SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOINFORMATICS

UNIT – 1 -INTRODUCTION TO BIOINFORMATICS– SBIA1101

UNIT 1

Cell

The **cell** (from Latin *cella*, meaning "small room") is the basic structural, functional, and biological unit of all known organisms. A cell is the smallest unit of life. Cells are often called the "building blocks of life". The study of cells is called cell biology, cellular biology, or cytology.

Cells consist of cytoplasm enclosed within a membrane, which contains many biomolecules such as proteins and nucleic acids. Most plant and animal cells are only visible under a microscope, with dimensions between 1 and 100 micrometres. Organisms can be classified as unicellular (consisting of a single cell such as bacteria) or multicellular (including plants and animals). Most unicellular organisms are classed as microorganisms.

The number of cells in plants and animals varies from species to species; it has been estimated that humans contain somewhere around 40 trillion (4×10^{13}) cells. The human brain accounts for around 80 billion of these cells.

Cells were discovered by Robert Hooke in 1665, who named them for their resemblance to cells inhabited by Christian monks in a monastery. Cell theory, first developed in 1839 by Matthias Jakob Schleiden and Theodor Schwann, states that all organisms are composed of one or more cells, that cells are the fundamental unit of structure and function in all living organisms, and that all cells come from pre-existing cells. Cells emerged on Earth at least 3.5 billion years ago

Cell theory

In biology, **cell theory** is the historic scientific theory, now universally accepted, that living organisms are made up of cells, that they are the basic structural/organizational unit of all organisms, and that all cells come from pre-existing cells. Cells are the basic unit of structure in all organisms and also the basic unit of reproduction.

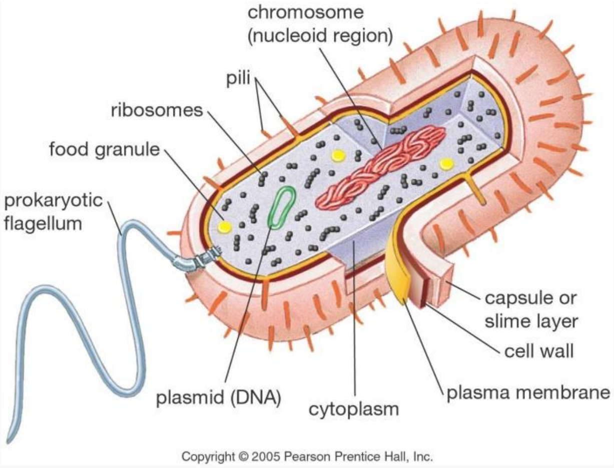
The three tenets to the cell theory are as described below:

1. All living organisms are composed of one or more cells.
2. The cell is the basic unit of structure and organization in organisms.
3. Cells arise from pre-existing cells.

There is no universally accepted definition of life. Some biologists consider non-cellular entities such as viruses living organisms,[1] and thus reasonably disagree with the first tenet.

Prokaryotic cells

Structure of a typical prokaryotic cell



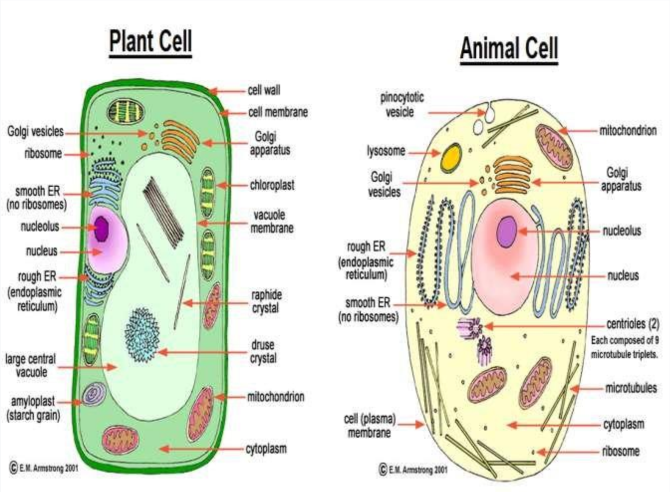
Copyright © 2005 Pearson Prentice Hall, Inc.

Prokaryotes include bacteria and archaea, two of the three domains of life. Prokaryotic cells were the first form of life on Earth, characterized by having vital biological processes including cell signaling. They are simpler and smaller than eukaryotic cells, and lack a nucleus, and other membrane-bound organelles. The DNA of a prokaryotic cell consists of a single circular chromosome that is in direct contact with the cytoplasm. The nuclear region in the cytoplasm is called the nucleoid. Most prokaryotes are the smallest of all organisms ranging from 0.5 to 2.0 µm in diameter.[13]

A prokaryotic cell has three regions:

- Enclosing the cell is the cell envelope – generally consisting of a plasma membrane covered by a cell wall which, for some bacteria, may be further covered by a third layer called a capsule. Though most prokaryotes have both a cell membrane and a cell wall, there are exceptions such as *Mycoplasma* (bacteria) and *Thermoplasma* (archaea) which only possess the cell membrane layer. The envelope gives rigidity to the cell and separates the interior of the cell from its environment, serving as a protective filter. The cell wall consists of peptidoglycan in bacteria, and acts as an additional barrier against exterior forces. It also prevents the cell from expanding and bursting (cytolysis) from osmotic pressure due to a hypotonic environment. Some eukaryotic cells (plant cells and fungal cells) also have a cell wall.
- Inside the cell is the cytoplasmic region that contains the genome (DNA), ribosomes and various sorts of inclusions. The genetic material is freely found in the cytoplasm. Prokaryotes can carry extrachromosomal DNA elements called plasmids, which are usually circular. Linear bacterial plasmids have been identified in several species of spirochete bacteria, including members of the genus *Borrelia* notably *Borrelia burgdorferi*, which causes Lyme disease.[14] Though not forming a nucleus, the DNA is condensed in a nucleoid. Plasmids encode additional genes, such as antibiotic resistance genes.

- On the outside, flagella and pili project from the cell's surface. These are structures (not present in all prokaryotes) made of proteins that facilitate movement and communication between cells.



Structure of a typical plant and animal cell

Eukaryotic cells

Plants, animals, fungi, slime moulds, protozoa, and algae are all eukaryotic. These cells are about fifteen times wider than a typical prokaryote and can be as much as a thousand times greater in volume. The main distinguishing feature of eukaryotes as compared to prokaryotes is compartmentalization: the presence of membrane-bound organelles (compartments) in which specific activities take place. Most important among these is a cell nucleus,[4] an organelle that houses the cell's DNA. This nucleus gives the eukaryote its name, which means "true kernel (nucleus)". Other differences include:

- The plasma membrane resembles that of prokaryotes in function, with minor differences in the setup. Cell walls may or may not be present.
- The eukaryotic DNA is organized in one or more linear molecules, called chromosomes, which are associated with histone proteins. All chromosomal DNA is stored in the cell nucleus, separated from the cytoplasm by a membrane.[4] Some eukaryotic organelles such as mitochondria also contain some DNA.
- Many eukaryotic cells are ciliated with primary cilia. Primary cilia play important roles in chemosensation, mechanosensation, and thermosensation. Each cilium may thus be "viewed as a sensory cellular antennae that coordinates a large number of cellular signaling pathways, sometimes coupling the signaling to ciliary motility or alternatively to cell division and differentiation."[15]
- Motile eukaryotes can move using motile cilia or flagella. Motile cells are absent in conifers and flowering plants.[16] Eukaryotic flagella are more complex than those of prokaryotes.

Comparison of features of prokaryotic and eukaryotic cells		
	Prokaryotes	**Eukaryotes**
Typical organisms	bacteria, archaea	protists, fungi, plants, animals
Typical size	~ 1–5 µm[18]	~ 10–100 µm[18]
Type of nucleus	nucleoid region; no true nucleus	true nucleus with double membrane
DNA	circular (usually)	linear molecules (chromosomes) with histone proteins
RNA/protein synthesis	coupled in the cytoplasm	RNA synthesis in the nucleus protein synthesis in the cytoplasm
Ribosomes	50S and 30S	60S and 40S
Cytoplasmic structure	very few structures	highly structured by endomembranes and a cytoskeleton
Cell movement	flagella made of flagellin	flagella and cilia containing microtubules; lamellipodia and filopodia containing actin
Mitochondria	none	one to several thousand
Chloroplasts	none	in algae and plants
Organization	usually single cells	single cells, colonies, higher multicellular organisms with specialized cells
Cell division	binary fission (simple division)	mitosis (fission or budding) meiosis
Chromosomes	single chromosome	more than one chromosome
Membranes	cell membrane	Cell membrane and membrane-bound organelles

Subcellular components

All cells, whether prokaryotic or eukaryotic, have a membrane that envelops the cell, regulates what moves in and out (selectively permeable), and maintains the electric potential of the cell. Inside the membrane, the cytoplasm takes up most of the cell's volume. All cells (except red blood cells which lack a cell nucleus and most organelles to accommodate maximum space for hemoglobin) possess DNA, the hereditary material of genes, and RNA, containing the information necessary to build various proteins such as enzymes, the cell's primary machinery. There are also other kinds of biomolecules in cells. This article lists these primary cellular components, then briefly describes their function.

Membrane

The cell membrane, or plasma membrane, is a biological membrane that surrounds the cytoplasm of a cell. In animals, the plasma membrane is the outer boundary of the cell, while in plants and prokaryotes it is usually covered by a cell wall. This membrane serves to separate and protect a cell from its surrounding environment and is made mostly from a double layer of phospholipids, which are amphiphilic (partly hydrophobic and partly hydrophilic). Hence, the layer is called a phospholipid bilayer, or sometimes a fluid mosaic membrane. Embedded within this membrane is a macromolecular structure called the porosome the universal secretory portal in cells and a variety of protein molecules that act as channels and pumps that move different molecules into and out of the cell.[4] The membrane is semi-permeable, and selectively permeable, in that it can either let a substance (molecule or ion) pass through freely, pass through to a limited extent or not pass through at all. Cell surface membranes also contain receptor proteins that allow cells to detect external signaling molecules such as hormones.

Cytoskeleton

The cytoskeleton acts to organize and maintain the cell's shape; anchors organelles in place; helps during endocytosis, the uptake of external materials by a cell, and cytokinesis, the separation of daughter cells after cell division; and moves parts of the cell in processes of growth and mobility. The eukaryotic cytoskeleton is composed of microfilaments, intermediate filaments and microtubules. There are a great number of proteins associated with them, each controlling a cell's structure by directing, bundling, and aligning filaments.[4] The prokaryotic cytoskeleton is less well-studied but is involved in the maintenance of cell shape, polarity and cytokinesis.[19] The subunit protein of microfilaments is a small, monomeric protein called actin. The subunit of microtubules is a dimeric molecule called tubulin. Intermediate filaments are heteropolymers whose subunits vary among the cell types in different tissues. But some of the subunit protein of intermediate filaments include vimentin, desmin, lamin (lamins A, B and C), keratin (multiple acidic and basic keratins), neurofilament proteins (NF–L, NF–M).

Genetic material

Two different kinds of genetic material exist: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Cells use DNA for their long-term information storage. The biological information contained in an organism is encoded in its DNA sequence.[4] RNA is used for information transport (e.g., mRNA) and enzymatic functions (e.g., ribosomal RNA). Transfer RNA (tRNA) molecules are used to add amino acids during protein translation.

Prokaryotic genetic material is organized in a simple circular bacterial chromosome in the nucleoid region of the cytoplasm. Eukaryotic genetic material is divided into different,[4] linear molecules called chromosomes inside a discrete nucleus, usually with additional genetic material in some organelles like mitochondria and chloroplasts (see endosymbiotic theory).

A human cell has genetic material contained in the cell nucleus (the nuclear genome) and in the mitochondria (the mitochondrial genome). In humans the nuclear genome is divided into 46 linear DNA molecules called chromosomes, including 22 homologous chromosome pairs and a pair of sex chromosomes. The mitochondrial genome is a circular DNA molecule distinct from the nuclear DNA. Although the mitochondrial DNA is very small compared to nuclear chromosomes,[4] it codes for 13 proteins involved in mitochondrial energy production and specific tRNAs.

Foreign genetic material (most commonly DNA) can also be artificially introduced into the cell by a process called transfection. This can be transient, if the DNA is not inserted into the cell's genome, or stable, if it is. Certain viruses also insert their genetic material into the genome.

Organelles

Organelles are parts of the cell which are adapted and/or specialized for carrying out one or more vital functions, analogous to the organs of the human body (such as the heart, lung, and kidney, with each organ performing a different function).[4] Both eukaryotic and prokaryotic cells have organelles, but prokaryotic organelles are generally simpler and are not membrane-bound.

There are several types of organelles in a cell. Some (such as the nucleus and golgi apparatus) are typically solitary, while others (such as mitochondria, chloroplasts, peroxisomes and lysosomes) can be numerous (hundreds to thousands). The cytosol is the gelatinous fluid that fills the cell and surrounds the organelles.

Eukaryotic

- **Cell nucleus**: A cell's information center, the cell nucleus is the most conspicuous organelle found in a eukaryotic cell. It houses the cell's chromosomes, and is the place where almost all DNA replication and RNA synthesis (transcription) occur. The nucleus is spherical and separated from the cytoplasm by a double membrane called the nuclear envelope. The nuclear envelope isolates and protects a cell's DNA from various molecules that could accidentally damage its structure or interfere with its processing. During processing, DNA is transcribed, or copied into a special RNA, called messenger RNA (mRNA). This mRNA is then transported out of the nucleus, where it is translated into a specific protein molecule. The nucleolus is a specialized region within the nucleus where ribosome subunits are assembled. In prokaryotes, DNA processing takes place in the cytoplasm.[4]
- **Mitochondria and Chloroplasts**: generate energy for the cell. Mitochondria are self-replicating organelles that occur in various numbers, shapes, and sizes in the cytoplasm of all eukaryotic cells.[4] Respiration occurs in the cell mitochondria, which generate the cell's energy by oxidative phosphorylation, using oxygen to release energy stored in cellular nutrients (typically pertaining to glucose) to generate ATP. Mitochondria multiply by binary fission, like prokaryotes. Chloroplasts can only be found in plants and algae, and they capture the sun's energy to make carbohydrates through photosynthesis.

- **Endoplasmic reticulum**: The endoplasmic reticulum (ER) is a transport network for molecules targeted for certain modifications and specific destinations, as compared to molecules that float freely in the cytoplasm. The ER has two forms: the rough ER, which has ribosomes on its surface that secrete proteins into the ER, and the smooth ER, which lacks ribosomes.[4] The smooth ER plays a role in calcium sequestration and release.
- **Golgi apparatus**: The primary function of the Golgi apparatus is to process and package the macromolecules such as proteins and lipids that are synthesized by the cell.
- **Lysosomes and Peroxisomes**: Lysosomes contain digestive enzymes (acid hydrolases). They digest excess or worn-out organelles, food particles, and

engulfed viruses or bacteria. Peroxisomes have enzymes that rid the cell of toxic peroxides. The cell could not house these destructive enzymes if they were not contained in a membrane-bound system.[4]

- **Centrosome**: the cytoskeleton organiser: The centrosome produces the microtubules of a cell – a key component of the cytoskeleton. It directs the transport through the ER and the Golgi apparatus. Centrosomes are composed of two centrioles, which separate during cell division and help in the formation of the mitotic spindle. A single centrosome is present in the animal cells. They are also found in some fungi and algae cells.
- **Vacuoles**: Vacuoles sequester waste products and in plant cells store water. They are often described as liquid filled space and are surrounded by a membrane. Some cells, most notably *Amoeba*, have contractile vacuoles, which can pump water out of the cell if there is too much water. The vacuoles of plant cells and fungal cells are usually larger than those of animal cells.

Eukaryotic and prokaryotic

- **Ribosomes**: The ribosome is a large complex of RNA and protein molecules.[4] They each consist of two subunits, and act as an assembly line where RNA from the nucleus is used to synthesise proteins from amino acids. Ribosomes can be found either floating freely or bound to a membrane (the rough endoplasmatic reticulum in eukaryotes, or the cell membrane in prokaryotes)

Structures outside the cell membrane

Many cells also have structures which exist wholly or partially outside the cell membrane. These structures are notable because they are not protected from the external environment by the semipermeable cell membrane. In order to assemble these structures, their components must be carried across the cell membrane by export processes.

Cell wall

Many types of prokaryotic and eukaryotic cells have a cell wall. The cell wall acts to protect the cell mechanically and chemically from its environment, and is an additional layer of protection to the cell membrane. Different types of cell have cell walls made up of different materials; plant cell walls are primarily made up of cellulose, fungi cell walls are made up of chitin and bacteria cell walls are made up of peptidoglycan.

Prokaryotic

Capsule

A gelatinous capsule is present in some bacteria outside the cell membrane and cell wall. The capsule may be polysaccharide as in pneumococci, meningococci or polypeptide as *Bacillus anthracis* or hyaluronic acid as in streptococci. Capsules are not marked by normal staining protocols and can be detected by India ink or methyl blue; which allows for higher contrast between the cells for observation.[21]:87

Flagella

Flagella are organelles for cellular mobility. The bacterial flagellum stretches from cytoplasm through the cell membrane(s) and extrudes through the cell wall. They are long and thick thread-like appendages, protein in nature. A different type of flagellum is found in archaea and a different type is found in eukaryotes.

Fimbriae

A fimbria (plural fimbriae also known as a pilus, plural pili) is a short, thin, hair-like filament found on the surface of bacteria. Fimbriae are formed of a protein called pilin (antigenic) and are responsible for the attachment of bacteria to specific receptors on human cells (cell adhesion). There are special types of pili involved in bacterial conjugation.

DNA

Deoxyribonucleic acid (/diːˈɒksɪˌraɪboʊnjuːˌkliːɪk, -ˌkleɪ-/ (🔊listen);[1] DNA) is a molecule composed of two polynucleotide chains that coil around each other to form a double helix carrying genetic instructions for the development, functioning, growth and reproduction of all known organisms and many viruses. DNA and ribonucleic acid (RNA) are nucleic acids. Alongside proteins, lipids and complex carbohydrates (polysaccharides), nucleic acids are one of the four major types of macromolecules that are essential for all known forms of life.

The two DNA strands are known as polynucleotides as they are composed of simpler monomeric units called nucleotides.[2][3] Each nucleotide is composed of one of four nitrogen-containing nucleobases (cytosine [C], guanine [G], adenine [A] or thymine [T]), a sugar called deoxyribose, and a phosphate group. The nucleotides are joined to one another in a chain by covalent bonds (known as the phospho-diester linkage) between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating sugar-phosphate backbone. The nitrogenous bases of the two separate polynucleotide strands are bound together, according to base pairing rules (A with T and C with G), with hydrogen bonds to make double-stranded DNA. The complementary nitrogenous bases are divided into two groups, pyrimidines and purines. In DNA, the pyrimidines are thymine and cytosine; the purines are adenine and guanine.

Both strands of double-stranded DNA store the same biological information. This information is replicated as and when the two strands separate. A large part of DNA (more than 98% for humans) is non-coding, meaning that these sections do not serve as patterns for protein sequences. The two strands of DNA run in opposite directions to each other and are thus antiparallel. Attached to each sugar is one of four types of nucleobases (informally, *bases*). It is the sequence of these four nucleobases along the backbone that encodes genetic information. RNA strands are created using DNA strands as a template in a process called transcription, where DNA bases are exchanged for their corresponding bases except in the case of thymine (T), for which RNA substitutes uracil (U).[4] Under the genetic code, these RNA strands specify the sequence of amino acids within proteins in a process called translation.

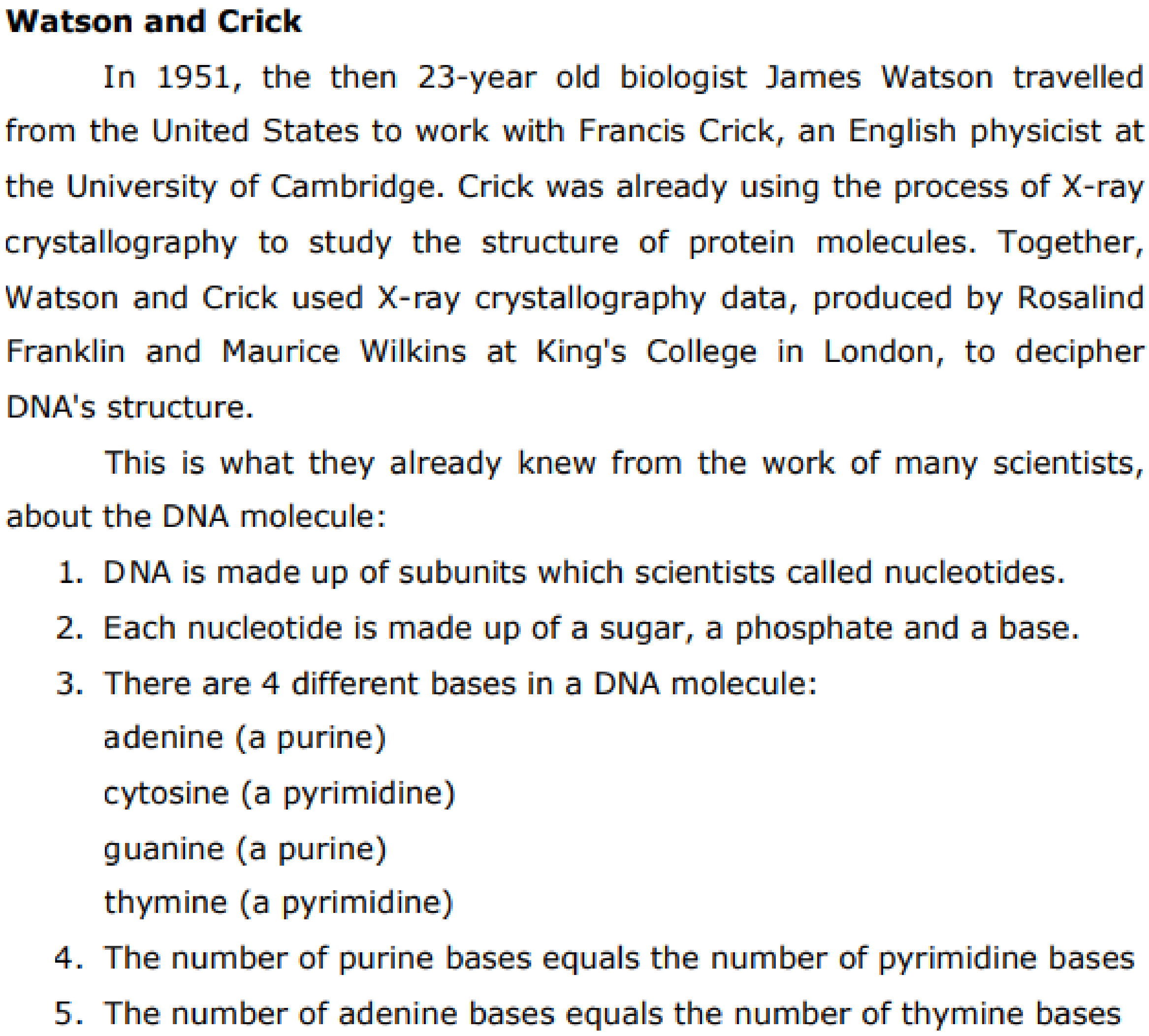
Within eukaryotic cells, DNA is organized into long structures called chromosomes. Before typical cell division, these chromosomes are duplicated in the process of DNA replication, providing a complete set of chromosomes for each daughter cell. Eukaryotic organisms (animals, plants, fungi and protists) store most of their DNA inside the cell nucleus as nuclear DNA, and some in the mitochondria as mitochondrial DNA or in chloroplasts as chloroplast DNA.[5] In contrast, prokaryotes (bacteria and archaea) store their DNA only in the cytoplasm, in circular chromosomes. Within eukaryotic chromosomes, chromatin proteins, such as histones, compact and organize DNA. These compacting structures guide the interactions between DNA and other proteins, helping control which parts of the DNA are transcribed.

DNA was first isolated by Friedrich Miescher in 1869. Its molecular structure was first identified by Francis Crick and James Watson at the Cavendish Laboratory within the University of Cambridge in 1953, whose model-building efforts were guided by X-ray diffraction data acquired by Raymond Gosling, who was a post-graduate student of Rosalind Franklin at King's College London. DNA is used by researchers as a molecular tool to explore physical laws and theories, such as the ergodic theorem and the theory of elasticity. The unique material properties of DNA have made it an attractive molecule for material scientists and engineers interested in micro- and nano-fabrication. Among notable advances in this field are DNA origami and DNA-based hybrid materials.

Sense and antisense

A DNA sequence is called a "sense" sequence if it is the same as that of a messenger RNA copy that is translated into protein. The sequence on the opposite strand is called the "antisense" sequence. Both sense and antisense sequences can exist on different parts of the same strand of DNA (i.e. both strands can contain both sense and antisense sequences). In both prokaryotes and eukaryotes, antisense RNA sequences are produced, but the functions of these RNAs are not entirely clear.[33] One proposal is that antisense RNAs are involved in regulating gene expression through RNA-RNA base pairing.

A few DNA sequences in prokaryotes and eukaryotes, and more in plasmids and viruses, blur the distinction between sense and antisense strands by having overlapping genes. In these cases, some DNA sequences do double duty, encoding one protein when read along one strand, and a second protein when read in the opposite direction along the other strand. In bacteria, this overlap may be involved in the regulation of gene transcription, while in viruses, overlapping genes increase the amount of information that can be encoded within the small viral genome



Watson and Crick

In 1951, the then 23-year old biologist James Watson travelled from the United States to work with Francis Crick, an English physicist at the University of Cambridge. Crick was already using the process of X-ray crystallography to study the structure of protein molecules. Together, Watson and Crick used X-ray crystallography data, produced by Rosalind Franklin and Maurice Wilkins at King's College in London, to decipher DNA's structure.

This is what they already knew from the work of many scientists, about the DNA molecule:

1. D NA is made up of subunits which scientists called nucleotides.
2. Each nucleotide is made up of a sugar, a phosphate and a base.
3. There are 4 different bases in a DNA molecule:
   adenine (a purine)
   cytosine (a pyrimidine)
   guanine (a purine)
   thymine (a pyrimidine)
4. The number of purine bases equals the number of pyrimidine bases
5. The number of adenine bases equals the number of thymine bases

6. The number of guanine bases equals the number of cytosine bases
7. The basic structure of the DNA molecule is helical, with the bases being stacked on top of each other

Components of DNA

DNA is a polymer. The monomer units of DNA are nucleotides, and the polymer is known as a "polynucleotide". Each nucleotide consists of a 5-carbon sugar (deoxyribose), a nitrogen containing base attached to the sugar, and a phosphate group. There are four different types of nucleotides found in DNA, differing only in the nitrogenous base. The four nucleotides are given one letter abbreviations as shorthand for the four bases.

- A is for adenine
- G is for guanine
- C is for cytosine
- T is for thymine

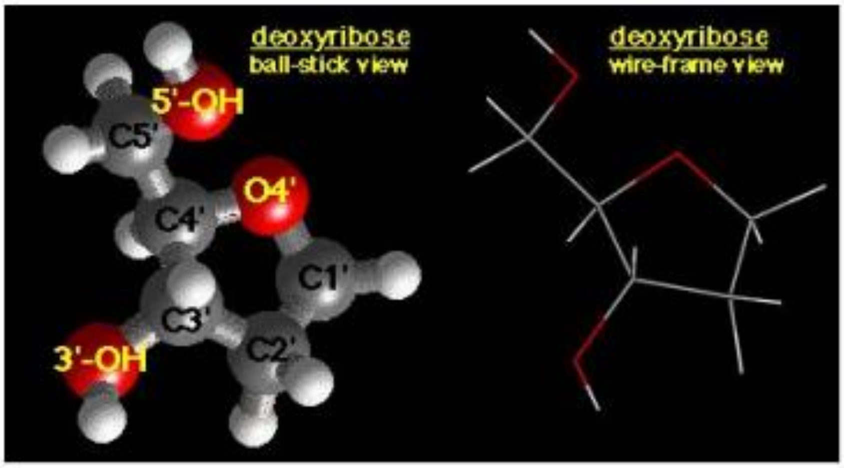
Purine Bases

Adenine and guanine are purines. Purines are the larger of the two types of bases found in DNA. Structures are shown below:

The 9 atoms that make up the fused rings (5 carbon, 4 nitrogen) are numbered 1-9. All ring atoms lie in the same plane.

Deoxyribose Sugar

The deoxyribose sugar of the DNA backbone has 5 carbons and 3 oxygens. The carbon atoms are numbered 1', 2', 3', 4', and 5' to distinguish from the numbering of the atoms of the purine and pyrmidine rings. The hydroxyl groups on the 5'- and 3'- carbons link to the phosphate groups to form the DNA backbone. Deoxyribose lacks an hydroxyl group at the 2'-position when compared to ribose, the sugar component of RNA.



Nucleosides

A nucleoside is one of the four DNA bases covalently attached to the C1' position of a sugar. The sugar in deoxynucleosides is 2'-deoxyribose. The sugar in ribonucleosides is ribose. Nucleosides differ from nucleotides in that they lack phosphate groups. The four different nucleosides of DNA are deoxyadenosine (dA), deoxyguanosine (dG), deoxycytosine (dC), and (deoxy)thymidine (dT, or T).

In dA and dG, there is an "N-glycoside" bond between the sugar C1' and N9 of the purine.

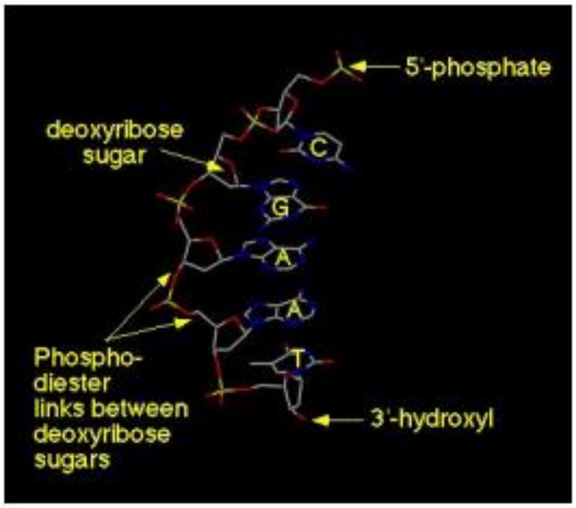
Nucleotides

A nucleotide is a nucleoside with one or more phosphate groups covalently attached to the 3'- and/or 5'-hydroxyl group(s).

DNA Backbone

The DNA backbone is a polymer with an alternating sugar-phosphate sequence. The deoxyribose sugars are joined at both the 3'-hydroxyl and 5'-hydroxyl groups to phosphate groups in ester links, also known as "phosphodiester" bonds.

Example of DNA Backbone: 5'-d (CGAAT)



Features of the 5'-d(CGAAT) structure:

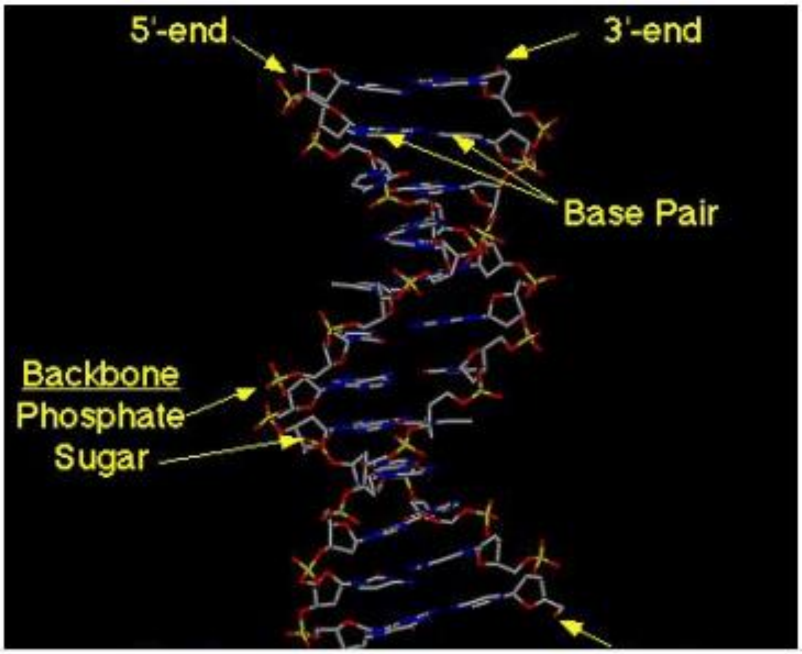
- Alternating backbone of deoxyribose and phosphodiester groups

- Chain has a direction (known as polarity), 5'- to 3'- from top to bottom
- Oxygens (red atoms) of phosphates are polar and negatively charged
- A, G, C, and T bases can extend away from chain, and stack atop each other
- Bases are hydrophobic

DNA Double Helix

DNA is a normally double stranded macromolecule. Two polynucleotide chains, held together by weak thermodynamic forces, form a DNA molecule.

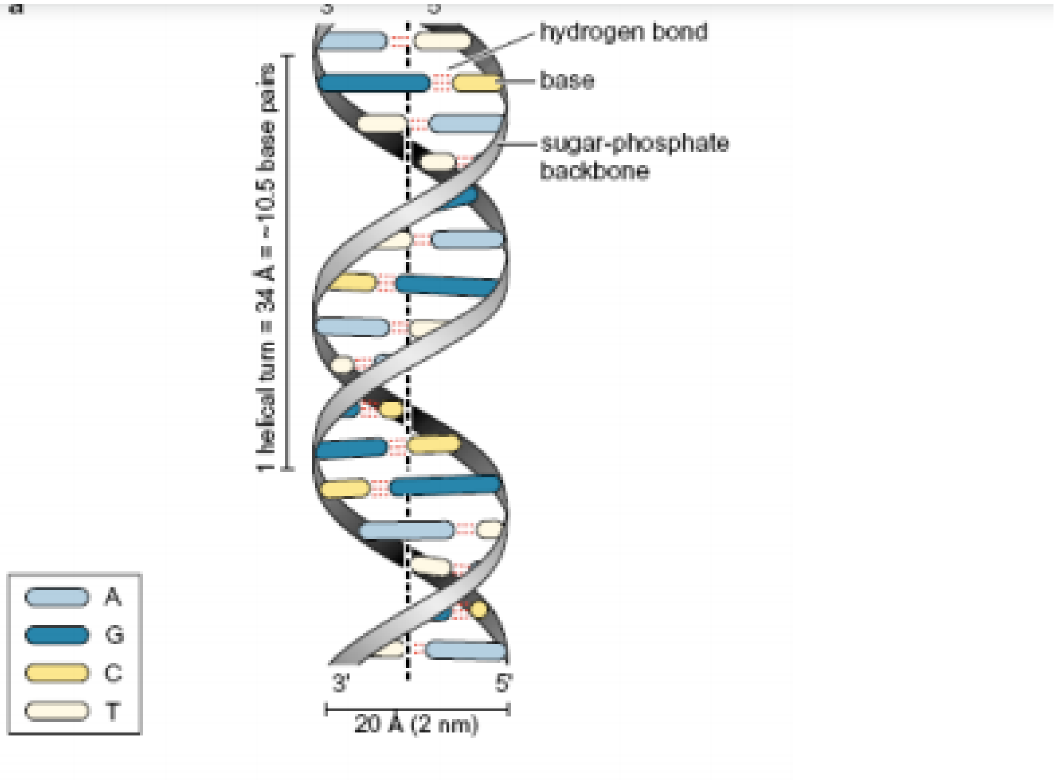
Structure of DNA Double Helix

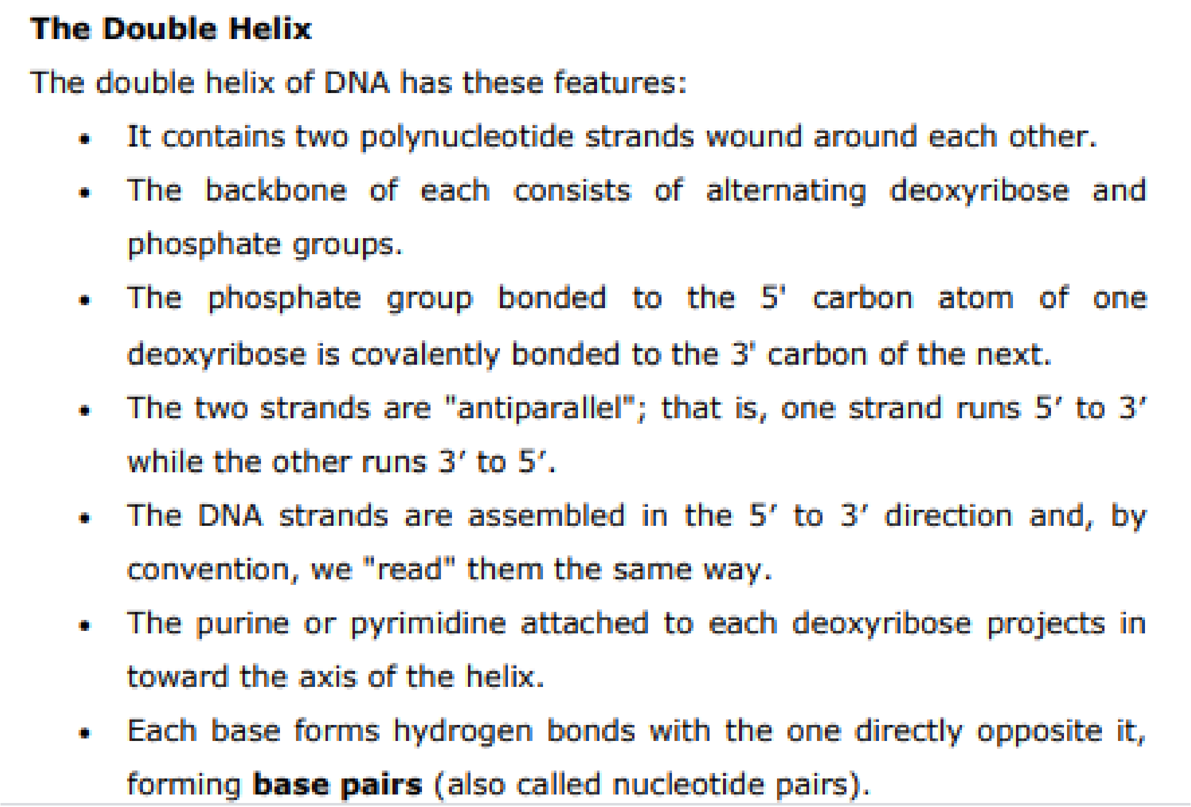


Features of the DNA Double Helix

- Two DNA strands form a helical spiral, winding around a helix axis in a right-handed spiral
- The two polynucleotide chains run in opposite directions
- The sugar-phosphate backbones of the two DNA strands wind around the helix axis like the railing of a sprial staircase

The bases of the individual nucleotides are on the inside of the helix, stacked on top of each other like the steps of a spiral staircase.

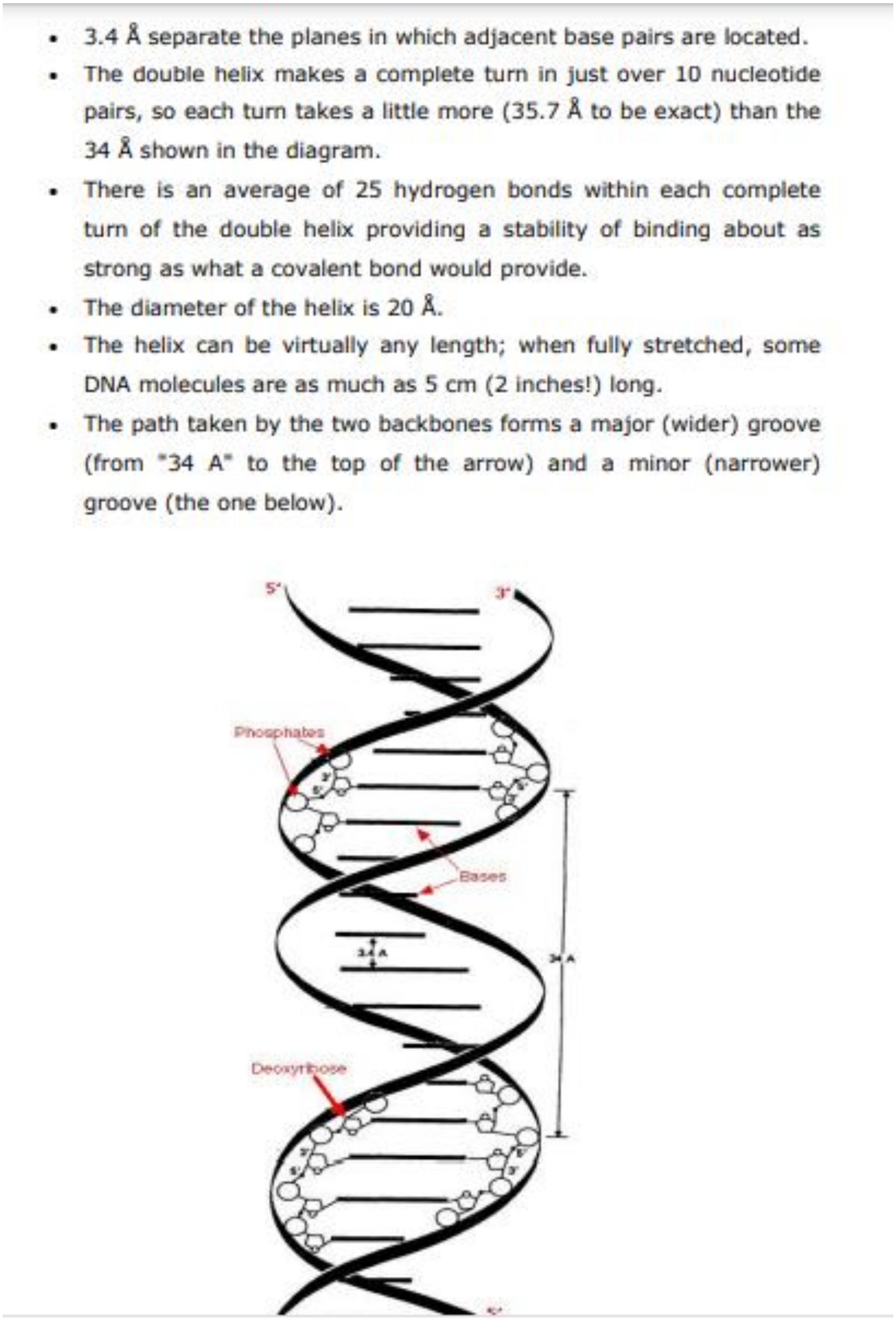




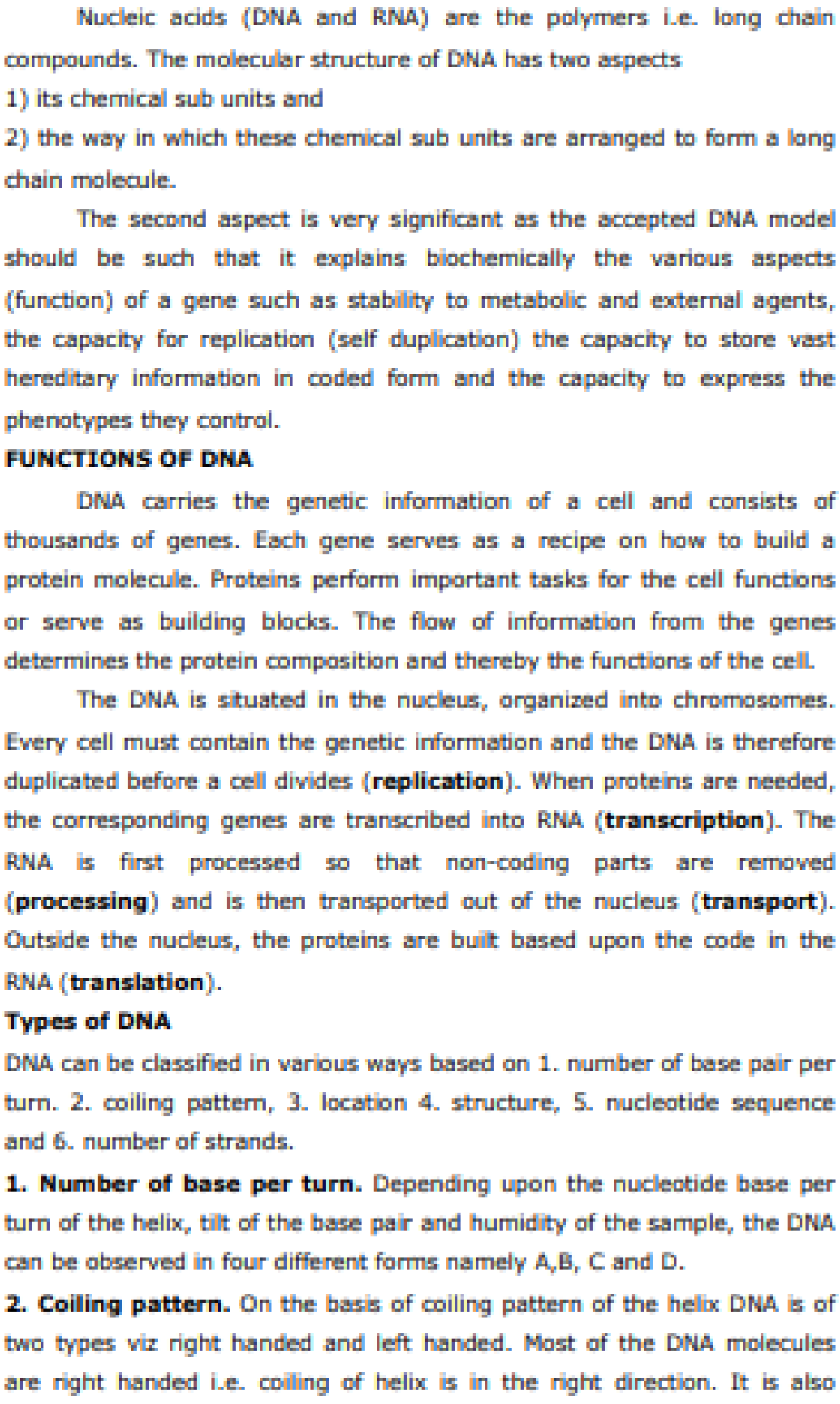
The Double Helix

The double helix of DNA has these features:

- It contains two polynucleotide strands wound around each other.
- The backbone of each consists of alternating deoxyribose and phosphate groups.
- The phosphate group bonded to the 5' carbon atom of one deoxyribose is covalently bonded to the 3' carbon of the next.
- The two strands are "antiparallel"; that is, one strand runs 5′ to 3′ while the other runs 3′ to 5′.
- The DNA strands are assembled in the 5′ to 3′ direction and, by convention, we "read" them the same way.
- The purine or pyrimidine attached to each deoxyribose projects in toward the axis of the helix.
- Each base forms hydrogen bonds with the one directly opposite it, forming **base pairs** (also called nucleotide pairs).



- 3.4 Å separate the planes in which adjacent base pairs are located.
- The double helix makes a complete turn in just over 10 nucleotide pairs, so each turn takes a little more (35.7 Å to be exact) than the 34 Å shown in the diagram.
- There is an average of 25 hydrogen bonds within each complete turn of the double helix providing a stability of binding about as strong as what a covalent bond would provide.
- The diameter of the helix is 20 Å.
- The helix can be virtually any length; when fully stretched, some DNA molecules are as much as 5 cm (2 inches!) long.
- The path taken by the two backbones forms a major (wider) groove (from "34 A" to the top of the arrow) and a minor (narrower) groove (the one below).



Nucleic acids (DNA and RNA) are the polymers i.e. long chain compounds. The molecular structure of DNA has two aspects

1) its chemical sub units and

2) the way in which these chemical sub units are arranged to form a long chain molecule.

The second aspect is very significant as the accepted DNA model should be such that it explains biochemically the various aspects (function) of a gene such as stability to metabolic and external agents, the capacity for replication (self duplication) the capacity to store vast hereditary information in coded form and the capacity to express the phenotypes they control.

FUNCTIONS OF DNA

DNA carries the genetic information of a cell and consists of thousands of genes. Each gene serves as a recipe on how to build a protein molecule. Proteins perform important tasks for the cell functions or serve as building blocks. The flow of information from the genes determines the protein composition and thereby the functions of the cell.

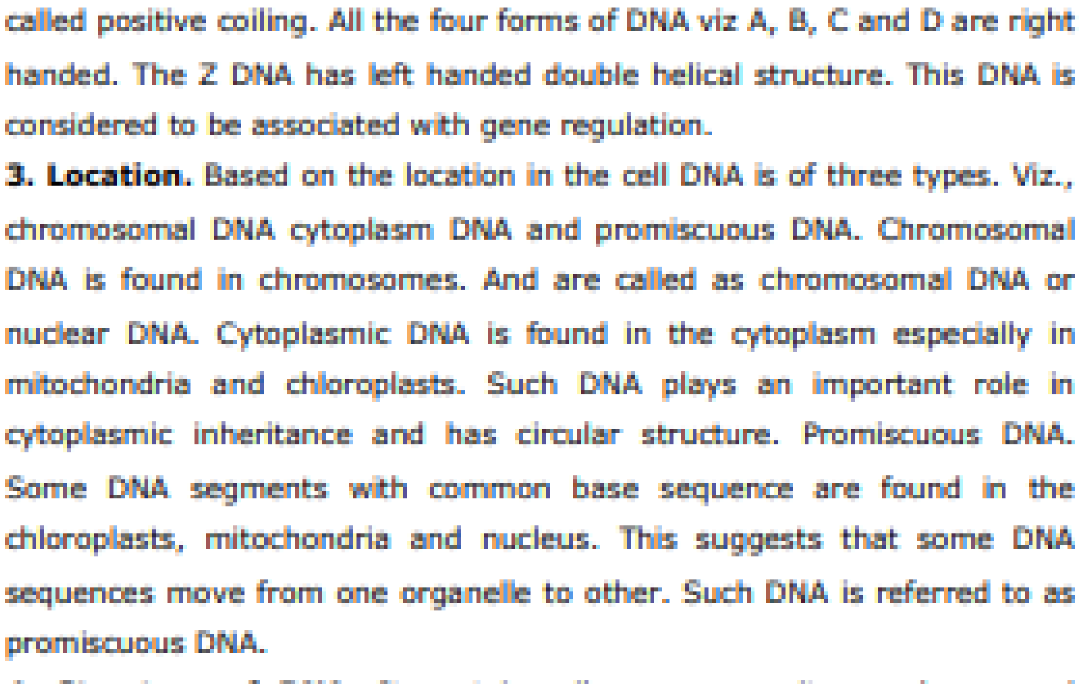
The DNA is situated in the nucleus, organized into chromosomes. Every cell must contain the genetic information and the DNA is therefore duplicated before a cell divides (**replication**). When proteins are needed, the corresponding genes are transcribed into RNA (**transcription**). The RNA is first processed so that non-coding parts are removed (**processing**) and is then transported out of the nucleus (**transport**). Outside the nucleus, the proteins are built based upon the code in the RNA (**translation**).

Types of DNA

DNA can be classified in various ways based on 1. number of base pair per turn. 2. coiling pattern, 3. location 4. structure, 5. nucleotide sequence and 6. number of strands.

1. Number of base per turn. Depending upon the nucleotide base per turn of the helix, tilt of the base pair and humidity of the sample, the DNA can be observed in four different forms namely A,B, C and D.

2. Coiling pattern. On the basis of coiling pattern of the helix DNA is of two types viz right handed and left handed. Most of the DNA molecules are right handed i.e. coiling of helix is in the right direction. It is also



called positive coiling. All the four forms of DNA viz A, B, C and D are right handed. The Z DNA has left handed double helical structure. This DNA is considered to be associated with gene regulation.

3. Location. Based on the location in the cell DNA is of three types. Viz., chromosomal DNA cytoplasm DNA and promiscuous DNA. Chromosomal DNA is found in chromosomes. And are called as chromosomal DNA or nuclear DNA. Cytoplasmic DNA is found in the cytoplasm especially in mitochondria and chloroplasts. Such DNA plays an important role in cytoplasmic inheritance and has circular structure. Promiscuous DNA. Some DNA segments with common base sequence are found in the chloroplasts, mitochondria and nucleus. This suggests that some DNA sequences move from one organelle to other. Such DNA is referred to as promiscuous DNA.

RNA- Properties, Structure, Types and Functions

- RNA or ribonucleic acid is a polymer of nucleotides which is made up of a ribose sugar, a phosphate, and bases such as adenine, guanine, cytosine, and uracil.
- It is a polymeric molecule essential in various biological roles in coding, decoding, regulation, and expression of genes.
-

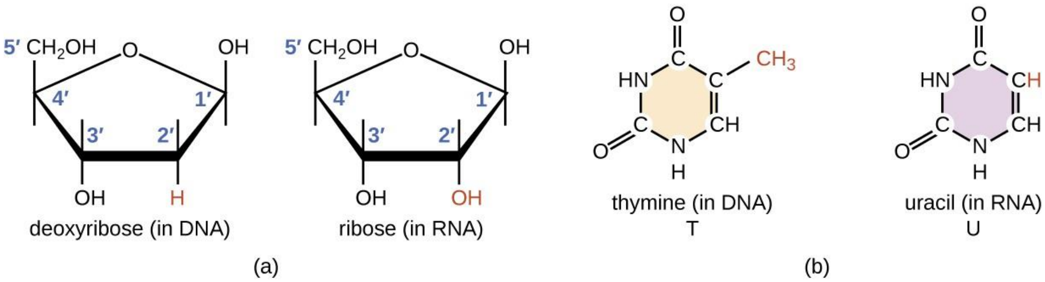
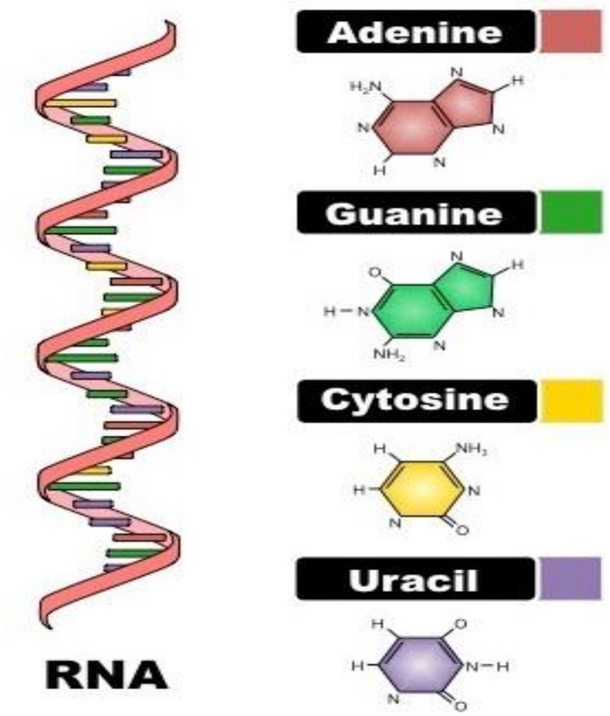


Figure: (a) Ribonucleotides contain the pentose sugar ribose instead of the deoxyribose found in deoxyribonucleotides. (b) RNA contains the pyrimidine uracil in place of thymine found in DNA.

RNA STRUCTURE



Like **DNA**, RNA is a long polymer consisting of nucleotides.

- RNA is a single-stranded helix.
- The strand has a 5′end (with a phosphate group) and a 3′end (with a hydroxyl group).
- It is composed of ribonucleotides.
- The ribonucleotides are linked together by 3′ –> 5′ phosphodiester bonds.
- The nitrogenous bases that compose the ribonucleotides include adenine, cytosine, uracil, and guanine.

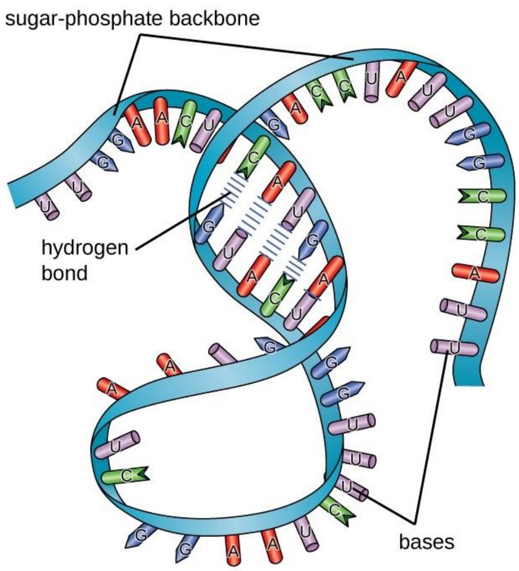
Thus, the difference in the structure of RNA from that of DNA include:

- The bases in RNA are adenine (abbreviated A), guanine (G), uracil (U) andcytosine (C).

Thus thymine in DNA is replaced by uracil in RNA, a different pyrimidine. However, like thymine, uracil can form base pairs with adenine.

- The sugar in RNA is ribose rather than deoxyribose as in DNA.
- The corresponding ribonucleosides are adenosine, guanosine, cytidine and uridine. The corresponding ribonucleotides are adenosine 5’-triphosphate (ATP), guanosine 5’-triphosphate (GTP), cytidine 5’-triphosphate (CTP) and uridine 5’-triphosphate (UTP).

RNA Secondary structure

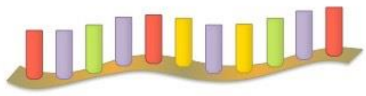


- Most RNA molecules are single-stranded but an RNA molecule may contain regions which can form complementary base pairing where the RNA strand loops back on itself.
- If so, the RNA will have some double-stranded regions.
- Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) exhibit substantial secondary structure, as do some messenger RNAs (mRNAs).

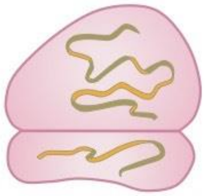
Types of RNA

In prokaryotes and eukaryotes, there are three main types of RNA.

- rRNA
- mRNA
- tRNA

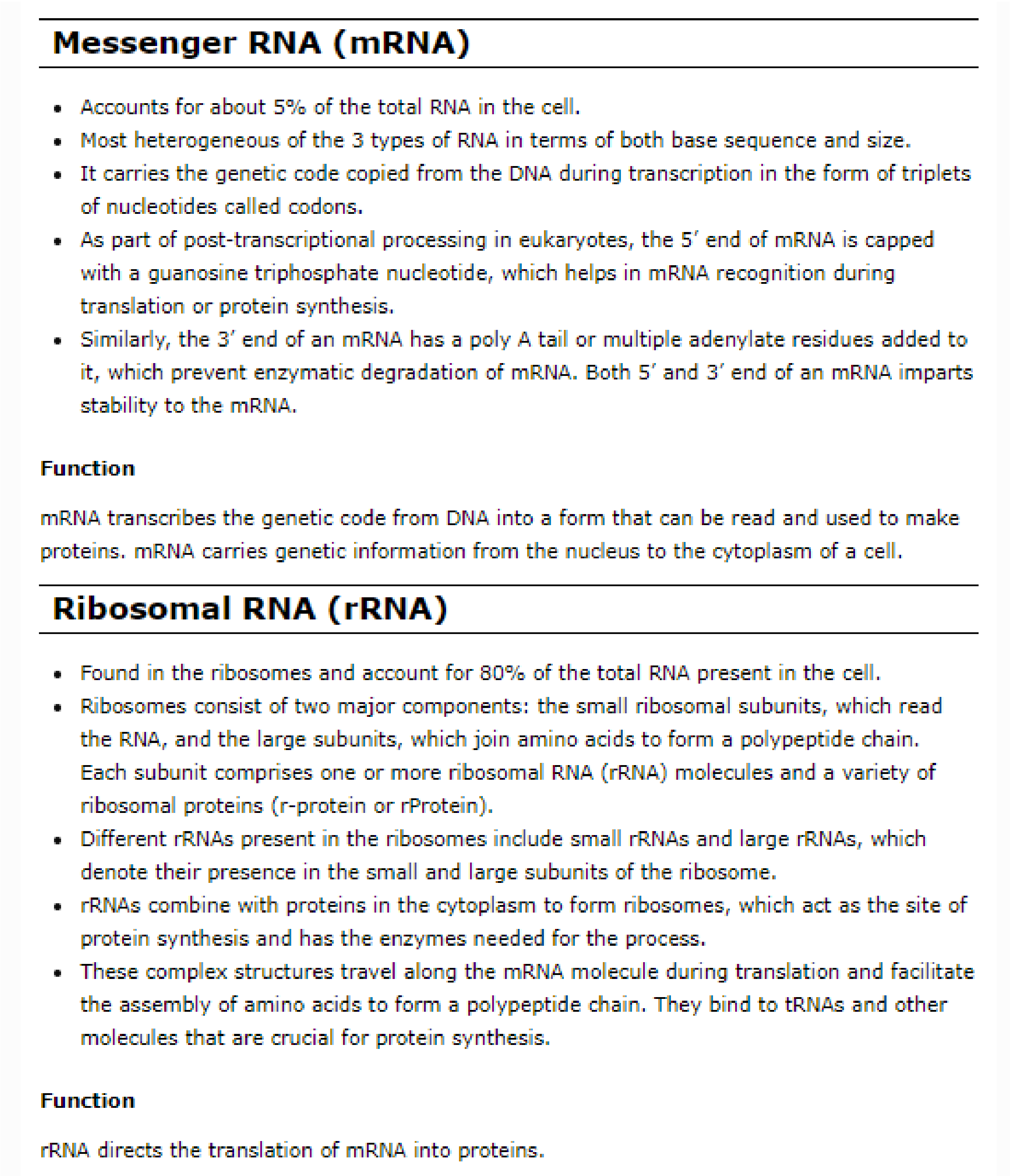


Messenger RNA (mRNA)



Ribosomal RNA (rRNA)

Transfer RNA (tRNA)



Messenger RNA (mRNA)

- Accounts for about 5% of the total RNA in the cell.
- Most heterogeneous of the 3 types of RNA in terms of both base sequence and size.
- It carries the genetic code copied from the DNA during transcription in the form of triplets of nucleotides called codons.
- As part of post-transcriptional processing in eukaryotes, the 5′ end of mRNA is capped with a guanosine triphosphate nucleotide, which helps in mRNA recognition during translation or protein synthesis.
- Similarly, the 3′ end of an mRNA has a poly A tail or multiple adenylate residues added to it, which prevent enzymatic degradation of mRNA. Both 5′ and 3′ end of an mRNA imparts stability to the mRNA.

Function

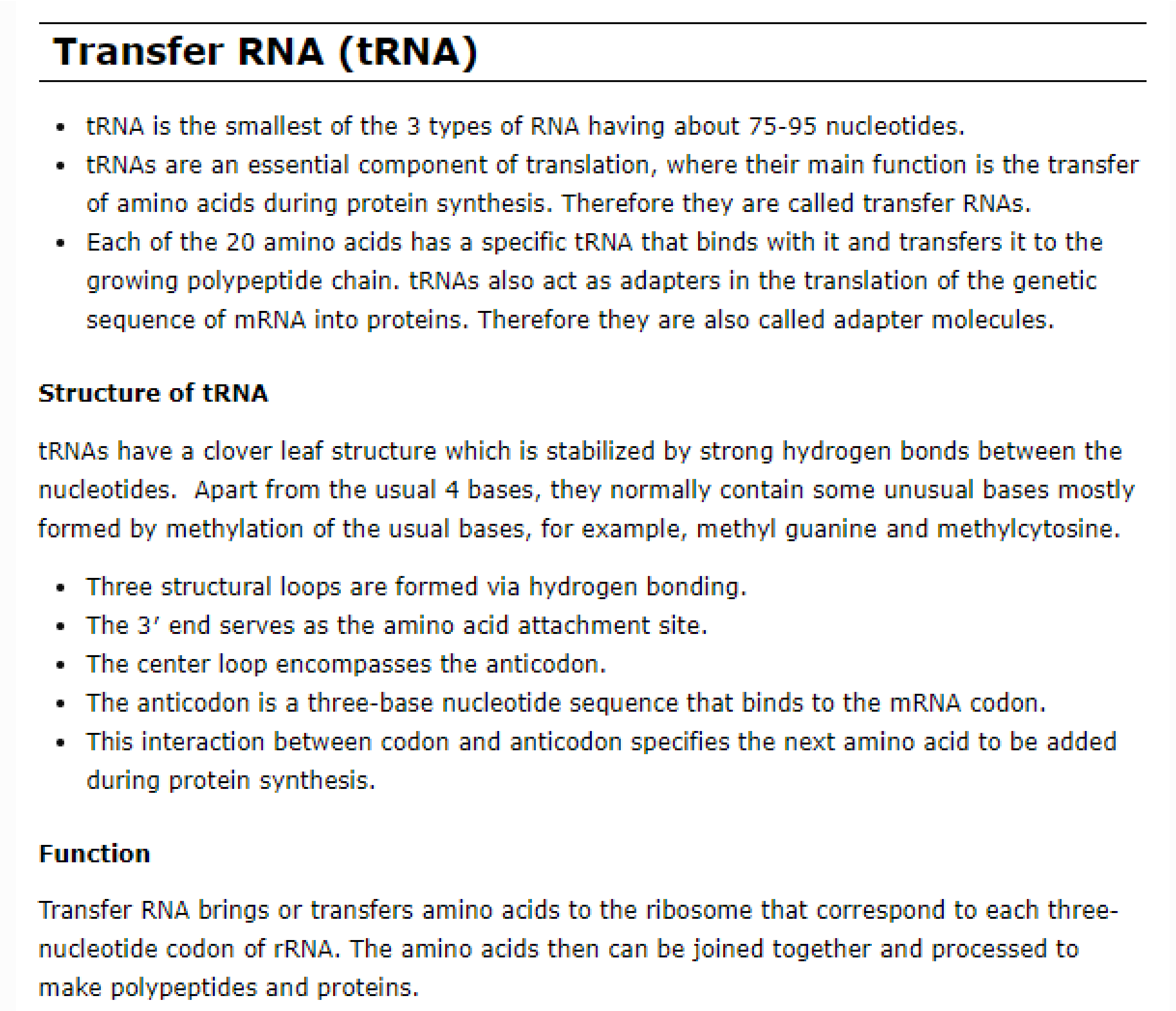
mRNA transcribes the genetic code from DNA into a form that can be read and used to make proteins. mRNA carries genetic information from the nucleus to the cytoplasm of a cell.

Ribosomal RNA (rRNA)

- Found in the ribosomes and account for 80% of the total RNA present in the cell.
- Ribosomes consist of two major components: the small ribosomal subunits, which read the RNA, and the large subunits, which join amino acids to form a polypeptide chain. Each subunit comprises one or more ribosomal RNA (rRNA) molecules and a variety of ribosomal proteins (r-protein or rProtein).
- Different rRNAs present in the ribosomes include small rRNAs and large rRNAs, which denote their presence in the small and large subunits of the ribosome.
- rRNAs combine with proteins in the cytoplasm to form ribosomes, which act as the site of protein synthesis and has the enzymes needed for the process.
- These complex structures travel along the mRNA molecule during translation and facilitate the assembly of amino acids to form a polypeptide chain. They bind to tRNAs and other molecules that are crucial for protein synthesis.

Function

rRNA directs the translation of mRNA into proteins.



Transfer RNA (tRNA)

- tRNA is the smallest of the 3 types of RNA having about 75-95 nucleotides.
- tRNAs are an essential component of translation, where their main function is the transfer of amino acids during protein synthesis. Therefore they are called transfer RNAs.
- Each of the 20 amino acids has a specific tRNA that binds with it and transfers it to the growing polypeptide chain. tRNAs also act as adapters in the translation of the genetic sequence of mRNA into proteins. Therefore they are also called adapter molecules.

Structure of tRNA

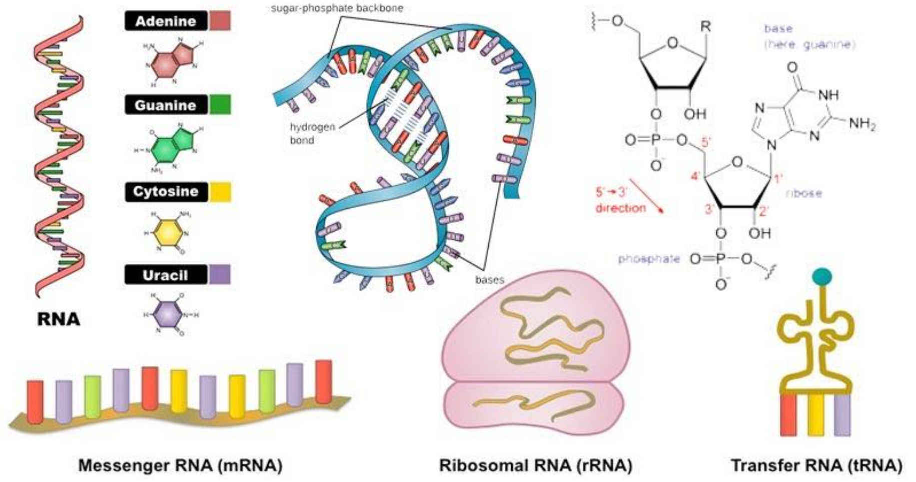
tRNAs have a clover leaf structure which is stabilized by strong hydrogen bonds between the nucleotides. Apart from the usual 4 bases, they normally contain some unusual bases mostly formed by methylation of the usual bases, for example, methyl guanine and methylcytosine.

- Three structural loops are formed via hydrogen bonding.
- The 3′ end serves as the amino acid attachment site.
- The center loop encompasses the anticodon.
- The anticodon is a three-base nucleotide sequence that binds to the mRNA codon.
- This interaction between codon and anticodon specifies the next amino acid to be added during protein synthesis.

Function

Transfer RNA brings or transfers amino acids to the ribosome that correspond to each three-nucleotide codon of rRNA. The amino acids then can be joined together and processed to make polypeptides and proteins.

Other Properties of RNA



- RNA forms in the nucleolus, and then moves to specialized regions of the cytoplasm depending on the type of RNA formed.
- RNA, containing a ribose sugar, is more reactive than DNA and is not stable in alkaline conditions. RNA's larger helical grooves mean it is more easily subject to attack by enzymes.
- RNA strands are continually made, broken down and reused.
- RNA is more resistant to damage from UV light than DNA.
- RNA's mutation rate is relatively higher.
- Unusual bases may be present.
- The number of RNA may differ from cell to cell.
- Rate of renaturation after melting is quick.
- RNA is more versatile than DNA, capable of performing numerous, diverse tasks in an organism.

FUNCTIONS OF RNA

- RNA is a nucleic acid messenger between DNA and ribosomes.
- It serves as the genetic material in some organisms (viruses).
- Some RNA molecules play an active role within cells by catalyzing biological reactions, controlling gene expression, or sensing and communicating responses to cellular signals.
- Messenger RNA (mRNA) copies DNA in the nucleus and carries the info to the ribosomes (in cytoplasm).
- Ribosomal RNA (rRNA) makes up a large part of the ribosome; reads and decodes mRNA.
- Transfer RNA (tRNA) carries amino acids to the ribosome where they are joined to form proteins.
- Certain RNAs are able to catalyse chemical reactions such as cutting and ligating other RNA molecules, and the catalysis of peptide bond formation in the ribosome; these are known as ribozymes.

Proteins

Proteins- Properties, Structure, Classification and Functions

- Proteins are the most abundant biological macromolecules, occurring in all cells.
- It is also the most versatile organic molecule of the living systems and occur in great variety; thousands of different kinds, ranging in size from relatively small peptides to large polymers.
- Proteins are the polymers of amino acids covalently linked by the peptide bonds.
- The building blocks of proteins are the twenty naturally occurring amino acids.
- Thus, proteins are the polymers of amino acids.

$$\text{Proteins} \xrightarrow{\text{hydrolysis}} \text{Peptides} \xrightarrow{\text{hydrolysis}} \text{Amino acids}$$

Properties of Proteins

Solubility in Water

- The relationship of proteins with water is complex.
- The secondary structure of proteins depends largely on the interaction of peptide bonds with water through hydrogen bonds.
- Hydrogen bonds are also formed between protein (alpha and beta structures) and water. The protein-rich static ball is more soluble than the helical structures.
- At the tertiary structure, water causes the orientation of the chains and hydrophilic radicals to the outside of the molecule, while the hydrophobic chains and radicals tend to react with each other within the molecule (hydrophobic effect).

Denaturation and Renaturation

- Proteins can be denatured by agents such as heat and urea that cause unfolding of polypeptide chains without causing hydrolysis of peptide bonds.
- The denaturing agents destroy secondary and tertiary structures, without affecting the primary structure.
- If a denatured protein returns to its native state after the denaturing agent is removed, the process is called renaturation.

Some of the denaturing agents include

Physical agents: Heat, radiation, pH

Chemical agents: Urea solution which forms new hydrogen bonds in the protein, organic solvents, detergents.

Isoelectric point

- The isoelectric point (pI) is the pH at which the number of positive charges equals the number of negative charges, and the overall charge on the amino acid is zero.
- At this point, when subjected to an electric field the proteins do not move either towards anode or cathode, hence this property is used to isolate proteins.

Molecular weight

- The average molecular weight of an amino acid is taken to be 110.
- The total number of amino acids in a protein multiplied by 110 gives the approximate molecular weight of that protein.
- Different proteins have different amino acid composition and hence their molecular weights differ.
- The molecular weights of proteins range from 5000 to 10^9 Daltons.

Posttranslational modifications

- It occurs after the protein has been synthesized on the ribosome.
- Phosphorylation, glycosylation, ADP ribosylation, methylation, hydroxylation, and acetylation affect the charge and the interactions between amino acid residues, altering the three-dimensional configuration and, thus, the function of the protein.

Protein structure

- The linear sequence of amino acid residues in a polypeptide chain determines the three-dimensional configuration of a protein, and the structure of a protein determines its function.
- All proteins contain the elements carbon, hydrogen, oxygen, nitrogen and sulfur some of these may also contain phosphorus, iodine, and traces of metals like ion, copper, zinc and manganese.
- A protein may contain 20 different kinds of amino acids. Each amino acid has an amine group at one end and an acid group at the other and a distinctive side chain.
- The backbone is the same for all amino acids while the side chain differs from one amino acid to the next.

The structure of proteins can be divided into four levels of organization:

1. Primary Structure

- The primary structure of a protein consists of the amino acid sequence along the polypeptide chain.
- Amino acids are joined by peptide bonds.
- Because there are no dissociable protons in peptide bonds, the charges on a polypeptide chain are due only to the N-terminal amino group, the C-terminal carboxyl group, and the side chains on amino acid residues.
- The primary structure determines the further levels of organization of protein molecules.

2. Secondary Structure

- The secondary structure includes various types of local conformations in which the atoms of the side chains are not involved.
- Secondary structures are formed by a regular repeating pattern of hydrogen bond formation between backbone atoms.
- The secondary structure involves α-helices, β-sheets, and other types of folding patterns that occur due to a regular repeating pattern of hydrogen bond formation.
- The secondary structure of protein could be :

1. **Alpha-helix**

2. **Beta-helix**

- The α-helix is a right-handed coiled strand.
- The side-chain substituents of the amino acid groups in an α-helix extend to the outside.
- Hydrogen bonds form between the oxygen of the C=O of each peptide bond in the strand and the hydrogen of the N-H group of the peptide bond four amino acids below it in the helix.
- The side-chain substituents of the amino acids fit in beside the N-H groups.
- The hydrogen bonding in a ß-sheet is between strands (inter-strand) rather than within strands (intra-strand).
- The sheet conformation consists of pairs of strands lying side-by-side.
- The carbonyl oxygens in one strand hydrogen bond with the amino hydrogens of the adjacent strand.
- The two strands can be either parallel or anti-parallel depending on whether the strand directions (N-terminus to C-terminus) are the same or opposite.
- The anti-parallel ß-sheet is more stable due to the more well-aligned hydrogen bonds.

3. Tertiary Structure

- Tertiary structure of a protein refers to its overall three-dimensional conformation.
- The types of interactions between amino acid residues that produce the three-dimensional shape of a protein include hydrophobic interactions, electrostatic interactions, and hydrogen bonds, all of which are non-covalent.
- Covalent disulfide bonds also occur.
- It is produced by interactions between amino acid residues that may be located at a considerable distance from each other in the primary sequence of the polypeptide chain.
- Hydrophobic amino acid residues tend to collect in the interior of globular proteins, where they exclude water, whereas hydrophilic residues are usually found on the surface, where they interact with water.

4. Quaternary Structure

- Quaternary structure refers to the interaction of one or more subunits to form a functional protein, using the same forces that stabilize the tertiary structure.
- It is the spatial arrangement of subunits in a protein that consists of more than one polypeptide chain.

Classification of proteins

Based on the chemical nature, structure, shape and solubility, proteins are classified as:

1. **Simple proteins**: They are composed of only amino acid residue. On hydrolysis these proteins yield only constituent amino acids. It is further divided into:
    - Fibrous protein: Keratin, Elastin, Collagen
    - Globular protein: Albumin, Globulin, Glutelin, Histones
2. **Conjugated proteins**: They are combined with non-protein moiety. Eg. Nucleoprotein, Phosphoprotein, Lipoprotein, Metalloprotein etc.
3. **Derived proteins**: They are derivatives or degraded products of simple and conjugated proteins. They may be :
    - Primary derived protein: Proteans, Metaproteins, Coagulated proteins
    - Secondary derived proteins: Proteosesn or albunoses, peptones, peptides.

Functions of proteins

Proteins are vital for the growth and repair, and their functions are endless. They also have enormous diversity of biological function and are the most important final products of the information pathways.

- Proteins, which are composed of amino acids, serve in many roles in the body (e.g., as enzymes, structural components, hormones, and antibodies).
- They act as structural components such as keratin of hair and nail, collagen of bone etc.
- Proteins are the molecular instruments through which genetic information is expressed.
- They execute their activities in the transport of oxygen and carbon dioxide by hemoglobin and special enzymes in the red cells.
- They function in the homostatic control of the volume of the circulating blood and that of the interstitial fluids through the plasma proteins.
- They are involved in blood clotting through thrombin, fibrinogen and other protein factors.
- They act as the defence against infections by means of protein antibodies.
- They perform hereditary transmission by nucleoproteins of the cell nucleus.
- Ovalbumine, glutelin etc. are storage proteins.
- Actin, myosin act as contractile protein important for muscle contraction

Aminoacids

Amino acids Amino acids are the building block of proteins. Amino acids are important organic compounds that contain amine (-NH2) and Carboxyl (-COOH) functional groups, along with a side-chain (R group) that is specific for each amino acid (Figure 1). Twenty different amino acids are commonly found in proteins. All of these 20 common amino acids are α-amino acids except proline and their general structure is shown below. They have a carboxyl group and amino group which are covalently bonded to a α-carbon atom. They differ from each other in their side chain R groups. Since, the remaining structure are same therefore properties of these amino acids are primarily determined by the side chain groups. The nature of these side chain maybe polar, nonpolar (aliphatic), hydrophilic, hydrophobic, acidic, basic and aromatic. These amino acids have been abbreviated using either three letter word or one letter word (Table 1).

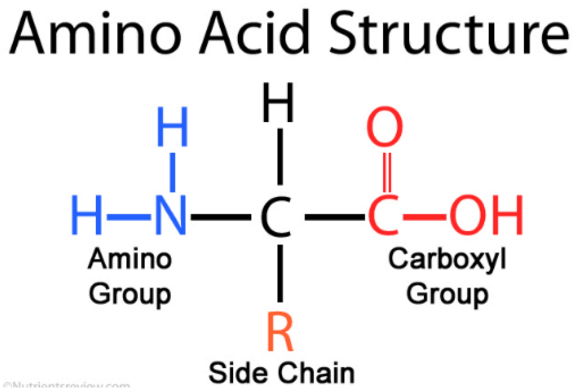


Figure 1: Structure of amino acid containing R side chain.

Classification

The 20 amino acids have been classified using different criteria by different scientists. For instance, they have been classified as polar, nonpolar, hydrophilic, hydrophobic, acidic, basic, aliphatic and aromatic. Here, we have classified all of these 20 common set of amino acids into six distinct classes. Nonpolar (Aliphatic) amino acids The R side chain in this class of amino acids including alanine, valine, leucine and isoleucine are hydrophobic in nature therefore they stabilize the protein structure through hydrophobic interactions. Glycine is also classified as nonpolar amino acids, but it has very small side chain. Therefore, it does not contribute to hydrophobic interactions. Glycine has the simplest structure. The side chain of proline has a distinctive cyclic structure which is an imino group held in a rigid conformation, therefore it reduces the structural flexibility of particularly that regions of polypeptide chain where it occurs.

Aromatic amino acids (Phenylalanine, Tyrosine and Tryptophan)

The side chain of aromatic amino acids contains an aromatic ring (Figure 3) which are relatively nonpolar (hydrophobic) in nature. These amino acids can participate in hydrophobic interaction. Tyrosine and tryptophan are much more polar than phenylalanine owing to their hydroxyl and nitrogen indole ring respectively. These amino acids show light absorption in the ultraviolet range due to the presence of conjugated double bond-single bond system.

Polar, uncharged amino acids

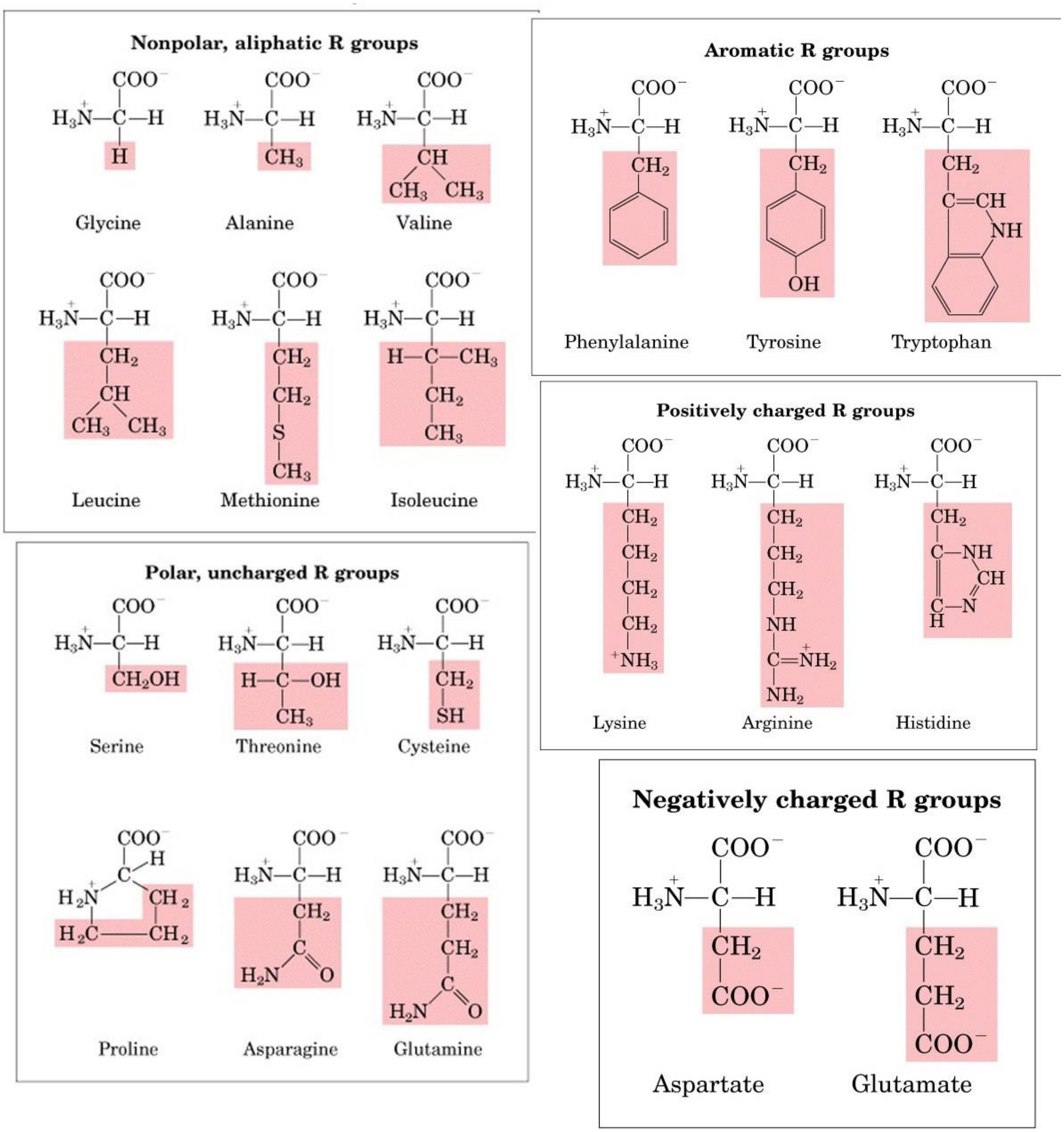
This class of amino acids includes serine, threonine, cysteine, asparagine and glutamine. The R group of these amino acids are more soluble in water or more hydrophilic than those of nonpolar amino acids because they contain functional groups (OH, SH, CONH2) that form hydrogen bonds with water. The polarity of serine and threonine is contributed by their hydroxyl groups, and that of cysteine and tryptophan by sulfhydryl and indole ring respectively which is weakly hydrogen bonded with oxygen and nitrogen respectively. Furthermore, polarity of asparagine and glutamine is contributed by their amide group.

Acidic amino acids

These amino acids contain two carboxyl groups, one α- carboxyl and other β- or γ-carboxyl group. Since they contain two acidic groups (one α-carboxyl group + one β or γ-carboxyl group) and one basic group (α-amino group), the net charge of these amino acids is therefore acidic and they are negatively charged at physiological pH.

Basic amino acids

The basic amino acid contains an α-amino group and the side chain contains second amino/ imino group (imidazole, ε-amino or guanidine group). These amino acids are histidine, lysine and arginine. Since these amino acids contain two basic groups one acidic group (α-carboxyl group), therefore the net behavior of these amino acids is basic and they are positively charged at physiological pH.



Acid-base character of amino acids

For explaining the acid-base character, let us consider neutral (aliphatic) amino acid alanine which is nonpolar that was discovered in 1923. Its carboxyl group can be deprotonated (Figure 7).

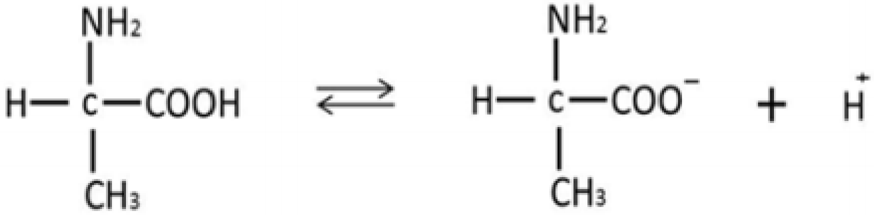


Figure 7: Carboxyl group of amino acid donates a proton.

Similarly, when amino acid accepts a proton (due to the presence of a basic amino group), the amino acid acquires a positive charge (Figure 8).

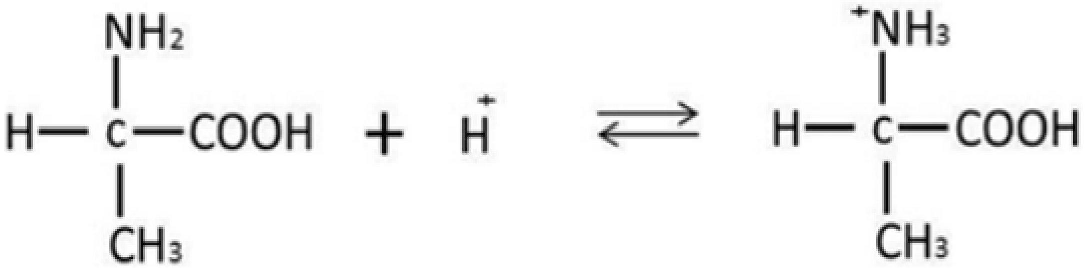


Figure 8: Amino group of amino acid accepts proton.

If we consider the acidity of amino acid, it releases proton which will be taken up by the solvent, water or by the basic amino group available on the amino acid. Since amino group is more basic, it takes the proton donated by carboxyl group. As a result carboxyl group acquires a negative charge whereas a positive charge develops on the amino group. Since this form of amino acid has both positive and negative charges, therefore the net charge of the amino acid is zero. This state of amino acid is known as zwitterionic state (Figure 9).

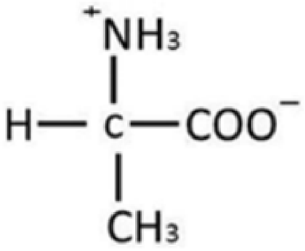


Figure 9: Zwitterionic state of alanine.

In zwitterions form, the carboxylate group acts as a base and the protonated amino group acts as an acid as shown below: -COO- + H+ → -COOH + -NH3 + OH→-NH2 + H2 O Since this type of amino acid is capable of acting as both acid and base, this implies that amino acid can act as buffer

Peptide bonds

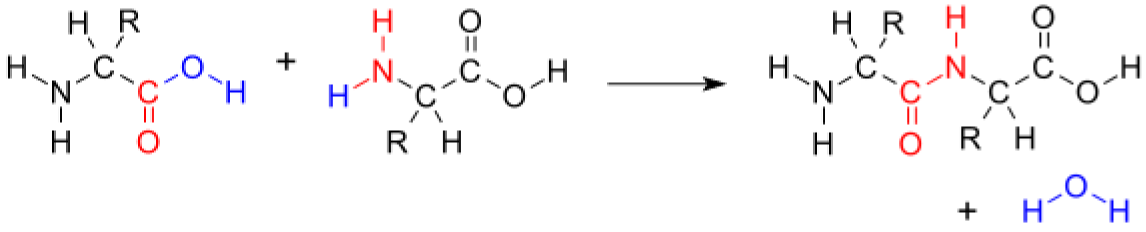
A **peptide bond** is an amide type of covalent chemical bond linking two consecutive alpha-amino acids from C1 (carbon number one) of one alpha-amino acid and N2 (nitrogen number two) of another, along a peptide or protein chain.

It can also be called an **eupeptide bond** to separate it from an isopeptide bond, a different type of amide bond between two amino acids.

Synthesis

When two amino acids form a *dipeptide* through a *peptide bond* it is a type of condensation reaction.[2] In this kind of condensation, two amino acids approach each other, with the non-side chain (C1) carboxylic acid moiety of one coming near the non-side chain (N2) amino moiety of the other. One loses a hydrogen and oxygen from its carboxyl group (COOH) and the other loses a hydrogen from its amino group (NH_2). This reaction produces a molecule of water (H_2O) and two amino acids joined by a peptide bond (-CO-NH-). The two joined amino acids are called a dipeptide.

The amide bond is synthesized when the carboxyl group of one amino acid molecule reacts with the amino group of the other amino acid molecule, causing the release of a molecule of water (H_2O), hence the process is a dehydration synthesis reaction.



The dehydration condensation of two amino acids to form a peptide bond (red) with expulsion of water (blue).

The formation of the peptide bond consumes energy, which, in organisms, is derived from ATP. Peptides and proteins are chains of amino acids held together by peptide bonds (and sometimes by a few isopeptide bonds). Organisms use enzymes to produce nonribosomal peptides, and ribosomes to produce proteins via reactions that differ in details from dehydration synthesis.

Some peptides, like alpha-amanitin, are called ribosomal peptides as they are made by ribosomes,[6] but many are nonribosomal peptides as they are synthesized by specialized enzymes rather than ribosomes. For example, the tripeptide glutathione is synthesized in two steps from free amino acids, by two enzymes: glutamate–cysteine ligase (forms an isopeptide bond, which is not a peptide bond) and glutathione synthetase (forms a peptide bond).

Degradation

A peptide bond can be broken by hydrolysis (the addition of water). In the presence of water they will break down and release 8–16 kilojoule/mol (2–4 kcal/mol) of Gibbs energy This process is extremely slow, with the half life at 25 °C of between 350 and 600 years per bond.

In living organisms, the process is normally catalyzed by enzymes known as peptidases or proteases, although there are reports of peptide bond hydrolysis caused by conformational strain as the peptide/protein folds into the native structure. This non-enzymatic process is thus not accelerated by transition state stabilization, but rather by ground state destabilization.

Web references

https://microbenotes.com/cell-organelles/

https://intl.siyavula.com/read/science/grade-10-lifesciences/cells-the-basic-units-of-life/02-cells-the-basic-units-of-life-04

https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3A_Biochemistry_Free_For_All_(Ahern_Rajagopal_and_Tan)/02%3A_Structure_and_Function/203%3A_Structure__Function-_Proteins_I

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOINFORMATICS

UNIT – 2 - INTRODUCTION TO BIOINFORMATICS– SBIA1101

Bioinformatics

Bioinformatics derives knowledge from computer analysis of biological data. These can consist of the information stored in the genetic code, but also experimental results from various sources, patient statistics, and scientific literature. Research in bioinformatics includes method development for storage, retrieval, and analysis of the data. Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. It has many practical applications in different areas of biology and medicine.

Fredj Tekaia at the Institut Pasteur offers this definition of bioinformatics: "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.

Bioinformatics definition - Organization / commitee

The NIH (National Instiute of Health) Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

The National Center for Biotechnology Information (NCBI 2001) defines bioinformatics as

Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline.

There are three important sub-disciplines within bioinformatics:

- the development of new algorithms and statistics with which to assess relationships among members of large data sets;
- the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures;
- and the development and implementation of tools that enable efficient access and management of different types of information."

Bioinformatics definition - other sources

- Bioinformatics or computational biology is the use of mathematical and informational techniques, including statistics, to solve biological problems, usually by creating or using computer programs, mathematical models or both. One of the main areas of bioinformatics is the data mining and analysis of the data gathered by the various genome projects. Other areas are sequence alignment, protein structure prediction, systems biology, protein-protein interactions and virtual evolution. (source: www.answers.com)

- Bioinformatics is the science of developing computer databases and algorithms for the purpose of speeding up and enhancing biological research. (source: www.whatis.com)
- "Biologists using computers, or the other way around. Bioinformatics is more of a tool than a discipline.(source: An Understandable Definition of Bioinformatics , The O'Reilly Bioinformatics Technology Conference, 2003) (4)
- The application of computer technology to the management of biological information. Specifically, it is the science of developing computer databases and algorithms to facilitate and expedite biological research.(source: Webopedia)
- Bioinformatics: a combination of Computer Science, Information Technology and Genetics to determine and analyze genetic information. (Definition from BitsJournal.com)
- Bioinformatics is the application of computer technology to the management and analysis of biological data. The result is that computers are being used to gather, store, analyse and merge biological data.(EBI - 2can resource)
- Bioinformatics is concerned with the creation and development of advanced information and computational technologies to solve problems in biology.
- Bioinformatics uses techniques from informatics, statistics, molecular biology and high-performance computing to obtain information about genomic or protein sequence data.

Even though the three terms: bioinformatics , computational biology and bioinformation infrastructure are often times used interchangeably, broadly, the three may be defined as follows:

- bioinformatics refers to database-like activities, involving persistent sets of data that are maintained in a consistent state over essentially indefinite periods of time;
- computational biology encompasses the use of algorithmic tools to facilitate biological analyses; while
- bioinformation infrastructure comprises the entire collective of information management systems, analysis tools and communication networks supporting biology.Thus, the latter may be viewed as a computational scaffold of the former two.

Bioinformatics is currently defined as the study of information content and information flow in biological systems and processes. It has evolved to serve as the bridge between observations (data) in diverse biologically related disciplines and the derivations of understanding (information) about how the systems or processes function, and subsequently the application (knowledge).

A Bioinformaticist versus a Bioinformatician (1999)

Bioinformatics has become a mainstay of genomics, proteomics, and all other *.omics (such as phenomics) that many information technology companies have entered the business or are considering entering the business, creating an IT (information technology) and BT (biotechnology) convergence.

A bioinformaticist is an expert who not only knows how to use bioinformatics tools, but also knows how to write interfaces for effective use of the tools.

A bioinformatician , on the other hand, is a trained individual who only knows to use bioinformatics tools without a deeper understanding.

The goal of bioinformatics thus is to provide scientists with a means to explain:

- Normal biological processes
- Malfunctions in these processes which lead to diseases
- Approaches to improving drug discovery.

To study how normal cellular activities are altered in different disease states, the biological data must be combined to form a comprehensive picture of these activities. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data. This includes nucleotide and amino acid sequences, protein domains, and protein structures.[18] The actual process of analyzing and interpreting data is referred to as computational biology. Important sub-disciplines within bioinformatics and computational biology include:

- Development and implementation of computer programs that enable efficient access to, management and use of, various types of information
- Development of new algorithms (mathematical formulas) and statistical measures that assess relationships among members of large data sets. For example, there are methods to locate a gene within a sequence, to predict protein structure and/or function, and to cluster protein sequences into families of related sequences.

The primary goal of bioinformatics is to increase the understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques to achieve this goal. Examples include: pattern recognition, data mining, machine learning algorithms, and visualization. Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein–protein interactions, genome-wide association studies, the modeling of evolution and cell division/mitosis.

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data.

Over the past few decades, rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. Bioinformatics is the name given to these mathematical and computing approaches used to glean understanding of biological processes.

Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning DNA and protein sequences to compare them, and creating and viewing 3-D models of protein structures.

History

Bioinformatics encompasses the use of tools and techniques from three separate disciplines; molecular biology (the source of the data to be analyzed), computer science (supplies the hardware for running analysis and the networks to communicate the results), and the data analysis algorithms which strictly define bioinformatics. For this reason, the editors have decided to incorporate events from these areas into a brief history of the field.

History of Bioinformatics
. 1860s Gregor Mendel
. 1950s double-helix structure of DNA sequencing.
. 1960s ARPANET.
. 1970s first recombinant DNA.
. Internet TCP.
. 1980s Personal computers.
. 1990s genome mapping.
. Http specification published.
. 2000 Human genome map completed.

Applications of Bioinformatics

Bioinformatics has not only become essential for basic genomic and molecular biology research, but is having a major impact on many areas of biotechnology and biomedical sciences. The main uses of bioinformatics include:

- Bioinformatics plays a vital role in the areas of structural genomics, functional genomics, and nutritional genomics.
- It covers emerging scientific research and the exploration of proteomes from the overall level of intracellular protein composition (protein profiles), protein structure, protein-protein interaction, and unique activity patterns (e.g. post-translational modifications).
- Bioinformatics is used for transcriptome analysis where mRNA expression levels can be determined.
- Bioinformatics is used to identify and structurally modify a natural product, to design a compound with the desired properties and to assess its therapeutic effects, theoretically.
- Cheminformatics analysis includes analyses such as similarity searching, clustering, QSAR modeling, virtual screening, etc.
- Bioinformatics is playing an increasingly important role in almost all aspects of drug discovery and drug development.
- Bioinformatics tools are very effective in prediction, analysis and interpretation of clinical and preclinical findings.

Applications of Bioinformatics in other fields

Its major applications include in the following fields:

Molecular medicine

- The human genome will have profound effects on the fields of biomedical research and clinical medicine.

- The completion of the human genome and the use of bioinformatic tools means that we can search for the genes directly associated with different diseases and begin to understand the molecular basis of these diseases more clearly.
- This new knowledge of the molecular mechanisms of disease will enable better treatments, cures and even preventative tests to be developed.

Personalised medicine

- Clinical medicine will become more personalised with the development of the field of pharmacogenomics.
- This is the study of how an individual's genetic inheritence affects the body's response to drugs.
- Today, doctors have to use trial and error to find the best drug to treat a particular patient as those with the same clinical symptoms can show a wide range of responses to the same treatment.
- In the future, doctors will be able to analyse a patient's genetic profile and prescribe the best available drug therapy and dosage from the beginning.

Preventative medicine

- With the specific details of the genetic mechanisms of diseases being unravelled, the development of diagnostic tests to measure a persons susceptibility to different diseases may become a distinct reality.

Gene therapy

- In the not too distant future with the use of bioinformatics tool, the potential for using genes themselves to treat disease may become a reality.
- Gene therapy is the approach used to treat, cure or even prevent disease by changing the expression of a person's genes.

Drug development

- At present all drugs on the market target only about 500 proteins.
- With an improved understanding of disease mechanisms and using computational tools to identify and validate new drug targets, more specific medicines that act on the cause, not merely the symptoms, of the disease can be developed.
- These highly specific drugs promise to have fewer side effects than many of today's medicines.

Microbial genome applications

- The arrival of the complete genome sequences and their potential to provide a greater insight into the microbial world and its capacities could have broad and far reaching implications for environment, health, energy and industrial applications.
- For these reasons, in 1994, the US Department of Energy (DOE) initiated the MGP (Microbial Genome Project) to sequence genomes of bacteria useful in energy production, environmental cleanup, industrial processing and toxic waste reduction.
- By studying the genetic material of these organisms, scientists can begin to understand these microbes at a very fundamental level and isolate the genes that give them their unique abilities to survive under extreme conditions.

Waste cleanup

- *Deinococcus radiodurans* is known as the world's toughest bacteria and it is the most radiation resistant organism known.
- Scientists are interested in this organism because of its potential usefulness in cleaning up waste sites that contain radiation and toxic chemicals.

Climate change Studies

- Increasing levels of carbon dioxide emission, mainly through the expanding use of fossil fuels for energy, are thought to contribute to global climate change.

- Recently, the DOE (Department of Energy, USA) launched a program to decrease atmospheric carbon dioxide levels.
- One method of doing so is to study the genomes of microbes that use carbon dioxide as their sole carbon source.

Alternative energy sources

- Scientists are studying the genome of the microbe *Chlorobium tepidum* which has an unusual capacity for generating energy from light

Biotechnology

- The archaeon *Archaeoglobus fulgidus* and the bacterium *Thermotoga maritima* have potential for practical applications in industry and government-funded environmental remediation.
- These microorganisms thrive in water temperatures above the boiling point and therefore may provide the DOE, the Department of Defence, and private companies with heat-stable enzymes suitable for use in industrial processes
- Other industrially useful microbes include, *Corynebacterium glutamicum* which is of high industrial interest as a research object because it is used by the chemical industry for the biotechnological production of the amino acid lysine.
- The substance is employed as a source of protein in animal nutrition.
- Biotechnologically produced lysine is added to feed concentrates as a source of protein, and is an alternative to soybeans or meat and bonemeal.
- *Lactococcus lactis* is one of the most important micro-organisms involved in the dairy industry.
- Researchers anticipate that understanding the physiology and genetic make-up of this bacterium will prove invaluable for food manufacturers as well as the pharmaceutical industry, which is exploring the capacity of *lactis* to serve as a vehicle for delivering drugs.

Antibiotic resistance

- Scientists have been examining the genome of *Enterococcus faecalis*-a leading cause of bacterial infection among hospital patients.
- They have discovered a virulence region made up of a number of antibiotic-resistant genes that may contribute to the bacterium's transformation from a harmless gut bacteria to a menacing invader.
- The discovery of the region, known as a pathogenicity island, could provide useful markers for detecting pathogenic strains and help to establish controls to prevent the spread of infection in wards.

Forensic analysis of microbes

- Scientists used their genomic tools to help distinguish between the strain of *Bacillus anthracis* that was used in the summer of 2001 terrorist attack in Florida with that of closely related anthrax strains.

The reality of bioweapon creation

- Scientists have recently built the virus poliomyelitis using entirely artificial means.
- They did this using genomic data available on the Internet and materials from a mail-order chemical supply.
- The research was financed by the US Department of Defence as part of a biowarfare response program to prove to the world the reality of bioweapons.
- The researchers also hope their work will discourage officials from ever relaxing programs of immunisation.
- This project has been met with very mixed feelings.

Evolutionary studies

- The sequencing of genomes from all three domains of life, eukaryota, bacteria and archaea means that evolutionary studies can be performed in a quest to determine the tree of life and the last universal common ancestor.

Crop improvement

- Comparative genetics of the plant genomes has shown that the organisation of their genes has remained more conserved over evolutionary time than was previously believed.
- These findings suggest that information obtained from the model crop systems can be used to suggest improvements to other food crops.
- At present the complete genomes of Arabidopsis thaliana (water cress) and Oryza sativa (rice) are available.

Insect resistance

- Genes from *Bacillus thuringiensis* that can control a number of serious pests have been successfully transferred to cotton, maize and potatoes.
- This new ability of the plants to resist insect attack means that the amount of insecticides being used can be reduced and hence the nutritional quality of the crops is increased.

Improve nutritional quality

- Scientists have recently succeeded in transferring genes into rice to increase levels of Vitamin A, iron and other micronutrients.
- This work could have a profound impact in reducing occurrences of blindness and anaemia caused by deficiencies in Vitamin A and iron respectively.
- Scientists have inserted a gene from yeast into the tomato, and the result is a plant whose fruit stays longer on the vine and has an extended shelf life.

Development of Drought resistance varieties

- Progress has been made in developing cereal varieties that have a greater tolerance for soil alkalinity, free aluminium and iron toxicities.
- These varieties will allow agriculture to succeed in poorer soil areas, thus adding more land to the global production base.
- Research is also in progress to produce crop varieties capable of tolerating reduced water conditions.

Veterinary Science

- Sequencing projects of many farm animals including cows, pigs and sheep are now well under way in the hope that a better understanding of the biology of these organisms will have huge impacts for improving the production and health of livestock and ultimately have benefits for human nutrition.

Comparative Studies

- Analysing and comparing the genetic material of different species is an important method for studying the functions of genes, the mechanisms of inherited diseases and species evolution.
- Bioinformatics tools can be used to make comparisons between the numbers, locations and biochemical functions of genes in different organisms.

Definitions of Fields Related to Bioinformatics

Bioinformatics has various applications in research in medicine, biotechnology, agriculture etc.
Following research fields has integral component of Bioinformatics

1. **Computational Biology:** The development and application of data-analytical andtheoretical methods, mathematical modeling and computational simulation techniquesto the study of biological, behavioral, and social systems.

2. **Genomics:** Genomics is any attempt to analyze or compare the entire genetic complement of a species or species (plural). It is, of course possible to compare genomes by comparing more-or-less representative subsets of genes within genomes.
3. **Proteomics:** Proteomics is the study of proteins - their location, structure and function. It is the identification, characterization and quantification of all proteins involved in a particular pathway, organelle, cell, tissue, organ or organism that can be studied in concert to provide accurate and comprehensive data about that system. Proteomics is the study of the function of all expressed proteins. The study of the proteome, called proteomics, now evokes not only all the proteins in any given cell, but also the set of all protein isoforms and modifications, the interactions between them, the structural description of proteins and their higher-order complexes, and for that matter almost everything 'post-genomic'." (5)
4. **Pharmacogenomics :** Pharmacogenomics is the application of genomic approaches and technologies to the identification of drug targets. In Short, pharmacogenomics is using genetic information to predict whether a drug will help make a patient well or sick. It Studies how genes influence the response of humans to drugs, from the population to the molecular level.
5. **Pharmacogenetics:** Pharmacogenetics is the study of how the actions of and reactions to drugs vary with the patient's genes. All individuals respond differently to drug treatments; some positively, others with little obvious change in their conditions and yet others with side effects or allergic reactions. Much of this variation is known to have a genetic basis. Pharmacogenetics is a subset of pharmacogenomics which uses genomic/bioinformatic methods to identify genomic correlates, for example SNPs (Single Nucleotide Polymorphisms), characteristic of particular patient response profiles and use those markers to inform the administration and development of therapies. Strikingly such approaches have been used to "resurrect" drugs thought previously to be ineffective, but subsequently found to work with in subset of patients or in optimizing the doses of chemotherapy for particular patients.
6. **Cheminformatics:** 'The mixing of those information resources [information technology and information management] to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the arena of drug lead identification and optimization.' (Frank K Brown 'Chemoinformatics: what is it and how does it impact drug discovery.' Ann. Rep. Med. Chem. 1998, 33 , 375-384.) (6)
   Related terms of cheminformatics are chemi-informatics, chemometrics, computational chemistry, chemical informatics, chemical information management/science, and cheminformatics.

   But we can distinguish chemoinformatics and chemical informatics as follows
   Chemical informatics: 'Computer-assisted storage, retrieval and analysis of chemical information, from data to chemical knowledge.' (Chem. Inf. Lett. 2003, 6 , 14.) This definition is distinct from ' Chemoinformatics ' (and the synonymous cheminformatics and chemiinformatics) which focus on drug design.
   chemometrics: The application of statistics to the analysis of chemical data (from organic, analytical or medicinal chemistry) and design of chemical experiments and simulations. [IUPAC Computational]

computational chemistry: A discipline using mathematical methods for the calculation of molecular properties or for the simulation of molecular behavior. It also includes, e.g., synthesis planning, database searching, combinatorial library manipulation (Hopfinger, 1981; Ugi et al., 1990). [IUPAC Computational]

7. **Structural genomics or structural bioinformatics** refers to the analysis of macromolecular structure particularly proteins, using computational tools and theoretical frameworks. One of the goals of structural genomics is the extension of idea of genomics , to obtain accurate three-dimensional structural models for all known protein families, protein domains or protein folds . Structural alignment is a tool of structural genomics.
8. **Comparative genomics:** The study of human genetics by comparisons with model organisms such as mice, the fruit fly, and the bacterium E. coli.
9. **Biophysics:**The British Biophysical Society defines biophysics as: "an interdisciplinary field which applies techniques from the physical sciences to understanding biological structure and function".
10. **Biomedical informatics / Medical informatics:** "Biomedical Informatics is an emerging discipline that has been defined as the study, invention, and implementation of structures and algorithms to improve communication, understanding and management of medical information."
11. **Mathematical Biology:** Mathematical biology also tackles biological problems, but the methods it uses to tackle them need not be numerical and need not be implemented in software or hardware. It includes things of theoretical interest which are not necessarily algorithmic, not necessarily molecular in nature, and are not necessarily useful in analyzing collected data.
12. **Computational chemistry:** Computational chemistry is the branch of theoretical chemistry whose major goals are to create efficient computer programs that calculate the properties of molecules (such as total energy, dipole moment, vibrational frequencies) and to apply these programs to concrete chemical objects. It is also sometimes used to cover the areas of overlap between computer science and chemistry.
13. **Functional genomics:** Functional genomics is a field of molecular biology that is attempting to make use of the vast wealth of data produced by genome sequencing projects to describe genome function. Functional genomics uses high-throuput techniques like DNA microarrays, proteomics, metabolomics and mutation analysis to describe the function and interactions of genes.
14. **Pharmacoinformatics:** Pharmacoinformatics concentrates on the aspects of bioinformatics dealing with drug discovery
15. **In silico ADME-Tox Prediction:**(Brief description)- Drug discovery is a complex and risky treasure hunt to find the most efficacious molecule which do not have toxic effects but at the same time have desired pharmacokinetic profile. The hunt starts when the researchers look for the binding affinity of the molecule to its target. Huge amount of research requires to be done to come out with a molecule which has the reliable binding profile. Once the molecules have been identified, as per the traditional methodologies, the molecule is further subjected to optimization with the aim of improving efficacy. The molecules which show better binding is

then evaluated for its toxicity and pharmacokinetic profiles. It is at this stage that most of the candidates fail in the race to become a successful drug.

16. **Agroinformatics / Agricultural informatics:** Agroinformatics concentrates on the aspects of bioinformatics dealing with plant genomes.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOINFORMATICS

UNIT – 3 - INTRODUCTION TO BIOINFORMATICS– SBIA1101

Biological databases

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

Biological databases can be broadly classified into sequence, structure and functional databases. Nucleic acid and protein sequences are stored in sequence databases and structure databases store solved structures of RNA and proteins. Functional databases provide information on the physiological role of gene products, for example enzyme activities, mutant phenotypes, or biological pathways. Model Organism Databases are functional databases that provide species-specific data. Databases are important tools in assisting scientists to analyze and explain a host of biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. This knowledge helps facilitate the fight against diseases, assists in the development of medications, predicting certain genetic diseases and in discovering basic relationships among species in the history of life.

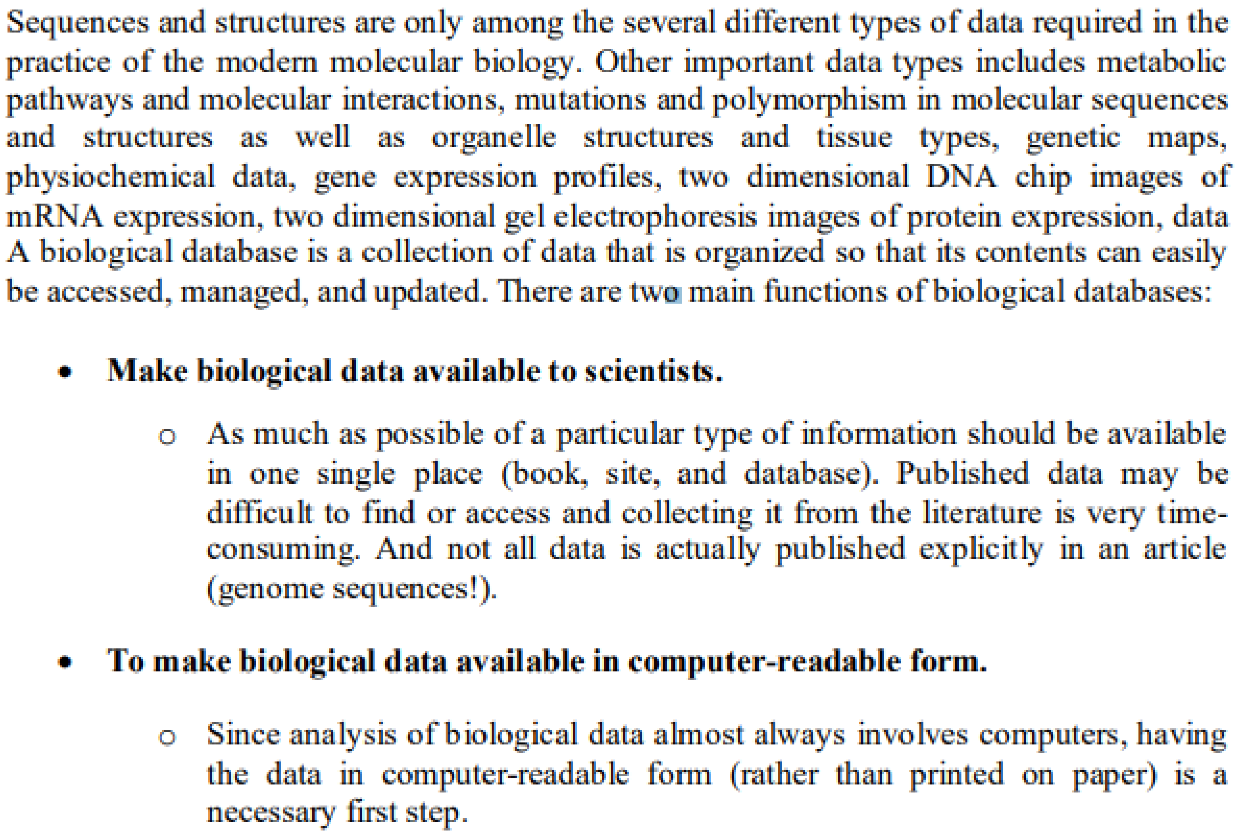
Biological knowledge is distributed among many different general and specialized databases. This sometimes makes it difficult to ensure the consistency of information. Integrative bioinformatics is one field attempting to tackle this problem by providing unified access. One solution is how biological databases cross-reference to other databases with accession numbers to link their related knowledge together.

Relational database concepts of computer science and Information retrieval concepts of digital libraries are important for understanding biological databases. Biological database design, development, and long-term management is a core area of the discipline of bioinformatics. Data contents include gene sequences, textual descriptions, attributes and ontology classifications, citations, and tabular data. These are often described as semi-structured data, and can be represented as tables, key delimited records, and XML structures

- One of the hallmarks of modern genomic research is the generation of enormous amounts of raw sequence data.
- As the volume of genomic data grows, sophisticated computational methodologies are required to manage the data deluge.
- Thus, the very first challenge in the genomics era is to store and handle the staggering volume of information through the establishment and use of computer databases.

- A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system.
- A simple database might be a single file containing many records, each of which includes the same set of information.
- The chief objective of the development of a database is to organize data in a set of structured records to enable easy retrieval of information.

Example. A few popular databases are GenBank from NCBI (National Center for Biotechnology Information), SwissProt from the Swiss Institute of **Bioinformatics** and PIR from the Protein Information Resource.



Sequences and structures are only among the several different types of data required in the practice of the modern molecular biology. Other important data types includes metabolic pathways and molecular interactions, mutations and polymorphism in molecular sequences and structures as well as organelle structures and tissue types, genetic maps, physiochemical data, gene expression profiles, two dimensional DNA chip images of mRNA expression, two dimensional gel electrophoresis images of protein expression, data A biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated. There are two main functions of biological databases:

- **Make biological data available to scientists.**
  - As much as possible of a particular type of information should be available in one single place (book, site, and database). Published data may be difficult to find or access and collecting it from the literature is very time-consuming. And not all data is actually published explicitly in an article (genome sequences!).
- **To make biological data available in computer-readable form.**
  - Since analysis of biological data almost always involves computers, having the data in computer-readable form (rather than printed on paper) is a necessary first step.

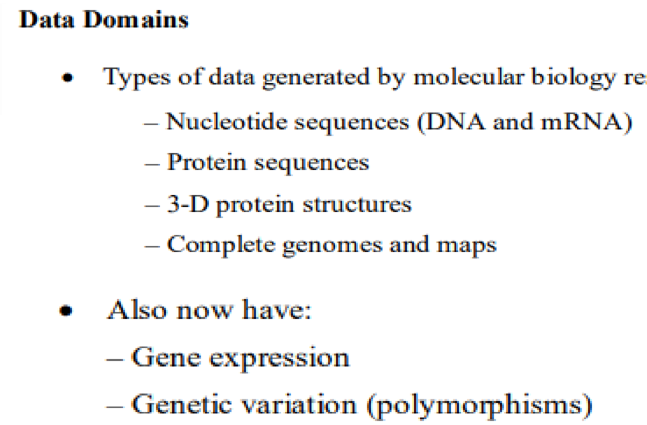
Importance of Databases

- Databases act as a store house of information.
- Databases are used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria.
- It allows knowledge discovery, which refers to the identification of connections between pieces of information that were not known when the information was first entered. This facilitates the discovery of new biological insights from raw data.

- Secondary databases have become the molecular biologist’s reference library over the past decade or so, providing a wealth of information on just about any gene or gene product that has been investigated by the research community.
- It helps to solve cases where many users want to access the same entries of data.
- Allows the indexing of data.
- It helps to remove redundancy of data.

Vocabulary:

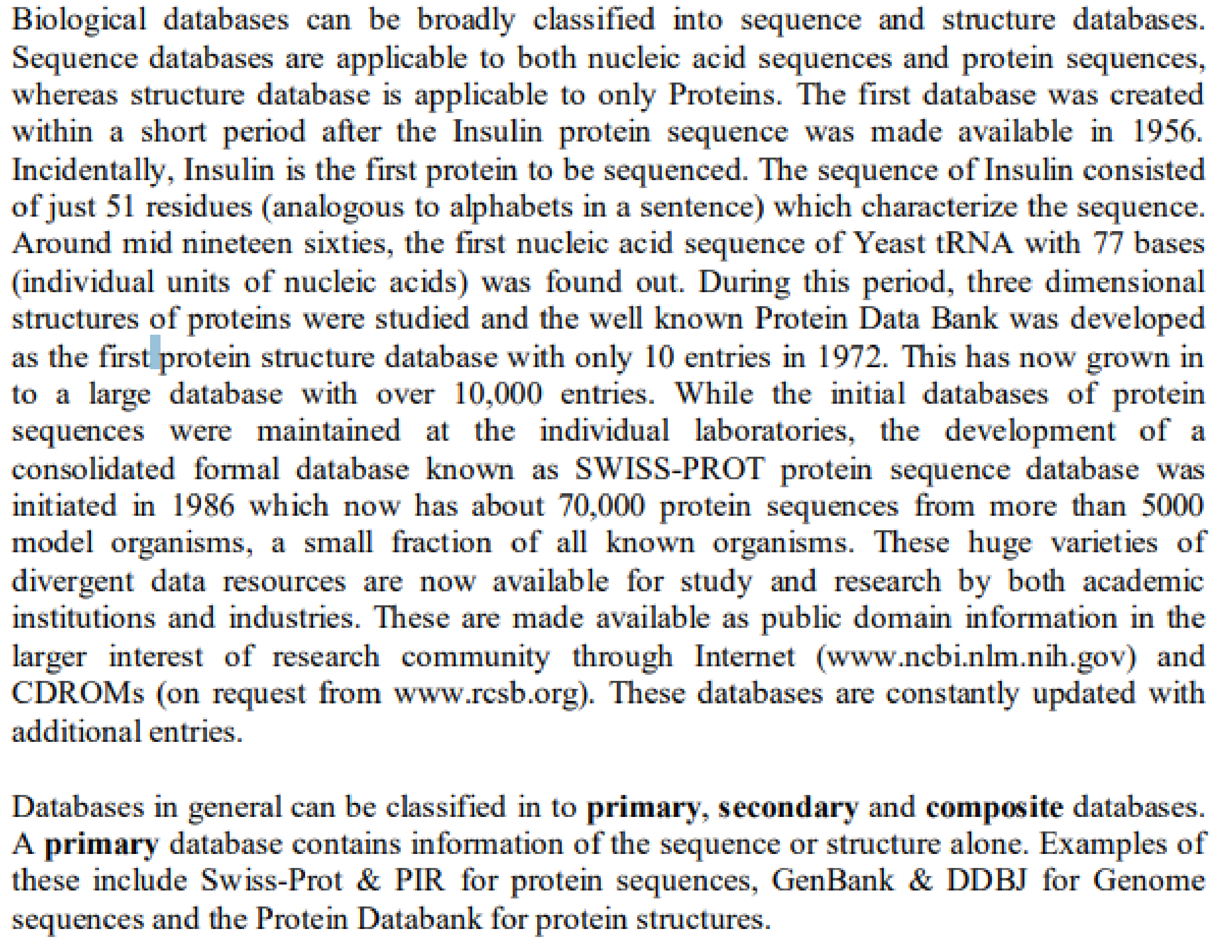
- *Entities*: The kind of things that we want to store in a database. E.g.: Genes, DNA sequences, bibliographical references.
- *Records*: The particular things stored in the database. E.g.: The gene BRCA1
- *Identifiers* or *key*: The unique name that identifies a record
- *Fields*: The properties that an entity has. E.g.: The name, sequence and mutations of the gene



Data Domains

- Types of data generated by molecular biology research:
  - – Nucleotide sequences (DNA and mRNA)
  - – Protein sequences
  - – 3-D protein structures
  - – Complete genomes and maps
- Also now have:
  - – Gene expression
  - – Genetic variation (polymorphisms)

Kinds of Biological Databases



Biological databases can be broadly classified into sequence and structure databases. Sequence databases are applicable to both nucleic acid sequences and protein sequences, whereas structure database is applicable to only Proteins. The first database was created within a short period after the Insulin protein sequence was made available in 1956. Incidentally, Insulin is the first protein to be sequenced. The sequence of Insulin consisted of just 51 residues (analogous to alphabets in a sentence) which characterize the sequence. Around mid nineteen sixties, the first nucleic acid sequence of Yeast tRNA with 77 bases (individual units of nucleic acids) was found out. During this period, three dimensional structures of proteins were studied and the well known Protein Data Bank was developed as the first protein structure database with only 10 entries in 1972. This has now grown in to a large database with over 10,000 entries. While the initial databases of protein sequences were maintained at the individual laboratories, the development of a consolidated formal database known as SWISS-PROT protein sequence database was initiated in 1986 which now has about 70,000 protein sequences from more than 5000 model organisms, a small fraction of all known organisms. These huge varieties of divergent data resources are now available for study and research by both academic institutions and industries. These are made available as public domain information in the larger interest of research community through Internet (www.ncbi.nlm.nih.gov) and CDROMs (on request from www.rcsb.org). These databases are constantly updated with additional entries.

Databases in general can be classified in to **primary**, **secondary** and **composite** databases. A **primary** database contains information of the sequence or structure alone. Examples of these include Swiss-Prot & PIR for protein sequences, GenBank & DDBJ for Genome sequences and the Protein Databank for protein structures.

Biological databases can be further classified as primary, secondary, and composite databases.

Primary databases

- Primary databases are also called as archieval database.
- They are populated with experimentally derived data such as nucleotide sequence, protein sequence or macromolecular structure.
- Experimental results are submitted directly into the database by researchers, and the data are essentially archival in nature.
- Once given a database accession number, the data in primary databases are never changed: they form part of the scientific record.

Primary databases contain information for sequence or structure only. Examples of primary biological databases include:

- Swiss-Prot and PIR for protein sequences
- GenBank and DDBJ for genome sequences
- Protein Databank for protein structures
- ENA, GenBank and DDBJ (nucleotide sequence)
- Array Express Archive and GEO (functional genomics data)
- Protein Data Bank (PDB; coordinates of three-dimensional macromolecular structures)

Secondary Databases

- Secondary databases comprise data derived from the results of analysing primary data.
- Secondary databases often draw upon information from numerous sources, including other databases (primary and secondary), controlled vocabularies and the scientific literature.
- They are highly curated, often using a complex combination of computational algorithms and manual analysis and interpretation to derive new knowledge from the public record of science.

Secondary databases contain information derived from primary databases. Secondary databases store information such as conserved sequences, active site residues, and signature sequences. Protein Databank data is stored in secondary databases. Examples include:

- SCOP at Cambridge University
- CATH at the University College of London
- PROSITE of the Swiss Institute of Bioinformatics
- eMOTIF at Stanford
- InterPro (protein families, motifs and domains)
- UniProt Knowledgebase (sequence and functional information on proteins)
- Ensembl (variation, function, regulation and more layered onto whole genome sequences)

Composite databases contain a variety of primary databases, which eliminates the need to search each one separately. Each composite database has different search algorithms and data structures. The NCBI hosts these databases, where links to the Online Mendelian Inheritance in Man (OMIM) is found.

However, many data resources have both primary and secondary characteristics. For example, UniProt accepts primary sequences derived from peptide sequencing experiments. However, UniProt also infers peptide sequences from genomic information, and it provides a wealth of additional information, some derived from automated annotation (TrEMBL), and even more from careful manual analysis (SwissProt).

There are also specialized databases are those that cater to a particular research interest. For example, Flybase, HIV sequence database, and Ribosomal Database Project are databases that specialize in a particular organism or a particular type of data.

Main sequence databases:
- NCBI
- EMBL

Main protein databases:
- Uniprot
- PDB
- MMDB

Some genome databases:
- ENSEMBL (Human, mouse and others)
- SGD (Yeast)
- TAIR (Arabidopsis)

Bibliography:
- Pubmed
- Web of Science

Human diseases:
- OMIM

Metabolic pathways:
- KEGG

Nucleotide sequence databases

- As biology has increasingly turned into a data-rich science, the need for storing and communicating large datasets has grown tremendously.
- The obvious examples are the nucleotide sequences, the protein sequences, and the 3D structural data produced by X-ray crystallography and macromolecular NMR.
- The biological information of nucleic acids is available as sequences while the data of proteins are available as sequences and structures. Sequences are represented in a single dimension whereas the structure contains the three-dimensional data of sequences.
- A biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated.
- The database is complemented with generalized software for processing, archiving, querying and distributing data.
- Such databases consisting of nucleotide sequences are called nucleic acid sequence databases.

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

1. Primary databases of nucleotide sequences

- There are three chief databases that store and make available raw nucleic acid sequences to the public and researchers alike: **GenBank, EMBL, DDBJ**.
- They are referred to as the primary nucleotide sequence databases since they are the repository of all nucleic acid sequences.
- GenBank is physically located in the USA and is accessible through the NCBI portal over the intern.
- EMBL (European Molecular Biology Laboratory) is in UK and DDJB (DNA databank of Japan) is in Japan.
- All three accept nucleotide sequence submissions and then exchange new and updated data on a daily basis to achieve optimal synchronization between them.
- These three databases are primary databases, as they house original sequence data.
- They collaborate with Sequence Read Archive (SRA), which archives raw reads from high-throughput sequencing instruments.

a. GenBank

The GenBank sequence database is open access, annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced and maintained by the National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration (INSDC). receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. GenBank has become an important database for research in biological fields and has grown in recent years at an exponential rate by doubling roughly every 18 months.

b. EMBL (European Molecular Biology Laboratory)

The European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database is a comprehensive collection of primary nucleotide sequences maintained at the European Bioinformatics Institute (EBI). Data are received from genome sequencing centers, individual scientists and patent offices.

c. DDBJ (DNA databank of Japan)

It is located at the National Institute of Genetics (NIG) in the Shizuoka prefecture of Japan. It is the only nucleotide sequence data bank in Asia. Although DDBJ mainly receives its data from Japanese researchers, it can accept data from contributors from any other country.

2.Secondary databases of nucleotide sequences

- Many of the secondary databases are simply sub-collection of sequences culled from one or the other of the primary databases such as GenBank or EMBL.
- There is also usually a great deal of value addition in terms of annotation, software, presentation of the information and the cross-references.

- There are other secondary databases that do not present sequences at all, but only information gathered from sequences databases.

a. Omniome Database:

Omniome Database is a comprehensive microbial resource maintained by TIGR (The Institute for Genomic Research). It has not only the sequence and annotation of each of the completed genomes, but also has associated information about the organisms (such as taxon and gram stain pattern), the structure and composition of their DNA molecules, and many other attributes of the protein sequences predicted from the DNA sequences.

It facilitates the meaningful multi-genome searches and analysis, for instance, alignment of entire genomes, and comparison of the physical proper of proteins and genes from different genomes etc.

b. FlyBase Database:

A consortium sequenced the entire genome of the fruit fly D. Melanogaster to a high degree of completeness and quality.

c. ACeDB:

It is a repository of not only the sequence but also the genetic map as well as phenotypic information about the C. Elegans nematode worm.

Protein Databases- Types and Importance

- As biology has increasingly turned into a data-rich science, the need for storing and communicating large datasets has grown tremendously.
- The obvious examples are the nucleotide sequences, the protein sequences, and the 3D structural data produced by X-ray crystallography and macromolecular NMR.
- The biological information of proteins is available as sequences and structures. Sequences are represented in a single dimension whereas the structure contains the three-dimensional data of sequences.
- A biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated.
- A protein database is one or more datasets about proteins, which could include a protein's amino acid sequence, conformation, structure, and features such as active sites.
- Protein databases are compiled by the translation of DNA sequences from different gene databases and include structural information. They are an important resource because proteins mediate most biological functions.

Importance

Huge amounts of data for protein structures, functions, and particularly sequences are being generated. Searching databases are often the first step in the study of a new protein. It has the following uses:

1. Comparison between proteins or between protein families provides information about the relationship between proteins within a genome or across different species and hence offers much more information that can be obtained by studying only an isolated protein.
2. Secondary databases derived from experimental databases are also widely available. These databases reorganize and annotate the data or provide predictions.
3. The use of multiple databases often helps researchers understand the structure and function of a protein.

Primary databases of Proteins

The PRIMARY databases hold the experimentally determined protein sequences inferred from the conceptual translation of the nucleotide sequences. This, of course, is not experimentally derived information, but has arisen as a result of interpretation of the nucleotide sequence information and consequently must be treated as potentially containing misinterpreted information. There is a number of primary protein sequence databases and each requires some specific consideration.

a. Protein Information Resource (PIR) – Protein Sequence Database (PIR-PSD):

- The PIR-PSD is a collaborative endeavor between the PIR, the MIPS (Munich Information Centre for Protein Sequences, Germany) and the JIPID (Japan International Protein Information Database, Japan).
- The PIR-PSD is now a comprehensive, non-redundant, expertly annotated, object-relational DBMS.
- A unique characteristic of the PIR-PSD is its classification of protein sequences based on the superfamily concept.
- The sequence in PIR-PSD is also classified based on homology domain and sequence motifs.
- Homology domains may correspond to evolutionary building blocks, while sequence motifs represent functional sites or conserved regions.
- The classification approach allows a more complete understanding of sequence function-structure relationship.

b. SWISS-PROT

- The other well known and extensively used protein database is SWISS-PROT. Like the PIR-PSD, this curated proteins sequence database also provides a high level of annotation.
- The data in each entry can be considered separately as core data and annotation.
- The core data consists of the sequences entered in common single letter amino acid code, and the related references and bibliography. The taxonomy of the organism from which the sequence was obtained also forms part of this core information.
- The annotation contains information on the function or functions of the protein, post-translational modification such as phosphorylation, acetylation, etc., functional and structural domains and sites, such as calcium binding regions, ATP-binding sites, zinc

fingers, etc., known secondary structural features as for examples alpha helix, beta sheet, etc., the quaternary structure of the protein, similarities to other protein if any, and diseases that may arise due to different authors publishing different sequences for the same protein, or due to mutations in different strains of an described as part of the annotation.

TrEMBL (for Translated EMBL) is a computer-annotated protein sequence database that is released as a supplement to SWISS-PROT. It contains the translation of all coding sequences present in the EMBL Nucleotide database, which have not been fully annotated. Thus it may contain the sequence of proteins that are never expressed and never actually identified in the organisms.

c. Protein Databank (PDB):

- PDB is a primary protein structure database. It is a crystallographic database for the three-dimensional structure of large biological molecules, such as proteins.
- In spite of the name, PDB archive the three-dimensional structures of not only proteins but also all biologically important molecules, such as nucleic acid fragments, RNA molecules, large peptides such as antibiotic gramicidin and complexes of protein and nucleic acids.
- The database holds data derived from mainly three sources: Structure determined by X-ray crystallography, NMR experiments, and molecular modeling.

Secondary Databases

- A **biological database** is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system.
- The chief objective of the development of a database is to organize data in a set of structured records to enable easy retrieval of information.
- Based on their contents, biological databases can be either primary database or secondary databases.
- Among the two, secondary databases have become a biologist's reference library over the past decade or so, providing a wealth of information on just any research or research product that has been investigated by the research community.
- Sequence annotation information in the primary database is often minimal.
- To turn the raw sequence information into more sophisticated biological knowledge, much post-processing of the sequence information is needed.
- This begs the need for secondary databases, which contain computationally processed sequence information derived from the primary databases.
- Thus, secondary databases comprise data derived from the results of analyzing primary data.
- Secondary databases often draw upon information from numerous sources, including other databases (primary and secondary), controlled vocabularies and the scientific literature.

- They are highly curated, often using a complex combination of computational algorithms and manual analysis and interpretation to derive new knowledge from the public record of science.
- The amount of computational processing work, however, varies greatly among the secondary databases; some are simple archives of translated sequence data from identified open reading frames in DNA, whereas others provide additional annotation and information related to higher levels of information regarding structure and functions.

Importance of secondary databases

- Secondary databases contain information derived from primary sequence data which are in the form of regular expressions (patterns), Fingerprints, profiles blocks or Hidden Markov Models.
- The type of information stored in each of the secondary databases is different. But in secondary databases, homologous sequences may be gathered together in multiple alignments.
- In multiple alignments, there are conserved regions that show little or no variation between the constituent sequences. These conserved regions are called motifs.
- Motifs reflect some vital biological role and are crucial to the structure of the function of the protein. This is the importance of the secondary database.
- So by concentrating on motifs, we can find out the common conserved regions in the sequences and study the functional and evolutionary details or organisms.

Some of the common secondary databases include:

Prosite

- It was the first secondary database developed.
- Protein families usually contain some most conserved motifs which can be encoded to find out various biological functions.
- So by using such a database tool, we can easily find out the family of proteins when a new sequence is searched. This is the importance of PROSITE.
- Within PROSITE motifs are encoded as a regular expression (called patterns).
- Entries are deposited in PROSITE in two distant files. The first file gives the pattern and lists all matches of pattern, whereas the second one gives the details of family, description of the biological role, etc.
- The process used to derive patterns involves the construction of multiple alignment and manual inspection.
- So PROSITE contains documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them.

- A set of databases collects together patterns found in protein sequences rather than the complete sequences. PROSITE is one such pattern database.
- The protein motif and pattern are encoded as “regular expressions”.
- The information corresponding to each entry in PROSITE is of the two forms – the patterns and the related descriptive text.

Prints

- Most protein families are characterized by several conserved motifs.
- All of these motifs can be an aid in constructing the `signatures’ of different families. This principle is highlighted in constructing PRINT database.
- Within PRINTS motifs are encoded as unweighted local alignments. So small initial multiple alignments are taken to identify conserved motifs.
- Then these regions are searched in the database to find out similarities.
- Results are analyzed to find out the sequences which matched all the motifs within the fingerprint.
- PROSITE and PRINTS are the only manually annotated secondary databases. The print is a diagnostic collection of protein fingerprints.
- In the PRINTS database, the protein sequence patterns are stored as ‘fingerprints’. A fingerprint is a set of motifs or patterns rather than a single one.
- The information contained in the PRINT entry may be divided into three sections. In addition to entry name, accession number and number of motifs, the first section contains cross-links to other databases that have more information about the characterized family.
- The second section provides a table showing how many of the motifs that make up the fingerprint occurs in the how many of the sequences in that family.
- The last section of the entry contains the actual fingerprints that are stored as multiple aligned sets of sequences, the alignment is made without gaps. There is, therefore, one set of aligned sequences for each motif.

Blocks

- The limitations of the above two databases led to the formation of Block database.
- In this database, the motifs (here called Blocks) are created automatically by highlighting and detecting the most conserved regions of each family of proteins.
- Block databases are fully automated.
- Keyword and sequence searching are the two important features of this type of database.
- Blocks are ungapped Multiple Sequence Alignment representing conserved protein regions.

Pfam

- Pfam contains the profiles used using Hidden Markov models.
- HMMs build the model of the pattern as a series of the match, substitute, insert or delete states, with scores assigned for alignment to go from one state to another.

- Each family or pattern defined in the Pfam consists of the four elements. The first is the annotation, which has the information on the source to make the entry, the method used and some numbers that serve as figures of merit.
- The second is the seed alignment that is used to bootstrap the rest of the sequences into the multiple alignments and then the family.
- The third is the HMM profile.
- The fourth element is the complete alignment of all the sequences identified in that family.

MHCPep:

- MHCPep is a database comprising over 13000 peptide sequences known to bind the Major Histocompatibility Complex of the immune system.
- Each entry in the database contains not only the peptide sequence, which may be 8 to 10 amino acid long but in addition has information on the specific MHC molecules to which it binds, the experimental method used to assay the peptide, the degree of activity and the binding affinity observed , the source protein that, when broken down gave rise to this peptide along with other, the positions along the peptide where it anchors on the MHC molecules and references and cross-links to other information.

Bibliographic databases

A bibliographic database is a database of bibliographic records, an organized digital collection of references to published literature, including journal and newspaper articles, conference proceedings, reports, government and legal publications, patents, books, etc. In contrast to library catalogue entries, a large proportion of the bibliographic records in bibliographic databases describe articles, conference papers, etc., rather than complete monographs, and they generally contain very rich subject descriptions in the form of keywords, subject classification terms, or abstracts.

OMIM

Online Mendelian Inheritance in Man (OMIM) is a continuously updated catalog of human genes and genetic disorders and traits, with a particular focus on the gene-phenotype relationship. As of 28 June 2019, approximately 9,000 of the over 25,000 entries in OMIM represented phenotypes; the rest represented genes, many of which were related to known phenotypes

OMIM is the online continuation of Dr. Victor A. McKusick's Mendelian Inheritance in Man (MIM), which was published in 12 editions between 1966 and 1998.[2][3][4] Nearly all of the 1,486 entries in the first edition of MIM discussed phenotypes.[2]

MIM/OMIM is produced and curated at the Johns Hopkins School of Medicine (JHUSOM). OMIM became available on the internet in 1987 under the direction of the Welch Medical Library at JHUSOM with financial support from the Howard Hughes Medical Institute. From 1995 to 2010, OMIM was available on the World Wide Web with informatics and financial support from the National Center for Biotechnology Information. The current OMIM website (OMIM.org), which was developed with funding from JHUSOM, is maintained by Johns Hopkins University with financial support from the National Human Genome Research Institute

The content of MIM/OMIM is based on selection and review of the published peer-reviewed biomedical literature. Updating of content is performed by a team of science writers and curators under the direction of Dr. Ada Hamosh at the McKusick-Nathans Institute of Genetic Medicine of Johns Hopkins University. While OMIM is freely available to the public, it is designed for use primarily by physicians and other health care professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine.[5]

The database may be used as a resource for locating literature relevant to inherited conditions,[7] and its numbering system is widely used in the medical literature to provide a unified index for genetic diseases

MIM classification system

MIM numbers

Each OMIM entry is given a unique six-digit identifier[9] as summarized below:

- 100000–299999: Autosomal loci or phenotypes (entries created before May 15, 1994)
- 300000–399999: X-linked loci or phenotypes
- 400000–499999: Y-linked loci or phenotypes
- 500000–599999: Mitochondrial loci or phenotypes
- 600000 and above: Autosomal loci or phenotypes (entries created after May 15, 1994)

In cases of allelic heterogeneity, the MIM number of the entry is followed by a decimal point and a unique 4-digit number specifying the variant.[9] For example, allelic variants in the HBB gene (141900) are numbered 141900.0001 through 141900.0538.[10]

Because OMIM has responsibility for the classification and naming of genetic disorders, these numbers are stable identifiers of the disorders.[5]

Symbols preceding MIM numbers

Symbols preceding MIM numbers[11] indicate the entry category:

- An asterisk (*) before an entry number indicates a gene.
- A number symbol (#) before an entry number indicates that it is a descriptive entry, usually of a phenotype, and does not represent a unique locus. The reason for the use of the number symbol is given in the first paragraph of the entry. Discussion of any gene(s) related to the phenotype resides in another entry (or entries) as described in the first paragraph.
- A plus sign (+) before an entry number indicates that the entry contains the description of a gene of known sequence and a phenotype.

- A percent sign (%) before an entry number indicates that the entry describes a confirmed Mendelian phenotype or phenotypic locus for which the underlying molecular basis is not known.
- No symbol before an entry number generally indicates a description of a phenotype for which the Mendelian basis, although suspected, has not been clearly established or that the separateness of this phenotype from that in another entry is unclear.
- A caret (^) before an entry number means the entry no longer exists because it was removed from the database or moved to another entry as indicated.

Pubmed

PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. The United States National Library of Medicine (NLM) at the National Institutes of Health maintain the database as part of the Entrez system of information retrieval.[1]

From 1971 to 1997, online access to the MEDLINE database had been primarily through institutional facilities, such as university libraries. PubMed, first released in January 1996, ushered in the era of private, free, home- and office-based MEDLINE searching.[2] The PubMed system was offered free to the public starting in June 1997

In addition to MEDLINE, PubMed provides access to:

- older references from the print version of *Index Medicus*, back to 1951 and earlier
- references to some journals before they were indexed in Index Medicus and MEDLINE, for instance *Science*, *BMJ*, and *Annals of Surgery*
- very recent entries to records for an article before it is indexed with Medical Subject Headings (MeSH) and added to MEDLINE
- a collection of books available full-text and other subsets of NLM records[4]
- PMC citations
- NCBI Bookshelf

Many PubMed records contain links to full text articles, some of which are freely available, often in PubMed Central[5] and local mirrors, such as Europe PubMed Central.[6]

Information about the journals indexed in MEDLINE, and available through PubMed, is found in the NLM Catalog.[7]

As of 27 January 2020, PubMed has more than 30 million citations and abstracts dating back to 1966, selectively to the year 1865, and very selectively to 1809. As of the same date, 20 million of PubMed's records are listed with their abstracts, and 21.5 million records have links to full-text versions (of which 7.5 million articles are available, full-text for free).[8] Over the last 10 years (ending 31 December 2019), an average of nearly 1 million new records were added each year. Approximately 12% of the records in PubMed correspond to cancer-related entries, which have grown from 6% in the 1950s to 16% in 2016.[9] Other significant proportion of records correspond to "chemistry" (8.69%), "therapy" (8.39%), and "infection" (5%).[*citation needed*]

In 2016, NLM changed the indexing system so that publishers are able to directly correct typos and errors in PubMed indexed articles.[10]

PubMed has been reported to include some articles published in predatory journals. MEDLINE and PubMed policies for the selection of journals for database inclusion are slightly different. Weaknesses in the criteria and procedures for indexing journals in PubMed Central may allow publications from predatory journals to leak into PubMed

PubMed Central (**PMC**) is a free digital repository that archives open access full-text scholarly articles that have been published in biomedical and life sciences journals. As one of the major research databases developed by the National Center for Biotechnology Information (NCBI), PubMed Central is more than a document repository. Submissions to PMC are indexed and formatted for enhanced metadata, medical ontology, and unique identifiers which enrich the XML structured data for each article.[1] Content within PMC can be linked to other NCBI databases and accessed via Entrez search and retrieval systems, further enhancing the public's ability to discover, read and build upon its biomedical knowledge.[2]

PubMed Central is distinct from PubMed.[3] PubMed Central is a free digital archive of full articles, accessible to anyone from anywhere via a web browser (with varying provisions for reuse). Conversely, although PubMed is a searchable database of biomedical citations and abstracts, the full-text article resides elsewhere (in print or online, free or behind a subscriber paywall).

As of December 2018, the PMC archive contained over 5.2 million articles,[4] with contributions coming from publishers or authors depositing their manuscripts into the repository per the NIH Public Access Policy. Earlier data shows that from January 2013 to January 2014 author-initiated deposits exceeded 103,000 papers during a 12-month period.[5] PMC identifies about 4,000 journals which participate in some capacity to deposit their published content into the PMC repository.[6] Some publishers delay the release of their articles on PubMed Central for a set time after publication, referred to as an "embargo period", ranging from a few months to a few years depending on the journal. (Embargoes of six to twelve months are the most common.) PubMed Central is a key example of "systematic external distribution by a third party"[7] which is still prohibited by the contributor agreements of many publishers

The PMCID (PubMed Central identifier), also known as the PMC reference number, is a bibliographic identifier for the PubMed Central database, much like the PMID is the bibliographic identifier for the PubMed database. The two identifiers are distinct however. It consists of "PMC" followed by a string of seven numbers. The format is:[27]

- PMCID: PMC1852221

Authors applying for NIH awards must include the PMCID in their application.

MEDLINE

MEDLINE (Medical Literature Analysis and Retrieval System Online, or MEDLARS Online) is a bibliographic database of life sciences and biomedical information. It includes bibliographic information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. MEDLINE also covers much of the literature in biology and biochemistry, as well as fields such as molecular evolution.

Compiled by the United States National Library of Medicine (NLM), MEDLINE is freely available on the Internet and searchable via PubMed and NLM's National Center for Biotechnology Information's Entrez system.

The database contains more than 26 million records[6] from 5,639 selected publications[7] covering biomedicine and health from 1950 to the present. Originally, the database covered articles starting from 1965, but this has been enhanced, and records as far back as 1950/51 are now available within the main index. The database is freely accessible on the Internet via the PubMed interface and new citations are added Tuesday through Saturday. For citations added during 1995-2003: about 48% are for cited articles published in the U.S., about 88% are published in English, and about 76% have English abstracts written by authors of the articles. The most common topic in the database is Cancer with around 12% of all records between 1950-2016, which have risen from 6% in 1950 to 16% in 2016

MEDLINE uses Medical Subject Headings (MeSH) for information retrieval. Engines designed to search MEDLINE (such as Entrez and PubMed) generally use a Boolean expression combining MeSH terms, words in abstract and title of the article, author names, date of publication, etc. Entrez and PubMed can also find articles similar to a given one based on a mathematical scoring system that takes into account the similarity of word content of the abstracts and titles of two articles.[9]

MEDLINE added a "publication type" term for “randomized controlled trial” in 1991 and a MESH subset “systematic review” in 2001

MEDLINE functions as an important resource for biomedical researchers and journal clubs from all over the world. Along with the Cochrane Library and a number of other databases, MEDLINE facilitates evidence-based medicine.[11][12][13] Most systematic review articles published presently build on extensive searches of MEDLINE to identify articles that might be useful in the review.[11][12] MEDLINE influences researchers in their choice of journals in which to publish

PubMed usage has been on the rise since 2008. In 2011, PubMed/MEDLINE was searched 1.8 billion times, up from 1.6 billion searches in the previous year.[16]

A service such as MEDLINE strives to balance usability with power and comprehensiveness. In keeping with the fact that MEDLINE's primary user community is professionals (medical scientists, health care providers), searching MEDLINE effectively is a learned skill; untrained users are sometimes frustrated with the large numbers of articles returned by simple searches.

Counterintuitively, a search that returns thousands of articles is not guaranteed to be comprehensive. Unlike using a typical Internet search engine, PubMed searching of MEDLINE requires a little investment of time. Using the MeSH database to define the subject of interest is one of the most useful ways to improve the quality of a search. Using MeSH terms in conjunction with limits (such as publication date or publication type), qualifiers (such as adverse effects or prevention and control), and text-word searching is another. Finding one article on the subject and clicking on the "Related Articles" link to get a collection of similarly classified articles can expand a search that otherwise yields few results.

For lay users who are trying to learn about health and medicine topics, the NIH offers MedlinePlus; thus, although such users are still free to search and read the medical literature themselves (via PubMed), they also have some help with curating it into something comprehensible and practically applicable for patients and family members.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOINFORMATICS

UNIT – 4 - INTRODUCTION TO BIOINFORMATICS– SBIA1101

File formats

In the field of bioinformatics there exists many different file formats that store DNA and protein sequence information. There is no one sequence format that is ideal: many are used in different contexts, and can often be converted from one to another for easier access or sharing. Below is a list of file formats and a link to their respective file format specs and descriptions for anyone wishing to get to know the file formats a little better. While there are many different formats out there used by commercial software, this list focuses mainly on open, non-propietary file formats.

What is a file format?

A *file format* is a way for computers (and humans) to standardize how data is organized. For example, this page was written on an .html extension. HTML files contain special *tags* that tell the browser what each block of text is, and how to display it on the page.
Additionally, computers are able to check file formats and immediately determine whether it should be opened in a text editor (for editing), a modern browser (for viewing) or some other software.
File types can also indicate which algorithm to use to view (or open) that file. For example, .gif, .jpg and .png all display images, but the level of compression, size and resolution differ.

- **Genbank** - quite possibly the standard in sequence file formats, the Genbank format is widely used by public databases such as NCBI. The Genbank file format is quite flexible and allows annotations, comments, and references to be included within the file. The file is plain text and thus can be read with a text editor. Genbank files often have the file extension '.gb' or '.genbank'.

- **EMBL** - similar in form to the Genbank file, the EMBL format is used by public databases such as European Molecular Biology Laboratory. The Genbank file format is quite flexible and allows annotations, comments, and references to be included within the file. The file is plain text and thus can be read with a text editor. Genbank files often have the file extension '.gb' or '.genbank'.

- **PDB** - the PDB file format is used to store both sequence information, but more importantly stores 3-dimensional structure information. This information can be used to visualize the crystal structure of a given molecule (typically a protein). PDB files are simply text files, thus can be viewed with a text editor, and often have the file extension '.pdb'.

- **MDL -** While not technically containing sequence data, the MDL file format is worth including in this list. The MDL mol file contains information regarding small molecules, the spec being quite similar to that of the PDB file format. The MDL mol file contains information regarding 2d (and possibly 3d) molecule structure, such as atom type and atom connectivity.

- **FASTA format**

  File format : FASTA

  File extensions : file.fa, file.fasta, file.fsa

  Example :

```
>XR_002086427.1 Candida albicans SC5314 uncharacterized ncRNA (SCR1), ncRNA
TGGCTGTGATGGCTTTTAGCGGAAGCGCGCTGTTCGCGTACCTGCTGTTTGTTGAA
AATTTAAGAGCAAAGTGTCCGGCTCGATCCCTGCGAATTGAATTCTGAACGCTAG
AGTAATCAGTGTCTTTCAAGTTCTGGTAATGTTTAGCATAACCACTGGAGGGAAG
CAATTCAGCACAGTAATGCTAATCGTGGTGGAGGCGAATCCGGATGGCACCTTGT
TTGTTGATAAATAGTGCGGTATCTAGTGTTGCAACTCTATTTTT
```

Fasta format is a simple way of representing nucleotide or amino acid sequences of nucleic acids and proteins. This is a very basic format with two minimum lines. First line referred as comment line starts with '>' and gives basic information about sequence. There is no set format for comment line. Any other line that starts with ';' will be ignored. Lines with ';' are not a common feature of fasta files. After comment line, sequence of nucleic acid or protein is included in standard one letter code. Any tabulators, spaces, asterisks etc in sequence will be ignored.

Plain sequence format

A sequence in plain format may contain only IUPAC characters and spaces (no numbers!).

Note: A file in plain sequence format may only contain **one** sequence, while most other formats accept several sequences in one file.

An example sequence in plain format is:

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

FASTQ format

A sequence file in FASTQ format can contain several sequences.
FASTQ is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. It is mainly used for storing the output of high-throughput sequencing instruments.
A FASTQ file usually uses four lines per sequence.

1. a '@' character, followed by a sequence identifier and an optional description
2. the raw sequence letters.
3. a '+' character, optionally followed by the same sequence identifier (and any description)
4. quality values for the sequence in Line 2

An example sequence in FASTQ format is:

```
@SEQUENCE_ID
GTGGAAGTTCTTAGGGCATGGCAAAGAGTCAGAATTTGAC
+
FAFFADEDGDBGEGGBCGGHE>EEBA@@=
```

For a detailed decription please see the Wikipedia entry.

EMBL format

A sequence file in EMBL format can contain several sequences.
One sequence entry starts with an identifier line ("ID"), followed by further annotation lines. The start of the sequence is marked by a line starting with "SQ" and the end of the sequence is marked by two slashes ("//").

An example sequence in EMBL format is:

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                                368
//
```

FASTA format

A sequence file in FASTA format can contain several sequences.
Each sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line must begin with a greater-than (">") symbol in the first column.

An example sequence in FASTA format is:

```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin like peptide,
complete cds.|len=368
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

GCG format

A sequence file in GCG format contains exactly one sequence, begins with annotation lines and the start of the sequence is marked by a line ending with two dot ("..") characters. This line also contains the sequence identifier, the sequence length and a checksum. This format should only be used if the file was created with the GCG package.

An example sequence in GCG format is:

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
AB000263  Length: 368  Check: 4514  ..
       1  acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
      61  ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
     121  caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
     181  aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
     241  gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
     301  agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
     361  gacctgaa
```

GCG-RSF (rich sequence format)

The new GCG-RSF can contain several sequences in one file. This format should only be used if the file was created with the GCG package.

GenBank format

A sequence file in GenBank format can contain several sequences.
One sequence in GenBank format starts with a line containing the word LOCUS and a number of annotation lines. The start of the sequence is marked by a line containing "ORIGIN" and the end of the sequence is marked by two slashes ("//").

An example sequence in GenBank format is:

```
LOCUS       AB000263                 368 bp    mRNA    linear   PRI 05-FEB-1999
DEFINITION  Homo sapiens mRNA for prepro cortistatin like peptide, complete
            cds.
ACCESSION   AB000263
ORIGIN
        1 acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
       61 ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
      121 caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
      181 aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
      241 gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
      301 agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
      361 gacctgaa
//
```

IG format

A sequence file in IG format can contain several sequences, each consisting of a number of comment lines that must begin with a semicolon (";"), a line with the sequence name (it may not contain spaces!) and the sequence itself terminated with the termination character '1' for linear or '2' for circular sequences.

An example sequence in IG format is:

```
; comment
; comment
AB000263
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA1
```

Genomatix annotation syntax

Some Genomatix tools, e.g. Gene2Promoter or GPD allow the extraction of sequences. Genomatix uses the following syntax to annotate sequence information: each information item is denoted by a keyword, followed by a "=" and the value. These information items are separated by a pipe symbol "|".
The keywords are the following:

loc	The **Genomatix Locus Id**, consisting of the string "GXL_" followed by a number.
sym	The **gene symbol**. This can be a (comma-separated) list.
geneid	The **NCBI Gene Id**. This can be a (comma-separated) list.
acc	A **unique identifier** for the sequence. E.g. for Genomatix promoter regions, the Genomatix Promoter Id is

	listed in this field.
taxid	The organism's **Taxon Id**
spec	The **organism name**
chr	The **chromosome** within the organism.
ctg	The **NCBI contig** within the chromosome.
str	**Strand**, (+) for sense, (-) for antisense strand.
start	**Start position** of the sequence (relative to the contig).
end	**End position** of the sequence (relative to the contig).
len	**Length** of the sequence in base pairs.
tss	A (comma-separated list of) **UTR-start/TSS position(s)**. If there are several TSS/UTR-starts, this means that several transcripts share the same promoter (e.g. when they are splice variants). The positions are relative to the promoter region.
probe	A (comma-separated list of) **Affymetrix Probe Id(s)**.
unigene	A (comma-separated list of) **UniGene Cluster Id(s)**.
homgroup	An identifier (a number) for the **homology group** (available for promoter sequences only). Orthologously related sequences have the same value in this field.
promset	If the sequence is a promoter region, the **promoter set** is denoted here.
eldorado	The **ElDorado version** from which the sequence has been extracted.
descr	The **gene description**. If several genes (i.e. NCBI gene ids) are associated with the sequence, the descriptions for all of the genes are listed, separated by ";"
comm	A **comment** field, used for additional annotation. For promoter sequences, this field contains information about the transcripts associated with the promoter. For each transcript the Genomatix Transcript Id, accession number, TSS position and quality is listed, separated by "/". For Genomatix CompGen promoters no transcripts are assigned, in this case the string "CompGen promoter" is denoted.

This syntax is currently used only for sequences in the FASTA and GenBank formats.

Example (a promoter sequence in GenBank format):

```
LOCUS       GXP_4405072(PAX6/human)    1105 bp    DNA
DEFINITION  loc=GXL_141121|sym=PAX6|geneid=5080|acc=GXP_4405072|
            taxid=9606|spec=Homo sapiens|chr=11|ctg=NC_000011|str=(-)|
            start=31806821|end=31807925|len=1105|tss=1001,1005|
            homgroup=-|promset=-|eldorado=E32R1605|descr=paired box 6|
            comm=GXT_25635656/ENST00000455099/1005/gold;
            GXT_27757207/NM_001310159/1001/bronze
ACCESSION   GXP_4405072
BASE COUNT    229 a  239 c  313 g  324 t
```

```
ORIGIN
        1 GACTTTTTTT TTTTTTCCTT TGGGAAAGGT AGGGAGGTGT TCGTACGGGA GCAGCCTCGG
       61 GGACCCCTGC ACTGGGTCAG GGCTTATGAA GCTAGAAGCG TCCCTCTGTT CCCTTTGTGA
      121 GTTGGTGGGT TGTTGGTACA TTTGGTTGGA AGCTGTGTTG CTGGTTAGGG AGACTCGGTT
      181 TTGCTCCTTG GGTTCGAGGA AAGCTGGAGA ATAGAAGCCA TTGTTTGCCG TCTGTCGGCT
      241 TTGTCGACCA CGCTCACCCC CTCCTGTTCG TACTTTTTAA AGCAGTGAGG CGAGGTAGAC
      301 AGGGTGTGTC ACAGTACAGT TAAAGGGGTG AAGATCTAAA CGCCAAAAGA GAAGTTAATC
      361 ACAATAAGTG AGGTTTGGGA TAAAAAGTTG GGCTTGCCCC TTTCAAAGTC CCAGAAAGCT
      421 GGGAGGTAGA TGGAGAGGGG GCCATTGGGA AGTTTTTTTG GTGTAGGGAG AGGAGTAGAA
      481 GATAAAGGGT AAGCAGAGTG TTGGGTTCTG GGGGTCTTGT GAAGTTCCTT AAGGAAGGAG
      541 GGAGTGTGGC CCTGCAGCCC TCCCAAACTG CTCCAGCCTA TGCTCTCCGG CACCAGGAAG
      601 TTCCAAGGTT CCCTTCCCCT GGTCTCCAAA CTTCAGGTAT TCCTCTCCCC TCACACCCCT
      661 TCAACCTCAG CTCTTGGCCT CTACTCCTTA CTCCACTGTT CCTCCTGTTT CCCCCTTCCC
      721 CTTTTCCTGG TTCTTTATAT TTTTGCAAAG TGGGATCCGA ACTTGCTAGA TTTTCCAATT
      781 CTCCCAAGCC AGACCAGAGC AGCCTCTTTT AAAGGATGGA GACTTCTGTG GCAGATGCCG
      841 CTGAAAATGT GGGTGTAATG CTGGGACTTA GAGTTTGATG ACAGTTTGAC TGAGCCCTAG
      901 ATGCATGTGT TTTTCCTGAG AGTGAGGCTC AGAGAGCCCA TGGACGTATG CTGTTGAACC
      961 ACAGCTTGAT ATACCTTTTT CTCCTTCTGT TTTGTCTTAG GGGGAAGACT TTAACTAGGG
     1021 GCGCGCAGAT GTGTGAGGCC TTTTATTGTG AGAGTGGACA GACATCCGAG ATTTCAGGCA
     1081 AGTTCTGTGG TGGCTGCTTT GGGCT
//
```

Sequence Submission
NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION

- Houses series of databases relevent to biotechnology and biomedicine.
- Mainly genbank for DNA and PubMed, a bibliographic database for biomedical literature, epigenomics database.
- Director: David Lipman
- Found: founded in 1988 through legislation sponsored by senetor Claude Pepper.

The sequence we want to submit will be added to one of these databases:
- GenBank
- Sequence Read Archive (SRA)
- dbSNP (single nucleotide polymorphism)
- dbVar (genomic variant)
- GEO (gene expression Omnibus)

GenBank Submission Types

Standard

GenBank accepts mRNA or genomic sequence data directly determined by the submitter. The submission must include information about the source organism and annotation provided by the submitter. More details about adding annotation and sample files can be found in the GenBank Submissions Handbook . If you have any questions about the best method for submitting your data, please contact our user services group at: info@ncbi.nlm.nih.gov.

The following data is not accepted by GenBank:

- Noncontiguous sequences
- Primer sequences
- Protein sequences with no underlying nucleotide submission
- Sequence containing a mix of genomic and mRNA sequence
- Sequences without a physical counterpart (consensus sequences)
- Sequences with length less than 200 nucleotides
- Raw sequence reads from next generation sequencing platforms should be submitted to the Sequence Read Archive (SRA).
- Sequence data not directly obtained by the submitter may be acceptable for the Third Party Annotation database.

Complete Microbial Genomes

The Bacterial Genome Submission Guidelines page provides a detailed guide to help bacterial genome submitters prepare their submissions.

Whole Genome Shotgun (WGS) Sequences

Genomic sequence read-overlap contig sequences and assemblies from ongoing Whole Genome Shotgun (WGS) sequencing projects of prokaryotic and eukaryotic genomes with or without annotations can be submitted and should be updated as sequencing progresses and new assemblies are computed. Detailed submission instructions can be found on the WGS submission guide.

Transcriptome Shotgun Assembly (TSA) Sequences

Transcriptomic sequence read-overlap contig sequences computationally assembled from primary data submitted to Sequence Read Archive (SRA) can be submitted to TSA. Detailed submission instructions can be found on the TSA submission guide.

High-Throughput Genomic (HTGs) Sequences

Clone-based High-Throughput Genomic Sequence (usually cosmids or BACs) submissions can be submitted to GenBank. The HTGs page provides detailed submission instructions for genome centers.

Third Party Annotation (TPA)

The TPA (Third Party Annotation) database accepts third party annotation of genomic sequences or computationally derived/assembled sequences. TPA submissions must include sequence data that is already represented in GenBank, and the analysis upon which the annotations are based must appear in a peer-reviewed scientific journal. Detailed requirements and submission instructions can be found on the TPA submission guide.

Targeted Locus Study (TLS)

A Targeted Locus Study (TLS) is a large scale targeted sequencing project (>2,500 sequences) for either a single gene locus from multiple organisms or multiple conserved elements derived from a single organism. Detailed submission instructions can be found on the TLS submission guide.

If we want to submit a single sequence and assume it in GenBank then we will be requiring BankIt or Sequin, these are sequence submitting tools.
NCBI: To submit a sequence in NCBI we need certain tools, which are easily found in the NCBI page itself

we use BankIt if,
• We have a single sequence, a simple set of sequences (for example:16S rRNA, matK, ITS/rRNA, amoE, tefB, cytb, or COI sets), or a small batch of different sequences
• we prefer to use a web-based submission tool
• the feature annotation for our sequences is not complicated
• we do not require advanced sequence analysis tools

we use Sequin if,
• we prefer to work on our submission off-line
• we have a sequence or sequences that are complex
• we would like graphical viewing and editing options, including an alignment editor
• we would like the option to have network access to related analytical tools

GenBank Sequence Submission Policy
• the GenBank database is intended for new sequence data that is determined by and annotated by the submitter
• sequences built or derived from other GenBank primary data intended for the Third Party Annotation (TPA) database may be submitted through BankIt
• the following types of submissions are NOT acceptable: – sequences less than 200 nucleotides long, unless they represent complete exons, non-coding RNAs (ncRNAs), microsatellites or ancient DNA – non-contiguous sequences that have been artificially joined; for example, multiple exons without their intervening introns or without a 'gap' representing any missing sequence – single sequences that are a mix of molecule types, such as mix of genomic and mRNA sequence data

THROUGH BankIt:
• registration through the MyNCBI Login.
• sequence data can be either cut-and-pasted as text or uploaded as file (multiple sequences must be in a FASTA format)
• date for public release (immediate or at a specified future date)
• basic information (authors and a working title) for a corresponding reference paper
• name(s) of the organism(s) from which the sequence data were isolated and any other related descriptive data
• sequence features (for example: CDS, gene, rRNA, tRNA, with nucleotide intervals and product names)

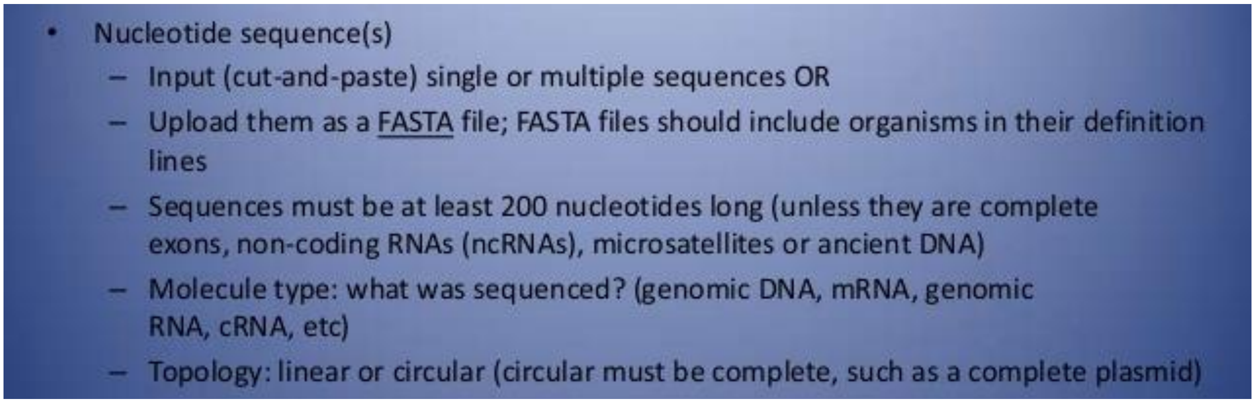
To Submit through BankIt we need to follow:

• Contact Information – Name, address, phone number, fax number and email address of the submitter must be entered when registering and submitting for the first time – Subsequent BankIt submissions will retain this information and display it once the submitter logs in

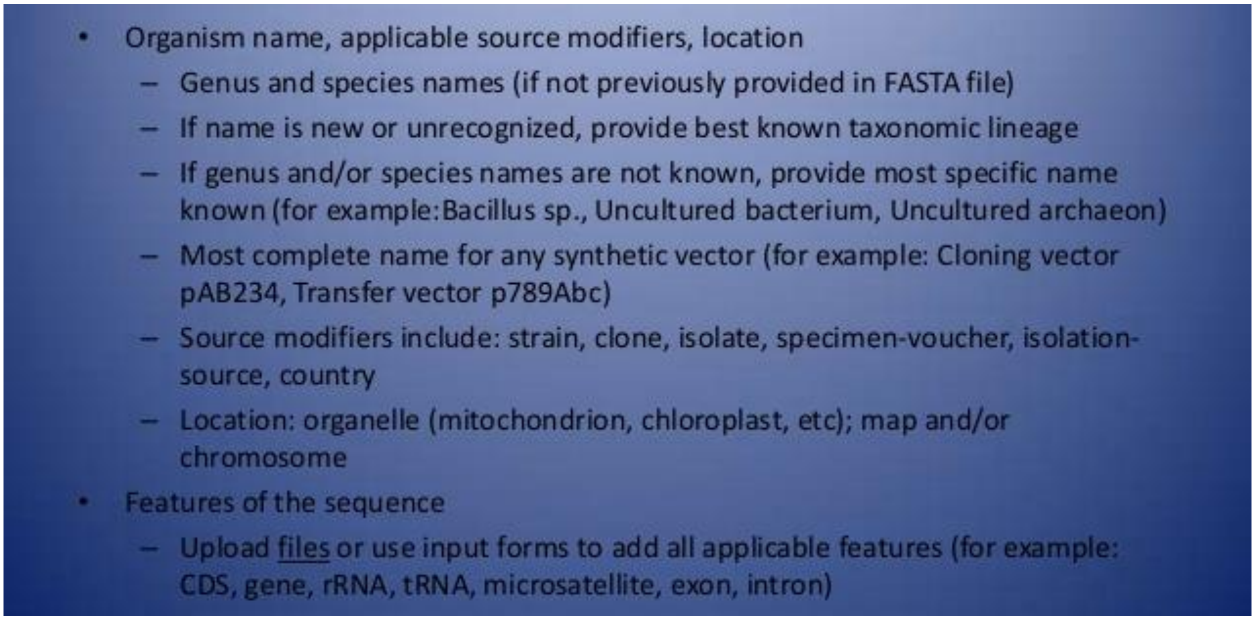
• Release date information – Immediately after it is processed at NCBI OR – On a date the submitter specifies

• Reference information – Sequence authors: names of the researchers who are credited with the sequence – Publication information: Unpublished, In-Press, or Published; and applicable citation information (paper's title, authors, journal title, volume, issue, year, pages)

• Submission Category and Type – Original sequencing or Third Party Annotation – Single sequence, sequence set (phylogenetic, population, environmental, etc), or batch



- Nucleotide sequence(s)
  - Input (cut-and-paste) single or multiple sequences OR
  - Upload them as a FASTA file; FASTA files should include organisms in their definition lines
  - Sequences must be at least 200 nucleotides long (unless they are complete exons, non-coding RNAs (ncRNAs), microsatellites or ancient DNA)
  - Molecule type: what was sequenced? (genomic DNA, mRNA, genomic RNA, cRNA, etc)
  - Topology: linear or circular (circular must be complete, such as a complete plasmid)



- Organism name, applicable source modifiers, location
  - Genus and species names (if not previously provided in FASTA file)
  - If name is new or unrecognized, provide best known taxonomic lineage
  - If genus and/or species names are not known, provide most specific name known (for example:Bacillus sp., Uncultured bacterium, Uncultured archaeon)
  - Most complete name for any synthetic vector (for example: Cloning vector pAB234, Transfer vector p789Abc)
  - Source modifiers include: strain, clone, isolate, specimen-voucher, isolation-source, country
  - Location: organelle (mitochondrion, chloroplast, etc); map and/or chromosome
- Features of the sequence
  - Upload files or use input forms to add all applicable features (for example: CDS, gene, rRNA, tRNA, microsatellite, exon, intron)

Sequin--A DNA Sequence Submission Tool

What Is Sequin

Sequin is a stand-alone software tool developed by the NCBI for submitting and updating entries to the GenBank sequence database. It is capable of handling simple submissions that contain a single short mRNA sequence, and complex submissions containing long sequences, multiple annotations, gapped sequences, or phylogenetic and population studies. A single Sequin file should contain less than 10,000 sequences for maximum performance. Larger submissions should be made with tbl2asn .

How to Get Sequin

Sequin 15.50 is currently available from the NCBI. Sequin runs on Macintosh, PC/Windows, and UNIX computers. Instructions for downloading and installing the program are provided. The program itself, along with its on-line help documentation, is available by anonymous FTP.

Sequin Help Documentation

A window containing the Sequin Help Documentation is opened when the Sequin program is launched. The contents of this scrolling window change as you move within the Sequin program, presenting you with help documentation appropriate for the section of Sequin you are presently visiting. This documentation is also available in a World Wide Web format. Detailed instructions for the various Sequin Wizards are also available.

Annotation Using A Table

A five-column, tab-delimited table of feature locations and qualifiers can be used to import annotation into an existing Sequin submission. This is the same table format that must be used to annotate features when creating a submission using tbl2asn.

Network-Aware Sequin

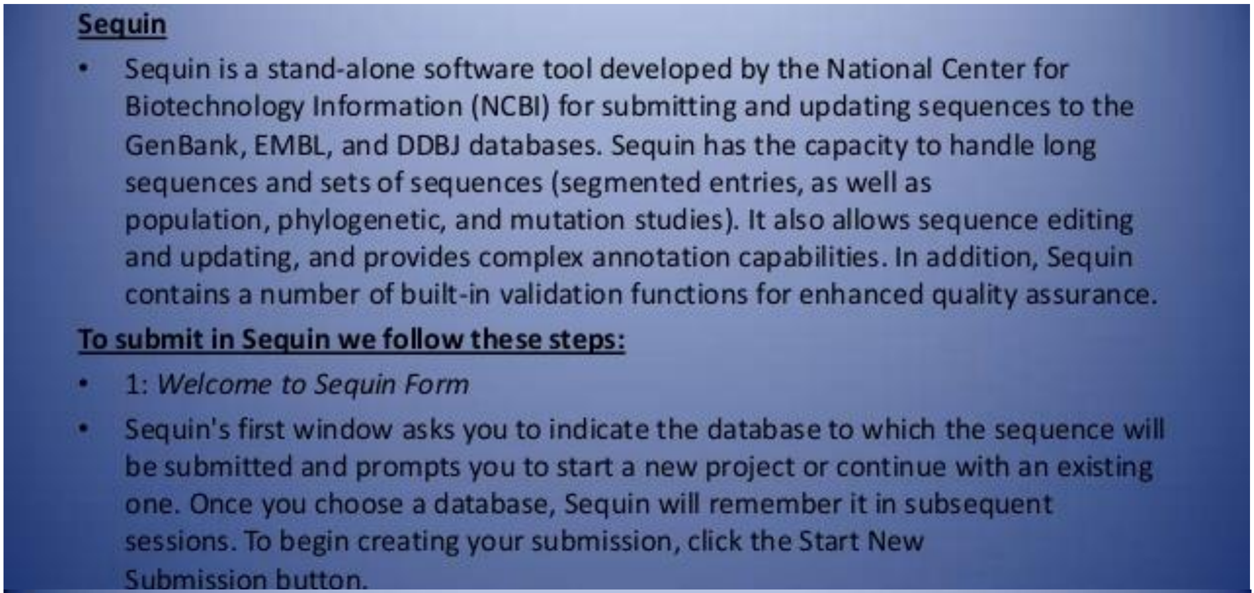
Sequin can be used in one of two modes, stand-alone or network-aware. In the network-aware mode, the program can exchange data between any computer connected to the Internet and the NCBI.

SequinMacroSend

The SequinMacroSend tool allows the submission of very large Sequin files directly. Those files that may be truncated during mailing with conventional maile rs, including large population sets or complete plasmids or small genomes, can be sent using this method.

Tbl2asn

The tbl2asn command line program is available via ftp and is designed as an alternative to the Sequin program for generating large single submissions (complete genomes) containing a great deal of annotation. It can also be used to generate a batch submission containing thousands of individual sequences. More detailed instructions about using this function are provided. When submitting a complete bacterial genome, please review the genome guidelines.

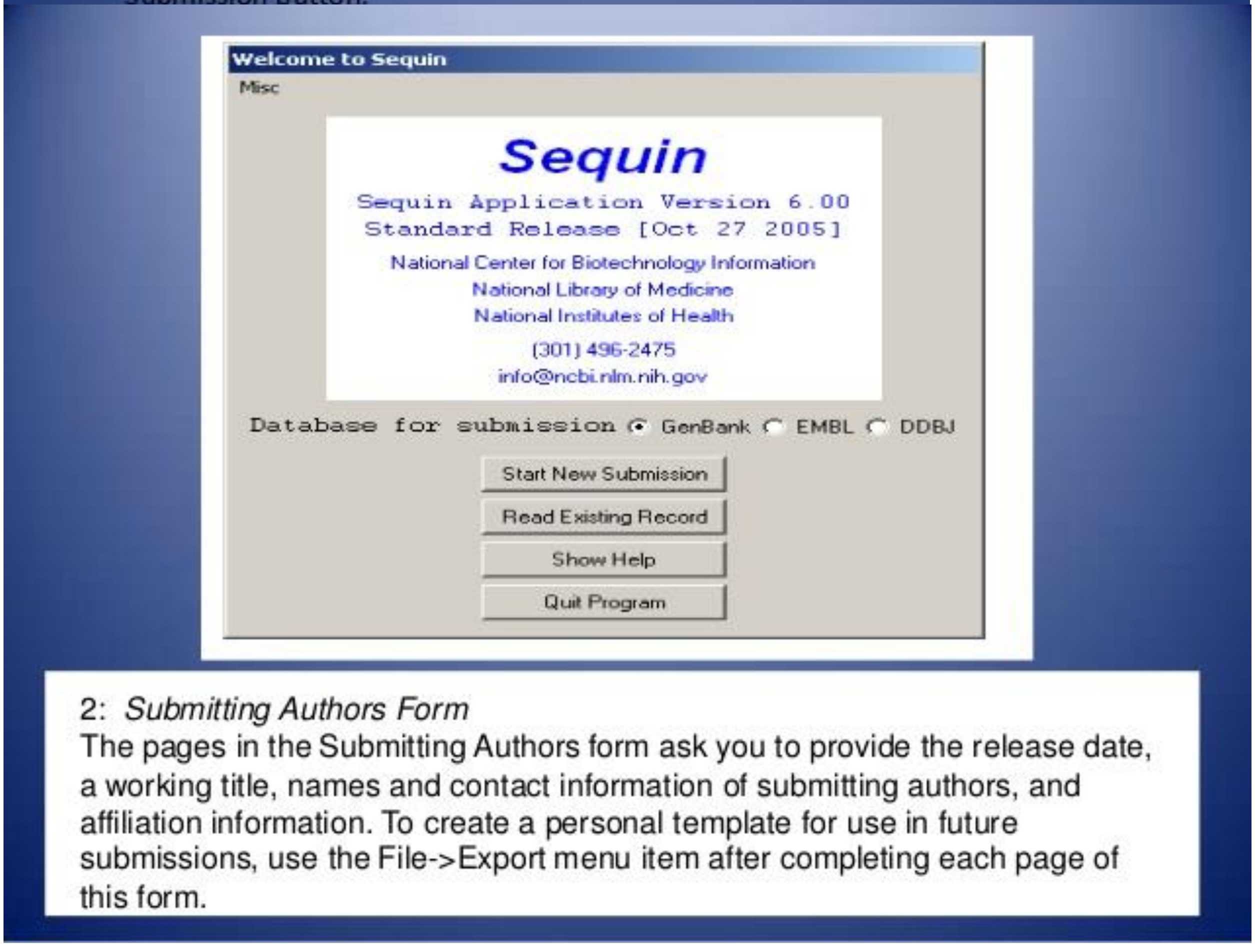


Sequin

- Sequin is a stand-alone software tool developed by the National Center for Biotechnology Information (NCBI) for submitting and updating sequences to the GenBank, EMBL, and DDBJ databases. Sequin has the capacity to handle long sequences and sets of sequences (segmented entries, as well as population, phylogenetic, and mutation studies). It also allows sequence editing and updating, and provides complex annotation capabilities. In addition, Sequin contains a number of built-in validation functions for enhanced quality assurance.

To submit in Sequin we follow these steps:

- 1: *Welcome to Sequin Form*
- Sequin's first window asks you to indicate the database to which the sequence will be submitted and prompts you to start a new project or continue with an existing one. Once you choose a database, Sequin will remember it in subsequent sessions. To begin creating your submission, click the Start New Submission button.

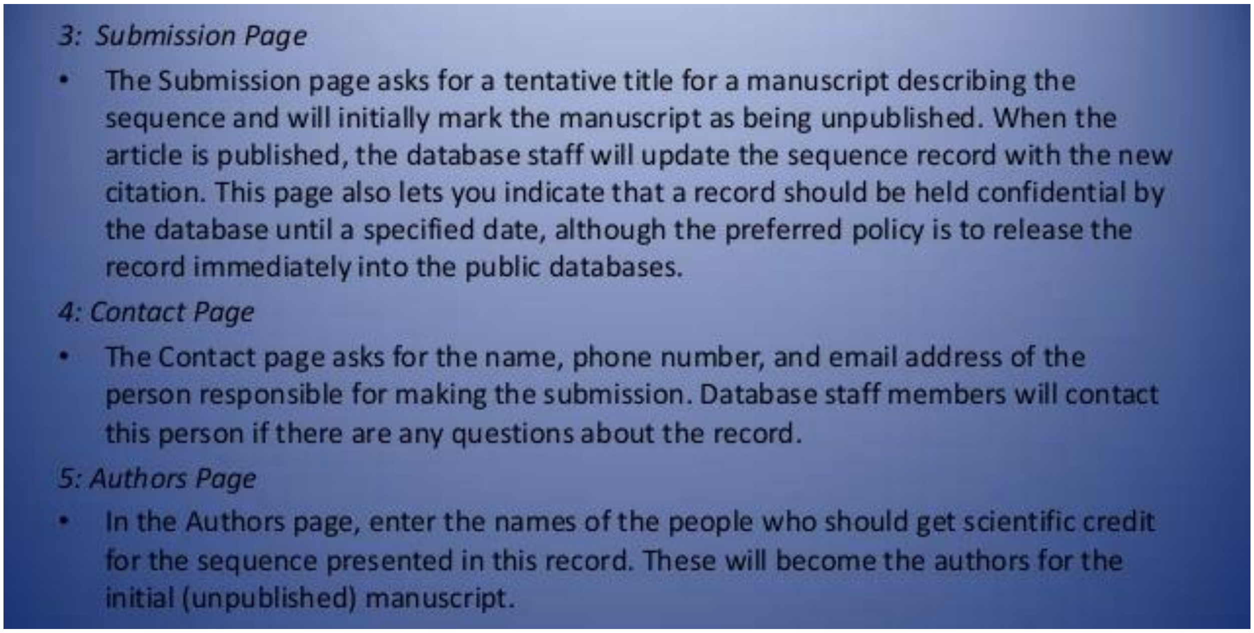


Welcome to Sequin
Misc
Sequin
Sequin Application Version 6.00
Standard Release [Oct 27 2005]
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
(301) 496-2475
info@ncbi.nlm.nih.gov
Start New Submission
Read Existing Record
Show Help
Quit Program

Database for submission ● GenBank ○ EMBL ○ DDBJ

2: *Submitting Authors Form*

The pages in the Submitting Authors form ask you to provide the release date, a working title, names and contact information of submitting authors, and affiliation information. To create a personal template for use in future submissions, use the File->Export menu item after completing each page of this form.



3: Submission Page

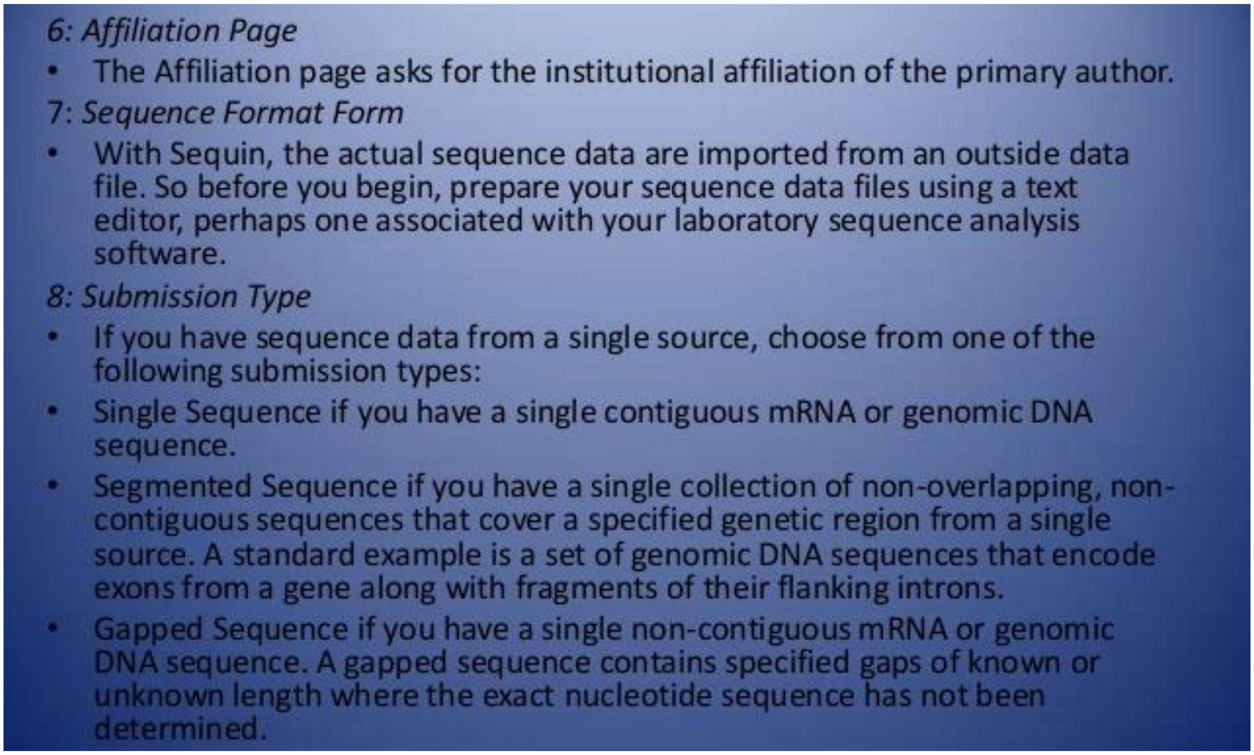
- The Submission page asks for a tentative title for a manuscript describing the sequence and will initially mark the manuscript as being unpublished. When the article is published, the database staff will update the sequence record with the new citation. This page also lets you indicate that a record should be held confidential by the database until a specified date, although the preferred policy is to release the record immediately into the public databases.

4: Contact Page

- The Contact page asks for the name, phone number, and email address of the person responsible for making the submission. Database staff members will contact this person if there are any questions about the record.

5: Authors Page

- In the Authors page, enter the names of the people who should get scientific credit for the sequence presented in this record. These will become the authors for the initial (unpublished) manuscript.



6: Affiliation Page

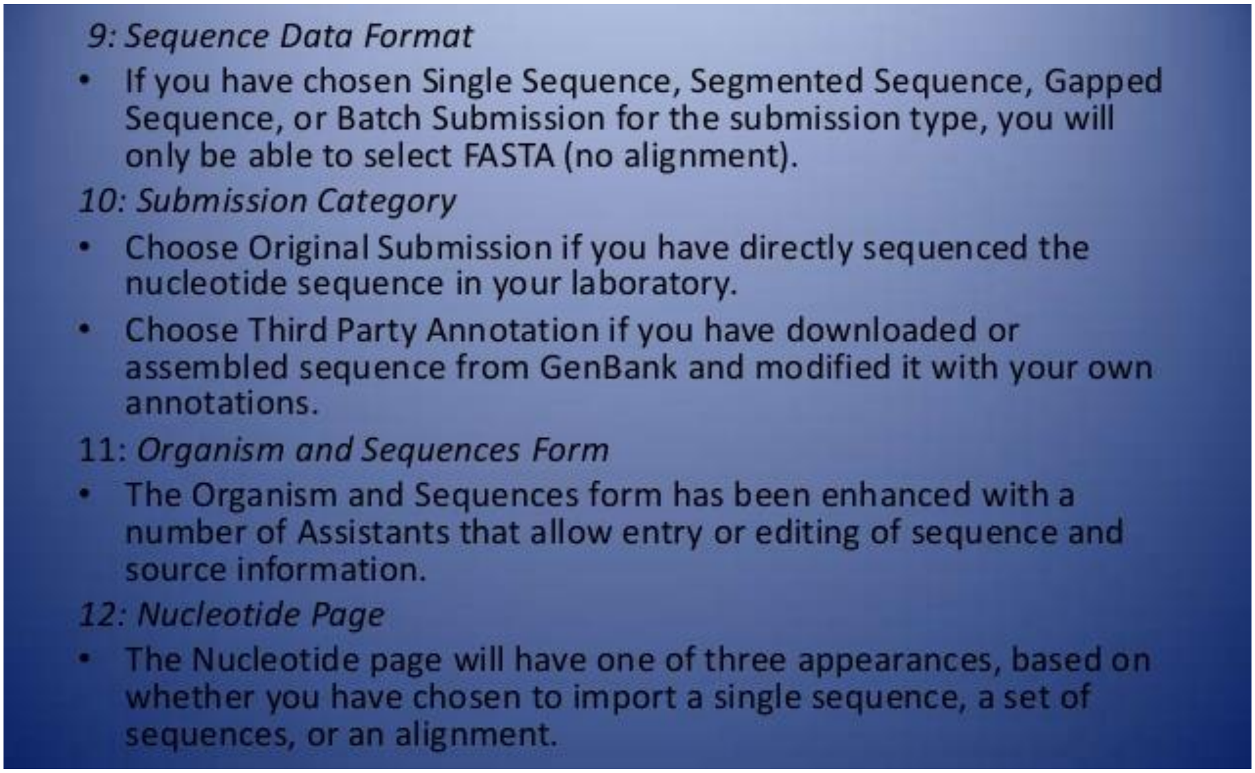
- The Affiliation page asks for the institutional affiliation of the primary author.

7: *Sequence Format Form*

- With Sequin, the actual sequence data are imported from an outside data file. So before you begin, prepare your sequence data files using a text editor, perhaps one associated with your laboratory sequence analysis software.

8: Submission Type

- If you have sequence data from a single source, choose from one of the following submission types:
- Single Sequence if you have a single contiguous mRNA or genomic DNA sequence.
- Segmented Sequence if you have a single collection of non-overlapping, non-contiguous sequences that cover a specified genetic region from a single source. A standard example is a set of genomic DNA sequences that encode exons from a gene along with fragments of their flanking introns.
- Gapped Sequence if you have a single non-contiguous mRNA or genomic DNA sequence. A gapped sequence contains specified gaps of known or unknown length where the exact nucleotide sequence has not been determined.



9: Sequence Data Format

- If you have chosen Single Sequence, Segmented Sequence, Gapped Sequence, or Batch Submission for the submission type, you will only be able to select FASTA (no alignment).

10: Submission Category

- Choose Original Submission if you have directly sequenced the nucleotide sequence in your laboratory.
- Choose Third Party Annotation if you have downloaded or assembled sequence from GenBank and modified it with your own annotations.

11: *Organism and Sequences Form*

- The Organism and Sequences form has been enhanced with a number of Assistants that allow entry or editing of sequence and source information.

12: Nucleotide Page

- The Nucleotide page will have one of three appearances, based on whether you have chosen to import a single sequence, a set of sequences, or an alignment.

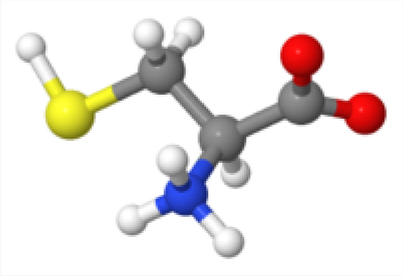
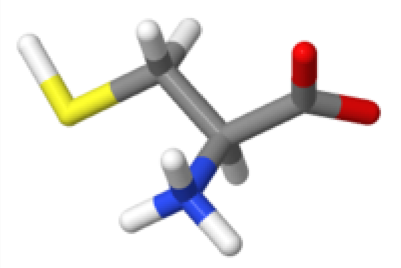
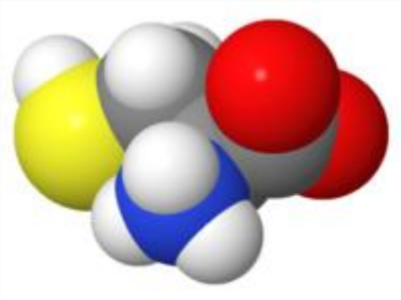
Visualization of Biomolecules

Introduction to molecular visualization

Molecular visualization means looking at molecular models in order to explore and understand them. Molecular visualization does not necessarily involve *molecular modeling*, which means creating molecular models, or changing the composition or configurations of existing models. Here we will be dealing primarily with models of macromolecules (protein, DNA, RNA, or their complexes).

Representations of Molecular Models

Atomic Representations

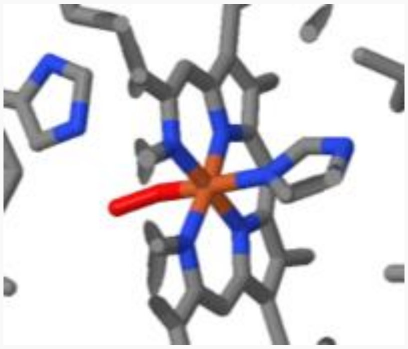
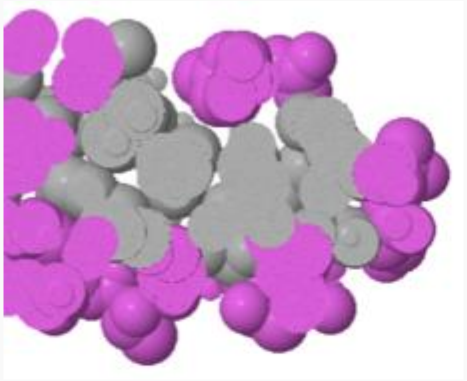
		
Ball & Stick	Stick (Wireframe)	Spacefilling
C H O N S		

Atomic representations (displays, renderings) include **ball-and-stick, stick (wireframe), and spacefilling**. The 20 amino acids are here represented in each of these 3 ways, and also illustrated in this page about Glycine. These representations show positions of atoms and covalent bonds. Hydrogen, shown in the images at right, is often missing in crystallographic models. Such representations are useful for looking at atomic detail, but become too cluttered to be useful for visualizing peptides or protein chains.

Ball and stick is one option in the *representations* tab of Proteopedia's Scene Authoring Tools. Another is *stick*, also called *wireframe*.

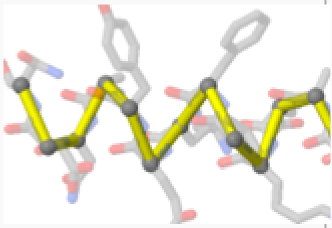
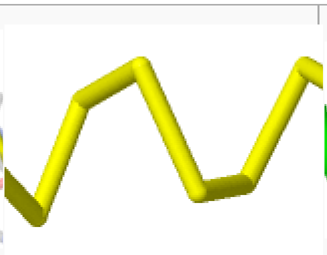
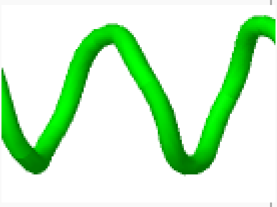
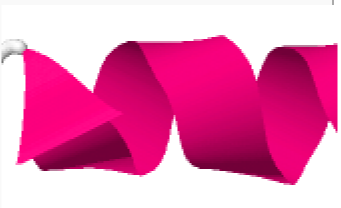
In FirstGlance in Jmol, you can isolate a small portion of a large structure, and then display it as sticks (*Vines/Sticks* in the *Views* tab). Or, rather than isolating it, which hides everything else, you can center it and then turn on slabbing.

Slabbing

	
Slab showing heme (1hho) C O N Fe	Slab showing **hydrophobic** core vs. **polar** (1pgb)

A useful way to see atomic details of a small part of a large macromolecular model is to center the moiety of interest, and then cut away the front and back portions of the molecule. This is called *slabbing* since one is, in effect, looking at a slab cut out of the larger model. Slabbing is also useful to see buried structures and their environments, such as the hydrophobic core of a protein domain. Slabbing can be done using FirstGlance in Jmol: In the *Views* tab, use *Center Atom* to center the region of interest, and then click the *Slab* button. Further instructions will appear automatically.

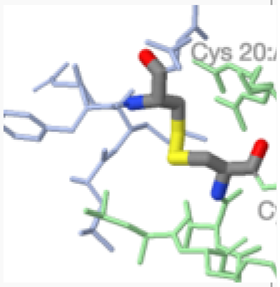
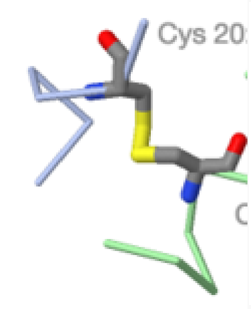
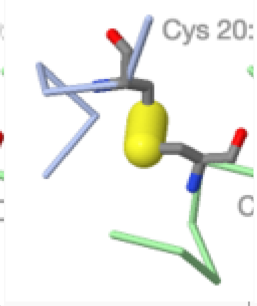
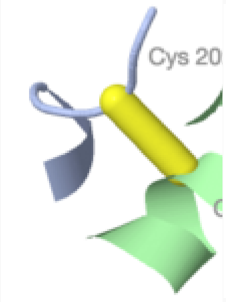
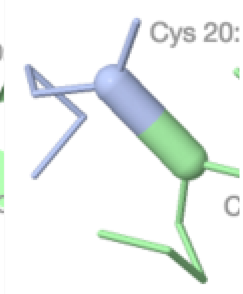
Simplified Schematic Representations

			
Alpha carbon backbone trace *simplifies*!	Alpha carbon backbone trace	Smoothed backbone trace	Ribbon backbone trace

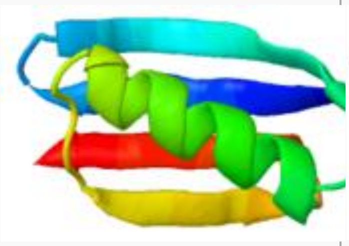
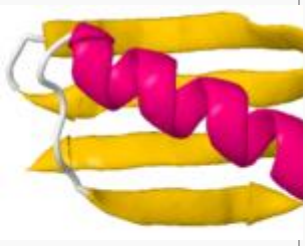
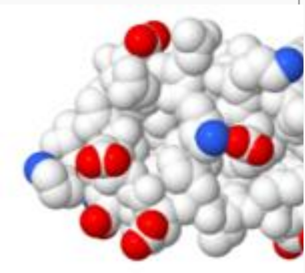
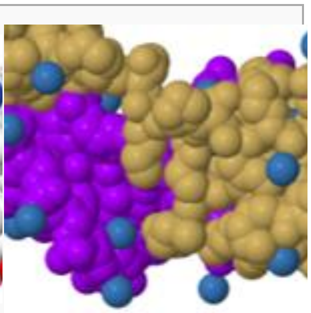
Backbones

Simplified representations of the polypeptide backbone or main chain, such as **backbone traces or ribbons/cartoons** are very helpful in understanding structure when it comes to large molecules such as proteins, DNA, RNA and their complexes. These representations are available under the *representations* tab in Proteopedia's Scene Authoring Tools, as well as in the *Views* tab in FirstGlance in Jmol.

Disulfide Bonds

				
Atomic detail of a between-chain disulfide bond in 9ins. C O N S	Protein chains A B simplified to backbone traces.	FirstGlance enlarges the sulfur-sulfur bond.	Schematic disulfide bridge connecting ribbon backbones.	Disulfide bridge colored by chain, an option in FirstGlance.

FirstGlance in Jmol highlights disulfide bonds in one click (in its *Tools* tab), and has several options for rendering and coloring them.

			
N->C Rainbow (1pgb) N 🟦🟦🟦🟩🟩🟨🟧 C	Secondary Structure **Alpha Helices**, **Beta Strands**, **Loops**.	Amino Acid Charge **Anionic (-) Cationic (+)**	Composition (1d66) **Protein**, **DNA**, **Solvent**

Color Schemes for Macromolecules

A set of standard color schemes for macromolecules, called DRuMS, was released in 2000. These color schemes are offered on buttons in Proteopedia's Scene Authoring Tools. They derive in part from physical ball and stick models (called Corey-Pauling-Kolton or CPK models) that pre-dated computer visualization. Those early colors for chemical elements (see examples above) were incorporated into early Molecular Visualization Software such as Kinemages, RasMol, and Chime. The CPK colors for chemical elements, and the DRuMS color schemes were incorporated into the color schemes built into Jmol, the visualization engine used in Proteopedia.

See Help:Color Keys for color-key templates in Proteopedia.

Visualizing Structural Features

Overall Features

Combinations of representations and color schemes are useful to highlight

1. Secondary, tertiary, and quaternary structure
2. N- and C-terminal ends of protein chains; 5' and 3' ends of nucleic acid chains.
3. Distribution of polar (charged or uncharged: hydrophilic) vs. hydrophobic amino acids on the surface and in the core (using slabbing).
4. The distribution of positive and negative charges on the surface of a protein.
5. Evolutionary conservation to identify functional regions of proteins.
6. Lipid bilayer boundaries for integral membrane proteins.

Features 1-4 can be most easily displayed in the *Views* tab of FirstGlance in Jmol. Evolutionary conservation can sometimes be seen with one click in Proteopedia, or in other cases requires submitting a job to the ConSurf server. Methods for visualizing lipid bilayer boundaries for integral membrane proteins are given in Jmol/Visualizing membrane position.

For a specific protein, such features can be displayed by using Proteopedia's Scene Authoring Tools and then attached to green links when authoring a page.

Covalent and Non-Covalent Interactions

FirstGlance in Jmol provides "one-click" routines, in its *Tools* tab, to highlight

- Disulfide bonds
- Salt bridges
- Cation-pi interactions
- Non-covalent interactions with any sub-structure that you select, using *Contacts & Non-Covalent Interactions* in the *Tools* tab.
- Distances between any two atoms, and angles or dihedral angles defined by 3 or 4 atoms, using *Distances/Angles* in the *Tools* tab.
- Crystal contacts.

Obtaining Molecular Models

You can browse for molecular models at the Atlas of Macromolecules, the Molecule of the Month, or Protein Spotlight.

Methods for searching the Protein Data Bank for published empirical 3D models are explained here. *Empirical models* are those determined by experimentation, notably X-ray diffraction, solution nuclear magnetic resonance, or electron cryomicroscopy. Empirical models are far more reliable than theoretical models, but one must pay attention to the quality of an empirical model since some are more reliable than others.

Empirical models are available for only a small fraction of all proteins, probably <10%. If an empirical model is not available, the next best thing would be a homology model. About one third of all proteins can be reliably homology modeled, but homology models have more uncertainties than do empirical models.

Softwares

Easy to use

The following visualization packages have extensive built-in help, and do not require that you learn any command language. All of these work in web browsers online.

- Proteopedia.Org - the easiest and most powerful[1] way to communicate 3D structure-function relationships online. It is a wiki that incorporates the Jmol applet (see below) for 3D interactive viewing. Molecular scenes can be imbedded in articles, and rotated and zoomed with the mouse. Generating new, custom molecular scenes is very easy using the built-in *Scene Authoring Tools* -- all menus, buttons, and forms. Molecular scenes are linked to the adjacent text describing them with *green links*.
- FirstGlance in Jmol (firstglance.jmol.org) - a free, open-source user-interface to Jmol utilized in the *3D View* links in papers in the journal Nature that report new macromolecular structures. FirstGlance in Jmol is probably the easiest-to-use[1] dedicated 3D macromolecular structure visualization software. It provides mostly "canned" views that reveal major structural features, but does not, for the most part, allow generation of customized molecular views. It is particularly strong in making it easy to visualize the noncovalent interactions between any moiety and the remainder of the structure. For more on its ease of use, and a comparison with other packages, see What Is FirstGlance in Jmol?.
- Polyview-3D (polyview.cchmc.org/polyview3d.html) - the easiest place to create publication-quality custom molecular views. It also creates high-quality animations suitable for Powerpoint® slides. Polyview-3D generates its images with PyMOL (see below).
- Protein Explorer (proteinexplorer.org) - an extensive and powerful open-source user-interface to the free MDL Chime browser plugin (see below) that enables users to create rotatable, zoomable customized molecular views. It is very easy to use, although taking full advantage of it requires many hours of experience because its power inevitably leads to some complexity. Protein Explorer has more help for beginners in macromolecular structure than do Proteopedia.Org or FirstGlance in Jmol. Because it depends on Chime, its use is now limited to MS Windows computers. Nevertheless, in 2009, there is no other package[1] that combines the ease of use with the power of Protein Explorer.
- Friend (ilyinlab.org/friend) - Integrated Front-End application for multiple structure visualization and multiple sequence alignment. Friend is a bioinformatics application designed for simultaneous analysis and visualization of multiple structures and sequences of proteins and/or DNA/RNA. The application provides functionalities such as: structure visualization with different rendering and coloring, sequence alignment, and simple phylogeny analysis, along with a number of extended features to perform more complex analyses of sequence structure relationships, including: structure alignment of proteins, investigation of specific interaction motifs, studies of protein-protein and protein-DNA interactions, and protein super-families.

More powerful, more complicated to use

Effective use of these packages requires learning some more complicated controls and/or command language. In return, they have considerably more power.

- SPADE (sites.google.com/view/spade) - the Structural Proteomics Application Development Environment. SPADE provides community tools for development and deployment of essential structure and sequence equipment. Includes a chemical probing suite to support experimental verification of predicted structural models. Written in Python with scripting tools available. Runs on Linux, Windows and Mac.
- Jmol (jmol.org) - a free, open-source java-based program available in stand-alone or applet forms. It uses a superset of the RasMol/Chime command language. It is widely accepted as a replacement for Chime. The Jmol applet is used in the Proteopedia.Org wiki. Jmol is extremely powerful both for small molecules (e.g. molecular orbitals) and macromolecules (e.g. symmetry operations, unit cells, crystal contacts, translucent surfaces and cavities, arbitrary objects, animations, etc. etc.).
- BALLView ([1]) - a powerful open-source molecular modeling and visualization tool. It is available for Windows, MacOS X and Windows. BALLView provides powerful visualization capabilities for proteins, nucleic acids and small molecules. The modeling functionality includes various molecular mechanics methods (molecular dynamics, geometry optimization) using various force fields (AMBER, CHARMM, MMFF94). The functionality of BALLView can be extended and scripted through a convenient Python interface.
- PyMol - a molecular graphics system with an embedded Python interpreter designed for real-time visualization and rapid generation of high-quality molecular graphics images and animations. PyMOL is extremely powerful and is very popular with crystallographers. A large percentage of macromolecular structure figures in scientific journal articles are made with PyMOL. Although it is open source, use of PyMOL requires a modest subscription fee, except for educational use.
- RasMol - a free, open-source stand-alone program first released in 1993, remains very popular. The program and reference manual are available from rasmol.org.
- MDL Chime (www.mdl.com) - a free browser plugin released in 1996. It was the best tool for free, web-browser based visualization from 1996 until about 2004, when it was superceded by Jmol (see above). It is not open-source, and development effectively ceased before 2000. Although hundreds of tutorials and other resources remain available only in Chime in 2009, Chime is now largely of historical interest.
- Friend (ilyinlab.org/friend) - a bioinformatics application designed for simultaneous analysis and visualization of multiple structures and sequences of proteins and/or DNA/RNA. The application provides basic functionalities such as: structure visualization with different rendering and coloring, sequence alignment, and simple phylogeny analysis, along with a number of extended features to perform *more complex analyses* of sequence structure relationships, including: structural alignment of proteins, investigation of specific interaction motifs, studies of protein-protein and protein-DNA interactions, and protein super-families. Friend is also useful for the functional annotation of proteins, protein modeling, and protein

folding studies. Friend provides three levels of usage; 1) an extensive GUI for a scientist with no programming experience, 2) a command line interface for scripting for a scientist with some programming experience, and 3) the ability to extend Friend with user written libraries for an experienced programmer. The application is linked and communicates with local and remote sequence and structure databases. Friend is also now availabe in Applet form, which empowers users with all the functionality currently found in Friend, and provides a new web-based presentation platform, with detailed organization and manipulation of structure/sequence information, at the press of a button.

Additional

- Chimera - excellent molecular graphics package with support for a wide range of operations, including flexible molecular graphics, high resolution images for publication, user-driven analysis, multiple sequence alignment analysis, multiple model analysis, docking
- Garlic - a free molecular visualization program
- Ghemical - a molecular modelling package; graphical user interface is built on GTK2; both quantum-mechanical and forcefield-based methods are supported, and it is also possible to add new methods
- Oslet - a molecular modeling and simulation environment in Java, mainly for education
- Spock - a full-featured molecular graphics program
- VEGA - developed to create a bridge between most of the molecular software packages, like BioDock, Quanta/CHARMm, Insight II, MoPac, etc.
- VMD - a molecular visualization program for displaying, animating, and analyzing large biomolecular systems using 3-D graphics and built-in scripting
- UGENE - visual suite allowing visualization of PDB and MMDB proteins

RASMOL

RasMol is a computer program written for molecular graphics visualization intended and used mainly to depict and explore biological macromolecule structures, such as those found in the Protein Data Bank. It was originally developed by Roger Sayle in the early 1990s.[1]

Historically, it was an important tool for molecular biologists since the extremely optimized program allowed the software to run on (then) modestly powerful personal computers. Before RasMol, visualization software ran on graphics workstations that, due to their cost, were less accessible to scholars. RasMol continues to be important for research in structural biology, and has become important in education.

RasMol has a complex licensing version history. Starting with the version 2.7 series,[2] RasMol source code is dual-licensed under a GNU General Public License (GPL), or custom license *RASLIC*.[3] Starting with version 2.7.5, a GPL is the only license valid for binary distributions.

RasMol includes a scripting language, to perform many functions such as selecting certain protein chains, changing colors, etc. Jmol and Sirius software have incorporated this language into their commands.

Protein Data Bank (PDB) files can be downloaded for visualization from members of the Worldwide Protein Data Bank (wwPDB). These have been uploaded by researchers who have characterized the structure of molecules usually by X-ray crystallography, protein NMR spectroscopy, or cryo-electron microscopy.

RasMol Features

RasMol is a molecular graphics program intended for the visualisation of proteins, nucleic acids and small molecules. The program is aimed at display, teaching and generation of publication quality images. RasMol runs on wide range of architectures and operating systems including Microsoft Windows, Apple Macintosh, UNIX and VMS systems. UNIX and VMS versions require an 8, 24 or 32 bit colour X Windows display (X11R4 or later). The X Windows version of RasMol provides optional support for a hardware dials box and accelerated shared memory communication (via the XInput and MIT-SHM extensions) if available on the current X Server.

The program reads in a molecule coordinate file and interactively displays the molecule on the screen in a variety of colour schemes and molecule representations. Currently available representations include depth-cued wireframes, 'Dreiding' sticks, spacefilling (CPK) spheres, ball and stick, solid and strand biomolecular ribbons, atom labels and dot surfaces.

The X Windows version of RasMol provides optional support for a hardware dials box and accelerated shared memory communication (via the XInput and MIT-SHM extensions) if available on the current X Server.

The program reads in molecular coordinate files and interactively displays the molecule on the screen in a variety of representations and colour schemes. Supported input file formats include Protein Data Bank (PDB), Tripos Associates' Alchemy and Sybyl Mol2 formats, Molecular Design Limited's (MDL) Mol file format, Minnesota Supercomputer Center's (MSC) XYZ (XMol) format, CHARMm format, CIF format and mmCIF format files. If connectivity information is not contained in the file this is calculated automatically. The loaded molecule can be shown as wireframe bonds, cylinder 'Dreiding' stick bonds, alpha-carbon trace, space-filling (CPK) spheres, macromolecular ribbons (either smooth shaded solid ribbons or parallel strands), hydrogen bonding and dot surface representations. Atoms may also be labelled with arbitrary text strings. Alternate conformers and multiple NMR models may be specially coloured and identified in atom labels. Different parts of the molecule may be represented and coloured independently of the rest of the molecule or displayed in several representations simultaneously. The displayed molecule may be rotated, translated, zoomed and z-clipped (slabbed) interactively using either the mouse, the scroll bars, the command line or an attached dial box. RasMol can read a prepared list of commands from a 'script' file (or via inter-process communication) to allow a given image or viewpoint to be restored quickly. RasMol can also create a script file containing the commands required to regenerate the current image. Finally, the rendered image may be written out in a variety of formats including either raster or vector PostScript, GIF, PPM, BMP, PICT, Sun rasterfile or as a MolScript input script or Kinemage.

The RasMol help facility can be accessed by typing "help <topic>" or "help <topic> <subtopic>" from the command line. A complete list of RasMol commands may be displayed by typing "help commands". A single question mark may also be used to abbreviate the keyword "help". Please type "help notices" for important notices.

SWISSPDB Viewer

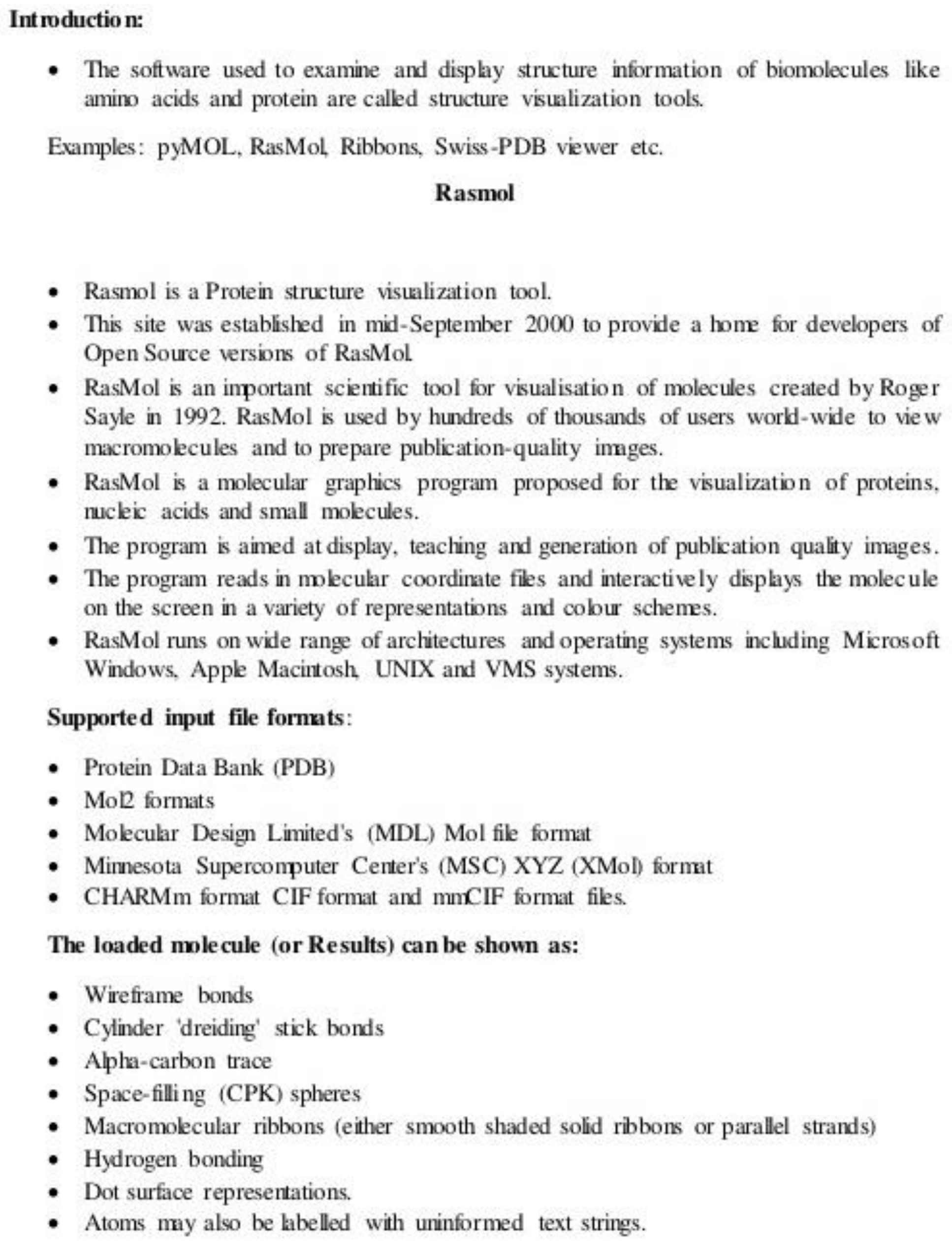
Swiss-PdbViewer (aka DeepView) is an application that provides a user friendly interface allowing to analyze several proteins at the same time. The proteins can be superimposed in order to deduce structural alignments and compare their active sites or any other relevant parts. Amino acid mutations, H-bonds, angles and distances between atoms are easy to obtain thanks to the intuitive graphic and menu interface.

Swiss-PdbViewer (aka DeepView) has been developped since 1994 by Nicolas Guex. Swiss-PdbViewer is tightly linked to SWISS-MODEL, an automated homology modeling server developed within the Swiss Institute of Bioinformatics (SIB) at the Structural Bioinformatics Group at the Biozentrum in Basel.

Working with these two programs greatly reduces the amount of work necessary to generate models, as it is possible to thread a protein primary sequence onto a 3D template and get an immediate feedback of how well the threaded protein will be accepted by the reference structure before submitting a request to build missing loops and refine sidechain packing.

Swiss-PdbViewer can also read electron density maps, and provides various tools to build into the density. In addition, various modeling tools are integrated and residues can be mutated.

Finally, as a special bonus, POV-Ray scenes can be generated from the current view in order to make stunning ray-traced quality images. An example can be found here.



Introduction:

- The software used to examine and display structure information of biomolecules like amino acids and protein are called structure visualization tools.

Examples: pyMOL, RasMol, Ribbons, Swiss-PDB viewer etc.

Rasmol

- Rasmol is a Protein structure visualization tool.
- This site was established in mid-September 2000 to provide a home for developers of Open Source versions of RasMol.
- RasMol is an important scientific tool for visualisation of molecules created by Roger Sayle in 1992. RasMol is used by hundreds of thousands of users world-wide to view macromolecules and to prepare publication-quality images.
- RasMol is a molecular graphics program proposed for the visualization of proteins, nucleic acids and small molecules.
- The program is aimed at display, teaching and generation of publication quality images.
- The program reads in molecular coordinate files and interactively displays the molecule on the screen in a variety of representations and colour schemes.
- RasMol runs on wide range of architectures and operating systems including Microsoft Windows, Apple Macintosh, UNIX and VMS systems.

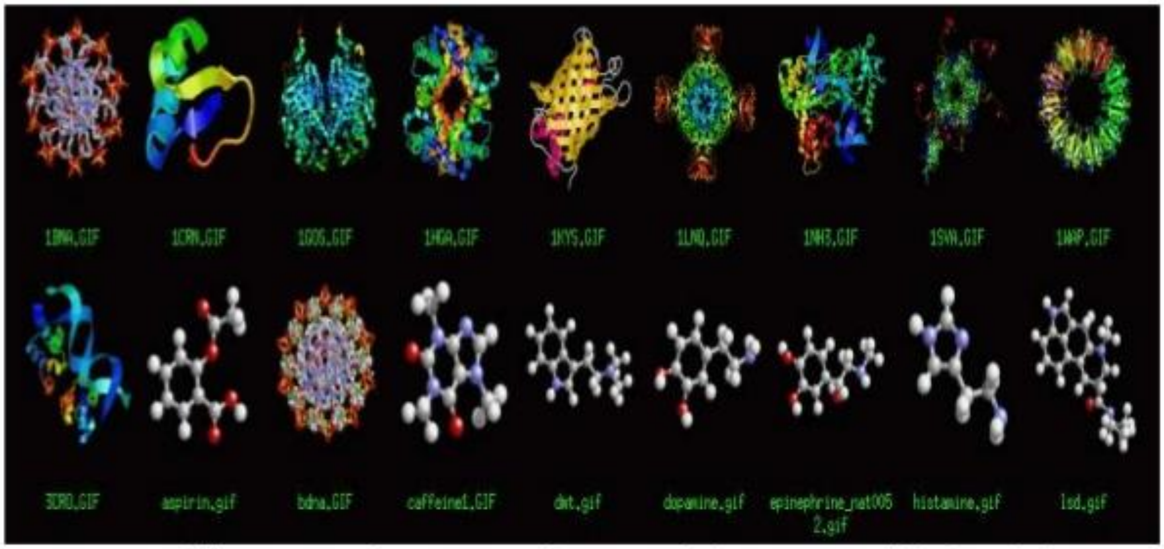
Supported input file formats:

- Protein Data Bank (PDB)
- Mol2 formats
- Molecular Design Limited's (MDL) Mol file format
- Minnesota Supercomputer Center's (MSC) XYZ (XMol) format
- CHARMm format CIF format and mmCIF format files.

The loaded molecule (or Results) can be shown as:

- Wireframe bonds
- Cylinder 'dreiding' stick bonds
- Alpha-carbon trace
- Space-filling (CPK) spheres
- Macromolecular ribbons (either smooth shaded solid ribbons or parallel strands)
- Hydrogen bonding
- Dot surface representations.
- Atoms may also be labelled with uninformed text strings.

- Alternate **conformers and multiple NMR models** may be specially coloured and identified in atom labels.
- **Different parts of the molecule may be represented and coloured independently** of the rest of the molecule or displayed in several representations simultaneously.
- The displayed molecule **may be rotated, translated, zoomed and z-clipped (slabbed) interactively using either the mouse, the scroll bars, the command line or an attached dial box.**
- RasMol can read a prepared list of **commands from a 'script' file** (or via inter-process communication) to allow a given image or viewpoint to be restored quickly.
- RasMol can also create a script file containing the commands required to regenerate the current image.
- Finally, the purified image may be written out in a variety of formats including either raster **or vector PostScript, GIF, PPM, BMP, PICT, Sun rasterfile or as a MolScript input script or Kinemage.**



- This software is freely available.Download link: http://www.openrasmol.org/#Software

RasMol Features:

- The ability to automatically mark non bonded atoms in wireframe and stick displays.
- The ability to report coordinates.

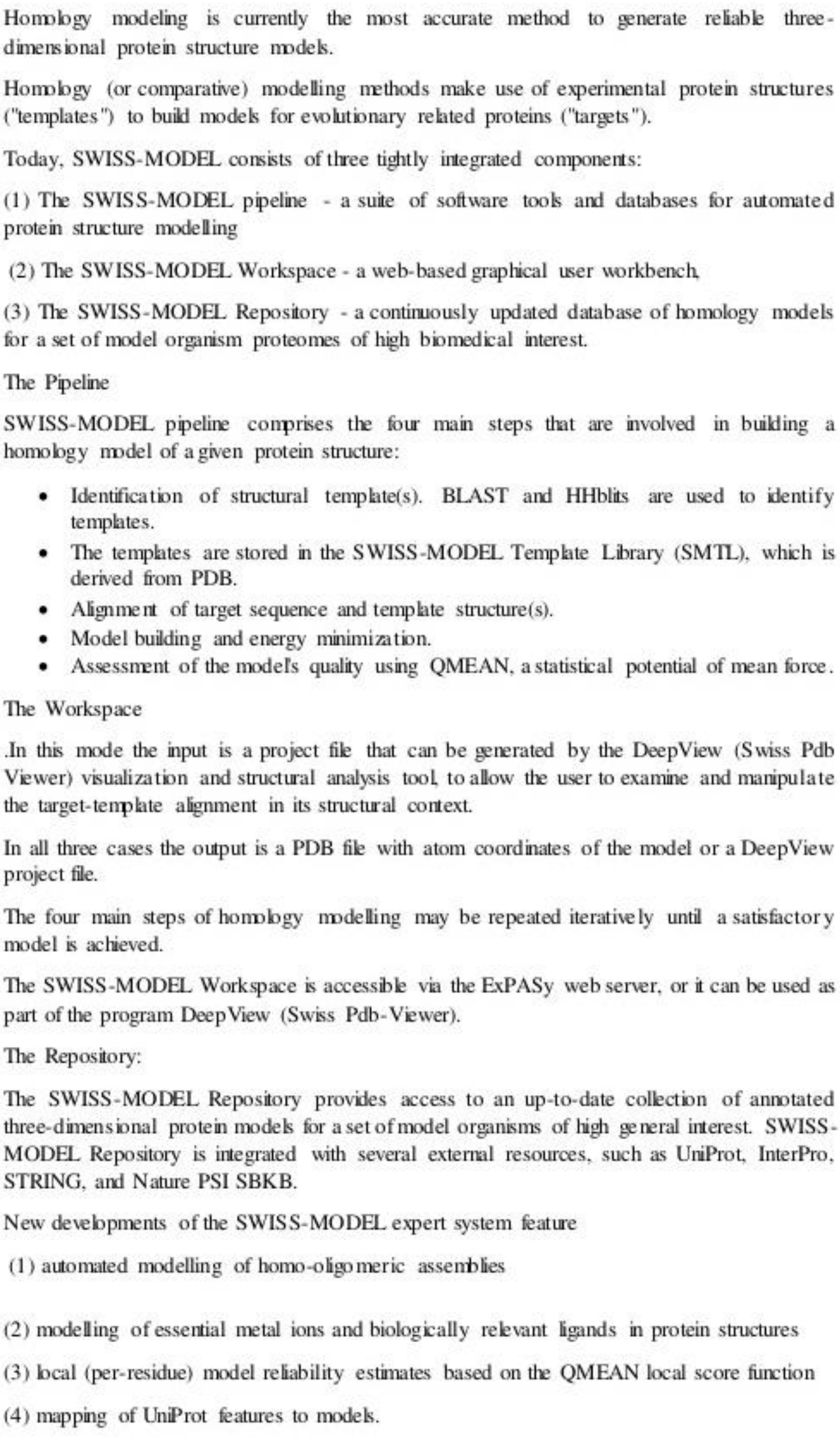
ARG465 .1HH2
3.05
ARG465 .2HH2
1.94
ARG465 .2HH1
2.09
2.70
ASP510 .OD1
3.10
ASP510 .O
2.13
G507 .H
SP510 .H
2:26
3.24
ARG507 .O
LYS508 .O

- Additions to the list of pre-defined colours.
- Improved accuracy of coordinates in pseudo-PDB output.
- Updating the picture title with the PDB ID code and EXPDTA information, so models will be clearly distinguished from experimental data.
- Introduction of a multilingual structure for RasMol.
- Population of messages and menu lists for English and Spanish.
- Correction of coordinate handling for Mol2 and XYZ coordinates
- An attempt to fix some of the chirality reversals in some of the output modes.

Swiss-PdbViewer

SWISS-MODEL is a structural bioinformatics web-server dedicated to homology modeling of protein 3D structures. Swiss-PdbViewer has been developed since 1994 by Nicolas Guex.

- Swiss-PdbViewer is tightly linked to SWISS-MODEL, an automated homology modeling server developed within the Swiss Institute of Bioinformatics (SIB) at the Structural Bioinformatics Group at the Biozentrum in Basel.
- Swiss-PdbViewer is an application that provides a user friendly interface allowing to analyze several proteins.
- The proteins can be superimposed in order to deduce structural alignments
- It is used to compare their active sites or any other relevant parts.
- Amino acid mutations, H-bonds, angles and distances between atoms are easy to visualize.
- Swiss-PdbViewer can also read electron density maps, and provides various tools to build into the density.
- Various modeling tools are integrated and residues can be mutated.



Homology modeling is currently the most accurate method to generate reliable three-dimensional protein structure models.

Homology (or comparative) modelling methods make use of experimental protein structures ("templates") to build models for evolutionary related proteins ("targets").

Today, SWISS-MODEL consists of three tightly integrated components:

(1) The SWISS-MODEL pipeline - a suite of software tools and databases for automated protein structure modelling

(2) The SWISS-MODEL Workspace - a web-based graphical user workbench,

(3) The SWISS-MODEL Repository - a continuously updated database of homology models for a set of model organism proteomes of high biomedical interest.

The Pipeline

SWISS-MODEL pipeline comprises the four main steps that are involved in building a homology model of a given protein structure:

- Identification of structural template(s). BLAST and HHblits are used to identify templates.
- The templates are stored in the SWISS-MODEL Template Library (SMTL), which is derived from PDB.
- Alignment of target sequence and template structure(s).
- Model building and energy minimization.
- Assessment of the model's quality using QMEAN, a statistical potential of mean force.

The Workspace

.In this mode the input is a project file that can be generated by the DeepView (Swiss Pdb Viewer) visualization and structural analysis tool, to allow the user to examine and manipulate the target-template alignment in its structural context.

In all three cases the output is a PDB file with atom coordinates of the model or a DeepView project file.

The four main steps of homology modelling may be repeated iteratively until a satisfactory model is achieved.

The SWISS-MODEL Workspace is accessible via the ExPASy web server, or it can be used as part of the program DeepView (Swiss Pdb-Viewer).

The Repository:

The SWISS-MODEL Repository provides access to an up-to-date collection of annotated three-dimensional protein models for a set of model organisms of high general interest. SWISS-MODEL Repository is integrated with several external resources, such as UniProt, InterPro, STRING, and Nature PSI SBKB.

New developments of the SWISS-MODEL expert system feature

(1) automated modelling of homo-oligomeric assemblies

(2) modelling of essential metal ions and biologically relevant ligands in protein structures

(3) local (per-residue) model reliability estimates based on the QMEAN local score function

(4) mapping of UniProt features to models.

Pymol

- PyMOL is an open source molecular visualization system created by Warren Lyford DeLano. It was commercialized initially by DeLano Scientific LLC, which was a private software company dedicated to creating useful tools that become universally accessible to scientific and educational communities. It is currently commercialized by Schrödinger, Inc. PyMOL can produce high-quality 3D images of small molecules and biological macromolecules, such as proteins. According to the original author, by 2009, almost a quarter of all published images of 3D protein structures in the scientific literature were made using PyMOL.
- PyMOL is one of the few open-source model visualization tools available for use in structural biology. The Py part of the software's name refers to the program having been written in the programming language Python.
- PyMOL uses OpenGL Extension Wrangler Library (GLEW) and FreeGLUT, and can solve Poisson–Boltzmann equations using the Adaptive Poisson Boltzmann Solver.[3] PyMOL used Tk for the GUI widgets and had native Aqua binaries for macOS through Schrödinger, which were replaced with a PyQt user interface on all platforms with the release of version 2.0.
- Early versions of PyMol were released under the Python License. On 1 August 2006, DeLano Scientific adopted a controlled-access download system for precompiled PyMOL builds (including betas) distributed by the company. Access to these executables is now limited to registered users who are paying customers; educational builds are available free to students and teachers. However, most of the current source code continues to be available for free, as are older precompiled builds. While the build systems for other platforms are open, the Windows API (WinAPI, Win32) build system is not, although unofficial Windows binaries are available online.[5] Anyone can either compile an executable from the Python-licensed source code or pay for a subscription to support services to obtain access to precompiled executables.
- On 8 January 2010, Schrödinger, Inc. reached an agreement to acquire PyMOL. The firm assumed development, maintenance, support, and sales of PyMOL, including all then-valid subscriptions. They also continue to actively support the PyMOL open-source community. In 2017, Schrödinger revamped the distribution system to unify the user interface under Qt and the package management under Anaconda, and released it as PyMol v2.[4] This version restricts some new functionalities and adds a watermark to the visualization if used unlicensed beyond the 30-day trial period; the overall license policy is similar to the DeLano system. The source code remains mostly available, this time under a BSD-like license.[6] As with the previous distribution, unofficial Windows binaries in the wheel format are available,[5] and indeed Linux distributions continue to provide their own builds of the open-source code.

Computers

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOINFORMATICS

UNIT – 5 - INTRODUCTION TO BIOINFORMATICS– SBIA1101

Overview of Computers

An electronic data processing device, which requires input raw data for processing and generates the output in desired form. It stores the data in its memory which can be accessed any number of times for reference from its memory. It is made up of a lot of electronics, software and mechanical parts.

A computer is divided into three basic units namely:

1. Input Unit
2. Central Processing Unit
3. Output Unit

These units are defined as below:

1) Input Unit

As the name suggests, this unit contains devices with the help of which the data is entered into the computer. This unit is a basic requirement for computer system. The input devices are of many types such as keyboard, mouse, joy stick, microphone, camera etc. Input devices give different set of input values converted into a form understandable to the computer.

2) Central Processing Unit (CPU)

Central Processing Unit (CPU) is known as the brain of the computer. It performs all types of data processing operations as required by a programmer. It stores all the data, intermediate results, and instructions as given by the programmer in the form of codes (program). Central Processing unit controls the operation of each part of the computer.

It has following three components:

1. Arithmetic Logic Unit (ALU)
2. Memory Unit
3. Control Unit

3) Output Unit

The devices with the help of which we get the information from the computer are known as the output devices. Output Unit is an interface between the computer and the user. Output devices notify the information displayed into a form which is understandable by the computer user.

Functions of a Computer

1. Data is entered into computer using Input Devices.

2. Data or Instructions are stored in the computer in its memory and processed or uses them as and when required.
3. Data is processed and converted into useful information.
4. Output is generated as per format.
5. Control Mechanism is established for controlling all the functions.

We can divide computer in Hardware and Software:

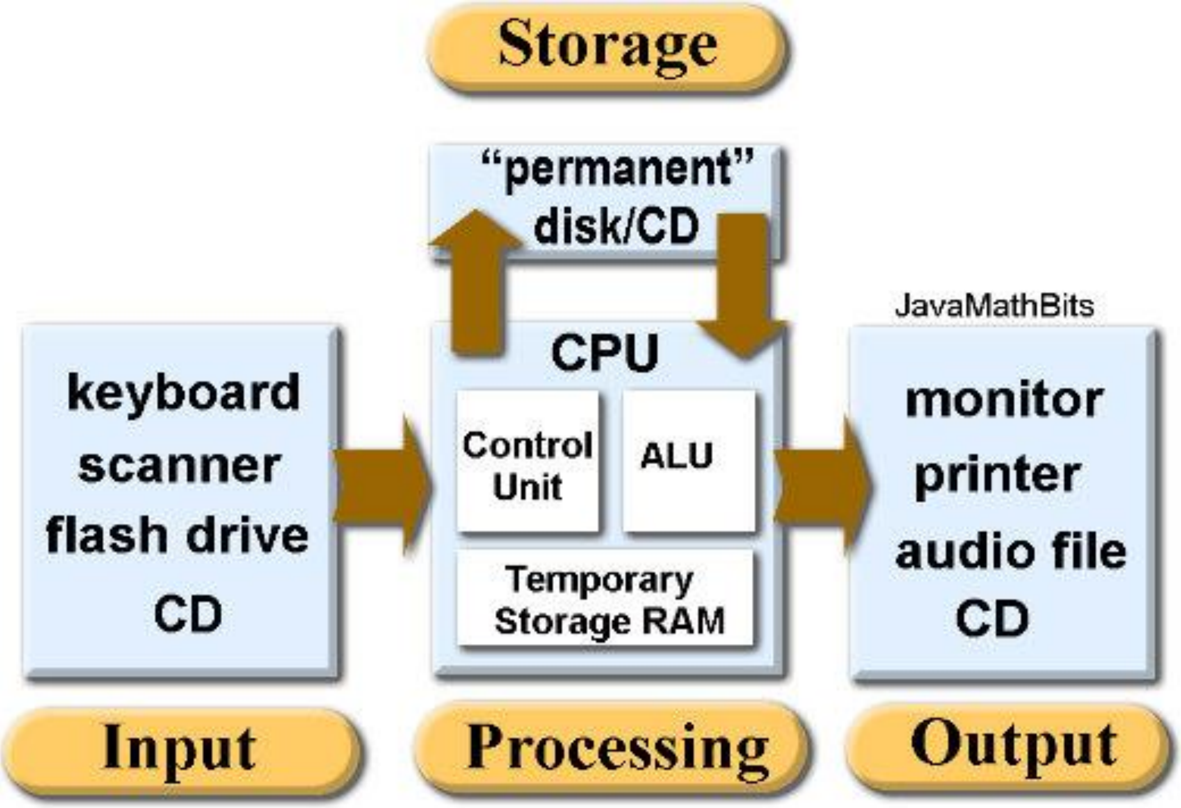
1. **Hardware:** Keyboard, mouse, joy stick, microphone, camera, printer, monitor, Hard disk, CD, DVD, CPU, motherboard, RAM etc are known as Hardware.
2. **Software:** System Software & Application Software.

Advantage of Computers

A Computer has a very High Speed of processing i.e can perform large amount of data very quickly. Computers are very accurate. Computers are very fast devices. Once the correct input is given to the computers, the output is 100% accurate. It has a large memory capacity. It can store a large amount of information for a large time. It is a reliable device.

Uses of Computers

Nowadays it is used in every walk of life. It has an important role industrial automation. Computers are playing very important role in Medical science, Engineering, General Education, Government and Private organizations, Film and Entertainment. It is at the top of making DIGITAL INDIA.



Functionalities of a Computer

If we look at it in a very broad sense, any digital computer carries out the following five functions −

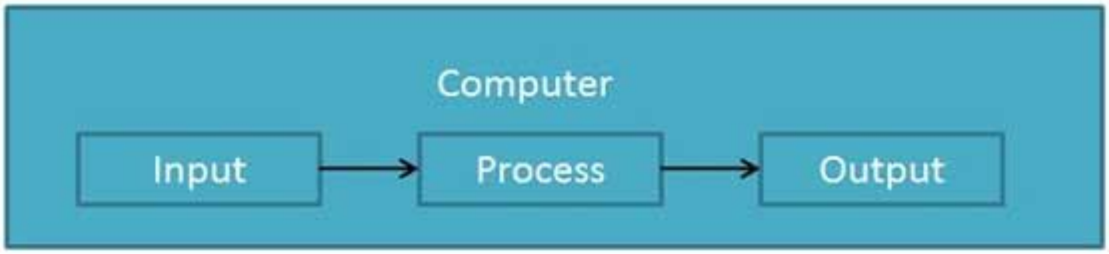
Step 1 − Takes data as input.

Step 2 − Stores the data/instructions in its memory and uses them as required.

Step 3 − Processes the data and converts it into useful information.

Step 4 − Generates the output.

Step 5 − Controls all the above four steps.



Advantages of Computers

Following are certain advantages of computers.

High Speed

- Computer is a very fast device.
- It is capable of performing calculation of very large amount of data.
- The computer has units of speed in microsecond, nanosecond, and even the picosecond.
- It can perform millions of calculations in a few seconds as compared to man who will spend many months to perform the same task.

Accuracy

- In addition to being very fast, computers are very accurate.
- The calculations are 100% error free.
- Computers perform all jobs with 100% accuracy provided that the input is correct.

Storage Capability

- Memory is a very important characteristic of computers.
- A computer has much more storage capacity than human beings.
- It can store large amount of data.
- It can store any type of data such as images, videos, text, audio, etc.

Diligence

- Unlike human beings, a computer is free from monotony, tiredness, and lack of concentration.
- It can work continuously without any error and boredom.
- It can perform repeated tasks with the same speed and accuracy.

Versatility

- A computer is a very versatile machine.
- A computer is very flexible in performing the jobs to be done.
- This machine can be used to solve the problems related to various fields.
- At one instance, it may be solving a complex scientific problem and the very next moment it may be playing a card game.

Reliability

- A computer is a reliable machine.

- Modern electronic components have long lives.
- Computers are designed to make maintenance easy.

Automation

- Computer is an automatic machine.
- Automation is the ability to perform a given task automatically. Once the computer receives a program i.e., the program is stored in the computer memory, then the program and instruction can control the program execution without human interaction.

Reduction in Paper Work and Cost

- The use of computers for data processing in an organization leads to reduction in paper work and results in speeding up the process.
- As data in electronic files can be retrieved as and when required, the problem of maintenance of large number of paper files gets reduced.
- Though the initial investment for installing a computer is high, it substantially reduces the cost of each of its transaction.

Disadvantages of Computers

Following are certain disadvantages of computers.

No I.Q.

- A computer is a machine that has no intelligence to perform any task.
- Each instruction has to be given to the computer.
- A computer cannot take any decision on its own.

Dependency

- It functions as per the user's instruction, thus it is fully dependent on humans.

Environment

- The operating environment of the computer should be dust free and suitable.

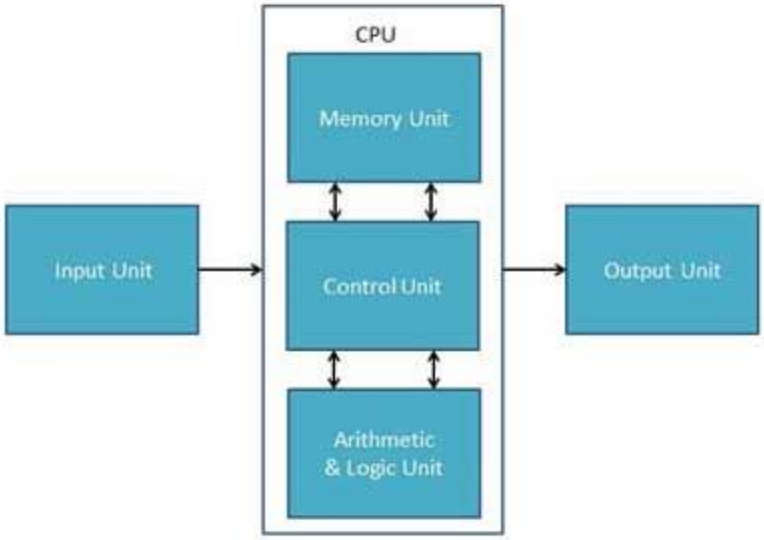
No Feeling

- Computers have no feelings or emotions.
- It cannot make judgment based on feeling, taste, experience, and knowledge unlike humans.
- Computers can be broadly classified by their speed and computing power.

S.No.	Type	Specifications
1	PC (Personal Computer)	It is a single user computer system having moderately powerful microprocessor
2	Workstation	It is also a single user computer system, similar to personal computer however has a more powerful microprocessor.
3	Mini Computer	It is a multi-user computer system, capable of supporting hundreds of users simultaneously.
4	Main Frame	It is a multi-user computer system, capable of supporting hundreds of users simultaneously. Software technology is different from minicomputer.
5	Supercomputer	It is an extremely fast computer, which can execute hundreds of millions of instructions per second.

All types of computers follow the same basic logical structure and perform the following five basic operations for converting raw input data into information useful to their users.

S.No.	Operation	Description
1	Take Input	The process of entering data and instructions into the computer system.
2	Store Data	Saving data and instructions so that they are available for processing as and when required.
3	Processing Data	Performing arithmetic, and logical operations on data in order to convert them into useful information.
4	Output Information	The process of producing useful information or results for the user, such as a printed report or visual display.
5	Control the workflow	Directs the manner and sequence in which all of the above operations are performed.



Input Unit

This unit contains devices with the help of which we enter data into the computer. This unit creates a link between the user and the computer. The input devices translate the information into a form understandable by the computer.

CPU (Central Processing Unit)

CPU is considered as the brain of the computer. CPU performs all types of data processing operations. It stores data, intermediate results, and instructions (program). It controls the operation of all parts of the computer.

CPU itself has the following three components −

- ALU (Arithmetic Logic Unit)
- Memory Unit
- Control Unit

Output Unit

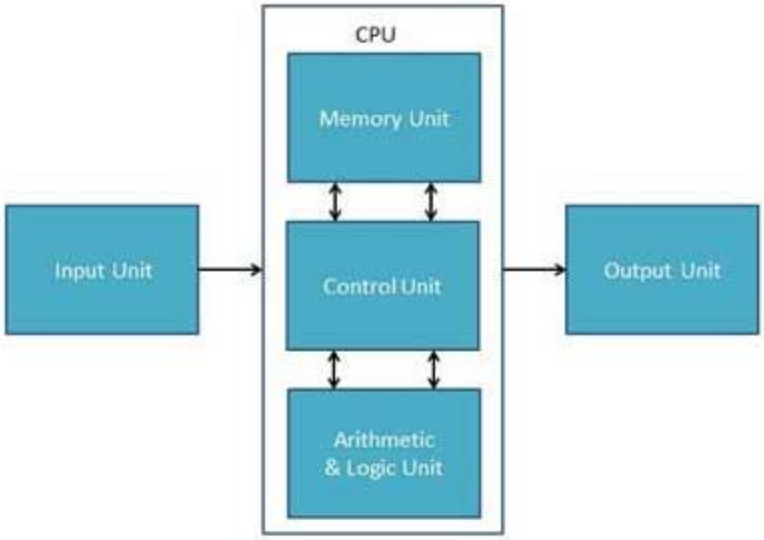
The output unit consists of devices with the help of which we get the information from the computer. This unit is a link between the computer and the users. Output devices translate the computer's output into a form understandable by the users.

Central Processing Unit (CPU) consists of the following features −

- CPU is considered as the brain of the computer.
- CPU performs all types of data processing operations.
- It stores data, intermediate results, and instructions (program).
- It controls the operation of all parts of the computer.

CPU itself has following three components.

- Memory or Storage Unit
- Control Unit
- ALU(Arithmetic Logic Unit)



Memory or Storage Unit

This unit can store instructions, data, and intermediate results. This unit supplies information to other units of the computer when needed. It is also known as internal storage unit or the main memory or the primary storage or Random Access Memory (RAM).

Its size affects speed, power, and capability. Primary memory and secondary memory are two types of memories in the computer. Functions of the memory unit are −

- It stores all the data and the instructions required for processing.
- It stores intermediate results of processing.
- It stores the final results of processing before these results are released to an output device.
- All inputs and outputs are transmitted through the main memory.

Control Unit

This unit controls the operations of all parts of the computer but does not carry out any actual data processing operations.

Functions of this unit are −

- It is responsible for controlling the transfer of data and instructions among other units of a computer.
- It manages and coordinates all the units of the computer.
- It obtains the instructions from the memory, interprets them, and directs the operation of the computer.
- It communicates with Input/Output devices for transfer of data or results from storage.
- It does not process or store data.

ALU (Arithmetic Logic Unit)

This unit consists of two subsections namely,

- Arithmetic Section
- Logic Section

Arithmetic Section

Function of arithmetic section is to perform arithmetic operations like addition, subtraction, multiplication, and division. All complex operations are done by making repetitive use of the above operations.

Logic Section

Function of logic section is to perform logic operations such as comparing, selecting, matching, and merging of data.

Following are some of the important input devices which are used in a computer −

- Keyboard
- Mouse
- Joy Stick
- Light pen
- Track Ball
- Scanner
- Graphic Tablet
- Microphone
- Magnetic Ink Card Reader(MICR)

- Optical Character Reader(OCR)
- Bar Code Reader
- Optical Mark Reader(OMR)

Keyboard

Keyboard is the most common and very popular input device which helps to input data to the computer. The layout of the keyboard is like that of traditional typewriter, although there are some additional keys provided for performing additional functions.

Keyboards are of two sizes 84 keys or 101/102 keys, but now keyboards with 104 keys or 108 keys are also available for Windows and Internet.

The keys on the keyboard are as follows −

S.No	Keys & Description
1	**Typing Keys** These keys include the letter keys (A-Z) and digit keys (09) which generally give the same layout as that of typewriters.
2	**Numeric Keypad** It is used to enter the numeric data or cursor movement. Generally, it consists of a set of 17 keys that are laid out in the same configuration used by most adding machines and calculators.
3	**Function Keys** The twelve function keys are present on the keyboard which are arranged in a row at the top of the keyboard. Each function key has a unique meaning and is used for some specific purpose.
4	**Control keys**

	These keys provide cursor and screen control. It includes four directional arrow keys. Control keys also include Home, End, Insert, Delete, Page Up, Page Down, Control(Ctrl), Alternate(Alt), Escape(Esc).
5	**Special Purpose Keys** Keyboard also contains some special purpose keys such as Enter, Shift, Caps Lock, Num Lock, Space bar, Tab, and Print Screen.

Mouse

Mouse is the most popular pointing device. It is a very famous cursor-control device having a small palm size box with a round ball at its base, which senses the movement of the mouse and sends corresponding signals to the CPU when the mouse buttons are pressed.

Generally, it has two buttons called the left and the right button and a wheel is present between the buttons. A mouse can be used to control the position of the cursor on the screen, but it cannot be used to enter text into the computer.

Advantages

- Easy to use
- Not very expensive
- Moves the cursor faster than the arrow keys of the keyboard.

Joystick

Joystick is also a pointing device, which is used to move the cursor position on a monitor screen. It is a stick having a spherical ball at its both lower and upper ends. The lower spherical ball moves in a socket. The joystick can be moved in all four directions.

The function of the joystick is similar to that of a mouse. It is mainly used in Computer Aided Designing (CAD) and playing computer games.

Light Pen

Light pen is a pointing device similar to a pen. It is used to select a displayed menu item or draw pictures on the monitor screen. It consists of a photocell and an optical system placed in a small tube.

When the tip of a light pen is moved over the monitor screen and the pen button is pressed, its photocell sensing element detects the screen location and sends the corresponding signal to the CPU.

Track Ball

Track ball is an input device that is mostly used in notebook or laptop computer, instead of a mouse. This is a ball which is half inserted and by moving fingers on the ball, the pointer can be moved.

Since the whole device is not moved, a track ball requires less space than a mouse. A track ball comes in various shapes like a ball, a button, or a square.

Scanner

Scanner is an input device, which works more like a photocopy machine. It is used when some information is available on paper and it is to be transferred to the hard disk of the computer for further manipulation.

Scanner captures images from the source which are then converted into a digital form that can be stored on the disk. These images can be edited before they are printed.

Digitizer

Digitizer is an input device which converts analog information into digital form. Digitizer can convert a signal from the television or camera into a series of numbers that could be stored in a computer. They can be used by the computer to create a picture of whatever the camera had been pointed at.

Digitizer is also known as Tablet or Graphics Tablet as it converts graphics and pictorial data into binary inputs. A graphic tablet as digitizer is used for fine works of drawing and image manipulation applications.

Microphone

Microphone is an input device to input sound that is then stored in a digital form.

The microphone is used for various applications such as adding sound to a multimedia presentation or for mixing music.

Magnetic Ink Card Reader (MICR)

MICR input device is generally used in banks as there are large number of cheques to be processed every day. The bank's code number and cheque number are printed on the cheques with a special type of ink that contains particles of magnetic material that are machine readable.

This reading process is called Magnetic Ink Character Recognition (MICR). The main advantages of MICR is that it is fast and less error prone.

Optical Character Reader (OCR)

OCR is an input device used to read a printed text.

OCR scans the text optically, character by character, converts them into a machine readable code, and stores the text on the system memory.

Bar Code Readers

Bar Code Reader is a device used for reading bar coded data (data in the form of light and dark lines). Bar coded data is generally used in labelling goods, numbering the books, etc. It may be a handheld scanner or may be embedded in a stationary scanner.

Bar Code Reader scans a bar code image, converts it into an alphanumeric value, which is then fed to the computer that the bar code reader is connected to.

Optical Mark Reader (OMR)

OMR is a special type of optical scanner used to recognize the type of mark made by pen or pencil. It is used where one out of a few alternatives is to be selected and marked.

It is specially used for checking the answer sheets of examinations having multiple choice questions.

Following are some of the important output devices used in a computer.

- Monitors
- Graphic Plotter
- Printer

Monitors

Monitors, commonly called as **Visual Display Unit** (VDU), are the main output device of a computer. It forms images from tiny dots, called pixels that are arranged in a rectangular form. The sharpness of the image depends upon the number of pixels.

There are two kinds of viewing screen used for monitors.

- Cathode-Ray Tube (CRT)
- Flat-Panel Display

Cathode-Ray Tube (CRT) Monitor

The CRT display is made up of small picture elements called pixels. The smaller the pixels, the better the image clarity or resolution. It takes more than one illuminated pixel to form a whole character, such as the letter 'e' in the word help.

A finite number of characters can be displayed on a screen at once. The screen can be divided into a series of character boxes - fixed location on the screen where a standard character can be placed. Most screens are capable of displaying 80 characters of data horizontally and 25 lines vertically.

There are some disadvantages of CRT −

- Large in Size
- High power consumption

Flat-Panel Display Monitor

The flat-panel display refers to a class of video devices that have reduced volume, weight and power requirement in comparison to the CRT. You can hang them on walls or wear them on your wrists. Current uses of flat-panel displays include calculators, video games, monitors, laptop computer, and graphics display.

The flat-panel display is divided into two categories −

- **Emissive Displays** − Emissive displays are devices that convert electrical energy into light. For example, plasma panel and LED (Light-Emitting Diodes).
- **Non-Emissive Displays** − Non-emissive displays use optical effects to convert sunlight or light from some other source into graphics patterns. For example, LCD (Liquid-Crystal Device).

Printers

Printer is an output device, which is used to print information on paper.

There are two types of printers −

- Impact Printers
- Non-Impact Printers

Impact Printers

Impact printers print the characters by striking them on the ribbon, which is then pressed on the paper.

Characteristics of Impact Printers are the following −

- Very low consumable costs
- Very noisy
- Useful for bulk printing due to low cost
- There is physical contact with the paper to produce an image

These printers are of two types −

- Character printers

- Line printers

Character Printers

Character printers are the printers which print one character at a time.

These are further divided into two types:

- Dot Matrix Printer(DMP)
- Daisy Wheel

Dot Matrix Printer

In the market, one of the most popular printers is Dot Matrix Printer. These printers are popular because of their ease of printing and economical price. Each character printed is in the form of pattern of dots and head consists of a Matrix of Pins of size (5*7, 7*9, 9*7 or 9*9) which come out to form a character which is why it is called Dot Matrix Printer.

Advantages

- Inexpensive
- Widely Used
- Other language characters can be printed

Disadvantages

- Slow Speed
- Poor Quality

Daisy Wheel

Head is lying on a wheel and pins corresponding to characters are like petals of Daisy (flower) which is why it is called Daisy Wheel Printer. These printers are generally used for word-processing in offices that require a few letters to be sent here and there with very nice quality.

Advantages

- More reliable than DMP
- Better quality
- Fonts of character can be easily changed

Disadvantages

- Slower than DMP
- Noisy
- More expensive than DMP

Line Printers

Line printers are the printers which print one line at a time.

These are of two types −

- Drum Printer
- Chain Printer

Drum Printer

This printer is like a drum in shape hence it is called drum printer. The surface of the drum is divided into a number of tracks. Total tracks are equal to the size of the paper, i.e. for a paper width of 132 characters, drum will have 132 tracks. A character set is embossed on the track. Different character sets available in the market are 48 character set, 64 and 96 characters set. One rotation of drum prints one line. Drum printers are fast in speed and can print 300 to 2000 lines per minute.

Advantages

- Very high speed

Disadvantages

- Very expensive
- Characters fonts cannot be changed

Chain Printer

In this printer, a chain of character sets is used, hence it is called Chain Printer. A standard character set may have 48, 64, or 96 characters.

Advantages

- Character fonts can easily be changed.
- Different languages can be used with the same printer.

Disadvantages

- Noisy

Non-impact Printers

Non-impact printers print the characters without using the ribbon. These printers print a complete page at a time, thus they are also called as Page Printers.

These printers are of two types −

- Laser Printers
- Inkjet Printers

Characteristics of Non-impact Printers

- Faster than impact printers
- They are not noisy
- High quality
- Supports many fonts and different character size

Laser Printers

These are non-impact page printers. They use laser lights to produce the dots needed to form the characters to be printed on a page.

Advantages

- Very high speed
- Very high quality output
- Good graphics quality
- Supports many fonts and different character size

Disadvantages

- Expensive
- Cannot be used to produce multiple copies of a document in a single printing

Inkjet Printers

Inkjet printers are non-impact character printers based on a relatively new technology. They print characters by spraying small drops of ink onto paper. Inkjet printers produce high quality output with presentable features.

They make less noise because no hammering is done and these have many styles of printing modes available. Color printing is also possible. Some models of Inkjet printers can produce multiple copies of printing also.

Advantages

- High quality printing
- More reliable

Disadvantages

- Expensive as the cost per page is high
- Slow as compared to laser printer

A memory is just like a human brain. It is used to store data and instructions. Computer memory is the storage space in the computer, where data is to be processed and instructions required for processing are stored. The memory is divided into large number of small parts called cells. Each location or cell has a unique address, which varies from zero to memory size minus one. For example, if the computer has 64k words, then this memory unit has 64 * 1024 = 65536 memory locations. The address of these locations varies from 0 to 65535.

Memory is primarily of three types −

- Cache Memory
- Primary Memory/Main Memory
- Secondary Memory

Cache Memory

Cache memory is a very high speed semiconductor memory which can speed up the CPU. It acts as a buffer between the CPU and the main memory. It is used to hold those parts of data and program which are most frequently used by the CPU. The parts of data and programs are transferred from the disk to cache memory by the operating system, from where the CPU can access them.

Advantages

The advantages of cache memory are as follows −

- Cache memory is faster than main memory.
- It consumes less access time as compared to main memory.
- It stores the program that can be executed within a short period of time.
- It stores data for temporary use.

Disadvantages

The disadvantages of cache memory are as follows −

- Cache memory has limited capacity.
- It is very expensive.

Primary Memory (Main Memory)

Primary memory holds only those data and instructions on which the computer is currently working. It has a limited capacity and data is lost when power is switched off. It is generally made up of semiconductor device. These memories are not as fast as registers. The data and instruction required to be processed resides in the main memory. It is divided into two subcategories RAM and ROM.

Characteristics of Main Memory

- These are semiconductor memories.
- It is known as the main memory.
- Usually volatile memory.
- Data is lost in case power is switched off.
- It is the working memory of the computer.
- Faster than secondary memories.
- A computer cannot run without the primary memory.

Secondary Memory

This type of memory is also known as external memory or non-volatile. It is slower than the main memory. These are used for storing data/information permanently. CPU directly does not access these memories, instead they are accessed via input-output routines. The contents of secondary memories are first transferred to the main memory, and then the CPU can access it. For example, disk, CD-ROM, DVD, etc.

Characteristics of Secondary Memory

- These are magnetic and optical memories.
- It is known as the backup memory.
- It is a non-volatile memory.
- Data is permanently stored even if power is switched off.
- It is used for storage of data in a computer.
- Computer may run without the secondary memory.
- Slower than primary memories.

RAM (Random Access Memory) is the internal memory of the CPU for storing data, program, and program result. It is a read/write memory which stores data until the machine is working. As soon as the machine is switched off, data is erased.

Access time in RAM is independent of the address, that is, each storage location inside the memory is as easy to reach as other locations and takes the same amount of time. Data in the RAM can be accessed randomly but it is very expensive.

RAM is volatile, i.e. data stored in it is lost when we switch off the computer or if there is a power failure. Hence, a backup Uninterruptible Power System (UPS) is often used with computers. RAM is small, both in terms of its physical size and in the amount of data it can hold.

RAM is of two types −

- Static RAM (SRAM)
- Dynamic RAM (DRAM)

Static RAM (SRAM)

The word **static** indicates that the memory retains its contents as long as power is being supplied. However, data is lost when the power gets down due to volatile nature. SRAM chips use a matrix of 6-transistors and no capacitors. Transistors do not require power to prevent leakage, so SRAM need not be refreshed on a regular basis.

There is extra space in the matrix, hence SRAM uses more chips than DRAM for the same amount of storage space, making the manufacturing costs higher. SRAM is thus used as cache memory and has very fast access.

Characteristic of Static RAM

- Long life
- No need to refresh
- Faster
- Used as cache memory
- Large size
- Expensive
- High power consumption

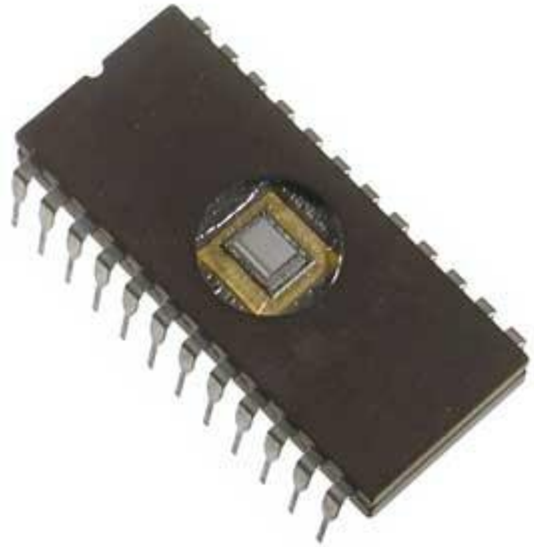
Dynamic RAM (DRAM)

DRAM, unlike SRAM, must be continually **refreshed** in order to maintain the data. This is done by placing the memory on a refresh circuit that rewrites the data several hundred times per second. DRAM is used for most system memory as it is cheap and small. All DRAMs are made up of memory cells, which are composed of one capacitor and one transistor.

Characteristics of Dynamic RAM

- Short data lifetime
- Needs to be refreshed continuously
- Slower as compared to SRAM
- Used as RAM
- Smaller in size
- Less expensive
- Less power consumption

ROM stands for **Read Only Memory**. The memory from which we can only read but cannot write on it. This type of memory is non-volatile. The information is stored permanently in such memories during manufacture. A ROM stores such instructions that are required to start a computer. This operation is referred to as **bootstrap**. ROM chips are not only used in the computer but also in other electronic items like washing machine and microwave oven.



Let us now discuss the various types of ROMs and their characteristics.

MROM (Masked ROM)

The very first ROMs were hard-wired devices that contained a pre-programmed set of data or instructions. These kind of ROMs are known as masked ROMs, which are inexpensive.

PROM (Programmable Read Only Memory)

PROM is read-only memory that can be modified only once by a user. The user buys a blank PROM and enters the desired contents using a PROM program. Inside the PROM chip, there are small fuses which are burnt open during programming. It can be programmed only once and is not erasable.

EPROM (Erasable and Programmable Read Only Memory)

EPROM can be erased by exposing it to ultra-violet light for a duration of up to 40 minutes. Usually, an EPROM eraser achieves this function. During programming, an electrical charge is trapped in an insulated gate region. The charge is retained for more than 10 years because the charge has no leakage path. For erasing this charge, ultra-violet light is passed through a quartz crystal window (lid). This exposure to ultra-violet light dissipates the charge. During normal use, the quartz lid is sealed with a sticker.

EEPROM (Electrically Erasable and Programmable Read Only Memory)

EEPROM is programmed and erased electrically. It can be erased and reprogrammed about ten thousand times. Both erasing and programming take about 4 to 10 ms (millisecond). In EEPROM, any location can be selectively erased and programmed. EEPROMs can be erased one byte at a time, rather than erasing the entire chip. Hence, the process of reprogramming is flexible but slow.

Advantages of ROM

The advantages of ROM are as follows −

- Non-volatile in nature
- Cannot be accidentally changed
- Cheaper than RAMs
- Easy to test
- More reliable than RAMs
- Static and do not require refreshing
- Contents are always known and can be verified

Hardware represents the physical and tangible components of a computer, i.e. the components that can be seen and touched.

Examples of Hardware are the following −

- **Input devices** − keyboard, mouse, etc.
- **Output devices** − printer, monitor, etc.
- **Secondary storage devices** − Hard disk, CD, DVD, etc.
- **Internal components** − CPU, motherboard, RAM, etc.

Relationship between Hardware and Software

- Hardware and software are mutually dependent on each other. Both of them must work together to make a computer produce a useful output.
- Software cannot be utilized without supporting hardware.
- Hardware without a set of programs to operate upon cannot be utilized and is useless.
- To get a particular job done on the computer, relevant software should be loaded into the hardware.
- Hardware is a one-time expense.

- Software development is very expensive and is a continuing expense.
- Different software applications can be loaded on a hardware to run different jobs.
- A software acts as an interface between the user and the hardware.
- If the hardware is the 'heart' of a computer system, then the software is its 'soul'. Both are complementary to each other.

Software is a set of programs, which is designed to perform a well-defined function. A program is a sequence of instructions written to solve a particular problem.

There are two types of software −

- System Software
- Application Software

System Software

The system software is a collection of programs designed to operate, control, and extend the processing capabilities of the computer itself. System software is generally prepared by the computer manufacturers. These software products comprise of programs written in low-level languages, which interact with the hardware at a very basic level. System software serves as the interface between the hardware and the end users.

Some examples of system software are Operating System, Compilers, Interpreter, Assemblers, etc.

Here is a list of some of the most prominent features of a system software −

- Close to the system
- Fast in speed
- Difficult to design
- Difficult to understand
- Less interactive
- Smaller in size
- Difficult to manipulate
- Generally written in low-level language

Application Software

Application software products are designed to satisfy a particular need of a particular environment. All software applications prepared in the computer lab can come under the category of Application software.

Application software may consist of a single program, such as Microsoft's notepad for writing and editing a simple text. It may also consist of a collection of programs, often called a software package, which work together to accomplish a task, such as a spreadsheet package.

Examples of Application software are the following −

- Payroll Software
- Student Record Software
- Inventory Management Software
- Income Tax Software
- Railways Reservation Software
- Microsoft Office Suite Software
- Microsoft Word
- Microsoft Excel
- Microsoft PowerPoint

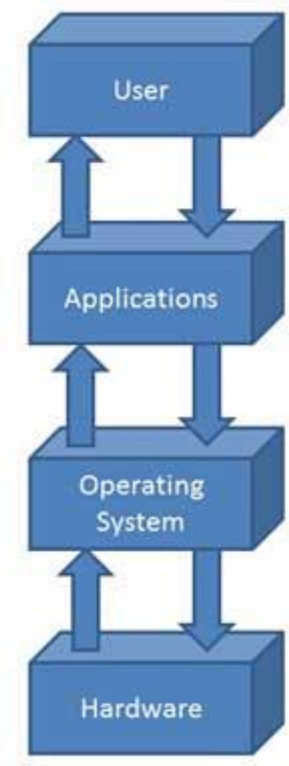
Features of application software are as follows −

- Close to the user
- Easy to design
- More interactive
- Slow in speed
- Generally written in high-level language
- Easy to understand
- Easy to manipulate and use
- Bigger in size and requires large storage space

The Operating System is a program with the following features −

- An operating system is a program that acts as an interface between the software and the computer hardware.
- It is an integrated set of specialized programs used to manage overall resources and operations of the computer.

- It is a specialized software that controls and monitors the execution of all other programs that reside in the computer, including application programs and other system software.



Objectives of Operating System

The objectives of the operating system are −

- To make the computer system convenient to use in an efficient manner.
- To hide the details of the hardware resources from the users.
- To provide users a convenient interface to use the computer system.
- To act as an intermediary between the hardware and its users, making it easier for the users to access and use other resources.
- To manage the resources of a computer system.
- To keep track of who is using which resource, granting resource requests, and mediating conflicting requests from different programs and users.
- To provide efficient and fair sharing of resources among users and programs.

Characteristics of Operating System

Here is a list of some of the most prominent characteristic features of Operating Systems −

- **Memory Management** − Keeps track of the primary memory, i.e. what part of it is in use by whom, what part is not in use, etc. and allocates the memory when a process or program requests it.

- **Processor Management** − Allocates the processor (CPU) to a process and deallocates the processor when it is no longer required.
- **Device Management** − Keeps track of all the devices. This is also called I/O controller that decides which process gets the device, when, and for how much time.
- **File Management** − Allocates and de-allocates the resources and decides who gets the resources.
- **Security** − Prevents unauthorized access to programs and data by means of passwords and other similar techniques.
- **Job Accounting** − Keeps track of time and resources used by various jobs and/or users.
- **Control Over System Performance** − Records delays between the request for a service and from the system.
- **Interaction with the Operators** − Interaction may take place via the console of the computer in the form of instructions. The Operating System acknowledges the same, does the corresponding action, and informs the operation by a display screen.
- **Error-detecting Aids** − Production of dumps, traces, error messages, and other debugging and error-detecting methods.
- **Coordination Between Other Software and Users** − Coordination and assignment of compilers, interpreters, assemblers, and other software to the various users of the computer systems.

Internet

It is a worldwide/global system of interconnected computer networks. It uses the standard Internet Protocol (TCP/IP). Every computer in Internet is identified by a unique IP address. IP Address is a unique set of numbers (such as 110.22.33.114) which identifies a computer's location.

A special computer DNS (Domain Name Server) is used to provide a name to the IP Address so that the user can locate a computer by a name. For example, a DNS server will resolve a name https://www.tutorialspoint.com to a particular IP address to uniquely identify the computer on which this website is hosted.

Internet is accessible to every user all over the world.

Intranet

Intranet is the system in which multiple PCs are connected to each other. PCs in intranet are not available to the world outside the intranet. Usually each organization has its own Intranet network and members/employees of that organization can access the computers in their intranet.

Each computer in Intranet is also identified by an IP Address which is unique among the computers in that Intranet.

Similarities between Internet and Intranet

- Intranet uses the internet protocols such as TCP/IP and FTP.

- Intranet sites are accessible via the web browser in a similar way as websites in the internet. However, only members of Intranet network can access intranet hosted sites.
- In Intranet, own instant messengers can be used as similar to yahoo messenger/gtalk over the internet.

Differences between Internet and Intranet

- Internet is general to PCs all over the world whereas Intranet is specific to few PCs.
- Internet provides a wider and better access to websites to a large population, whereas Intranet is restricted.
- Internet is not as safe as Intranet. Intranet can be safely privatized as per the need.