaminoglycans (GAG) are negatively charged unbranched heteropolysaccharides. They are composed of repeating units of disaccharides with the general structural formula n. Amino acids like N-acetylglucosamine or N-acetylgalactosamine and uronic acid (like glucuronic acid) are normally present in the GAG structure.
- A list containing major GAGs is mentioned below:

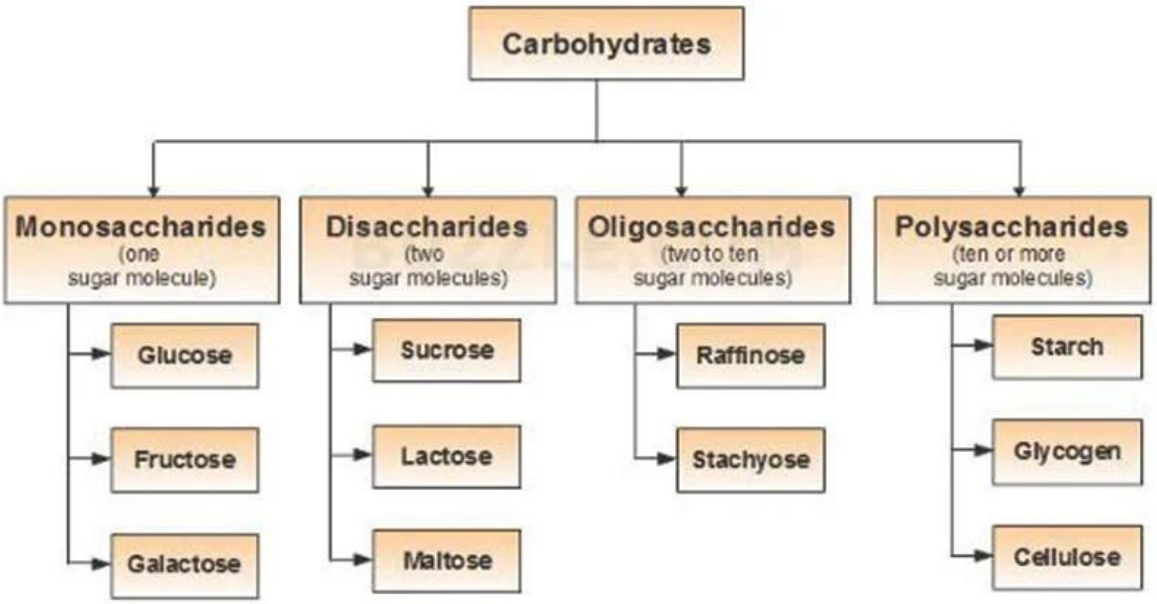
GAGs	Acidic sugar
Hyaluronic acid	D-Glucuronic acid
Chondroitin sulfate	D-Glucuronic acid
Heparan sulfate	D-Glucuronic acid or L-iduronic acid
Heparin	D-Glucuronic acid or L-iduronic acid
Dermatan sulfate	D-Glucuronic acid or L-iduronic acid
Keratan sulfate	D-Galactose

- Peptidoglycan is a heteropolymer of alternating units of N-acetylglucosamine (NAG) and N-acetylmuramic acids (NAM), linked together by beta-1,4-glycosidic linkage.
- Agarose is a polysaccharide composed of repeating units of a disaccharide, agarobiose, consisting of D-galactose and 3,6-anhydro-L-galactopyranose.



Functions of Polysaccharides

- **Structural polysaccharide:** They provide mechanical stability to cells, organs, and organisms. Examples include chitin and cellulose. Chitin is involved in the synthesis of fungal cell walls, while cellulose is an important constituent of diet for ruminants.
- **Storage polysaccharides:** They are carbohydrate storage reserves that release sugar monomers when required by the body. Examples include starch, glycogen, and inulin. Starch stores energy for plants, and in animals, it is catalyzed by the enzyme amylase (found in saliva) to fulfill the energy requirement. Glycogen is a polysaccharide food reserve of animals, bacteria, and fungi, while inulin is a storage reserve in plants.
- Agarose provides a supporting structure in the cell wall of marine algae.
- Peptidoglycan is an essential component of bacterial cell walls. It provides strength to the cell wall and participates in binary fission during bacterial reproduction.
- Peptidoglycan protects bacterial cells from bursting by counteracting the osmotic pressure of the cytoplasm.
- Hyaluronic acids are an essential component of the vitreous humor in the eye and synovial fluid (a lubricant fluid present in the body's joints). It's also involved in other developmental processes like tumor metastasis, angiogenesis, and blood coagulation.
- Heparin acts as a natural anticoagulant that prevents blood from clotting.
- Keratan sulfate is present in the cornea, cartilage, and bones. In joints, it acts as a cushion to absorb mechanical shocks.
- Chondroitin is an essential component of cartilage that provides resistance against compression.
- Dermatan sulfate is involved in wound repair, blood coagulation regulation, infection responses, and cardiovascular diseases.


Carbohydrates
Monosaccharides
(one
sugar molecule)
Disaccharides
(two
sugar molecules)
Oligosaccharides
(two to ten
sugar molecules)
Polysaccharides
(ten or more
sugar molecules)
Glucose
Fructose
Galactose
Sucrose
Lactose
Maltose
Raffinose
Stachyose
Starch
Glycogen
Cellulose

UNIT- III

LIPIDS

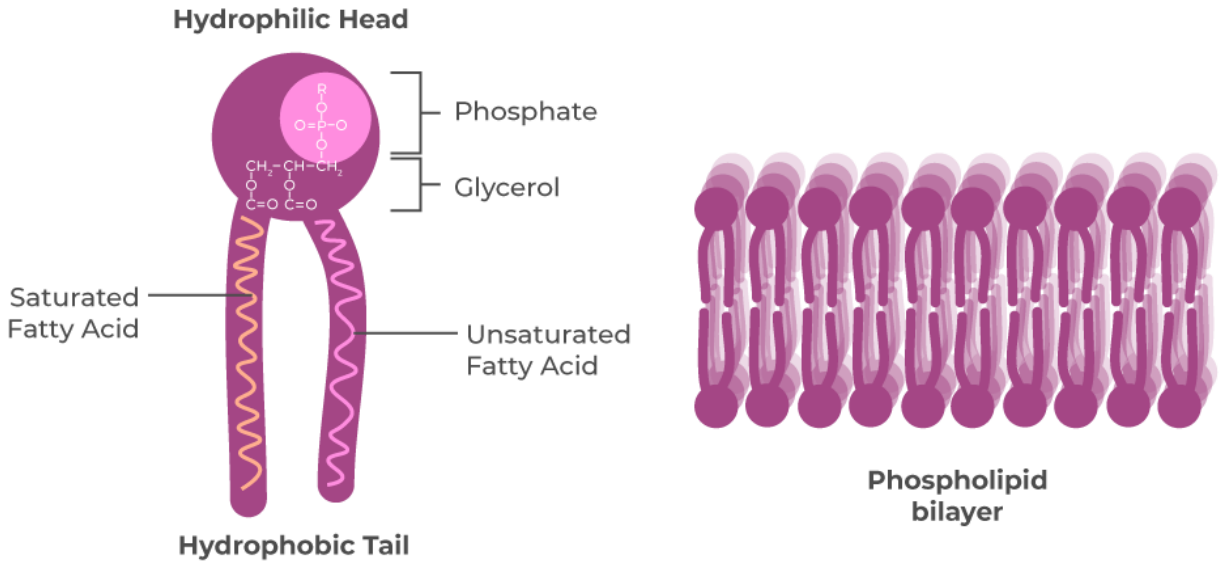
Lipids are a group of organic molecules that play essential roles in the structure and function of living organisms. They are characterized by their hydrophobic (water-repellent) nature and include compounds such as fats, oils, phospholipids, and steroids. Lipids, commonly known as fats and oils, are an important source of energy and form an important part of the structure of cell membranes, as well as are involved in cell signaling and hormone production. In this article, we will discuss lipids classification and lipids structure & function.

Lipids Definition

Lipids are organic molecules consisting of carbon, hydrogen, and oxygen atoms and serve as energy storage, structural support, and cell membrane composition in living organisms. Lipids include fats, oils, phospholipids, and steroids.

Lipids are group of heterogeneous organic compounds which are soluble in non-polar solvents. Lipids naturally occur in most plants, animals, and microorganisms. They include a variety of compounds such as fatty acids, phospholipids, sterols, sphingolipids, terpenes, and others. Structurally, they are esters or amides of fatty acids. These molecules can be soluble in non-polar solvents but not soluble in water.

Beyond their structural roles, lipids function as insulators, assisting in the maintenance of body temperature, and steroid hormones play a vital role in regulating various physiological processes. In the diet, lipids provide essential fatty acids and facilitate the absorption of fat-soluble vitamins.



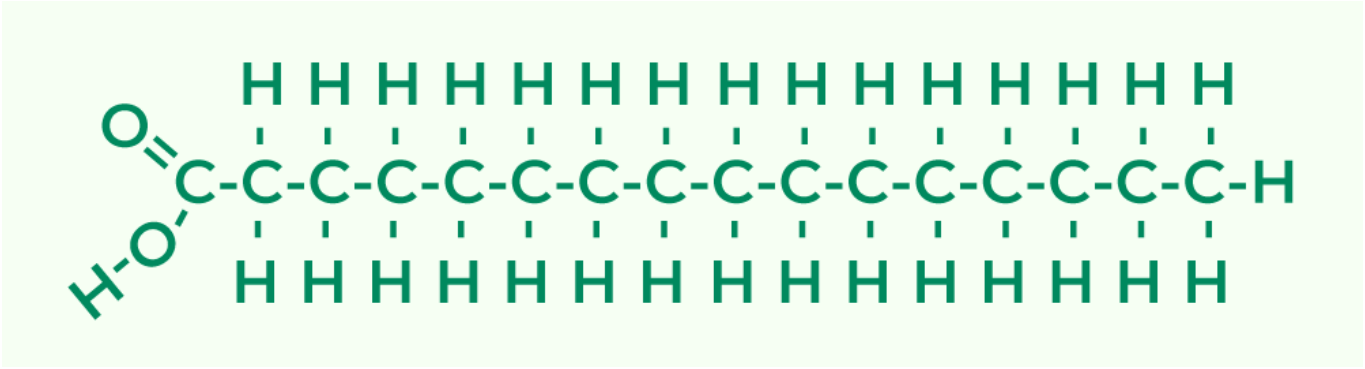
Properties of Lipids

They are organic compounds formed of fats and oils. Lipids produce high energy and perform different functions within a living organism, such as:

- Lipids stored in kidney.
- Lipids are generally hydrophobic, meaning they repel water and do not dissolve in it.
- Lipids are formed from hydrocarbon chains, and they are heterogeneous in nature.
- Fats and oils, in the form of triglycerides, are efficient energy storage molecules, providing a concentrated source of energy when broken down.
- Phospholipids are essential components of cell membranes, forming the lipid bilayer that defines cellular boundaries. They help in the selective permeability of a cell membrane.
- Lipids like cholesterol and steroid hormones consists of four-ring structure and function in membrane fluidity and cellular signaling.
- Lipids provide essential fatty acids that the body cannot produce on its own and allow the absorption of fat-soluble vitamins.

Lipids Structure

Lipid monomers are **glycerol and fatty acids**. The lipid structure is as follows:



- Fatty acids are a type of lipids that consists of long hydrocarbon chains with a carboxyl group (COOH) at one end.
- In lipids, such as triglycerides, the glycerol molecule function as a backbone. Glycerol molecule consists of three carbon atoms with a hydroxyl group attached to them.
- Glycerol are linked to the fatty acid through ester bonds, that forms triglycerides.
- The hydrocarbon chains of fatty acids are hydrophobic, that is repelling water.
- In lipids like phospholipids, a hydrophilic phosphate group is attached to the glycerol, while the fatty acid chains remain hydrophobic, resulting in an **amphipathic molecule**.

Classification of Lipids

Broadly, lipids classification is based on their chemical reactivity and the nature of their constituent molecules into two groups as follows:

1. Saponifiable Lipids
2. Nonsaponifable Lipids

Nonsaponifiable Lipids

- These lipids cannot be hydrolyzed or saponified using alkaline hydrolysis.
- They are often complex and structurally diverse.
- Examples of nonsaponifiable lipids include cholesterol (a steroid) and carotenoids (found in pigments like beta-carotene).

Saponifable Lipids

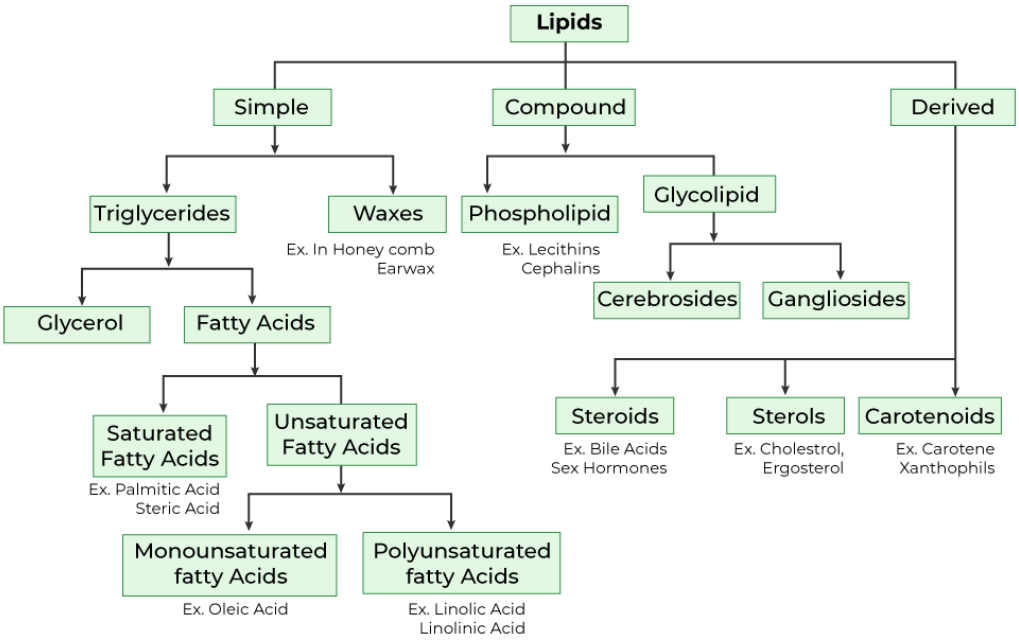
- Saponifiable lipids can be hydrolyzed or saponified using alkaline hydrolysis.
- They consist of fatty acids and other components that can be broken down into simpler compounds.
- The most common saponifiable lipids are triglycerides (fats and oils), which consist of glycerol and fatty acids esterified together.
- When saponified, these lipids break down into glycerol and fatty acids.
- Saponifiable are further divided into Polar and non-Polar lipids.

Polar Lipids:

Polar lipids are also known as amphipathic lipids because they have both hydrophilic (water-attracting) and hydrophobic (water-repellent) regions within their molecular structure. Examples of polar lipids include phospholipids and glycolipids.

Non- Polar Lipids:

Non-polar lipids are hydrophobic and do not have a significant hydrophilic component in their structure. They are primarily involved in energy storage and insulation. For example, Triglycerides (fats and oils).



Types of Lipids

Lipids subunits are – Glycerol and Fatty acids. Lipids are mainly classified into **three types**. They are simple, complex, and derived lipids.

- **Simple Lipids**: Simple lipids are triglycerides, esters of fatty acids, and wax esters. The hydrolysis of these lipids gives glycerol and fatty acids.
- **Complex Lipids:** Complex or compound lipids are the esters of fatty acids with groups along with alcohol and fatty acids. Examples are Phospholipids and Glycolipids.
- **Derived lipids:** Derived lipids are the hydrolyzed compounds of simple and complex lipids. Examples are fatty acids, steroids, fatty aldehydes, ketone bodies, lipid-soluble vitamins, and hormones.

Simple Lipids

Simple lipids are triglycerides, esters of fatty acids, and wax esters. The hydrolysis of these lipids gives glycerol and fatty acids. Simple lipids are classified into Triglycerides and Waxes.

1. **Fats**: Fatty acids join with glycerol via ester bonds.
2. **Waxes:** Fatty acid jig with a large molecular weight monohydric alcohol with an ester bond.

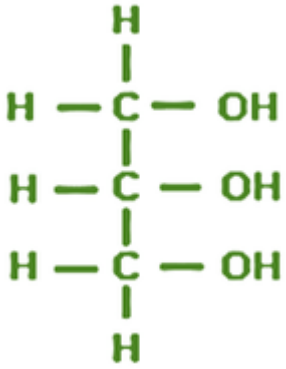
Triglycerides

- Triglycerides are the most common type of simple lipids.
- They consist of glycerol molecules linked to three fatty acid chains through ester bonds.
- Triglycerides are found in adipose tissue (body fat) and serve as a long-term energy reserve.

- They are the constituents of fats and oils. Lipids that are solid at room temperature are fats, and lipids that are liquid at room temperature are oils.

Glycerol

It is a colorless, odorless, viscous liquid that is sweet-tasting and non-toxic. The glycerol backbone is found in those lipids known as glycerides. It is a simple polyol compound.

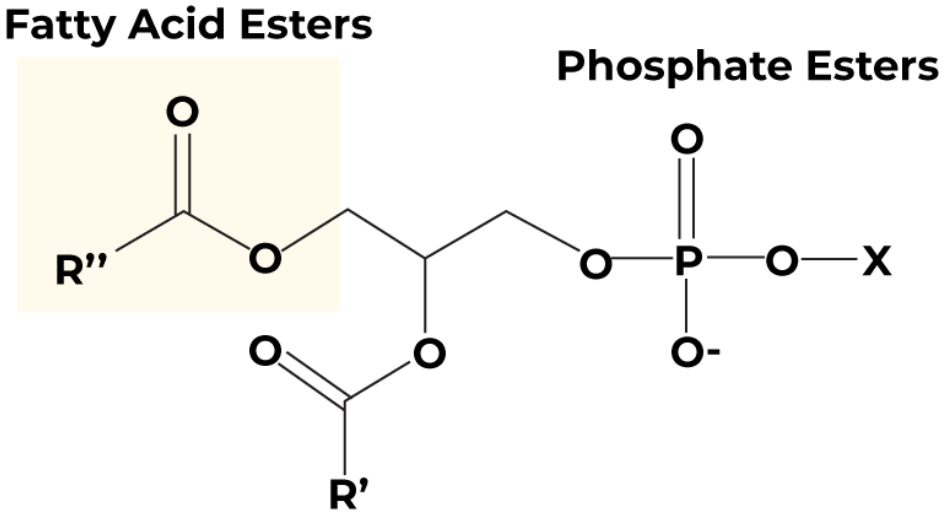


Complex Lipids

Complex lipids are a type of lipids that have more complex structures compared to simple lipids (triglycerides and waxes). They contain additional molecules, such as phosphates, carbohydrates, proteins, fatty acids and glycerol. Complex lipids are involved in various biological functions, including cell structure, energy storage, and cell signaling. Examples of complex lipids are Phospholipids and glycolipids.

Phospholipids

Phospholipids are constituents of cellular membranes. An ester is formed when a hydroxyl reacts with a carboxylic acid and loses H_2O. Phospholipids, also known as phosphatides, are classes of lipids whose molecule has a hydrophilic head and two hydrophobic tails. A head containing a phosphate group and tails derived from fatty acids joined by a glycerol molecule. They serve as emulsifiers.



There are two types of phospholipids:

- **Glycerophospholipids:** Glycerophospholipids are the class of phospholipids containing glycerol as alcohol, two fatty acids, and phosphate. It is the most abundant lipid in the cell membrane.
- **Sphingophospholipids:** Sphingophospholipids are the class of phospholipids containing sphingosine as alcohol. It produces ceramide by an amide linkage to a fatty acid. Ceramide is an important component of skin. It acts as a second messenger to regulate **programmed cell death**.

Glycolipid

It is a structural lipid, an essential part of the cell membrane. They are lipids with a carbohydrate attached by a glycosidic bond. They act as receptors at the surface of the red blood cell. It helps in the determination of an individual blood group. It has an important role in maintaining of the stability of the cell membrane. It kills pathogens to help the immune system of the body. Cerebrosides and Gangliosides are the two types of Glycolipids.

Precursor Lipids

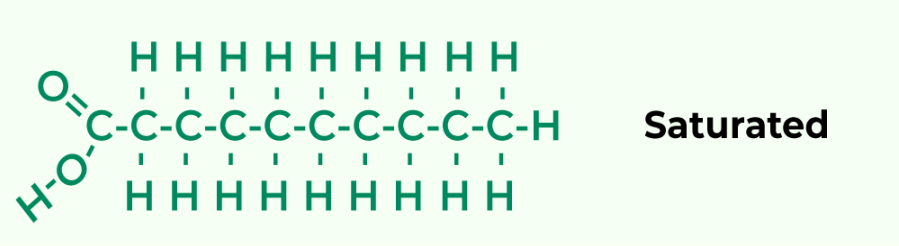
Precursor lipids are the building blocks from which other lipid molecules are synthesized or derived. They serve as starting points for the biosynthesis of more complex lipids. Some examples are- Fatty acids, Glycerol, and alcohol.

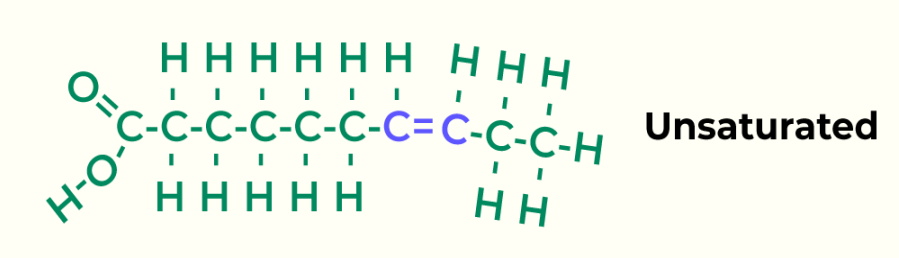
Fatty Lipids

Fatty acids are carboxylic acids; they are long chains of hydrocarbons with a carboxylic group at the end. Fatty acids are an important component of lipids, they are the building blocks of fat in the body. There are two types of fatty acids, saturated fatty acids and unsaturated fatty acids.

Saturated Fatty Acids

It consists of single C-C single bonds. These molecules fit closely together in a regular pattern and have strong attractions between fatty acid chains. These fatty acids have high melting points, which makes them solid at room temperature. Examples of saturated fatty acids are palmitic acid and stearic acid.





Unsaturated Fatty Acids

Unsaturated fatty acids are fatty acids that consist of one or more C=C double bonds. An unsaturated fatty acid is divided into two types.

1. **Mono polyunsaturated fatty acids:** Example: oleic acid.
2. **Polyunsaturated fatty acids:** Example: linoleic acid.

Role of Fats

Fats play an essential role in the body, including:

- Fats help our body by absorbing and transporting important fat-soluble vitamins.
- They are an important source of essential fatty acids.
- They insulate and protect our vital body organs.
- Fats produce energy in the form of carbohydrates.
- Fats are the structural component of cells.
- They help the body produce and regulate hormones.
- Fats support cell growth.
- They maintain our core temperature.
- Maintains blood pressure and cholesterol.

Derived Lipids

Derived lipids are the hydrolyzed compounds of simple and complex lipids. Examples are fatty acids, steroids, fatty aldehydes, ketone bodies, lipid-soluble vitamins, and hormones.

- **Steroids**

Steroids are found in the cell membrane and have fused ring structures. Many steroids have -OH functional groups, they are also hydrophobic and insoluble in water. All the steroids have 4 linked carbon rings and most of them have a short tail. Steroids also act as hormones in the body.

- **Sterols**

Sterols are solid steroid alcohols that are widely present in plants and animals such as cholesterol and ergosterol. They are the subgroup of steroids, which naturally occur in most eukaryotes. They are found in animal products. They are used to make bile for digestion in the body. Sterols can have greater than half of the membrane lipid content in cells and they are known to alter membrane structure and fluidity.

- **Carotenoids**

Carotenoids are lipid-soluble compounds. They are pigments that are mainly responsible for the yellow and red colors of plant and animal products. Carotenoids consist of carotenes and xanthophylls. A class of hydrocarbons is carotenes and its oxygenated derivatives are xanthophylls. They give color to many fruits and vegetables. They have antioxidant and anti-inflammatory properties for humans. Carotenoids are important for the health of the human eye.

Lipids Function

Functions of lipids are mentioned below:

- Lipids, like adipose tissue, act as insulators and help to maintain body temperature by reducing heat loss.
- Lipids, especially triglycerides, act as energy storage in organisms, providing a reserve of metabolic fuel.
- Phospholipids form the lipid bilayers of cell membranes and regulate the passage of molecules in and out of cells.
- Protecting the plant leaves from direct heat and drying.
- Steroid hormones, derived from cholesterol, play vital roles in regulating various physiological processes, including metabolism, growth, and reproduction.
- It acts as the structural component of the body and also acts as the hydrophobic barrier.
-

- In plants, lipids can be stored as oils in seeds, providing a source of energy for germination and early growth.
- Lipids form waterproofing structures, such as the waxy cuticle on plant leaves or the oil on the feathers of water birds.
- It provides color to many fruits and vegetables with the presence of carotenoid pigment.

Examples of Lipids

Lipid example are Ghee, Butter, Oil, Cheese, Cholesterol, waxes, etc. All these compounds have one thing in common i.e., they are insoluble in water and are soluble in organic solvents. Examples of lipids are as follows:

- **Triglycerides:** Found in fats and oils, these store energy in cells.
- **Phospholipids:** Key components of cell membranes, forming lipid bilayers. Both TAG and phosphoacylglycerol structures are almost the same just, phosphoacylglycerol-OH and phosphoric acid are attached with an ester bond and form the phosphatidic acid.
- **Steroids:** Include cholesterol, a membrane component, and steroid hormones.
- **Waxes:** Provide waterproofing in plants, animals, and microorganisms. Waxes are usually saturated with long-chain monohydric alcohols. They are the simple esters of fatty acids. Here are some examples:
    - **Beeswax:** Beeswax includes fatty acids and some free alcohol.
    - **Carnauba wax:** It is a hard wax used on cars and boats.
    - **Spermaceti**: it consists of cetyl palmitate. Used for pharmaceuticals.
- **Glycolipids:** Contain carbohydrates and are important in cell recognition.
- **Sphingolipids:** A diverse group of lipids involved in signaling and structure.
- **Lipoproteins:** Complexes of lipids and proteins, transport lipids in the bloodstream.
- **Eicosanoids:** Signaling molecules derived from fatty acids, regulate inflammation.
- **Isoprenoids:** Include vitamins (e.g., vitamin A) and carotenoids, with various functions

Unit-IV

Classification of nucleic acids

Nucleic acid, naturally occurring chemical compound that is capable of being broken down to yield phosphoric acid, sugars, and a mixture of organic bases (purines and pyrimidines). Nucleic acids are the main information-carrying molecules of the cell, and, by directing the process of protein synthesis, they determine the inherited characteristics of every living thing. The two main classes of nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA is the master blueprint for life and constitutes the genetic material in all free-living organisms and most viruses. RNA is the genetic material of certain viruses, but it is also found in all living cells, where it plays an important role in certain processes such as the making of proteins.

This article covers the chemistry of nucleic acids, describing the structures and properties that allow them to serve as the transmitters of genetic information. For a discussion of the genetic code, see heredity, and for a discussion of the role played by nucleic acids in protein synthesis, see metabolism.

Nucleotides: building blocks of nucleic acids

Basic structure

Nucleic acids are polynucleotides—that is, long chainlike molecules composed of a series of nearly identical building blocks called nucleotides. Each nucleotide consists of a nitrogen-containing aromatic base attached to a pentose (five-carbon) sugar, which is in turn attached to a phosphate group. Each nucleic acid contains four of five possible nitrogen-containing bases: adenine (A), guanine (G), cytosine (C), thymine (T), and uracil (U). A and G are categorized as purines, and C, T, and U are collectively called pyrimidines. All nucleic acids contain the bases A, C, and G; T, however, is found only in DNA, while U is found in RNA. The pentose sugar in DNA (2′-deoxyribose) differs from the sugar in RNA (ribose) by the absence of a hydroxyl group (—OH) on the 2′ carbon of the sugar ring. Without an attached phosphate group, the sugar attached to one of the bases is known as a nucleoside. The phosphate group connects successive sugar residues by bridging the 5′-hydroxyl group on one sugar to the 3′-hydroxyl group of the next sugar in the chain. These nucleoside linkages are called phosphodiester bonds and are the same in RNA and DNA.

Biosynthesis and degradation

Nucleotides are synthesized from readily available precursors in the cell. The ribose phosphate portion of both purine and pyrimidine nucleotides is synthesized from glucose via the pentose phosphate pathway. The six-atom pyrimidine ring is synthesized first and subsequently attached to the ribose phosphate. The two rings in purines are synthesized while attached to the ribose phosphate during the assembly of adenine or

guanine nucleosides. In both cases the end product is a nucleotide carrying a phosphate attached to the 5′ carbon on the sugar. Finally, a specialized enzyme called a kinase adds two phosphate groups using adenosine triphosphate (ATP) as the phosphate donor to form ribonucleoside triphosphate, the immediate precursor of RNA. For DNA, the 2′-hydroxyl group is removed from the ribonucleoside diphosphate to give deoxyribonucleoside diphosphate. An additional phosphate group from ATP is then added by another kinase to form a deoxyribonucleoside triphosphate, the immediate precursor of DNA.

During normal cell metabolism, RNA is constantly being made and broken down. The purine and pyrimidine residues are reused by several salvage pathways to make more genetic material. Purine is salvaged in the form of the corresponding nucleotide, whereas pyrimidine is salvaged as the nucleoside.
Deoxyribonucleic acid (DNA)

DNA structure

DNA structure, showing the nucleotide bases cytosine (C), thymine (T), adenine (A), and guanine (G) linked to a backbone of alternating phosphate (P) and deoxyribose sugar (S) groups. Two sugar-phosphate chains are paired through hydrogen bonds between A and T and between G and C, thus forming the twin-stranded double helix of the DNA molecule.

Learn about Watson and Crick's double-helix DNA structure, composed of two intertwined chains of nucleotides resembling a spiral ladder

The animated structure of a DNA molecule. Deoxyribose sugar molecules and phosphate molecules form the outer edges of the DNA double helix, and base pairs bind the two strands to one another.(more)

DNA is a polymer of the four nucleotides A, C, G, and T, which are joined through a backbone of alternating phosphate and deoxyribose sugar residues. These nitrogen-containing bases occur in complementary pairs as determined by their ability to form hydrogen bonds between them. A always pairs with T through two hydrogen bonds, and G always pairs with C through three hydrogen bonds. The spans of A:T and G:C hydrogen-bonded pairs are nearly identical, allowing them to bridge the sugar-phosphate chains uniformly. This structure, along with the molecule’s chemical stability, makes DNA the ideal genetic material. The bonding between complementary bases also provides a mechanism for the replication of DNA and the transmission of genetic information.
Chemical structure

The initial proposal of the structure of DNA by James Watson and Francis Crick, which was accompanied by a suggestion on the means of replication.(more)

In 1953 James D. Watson and Francis H.C. Crick proposed a three-dimensional structure for DNA based on low-resolution X-ray crystallographic data and on Erwin Chargaff's observation that, in naturally occurring DNA, the amount of T equals the amount of A and the amount of G equals the amount of C. Watson and Crick, who shared a Nobel Prize in 1962 for their efforts, postulated that two strands of polynucleotides coil around each other, forming a double helix. The two strands, though identical, run in opposite directions as determined by the orientation of the 5′ to 3′ phosphodiester bond. The sugar-phosphate chains run along the outside of the helix, and the bases lie on the inside, where they are linked to complementary bases on the other strand through hydrogen bonds.

The double helical structure of normal DNA takes a right-handed form called the B-helix. The helix makes one complete turn approximately every 10 base pairs. B-DNA has two principal grooves, a wide major groove and a narrow minor groove. Many proteins interact in the space of the major groove, where they make sequence-specific contacts with the bases. In addition, a few proteins are known to make contacts via the minor groove.

Several structural variants of DNA are known. In A-DNA, which forms under conditions of high salt concentration and minimal water, the base pairs are tilted and displaced toward the minor groove. Left-handed Z-DNA forms most readily in strands that contain sequences with alternating purines and pyrimidines. DNA can form triple helices when two strands containing runs of pyrimidines interact with a third strand containing a run of purines.

B-DNA is generally depicted as a smooth helix; however, specific sequences of bases can distort the otherwise regular structure. For example, short tracts of A residues interspersed with short sections of general sequence result in a bent DNA molecule. Inverted base sequences, on the other hand, produce cruciform structures with four-way junctions that are similar to recombination intermediates. Most of these alternative DNA structures have only been characterized in the laboratory, and their cellular significance is unknown.

Biological structures

Naturally occurring DNA molecules can be circular or linear. The genomes of single-celled bacteria and archaea (the prokaryotes), as well as the genomes of mitochondria and chloroplasts (certain functional structures within the cell), are circular molecules. In addition, some bacteria and archaea have smaller circular DNA molecules called plasmids that typically contain only a few genes. Many plasmids are readily transmitted from one cell to another. For a typical bacterium, the genome that encodes all of the genes of the organism is a single contiguous circular molecule

that contains a half million to five million base pairs. The genomes of most eukaryotes and some prokaryotes contain linear DNA molecules called chromosomes. Human DNA, for example, consists of 23 pairs of linear chromosomes containing three billion base pairs.

DNA wrapped around clusters of histone proteins to form nucleosomes, which are coiled to form solenoids, the basis of the chromatin fibre that makes up chromosomes.

In all cells, DNA does not exist free in solution but rather as a protein-coated complex called chromatin. In prokaryotes, the loose coat of proteins on the DNA helps to shield the negative charge of the phosphodiester backbone. Chromatin also contains proteins that control gene expression and determine the characteristic shapes of chromosomes. In eukaryotes, a section of DNA between 140 and 200 base pairs long winds around a discrete set of eight positively charged proteins called a histone, forming a spherical structure called the nucleosome. Additional histones are wrapped by successive sections of DNA, forming a series of nucleosomes like beads on a string. Transcription and replication of DNA is more complicated in eukaryotes because the nucleosome complexes have to be at least partially disassembled for the processes to proceed effectively.

Most prokaryote viruses contain linear genomes that typically are much shorter and contain only the genes necessary for viral propagation. Bacterial viruses called bacteriophages (or phages) may contain both linear and circular forms of DNA. For instance, the genome of bacteriophage λ (lambda), which infects the bacterium Escherichia coli, contains 48,502 base pairs and can exist as a linear molecule packaged in a protein coat. The DNA of phage λ can also exist in a circular form (as described in the section Site-specific recombination) that is able to integrate into the circular genome of the host bacterial cell. Both circular and linear genomes are found among eukaryotic viruses, but they more commonly use RNA as the genetic material.

Biochemical properties

➢ Denaturation

The strands of the DNA double helix are held together by hydrogen bonding interactions between the complementary base pairs. Heating DNA in solution easily breaks these hydrogen bonds, allowing the two strands to separate—a process called denaturation or melting. The two strands may reassociate when the solution cools, reforming the starting DNA duplex—a process called renaturation or hybridization. These processes form the basis of many important techniques for manipulating DNA. For example, a short piece of DNA called an oligonucleotide can be used to test whether a very long DNA sequence has the complementary sequence of the oligonucleotide embedded within it. Using hybridization, a single-stranded

DNA molecule can capture complementary sequences from any source. Single strands from RNA can also reassociate. DNA and RNA single strands can form hybrid molecules that are even more stable than double-stranded DNA. These molecules form the basis of a technique that is used to purify and characterize messenger RNA (mRNA) molecules corresponding to single genes.

- **Ultraviolet absorption**

DNA melting and reassociation can be monitored by measuring the absorption of ultraviolet (UV) light at a wavelength of 260 nanometres (billionths of a metre). When DNA is in a double-stranded conformation, absorption is fairly weak, but when DNA is single-stranded, the unstacking of the bases leads to an enhancement of absorption called hyperchromicity. Therefore, the extent to which DNA is single-stranded or double-stranded can be determined by monitoring UV absorption.

- **Chemical modification**

After a DNA molecule has been assembled, it may be chemically modified—sometimes deliberately by special enzymes called DNA methyltransferases and sometimes accidentally by oxidation, ionizing radiation, or the action of chemical carcinogens. DNA can also be cleaved and degraded by enzymes called nucleases.

- **Methylation**

Three types of natural methylation have been reported in DNA. Cytosine can be modified either on the ring to form 5-methylcytosine or on the exocyclic amino group to form N^4-methylcytosine. Adenine may be modified to form N^6-methyladenine. N^4-methylcytosine and N^6-methyladenine are found only in bacteria and archaea, whereas 5-methylcytosine is widely distributed. Special enzymes called DNA methyltransferases are responsible for this methylation; they recognize specific sequences within the DNA molecule so that only a subset of the bases is modified. Other methylations of the bases or of the deoxyribose are sometimes induced by carcinogens. These usually lead to mispairing of the bases during replication and have to be removed if they are not to become mutagenic.

Natural methylation has many cellular functions. In bacteria and archaea, methylation forms an essential part of the immune system by protecting DNA molecules from fragmentation by restriction endonucleases. In some organisms, methylation helps to eliminate incorrect base sequences introduced during DNA replication. By marking the parental strand with a methyl group, a cellular mechanism known as the mismatch repair system distinguishes between the newly replicated strand where the errors occur and the correct sequence on the template strand. In higher eukaryotes, 5-methylcytosine controls many cellular phenomena by preventing DNA transcription. Methylation is also believed to signal imprinting, a process whereby some genes inherited from one parent are selectively inactivated. Correct methylation may also

repress or activate key genes that control embryonic development. On the other hand, 5-methylcytosine is potentially mutagenic because thymine produced during the methylation process converts C:G pairs to T:A pairs. In mammals, methylation takes place selectively within the dinucleotide sequence CG—a rare sequence, presumably because it has been lost by mutation. In many cancers, mutations are found in key genes at CG dinucleotides.

- **Nucleases**

Nucleases are enzymes that hydrolytically cleave the phosphodiester backbone of DNA. Endonucleases cleave in the middle of chains, while exonucleases operate selectively by degrading from the end of the chain. Nucleases that act on both single- and double-stranded DNA are known.

Restriction endonucleases are a special class that recognize and cleave specific sequences in DNA. Type II restriction endonucleases always cleave at or near their recognition sites. They produce small, well-defined fragments of DNA that help to characterize genes and genomes and that produce recombinant DNAs. Fragments of DNA produced by restriction endonucleases can be moved from one organism to another. In this way it has been possible to express proteins such as human insulin in bacteria.

- **Mutation**

Chemical modification of DNA can lead to mutations in the genetic material. Anions such as bisulfite can deaminate cytosine to form uracil, changing the genetic message by causing C-to-T transitions. Exposure to acid causes the loss of purine residues, though specific enzymes exist in cells to repair these lesions. Exposure to UV light can cause adjacent pyrimidines to dimerize, while oxidative damage from free radicals or strong oxidizing agents can cause a variety of lesions that are mutagenic if not repaired. Halogens such as chlorine and bromine react directly with uracil, adenine, and guanine, giving substituted bases that are often mutagenic. Similarly, nitrous acid reacts with primary amine groups—for example, converting adenosine into inosine—which then leads to changes in base pairing and mutation. Many chemical mutagens, such as chlorinated hydrocarbons and nitrites, owe their toxicity to the production of halides and nitrous acid during their metabolism in the body.

- **Supercoiling**

Circular DNA molecules such as those found in plasmids or bacterial chromosomes can adopt many different topologies. One is active supercoiling, which involves the cleavage of one DNA strand, its winding one or more turns around the complementary strand, and then the resealing of the molecule. Each complete rotation leads to the introduction of one supercoiled turn in the DNA, a process that can continue until the DNA is fully wound and collapses on itself in a tight ball. Reversal is also possible. Special enzymes called gyrases and topoisomerases catalyze the winding and relaxation of supercoiled DNA. In the linear chromosomes of eukaryotes, the DNA is usually tightly constrained at various points by proteins, allowing

the intervening stretches to be supercoiled. This property is partially responsible for the great compaction of DNA that is necessary to fit it within the confines of the cell. The DNA in one human cell would have an extended length of between two and three metres, but it is packed very tightly so that it can fit within a human cell nucleus that is 10 micrometres in diameter.

Sequence determination

Methods to determine the sequences of bases in DNA were pioneered in the 1970s by Frederick Sanger and Walter Gilbert, whose efforts won them a Nobel Prize in 1980. The Gilbert-Maxam method relies on the different chemical reactivities of the bases, while the Sanger method is based on enzymatic synthesis of DNA in vitro. Both methods measure the distance from a fixed point on DNA to each occurrence of a particular base—A, C, G, or T. DNA fragments obtained from a series of reactions are separated according to length in four "lanes" by gel electrophoresis. Each lane corresponds to a unique base, and the sequence is read directly from the gel. The Sanger method has now been automated using fluorescent dyes to label the DNA, and a single machine can produce tens of thousands of DNA base sequences in a single run.

Ribonucleic acid (RNA)

RNA is a single-stranded nucleic acid polymer of the four nucleotides A, C, G, and U joined through a backbone of alternating phosphate and ribose sugar residues. It is the first intermediate in converting the information from DNA into proteins essential for the working of a cell. Some RNAs also serve direct roles in cellular metabolism. RNA is made by copying the base sequence of a section of double-stranded DNA, called a gene, into a piece of single-stranded nucleic acid. This process, called transcription (see below RNA metabolism), is catalyzed by an enzyme called RNA polymerase.

Chemical structure

Whereas DNA provides the genetic information for the cell and is inherently quite stable, RNA has many roles and is much more reactive chemically. RNA is sensitive to oxidizing agents such as periodate that lead to opening of the 3′-terminal ribose ring. The 2′-hydroxyl group on the ribose ring is a major cause of instability in RNA, because the presence of alkali leads to rapid cleavage of the phosphodiester bond linking ribose and phosphate groups. In general, this instability is not a significant problem for the cell, because RNA is constantly being synthesized and degraded.

Interactions between the nitrogen-containing bases differ in DNA and RNA. In DNA, which is usually double-stranded, the bases in one strand pair with complementary bases in a second DNA strand. In RNA, which is usually single-stranded, the bases pair with other bases within the same molecule, leading to

complex three-dimensional structures. Occasionally, intermolecular RNA/RNA duplexes do form, but they form a right-handed A-type helix rather than the B-type DNA helix. Depending on the amount of salt present, either 11 or 12 base pairs are found in each turn of the helix. Helices between RNA and DNA molecules also form; these adopt the A-type conformation and are more stable than either RNA/RNA or DNA/DNA duplexes. Such hybrid duplexes are important species in biology, being formed when RNA polymerase transcribes DNA into mRNA for protein synthesis and when reverse transcriptase copies a viral RNA genome such as that of the human immunodeficiency virus (HIV).

Single-stranded RNAs are flexible molecules that form a variety of structures through internal base pairing and additional non-base pair interactions. They can form hairpin loops such as those found in transfer RNA (tRNA), as well as longer-range interactions involving both the bases and the phosphate residues of two or more nucleotides. This leads to compact three-dimensional structures. Most of these structures have been inferred from biochemical data, since few crystallographic images are available for RNA molecules. In some types of RNA, a large number of bases are modified after the RNA is transcribed. More than 90 different modifications have been documented, including extensive methylations and a wide variety of substitutions around the ring. In some cases these modifications are known to affect structure and are essential for function.

Types of RNA

Messenger RNA (mRNA)

Messenger RNA (mRNA) delivers the information encoded in one or more genes from the DNA to the ribosome, a specialized structure, or organelle, where that information is decoded into a protein. In prokaryotes, mRNAs contain an exact transcribed copy of the original DNA sequence with a terminal 5′-triphosphate group and a 3′-hydroxyl residue. In eukaryotes the mRNA molecules are more elaborate. The 5′-triphosphate residue is further esterified, forming a structure called a cap. At the 3′ ends, eukaryotic mRNAs typically contain long runs of adenosine residues (polyA) that are not encoded in the DNA but are added enzymatically after transcription. Eukaryotic mRNA molecules are usually composed of small segments of the original gene and are generated by a process of cleavage and rejoining from an original precursor RNA (pre-mRNA) molecule, which is an exact copy of the gene (as described in the section Splicing). In general, prokaryotic mRNAs are degraded very rapidly, whereas the cap structure and the polyA tail of eukaryotic mRNAs greatly enhance their stability.

Ribosomal RNA (rRNA)

Ribosomal RNA (rRNA) molecules are the structural components of the ribosome. The rRNAs form extensive secondary structures and play an active role in recognizing conserved portions of mRNAs and

tRNAs. They also assist with the catalysis of protein synthesis. In the prokaryote E. coli, seven copies of the rRNA genes synthesize about 15,000 ribosomes per cell. In eukaryotes the numbers are much larger. Anywhere from 50 to 5,000 sets of rRNA genes and as many as 10 million ribosomes may be present in a single cell. In eukaryotes these rRNA genes are looped out of the main chromosomal fibres and coalesce in the presence of proteins to form an organelle called the nucleolus. The nucleolus is where the rRNA genes are transcribed and the early assembly of ribosomes takes place.

Transfer RNA (tRNA)

Transfer RNA (tRNA) carries individual amino acids into the ribosome for assembly into the growing polypeptide chain. The tRNA molecules contain 70 to 80 nucleotides and fold into a characteristic cloverleaf structure. Specialized tRNAs exist for each of the 20 amino acids needed for protein synthesis, and in many cases more than one tRNA for each amino acid is present. The nucleotide sequence is converted into a protein sequence by translating each three-base sequence (called a codon) with a specific protein. The 61 codons used to code amino acids can be read by many fewer than 61 distinct tRNAs (as described in the section Translation). In E. coli a total of 40 different tRNAs are used to translate the 61 codons. The amino acids are loaded onto the tRNAs by specialized enzymes called aminoacyl tRNA synthetases, usually with one synthetase for each amino acid. However, in some organisms, less than the full complement of 20 synthetases are required because some amino acids, such as glutamine and asparagine, can be synthesized on their respective tRNAs. All tRNAs adopt similar structures because they all have to interact with the same sites on the ribosome.

Ribozymes

Not all catalysis within the cell is carried out exclusively by proteins. Thomas Cech and Sidney Altman, jointly awarded a Nobel Prize in 1989, discovered that certain RNAs, now known as ribozymes, showed enzymatic activity. Cech showed that a noncoding sequence (intron) in the small subunit rRNA of protozoans, which had to be removed before the rRNA was functional, can excise itself from a much longer precursor RNA molecule and rejoin the two ends in an autocatalytic reaction. Altman showed that the RNA component of an RNA protein complex called ribonuclease P can cleave a precursor tRNA to generate a mature tRNA. In addition to self-splicing RNAs similar to the one discovered by Cech, artificial RNAs have been made that show a variety of catalytic reactions. It is now widely held that there was a stage during evolution when only RNA catalyzed and stored genetic information. This period, sometimes called “the RNA world,” is believed to have preceded the function of DNA as genetic material.

Antisense RNAs

Most antisense RNAs are synthetically modified derivatives of RNA or DNA with potential therapeutic value. In nature, antisense RNAs contain sequences that are the complement of the normal coding sequences found in mRNAs (also called sense RNAs). Like mRNAs, antisense RNAs are single-stranded, but they cannot be translated into protein. They can inactivate their complementary mRNA by forming a double-stranded structure that blocks the translation of the base sequence. Artificially introducing antisense RNAs into cells selectively inactivates genes by interfering with normal RNA metabolism.

Viral genomes

Many viruses use RNA for their genetic material. This is most prevalent among eukaryotic viruses, but a few prokaryotic RNA viruses are also known. Some common examples include poliovirus, human immunodeficiency virus (HIV), and influenza virus, all of which affect humans, and tobacco mosaic virus, which infects plants. In some viruses the entire genetic material is encoded in a single RNA molecule, while in the segmented RNA viruses several RNA molecules may be present. Many RNA viruses such as HIV use a specialized enzyme called reverse transcriptase that permits replication of the virus through a DNA intermediate. In some cases this DNA intermediate becomes integrated into the host chromosome during infection; the virus then exists in a dormant state and effectively evades the host immune system.

Other RNAs

Many other small RNA molecules with specialized functions are present in cells. For example, small nuclear RNAs (snRNAs) are involved in RNA splicing (see below), and other small RNAs that form part of the enzymes telomerase or ribonuclease P are part of ribonucleoprotein particles. The RNA component of telomerase contains a short sequence that serves as a template for the addition of small strings of oligonucleotides at the ends of eukaryotic chromosomes. Other RNA molecules serve as guide RNAs for editing, or they are complementary to small sections of rRNA and either direct the positions at which methyl groups need to be added or mark U residues for conversion to the isomer pseudouridine.

RNA processing

Cleavage

Following synthesis by transcription, most RNA molecules are processed before reaching their final form. Many rRNA molecules are cleaved from much larger transcripts and may also be methylated or enzymatically modified. In addition, tRNAs are usually formed as longer precursor molecules that are cleaved by ribonuclease P to generate the mature 5′ end and often have extra residues added to their 3′ end to form the sequence CCA. The hydroxyl group on the ribose ring of the terminal A of the 3′-CCA sequence acts as the amino acid acceptor necessary for the function of RNA in protein building.

Splicing

In prokaryotes the protein coding sequence occupies one continuous linear segment of DNA. However, in eukaryotic genes the coding sequences are frequently "split" in the genome—a discovery reached independently in the 1970s by Richard J. Roberts (the author of this article) and Phillip A. Sharp, whose work won them a Nobel Prize in 1993. The segments of DNA or RNA coding for protein are called exons, and the noncoding regions separating the exons are called introns. Following transcription, these coding sequences must be joined together before the mRNAs can function. The process of removal of the introns and subsequent rejoining of the exons is called RNA splicing. Each intron is removed in a separate series of reactions by a complicated piece of enzymatic machinery called a spliceosome. This machinery consists of a number of small nuclear ribonucleoprotein particles (snRNPs) that contain small nuclear RNAs (snRNAs).

RNA editing

Some RNA molecules, particularly those in protozoan mitochondria, undergo extensive editing following their initial synthesis. During this editing process, residues are added or deleted by a posttranscriptional mechanism under the influence of guide RNAs. In some cases as much as 40 percent of the final RNA molecule may be derived by this editing process, rather than being coded directly in the genome. Some examples of editing have also been found in mRNA molecules, but these appear much more limited in scope.

Nucleic acid metabolism

DNA metabolism

Replication, repair, and recombination—the three main processes of DNA metabolism—are carried out by specialized machinery within the cell. DNA must be replicated accurately in order to ensure the integrity of the genetic code. Errors that creep in during replication or because of damage after replication must be repaired. Finally, recombination between genomes is an important mechanism to provide variation within a species and to assist the repair of damaged DNA. The details of each process have been worked out in prokaryotes, where the machinery is more streamlined, simpler, and more amenable to study. Many of the basic principles appear to be similar in eukaryotes.

Replication

Basic mechanisms

DNA replication is a semiconservative process in which the two strands are separated and new complementary strands are generated independently, resulting in two exact copies of the original DNA molecule. Each copy thus contains one strand that is derived from the parent and one newly synthesized strand. Replication begins at a specific point on a chromosome called an origin, proceeds in both

directions along the strand, and ends at a precise point. In the case of circular chromosomes, the end is reached automatically when the two extending chains meet, at which point specific proteins join the strands. DNA polymerases cannot initiate replication at the end of a DNA strand; they can only extend preexisting oligonucleotide fragments called primers. Therefore, in linear chromosomes, special mechanisms initiate and terminate DNA synthesis to avoid loss of information. The initiation of DNA synthesis is usually preceded by synthesis of a short RNA primer by a specialized RNA polymerase called primase. Following DNA replication, the initiating primer RNAs are degraded.

The two DNA strands are replicated in different fashions dictated by the direction of the phosphodiester bond. The leading strand is replicated continuously by adding individual nucleotides to the 3′ end of the chain. The lagging strand is synthesized in a discontinuous manner by laying down short RNA primers and then filling the gaps by DNA polymerase, such that the bases are always added in the 5′ to 3′ direction. The short RNA fragments made during the copying of the lagging strand are degraded when no longer needed. The two newly synthesized DNA segments are joined by an enzyme called DNA ligase. In this way, replication can proceed in both directions, with two leading strands and two lagging strands proceeding outward from the origin.

Enzymes of replication

DNA polymerase adds single nucleotides to the 3′ end of either an RNA or a DNA molecule. In the prokaryote E. coli, there are three DNA polymerases; one is responsible for chromosome replication, and the other two are involved in the resynthesis of DNA during damage repair. DNA polymerases of eukaryotes are even more complicated. In human cells, for instance, more than five different DNA polymerases have been characterized. Separate polymerases catalyze the synthesis of the leading and lagging strands in human cells, and a separate polymerase is responsible for replication of mitochondrial DNA. The other polymerases are involved in the repair of DNA damage.

A number of other proteins are also essential for replication. Proteins called DNA helicases help to separate the two strands of DNA, and single-stranded DNA binding proteins stabilize them during opening prior to being copied. The opening of the DNA helix introduces considerable strain in the form of supercoiling, a movement that is subsequently relaxed by enzymes called topoisomerases (see above Supercoiling). A special RNA polymerase called primase synthesizes the primers needed at the origin to begin transcription, and DNA ligase seals the nicks formed between individual fragments.

The ends of linear eukaryotic chromosomes are marked by special sequences called telomeres that are synthesized by a special DNA polymerase called telomerase. This enzyme contains an RNA component that

serves as a template for the exact sequence found at the ends of chromosomes. Multiple copies of a short sequence within the telomerase-associated RNA are made and added to the telomere ends. This has the effect of preventing shortening of the DNA chain that would otherwise occur during replication.

Single-stranded viral genomes, mitochondrial genomes, and some viral genomes are replicated in specialized ways. Several viruses such as adenovirus use a nucleotide covalently bound to a protein as a primer, and the protein remains covalently bound to the DNA after replication. Many single-stranded viruses use a rolling circle mechanism of replication whereby a double-stranded copy of the virus is first made. The replicating machinery then copies the nonviral strand in a continuous fashion, generating long single-stranded DNA from which full-length viral DNA strands are excised by specialized nucleases.

Recombination

Recombination is the principal mechanism through which variation is introduced into populations. For example, during meiosis, the process that produces sex cells (sperm or eggs), homologous chromosomes—one derived from the mother and the equivalent from the father—become paired, and recombination, or crossing-over, takes place. The two DNA molecules are fragmented, and similar segments of the chromosome are shuffled to produce two new chromosomes, each being a mosaic of the originals. The pair separates so that each sperm or egg receives just one of the shuffled chromosomes. When sperm and egg fuse, the normal set of two copies of each chromosome is restored.

There are two forms of recombination, general and site-specific. General recombination typically involves cleavage and rejoining at identical or very similar sequences. In site-specific recombination, cleavage takes place at a specific site into which DNA is usually inserted. General recombination occurs among viruses during infection, in bacteria during conjugation, during transformation whereby DNA is directly introduced into cells, and during some types of repair processes. Site-specific recombination is frequently involved in the parasitic distribution of DNA segments throughout genomes. Many viruses, as well as special segments of DNA called transposons, rely on site-specific recombination to multiply and spread. The two processes are described in greater detail below.

General recombination

General recombination, also called homologous recombination, involves two DNA molecules that have long stretches of similar base sequences. The DNA molecules are nicked to produce single strands; these subsequently invade the other duplex, where base pairing leads to a four-stranded DNA structure. The cruciform junction within this structure is called a Holliday junction, named after Robin Holliday, who proposed the original model for homologous recombination in 1964. The Holliday junction travels along the

DNA duplex by "unzipping" one strand and reforming the hydrogen bonds on the second strand. Following this branch migration, the two duplexes can be nicked again, allowing them to separate. Finally, the nicks are repaired by DNA ligase. The result is two DNA duplexes in which the segment between the two nicks has been replaced. The enzymes involved in recombination have been characterized best in the prokaryote E. coli. A key enzyme is RecA, which catalyzes the strand invasion process. RecA coats single-stranded DNA and facilitates its pairing with a double-stranded DNA molecule containing the same sequence, which produces a loop structure.

Another protein, known as RecBC, is important for the recombination process. Functioning at free ends of DNA, RecBC catalyzes an unwinding-rewinding reaction as it traverses the length of the molecule. Since unwinding is faster than rewinding, a loop is produced behind the enzyme that facilitates subsequent pairing with another DNA molecule. A number of other proteins are also important for recombination, including single-stranded DNA binding proteins that stabilize single-stranded DNA, DNA polymerase to repair any gaps that might be formed, and DNA ligase to reseal the nicks after recombination is complete. The details of eukaryotic recombination are expected to parallel those found in E. coli, although the highly compact chromatin structure in eukaryotes makes the process more complicated.

It is important to note that the initial product of recombination between two regions of DNA that are similar but not identical will be a "heteroduplex"—that is, a molecule in which mismatched bases will be present at some positions in the helix. Thus, in the specialized recombination that takes place during meiosis, one round of replication is necessary before the mosaic chromosomes produced by recombination are properly matched. Enzymes are present in cells that specifically recognize and repair mismatches, so that the initial products of recombination can sometimes be repaired before they are replicated. In such cases the final products of replication will not be true reciprocal events, but rather one of the original parental molecules will appear to have been maintained to the exclusion of the other—a process called gene conversion.

Recombination also functions occasionally to repair lesions in DNA. If one chromosome of a pair becomes irreversibly damaged, the information from the other chromosome can be copied and inserted by recombination to provide a correct replacement of the damaged section. The key idea here is that sequences flanking the damage from a sister chromosome can base-pair with the corresponding sequences on the damaged chromosome, thus allowing replication to copy the correct sequence and repair the lesion.

Site-specific recombination

Site-specific recombination involves very short specific sequences that are recognized by proteins. Long DNA sequences such as viral genomes, drug-resistance elements, and regulatory sequences such as the

mating type locus in yeast can be inserted, removed, or inverted, having profound regulatory effects. More than any other mechanism, site-specific recombination is responsible for reshaping genomes. For example, the genomes of many higher organisms, including plants and humans, show evidence that transposable elements have been constantly inserted throughout the genome and even into one another from time to time.

One example of site-specific recombination is the integration of DNA from bacteriophage λ into the chromosome of E. coli. In this reaction, bacteriophage λ DNA, which is a linear molecule in the normal phage, first forms a circle and then is cleaved by the enzyme λ-integrase at a specific site called the phage attachment site. A similar site on the bacterial chromosome is cut by integrase to give ends with the identical extension. Because of the complementarity between these two ends, they can be rejoined so that the original circular λ chromosome is inserted into the chromosome of the E. coli bacterium. Once integrated, the phage can be held in an inactive state until signals are generated that reverse the process, allowing the phage genome to escape and resume its normal life cycle of growth and spread into other bacteria. This site-specific recombination process requires only λ-integrase and one host DNA binding protein called the integration host factor. A third protein, called excisionase, recognizes the hybrid sites formed on integration and, in conjunction with integrase, catalyzes an excision process whereby the λ chromosome is removed from the bacterial chromosome.

A similar but more widespread version of DNA integration and excision is exhibited by the transposons, the so-called jumping genes. These elements range in size from fewer than 1,000 to as many as 40,000 base pairs. Transposons are able to move from one location in a genome to another, as first discovered in corn (maize) during the 1940s and ’50s by Barbara McClintock, whose work won her a Nobel Prize in 1983. Most, if not all, transposons encode an enzyme called transposase that acts much like λ-integrase by cleaving the ends of the transposon as well as its target site. Transposons differ from bacteriophage λ in that they do not have a separate existence outside of the chromosome but rather are always maintained in an integrated site. Two types of transposition can occur—one in which the element simply moves from one site in the chromosome to another and a second in which the transposon is replicated prior to moving. This second type of transposition leaves behind the original copy of the transposon and generates a second copy that is inserted elsewhere in the genome. Known as replicative transposition, this process is the mechanism responsible for the vast spread of transposable elements in many higher organisms.

The simplest kinds of transposons merely contain a copy of the transposase with no additional genes. They behave as parasitic elements and usually have no known associated function that is advantageous to the host. More often, transposable elements have additional genes associated with them—for example, antibiotic

resistance factors. Antibiotic resistance typically occurs when an infecting bacterium acquires a plasmid that carries a gene encoding resistance to one or more antibiotics. Typically, these resistance genes are carried on transposable elements that have moved into plasmids and are easily transferred from one organism to another. Once a bacterium picks up such a gene, it enjoys a great selective advantage because it can grow in the presence of the antibiotic. Indiscriminate use of antibiotics actually promotes the buildup of these drug-resistant plasmids and strains.

Repair

It is extremely important that the integrity of DNA be maintained in order to ensure the accurate workings of a cell over its lifetime and to make certain that genetic information is accurately passed from one generation to the next. This maintenance is achieved by repair processes that constantly monitor the DNA for lesions and activate appropriate repair enzymes. As described in the section General recombination, serious lesions in DNA such as pyrimidine dimers or gaps can be repaired by recombination mechanisms, but there are many other repair mechanisms.

One important mechanism is that of mismatch repair, which has been studied extensively in E. coli. The system is directed by the presence of a methyl group within the sequence GATC on the template strand. Comparable systems for mismatch repair also operate in eukaryotes, though the template strand is not marked by methyl groups. In fact, lesions within the genes for human mismatch repair systems are known to be responsible for many cancers. Loss of the mismatch repair system allows mutations to build up quickly and eventually to affect the genes that cause cells to divide. As a result, cells divide in an uncontrolled manner and become cancerous.

Once replication is complete, the most common kind of damage to nucleic acids is one in which the normal A, C, G, and T bases are changed into chemically modified bases that usually differ significantly from their natural counterparts. The only exceptions are the deamination of cytosine to uracil and the deamination of 5-methylcytosine to thymine. In these cases the product is a G:U or G:T mismatch. Specific enzymes called DNA glycosylases can recognize uracil in DNA or the thymine in a G:T mismatch and can selectively remove the base by cleaving the bond between the base and the deoxyribose sugar. Many of these enzymes are specific for the different chemically modified bases that may be present in DNA.

Another common means of repairing DNA lesions is by an excision repair pathway. Enzymes recognize damage within DNA, probably by detecting an altered conformation of DNA, and then nick the strand on either side of the lesion, allowing a small single-stranded DNA to be excised. DNA polymerase and DNA ligase then repair the single-stranded gap. In all of these systems, the presence of an abnormal base signifies which strand is to be repaired, and the complementary strand is used as the template to ensure the accuracy of repair.

RNA metabolism

RNA provides the link between the genetic information encoded in DNA and the actual workings of the cell. Some RNA molecules such as the rRNAs and the snRNAs (described in the section Types of RNA) become part of complicated ribonucleoprotein structures with specialized roles in the cell. Others such as tRNAs play key roles in protein synthesis, while mRNAs direct the synthesis of proteins by the ribosome. Three distinct phases of RNA metabolism occur. First, selected segments of the genome are copied by transcription to produce the precursor RNAs. Second, these precursors are processed to become functionally mature RNAs ready for use. When these RNAs are mRNAs, they are then used for translation. Third, after use the RNAs are degraded, and the bases are recycled. Thus, transcription is the process where a specific segment of DNA, a gene, is copied into a specific RNA that encodes a single protein or plays a structural or catalytic role. Translation is the decoding of the information within mRNA molecules that takes place on a specialized structure called a ribosome. There are important differences in both transcription and translation between prokaryotic and eukaryotic organisms.

Transcription

Small segments of DNA are transcribed into RNA by the enzyme RNA polymerase, which achieves this copying in a strictly controlled process. The first step is to recognize a specific sequence on DNA called a promoter that signifies the start of the gene. The two strands of DNA become separated at this point, and RNA polymerase begins copying from a specific point on one strand of the DNA using a ribonucleoside 5′-triphosphate to begin the growing chain. Additional ribonucleoside triphosphates are used as the substrate, and, by cleavage of their high-energy phosphate bond, ribonucleoside monophosphates are incorporated into the growing RNA chain. Each successive ribonucleotide is directed by the complementary base pairing rules of DNA. Thus, a C in DNA directs the incorporation of a G into RNA, G is copied into C, T into A, and A into U. Synthesis continues until a termination signal is reached, at which point the RNA polymerase drops off the DNA, and the RNA molecule is released. In some cases this RNA molecule is the final mRNA. In other cases it is a pre-mRNA and requires further processing before it is ready for translation by the ribosome. Ahead of many genes in prokaryotes, there are signals called “operators” where specialized proteins called repressors bind to the DNA just upstream of the start point of transcription and prevent access to the DNA by RNA polymerase. These repressor proteins thus prevent transcription of the gene by physically blocking the action of the RNA polymerase. Typically, repressors are released from their blocking action when they receive signals from other molecules in the cell indicating that the gene needs to be expressed. Ahead of some prokaryotic genes are signals to which activator proteins bind that positively induce transcription.

Transcription in higher organisms is more complicated. First, the RNA polymerase of eukaryotes is a more complicated enzyme than the relatively simple five-subunit enzyme of prokaryotes. In addition, there are many more accessory factors that help to control the efficiency of the individual promoters. These accessory proteins are called transcription factors and typically respond to signals from within the cell that indicate whether transcription is required. In many human genes, several transcription factors may be needed before transcription can proceed efficiently. A transcription factor can cause either repression or activation of gene expression in eukaryotes.

During transcription, only one strand of the DNA is usually copied. This is called the template strand, and the RNA molecules produced are single-stranded. The DNA strand that would correspond to the mRNA is called the coding or sense strand, and it is not unusual for this to change from one gene to the next. In eukaryotes the initial product of transcription is called a pre-mRNA, which is extensively spliced before the mature mRNA is produced, ready for translation by the ribosome.

Translation

The process of translation uses the information present in the nucleotide sequence of mRNA to direct the synthesis of a specific protein for use by the cell. Translation takes place on the ribosomes—complex particles in the cell that contain RNA and protein. In prokaryotes the ribosomes are loaded onto the mRNA while transcription is still ongoing. Near the 5′ end of the mRNA, a short sequence of nucleotides signals the starting point for translation. It contains a few nucleotides called a ribosome binding site, or Shine-Dalgarno sequence. In E. coli the tetranucleotide GAGG is sufficient to serve as a binding site. This typically lies five to eight bases upstream of an initiation codon. The mRNA sequence is read three bases at a time from its 5′ end toward its 3′ end, and one amino acid is added to the growing chain from its respective aminoacyl tRNA, until the complete protein chain is assembled. Translation stops when the ribosome encounters a termination codon, normally UAG, UAA, or UGA. Special release factors associate with the ribosome in response to these codons, and the newly synthesized protein, tRNAs, and mRNA all dissociate. The ribosome then becomes available to interact with another mRNA molecule.

In eukaryotes the essence of protein synthesis is the same, although the ribosomes are more complicated. As with prokaryotic initiation, the signal sequence interacts with the 3′ end of the small subunit rRNA during formation of the initiation complex.

The issue of fidelity is important during protein synthesis, but it is not as crucial as fidelity during replication. One mRNA molecule can be translated repeatedly to give many copies of the protein. When an occasional protein is mistranslated, it usually does not fold properly and is then degraded by the cellular

machinery. However, proofreading mechanisms exist within the ribosome to ensure accurate pairing between the codon in the mRNA and the anticodon in the tRNA.

One of the crowning achievements of molecular biology was the elucidation during the 1960s of the genetic code. Principals in this effort were Har G. Khorana and Marshall W. Nirenberg, who shared a Nobel Prize in 1968. Khorana and Nirenberg used artificial templates and protein synthesizing systems in the test tube to determine the coding potential of all 64 possible triplet codons (see the Click Here to see full-size t