Module 1: Unit of Life

Text (6 minutes)

Cells as the Basic Unit of Life

A cell is the smallest unit of a living thing and is the basic building block of all organisms.

Cells as Building Blocks

A cell is the smallest unit of a living thing. A living thing, whether made of one cell (like bacteria) or many cells (like a human), is called an organism. Thus, cells are the basic building blocks of all organisms. Several cells of one kind that interconnect with each other and perform a shared function form tissues; several tissues combine to form an organ (your stomach, heart, or brain); and several organs make up an organ system (such as the digestive system, circulatory system, or nervous system). Several systems that function together form an organism (like a human being). There are many types of cells all grouped into one of two broad categories: prokaryotic and eukaryotic. For example, both animal and plant cells are classified as eukaryotic cells, whereas bacterial cells are classified as prokaryotic.

In general, the classification of cells is associated with the presence or absence of a nucleus, the biggest and at one time until the dawn of the electron microscopy age, the only visible organelle found exclusively in eukaryotic cells. Although the nucleus or "karyon" is the major identifiable characteristic of eukaryotic cells, simple possession of this organelle is not the standalone attribute setting it apart from prokaryotic cells. Organelles are membrane-bound compartments optimized for a function so a cellular business can be more efficiently conducted.

Macromolecules are made up of basic molecular units. They include the proteins (polymers of amino acids), nucleic acids (polymers of nucleotides), carbohydrates (polymers of sugars) and lipids (with a variety of modular constituents). The biosynthesis and degradation of biological macromolecules involves linear polymerization, breakdown steps (proteins, nucleic acids and lipids) and may also involve branching/debranching (carbohydrates). These processes may involve multi-protein complexes (e.g. ribosome, proteasome) with complex regulation.

Homeostasis: A property of cells, tissues, and organisms that allows the maintenance and regulation of the stability and constancy needed to function properly. Homeostasis is a healthy state that is maintained by the constant adjustment of biochemical and physiological pathways

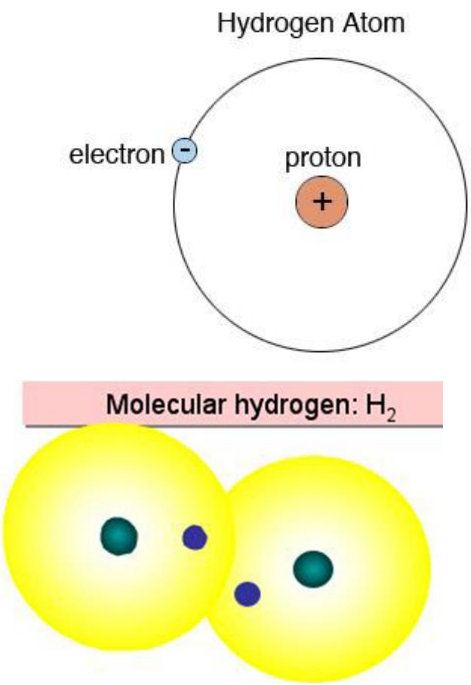
Water is needed to maintain homeostasis. Cells are also made up of macromolecules—nucleic acids, lipids, proteins, carbohydrates. These macromolecules help maintain a cell's structure, help cells communicate with each other, aid in energy storage, etc.

Module 2: Composition of matter

Text (6 minutes)

Matter is made of combinations of elements substances such as hydrogen or carbon that cannot be broken down or converted into other substances by chemical means. The smallest particle of an element that still retains its distinctive chemical properties is an atom. However, the characteristics of substances other than pure elements—including the materials from which living cells are made—depend on the way their atoms are linked together in groups to form molecules. In order to understand how living organisms are built from inanimate matter, therefore, it is crucial to know how all of the chemical bonds that hold atoms together in molecules are formed.

Each atom has at its center a positively charged nucleus, which is surrounded at some distance by a cloud of negatively charged electrons, held in a series of orbitals by electrostatic attraction to the nucleus. The nucleus in turn consists of two kinds of subatomic particles: protons, which are positively charged, and neutrons. which are electrically neutral. The number of protons in the atomic nucleus gives the atomic number. An atom of hydrogen has a nucleus composed of a single proton; so hydrogen, with an atomic number of 1, is the lightest element.



Electronegativity is the property of an atom which increases with its tendency to attract the electrons of a bond. If two bonded atoms have the same electronegativity values as each other, they share electrons equally in a covalent bond. Usually, the electrons in a chemical bond are more attracted to one atom (the more electronegative one) than to the other. If the electronegativity values are very different, the electrons aren't shared at all. One atom essentially takes the bond electrons from the other atom, forming an **ionic bond**

Electronegativity Example

The chlorine atom has a higher electronegativity than the hydrogen atom, so the bonding electrons will be closer to the Cl than to the H in the HCl molecule

In the O2 molecule, both atoms have the same electronegativity. The electrons in the covalent bond are shared equally between the two oxygen atoms.

Module 3: Molecules of life

Text (4 minutes)

Living organisms are composed of several types of substances called biomolecules. According to their molecular weight, substances in living organisms are divided into two groups:

1. Low molecular substances (Mr < 10 000)

- Water
- Inorganic (mineral) substances
- Intermediates of metabolic pathways (carboxylic acids etc.)
- Final products of metabolic pathways (amino acids, monosaccharides, lipids, nucleotides)

2. High molecular substances (Mr > 10 000)

- Proteins
- Polysaccharides
- Nucleic acids

High molecular substances, which are present in living organisms, are also named as biological **macromolecules or biopolymers.** The building units of proteins are amino acids, the building units of polysaccharides are monosaccharides, and the building units of nucleic acids are nucleotides.

According to their origin, the substances included in the living organisms are divided into inorganic substances (water, carbon dioxide, mineral substances) and organic substances (the most important are nucleic acids, proteins, saccharides, lipids).

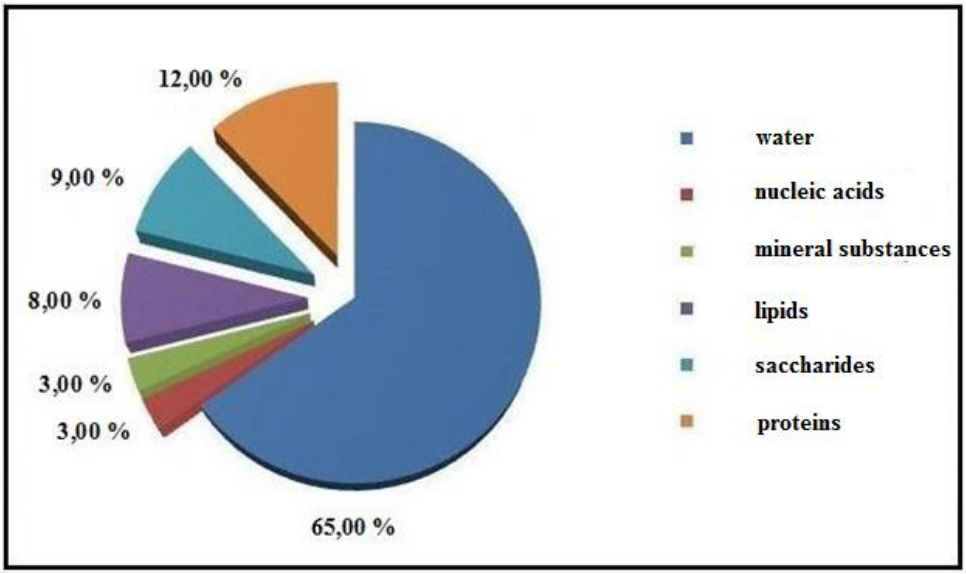


Fig.: Average representation of the main groups of substances in organisms

Condensation is a chemical process by which 2 molecules are joined together to make a larger, more complex, molecule, with the loss of water.

It is the basis for the synthesis of all the important biological macromolecules (carbohydrates, proteins, lipids, nucleic acids) from their simpler sub-units.

In all cases of condensation, molecules with projecting -H atoms are linked to other molecules with projecting -OH groups, producing H2O, (H.OH) also known as water, which then moves away from the original molecules.

A-H + B-OH --> A-B + H2O

Hydrolysis is the opposite to condensation. A large molecule is split into smaller sections by breaking a bond, adding -H to one section and -OH to the other.

The products are simpler substances. Since it involves the addition of water, this explains why it is called hydrolysis, meaning splitting by water.

A-B + H2O --> A-H + B-OH

Module 4: Journey into the cell

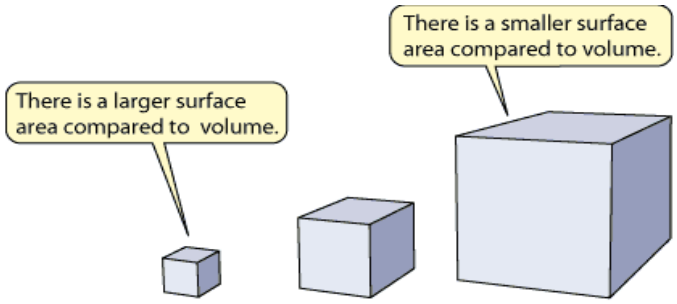
Text (7 minutes)

Module 5; Size Matters

Text (8 minutes)

Cell Structure:

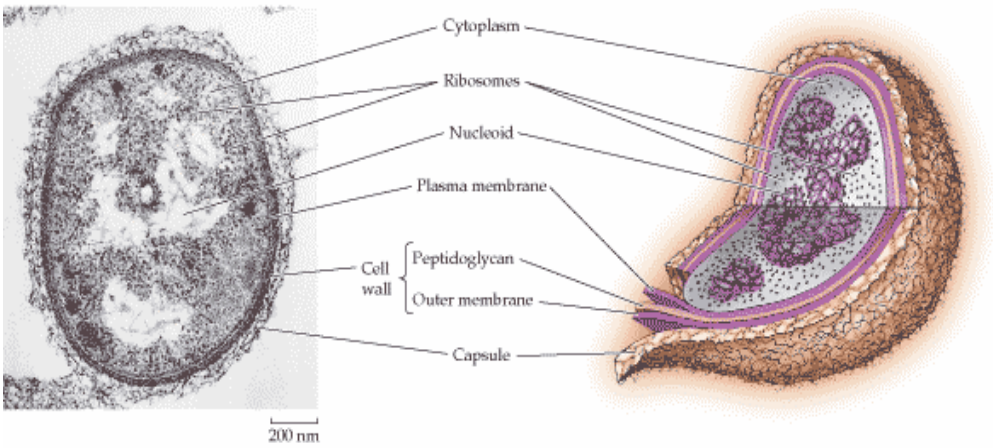
Cell are very small to maintain the large surface area to volume ratio. A smaller cell is greatly powerful and having more transporting materials; including waste products than a larger cell.



	1-mm cube	2-mm cube	4-mm cube
Surface area	6 sides × 1^2 = 6 mm^2	6 sides × 2^2 = 24 mm^2	6 sides × 4^2 = 96 mm^2
Volume	1^3 = 1 mm^3	2^3 = 8 mm^3	4^3 = 64 mm^3
Surface area-to-volume ratio	6/1	3/1	1.5/1

<u>Prokaryotic cells:</u>

- No membrane enclosed internal compartments.
- Plasma membrane regulates traffic (barrier).
- Nucleoid region contains DNA.
- Most have cell wall.



Special Features of prokaryotic cells;

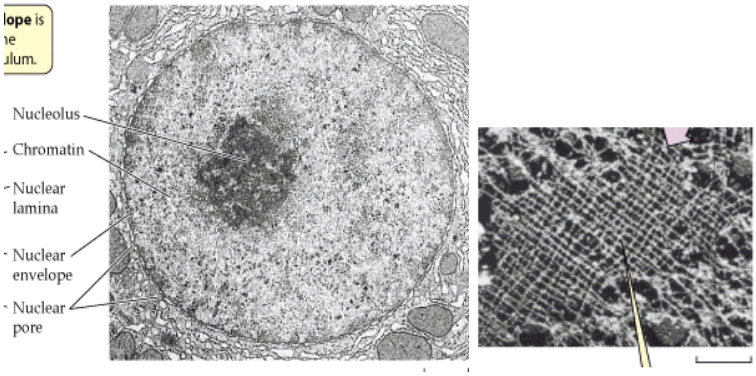
- Cyanobacteria Chlorophyll containing have folds of plasma membrane, other have mesosomes (energy).
- Some have actin like filaments and other have Flagella made-up of Flagellin.

Module 6: The Nucleus

Text (7 minutes)

Nucleus:

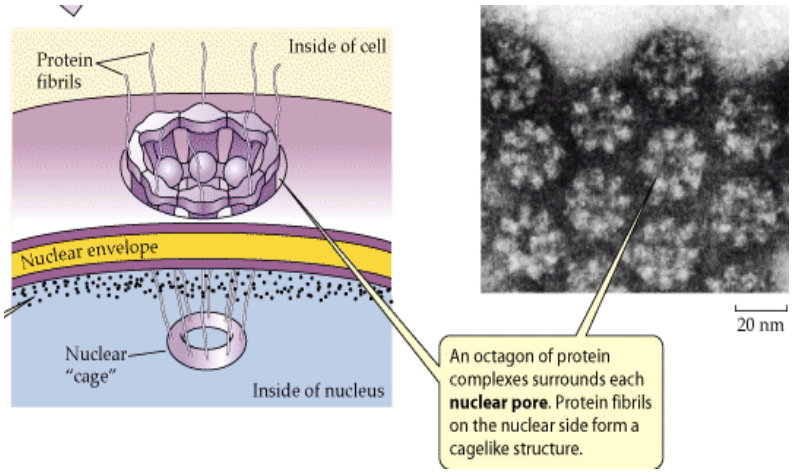
The nucleus contains most of the cell's genetic material (DNA). The duplication of the genetic material and the first steps in decoding genetic information take place in the nucleus.



The single nucleus is usually the largest organelle in a cell The nucleus of most animal cells is approximately 5 µm in diameter—substantially larger than most entire prokaryotic cells. The nucleus has several roles in the cell: The nucleus is the site of DNA duplication. The nucleus is the site of genetic control of the cell's activities. A region within the nucleus, the nucleolus, begins the assembly of ribosomes from specific proteins and RNA.

The nucleus is surrounded by two membranes, which together form the nuclear envelope. The two membranes of the nuclear envelope are separated by 10–20 nm and are perforated by nuclear pores approximately 9 nm in diameter, which connect the interior of the nucleus with the cytoplasm. At these pores, the outer membrane of the nuclear envelope is continuous with the inner membrane.Each pore is surrounded by a pore complex made up

of eight large protein granules arranged in an octagon where the inner and outer membranes merge. RNA and proteins pass through these pores to enter or leave the nucleus.



Module 7: Introduction to Molecular Biology

Text (6 minutes)

Introduction

Molecular Biology is the study of biological molecules related to genes, gene products and heredity. In the present age, world is in the midst of two scientific revolutions. One is information technology and the other is Molecular Biology. Both deal with the handling of large amounts of information. Molecular Biology has revolutionized the biological sciences as well especially in the fields of Health Sciences and Agricultural Sciences.

Contribution of Molecular Biology:

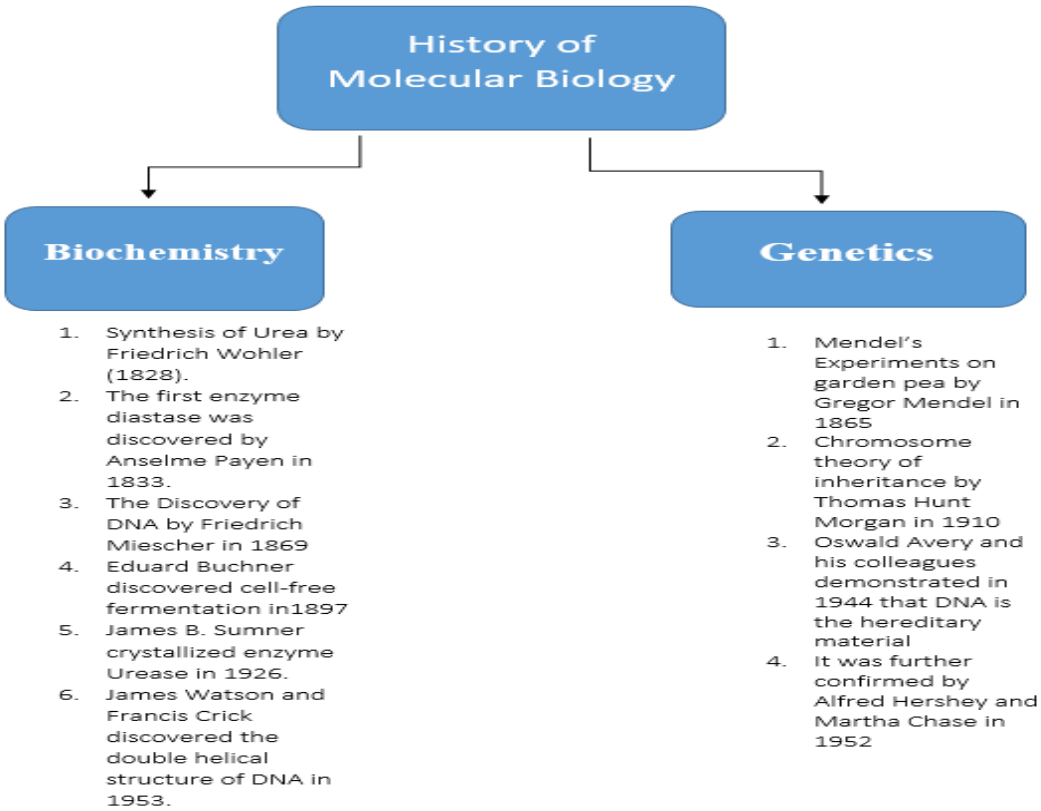
The almost complete sequence of the DNA molecules comprising the human genome was revealed in the year 2003. So, in theory, science has made available all of the genetic information needed to make a human being. However, the function of most of a human's approximately 35,000 genes remains a mystery.

The other main arena where molecular biology has a massive impact is agriculture. New varieties of genetically engineered plants and animals have already been made and some are in agricultural use. So, you can well imagine that how much important is this subject for you and for the economy of Pakistan.

Module 8: History of Molecular Biology

Text (6 minutes)

Molecular Biology is a molecular mechanism that perform various cellular functions, the advances in molecular biology is very closely related to the new technology development. The work of molecular biology is done by many scientists, SO the history of molecular biology depends on the work of scientists and their experiment.
A list of scientists work are given below;



Conclusion:

It is the study of biological phenomena. Molecular biologists concluded that how the molecules interact to one another in living organisms that performs the important functions of life. molecular biologist studies how molecules interact with one another in living organisms to perform the functions of life.

Module: 9 Achievements of molecular biology

Text (7minutes)

Achievements of Molecular Biology

1. In 1957, Francis Crick laid out the "Central dogma of molecular biology" which foretold the relationship between DNA, RNA, and proteins
2. In 1958, Mathew Meselson & Franklin Stahl proved that DNA replication was semi-conservative
3. Marshall Nirenberg and Go bind Khorana working independently cracked the code in the early 1960s
4. The Human Genome Project (HGP) was launched in 1990 and completed in 2003

Module 10: Nucleic Acid

Text (7 minutes)

Nucleotide

Nucleic acids are important group of biomolecules which are responsible for storage & transmission of hereditary

information. Like proteins and polysaccharides, nucleic acids are also polymeric compounds.

The repeating units in the nucleic acids are Nucleotides. There are two main types of nucleic acids, Deoxyribonucleic acids (DNA) and ribonucleic acids (RNA)

	DNA	RNA
Pentose sugar	Deoxyribose	Ribose
Base Composition	Adenine (A) Guanine (G) Cytosine (C) **Thymine (T)**	Adenine (A) Guanine (G) Cytosine (C) **Uracil (U)**
Number of strands	Double stranded (forms a double helix)	Single stranded

Module11: Chemical composition of DNA

Text (6 minutes)

The chemical structure of DNA

Deoxyribonucleotides is an organic chemical that give instructions and genetic information about the synthesis of protein.

DNA is a polymer of Deoxyribonucleotides. It is composed of three components. Deoxyribose, Nitrogenous Base, Phosphoric acid.

DNA has three parts such as; a phosphate group, a sugar group and one to four types of nitrogen bases. DNA is also composed of chemical building blocks that is called Nucleotides. In DNA strand the nucleotides are linked into chain.

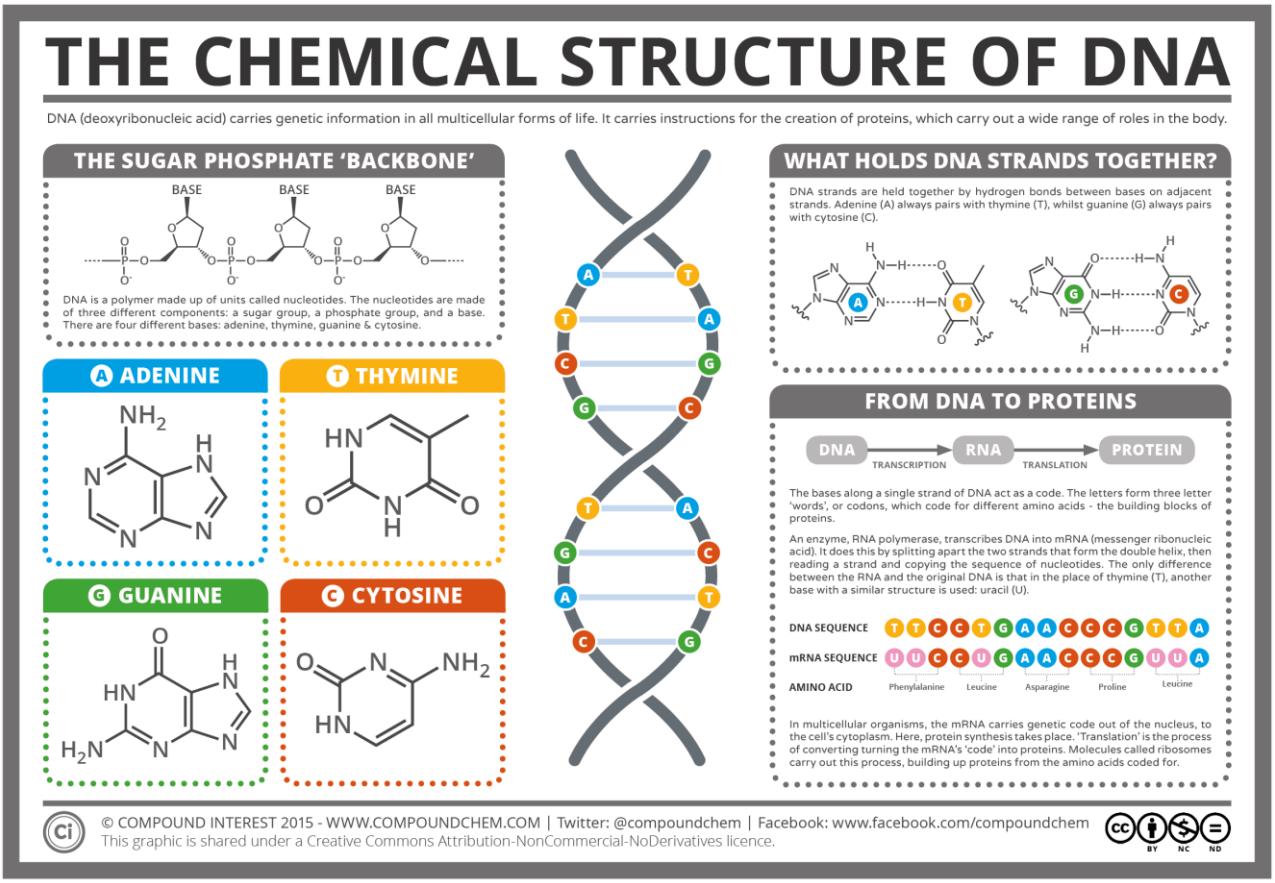
There are four bases in DNA such as

- Adenine
- Guanine
- Cytosine
- Thymine

These bases forms pairs (Adenine A with thymine T) and (Guanine G with Cytosine C)

The chemical DNA was first discovered in 1869, but its role in genetic inheritance was not demonstrated until 1943. In 1953James Watson and Francis Crick determined that the structure of DNA is a double-helix polymer, a spiral consisting of two DNA strands wound around each other. Each strand is composed of a long chain of monomer nucleotides. The nucleotide of DNA consists of a deoxyribose sugar molecule which is attached a phosphate group and one of four nitrogenous

bases: two purines (adenine and guanine) and two pyrimidines (cytosine and thymine). The nucleotides are joined together by covalent bonds between the phosphate of one nucleotide and the sugar of the next, forming a phosphate-sugar backbone from which the nitrogenous bases protrude. One strand is held to another by hydrogen bonds between the bases; the sequencing of this bonding is specific—i.e., adenine bonds only with thymine, and cytosine only with guanine.



Module:12 Nucleoside and Nucleotide

Text (7 minutes)

Nucleotide:

A molecule that contains phosphate group, pentose sugar and nitrogenous bases is called nucleotide.

Nucleoside: A molecule that contain pentose sugar and nitrogenous bases but lack phosphate group that's called nucleoside.

Nucleoside	Nucleotide
It is a combination of base and sugar.	It is a combination of nucleoside and phosphoric acid.
Examples	**Examples**
Adenosine = Adenine + Ribose	Adenylic acid = Adenosine + Phosphoric acid
Guanosine = Guanine + Ribose	Guanylic acid = Guanosine + Phosphoric acid
Cytidine = Cytosine + Ribose	Cytidylic acid = Cytidine + Phosphoric acid
Deoxythymidine = Thymine + Deoxyribose	Uridylic acid = Uridine + Phosphoric acid

Nucleoside play important role in the metabolism, macromolecule biosynthesis of and cell signaling. Nucleoside also help in transmitting, encoding and expressing genetic information in living organism.

Conclusion:

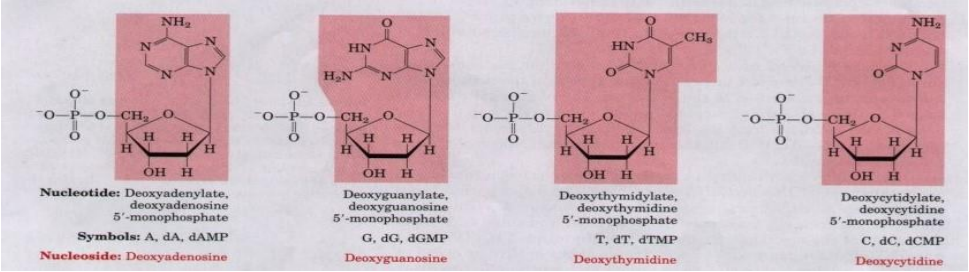
Many nucleosides and nucleotides inhibits the enzyme reverse transcriptase that control the replication of retroviruses and most importantly human immunodeficiency virus.

Module:13 Types of Deoxyribonucleotides

Text (7 minutes)

Types of Deoxyribonucleotides

There are four types of Deoxyribonucleotides such as dCTP (Deoxycytidine Triphosphate), dATP (deoxyadenosine Triphosphate), dGTP (deoxyguanine triphosphate) and dTTP (deoxythymine triphosphate).



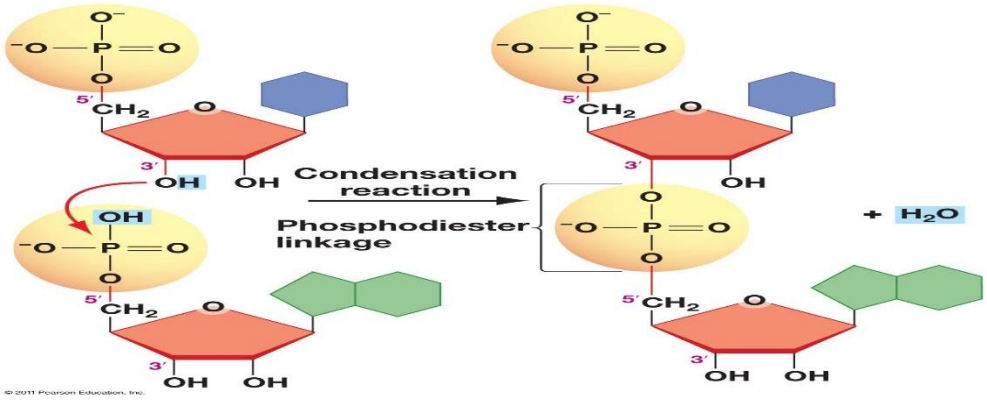
Module 14: How Deoxyribunucleotides join?

Text (7 minutes)

Deoxyribonucleotide joining:

Nucleotides are joined by bond that name as covalent bond, the covalent bond between phosphate group of one nucleotide and carbon atom 3 of pentose sugar in the next nucleotide that called

phospho diester bond.



That produces alternating backbone of sugar – phosphate – sugar-phosphate all the polynucleotide chain.

A deoxyribonucleotide is the monomer, or single unit, of DNA, or deoxyribonucleic acid. Each deoxyribonucleotide comprises three parts: a nitrogenous base, a deoxyribose sugar, and one phosphate group. The nitrogenous base is always bonded to the 1' carbon of the deoxyribose, which is distinguished from ribose by the presence of a proton on the 2' carbon rather than an OH group. The phosphate groups bind to the 5' carbon of the sugar. When deoxyribonucleotides polymerize to form DNA, the phosphate group from one nucleotide will bond to the 3' carbon on another nucleotide, forming a phosphodiester bond via dehydration synthesis. New nucleotides are always added to the 3' carbon of the last nucleotide, so synthesis always proceeds from 5' to 3'.

Module:15 Structure of DNA

Text (7 minutes)

DNA Structure

Nucleic Acids

Nucleic acids are biopolymers, or large biomolecules, essential for all known forms of life. Nucleic acids, which include DNA (deoxyribonucleic acid) and RNA (ribonucleic acid), are made from monomers known as nucleotides. Each nucleotide has three components: a 5-carbon sugar, a phosphate group, and a nitrogenous base. If the sugar is deoxyribose, the polymer is DNA. If the sugar is ribose, the polymer is RNA. When all three components are combined, they form a nucleotide. Nucleotides are also known as phosphate nucleotides.

Nucleic acids are among the most important biological macromolecules (others being amino acids/proteins, sugars/carbohydrates, and lipids/fats). They are found in abundance in all living things, where they function in encoding, transmitting and expressing genetic information in other words, information is conveyed through the nucleic acid sequence, or the order of nucleotides within a DNA or RNA molecule. Strings of nucleotides strung together in a specific sequence are the mechanism for storing and transmitting hereditary, or genetic information via protein synthesis.

Nucleic acids were discovered by Friedrich Miescher in 1869.

Deoxyribonucleic acid

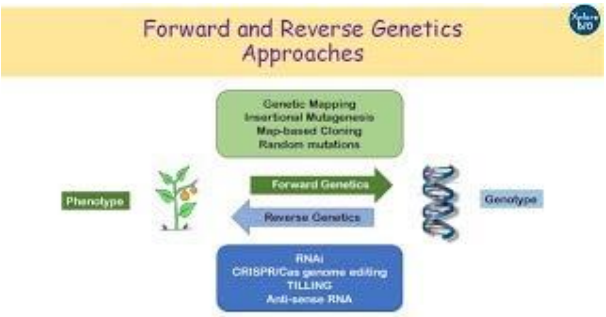
Deoxyribonucleic acid (DNA) is a nucleic acid containing the genetic instructions used in the development and functioning of all known living organisms. The DNA segments carrying this genetic information are called genes. Likewise, other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information. Along with RNA and proteins, DNA is one of the three major macromolecules that are essential for all known forms of life. DNA consists of two long polymers of simple units called nucleotides, with backbones made of sugars and phosphate groups joined by ester bonds. These two strands run in opposite directions to each other and are, therefore, anti-parallel. Attached to each sugar is one of four types of molecules called nucleobases (informally, bases). It is the sequence of these four nucleobases along the backbone that encodes information. This information is read using the genetic code, which specifies the sequence of the amino acids within proteins. The code is read by copying stretches of DNA into the related nucleic acid RNA in a process called transcription. Within cells DNA is organized into long structures called chromosomes

Module 16: What is genetics?

Text (7 minutes)

Definition:

- Genetic is the study of genes, heredity and variation
- Field of biology.
- The principals of heredity
- Mandal unaware chromosomes, gene

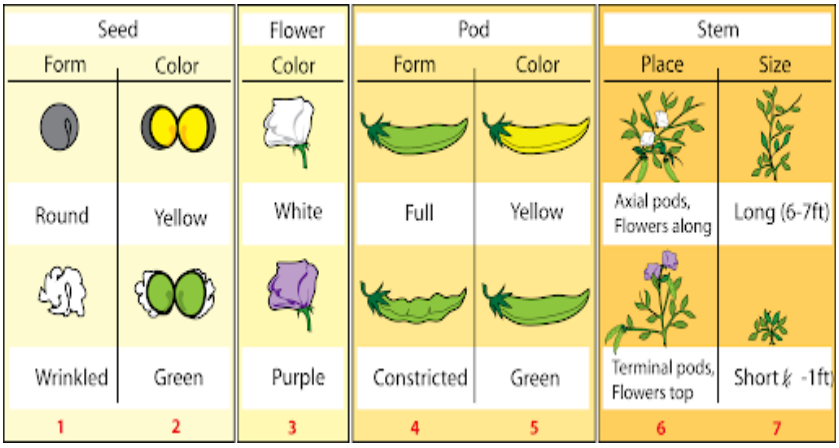


Garden Pea:

- Seeds in a variety of shapes and colors.
- Self and cross pollinate
- Takes up little space
- Short generation time
- And produce more offspring

Mendel was fortunate:

- Peas in many varieties
- Strict over which plant mated
- The pea traits are distinct and contrasting



Conclusion:

Study of genes, chromosomes and heredity s called Genetics.

Module17: Sub disciplines of genetics

Text (11:00)

Sub disciplines:

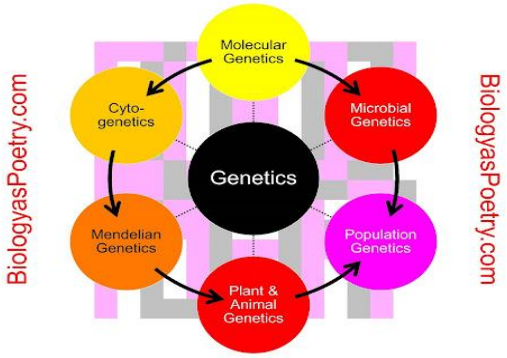
There are four sub disciplines of genetics;

1-Transmission genetics

2- Population genetics

3-Quantitative genetics

4-Molecular genetics



Development in genetics:

Historically transmission genetics developed first by followed by population genetics, quantitative and finally molecular genetics.

Transmission or classical genetics:

- Deals with movement of genes and genetic traits from parents to offspring
- Deals with genetic recombination

Population genetics:

- Study traits in a group of population
- Study heredity in groups for traits determined by one or few genes

Quantitative genetics:

- Studies group hereditary for traits determined by many genes simultaneously
- Skin color, height and eye color

Molecular genetics:

Deals with molecular structure and function of genes

Module 18: Genetic terminologies

Text (9:00)

Common Genetics Terminologies:

What is Character:

A heritable feature (skin color, height etc.).

What is Trait: variant for a character (i.e. brown, black, white etc.).

What is True-breed: all offspring of same variety.

✓ Different generations of a cross can be P generation (parents) F1 generation (1st filial generation) F2 generation (2nd filial generation)

✓ Pure Cross: A cross between a true breed plant/animal with another true breeds plant/animal is called pure cross True Breeding X True breeding WW X ww

✓ Hybrid Cross: F1 generation X F1 generation Ww X Ww

✓ Genotype and Phenotype: Genetic make-up of an organism is called Genotype while physical appearance of an organism is called Phenotype.

✓ Dominant and Recessive: when one characteristic expresses itself over the other i.e. round over wrinkled was dominant in Gregor Mendel experiments while the trait that does not show through in the first generation is called as recessive trait i.e. wrinkled.

Module 19: Genome informatics

Text (9)

Genome Informatics:

Genome Informatics (also Geno informatics) is the field of study of information processing and flow in genomes

More than 20,000 genes are there in the human genome. Comparing to the annotation of genes, how their expression is regulated are largely unknown. Moreover, identification of these regulatory regions in the genome seems to be very important for molecular medicine because mutations in these regions might be responsible for many diseases.

Today, more than 40 genomic sequences of various vertebrates are available, and comparative genome analyses are necessary to understand the changes in genomic structures and to identify functional regions. The main focus is to get most insights into the regulation of gene expression, which may cause some diseases, by using bioinformatics means. The current research topics are

(1) identification of cis-regulatory elements for transcription and splicing,

(2) comparative genome analyses to understand gene duplications and genome rearrangements,

(3) genome informatics analysis of sex differences, and

(4) gene expression in cancer tissues. With the progress of high-throughput analyses, such as microarrays and next-generation sequence technologies, interpretation of the data is not possible without computational analysis

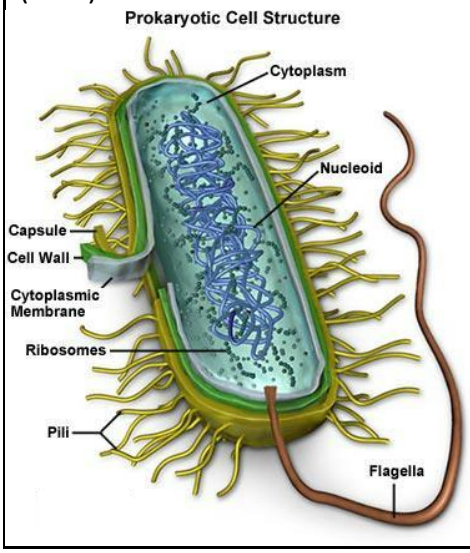
Massive data analysis is not only for bioinformaticians, but the researchers working at the bench are also required to master to use some basic tools and databases. To support those researchers, genome sequence analyses and the high-throughput data analyses are required.

An interactive process between experimental molecular biologists and bioinformaticians is necessary to fully facilitate genome data and high-throughput data. Such process includes feedback-loop between hypothesis making and experimental verification.

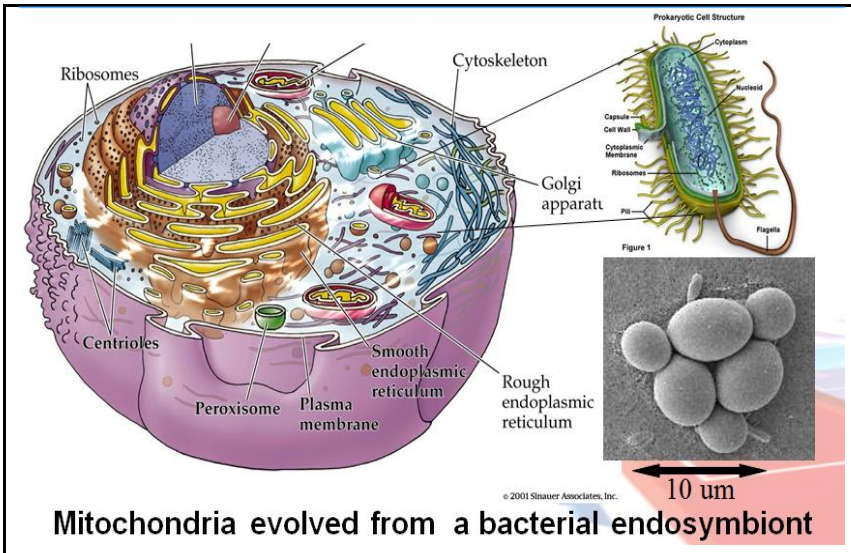
Module 20: Prokaryotic genome

Text (10)

Now, we study the ***prokaryotic genome***, prokaryotes are the organisms whose Genetic material (DNA) is not enclosed in a nuclear membrane, so there is no nucleus in them. As there is no nucleus in prokaryotes, there is no justification to have other membrane bound organelles. These are relatively simple cells.



Here, in this diagram we see a prokaryotic cell which is a bacterium (here). We have a genome (DNA) in the shape of a big chromosome in the middle, and ribosomes (small structures important for protein synthesis that occurs in every other organism so ribosomes can also be seen here). It's relatively simple cell, having cell wall with different layers.



Here, in this diagram we see a comparison between a eukaryotic cell and a prokaryotic cell. We can clearly see the membrane bounded organelles in the eukaryotic cell, like mitochondria (involved with the respiration process; food is broken down into the energy. There is a hypothesis that mitochondria actually evolved from bacteria and is known as endosymbiont hypothesis). Here we can also see the difference in the size of both cells, so eukaryotic cells are complex and bigger than prokaryotes.

The first prokaryotic genome sequenced was that of *Hemophilic influenza* (we have seen in the previous section) and this organism was sequenced in a relatively moderate cost and with an efficient pace that paved the way for sequencing of many other organisms. So study of those prokaryotic organisms is important.

Hemophilia's a bacterium with genome size of 1.83 Mbp (1743 protein encoding genes) and is a human pathogen. *Mycoplasma* is another bacterium with genome size of 0.82 Mbp (676 protein encoding genes) and is also a human pathogen that grown inside cells; metabolically weak.

Conclusions:

We conclude that:

- Prokaryotes are simple Genomes.
- They are easy models to study Biochemistry, physiology and Molecular biology of life processes.

  Sequencing is done on economically important organisms (i.e. first it's implemented on simpler genome which is then used to explore complex genomes).

Module21: Eukaryotic genome

Text (9:00)

Eukaryotic Genome:

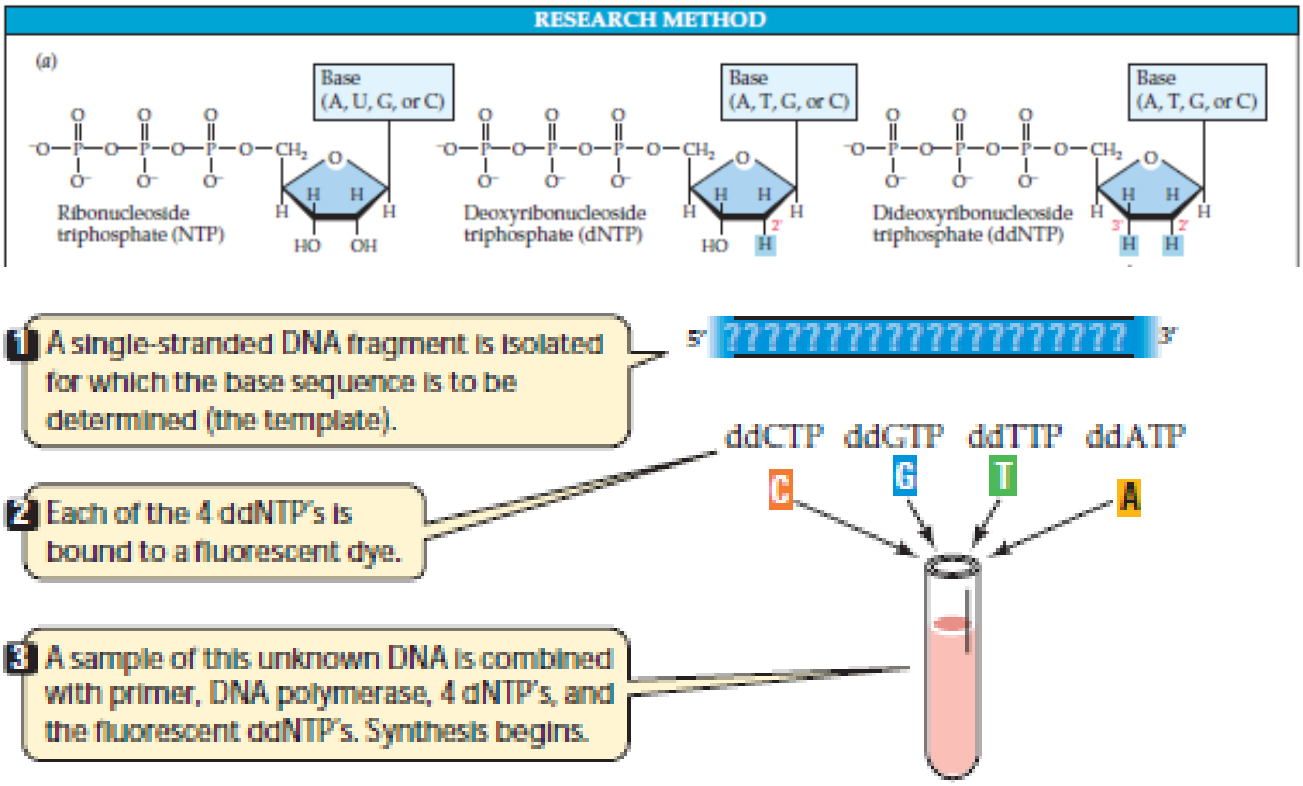
The genomes of most eukaryotes are larger and more complex than those of prokaryotes. This larger size of eukaryotic genomes is not inherently surprising, since one would expect to find more genes in organisms that are more complex. However, the genome size of many eukaryotes does not appear to be related to genetic complexity. For example, the genomes of salamanders and lilies contain more than ten times the amount of DNA that is in the human genome, yet these organisms

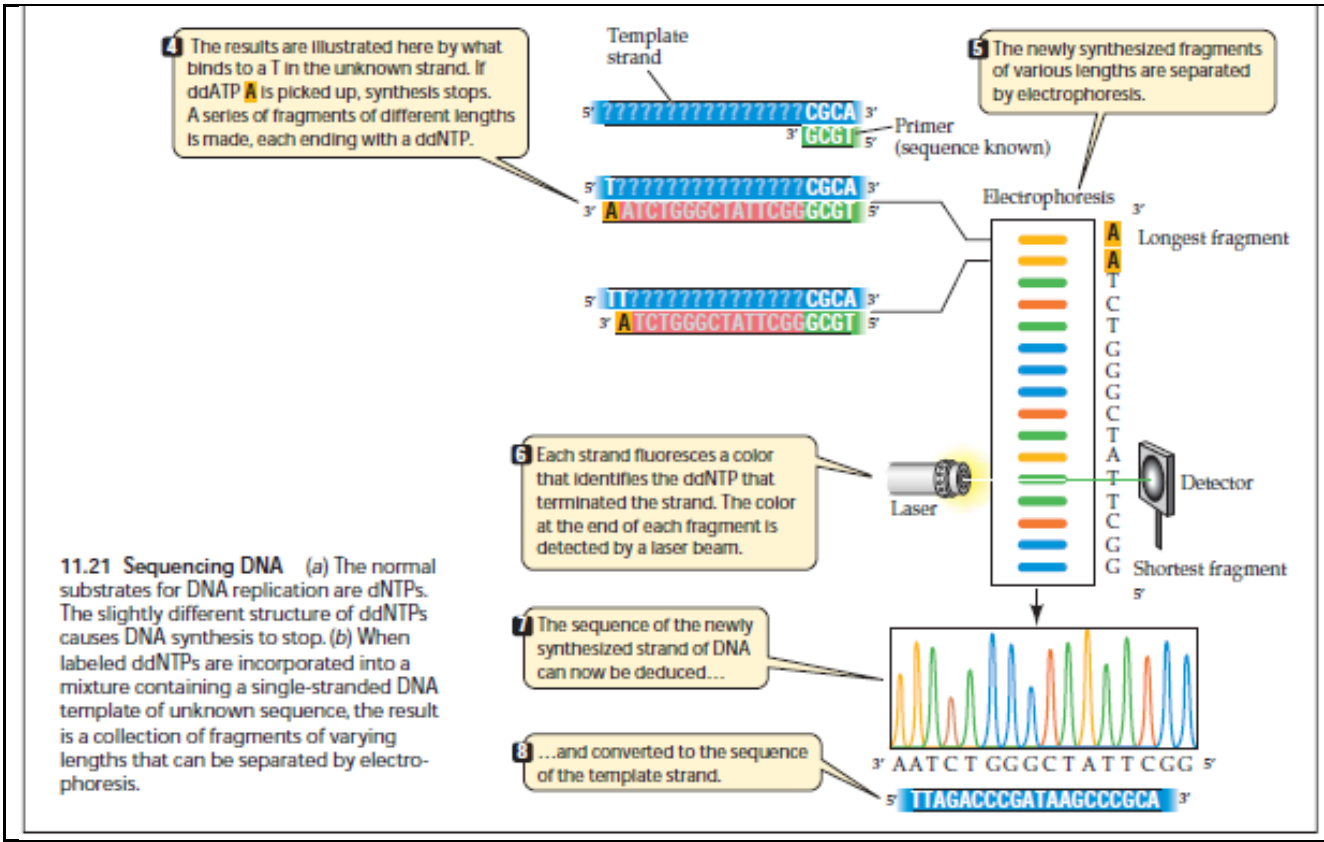
are clearly not ten times more complex than humans.

This apparent paradox was resolved by the discovery that the genomes of most eukaryotic cells contain not only functional genes but also large amounts of DNA sequences that do not code for proteins. The difference in the sizes of the salamander and human genomes thus reflects larger amounts of non-coding DNA, rather than more genes, in the genome of the salamander. The presence of large amounts of noncoding sequences is a general property of the genomes of complex eukaryotes. Thus, the thousand fold greater size of the human genome compared to that of *E. coli* is not due solely to a larger number of human genes. The human genome is thought to contain approximately 100,000 genes—only about 25 times more than *E. coli* has. Much of the complexity of eukaryotic genomes thus results from the abundance of several different types of noncoding sequences, which constitute most of the DNA of higher eukaryotic cells

Module 22: DNA Sequencing

Text (9:00)





11.21 Sequencing DNA (*a*) The normal substrates for DNA replication are dNTPs. The slightly different structure of ddNTPs causes DNA synthesis to stop. (*b*) When labeled ddNTPs are incorporated into a mixture containing a single-stranded DNA template of unknown sequence, the result is a collection of fragments of varying lengths that can be separated by electrophoresis.

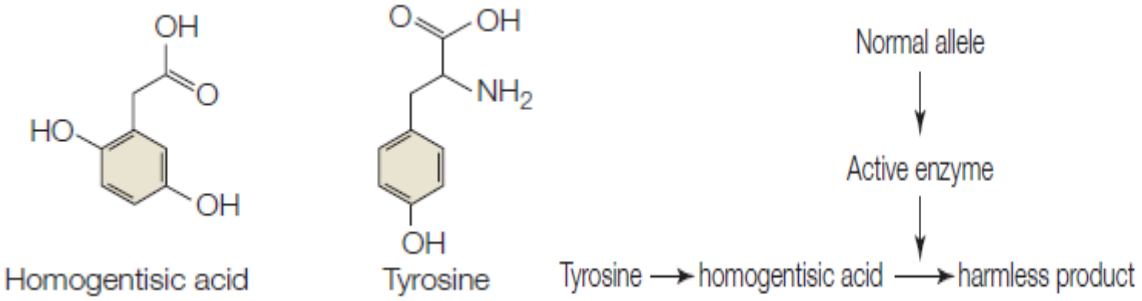
Module 23: Genetic information is converted into proteins

Text (8)

- **Alkaptonuria ("black urine"):**

The molecular basis of phenotypes was actually discovered before it was known that DNA was the genetic material.

He linked the biochemical phenotype of the disease to an abnormal gene and a missing enzyme**.**



Most common in children whose parents were first cousins 12.5%

- **Neurospora:**

An altered gene resulted in an altered phenotype, associated with an altered enzyme

Neurospora haploid (n) recessive alleles.

X-rays, which act as a mutagen prototrophs (original) converted to auxotrophs (increase)

Some mutant strains could no longer grow on the

minimal medium. One group could grow if supplemented with the arginine (arg mutants)

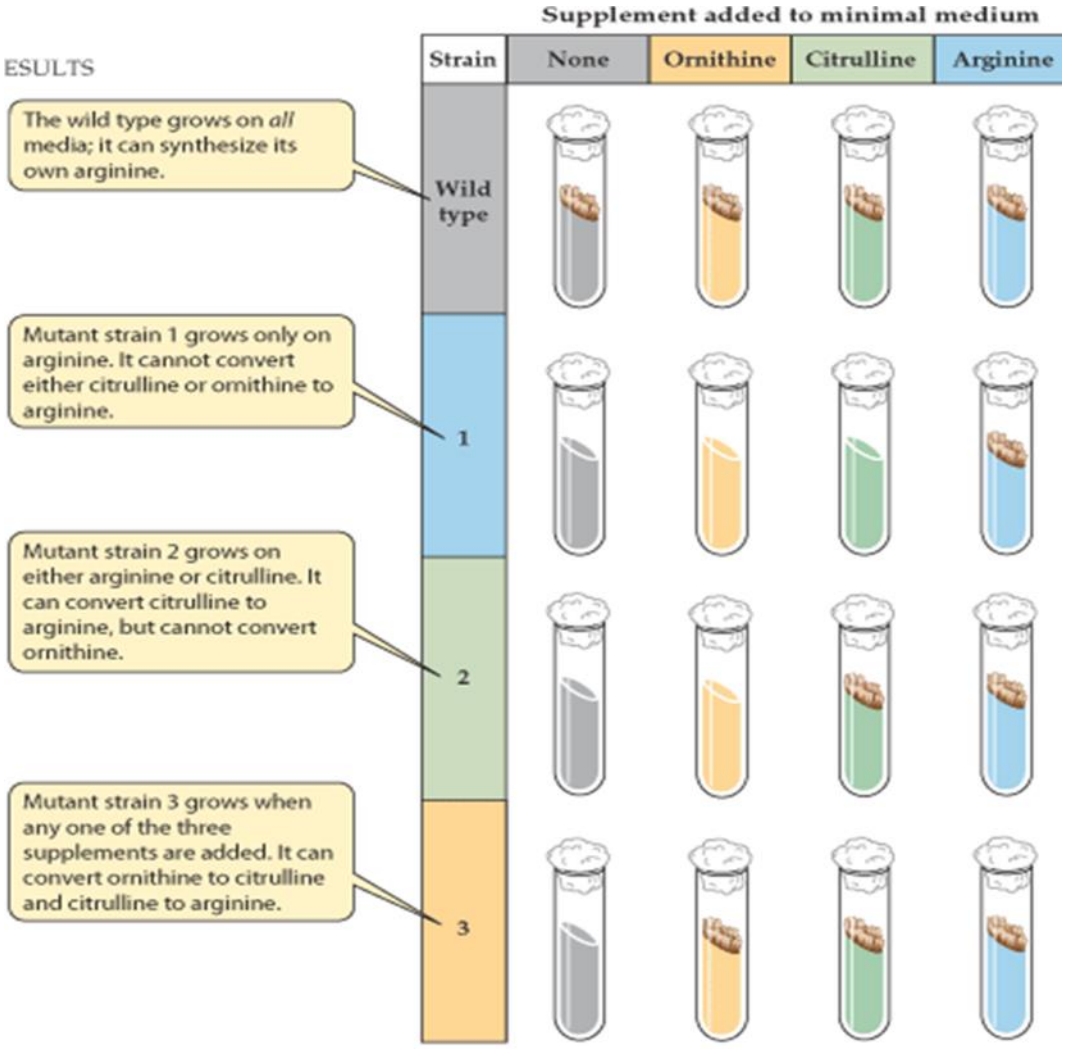
arg mutants were grown in presence of various

compounds *suspected intermediates* in the synthetic metabolic pathway for arginine,

B & T classified each mutation as affecting one enzyme or another.

wild-type and mutant cells examined for enzyme activities.

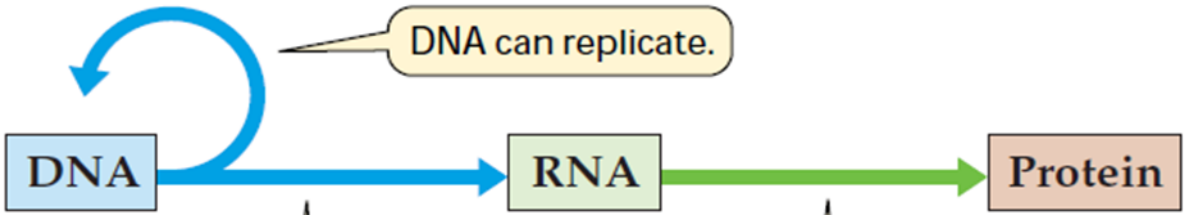
results confirmed: Each mutant strain was indeed missing a single active enzyme in the pathway.



Module24: Central dogma of molecular biology

Text (9)

Info flows from DNA to RNA to proteins:



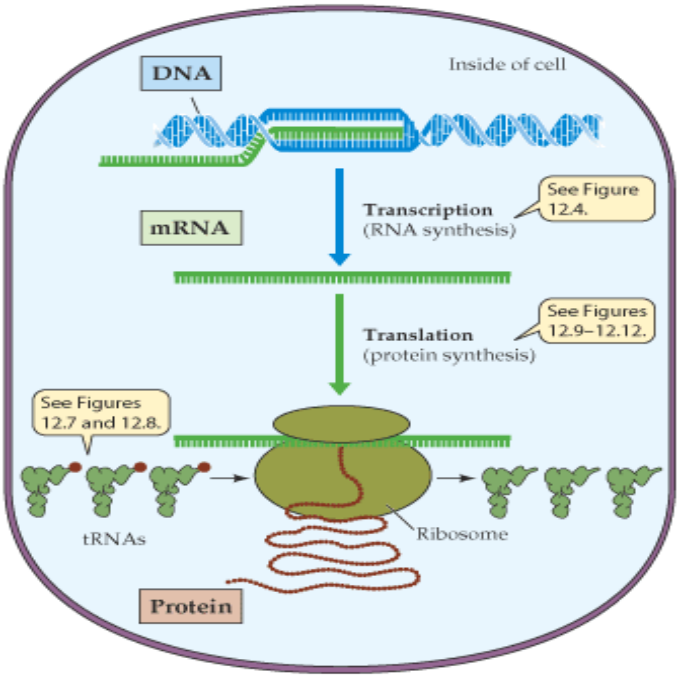
- How does genetic information get from the nucleus to the cytoplasm?
- What is the relationship between a specific nucleotide sequence in DNA and a specific amino acid sequence in a protein?

➢ **THE MESSENGER HYPOTHESIS AND TRANSCRIPTION**

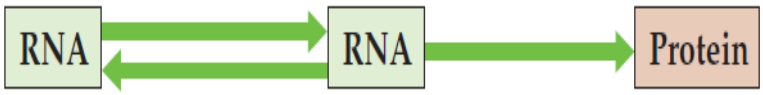
RNA molecule forms complementary copy of one DNA strand, moves to cytoplasm serves as template for protein synthesis.

➢ **THE ADAPTER HYPOTHESIS AND TRANSLATION**

Adapter molecule binds a specific amino
acid with one region and recognize a sequence of nucleotides with another region to translate the language of DNA into the language of proteins

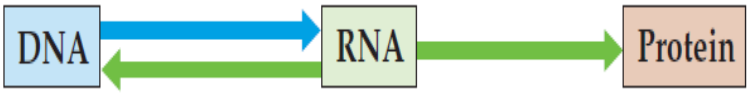


- **Influenza virus, & poliovirus**



Transcribing from RNA to RNA

- **HIV**



make a DNA copy of their genome

Module 25: Background of Bioinformatics

Text (8)

1. BACKGROUND

The term bioinformatics was first introduced in 1990s. Originally, it dealt with the management and analysis of the data pertaining to DNA, RNA and protein sequences. As the biological data is being produced at an unprecedented rate, its management and interpretation invariably requires bioinformatics. Bioinformatics is an interdisciplinary science at the cross-roads of biology, mathematics, computer science, chemistry and physics. With the digitalization of the biological information, doors have been wide opened towards the analysis of this information using computer algorithms and software.

Now we know well that the human genome has over 25,000 genes and these genes code for thousands of different proteins which perform day-to-day functions in the living cell. Furthermore, these proteins may take on various post-translational modifications leading to a very large number of functionally unique molecules. This presents us with a huge challenge in identification of genes and proteins.

1.1. EXPERIMENTS IN BIOLOGY

With the advancements in experimental protocols, now we have several next generation instruments and techniques available for obtaining digitalized biological information on genes and proteins etc. These instruments include:

- Next Generation Sequencers (NGS) for whole genome sequencing
- High Resolution Mass Spectrometry for whole proteome profiling
- Nuclear Magnetic Resonance Spectroscopy for structural studies

1.2. DIGITALIZATION OF BIOLOGY

In today's world, when a biologist performs an experiment in the wet lab, he or she in fact produces digital data which is continuously being stored on computer disks. The data may include text, numbers, symbols or images.

The study of the fundamental computation performed by biological processes, from gene regulatory systems to ecosystems and from neural networks to swarming systems. The increasing amount of data that are being acquired, stored, and processed in the life sciences and health sector makes the development of new information technologies one of the key factors for advancing the current state of knowledge in biomedical and health research.

1.3. SPEED OF DATA GROWTH

Due to advancement in instrumentation used in biological experiments, data is being accumulated at exponentially increasing rates. For example, genome sequences in genome databases are doubling every few years.

2. CONCLUSION

Human brain is limited in recalling information from memory. First, we should commit all

information to our memory followed by its recall. To overcome our ability to memorize and recall, computers can come to our rescue. This is because computers have an infinite ability to recall this information and process it quickly towards results.

Module 26: Introduction to bioinformatics

Text (6 minutes)

1. Bioinformatics is an interdisciplinary field mainly involving molecular biology and genetics, computer science, mathematics, and statistics. Data intensive, large-scale biological problems are addressed from a computational point of view. The most common problems are modeling biological processes at the molecular level and making inferences from collected data.

2. MOTIVATION

- Bioinformatics is a becoming a popular science due to several reasons.
- It is an i**nterdisciplinary field** as it covers the information of biological digital information including human, plants, animals, and microorganisms
- Although it is a new field, but it is rapidly **developing field**
- It demands a very **low-cost** infrastructure and hardly any lab equipment
- As bioinformatics data concerns a wide range of species such as humans, plants and micro-organisms, it presents us with **plenty of opportunities** in scientific discovery.

2.2. SCOPE OF BIOINFORMATICS

Bioinformatics primarily deals with digitalized biological information as well as data reported from biology experiments. Computational methods, data processing techniques and algorithms are employed in addressing the following issues:

- Storage of data
- Organization data
- Analysis of many experiments
- For representation of biological information

Bioinformatics is the application of computer technology to get the information that's stored in certain types of biological data. Bioinformatics provides central, globally accessible databases that enable scientists to submit, search and analyses information.

It offers analysis software for data studies and comparisons and provides tools for modelling, visualizing, exploring and interpreting data. The main goal is to convert a multitude of complex data into useful information and knowledge.

Bioinformatics approaches are used to understand the function of genes, the regulation of cells, drug target selection, drug design, and disease. Without quantitative analysis of the massive amounts of biological data generated by various systems, biology and -omics data cannot be interpreted or exploited.

2.3. ACTIVITIES

In modern biological sciences, bioinformatics is used for activities such as:

- Developing algorithms for organizing data collected from experiments
- Writing software and tools for data analysis
- Data processing to determine the role of underlying biomolecules
- Statistical evaluation of data using methods such as t-test and ANOVA
- Data visualization for meaningful presentation of biological information

3. CONCLUSION

In Pakistan, the field of biology is undergoing a rapid change due to the onset of bioinformatics. New research and educational programs are being constructed which is opening new door of opportunities for our future generations.

Module 27: NEED OF BIOINFORMATICS-I

Text (4 minutes)

1. NEED FOR BIOINFORMATICS –I

Our body's made up of trillions of cells. According to human genome project, the number of genes in each cell is approximately 20,000. This microscopic cell has an ultramicroscopic commanding center called nucleus in which DNA is packaged. The number of nucleotides is ~3*10^9. That much enormous data in a cell. How could we this, access this data, analyze this data. Here comes the use of computers. We developed and use computers for the same purpose, efficient data storage retrieval and analysis. With the advancement in sequencing technology, each day thousands of nucleotides of different organisms are sequenced and submitted to the databases worldwide. In bioinformatics, the use of computer is same as previously but the data is biological data, the letters of life. Actually, we are now facing an information load. Loads pf sequence data but the real challenge is to make sense of this data.

Listed are some of the major needs of bioinformatics

- To store ans retrieve biological data
- To analyses the biological data like sequence patterns
- To interpret biological data
- To predict 3D structures of bio molecules
- To construct evolutionary trees that help us to find ancestry of different organism.

If we look at the pace of development in bioinformatics then we can easily observe that from year's 2000 to 2015, the number of online tools for processing genomics and proteomics information are rapidly increasing. This is just a reflection of the need for bioinformatics in modern day biology.

The field of Bioinformatics and Computational Biology is characterized by a highly diverse confluence of traditional academic disciplines. Informatics and Bio-science are the umbrella terms given to a set of allied disciplines which make up the field, but a much larger array of traditional areas **contributes** to the set of tools needed by individuals training for this new and expanding interdisciplinary field. Biomedical Engineering, Electrical and Computer Engineering, Computer Science, Applied Mathematics, Genetics, Biology, Anatomy and Cell Biology, Microbiology, and Biostatistics are the principal allied disciplines.

2. **CONCLUSION**
   The need for bioinformatics is on a rapid rise as biological data is rapidly increasing and becoming available online, free of any cost.

Module 28: NEED FOR BIOINFORMATICS –II

Text (7 minutes)

If we observe the growth of gene bank than from 1982 it comprised of 2 billion base pairs but by year 2002 it had risen to 56 billion base pairs. With the data in our hands, there is an urgent need to interpret this data. For instance, analysis of this data can help us in developing an understanding of the phylogenetic "tree of life" which consist of:

- Bacteria
- Archaea
- Eucarya

Towards exploring the possible benefits of using bioinformatics, one needs to answer the following question:

1. **WHAT IS IT THAT BIOINFORMATICS CAN DELEIVER?**
   The simple answer to that bioinformatics is:
   - Provide us better understanding of life, evolution, molecular mechanisms as well as disease.
   - Moreover, we can make better drugs with the availability of an enhanced molecular understanding of disease.

1.1. POSSIBLE CONTRIBUTIONS

- It can help us to organize the large datasets from new experiments instruments.
- Bioinformatics can help store and process this data as well.
- It can provide insights into the meanings of our research results and findings.
- Overall, it can help us to better understand paradoxes defining the life forms.

2. **CONCLUSION**
   From gene sequencing to protein sequencing, bioinformatics is providing us with an improved understanding of the genes, proteins, protein interaction and signaling pathways involved in biological functioning and disease.

Module 29: APPLICATIONS OF BIOINFORMATICS – I

Text (8 minutes)

There is a tremendous application of bioinformatics in the field of homology and similarity tools, protein function analysis, personalized medicine, Gene therapy, Drug development, Comparative Studies and also climate change studies. Computational methodologies have turn into a noteworthy piece of structure-based medication outline. Structure-based medication outline uses the three-dimensional structure of a protein focus to plan hopeful medications that are anticipated to tie with high natural inclination and selectivity to the objective.

For comprehensive study please see the link: https://microbenotes.com/bioinformatics-introduction-and-applications/

When we look at bioinformatics, it seems to be a very complex and abstract field. How and where can bioinformatics be applied specifically? How does it improve the fundamental understanding of biological phenomenon? Most importantly, how can its benefits be delivered to the society at large?

The answers to these questions are categorized as follows:

1. **GENOMICS**
   - Bioinformatics can help in assembling DNA sequencing data
   - It can help in gene finding (markers)
   - Gene assembly can be performed using bioinformatics tools (nucleotide alignments)
   - It can help transcribe the gene data to RNA data
   - Also, databases can be generated from such data

2. **EVOLUTIONARY STUDIES**
   - Evolutionary relationships between different organisms can be derived from data.
   - Evolutionary distance among species can be computed by using bioinformatics tools
   - Phylogenetic trees can be constructed to find relationships between species
   - Ancestry can be better understood between several species and organisms

3. **PROTEOMICS**
   - Bioinformatics can help us in decoding protein sequences
   - It can also help us in understanding protein structure
   - We can also understand post translational changes in proteins with the help of bioinformatics
   - We can better understand the protein-protein interaction in different biological reactions
   - It can also help us in generating databases of these sequences and structures

4. **SYSTEMS BIOLOGY**
   - Bioinformatics can assist us in modelling regulatory mechanisms in gene and protein networks
   - Such models can be analyzed to identify the key regulators in these networks
   - Moreover, the models can help evaluate drugs to treat these key regulators

5. **CONCLUSION**

Bioinformatics can be applied to life in many ways it helps us to understand the sequence and function of biomolecules and their relationships. Recent trends in bioinformatics involve development of personalized therapeutics for cancer and diabetes.

Module 30: Applications of Bioinformatics - II

Text (7 minutes)

1. INTRODUCTION

Bioinformatics is now being commonly applied in routine research and analysis. Most significantly, its salient applications include **Genomics, Transcriptomics, Proteomics, Metabolomics, Structural Proteomics, Designing Drugs, System Biology** and in personalization of medicines for cure.

Except this applications Bioinformatics introduced us the techniques which enabled us to generate the large data regarding biology and its use. And step by step the applications of bioinformatics increased from genomic level to entire system level.

1.1. SMALL TO BIG

- Bioinformatics helps us to understand the systems from small to big like from gene findings to entire system prediction
- In structure findings and modeling of many biological system to understand them in better ways
- Bioinformatics helped the human to understand the protein, protein interaction in many biological systems
- And provide us the concept how these biological processes are interconnected with each other and how they affect each other
- Now we can understand the modeling of molecules and genome at cell level
- Signaling pathways are easy just because of bioinformatics
- Now morphology of tissue can be understanding by creating the models with help of bioinformatics tools

2. CONCLUSION

Bioinformatics not only just collect, analyze and store the data it processes it in very authentic way and validates our hypothesis and very soon in future it will help us to understand that which disease is coming in future and how to tackle it with personalize medicine.

Module 31: Frontiers in Bioinformatics - I

Text (6 minutes)

1. INTROCDUCTION

Bioinformatics is new and emerging field of science having vast opportunities and with innovation in tools it is increasing the scale of biological data, but still there are many unsolved challenges which are pending in the field of life science and for which bioinformatics is doing new innovative ideas.

1.1. FRONTIER IN GENOMICS

- Now we can sequence the whole genome with the bioinformatics tool i.e. Next generation sequencing (NGS)
- We can save, store, and analyze the massive amount of biological data which is in (Terabyte files)
- We can handle the large number of data easily and can process it as well in easy way
- Whole genome can be assembled in sequence and flaws can be identified easily

1.2. FRONTIER IN TRANSCRIPTOMICS

- Now in genomics we can identify those matters which are unknown yet or under discussion
- Role of RNA in making proteins and its dynamics can be understood easily
- Interactions of RNA molecule can be easily understood by simple model

1.3. FRONTIER IN PROTEOMICS

- We can identify the deficiency of low proteins in any patient's body tissue
- We can identify production of protein in large molecular level in any organism
- Pathways before and after any biological reaction are easy to design

2. CONCLUSION

Bioinformatics is literally a science of full of challenges and opportunities having a revolution in field of biology and routine life.

Module 32: FRONTIERS IN BIOINFORMATICS-II

Text (6 minutes)

1. INTRODUCTION

Frontier in Bioinformatics includes

- Next generation genomics
- Transcriptomics
- Proteomics

1.1. FRONTIER IN PROTEIN STURUCTURE

Bioinformatics helps us to understand the layer folding of proteins that how they are proceed and helps us to know that how protein interact with each other and how a drug can affect or stimulate a protein. Protein structure also relate with

1.2. FRONTIER IN SYSTEM BIOLOGY

It helps us to understand the whole system of a single cell, in that cell how organelles, gene, proteins and metabolites are interconnected in a single unified system (cell). And bioinformatics also gives us the idea how these models can be applied to real-time.

1.3. FRONTIER IN PERSONALIZED MEDICINE

This is the important thing for this century and upcoming generation that personalize the medicine for exact cure of a disease. Because all the medicine cannot work exact some effect patient badly therefor with the help of Bioinformatics, we are now able to personalize some medicines for some diseases. And bioinformatics helps us to evaluate the medicine.

2. CONCLUSION

If we talk about the 21st century than it's the century of bioinformatics it will enable the human to cure many diseases with one drug by personalizing it.

Module 33: The Central Dogma

Text (13)

The Central Dogma:

- The central dogma outlines the flow of genetic information during growth and division of the cells.
- Genetic information flows from DNA to RNA to protein during cell growth.
- In addition, all living cells must replicate their DNA when they divide.
- During cell division each daughter cell receives a copy of the genome of the parent cell.
- Replication is the process by which two identical copies of DNA are made from an original molecule of DNA.
- So Replication occurs in the cells prior to cell division.
- An important point is that information does not flow from protein to RNA or DNA.
- However, flow of information from RNA "backwards" to DNA is possible in certain special circumstances due to the operation of reverse transcriptase
- By the end of 1953, the working hypothesis was adopted that chromosomal DNA functions as the template for the synthesis of RNA molecules.
- These RNA molecules, the subsequently move to the cytoplasm, where they determine the arrangement of amino acids within proteins
- In 1956, Francis Crick referred to this pathway for the flow of genetic information as the Central Dogma.

- An important point in the above equation is that the two arrows are unidirectional which means that RNA sequences are never determined by protein templates nor was DNA then imagined ever to be made on RNA templates.
- The idea that proteins never serve as templates for RNA has stood the test of time.
- However, RNA chains sometimes do act as templates for the synthesis of DNA chains of complementary sequence.
- Such reversals of the normal flow of information are very rare events compared with the enormous number of RNA molecules made on DNA templates.
- Thus, the central dogma as originally proclaimed more than 50 years ago still remains essentially valid.

Module 34: Gene, mRNA and Protein Sequences

Text (12 minutes)

1. INTRODUTION

We know that all living things are composed of cells. Here, a question arises on how these cells came into being? For composition of cell DNA has blueprints for building cells along with the information of cell's protein, carbohydrate and vitamins production.

And transfer of this information from DNA to these molecules is termed as **"Central Dogma"** which is:

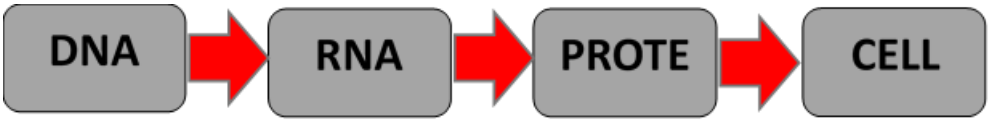


Figure 3.1.1: Flowchart of Central Dogma

1.1. DNA

DNA is a hereditary material present in all living organisms. It passes from one generation to the other via cell division. All cells have DNA. In eukaryotic cell's DNA presents in nucleus whereas in prokaryotic cell's DNA present in spreader form in cytoplasm. Due to DNA nucleus is known as the brain of cell. DNA molecule is double helix structure contains base pairs composed of nucleotides and these nucleotides are composed of sugar phosphate group and are bind with each other with hydrogen bonds.

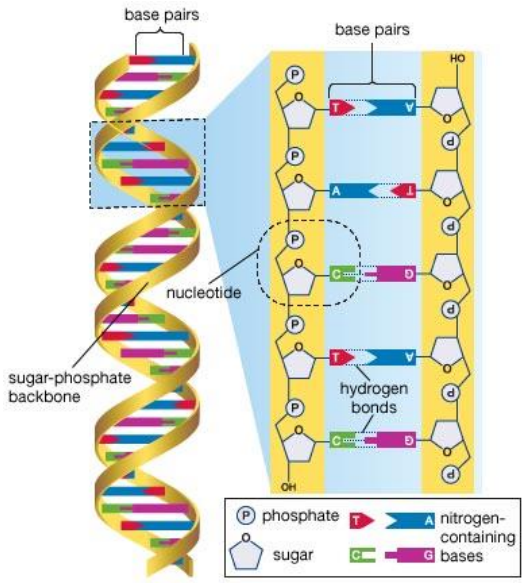


Figure 3.1.2: DNA Double helix (Courtesy Britannica)

DNA and RNA are different from each other. DNA has double strand whereas RNA has single strand. Normally all the nucleotides are same in both DNA and RNA except one position in RNA which is U (Uracil) and in DNA it is T (Thiamin). We will briefly discuss the difference between RNA and DNA in coming MODULEs.

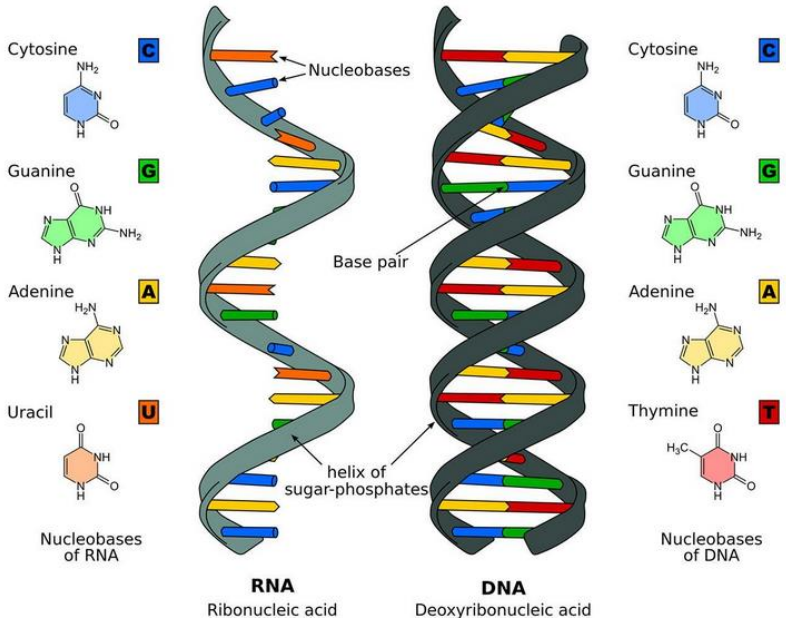


Figure 3.1.3: RNA vs. DNA

2. MECHANISM

DNA sends the information to cell via **mRNA** and that sequence the amino acids according to coded information and **protein** structure is formed and that protein forms a **cell**.

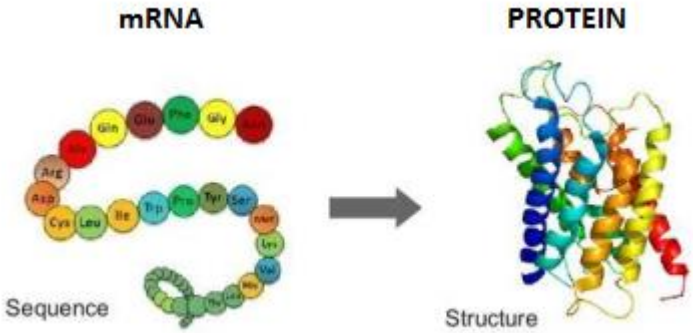


Figure 3.1.4: mRNA sequence to protein sequence

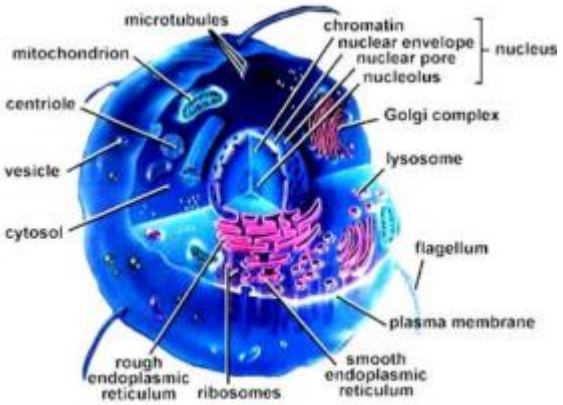


Figure 3.1.5: Structure of animal cell

3. CONCLUSION

According to the central dogma DNA codes information for RNA and RNA makes the Protein and that protein along with some biomolecules make cells and its systems.

Module 35: NUCLEOTIDES

Text (7 minutes)

1. INTRODUCTION

A nucleotide is **the basic building block of nucleic acids**. RNA and DNA are polymers made of long chains of nucleotides. A nucleotide consists of a sugar molecule (either ribose in RNA or deoxyribose in DNA) attached to a phosphate group and a nitrogen-containing base.

1.1. TYPES OF NITROGENOUS BASES

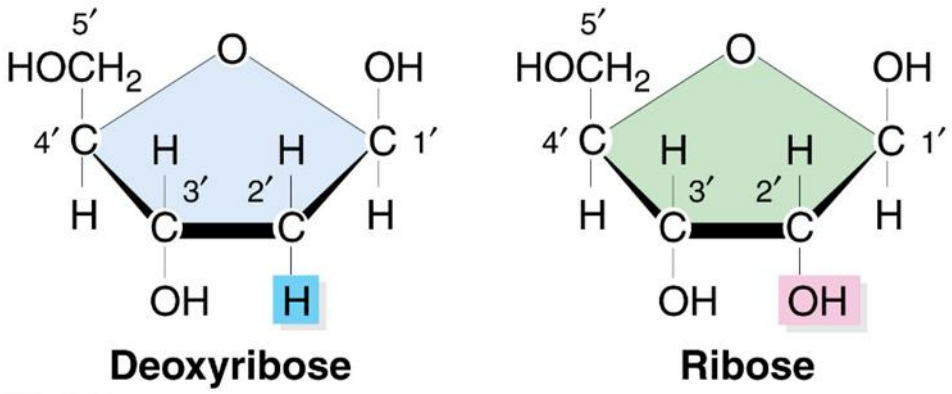
There are five types of nitrogenous bases.

- Adenine (A)
- Cytosine (C)
- Guanine (G)
- Thymine (T) &
- Uracil (U)

1.2. STRUCTURE OF DNA AND RNA

DNA molecule although is double stranded and RNA is single stranded but there is difference in sugar composition and in one nitrogenous base i.e., in DNA Thymine present while in RNA Uracil present.

RNA has Ribose sugar and DNA has De-oxyribose sugar:



DNA sugar RNA sugar

Figure 3.3.1: Difference between RNA and DNA sugar (Courtesy Pearson Education)

Adenine and Guanine collectively called Purines while Cytosine, Uracil, and Thymine are called as Pyrimidine.

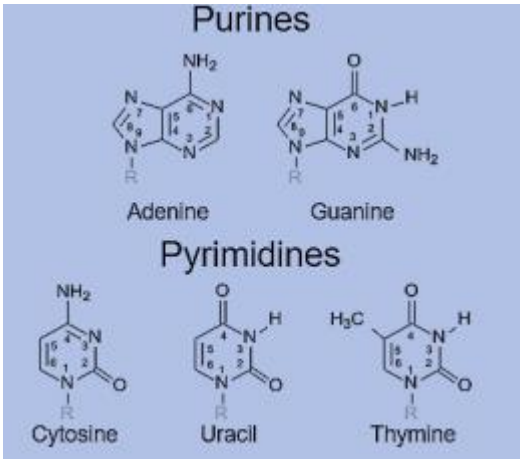


Figure 2.3.2: Structures of purines and pyrimidines (Courtesy Wikipedia)

When phosphate, nitrogen base and sugar come together if there is (OH) than molecule is RNA and if there is (H) in sugar than molecule is DNA.

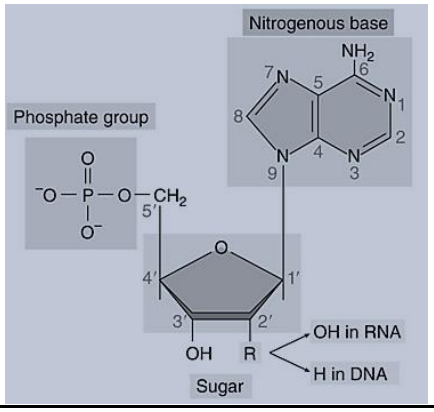


Figure 2.3.3: Detail view of one Nucleotide

2. CONCLUSION

DNA molecule make RNA and RNA make the protein and DNA differ from RNA in nature due to sugar and nitrogenous base. DNA just codes the information for protein, but RNA helps in making protein.

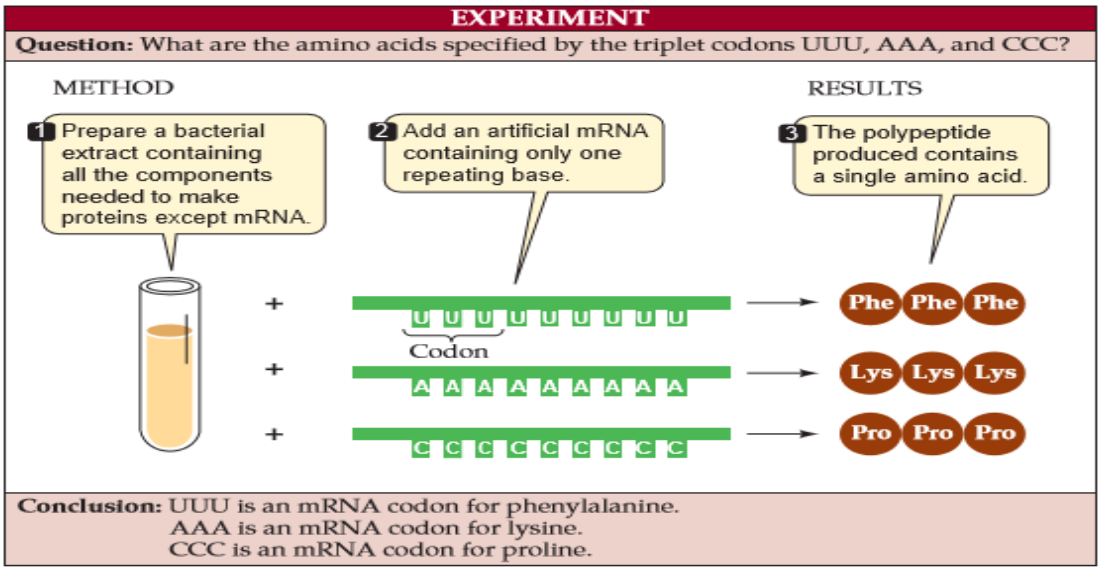
Module 36: Genetic Code

Text (9)

Genetic Code:

How do transcription and translation produce specific and functional protein products? These processes require a genetic code that relates genes (DNA) to mRNAand mRNAto the amino acids of proteins. The genetic code specifies which amino acids will be used to build a protein. You can think of the genetic information in an mRNAmolecule as a series of sequential, nonoverlapping three-letter "words." Each sequence of three nucleotide bases (the three "letters") along the chain specifies a particular amino acid. Each three-letter "word" is called a codon. Each codon is complementary to the corresponding triplet in the DNAmolecule from which it was transcribed. Thus, the genetic code is the means of relating codons to their specific amino acids. The complete genetic code is shown in Figure 12.5. Notice that there are many more codons than there are different amino acids in proteins. Combinations of the four available "letters" (the bases) give 64 (43) different three-letter codons, yet these codons determine only 20 amino acids. AUG, which codes for methionine, is also the start codon, the initiation signal for translation. Three of the codons (UAA, UAG, UGA) are stop codons, or termination signals for translation; when the translation machinery reaches one of these codons, translation stops, and the polypeptide is released from the translation complex.

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC Phenyl-alanine; UUA, UUG Leucine	UCU, UCC, UCA, UCG Serine	UAU, UAC Tyrosine; UAA Stop codon; UAG Stop codon	UGU, UGC Cysteine; UGA Stop codon; UGG Tryptophan	U, C, A, G
C	CUU, CUC, CUA, CUG Leucine	CCU, CCC, CCA, CCG Proline	CAU, CAC Histidine; CAA, CAG Glutamine	CGU, CGC, CGA, CGG Arginine	U, C, A, G
A	AUU, AUC, AUA Isoleucine; AUG Methionine; start codon	ACU, ACC, ACA, ACG Threonine	AAU, AAC Asparagine; AAA, AAG Lysine	AGU, AGC Serine; AGA, AGG Arginine	U, C, A, G
G	GUU, GUC, GUA, GUG Valine	GCU, GCC, GCA, GCG Alanine	GAU, GAC Aspartic acid; GAA, GAG Glutamic acid	GGU, GGC, GGA, GGG Glycine	U, C, A, G



the Genetic Code Nirenberg and Matthaei used a test-tube protein synthesis system to determine the amino acids specified by synthetic mRNAs of known codon composition

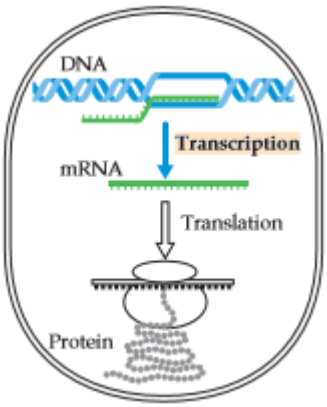
Module 37: Transcription

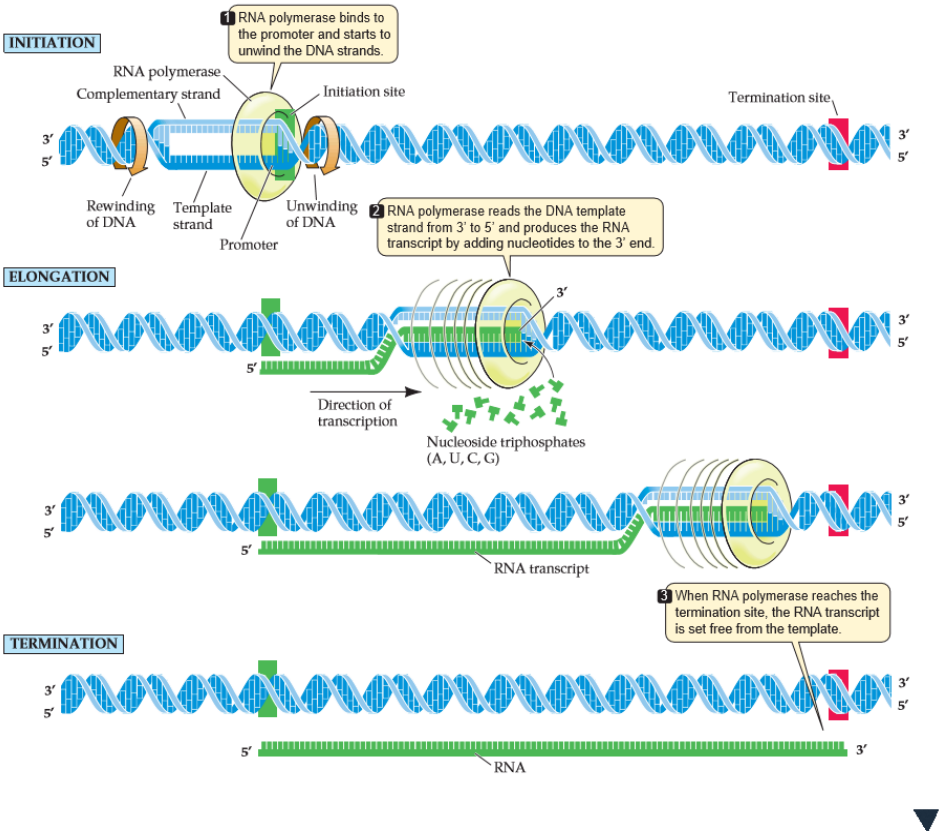
Text (8)

Transcription—the formation of a specific RNA from a specific DNA—requires several components:

- A DNA template for complementary base pairing
- The appropriate ribonucleoside triphosphates (ATP, GTP, CTP, and UTP) to act as substrates
- An enzyme, RNA polymerase

Within each gene, only one of the two strands of DNA— the template strand—is transcribed. The other, complementary DNA strand, referred to as the non-template strand, remains untranscribed. For different genes in the same DNA molecule, different strands may be transcribed. That is, the strand that is the non-template strand in one gene may be the template strand in another.





Module 38: TRANSCRIPTION

Text (7minutes)

1. **BACKGROUND**

   All cells are made up of proteins, carbohydrates, and lipids and for these cells DNA codes the information which makes the RNA and protein both.

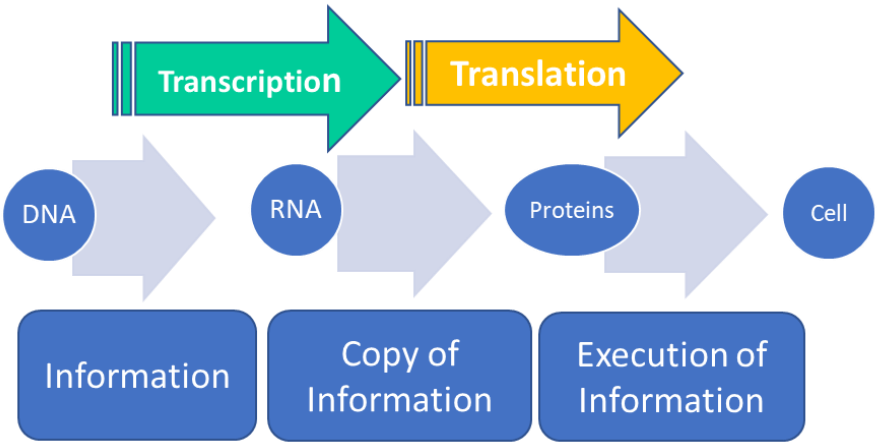


Figure 3.2.1: Flow of information from DNA to Proteins

TranscriptionTranscription is the process by which the information in a strand of DNA is copied into a new molecule of messenger RNA (mRNA). DNA safely and stably stores genetic material in the nuclei of cells as a reference, or template. Meanwhile, mRNA is comparable to a copy from a reference book because it carries the same information as DNA but is not used for long-term storage and can freely exit the nucleus. Although the mRNA contains the same information, it is not an identical copy of the DNA segment, because its sequence is complementary to the DNA template.

Transcription is carried out by an enzyme called RNA polymerase and a number of accessory proteins called transcription factors. Transcription factors can bind to specific DNA sequences called enhancer and promoter sequences in order to recruit RNA polymerase to an appropriate transcription site. Together, the transcription factors and RNA polymerase form a complex called the transcription initiation complex. This complex initiate transcription, and the RNA polymerase begins mRNA synthesis by matching complementary bases to the original DNA strand. The mRNA molecule is elongated and, once the strand is completely synthesized, transcription is terminated. The newly formed mRNA copies of the gene then serve as blueprints for protein synthesis during the process of translation.

2. MECHANISM

The above mechanism explains the process of **transcription** in very simple way, DNA codes the information and converted into RNA where mRNA copies the information and it execute the information in cell and amino acids combine with each other according to coded information of DNA and protein formation takes place. This is known as **translation**.

3. RNA VS. DNA

Molecule of DNA contains only four base pairs (A, T, C, and G) which are repeated thousands of time and Adenine “A” pairs with Thymine “T” by two Hydrogen bonding, While Cytosine “C” binds with Guanine “G” by three Hydrogen bonding.

Same like DNA, the RNA contains four base pairs but Thymine is replaced with Uracil “U” and RNA is single stranded.

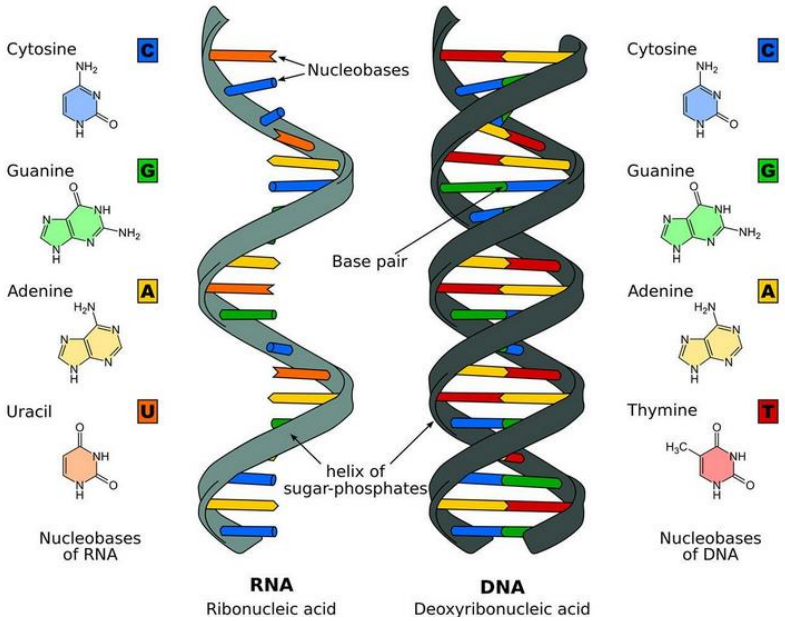


Figure 3.2.2: RNA vs. DNA

4. CONCLUSION

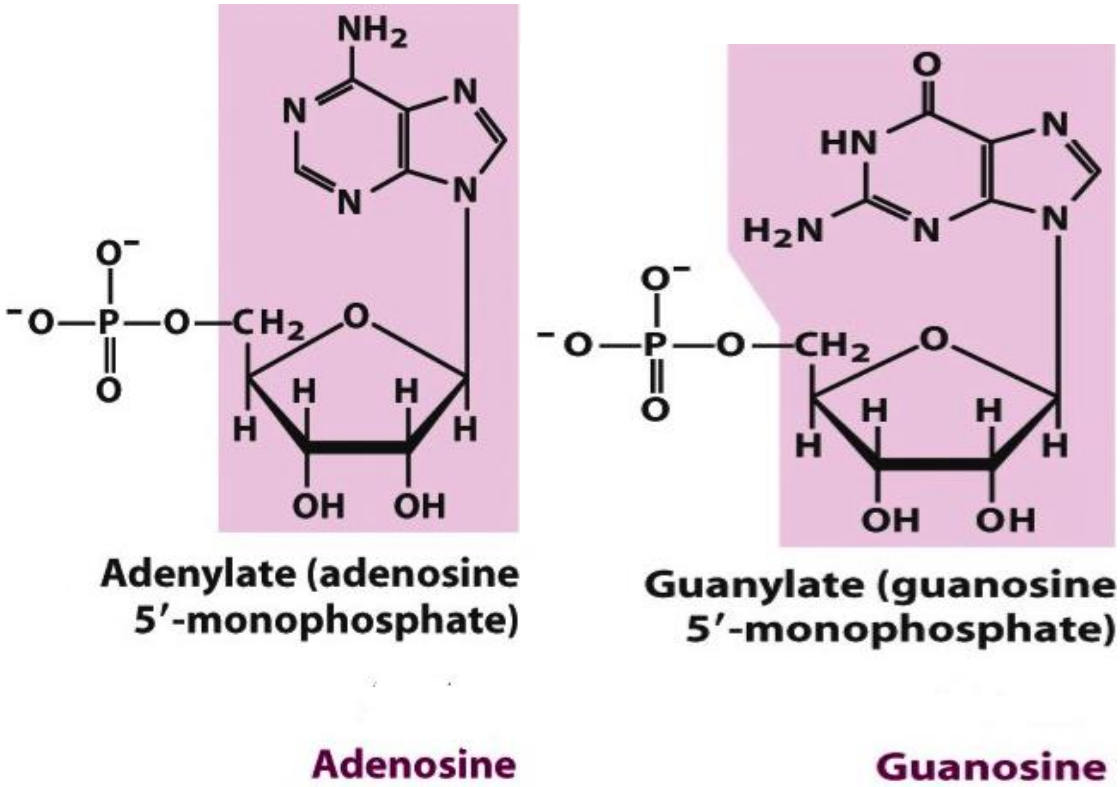
DNA has four bases **A**, **C**, **G**, **T** and RNA also has four bases **A**, **C**, **G**, **U**. DNA is double stranded whereas RNA is single stranded. DNA just codes the information for protein, but RNA helps in making protein.

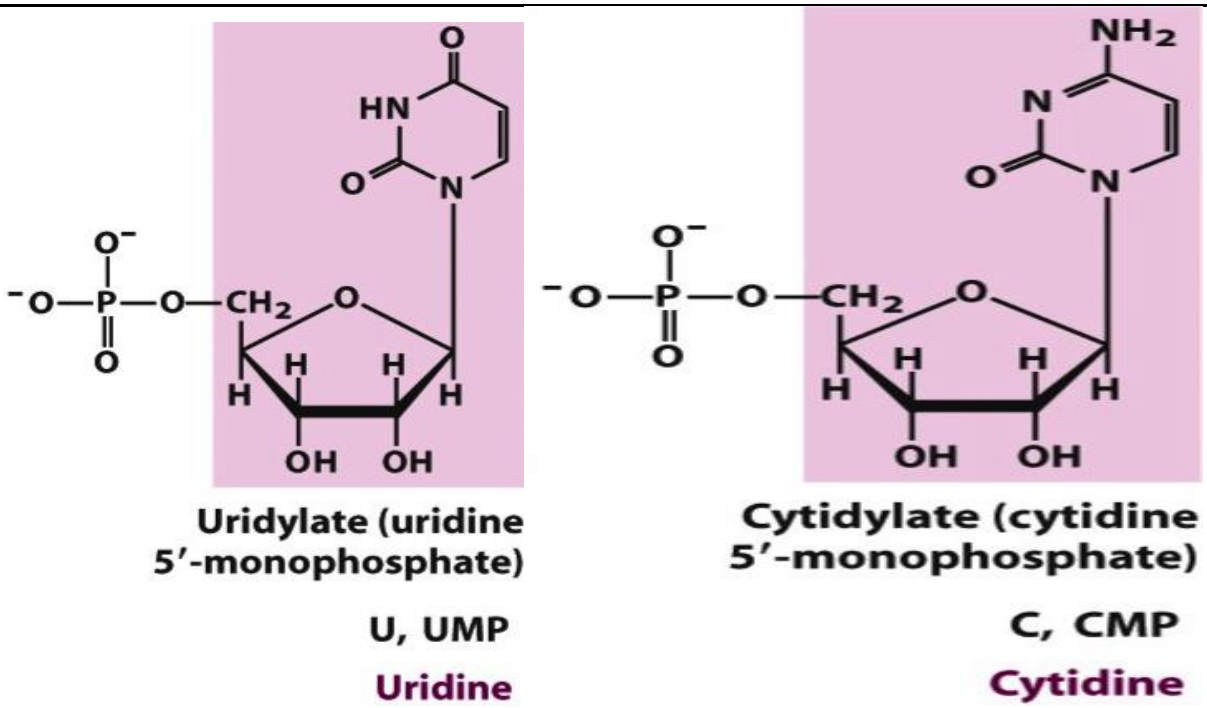
Module39: Types of Ribonucleotides

Text (10)

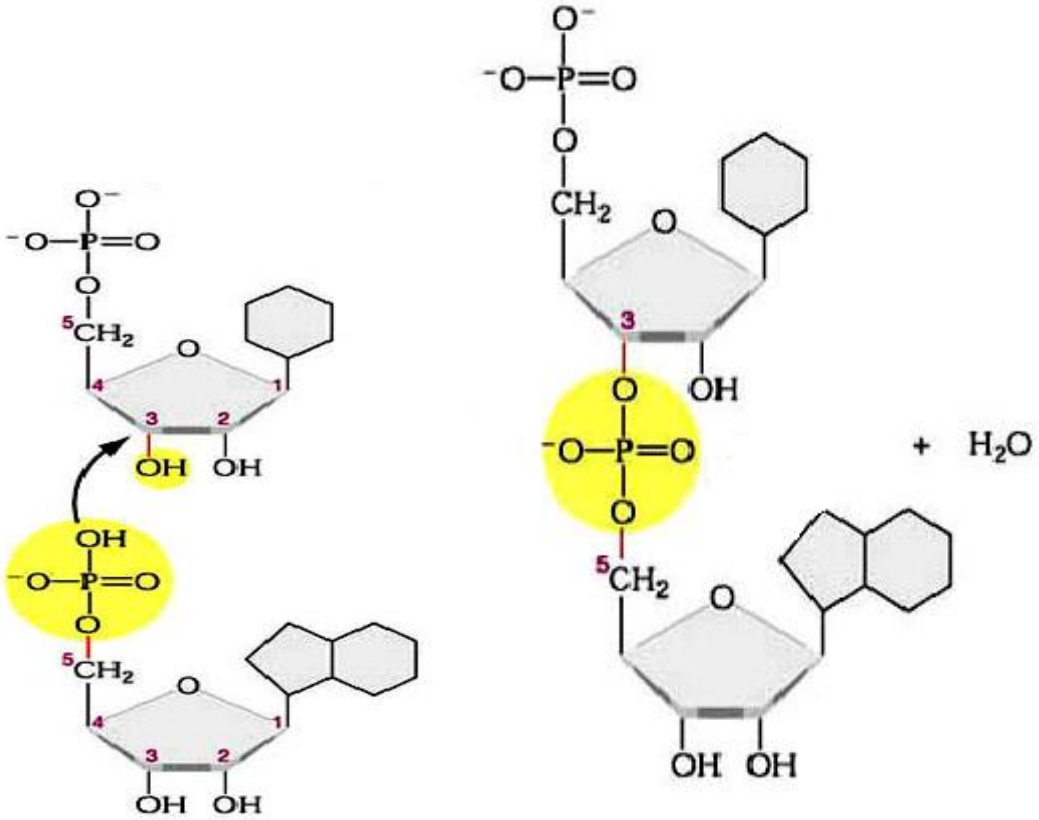
Types of Ribonucleotides:

- There are mainly four types of ribonucleotides depending upon the types of nitrogenous bases present in RNA.

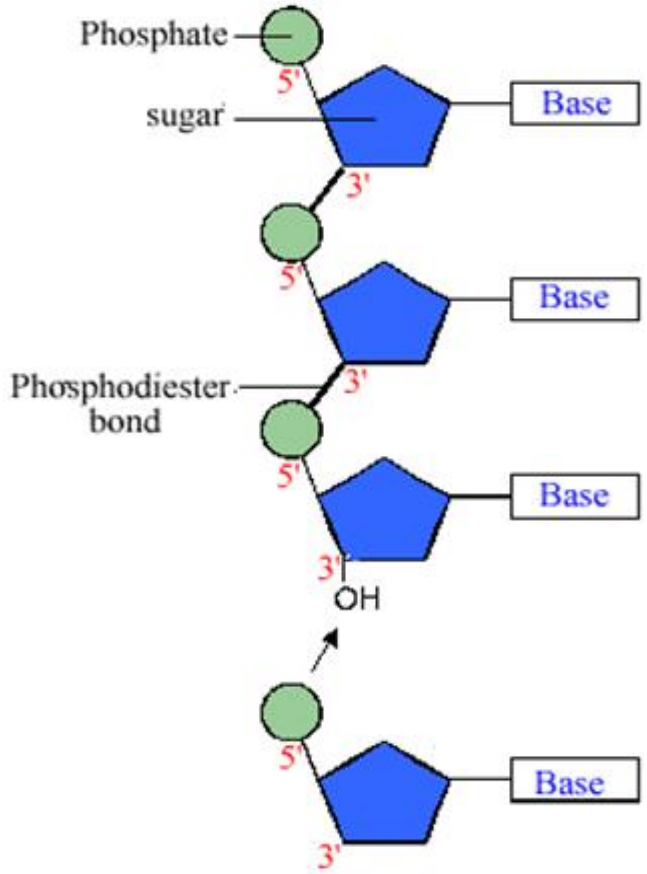




How do Ribonucleotides Join?



A Poly-Ribonucleotide



Module40: Types of RNA

Text (7)

There are mainly three types of Ribonucleic acids (RNAs) present in the cells of living organisms.

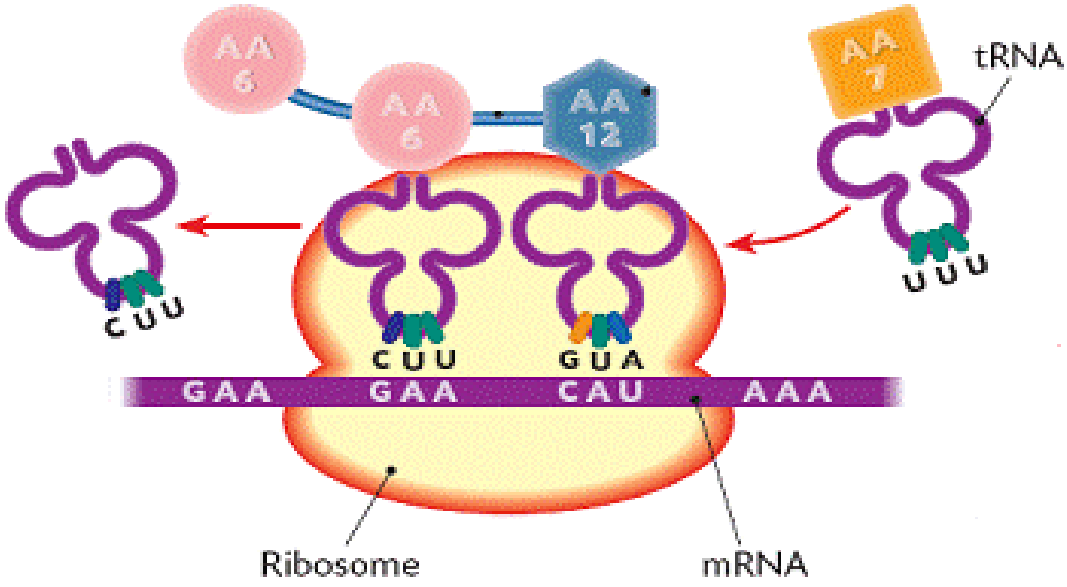
- Messenger RNA (mRNA)
- Transfer RNA (tRNA)
- Ribosomal RNA (rRNA)

Messenger RNA (mRNA)

- It is the type of RNA that carries genetic information from DNA to the protein biosynthetic machinery of the ribosome.
- It provides the templates that specify amino acid sequences in polypeptide chains.
- The process of forming mRNA on a DNA template is known as transcription.
- It may be monocistronic or polycistronic.
- The length of mRNA molecules is variable and it depends on the length of gene.

Transfer RNA (tRNA)

- Transfer RNAs serve as adapter molecules in the process of protein synthesis.
- They are covalently linked to an amino acid at one end.
- They pair with the mRNA in such a way that amino acids are joined to a growing polypeptide in the correct sequence.



Ribosomal RNA (rRNA)

- Ribosomal RNAs are components of ribosomes.

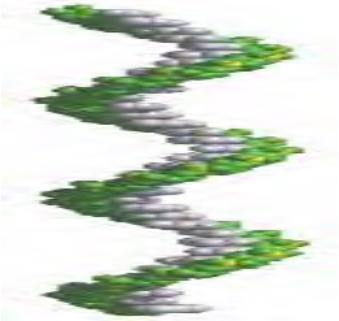
- rRNA is a predominant material in the ribosomes constituting about 60% of its weight.
- It has a number of functions to perform in the ribosomes.

Module41:Structure of RNA

Text (9)

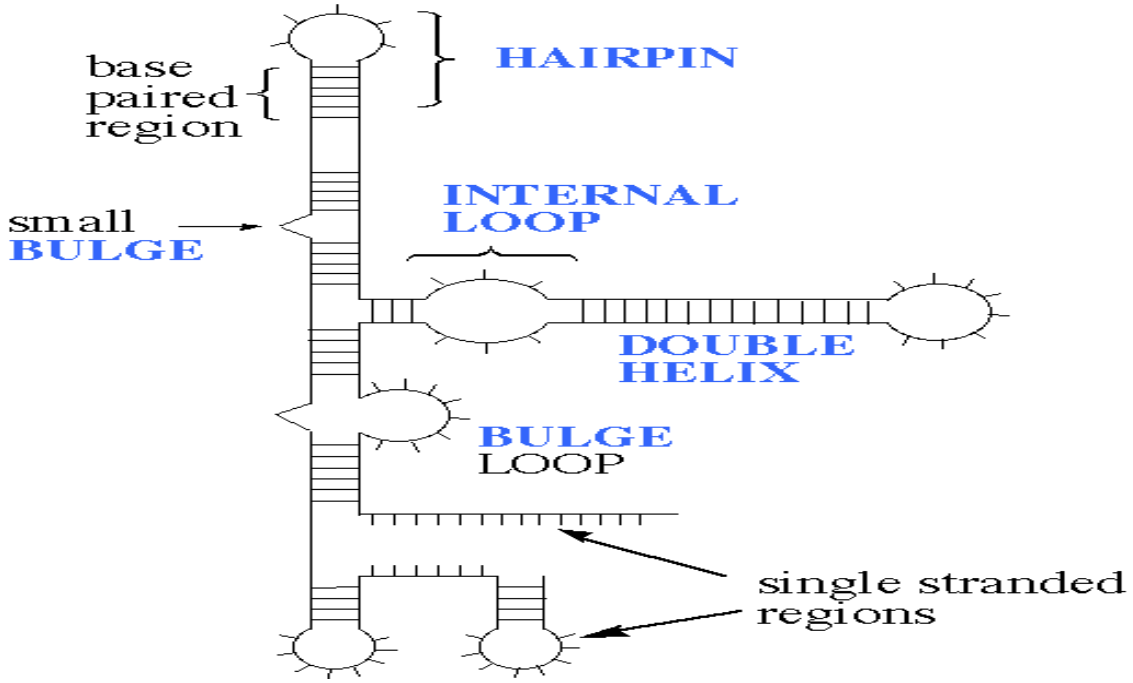
- mRNA is always single stranded when it is formed from DNA.
- But this single strand assumes a double helical conformation soon after its formation.
- This confirmation is achieved mainly due to base stacking interactions.

Messenger RNA (mRNA)

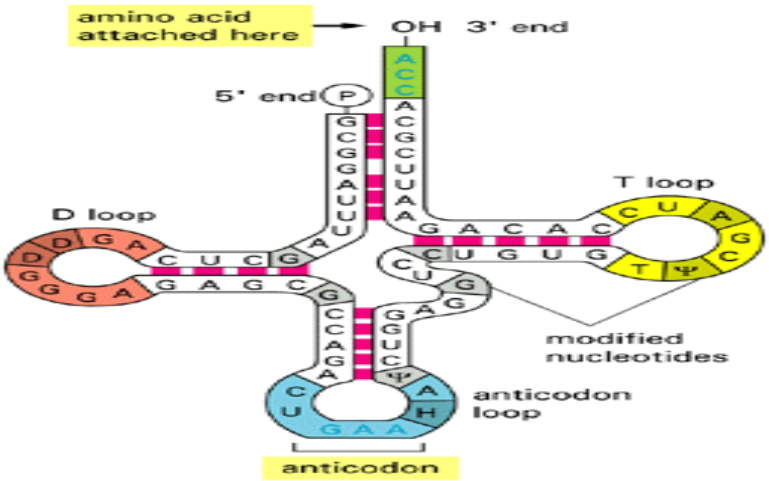


- Self-complementary sequences may occur in the RNA molecules which produce more complex structures.
- So RNA can base-pair with complementary regions of either RNA or DNA.
- RNA has no any regular secondary structure that serves as a reference point. The three-dimensional structures of many RNAs are complex and unique.
- Breaks in the helix caused by mismatched or unmatched bases in one or both strands are common and result in bulges or internal loops.

- Hairpin loops form between nearby self-complementary sequences.



Transfer RNA (tRNA)



Modul42: Messenger RNA

Text (9)

1. The protein-coding region(s) of each mRNA is composed of a contiguous, non-overlapping string of codons called an open reading frame (commonly known as an ORF).
2. Each ORF specifies a single protein and starts and ends at internal sites within the mRNA. That is, the ends of an ORF are distinct from the ends of the mRNA
3. Translation starts at the 5' end of the ORF and proceeds one codon at a time to the 3' end. The first and last codons of an ORF are known as the start and stop codons
4. In bacteria, the start codon is usually 5'-AUG-3', but 5'-GUG-3' and sometimes even 5'-UUG-3' are also used.
5. Eukaryotic cells always use 5'-AUG-3' as the start codon
6. The start codon has two important functions.
    a. First, it specifies the first amino acid to be incorporated into the growing polypeptide chain.

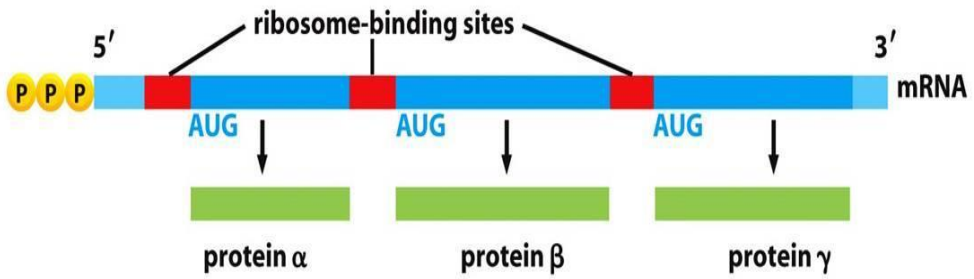
b. Second, it defines the reading frame for all subsequent codons

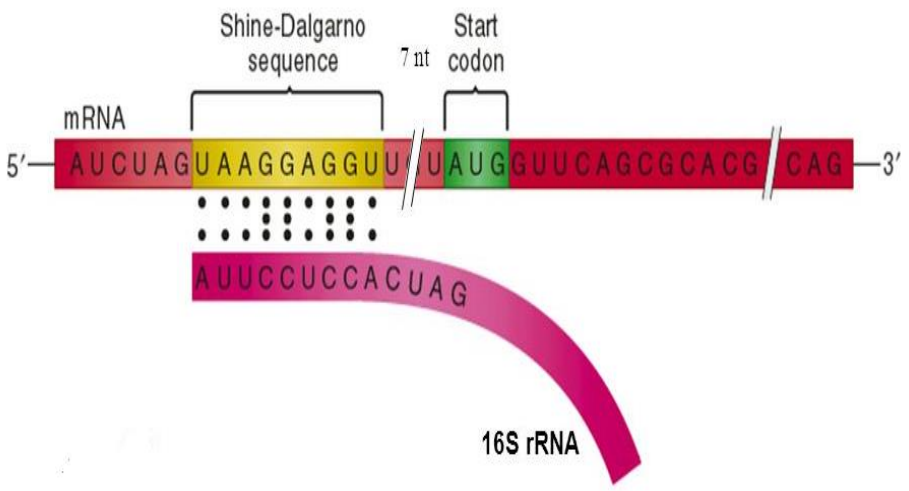
7. Stop codons, of which there are three (5'-UAG-3', 5'-UGA-3', and 5'-UAA-3'), define the end of the ORF and signal termination of polypeptide synthesis
8. You can now understand the origin of the term open reading frame. It is a contiguous stretch of codons "read" in a particular frame (as set by the first codon) that is "open" to translation because it lacks a stop codon
9. mRNAs contain at least one ORF. The number of ORFs per mRNA is different between eukaryotes and prokaryotes. Eukaryotic mRNAs almost always contain a single ORF
10. In contrast, prokaryotic mRNAs frequently contain two or more ORFs. mRNAs containing multiple ORFs are known as polycistronic RNAs and those encoding a single ORF are known as monocistronic RNAs
11. The polycistronic mRNAs found in bacteria often encode proteins that perform related functions, such as different steps in the biosynthesis of an amino acid or nucleotide

Module43:Prokaryotic mRNA

Text (09:00)

1. For translation to occur, the ribosome must be recruited to the mRNA. Prokaryotic mRNAs have a ribosome-binding site that recruits the translational machinery
2. To facilitate binding by a ribosome, many prokaryotic ORFs contain a short sequence upstream (on the 5' side) of the start codon called the ribosome-binding site (RBS)
3. This element is also referred to as a Shine Őalgarno sequence after the scientists who discovered it by comparing the sequences of multiple mRNAs
4. The extent of complementarity and the spacing between the RBS and the start codon has a strong influence on how actively a particular ORF is translated
5. Some prokaryotic ORFs lack a strong RBS but are nonetheless actively translated. These ORFs are not the first ORF in an mRNA but instead are located just after another ORF in a polycistronic message
6. phenomenon of linked translation between overlapping ORFs is known as translational coupling. So in this situation translation of the downstream ORF requires translation of the upstream ORF





Module44: Eukaryotic mRNA

Text (13:00)

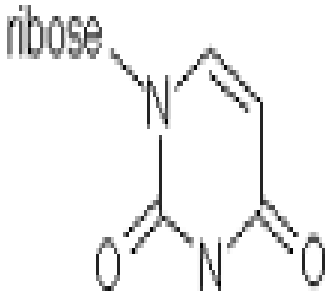
1. Unlike their prokaryotic counterparts, eukaryoticmRNAs recruit ribosomes using a specific chemical modification called the 5’ cap, which is located at the extreme 5’ end of the mRNA
2. The 5’ cap is a methylated guanine nucleotide that is joined to the 5’ end of the mRNA via an unusual 5’-to-5’ linkage
3. Two other features of eukaryotic mRNAs stimulate translation. One feature is the presence, in some mRNAs, of a purine three bases upstream of the start codon and a guanine immediately downstream
4. A second feature that contributes to efficient translation is the presence of a poly-A tail at the extreme 3’ end of the mRNA. This tail is added enzymatically by the enzyme poly-A polymerase
5. Despite its location at the 3’ end of the mRNA, the poly-A tail enhances the level of translation of them RNA by enhancing the recruitment of key translation initiation factors
6. Importantly, in addition to their roles in translation, these 5’- and 3’-end modifications also protect eukaryotic mRNAs from rapid degradation

Module45:Transfer RNA

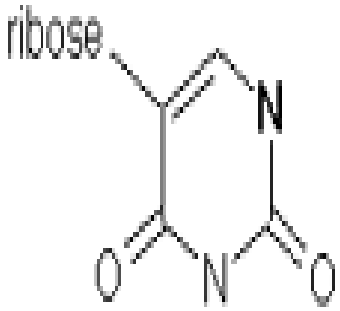
Text (08:00)

1. The heart of protein synthesis is the “translation” of nucleotide sequence information (in the form of codons) into amino acids
2. This is accomplished by tRNA molecules, which act as adaptors between codons and the amino acids they specify
3. There are many types of tRNA molecules, but each is attached to a specific amino acid, and each recognizes a particular codon, or codons, in the mRNA (most tRNAs recognize more than one codon).
4. A striking aspect of tRNAs is the presence of several unusual bases in their primary structure. These unusual features are created post- transcriptionally by enzymatic modification of normal bases in the polynucleotide chain
5. Unusual bases found in tRNA include hypoxanthine, thymine, and methylguanine. These

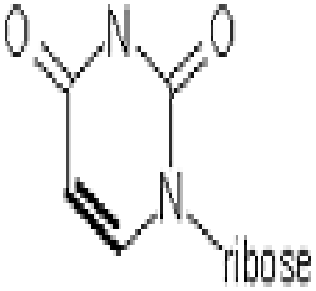
modified bases are not essential for tRNA function, but cells lacking these modified bases show reduced rates of growth



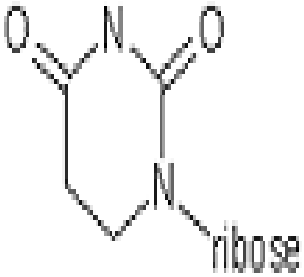
uridine



Pseudouridine



uridine



dihydrouridine

Module46:Secondary Structure of tRNA

Text (9:00)

1. RNA molecules typically contain regions of self complementarity that enable them to form limited stretches of double helix that are held together by base pairing
2. tRNA molecules show a characteristic and highly conserved pattern of single-stranded and double stranded regions (secondary structure) that can be illustrated as a cloverleaf
3. The principal features of the tRNA cloverleaf are an acceptor stem, three stem-loops (referred to as the yU loop, the D loop, and the anticodon loop), and a fourth variable loop
4. The Acceptor Stem: It is so-named because it is the site of attachment of the amino acid, is formed by pairing between the 5’ and 3’ ends of the tRNA molecule
5. The yU Loop: It is is so-named because of the characteristic presence of the unusual base yU in the loop. The modified base is often found within the sequence 5’-TCUCG-3’
6. The D Loop: It takes its name from the characteristic presence of dihydrouridines in the loop
7. The Anticodon Loop: As its name implies, contains the anticodon, a three-nucleotide-long sequence that is responsible for recognizing the codon by base pairing with the mRNA
8. The Variable Loop: It sits between the anticodon loop and the yU loop and, as its name implies, varies in size from 3 to 21 bases

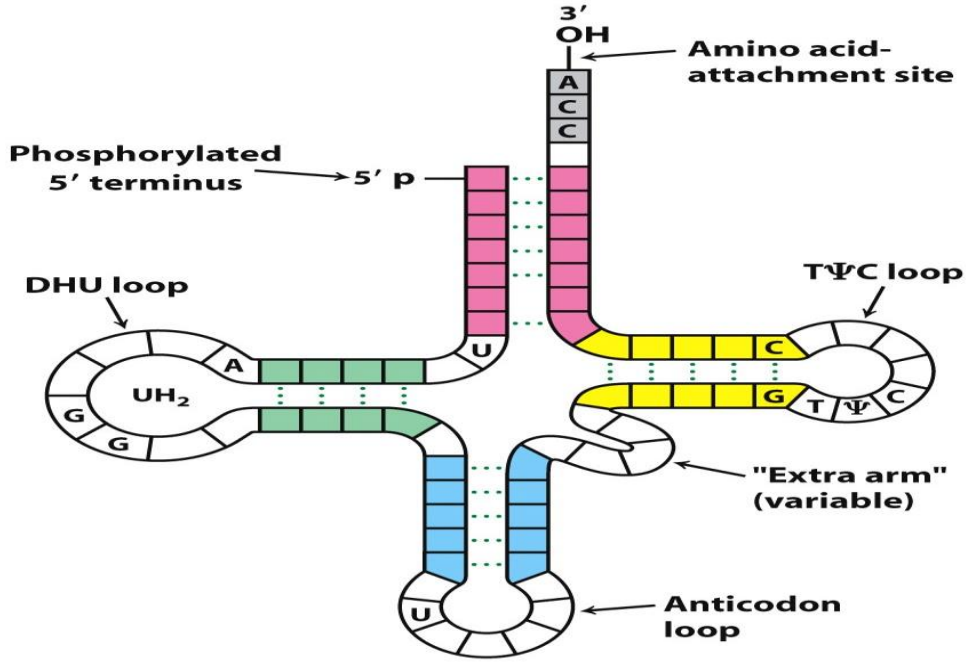


Figure 30.3
Biochemistry, Seventh Edition

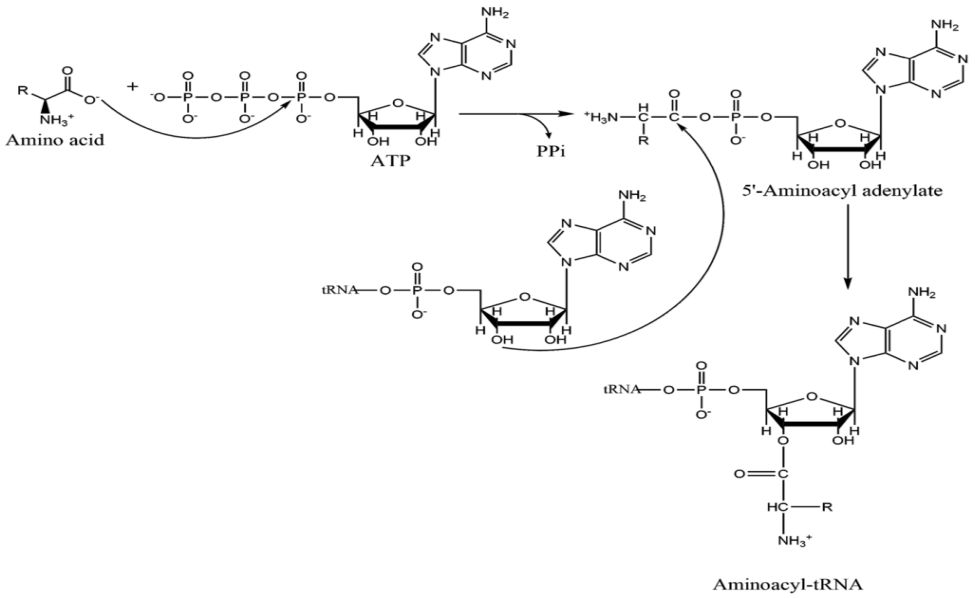
Module47: Attachment of Amino acid to tRNAs

Text (9:00)

1. tRNA molecules to which an amino acid is attached are said to be charged, and tRNAs that lack an amino acid are said to be uncharged
2. Charging requires an acyl linkage between the carboxyl group of the amino acid and the 2'- or 3'-hydroxyl group of the adenosine nucleotide that protrudes from the acceptor stem at the 3' end of the tRNA
3. This acyl linkage is a high-energy bond because its hydrolysis results in a large change in free energy
4. This is significant for protein synthesis: the energy released when this acyl bond is broken is coupled to the formation of the peptide bonds that link amino acids to each other in polypeptide chains
5. All aminoacyl-tRNA synthetases attach an amino acid to a tRNA in two enzymatic steps:
    a. Adenylation
    b. tRNA charging
6. Step one is adenylation in which the amino acid reacts with ATP to become adenylylated with the concomitant release of pyrophosphate
    a. Adenylylation refers to transfer of AMP, as opposed to adenylation, which would indicate the transfer of adenine
    b. The principal driving force for the adenylylation reaction is the subsequent hydrolysis of pyrophosphate by pyrophosphatase
    c. As a result of adenylylation, the amino acid is attached to adenylic acid via a high-energy ester bond in which the carbonyl group of the amino acid is joined to the phosphoryl group of AMP
7. Step two is tRNA Charging in which the adenylylated amino acid, which remains tightly bound to the synthetase, reacts with tRNA
    a. This reaction results in the transfer of the amino acid to the 3' end of the tRNA via the 2'- or 3'-hydroxyl and the release of AMP

b. There are two classes of tRNA synthetases:
   i. Class I enzymes attach the amino acid to the 20-OH of the tRNA and are generally monomeric
   ii. Class II enzymes attach the amino acid to the 30-OH of the tRNA and are typically dimeric or tetrameric

8. Each of the 20 amino acids is attached to the appropriate tRNA by a single, dedicated tRNA synthetase
9. Because most amino acids are specified by more than one codon, it is not uncommon for one synthetase to recognize and charge more than one tRNA (known as isoaccepting tRNAs)
10. the same tRNA synthetase is responsible for charging all tRNAs for a particular amino acid.
11. Thus, one and only one tRNA synthetase attaches each amino acid to all of the appropriate tRNAs



Module48: The Ribosomes

Text (10:00)

- The ribosome is the macromolecular machine that directs the synthesis of proteins.
- The ribosome is larger and more complex than the minimal machinery required for DNA or RNA synthesis.
- The machinery for polymerizing amino acids is composed of at least three RNA molecules and more than 50 different proteins, with an overall molecular mass of >2.5 MDa.
- Compared with the speed of DNA replication i.e., 200 –1000 nucleotides per second; translation takes place at a rate of only two to 20 amino acids per second.
- In prokaryotes, the transcription machinery and the translation machinery are located in the same compartment. Thus, the ribosome can commence translation of the mRNA as it emerges from the RNA polymerase.
- This situation allows the ribosome to proceed in tandem with the RNA polymerase as it elongates the transcript.
- Recall that the 5’ end of an RNA is synthesized first, and thus the ribosome, which begins translation at the 5’ end of the mRNA, can start translating a nascent transcript as soon as it emerges from the RNA polymerase.

Module 49: Structure Of Peptide Bond

Text (11:00)

1. Each new amino acid is added to the carboxyl terminus of the growing polypeptide chain (often referred to as synthesis in the amino- to carboxy-terminal direction)
2. The ribosome catalyzes a single chemical reaction —the formation of a peptide bond
3. This reaction occurs between the amino acid residue at the carboxy-terminal end of the growing polypeptide and the incoming amino acid to be added to the chain
4. Both the growing chain and the incoming amino acid are attached to tRNAs; as a result, during peptide-bond formation, the growing polypeptide is continuously attached to a tRNA
5. The actual substrates for each round of amino acid addition are two charged species of tRNAs —an aminoacyl-tRNA and a peptidyl-tRNA
6. aminoacyl-tRNA is attached at its 3' end to the carboxyl group of the amino acid. The peptidyl-tRNA is attached in exactly the same manner (at its 3' end) to the carboxyl terminus of the growing polypeptide chain
7. The bond between the aminoacyl-tRNA and the amino acid is not broken during the formation of the next peptide bond
8. Instead, the bond between the peptidyl-tRNA and the growing polypeptide chain is broken as the growing chain is attached to the amino group of the amino acid attached to the aminoacyl-tRNA to form a new peptide bond
9. To catalyze peptide-bond formation, the 3' ends of these two tRNAs are brought into close proximity by the ribosome
10. The resulting tRNA positioning allows the amino group of the amino acid attached to aminoacyl-tRNA to attack the carbonyl group of the most carboxy-terminal amino acid attached to the peptidyl-tRNA
11. The result of this nucleophilic attack is the formation of a new peptide bond between the amino acids attached to the tRNAs and the release of the polypeptide chain from the peptidyl tRNA. There are two consequences of this method of polypeptide synthesis.
    a. First, this mechanism of peptide-bond formation requires that the amino terminus of the protein be synthesized before the carboxyl terminus
    b. Second, the growing polypeptide chain is transferred from the peptidyl-tRNA to the aminoacyl-tRNA. For this reason, the reaction to form a new peptide bondis called the peptidyl transferase reaction
12. Interestingly, peptide-bond formation takes place without the simultaneous hydrolysis of a nucleoside triphosphate
13. This is because peptide-bond formation is driven by breaking the high-energy acyl bond that joins the growing polypeptide chain to the tRNA

Module50: Bonding site on the Ribosomes on tRNA

Text (9:00)

1. The ribosome is composed of two subassemblies of RNA and protein known as the large and small subunits
2. The large subunit contains the peptidyl transferase center, which is responsible for the formation of peptide bonds
3. The small subunit contains the decoding center in which charged tRNAs read or "decode" the codon units of the mRNA
4. Both the decoding center and the peptidyl transferase center are buried within the intact ribosome
5. Yet, mRNA must be threaded through the decoding center during translation, and the nascent

polypeptide chain must escape from the peptidyl transferase center

6. There are "tunnels" in and out of the ribosome. polymers enter through these tunnels
7. To perform the peptidyl transferase reaction, the ribosome must be able to bind at least two tRNAs simultaneously.
8. In fact, the ribosome contains three tRNA-binding sites, called the A-, P-, and E-sites
    a. The A-site is the binding site for the aminoacylated-tRNA
    b. the P-site is the binding site for the peptidyl-tRNA
    c. The E-site is the binding site for the tRNA that is released after the growing polypeptide chain has been transferred to the aminoacyl-tRNA (E is for "exiting").
9. Each tRNA binding site is formed at the interface between the large and the small subunits of the ribosome
10. In this way, the bound tRNAs can span the distance between the peptidyl transferase center in the large subunit and the decoding center in the small subunit
11. The 3' ends of the tRNAs that are coupled to the amino acid or to the growing peptide chain are adjacent to the large subunit.
12. The anticodon loops of the bound tRNAs are located adjacent to the small subunit

Module51:Initiation of Translation

Text (10:00)

- For translation to be successfully initiated, three events must occur:-
  i) the ribosome must be recruited to the mRNA.
  ii) a charged tRNA must be placed into the P-site of the ribosome.
  iii) the ribosome must be precisely positioned over the start codon.
- The correct positioning of the ribosome over the start codon is critical because this establishes the reading frame for the translation of the mRNA.
- In prokaryotes, the assembly of the ribosome on an mRNA occurs one subunit at a time. The small subunit associates with the mRNA first.
- For ideally positioned RBSs, the small subunit is positioned on the mRNA such that the start codon will be in the P-site when the large subunit joins the complex.
- The large subunit joins its partner only at the very end of the initiation process, just before the formation of the first peptide bond.
- Thus, many of the key events of translation initiation occur in the absence of the full ribosome.
- Translation initiation is the only time a tRNA binds to the P-site without previously occupying the A-site. This event requires a special tRNA known as the initiator tRNA.
- The initiator tRNA base-pairs with the start codon (AUG or GUG). AUG and GUG have a different meaning when they occur within an ORF, where they are read by tRNAs for methionine and valine, respectively.
- Although the initiator tRNA is first charged with a methionine, a formyl group is rapidly added to the methionine amino group by a separate enzyme (Met-tRNA transformylase).
- Thus rather than valine or methionine, the initiator tRNA is coupled to N-formyl methionine. The charged initiator tRNA is referred to as fMet-tRNA^{fMet}.
- Because N-formyl methionine is the first amino acid to be incorporated into a polypeptide chain, one might think that all prokaryotic proteins have a formyl group at their amino

termini.
This is not the case, however, because an enzyme known as a deformylase removes the formyl group from the amino terminus during or after the synthesis of the polypeptide chain.

- In fact, many mature prokaryotic proteins do not even start with a methionine; aminopeptidases often remove the amino-terminal methionine as well as one or two additional amino acids.

Module52:The Initiation factor

Text (7:00)

1. The initiation of prokaryotic translation commences with the small subunit and is catalyzed by three translation initiation factors called IF1, IF2, and IF3. Each factor facilitates a key step in the initiation process
2. **IF1:** It prevents tRNAs from binding to the portion of the small subunit that will become part of the A-site.
3. **IF2:** It is a GTPase that interacts with three key components of the initiation machinery: the small subunit, IF1, and the charged initiator tRNA (fMet-tRNA ^{fMet}).
4. **IF3:** It binds to the small subunit and blocks it from re-associating with a large subunit. Because initiation requires a free small subunit, the binding of IF3 is critical for a new cycle of translation
5. Each of the initiation factors binds at, or near, one of the three tRNA binding sites on the small subunit
6. From the three potential tRNA-binding sites on the small subunit, only the P-site is capable of binding a tRNA in the presence of the initiation factors
7. With all three initiation factors bound, the small subunit is prepared to bind to the mRNA and the initiator tRNA . These two RNAs can bind in either order and independently of each other
8. The last step of initiation involves the association of the large subunit to create the 70S initiation complex.
9. When the start codon and fMet-tRNA^{fMet} base-pair, the small subunit undergoes a change in conformation
10. IF2 bound to GDP has reduced affinity for the ribosome and the initiator tRNA, leading to the release of IF2.GDP as well as IF1 from the ribosome

Module53: Translation elongation

Text (9:00)

1. Once the ribosome is assembled with the charged initiator tRNA in the P site, polypeptide synthesis can begin. There are three key events that must occur for the correct addition of each amino acid:
    a. First, the correct aminoacyl-tRNA is loaded into the A site of the ribosome as dictated by the A-site codon
    b. Second, a peptide bond is formed between the aminoacyl-tRNA in the A site and the peptide chain that is attached to the peptidyl-tRNA in the P site. This peptidyl transferase reaction results in the transfer of the growing polypeptide from the tRNA

in the P site to the amino acid moiety of the charged tRNA in the A site

c. Third, the resulting peptidyl-tRNA in the A site and its associated codon must be translocated to the P site so that the ribosome is poised for another cycle of codon recognition and peptide bond formation

2. As with the original positioning of the mRNA, this shift must occur precisely to maintain the correct reading frame of the message.
3. Two auxiliary proteins known as elongation factors control these events.
4. Both of these factors use the energy of GTP binding and hydrolysis to enhance the rate and accuracy of ribosome function
5. Unlike the initiation of translation, the mechanism of elongation is highly conserved between prokaryotic and eukaryotic cells
6. Aminoacyl-tRNAs do not bind to the ribosome on their own. Instead, they are "escorted " to the ribosome by the elongation factor EF-Tu.
7. Like the initiation factor IF2, the elongation factor EF-Tu binds and hydrolyzes GTP and the type of guanine nucleotide bound governs its function
8. The trigger that activates the EF-Tu GTPase is the same domain on the large subunit of the ribosome that activates the IF2 GTPase when the large subunit joins the initiation complex. This domain is known as the factor binding center
9. EF-Tu only interacts with the factor binding center after the tRNA is loaded into the A site and a correct codon-anticodon match is made
10. The error rate of translation is between 10^{-3} to 10^{-4}.
11. The ultimate basis for the selection of the correct aminoacyl-tRNA is the base pairing between the charged tRNA and the codon displayed in the A site of the ribosome.
12. However, in some cases, the base pairing in the anticodon-codon interaction may be mismatched, yet the ribosome rarely allows such mismatched aminoacyl-tRNAs to continue in the translation process

Module54:The Ribosome is a ribozyme

Text (12:00)

Once the correctly charged tRNA has been placed in the A site and has rotated into the peptidyl transferase center, peptide bond formation takes place.

This reaction is catalyzed by RNA, specifically the 23 S rRNA component of the large subunit.

- Early evidence for this came from experiments in which it was shown that a large subunit that had been largely stripped of its proteins was still able to carry out peptide bond formation.
- Proof that the peptidyl transferase is entirely composed of RNA has come from the high-resolution, three-dimensional structure of the ribosome, which reveals that no amino acid is located closer than 18 Aº from the active site.
- Because catalysis requires distances in the 1 - 3 Aº range, it is clear that the peptidyl transferase center is a ribozyme. That is an enzyme composed of RNA.
- How does the 23 S rRNA catalyze peptide bond formation?
- The exact mechanism remains to be determined, but some answers to this question are beginning to emerge

- First, base-pairing between the 23 S rRNA and the CCA ends of the tRNAs in the A and the P sites help to position the alpha-amino group of the aminoacyl-tRNA to attack the carbonyl group of the growing polypeptide attached to the peptidyl-tRNA.
- These interactions are also likely to stabilize the aminoacyl-tRNA after accommodation.
- Because close proximity of substrates is rarely sufficient to generate high levels of catalysis, it is hypothesized that other elements of the ribosomal RNA change the chemical environment of the peptidyl transferase active site.
- For example, it has been proposed that nucleotides in the peptidyl transferase center accept a hydrogen from the alpha amino group of the aminoacyl-tRNA, making the associated nitrogen a stronger nucleophile
- This is a common mechanism used by many proteins to stimulate nucleophilic attack of carbonyl groups.
-

Module55:The Translation in the large subunits

Text (10:00)

1. Once the peptidyl transferase reaction has occurred, the tRNA in the P-site is deacetylated (no longer attached to an amino acid), and the growing polypeptide chain is linked to the tRNA in the A-site
2. For a new round of peptide chain elongation to occur, the P-site tRNA must move to the E-site and the A-site tRNA must move to the P-site. At the same time, the mRNA must move by three nucleotides to expose the next codon
3. These movements are coordinated within the ribosome and are collectively referred to as translocation.
4. The initial steps of translocation are coupled to the peptidyl transferase reaction
5. Once the growing peptide chain has been transferred to the A-site tRNA, the A- and P-site tRNAs have a preference to occupy new positions in the large subunit
6. The 3’ end of the A-site tRNA is bound to the growing polypeptide chain and prefers to bind in the P-site of the large subunit
7. The now deacetylated P-site tRNA is no longer attached to the growing polypeptide chain and prefers to bind in the E-site of the large subunit
8. In contrast, at this time, the anticodons of these tRNAs remain in their initial location in the small subunit bound to the mRNA
9. Translocation is initiated in the large subunit before the small subunit, and the tRNAs are said to be in “hybrid states.”
10. The completion of translocation requires the action of a second elongation factor called EF-G. Initial binding of EF-G to the ribosome occurs when associated with GTP
11. After the peptidyl transferase reaction, EF-G–GTP binds to and stabilizes the ribosome in the rotated, hybrid state
12. When EF-G–GTP binds, it contacts the factor-binding center of the large subunit, which stimulates GTP hydrolysis. GTP hydrolysis changes the conformation of EF-G with two consequences
    a. First, interactions between EF-G–GDP and the ribosome are thought to “unlock” the ribosome
    b. Second, the changed EF-G–GDP conformation binds to the A-site of the decoding

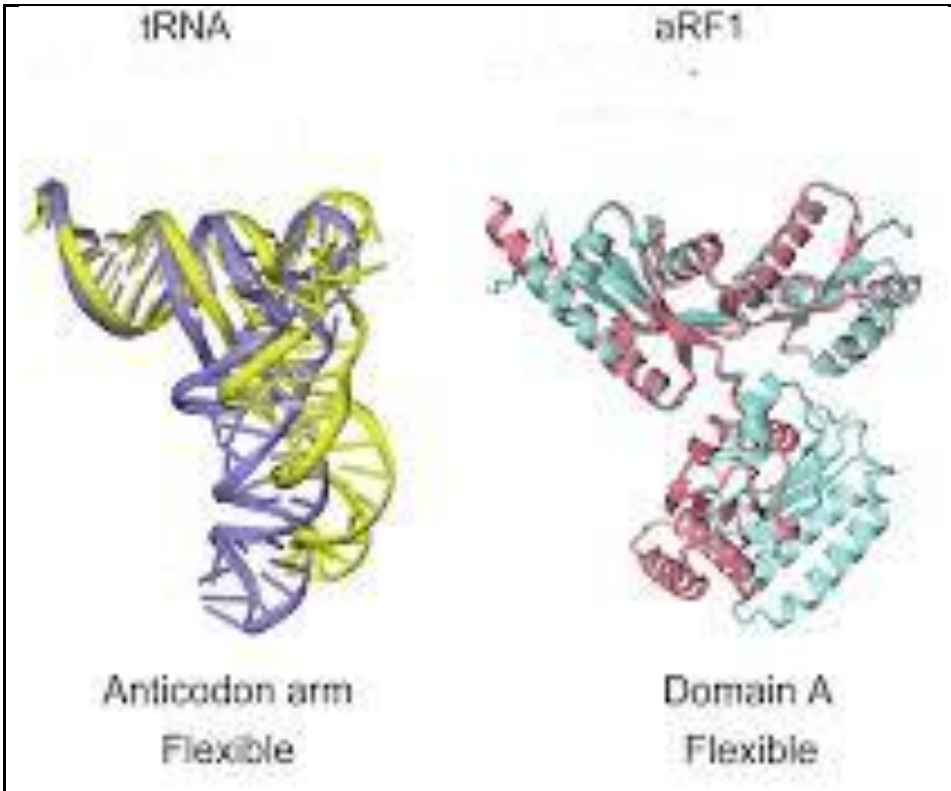
center. This interaction competes with the tRNA for binding to the A-site of the decoding center

13. Release of EF-G results in the return of the ribosome to a “locked” state in which the tRNAs and mRNA are once again tightly associated with the small subunit decoding center and the gates between the A-, P- and E-sites are closed
14. Together, these events result in the translocation of the A-site tRNA into the P-site, the P-site tRNA into the E-site, and the movement of the mRNA by exactly 3 bp. The ribosome is now ready for a new cycle of amino acid addition to begin.

Module56:Termination of the Translation

Text (10:00)

1. The ribosome’s cycle of aminoacyl-tRNA binding, peptide-bond formation, and translocation continues until one of the three stop codons enters the A-site
2. Stop codons are recognized by proteins called release factors (RFs) that activate the hydrolysis of the polypeptide from the peptidyl-tRNA. There are two classes of release factors.
    a. Class I release factors recognize the stop codons and trigger hydrolysis of the peptide chain from the tRNA in the P-site
    b. Class II release factors stimulate the dissociation of the class I factors from the ribosome after release of the polypeptide chain
3. Prokaryotes have two class I release factors called RF1 and RF2.
4. RF1 recognizes the stop codon UAG and RF2 recognizes the stop codon UGA.
5. The third stop codon, UAA, is recognized by both RF1 and RF2
6. In eukaryotic cells, there is a single class I release factor called eRF1 that recognizes all three stop codons
7. Prokaryotes and eukaryotes have only one class II factor called RF3 and eRF3, respectively.
8. Like EF-G, IF2, and EF-Tu, class II release factors are regulated by GTP binding and hydrolysis
9. Release factors are composed entirely of protein, protein–RNA interaction must mediate stop codon recognition



Module57: Termination of the translation part 2

Text (9:00)

1. The ribosome's cycle of aminoacyl-tRNA binding, peptide-bond formation, and translocation continues until one of the three stop codons enters the A-site
2. Stop codons are recognized by proteins called release factors (RFs) that activate the hydrolysis of the polypeptide from the peptidyl-tRNA. There are two classes of release factors.
    a. Class I release factors recognize the stop codons and trigger hydrolysis of the peptide chain from the tRNA in the P-site
    b. Class II release factors stimulate the dissociation of the class I factors from the ribosome after release of the polypeptide chain
3. Prokaryotes have two class I release factors called RF1 and RF2.
4. RF1 recognizes the stop codon UAG and RF2 recognizes the stop codon UGA.
5. The third stop codon, UAA, is recognized by both RF1 and RF2
6. In eukaryotic cells, there is a single class I release factor called eRF1 that recognizes all three stop codons
7. Prokaryotes and eukaryotes have only one class II factor called RF3 and eRF3, respectively.
8. Like EF-G, IF2, and EF-Tu, class II release factors are regulated by GTP binding and hydrolysis
9. Release factors are composed entirely of protein, protein–RNA interaction must mediate stop codon recognition

Module58: AMINO ACID

Text (08:00)

1. **BACKGROUND**

    RNA decodes the information at ribosomes in form of codons. Each codon consists of three

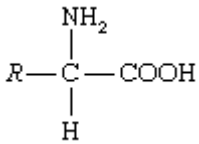
nucleotides which forms one amino acid.

Table 4.1.1: Different combinations of codons

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC } Phe UUA, UUG } Leu	UCU, UCC, UCA, UCG } Ser	UAU, UAC } Tyr UAA Stop UAG Stop	UGU, UGC } Cys UGA Stop UGG Trp	U C A G
C	CUU, CUC, CUA, CUG } Leu	CCU, CCC, CCA, CCG } Pro	CAU, CAC } His CAA, CAG } Gln	CGU, CGC, CGA, CGG } Arg	U C A G
A	AUU, AUC, AUA } Ile AUG Met	ACU, ACC, ACA, ACG } Thr	AAU, AAC } Asn AAA, AAG } Lys	AGU, AGC } Ser AGA, AGG } Arg	U C A G
G	GUU, GUC, GUA, GUG } Val	GCU, GCC, GCA, GCG } Ala	GAU, GAC } Asp GAA, GAG } Glu	GGU, GGC, GGA, GGG } Gly	U C A G

2. INTRODUCTION

amino acid, any of a group of organic molecules that consist of a basic amino group (—NH_2), an acidic carboxyl group (—COOH), and an organic *R* group (or side chain) that is unique to each amino acid. The term amino acid is short for α-amino [alpha-amino] carboxylic acid. Each molecule contains a central carbon (C) atom, called the α-carbon, to which both an amino and a carboxyl group are attached. The remaining two bonds of the α-carbon atom are generally satisfied by a hydrogen (H) atom and the *R* group. The formula of a general amino acid is:



The amino acids differ from each other in the particular chemical structure of the *R* group.

There are 20 different amino acids in nature therefore they fold together and make a protein structure by polymerizing themselves. If we observe the structure of amino acid it contains nitrogen, hydrogen, oxygen and two carbon atoms. And a variable group R (alkyl group).

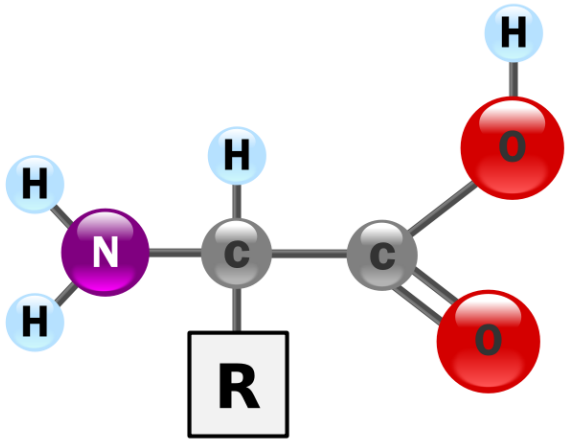


Figure 4.1.1: Structure of amino acid (Courtesy Wikipedia)

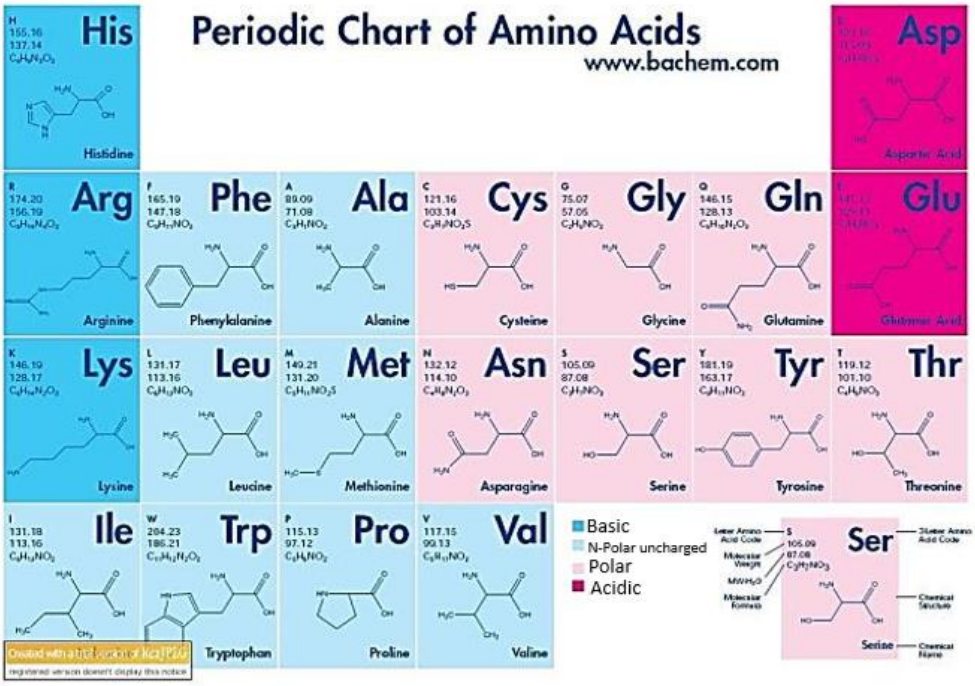


Figure 4.1.2: Periodic chart of amino acids

3. MECHANISM

When polymerizations take place then, water is formed. On the removal of water, peptide bond formed between two amino acids. Long chain of peptide bonds formed protein.

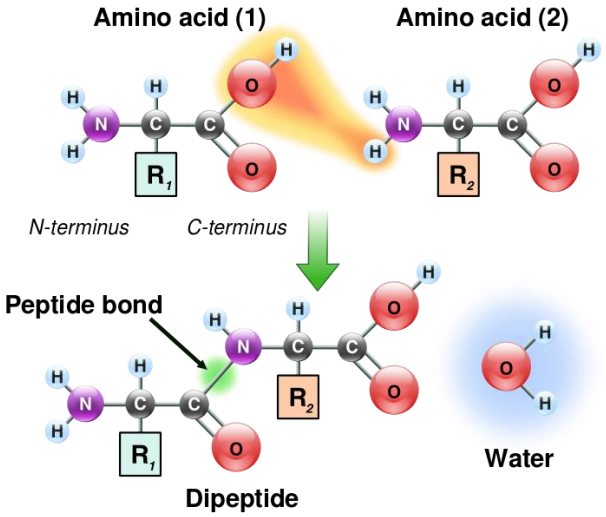


Figure 4.1.3: Structure of amino acid (Courtesy Wikipedia)

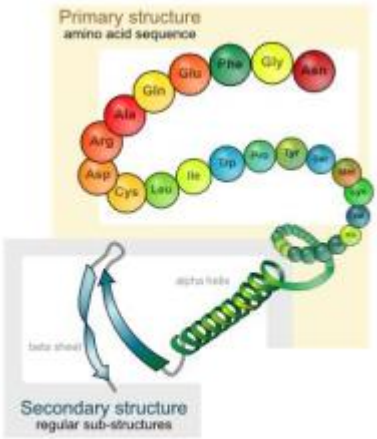


Figure 4.1.4: Polymerization of amino acids (Courtesy Wikipedia)

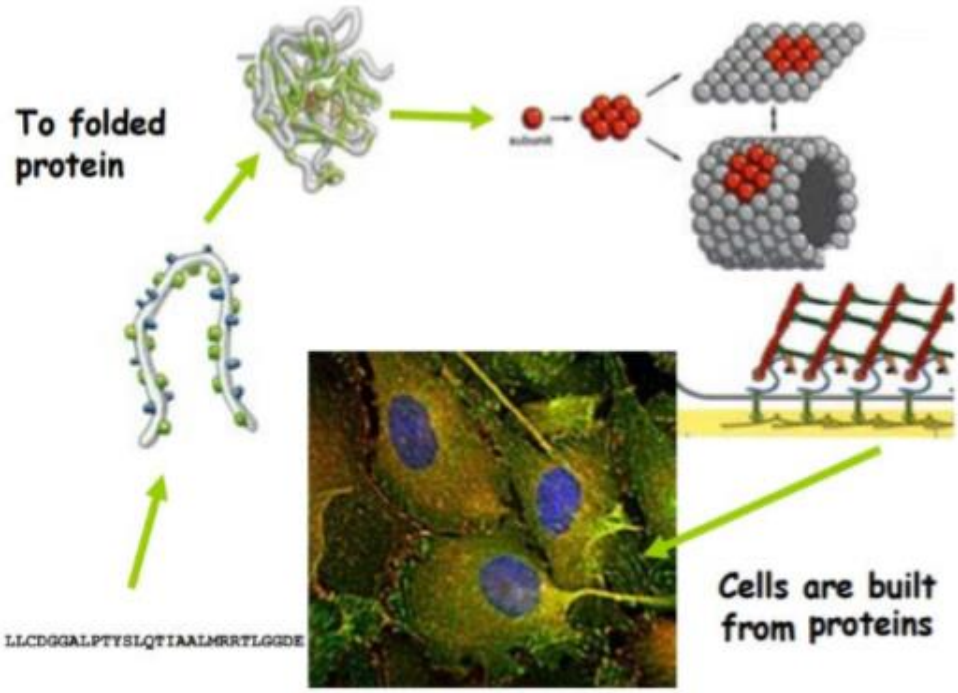


Figure 4.2.5: Primary structure of protein to cell formation (Courtesy Wikipedia)

4. CONCLUSION

Amino acids polymerized and form peptide bonds also known as peptide linkage. Then, proteins formed. Protein fold itself in 3-D structure to acquire more stability. And these proteins formed cells with the collaboration of other biomolecules.

Module59:Proteins:Polymers of amino acid

Text (14:00)

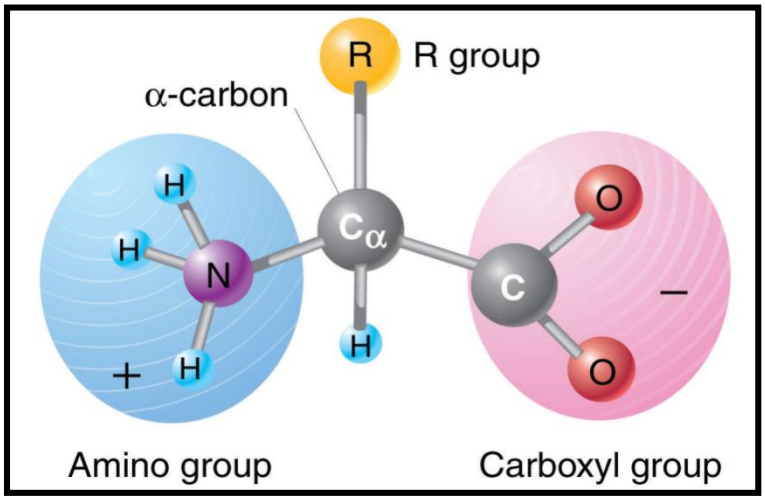
Protein Functions:

Most Abundant Polymer in the cell

Support, protection, catalysis, transport, defense, regulation & movement

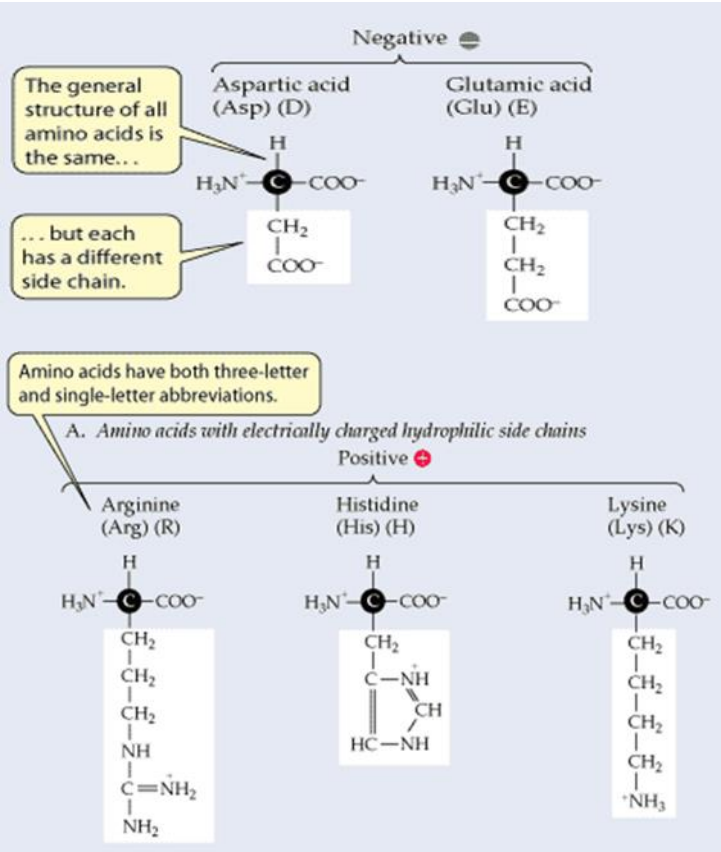
Do not store genetic information

Amino Acid Structure

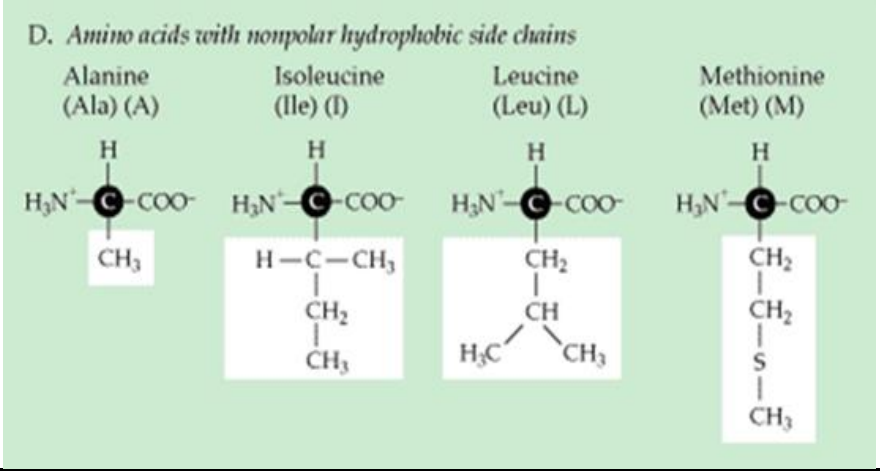


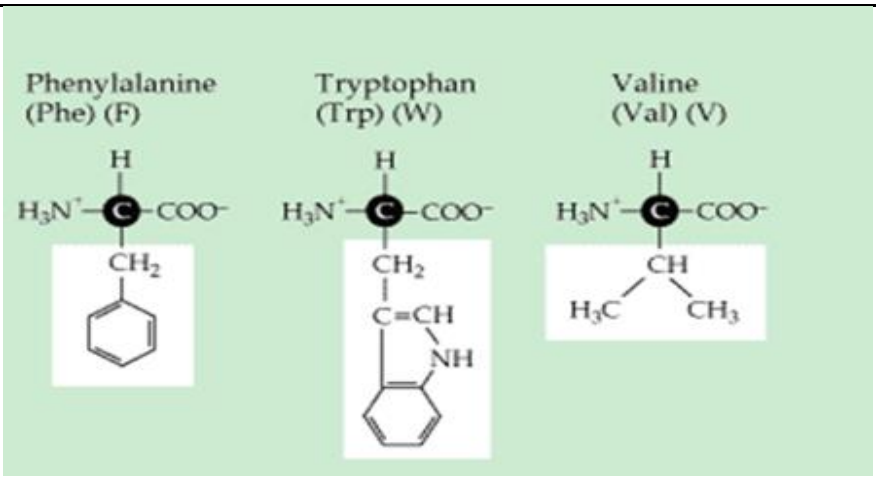
The side chains or R groups give different properties to each of the amino acids

Charged R Groups

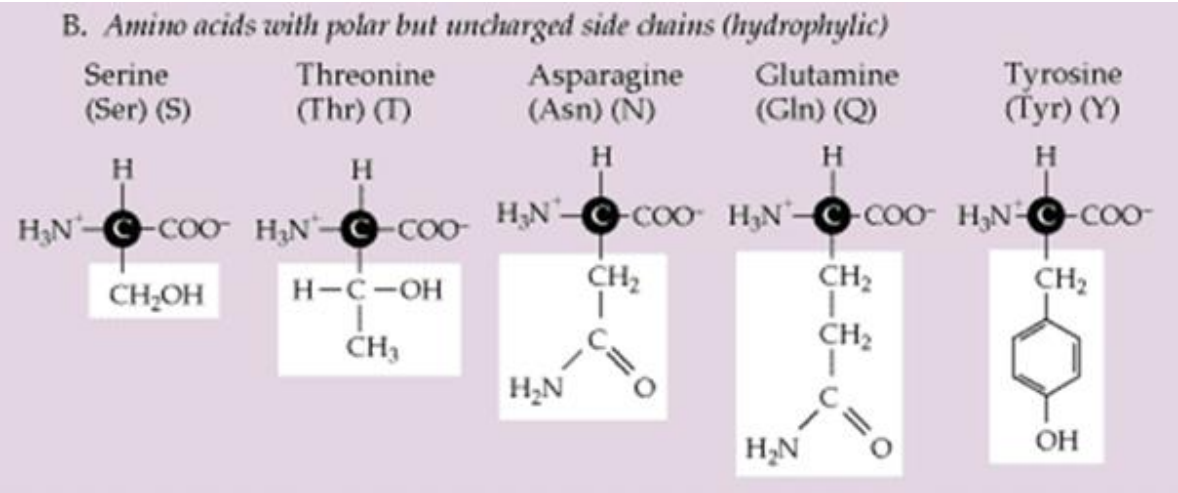


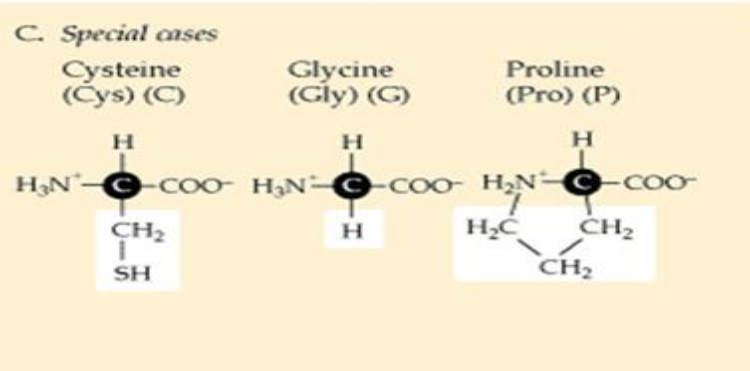
Nonpolar R Groups





Polar & Special R Groups

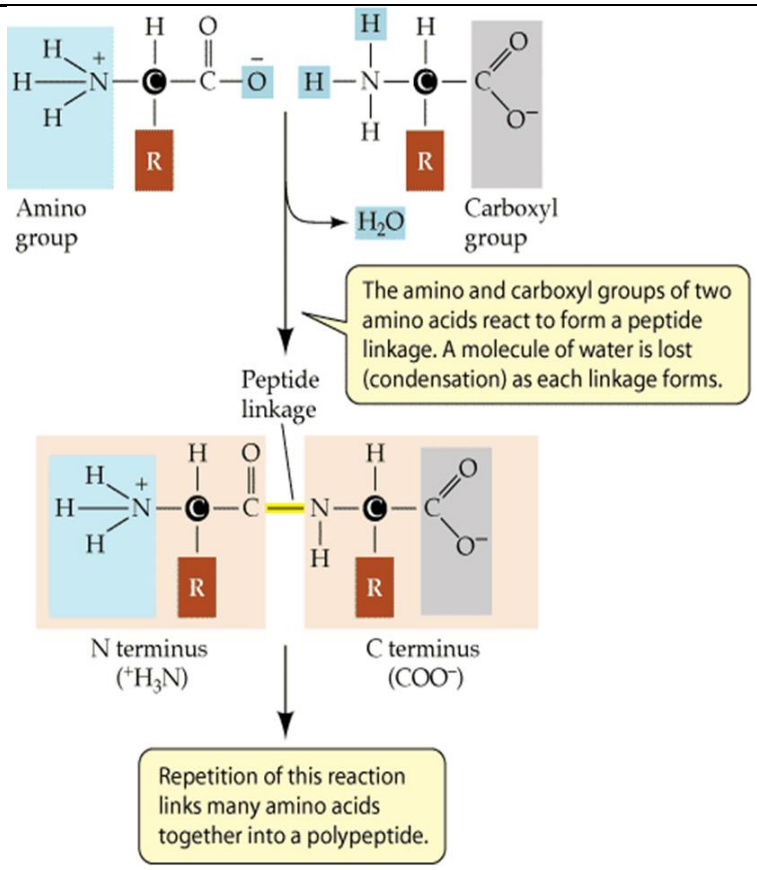




Module60:Protein structure

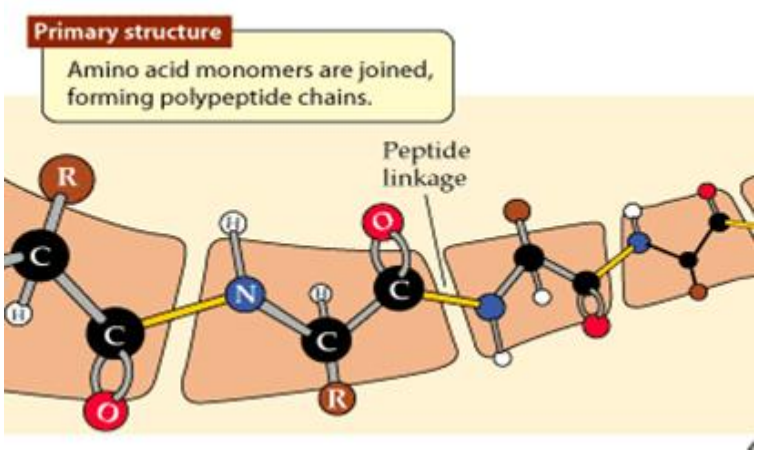
Text (10:00)

The Peptide Bond



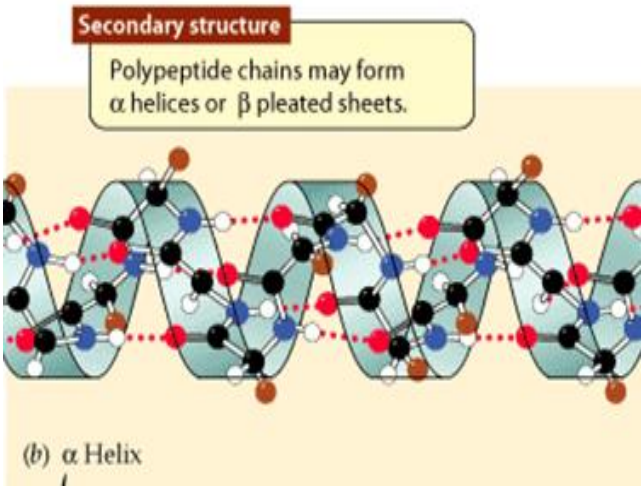
Primary Structure

Sequence of amino acids bonded by peptide linkages (Diversity 20n)



Secondary Structure

α helices (hydrogen bonds between amino acid residues)



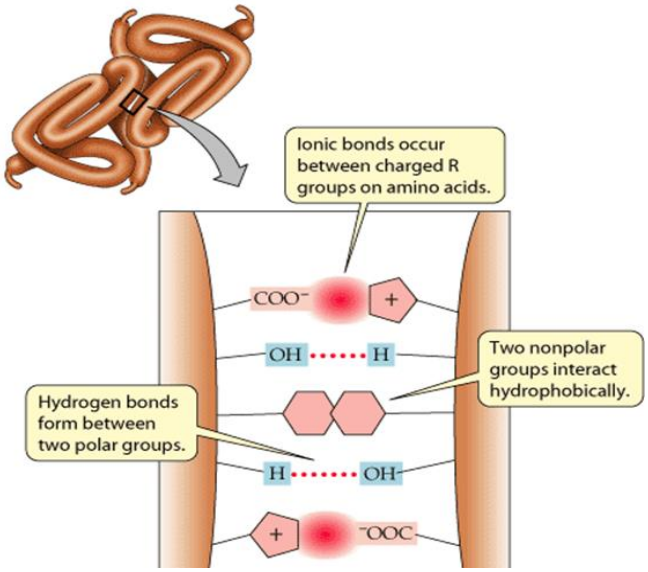
Module61: Protein interaction

Text (10:00)

Protein Interaction:

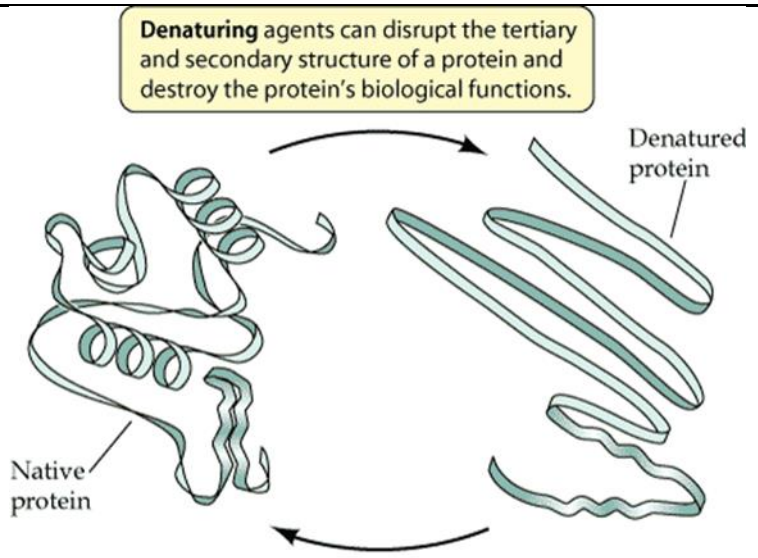
Weak interactions

3D structure



Protein: Denaturation

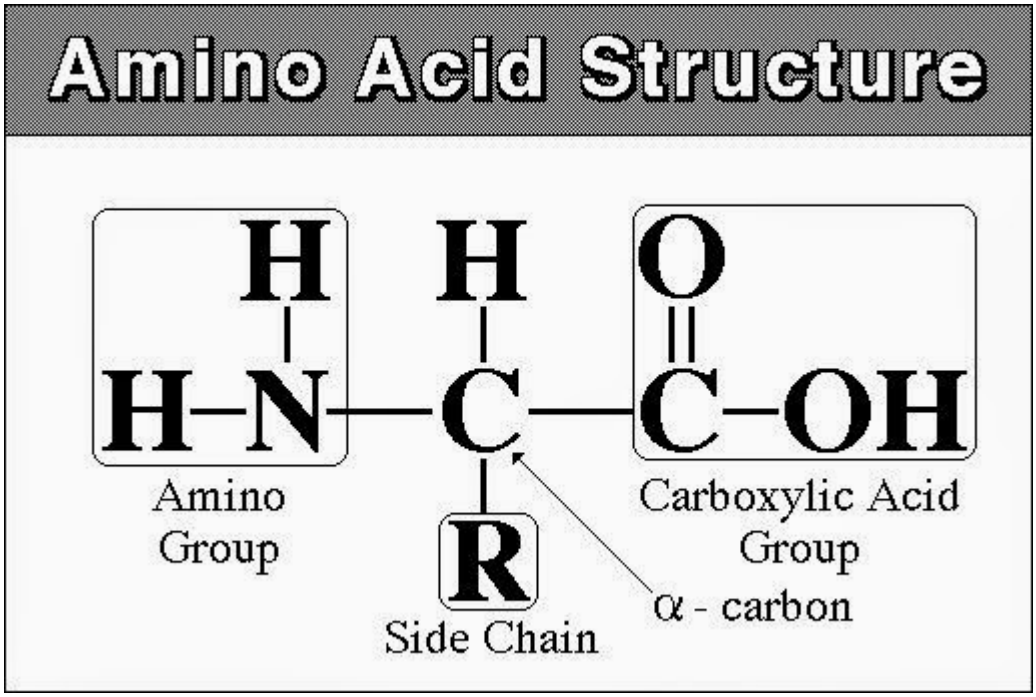
Heat, pH change: Tertiary and secondary structure as well as biological function.

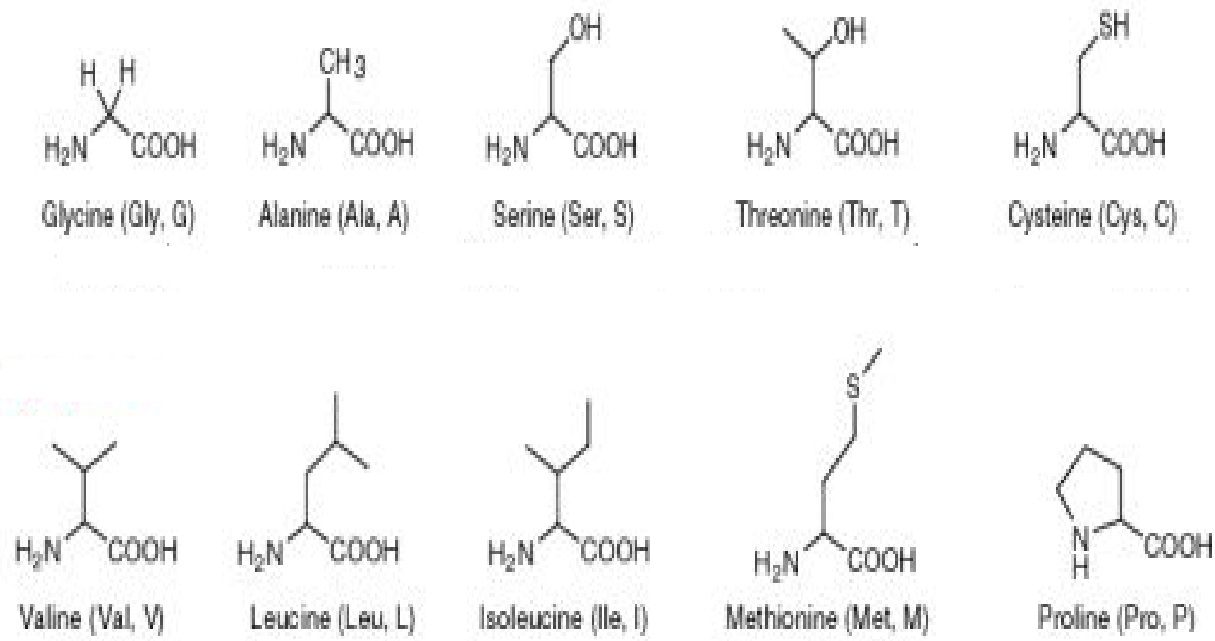


Module62: Chemical composition of protein

Text (9:00)

- Proteins are polymers of amino acids
- They range in size from small to very large
- All the proteins are made up of Twenty different types of amino acids. So these amino acids are called standard amino acids
- In a protein molecule, each amino acid residue is joined to its neighbour by a specific type of covalent bond which is called Peptide Bond
- Amino acids can successively join to form dipeptides, tripeptides, tetrapeptides, oligo peptides and polypeptides





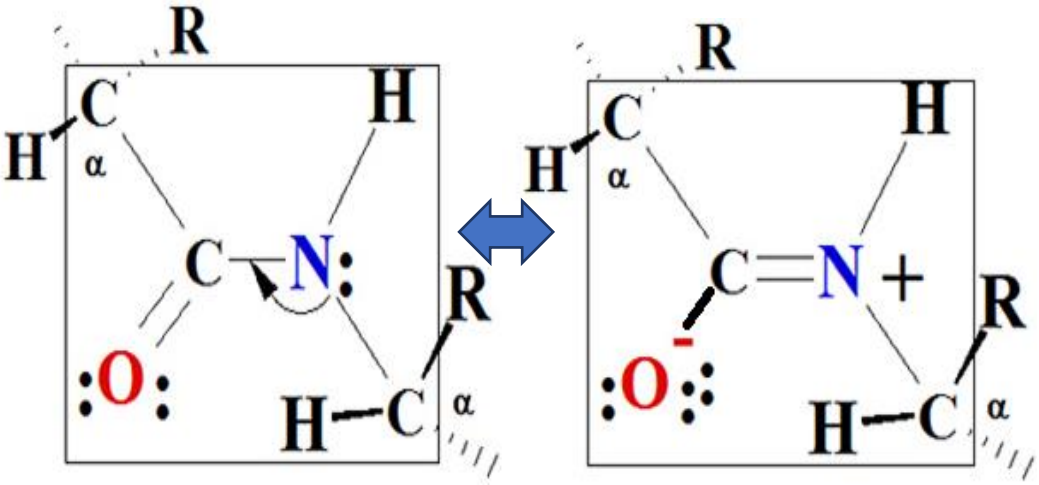
Module63: Primary Structure of proteins

Text (8:00)

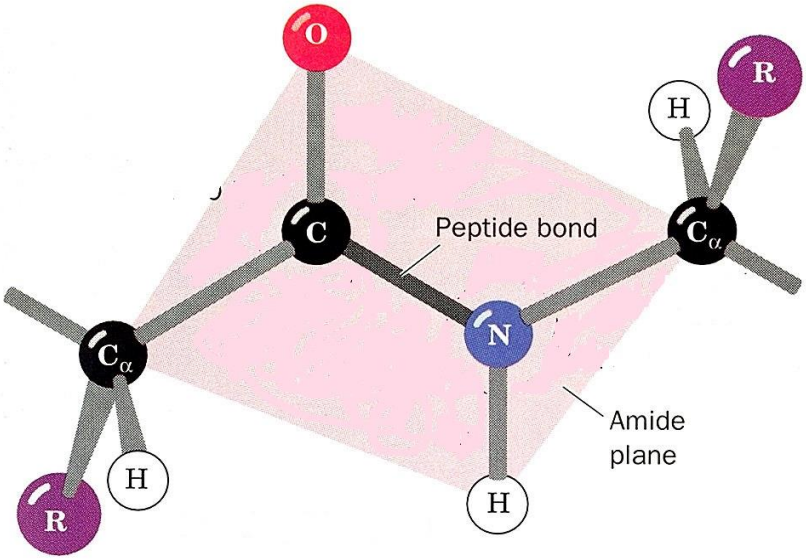
- Primary structure or covalent structure of protein refers to the amino acid sequence of its polypeptide chain.
- Each type of protein has a unique amino acid sequence.

Peptide Bond Is Rigid and Planar

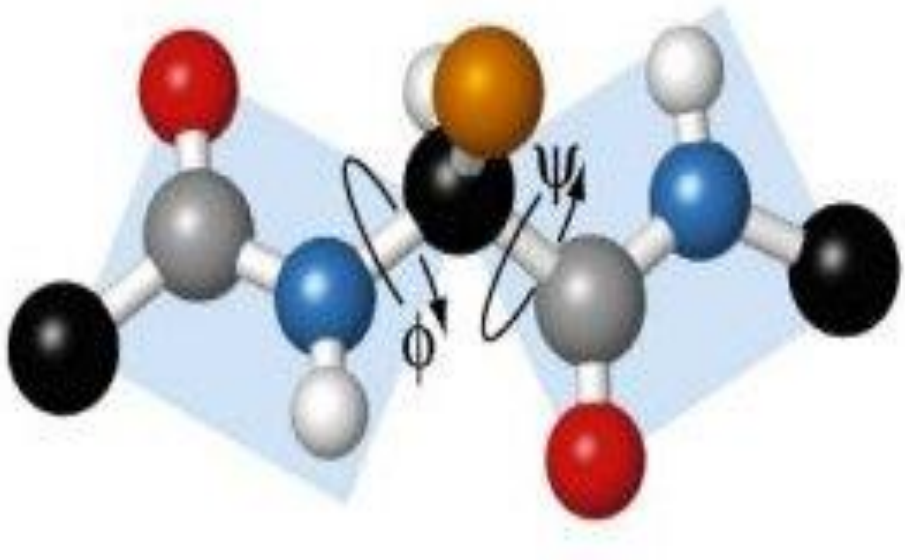
- They demonstrated that the peptide C - N bond is somewhat shorter than the C - N bond in a simple amine.



- The six atoms of the peptide group are co-planar i.e., lie in a single plane, with the oxygen atom of the carbonyl group and the hydrogen atom of the amide nitrogen trans to each other.



- Pauling and Corey concluded that the peptide C - N bonds are unable to rotate freely because of their partial double-bond character.
- Rotation is permitted about the N - αC and the αC - C bonds.

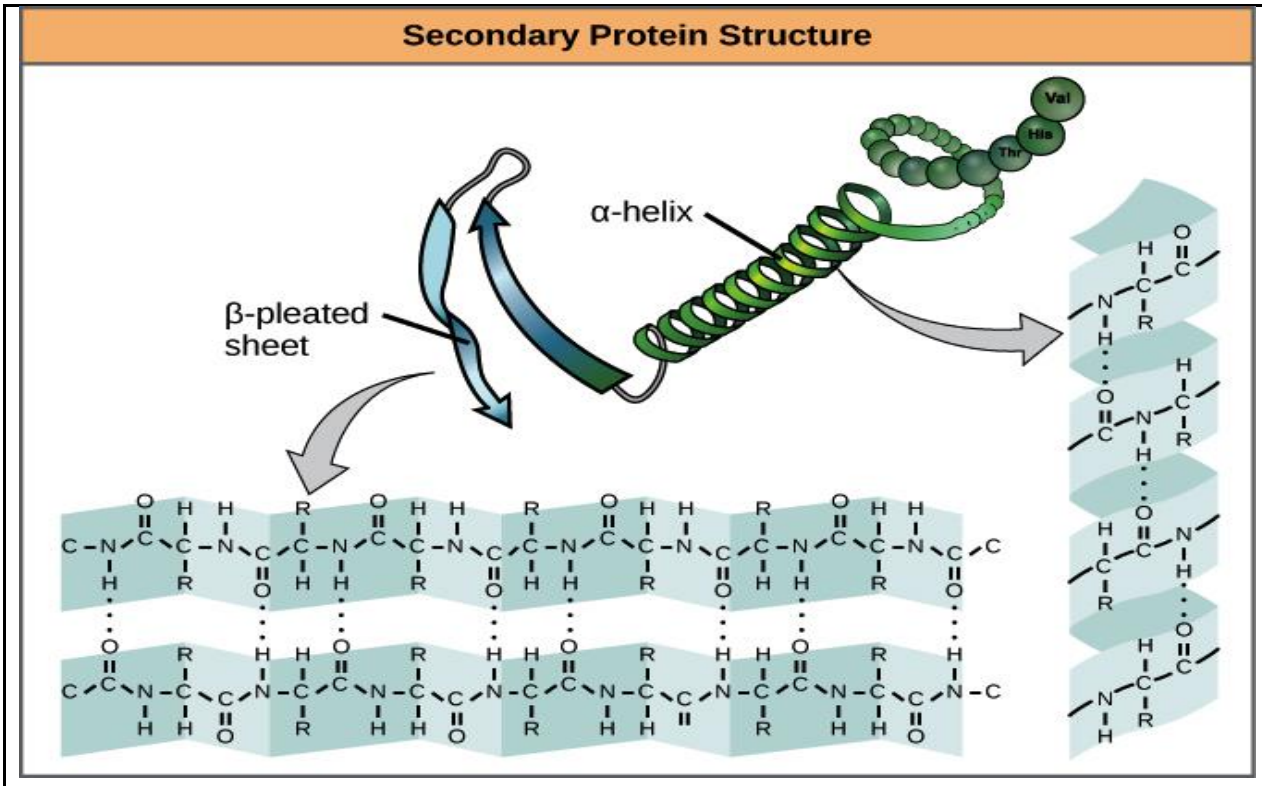


- The bond angles resulting from rotations at C are labelled ϕ (phi) for the N - αC bond and ψ(psi) for the αC - C bond.
- In principle, ϕ and ψ can have any value between +180 & -180.

Module64: Secondary structure of protein

Text (7:00)

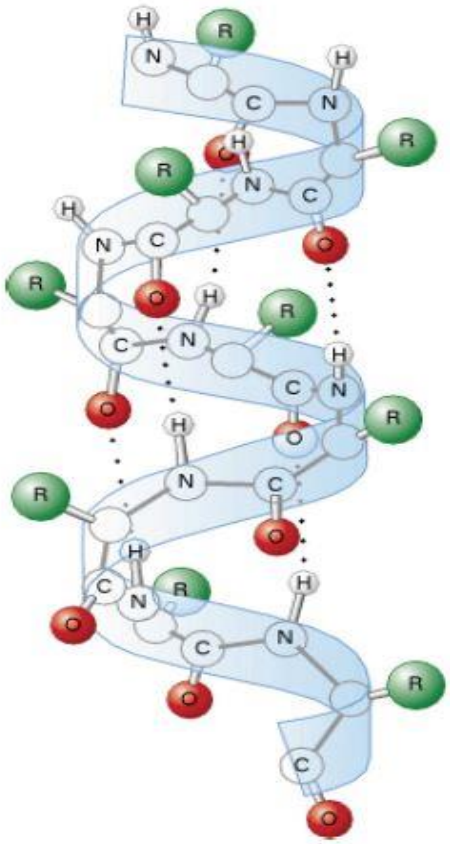
1. Secondary structure of proteins refers to the local conformation of some part of a polypeptide
2. A few types of secondary structures are particularly stable and occur widely in proteins
3. The most prominent are:- Alpha helix and Beta conformation



Module65:Alpha Helix

Text (3:49)

1. The simplest arrangement which a polypeptide chain could assume with its rigid peptide bonds is a helical structure, which Pauling and Corey called the **α-helix**.
2. The helical twist of the *α*-helix found in all proteins is right-handed.
3. The repeating unit is a single turn of the helix, which extends about 5.4 Å (includes 3.6 amino acid residues) along the long axis
4. The amino acid residues in an helix have conformations with psi = –45 to –50 and phi = –60.
5. An helix makes optimal use of internal hydrogen bonds.
6. About one-fourth of all amino acid residues in polypeptides are found in α-helices while in some proteins it is the predominant structure



Module66: Beta Pleated Sheet

Text (8:00)

1. Pauling and Corey predicted a second type of secondary structure which they called **β-sheets**
2. This is a more extended conformation of polypeptide chains
3. The backbone of the polypeptide chain is extended into a zigzag structure
4. The zigzag polypeptide chains are arranged side by side to form a structure resembling a series of pleats
5. The R groups of adjacent amino acids protrude from the zigzag structure in opposite directions
6. Hydrogen bonds are formed between adjacent segments of polypeptide chain
7. The adjacent polypeptide chains in a sheet can be either parallel or antiparallel

Module67: Tertiary structure of protein

Text (8:00)

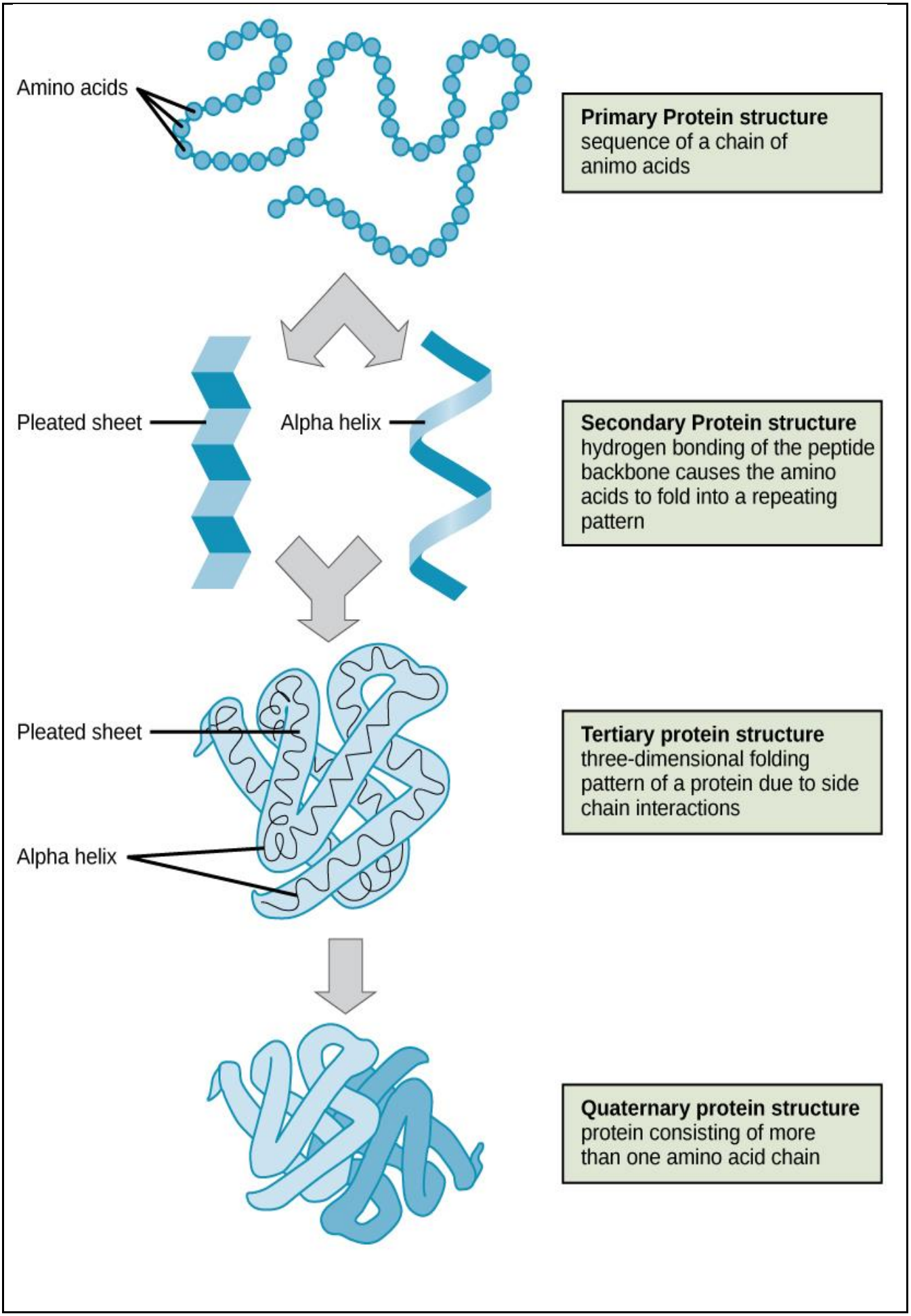
- The overall three-dimensional arrangement of all atoms in a protein is referred to as the protein's tertiary structure.

- It includes longer-range aspects of amino acid sequence.
- Amino acids that are far apart in the polypeptide chain may interact within the completely folded structure of a protein.
- Interacting segments of polypeptide chains are held in their characteristic tertiary positions by different kinds of weak interactions (and sometimes by covalent bonds) between the segments.
- Large polypeptide chains usually fold into two or more globular clusters known as domains, which often give these proteins a bi- or multilobal appearance.

Module68: Quaternary structure of protein

Text (11:00)

1. Some proteins contain two or more separate polypeptide chains or subunits, which may be identical or different.
2. The spatial arrangement of these subunits is known as a protein's quaternary structure
3. A multi-subunit protein is also referred to as a multimer
4. A multimer with just a few subunits is called as oligomer and a single subunit or a group of subunits, is called a protomer
5. Identical subunits of multimeric proteins are generally arranged in a symmetric patterns
6. Oligomers can have either rotational symmetry or helical symmetry
7. There are several forms of rotational symmetry. The simplest is cyclic symmetry, involving rotation about a single axis
8. A somewhat more complicated rotational symmetry is dihedral symmetry, in which a twofold rotational axis is present
9. More complex rotational symmetries include icosahedral symmetry
10. An icosahedron is a regular 12-cornered polyhedron having 20 triangular faces
11. The other major type of symmetry found in oligomers is helical symmetry


Amino acids
Primary Protein structure
sequence of a chain of
animo acids
Pleated sheet
Alpha helix
Secondary Protein structure
hydrogen bonding of the peptide
backbone causes the amino
acids to fold into a repeating
pattern
Pleated sheet
Alpha helix
Tertiary protein structure
three-dimensional folding
pattern of a protein due to side
chain interactions
Quaternary protein structure
protein consisting of more
than one amino acid chain

Module69; Storage of biological sequence

Text (8:00)

1. **BACKGROUND**

   We know that sequence of DNA contain A, C, T & G nucleotides and sequence of RNA contains A, C, U & G while sequence of protein contains A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y & P these are actually 20 different amino acids in nature which compose a protein.

2. **INTRODUCTION**

   When both DNA and RNA are sequenced in lab their sequences contain larger number of nucleotides with a huge number of varieties. And when we talk about protein its sequence contains large number of bases as they are complex in nature. So, at this stage we required some extra methods or techniques which will help us to restore the data of DNA or RNA.

3. **SOLUTIONS DATABASES**

   This large number of sequence or bases cannot be stored in a single computer that's why there are many public sequence data bases for DNA & RNA such as **GenBank (by NIH).** For proteins, the public database is **UniProt (by Uniprot Consortium)**. Both **GenBank** and **UniProt** are online database and the DNA, RNA and Protein sequences are available here online for public and researchers.

4. **CONCLUSION**

We stored the information of nucleotide and amino acids regarding sequences and other properties via online databases. These databases offer readily available sequences.

Module 70: COMPARING SEQUENCES

Text (7 minutes)

Sequence comparison is the process of comparing and detecting similarities between biological sequences. What "similarities" are being detected will depend on the goals of the particular alignment process. Sequence alignment appears to be extremely useful in a number of bioinformatics applications.

For example, the simplest way to compare two sequences of the same length is to calculate the number of matching symbols. The value that measures the degree of sequence similarity is called the alignment score of two sequences.

1. **BACKGROUND**

   There are millions of sequences in GenBank and UniProt (online databases) and if we will compare them so, we can acquire knowledge.

2. **INTRODUCTION**

   By comparing sequences of DNA, RNA and Proteins we can get

   - Similarity among sequences

- There might be some specific difference due to some disease or mutation
- There may be some evolutionary history
- Relationship among species

As their nucleotides, can be similar or differ from each other. While UniProt and blastp etc. are used in case of amino acids sequence comparison.

3. CONCLUSION

By comparison of nucleotides and amino acids of any DNA, RNA and protein sequence we can find similarities, difference, evolutionary facts and relations among species.

Module72: Pairwise Sequence Alignment – I

Text (9:00)

1. BACKGROUND

A sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

Pairwise sequence alignment methods are used to find the best-matching piecewise (local or global) alignments of two query sequences. Pairwise alignments can only be used between two sequences at a time, but they are efficient to calculate and are often used for methods that do not require extreme precision (such as searching a database for sequences with high similarity to a query). The three primary methods of producing pairwise alignments are dot-matrix methods, dynamic programming, and word methods;[1] however, multiple sequence alignment techniques can also align pairs of sequences. Although each method has its individual strengths and weaknesses, all three pairwise methods have difficulty with highly repetitive sequences of low information content - especially where the number of repetitions differ in the two sequences to be aligned.

Exact matches mean all nucleotides or amino acids are same whereas inexact matches mean there is some or more difference between the sequences in form of change in nucleotides or amino acids.

2. INTRODUCTION

Before the matching of nucleotide or amino acid we determine the conserved residue. A residue which is exactly same in sequences known as **conserved residue**. If we compare two or more than two sequences and find match and mismatch this process is known as **alignment**. If we compare two sequences then, this alignment is known as **pair-wise alignment** of nucleotides or amino acids. In pair wise alignment we study functional, structural and/or evolutionary relationships.

In pair wise alignment, matches show in colored form whereas mis matches are shown by "**.**"

Which is known as gaps.

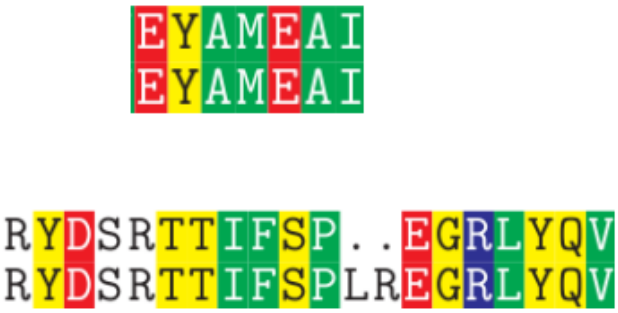


Figure 5.2.1.

3. **CONCLUSION**

   In pair wise alignment the nucleotides (or amino acids) come in pairs and matching are colored while missing nucleotides (or amino acids) are indicated with empty space which is known as gap.

Module73: Pairwise Sequence Alignment – II

Text (9)

There are two types of pairwise alignments: local and global alignments.

***A Local Alignment*. A local alignment is an alignment of two sub-regions of a pair of sequences. This type of alignment is appropriate when aligning two segments of genomic DNA that may have local regions of similarity embedded in a background of a non-homologous sequence.**

***A Global Alignment*. A global alignment is a sequence alignment over the entire length of two or more nucleic acid or protein sequences. In a global alignment, the sequences are assumed to be homologous along their entire length.**

1. **BACKGROUND**

   In pairwise sequence alignment, we align two sequences of nucleotides or amino acids. Matches are shown in colored form whereas mismatches are denoted by "**.**".

2. **SALIENT POINTS**

   Sometimes sequences are slides on each other to maximizing the matches and mis matches whereas gaps are used for deletions and insertions. Gaps are reducing overall score of alignment.

3. **TYPES OF PAIRWISE ALIGNMENTS**

   Mainly there are two types of pair wise alignments.

- Global alignment
- Local alignment

1.1. GLOBAL ALIGNMENT

A type of alignment in which we introduce gaps for getting maximum matching score is known as **global alignment.** In this alignment, we align whole sequences.

1.2. LOCAL ALIGNMENT

A type of alignment in which we find those regions which have strongest matching score and it ignores the less similar matchings as per threshold, known as **local alignment**. This type of alignment used to find the evolutionary behavior.

Another type of alignment; **optimal alignment** that exhibits the most correspondence between the query and the source sequence. It is the alignment with the highest score.

2. CONCLUSION

Gaps are introduced in sequences for maximum matching. We can use both global and local alignment, but it depends upon our conditions.

Module74: Pairwise Sequence Alignment – III

Text (10)

If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations and gaps as indels (that is, insertion or deletion mutations) introduced in one or both lineages in the time since they diverged from one another. In sequence alignments of proteins, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region or sequence motif is among lineages. The absence of substitutions, or the presence of only very conservative substitutions (that is, the substitution of amino acids whose side chains have similar biochemical properties) in a particular region of the sequence, suggest that this region has structural or functional importance. Although DNA and RNA nucleotide bases are more similar to each other than are amino acids, the conservation of base pairs can indicate a similar functional or structural role.

To compare two or more sequences, it is necessary to align the conserved and unconserved residues across all the sequences (identification of locations of insertions and deletions that have occurred since the divergence of a common ancestor). These residues form a pattern from which the relationship between sequences can be determined with phylogenetic programs. When the sequences are aligned, it is possible to identify locations of insertions or deletions since their divergence from their common ancestor. There are three possibilities :

- The bases match : this means that there is no change since their divergence.
- The bases mismatch : this means that there is a substitution since their divergence.

- There is a base in one sequence, no base in the other : there is an insertion or a deletion since their divergence.

Figure 12 : The comparison of sequences.

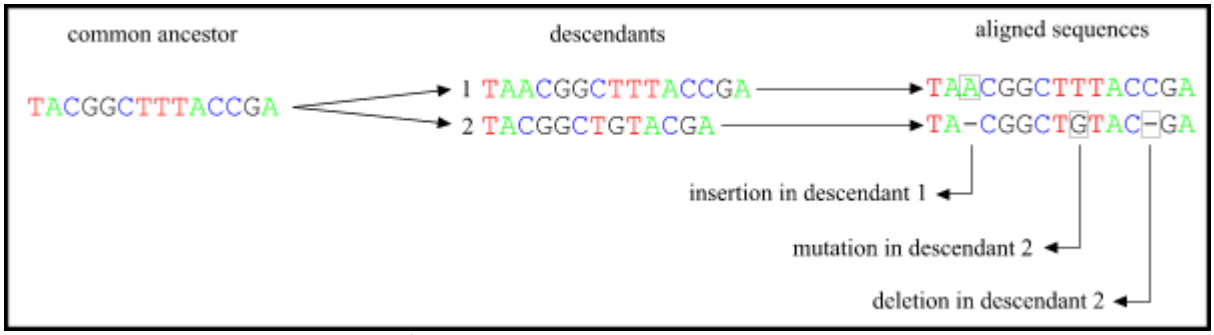
INTRODUCTION

Pair wise alignment helps us to find the similarity and differences there are three ways according to which sequences can differ from each other.
These are

- Substitutions A**C**GA → A**G**GA
- Insertions ACGA → AC**C**GA
- Deletions A**C**GA → AGA

By applying all above ways to any sequence, the matching and mismatching can be increased or decreased between to different comparing sequencing.

MATCH → **NO CHANGE BETWEEN/AMONG SEQUENCES**



MISMATCH → **SUBSTITUTIONS IN SEQUENCES**

GAP → **INSERTIONS OR DELETIONS IN SEQUENCES**

1. **OPTIMAL ALIGNMENT**

The optimal alignment of two protein sequences is the alignment that maximizes the sum of pair-scores less any penalty for introduced gaps.

Given below is the example to study the concept of optimal alignment.

Sequence 1 & 2

1. THIS IS A RATHER LONGER SENTENCE THAN THE NEXT.

2. THIS IS A SHORT SENTENCE.

1: THIS IS A RATHER LONGER SENTENCE THAN THE NEXT.
2: THIS IS A .....SHORT.....SENTENCE.....
OR
1: THIS IS A RATHER LONGER SENTENCE THAN THE NEXT.
2: THIS IS A .....SHORT.....SENT..EN.....CE.....

Optimal aligned sequence is first one. Because there are **16** matches and in second sequence there are only **14** matches.

3. **CONCLUSION**
   Both local and Global alignments give us different results. Indels (Insertion or deletion) gives gaps in alignment whereas substitution increases mismatch of sequence.

Module76: Introduction to Multiple Sequences Alignment

Text (7:00)

1. **BACKGROUND**
   In pair-wise sequence alignments, we use pairs of sequence to compare them. And scoring matrices were used to score the sequence ranks. Pair-wise alignment used for either local or global alignment.

2. **INTRODUCTION**
   In **M**ultiple **S**equence **A**lignments (**MSA**) we compare multiple numbers of protein and DNA sequences to identify the matches and mismatches. Multiple sequence alignment is mostly use for global alignment.

```
M Q V K L F T P L H D K S D H G K Y H
M Q V K I F T P L H D K S - H G K S H
M Q V H L Y - P L H D K S - T G K S H
M Q V H L F - P L H D K S D T G K S H
M Q V K L Y T P L H D K S D H G K Y H
```

Figure 11.5.1: A block of MSA

3. **MSA VS. PAIR-WISE ALIGNMENT**

   For pair-wise alignment we use Dynamic programming but for multiple alignments it would be very expensive, computationally. So, solution for this is "**progressive alignment**".

4. **CONCLUSION**

MSA can help us to align multiple sequences. For MSA we can use progressive alignment so, we can use CLUSTAL (online software tool).

Module77: More on Multiple Sequence Alignment

Text (9:00)

1. **BACKGROUND**

   MSA helps us to compare several sequences by aligning them. MSA can extract consensus sequences from several aligned sequences. Characterize protein families based on homologous regions.

2. **APPLICATION OF MSA**

   There are many applications of MSA but here are applications which are most important.

   - We can predict secondary and tertiary structures of new protein sequences
   - We can also determine the evolutionary order of species or "Phylogeny"

3. **METHODOLOGY**

   - Pair-wise alignment is the alignment of two sequences
   - MSA can be performed by repeated application of pair-wise alignment

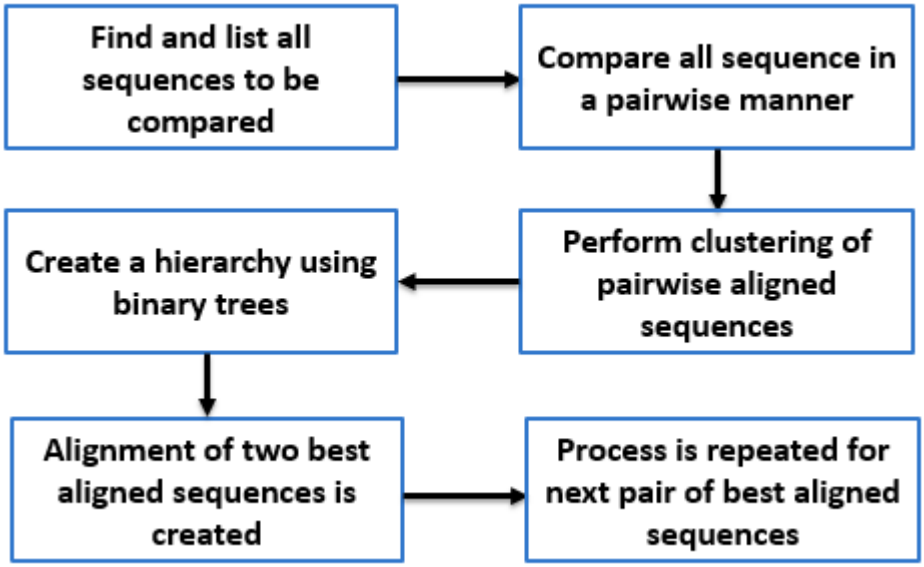


Figure 12.1.1: Methodology

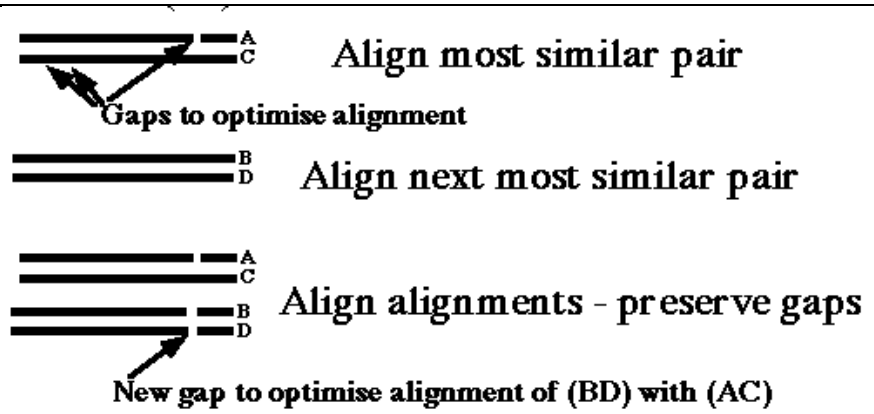


Figure 12.1.2: Progressive alignment

4. CONCLUSION

MSA can help us to align multiple sequences. Progressive alignment used to perform MSA. Need to remove sequences with >80% similarity. For this purpose, there are many online tools, but we will use CLUSTAL.

Module78: Progressive Alignment for MSA

Text (12:00)

1. BACKGROUND

MSA uses progressive alignment of sequences. But numerous progressive alignments can be slow the whole process. And then, it will be computationally expensive.

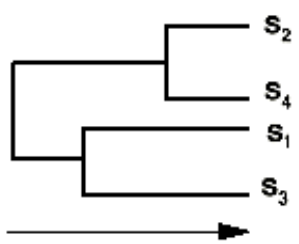
2. HOW IT WORKS?

Step 1: Pairwise Alignment of all sequences

Example: S_1, S_2, S_3, S_4, so that is 6 pairwise comparisons i.e. S_1-S_2 and S_1-S_3 etc.

Step 2: Construct a Guide Tree (dendogram) using a Distance Matrix

Step 3: Progressive alignment following branching order in tree



$$Distance = 1 - \frac{Matches}{Comparisons}$$

$$Similarity = \frac{Matches}{Comparisons}$$

Figure 12.2.1: (a) Dendogram (b) Formulae

	v_1	v_2	v_3	v_4
v_1	-			
v_2	.17	-		
v_3	.87	.28	-	
v_4	.59	.33	.62	-

Figure 12.2.2: (a) Dendogram (b)Results

3. **SHORTCOMING OF THIS APPROACH**
   - It depends upon initial alignments
   - If sequences are dissimilar, errors in alignment are propagated

4. **SOLUTION: SHORTCOMING OF THIS APPROACH**
   Begin by using an initial alignment and refine it repeatedly.

5. **CONCLUSION**

Progressive alignments are used in aligning multiple sequences. Iteratively, refining of results from progressive alignments make this approach effective.

Module80: CLUSTAL

Text (10:00)

1. **BACKGROUND**
   MSA uses progressive alignment of sequences. But numerous progressive alignments can be slow the whole process. And then, it will be computationally expensive. CLUSTALW is an online tool to perform MSA.

2. **INTRODUCTION**
   CLUSTALW is developed by European Molecular Biology Laboratory & European Bioinformatics Institute. It uses multiple file formats as an input which is EMBL/SwissProt, Pearson (FASTA) etc. It performs alignment under two types of options:
   - Slow/accurate
   - Fast/approximate

3. **SCOPE**
   - Create multiple alignments
   - Optimize existing alignments
   - Profile analysis
   - Create phylogenetic trees

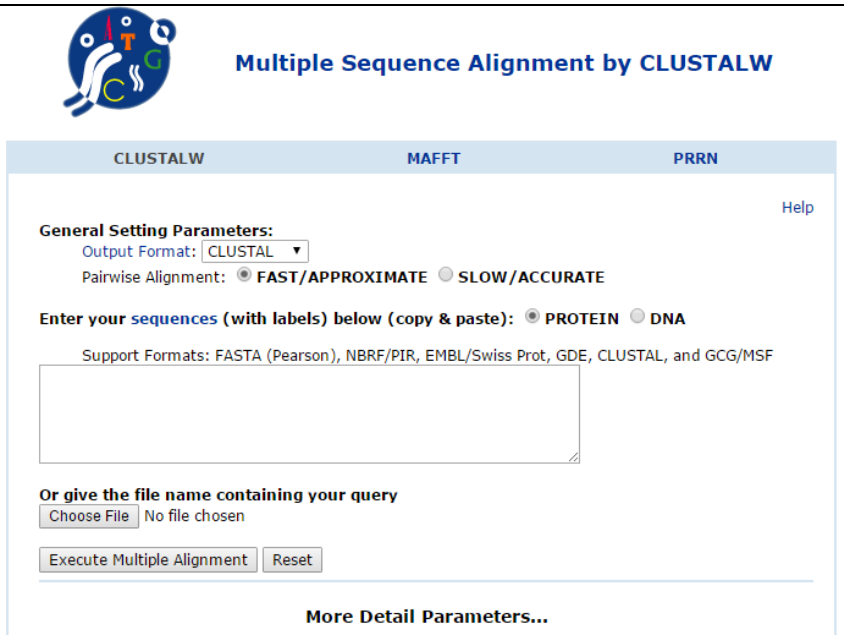


Figure 12.4.1: Homepage of CLUSTALW

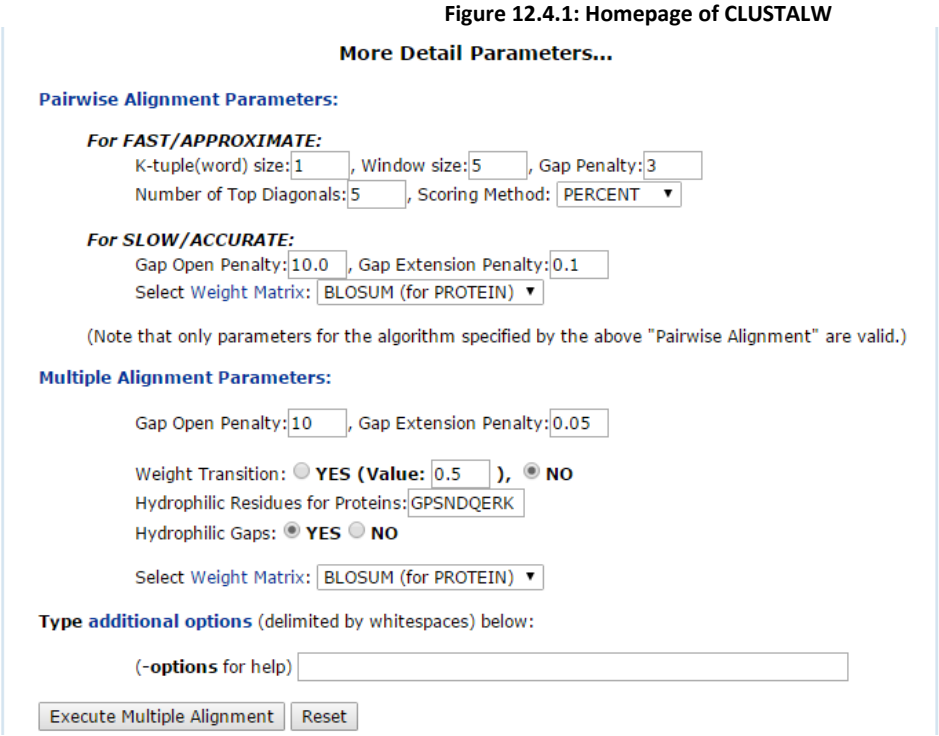


Figure 12.4.2: Inputting in CLUSTALW

Table 12.4.1: Some recent & less recent methods for MSAs

Name	Algorithm	URL
MSA	Exact	http://www.ibc.wustl.edu/ibc/msa.html
DCA	Exact (requires MSA)	http://bibiserv.techfak.uni-biefield.de/dca
OMA	Iterative DCA	http://bibiserv.techfak.uni-biefield.de/oma
ClustalW, ClustalX	Progressive	ftp://ftp-igbmc.u-strasbg.fr/pub/clustalW or clustalX
MultAlin	Progressive	http://www.toulouse.inra.fr/multalin.html
DiAlign	Consistency-based	http://www.gsf.de/biodv/dialign.html
ComAlign	Consistency-based	http://www.daimi.au.df/~ ocaprani
T-Coffee	Consistency-based/progressive	http://igs-server.cnrs-mrs.fr/~ cnotred
Praline	Iterative/progressive	jhering@nimr.mrc.ac.uk
IterAlign	Iterative	http://giotto.Stanford.edu/~ luciano/iteralign.html
Prrp	Iterative/Stochastic	ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/
SAM	Iterative/Stochastic/HMM	rph@cse.ucsc.edu
HMMER	Iterative/Stochastic/HMM	http://hmmer.wustl.edu/
SAGA	Iterative/Stochastic/GA	http://igs-server.cnrs-mrs.fr/~ cnotred
GA	Iterative/Stochastic/GA	czhang@watnow.uwaterloo.ca

4. CONCLUSION

CLUSTALW can be used to perform MSA. It has two types of modes for alignment which are fast and slow. CLUSTAL Omega is now available which includes several upgrades.

Module81: Introduction to BLAST - I

Text (9:00)

1. INTRODUCTION

Basic Local Alignment Search Tool (BLAST) developed by National Center for the Biotechnology Information (NCBI) – USA in 1990. It searches databases for query protein and nucleotide sequences. Also, searches for translational products etc.

Online availability at www.blast.ncbi.nlm.nih.gov/Blast.cgi

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information

BLAST®

BLAST finds regions of similarity between biological sequences. more...

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested GO

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- Microbes

Basic BLAST

Choose a BLAST program to run.

nucleotide blast | Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontiguous megablast

Figure 12.5.1: Homepage of BLAST

BLAST having five distinct features which can be choose based on desired operation.

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Figure 12.5.2: Features of BLAST (Courtesy NCBI)

You can choose your desired feature by clicking on relevant button.

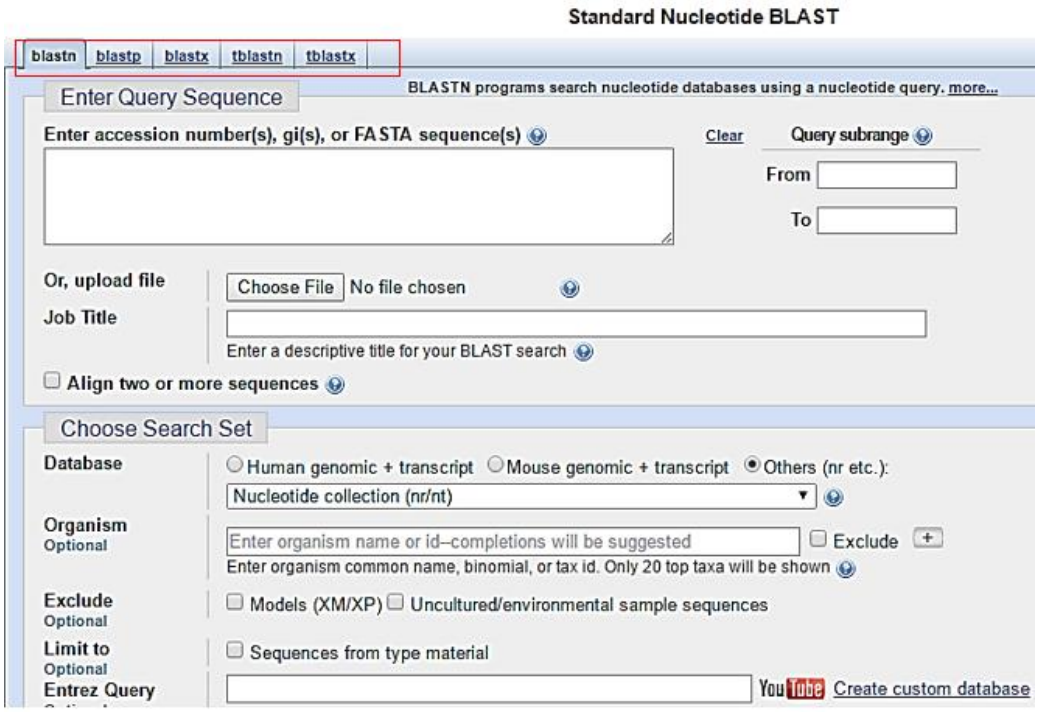


Figure 12.5.3: Homepage of standard Nucleotide blast (blastn) (Courtesy NCBI)

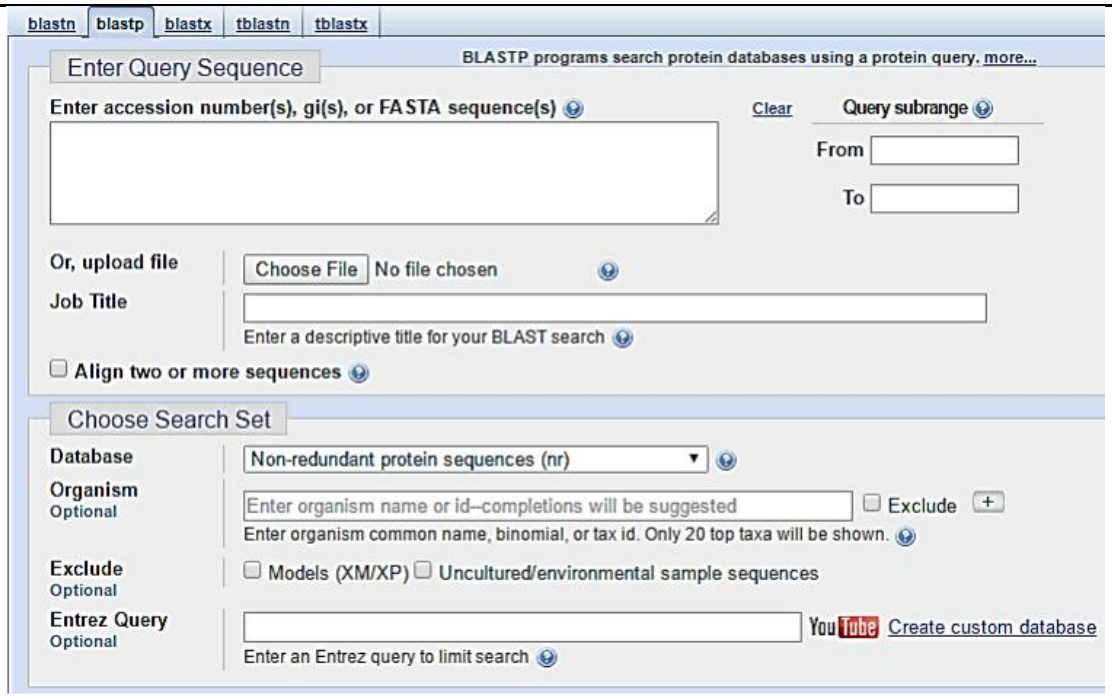


Figure 12.5.4: Homepage of standard Protein blast (blastp) (Courtesy NCBI)

2. CONCLUSION

BLAST can be used to search for local alignment of protein and nucleotide sequences. It is free of cost and available online. BLAST can perform searches across species and organisms.

Module82: Introduction to BLAST - II

Text (08:00)

1. BACKGROUND

Basic **L**ocal **A**lignment **S**earch **T**ool (**BLAST**) developed by **N**ational **C**enter for the **B**iotechnology **I**nformation (**NCBI**) – USA in 1990. It searches databases for query protein and nucleotide sequences. Also, searches for translational products etc.

Online availability at **www.blast.ncbi.nlm.nih.gov/Blast.cgi**

2. INTRODUCTION

Smith Waterman algorithm can align complete sequences. BLAST work on it as an approximate way. Hence, BLAST is faster, but it does not ensure optimal alignment. BLAST provides for approximate sequence matching. For input, we used FASTA formatted sequence in a BLAST and a set of search parameters.

3. HOW IT WORKS?

These steps are used to acquire query sequence of genes or proteins.

- **Step 1:** You can go to NCBI homepage by typing URL (enclosed in red box)
- **Step 2:** By clicking small black arrow (enclosed in black box) you can select your database
- **Step 3:** Then you can enter your query (enclosed in green box) and hit on enter

Similar case is in UniProt, a protein database.

After acquiring query sequences, we used blastn (for nucleotide blast) or blastp (for protein blast) etc. as per desired job.

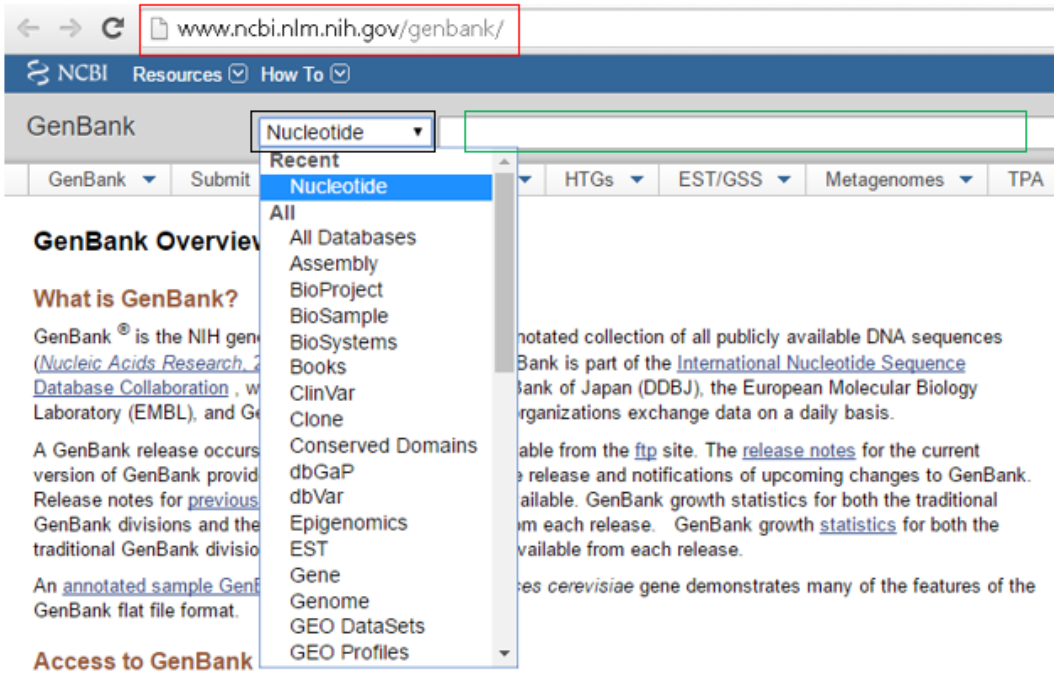


Figure 13.1.1: Input to BLAST: Gene IDs

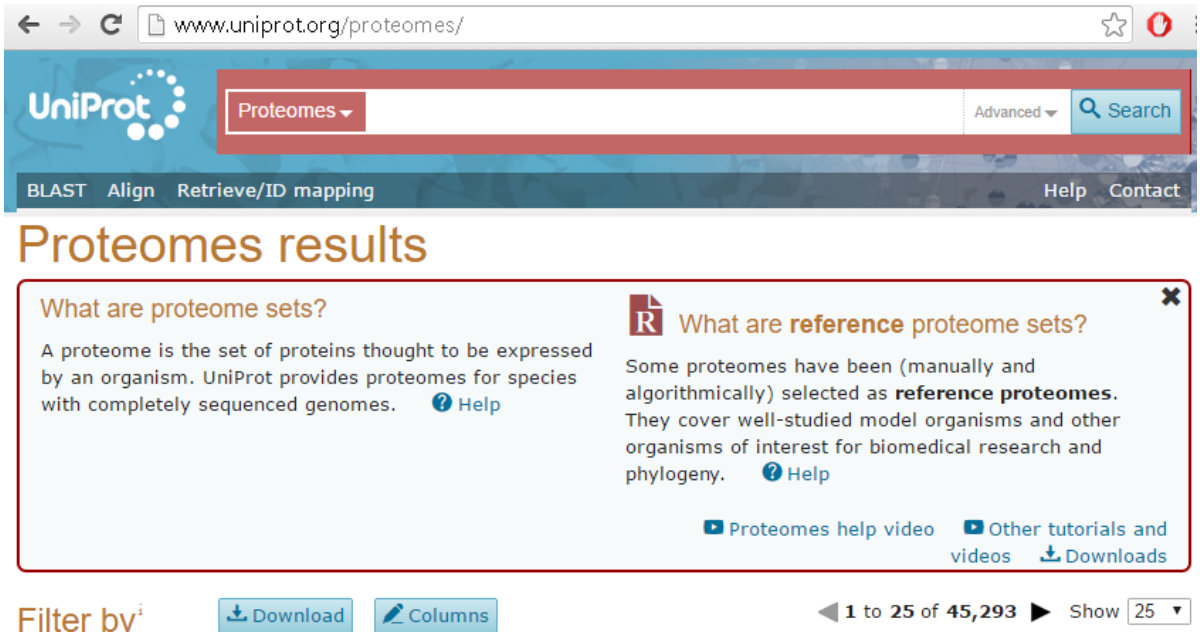


Figure 13.1.2: Input to BLAST: Protein IDs

4. OUTPUT OF BLAST

Results are shown in HTML, plain text, and XML formats. A table lists the sequence hits found along with scores. Users can read this table off and evaluate results.

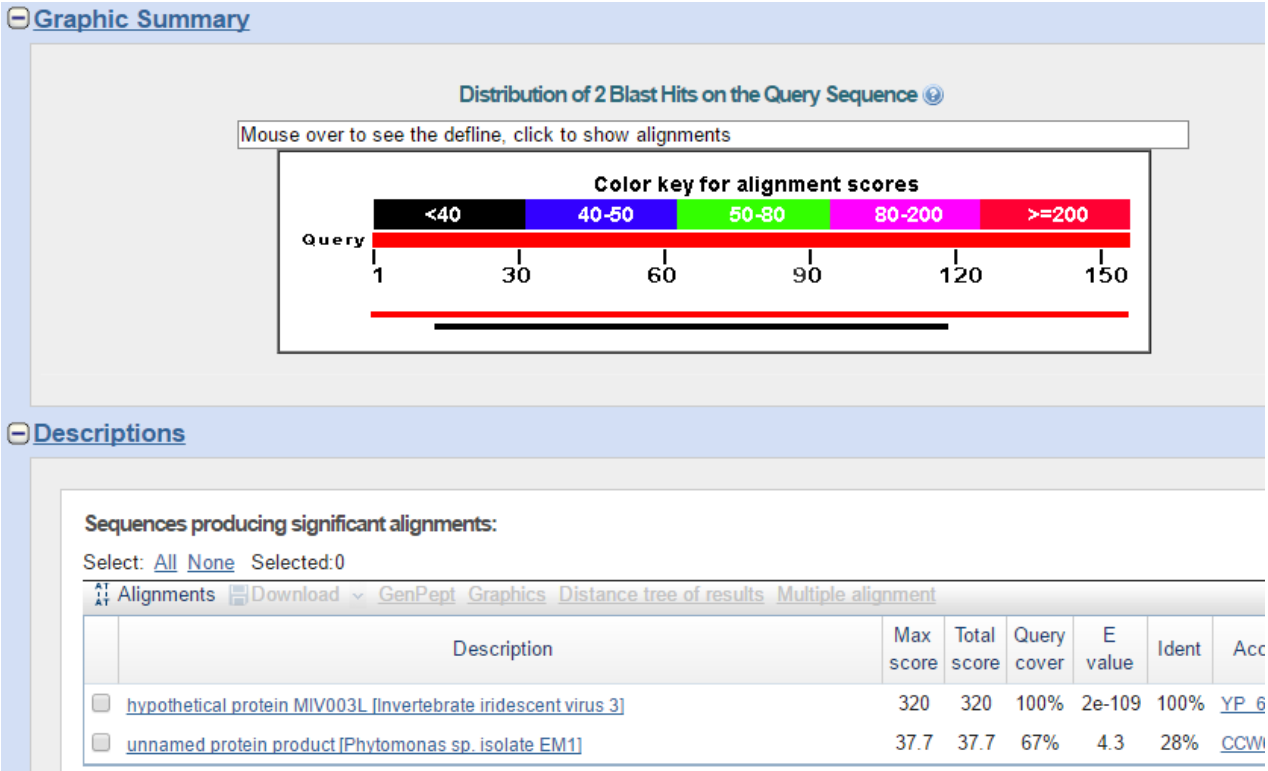


Figure 13.1.3: Results from BLAST

5. CONCLUSION

BLAST is online tool used for both local and global alignment. It is faster but not gives optimal alignment.

Module83: BLAST Algorithm

Text (14:00)

1. INTRODUCTION

BLAST can search sequence databases and then, identify unknown sequences by comparing them to the known sequences. This can help us to identify the parent organism, function and evolutionary history etc.

2. HOW IT WORKS?

Here we are using example for better understanding the working of BLAST.

For example:

Query sequence: **PQGELV**

Make list of all possible words (length 3 for proteins). This list give us alignment score on matching.

PQG (score 15), QGE (score 9)
GEL (score 12), ELV (score 10)

Assign scores from Blosum62, use those with score> 11: PQG & GEL. Score is set to avoid low scoring. And now mutate words such that score still > 11.

PQG (score 15) similar to PEG (score 13)

RESULT: PQG, GEL and PEG

Find all database sequences that have at least 2 matches among our 3 words: PQG, GEL & PEG. Find database hits and extend alignment (High-scoring Segment Pair):

```
Query:    M E T P Q G I A V
Database: - - - P Q G E L V
                8 5 5 2 0 8
```

2.1. **H**igh **S**coring **P**air(**HSP**): PQGL (score 8+5+5+2)

If 2 HSP in query sequence are < 40 positions away
Full dynamic alignment on query and hit sequences. BLAST performs quick alignments on sequences.

3. CONCLUSION

BLAST performs quick alignments on sequences. The results of BLAST are tabulated with alignment regions overlapping each other. Statistical evaluation is also provided.

Module84: Types of BLAST

Text (10:00)

1. INTRODUCTION

BLAST can search sequence databases and identify unknown sequences by comparing them to the known sequences. BLAST This can help identify the parent organism, function and evolutionary history.

2. TWO MAIN TYPES OF BLAST

There are two main types of BLAST.

2.1. Nucleotide BLAST

- **Blastn:** Compares a nucleotide query sequence against a nucleotide database

2.2. Protein BLAST

- **Blastp:** Compares an amino acid query sequence against a protein database

2.3. OTHER TYPES OF BLAST

There are also many other types of BLAST:

- **Blastx:** Compares a nucleotide query sequence against a protein sequence database. It is used to find potential translation products of unknown nucleotide sequences.

- **tblastn:** Compares a protein query sequence against a nucleotide sequence database Nucleotide sequence dynamically translated into all reading frames

- **tblastx:** Compares the six-frame (**O**pen **R**eading **F**rame) translated proteins of a nucleotide query sequence against the six-frame (**O**pen **R**eading **F**rame) translated proteins of a nucleotide sequence database.

3. CONCLUSION

BLAST performs quick alignments on biological sequences. Several types of BLAST exist which can assist in comparing nucleotide sequences with amino acids and vice versa.

Module85: Summary of BLAST

Text (10:00)

1. INTRODUCTION

BLAST can search sequence databases and identify unknown sequences by comparing them to the known sequences. This can help to identify the parent organism, function and evolutionary history. Different types of BLAST program exist for aligning purpose. User can select program according to its requirement.

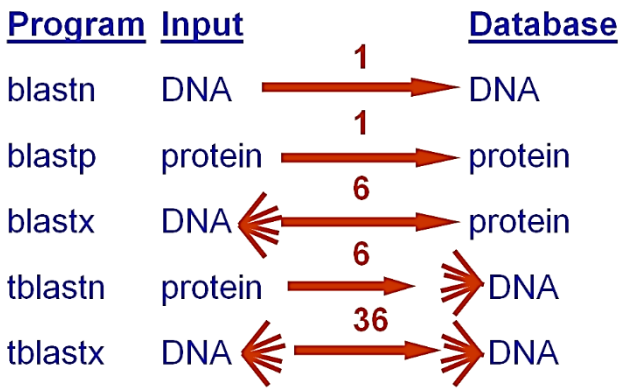


Figure 13.4.1: Types of BLAST with description

2. HOW IT WORKS?

- **Step 1:** Obtained a query of a sequence
- **Step 2:** Choose a type of BLAST according to your goal
- **Step 3:** Entered your query and search parameters
- **Step 4:** Processing by BLAST (Auto performed by BLAST)
- **Step 5:** Result page opened after alignment process
- **Step 6:** Tabulated search results

NCBI Resources How To My NCBI Sign

Protein
Translations of Life

Search: Protein Limits Advanced search Help

Search Clear

Display Settings: FASTA Send to: Change region shown

hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

GenPept Graphics

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

Analyze this sequence

Run BLAST

Identify Conserved Domains

Find in this Sequence

Figure 13.4.2: Obtain query sequence (Step 1)

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Figure 13.4.3: Choose type of BLAST (Step 2)

BLAST Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastp suite

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

NP_000509.1

Clear

Query subrange

From

To

Or, upload file

Browse...

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested

Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

blastp (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters

Figure 13.4.4: Enter search parameters (Step 3)

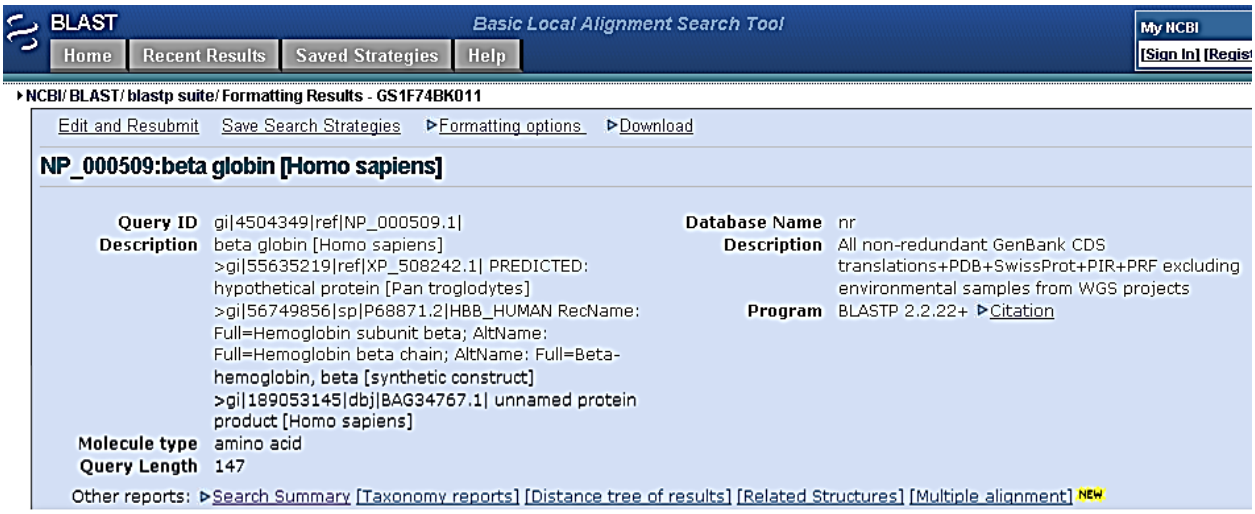


Figure 13.4.4: Processing by BLAST (Auto performed by BLAST) (Step 4)

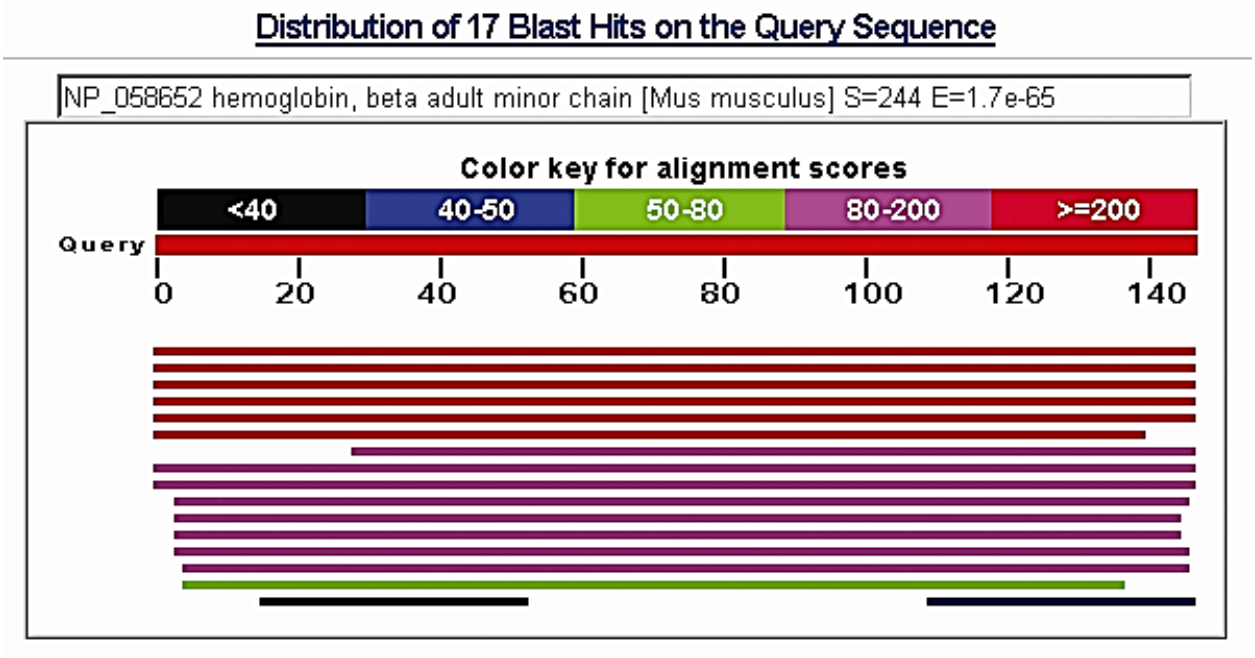


Figure 13.4.5: Result page opened after alignment process (Step 5)

Distance tree of results NEW

Sequences producing significant alignments:		Score (Bits)	E Value	
ref|NP_058652.1|	hemoglobin, beta adult minor chain [Mus musculu	244	2e-65	U G
ref|NP_032246.2|	hemoglobin, beta adult major chain [Mus musculu	228	2e-60	U G
ref|XP_978992.1|	PREDICTED: similar to Hemoglobin epsilon-Y2 ...	226	3e-60	G
ref|NP_032247.1|	hemoglobin Y, beta-like embryonic chain [Mus mu	223	4e-59	U G
ref|NP_032245.1|	hemoglobin Z, beta-like embryonic chain [Mus mu	223	6e-59	U G
ref|XP_998314.1|	PREDICTED: similar to Hemoglobin beta-H1 sub...	203	4e-53	G
ref|XP_978924.1|	PREDICTED: similar to Hemoglobin epsilon-Y2 ...	187	2e-48	G
ref|XP_912634.1|	PREDICTED: similar to Hemoglobin beta-2 subu...	161	2e-40	G
ref|XP_488069.1|	PREDICTED: similar to Hemoglobin beta-2 subu...	154	3e-38	U G
ref|NP_032244.1|	hemoglobin alpha 1 chain [Mus musculus]	105	1e-23	U G
ref|XP_994669.1|	PREDICTED: similar to Hemoglobin alpha subun...	101	3e-22	G
ref|XP_356935.3|	PREDICTED: similar to Hemoglobin alpha subun...	100	4e-22	U G
ref|NP_034535.1|	hemoglobin X, alpha-like embryonic chain in ...	94.0	4e-20	U G
ref|NP_001029153.1|	similar to hemoglobin, theta 1 [Mus musculus	88.2	2e-18	U G
ref|NP_778165.1|	hemoglobin, theta 1 [Mus musculus]	73.9	5e-14	U G
ref|XP_978150.1|	PREDICTED: similar to hemoglobin, beta adult...	41.6	2e-04	G
ref|NP_795942.2|	5'-nucleotidase, cytosolic II-like 1 protein [M	28.9	1.5	U G

Figure 13.4.6: Tabulated search results (Step 6)

3. CONCLUSION

BLAST is an online tool which performs quick alignment of biological sequences. According to user's need BLAST offers five different types of features.

Module86: Introduction to FastA-I

Text (9:00)

1. BACKGROUND

For comparing two sequences we use pair-wise sequencing and for the comparison of many sequences we use multiple sequence alignment.

MSA is a progressive pair-wise alignment. To handle the multiple sequences, we perform alignment through Smith-Waterman algorithm for local alignment. And for global alignment we use Needleman-Wunsch algorithm.

Both local and global alignments are the dynamic approaches. Many of the sequences are compared, which takes time and we use BLAST which is an approximate local alignment search tool BLAST compares a large number of sequences, quickly. FASTA took a similar approach.

2. INTRODUCTION

FASTA stands for Fast Alignment, developed in 1988. It does fast alignment. It searches databases for query protein and nucleotide sequences. FATSA was later improved upon in BLAST.

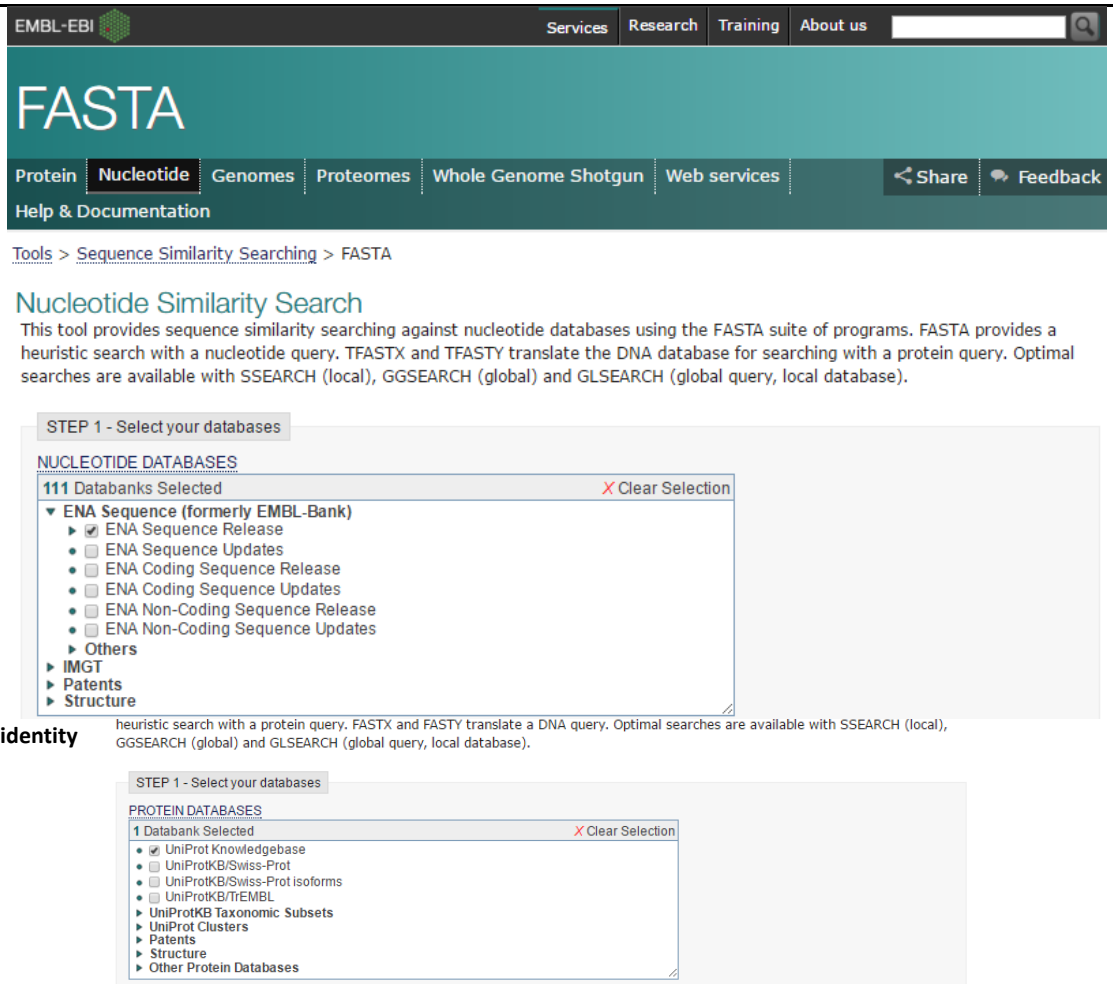


Figure 13.5.1: Regions of absolute identity

Figure 13.5.2: Protein FATSA homepage of EBI

Figure 13.5.3: Nucleotide FATSA homepage of EBI (Courtesy EBI)

3. **CONCLUSION**
   FATSA can perform quick comparison of protein and nucleotide sequences. It can also perform genome and proteome similarity search. It is available online.

Module87: Introduction to FastA-II

Text (8:00)

1. **INTRODUCTION**
   FASTA - Fast Alignment Algorithm. It can search DNA and protein databases with statistically significant similarity. FASTA achieves alignment by using short lengths of exact matches. It is not guaranteed that FASTA can find best alignment between query and alignment because it prefers speed.

2. **USES OF FASTA**
   FASTA relies on aligning subsequences of absolute identity. FASTA can take input for search FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP or UniProt formats.

Nucleotide Similarity Search

This tool provides sequence similarity searching against nucleotide databases using the FASTA suite of programs. FASTA provides a heuristic search with a nucleotide query. TFASTX and TFASTY translate the DNA database for searching with a protein query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

NUCLEOTIDE DATABASES

111 Databanks Selected — X Clear Selection

- ENA Sequence (formerly EMBL-Bank)
  - ENA Sequence Release
  - ENA Sequence Updates
  - ENA Coding Sequence Release
  - ENA Coding Sequence Updates
  - ENA Non-Coding Sequence Release
  - ENA Non-Coding Sequence Updates
  - Others
- IMGT
- Patents
- Structure

STEP 2 - Enter your input sequence

Enter or paste a DNA sequence in any supported format:

```
CAGTCAGTCATAGTCGTAGATGTACGTAGCTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCT
AGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAG
TCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAG
TAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGT
```

or upload a file: Choose File No file chosen

Figure 14.1.1: Input to FASTA: Nucleotide Sequence/Gene IDs

Tools > Sequence Similarity Searching > FASTA

Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected — X Clear Selection

- UniProt Knowledgebase
- UniProtKB/Swiss-Prot
- UniProtKB/Swiss-Prot isoforms
- UniProtKB/TrEMBL
- UniProtKB Taxonomic Subsets
- UniProt Clusters
- Patents
- Structure
- Other Protein Databases

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:

```
MDASSSPSPSEESLKLELDDLQKQLNKKLRFEASVCSIHNLLRDHYSSSSPSLRKQFYIV
VSRVATVLKTRYTATGFWVAGLSLFEEAERLVSDASEKKHLKSCVAQAKEQLSEVDNQPT
ESSQGYLFEGHLTVDREPPQPQWLVQQNLMSAFASIVGGESSNGPTENTIGETANLMQEL
INGLDMIIPDILDDGGPPRAPPASKEVVEKLPVIIFTEELLKKFGAEAECCICKENLVIG
DKMQELPCKHTFHPPCLKPWLDEHNSCPICRHELPTDDQKYENWKEREKEAEEERKGAEN
AVRGGEYMYV
```

or Upload a file: Choose File No file chosen

Figure 14.1.2: Input to FASTA: Protein Sequence

3. **OUTPUT OF FASTA**

   Results are given in visual format along with functional prediction. It makes tabular list with the sequence hits, found along with scores. Users can click on each reported match to look at the details.

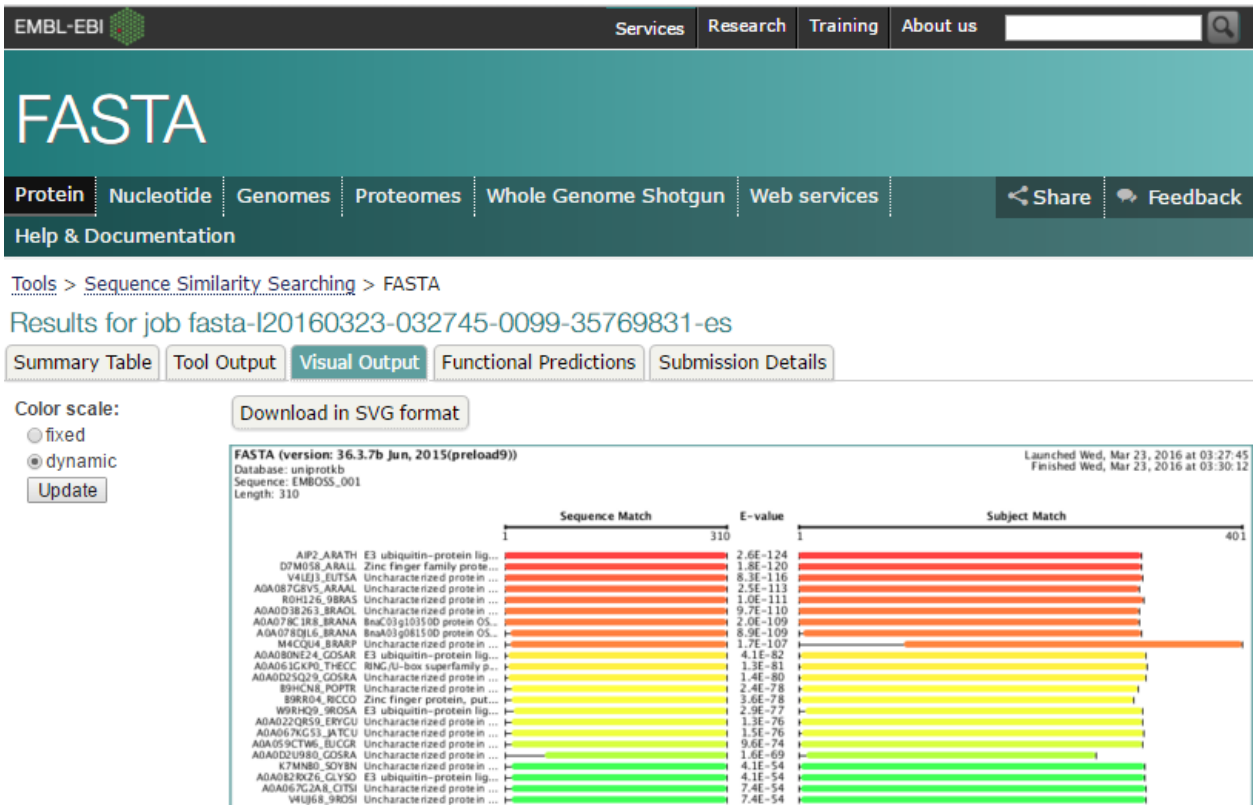


Figure 14.1.3: Results from FASTA

Module88: FASTA Algorithm

Text (7:00)

1. **INTRODUCTION**

   FASTA can search sequence databases and identify unknown sequences by comparing them to the known sequence databases by following Smith-Waterman algorithm. Smith-Waterman algorithm is also used for local alignment. This can help obtain information on the parent organism, function and evolutionary history.

2. **HOW IT WORKS?**

   - **Step 1:** Local regions of identity are found
   - **Step 2:** Rescore the local regions using PAM or BLOSUM matrix
   - **Step 3:** Eliminate short diagonals below a cutoff score
   - **Step 4:** Create a gapped alignment in a narrow segment and then perform Smith Waterman alignment

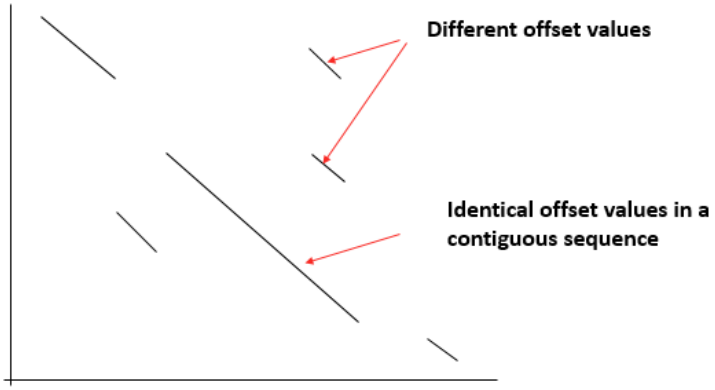


Figure 14.2.1: Identifying local regions (Step 1)

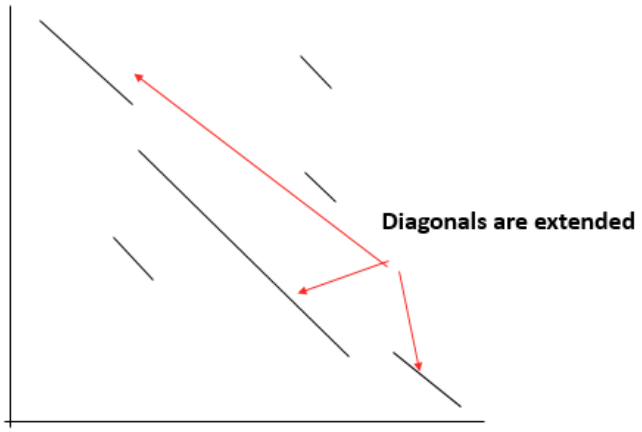


Figure 14.2.2: Rescoring using PAM or BLOSUM (Step 2)

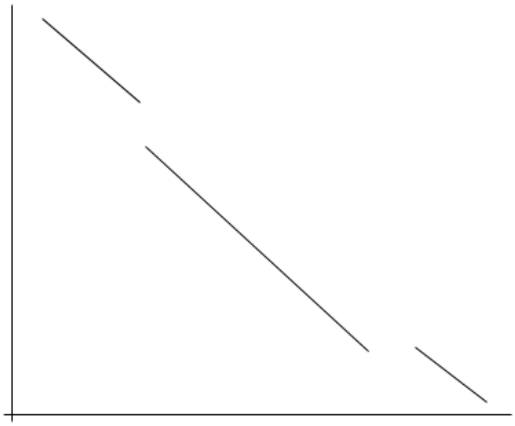


Figure 14.2.3: Eliminate short diagonals (Step 3)

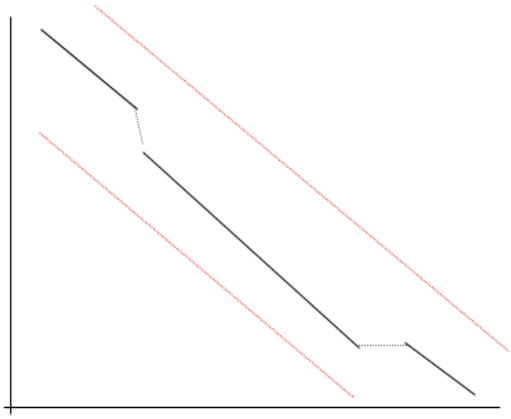


Figure 14.2.4: Perform alignment using Smith-Waterman (Step 4)

3. CONCLUSION

FASTA performs preliminary alignments quickly followed by Smith-Waterman. Results are given by visual format along with statistical scores.

Module89: Types of FASTA

Text (3:49)

1. INTRODUCTION

FASTA can search sequence databases and compare them against known DNA, RNA and protein sequences. It functionality based upon Smith-Waterman algorithm. There are many types of FASTA programs some are:

- **fasts35:** Compare unordered peptides to a protein sequence database
- **fastm35:** Compare ordered peptides (or short DNA sequences) to a protein (DNA) sequence database
- **Fasta35:** Scan a protein or DNA sequence library for similar sequences
- **Fastx35:** Compare a translated DNA sequence (6 ORFs) to a protein sequence database
- **tfastx35:** Compare a protein sequence to a DNA sequence database (6 ORFs)
- **fasty35:** Compare a DNA sequence (6 ORFs) to a protein sequence database

2. CONCLUSION

FASTA performs quick alignments on biological sequences. Several types of FASTA exist which can assist in comparing DNA/RNA/protein sequences with each other.

Module90: Summary of FASTA

Text (8:00)

1. INTRODUCTION

FASTA can briskly perform sequence search databases if given a query sequence. Multiple types of FASTA exist which assist in aligning DNA/RNA/protein sequences.

2. HOW IT WORK

- Obtain a query sequence
  For known sequence: Use NCBI, UCSC etc.
  For unknown sequence: Use NGS or Mass Spectrometry. It's a whole process by which we can find unknown sequence (we will briefly be discussed it in later chapters).
- Choose a type of FASTA
- Enter your input (query) sequence & set your parameters if you want otherwise use default parameters.
- Tabulated search results found (this step auto performed by FATSA)
- Select your desired align result

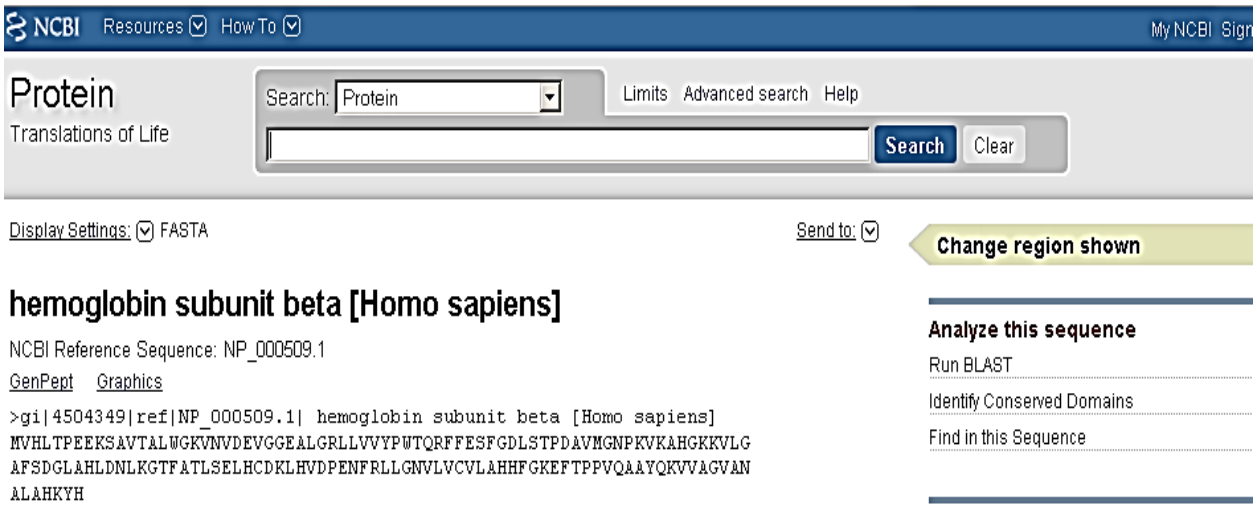


Figure 14.4.1: Obtain a query sequence (Step 1)

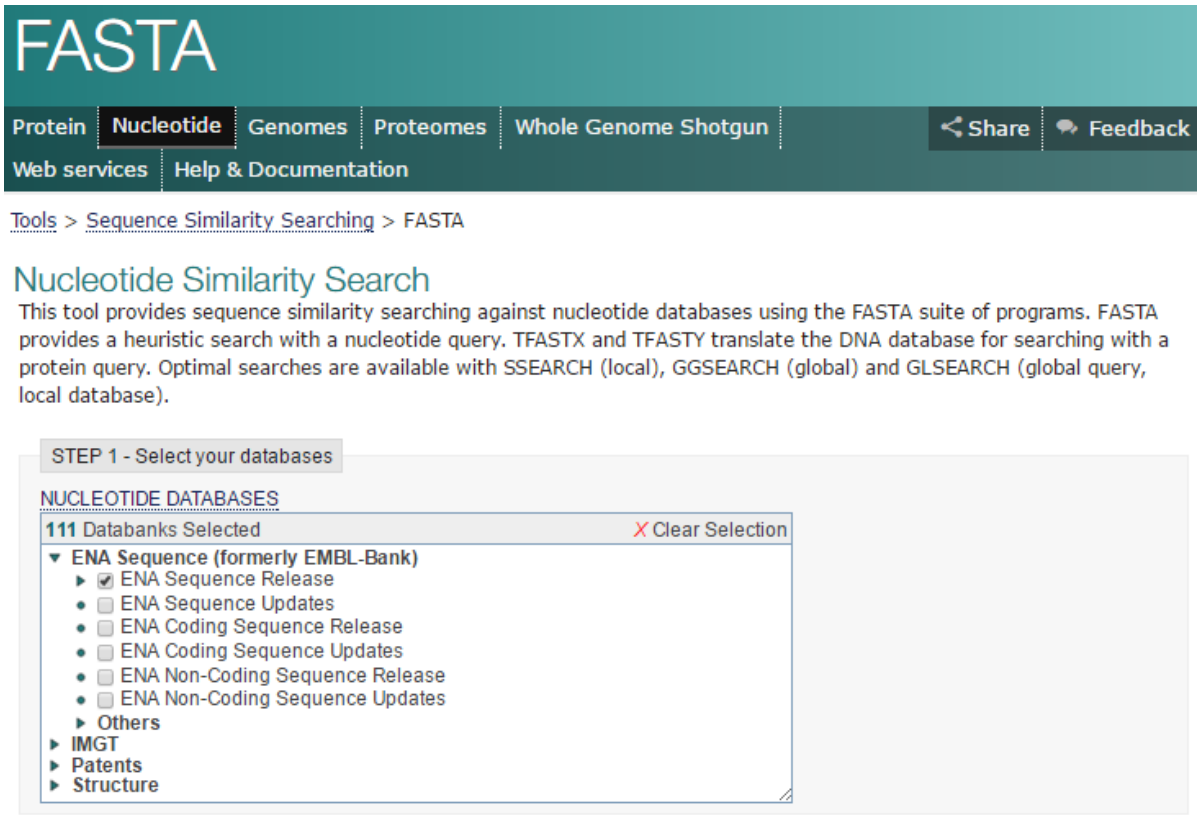


Figure 14.4.2: Choose a type of FASTA (Step 2)

fasta35, fasta35_t*	scan a protein or DNA sequence library for similar sequences
fastx35, fastx35_t	compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.
tfastx35, tfastx35_t	compare a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations.
fasty35, fasty35_t	compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.
tfasty35, tfasty35_t	compare a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations.
fasts35, fasts35_t	compare unordered peptides to a protein sequence database
fastm35, fastm35_t	compare ordered peptides (or short DNA sequences) to a protein (DNA) sequence database
tfasts35, tfasts35_t	compare unordered peptides to a translated DNA sequence database
fastf35, fastf35_t	compare mixed peptides to a protein sequence database
tfastf35, tfastf35_t	compare mixed peptides to a translated DNA sequence database
ssearch35, ssearch35_t	compare a protein or DNA sequence to a sequence database using the Smith-Waterman algorithm.

Figure 14.4.3: Type of FASTA, with it uses

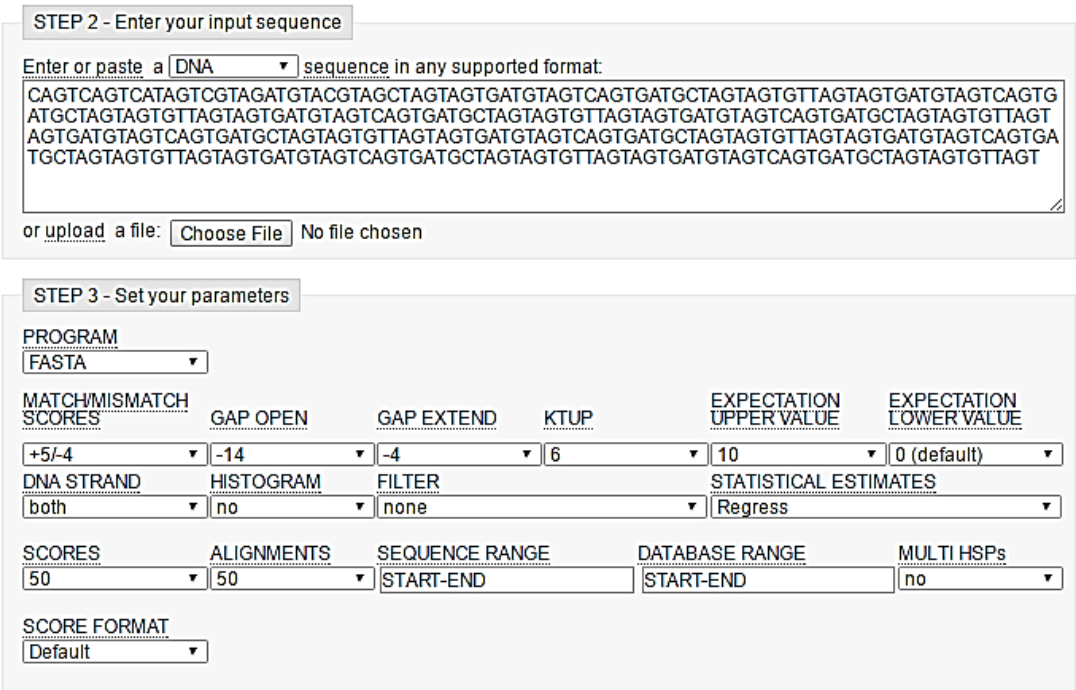


Figure 14.4.4: Setup Search Parameters (Step 3)

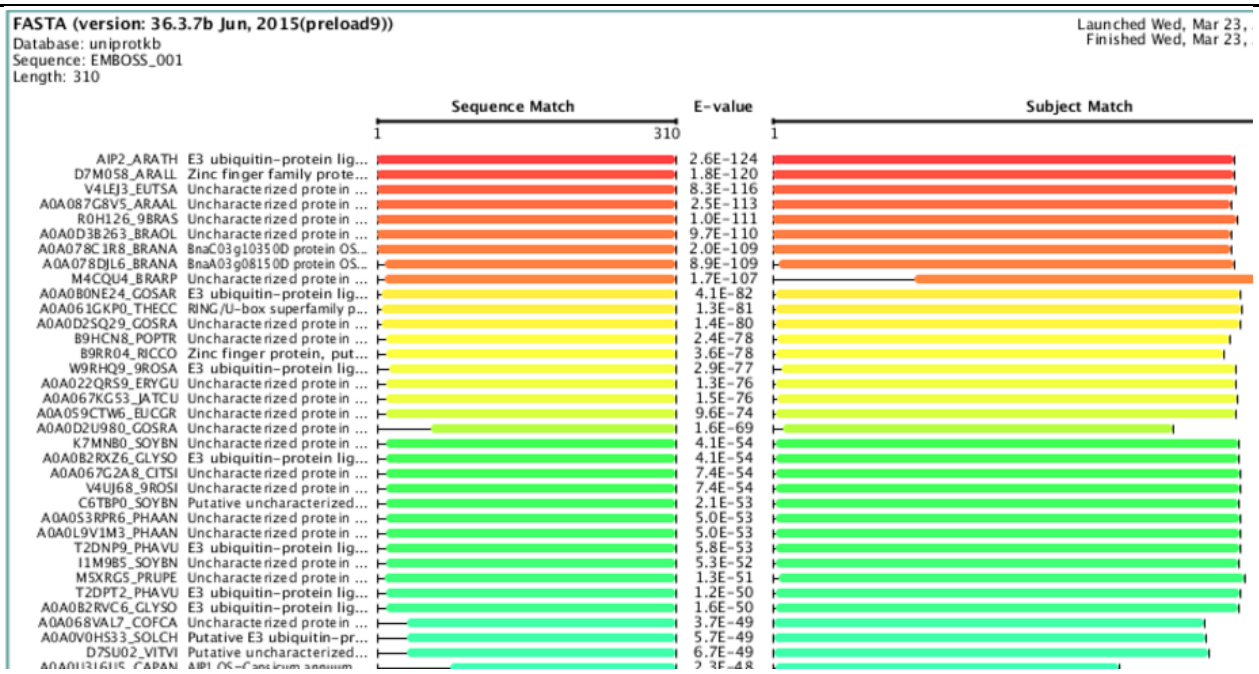


Figure 14.4.5: Tabulated Search Results (Step 4)

Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
☑ 1	SP:AIP2_ARATH	E3 ubiquitin-protein ligase AIP2 OS=Arabidopsis thaliana GN=AIP2 PE=1 SV=1 *Cross-references and related information in:* ▶ Gene expression ▶ Small molecules ▶ Nucleotide sequences ▶ Genomes & metagenomes ▶ Enzymes ▶ Samples & ontologies ▶ Molecular interactions ▶ Protein families ▶ Literature ▶ Protein sequences	310	453.0	100.0	100.0	2.6E-124
☑ 2	TR:D7M058_ARALL	Zinc finger family protein OS=Arabidopsis lyrata subsp. lyrata GN=ARALYDRAFT_488997 PE=4 SV=1 *Cross-references and related information in:* ▶ Nucleotide sequences ▶ Genomes & metagenomes ▶ Samples & ontologies	310	440.3	96.5	99.7	1.8E-120

Figure 14.4.6: Tabulated align data (Step 5)

3. CONCLUSION

FASTA is freely available online tool which performs quick alignments on biological sequences. Depending upon your need you can choose a specific type of FASTA to compare and score alignments.

Module91: Database

Text (7:00)

Defination 1: A shared collection of logically related data design to meet the information needs of multiple users in an organization the term database is often in erronously reffered to as a synonym for a ''database management system (DBMS)''

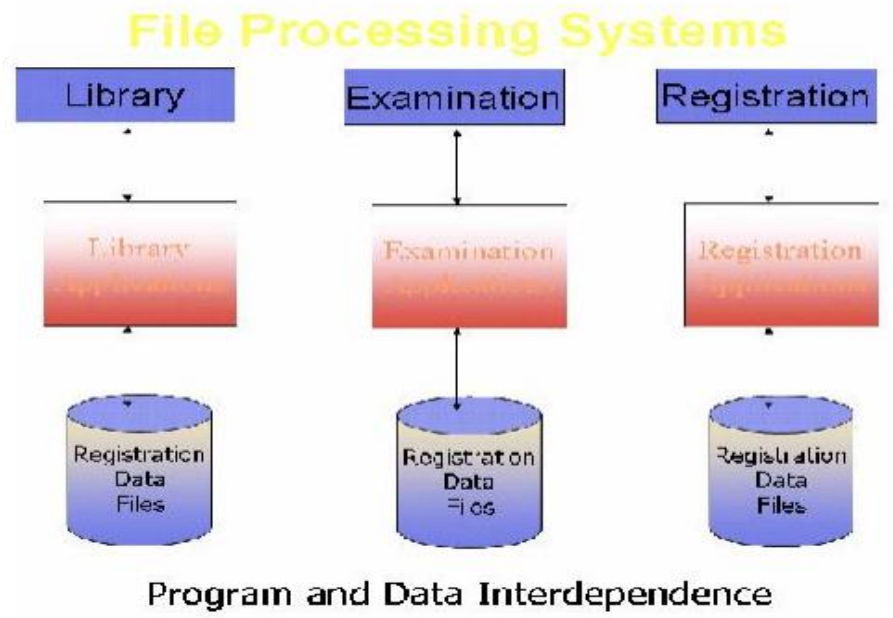
Database is not only being used in the commercial applications rather today many of the scientific engineering applications are also using databases less or more. Databases are concern of the respectively ladder form of appliations are more commercial applications .

Databases and Traditional file processing system:

Traditional file processing system is a simple file processing system refers to the first computer based approach of handling the commercial or bussiness applications. That's is why it is also called a replacement of the manual file system.

It is not necessary that we understand the working of the file processing enviroment for the understanding of the database and its working.

However a comprison between the characteristics of the two definatly helps to understand advantages of the databases and their working approach.

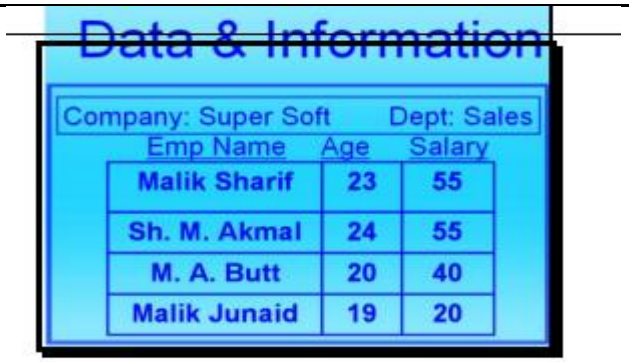


Module92: Database Advantages

Text (08:00)

Data and information:

Data is the collection of raw facts collected from any specific environment for a specific purpose. Data in itself does not show anything about its environment, so to get desired types of results from the data we transform it into information by applying certain processing on it . once we have processed data using different methods . Data is converted into meaningful form and that form of



the data is called information.

Database applications:

Database application is the group of programme which is used for performing certain operations on the data stored in the database . These operations may contain insertion of data into a database or extracting some data from the database based on a certain condition , updating data in the database, producing the data as output on any device such as Screen, disk or printer.

Database management system:

Database management system is software of collection of small programs to perform certain operation on data and manage the data.

Two basic operations performed by DBMS are;

- Management of data in the Database
- Management of users associated with the database

Module 93:Data Management

Data Base

Data Management

- Keeping track of a few dozen data items is straight forward
- However, dealing with situations that involve significant number of data items, requires more attention to the data handling process
- Dealing with millions - even billions - of inter-related data items requires even more careful thought
- Interactive software designed to improve the decision-making capability of their users
- The do not make decisions - just assist in the process

Issues in Data Management

- Data entry
- Data updates
- Data integrity
- Data security
- Data accessibility

DBMSes are popularly, but incorrectly, also known as 'Databases'

- A DBMS is the SW system that operates a database, and is not the database itself
- Some people even consider the database to be a component of the DBMS, and not an entity outside the DBMS

Module94: Database Software

Text (9:00)

Relational Databases (1)

- Databases consisting of two or more related tables are called *relational databases*
- A typical relational database may have anywhere from 10 to over a thousand tables
- Each column of those tables can contain only a single type of data (contrast this with spreadsheet columns!)
- Table rows are called records; row elements are called fields
- A relational database stores all its data inside tables, and nowhere else
- All operations on data are done on those tables or those that are generated by table operations
- Tables, tables, and nothing but tables!

RDBMS

- Relational DBMS software
- Contains facilities for creating, populating, modifying, and querying relational databases
- Examples:
  - Access
  - FileMaker Pro
  - SQL Server
  - Oracle

Module 95: Nucleotide sequence data base

Text (9:00)

Biological Databases:

Biological databases in general store biological data and their main goals are

- **Data storage**
- **Information retrieval**
- **Knowledge discovery**

Classification:

Biological databases can be classified as

- **Primary databases (**that stores the Primary Sequences**)**
- **Secondary databases (**the primary sequences are annotated and kept in Secondary Databases**)**
- **Specialized Databases (**they are dedicated towards some specific organism or can have some disease data**)**

Biological databases can also be classified on the bases of types of data which they contain, such as:

- **Nucleotide databases**
- **Protein databases**
- **RNA databases**
- **Genome databases**
- **Expression databases (**Gene Expression Databases**)**

Issues:

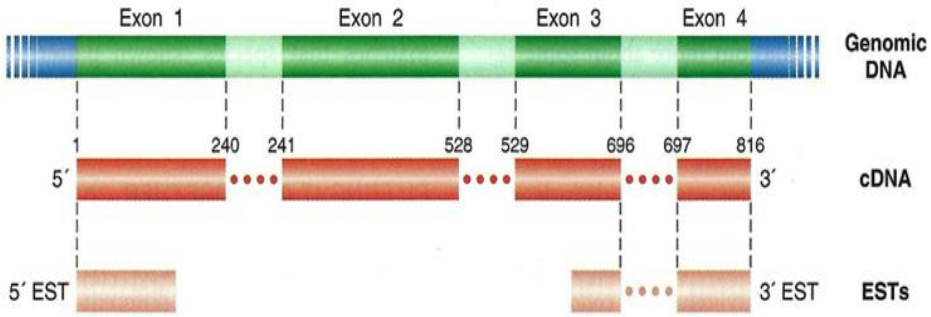
The issues which are present generally in other databases are also found to be in Biological databases that may be co-related with the relatively slow pace of quality assurance techniques as compared to the pace with which new data is emerging, so the issues are similar and are as follows:

Due to limited **Q/A**

- **Redundancy**
- **Inconsistency**
- **Incompatibility (format, terminology, data types, etc.)**

Nucleotide Sequence databases:

The Nucleotide Sequence Databases are one of the types of Biological Databases that contains nucleotide sequences in it, which can be DNA and cDNA or EST sequences.



Here, we have a diagram where we have a genomic DNA which has different *Exons*(we know that in Eukaryotes, we have exons and introns). So *exons*gets transcribed into mRNA and we can get cDNA from this mRNA through ***reverse***

transcriptionand then we can store this cDNA into our databases whereas the ESTs are the subsets within those cDNA's.

Conclusions:

In the end, we conclude some of the followings:

- Biological databases store biological data.
- **INSDC** is joint venture of NCBI, EMBL and DDBJ.
- Growth of bases in **GeneBank** is exponential, doubling every 18 months.

Module 96: Protein Databases

Text (9:00)

Introduction:

Protein databases store protein data which may include the following:

- **Protein sequences**
- **Motif (**patterns of amino acids**)**
- **Structure**
- **Structure alignments (**aligned structures**)**

Origin:

First sequences to be collected were Proteins (before Nucleotide Sequences) using **Sanger and Tupy's**methods (1951) where Common protein families like cytochromes were sequenced (as in that era people were focusing on the sequences made from cytochrome molecules).

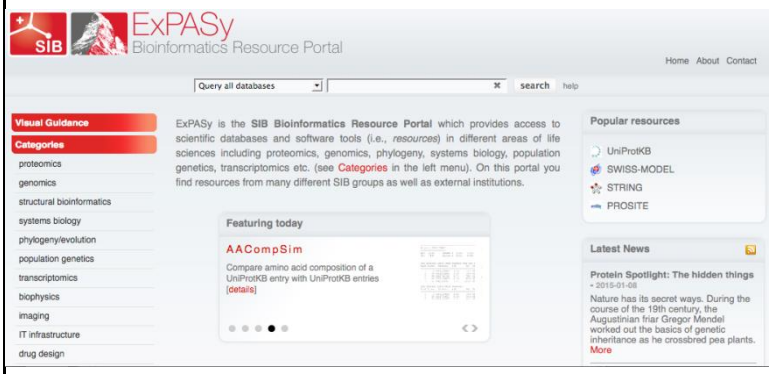
Atlas of protein sequences (mainly cytochromes) was assembled by Margret Dayhoff and her collaborators at **N**ational **B**iomedical **R**esearch **F**oundation (NBRF) in 1960s.

PIR (Protein Information Resource):

The collection (of ***Dayhoff*** and co) became **PIR (P**rotein **I**nformation **R**esource**)** which is now a collaboration of **NBRF**, **M**unich **C**enter for **P**rotein **S**equences (**MIPS**) and **J**apan **I**nternational **P**rotein **I**nformation **D**atabase (**JIPID**).

Protein Sequences:

Swiss-Protis a Collaboration between the SIB (**S**wiss**I**nstitue of **B**ioinformatics) and EBI (**E**uropean**B**ioinformatics**I**nstitute) and it weekly releases from about 50 servers across the world, the main source beingExPASy in Geneva (i.e. it'smainlycontrolled by ExPASywhichis the main server located in Geneva).



Here, is the page for ExPASy, and you can find different structural alignments, proteomic data, genomic data.

Conclusions:

We conclude that

- First sequences to be collected were Protein sequences.
- Protein databases are classified on the basis of sequences, motifs, structures and different structural alignments.
- Growth of Sequence in Databases is exponential (just like as in Nucleotide Databases the growth of sequence is higher)

Module 97: Genome and organisms specific data

Text (08:00)

Origin:

First attempt to sequence free living organism was launched in late 1990's (Blattner et al. 1997) and Viruses had already been sequenced (Fleischmann et al. 1995).

Haemophilusinfluenzaewas the first genome that was published and the project was initiated at **T**he **I**nstitute of **G**enome **R**esearch **(TIGR)** under the leadership of **Craig Ventor (**the same person who's name we'll see in the human genome project**).** At that time, a method which was already established known as **shotgun sequencing method** was being tested by this project to verify its reliability and efficiency. And by utilizing this method they sequenced the genome which was about 1.8 million base pairs (bp), it took 9-months and the cost was around 1 million US dollars and this project Paved the way for sequencing of many other organisms.

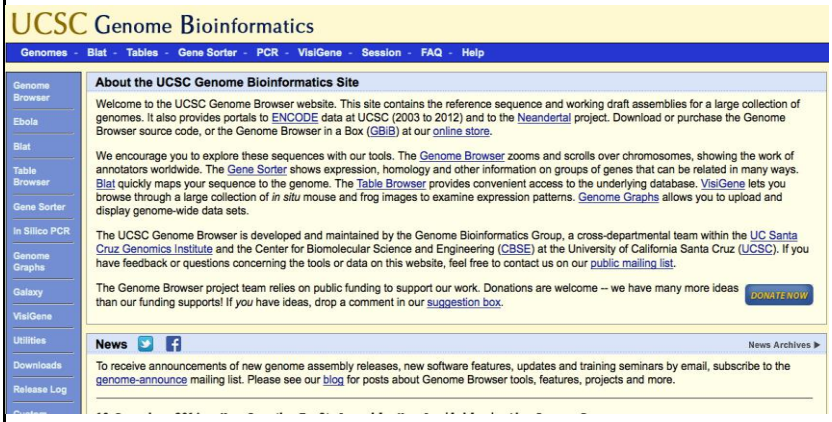
Examples

- **AceDB** (**A*C. e*legans*D**ata**B**ase) was the first genome database for genome sequences was developed in 1989 and was established by **Richard durbin**and**Thierry-Miegi.**

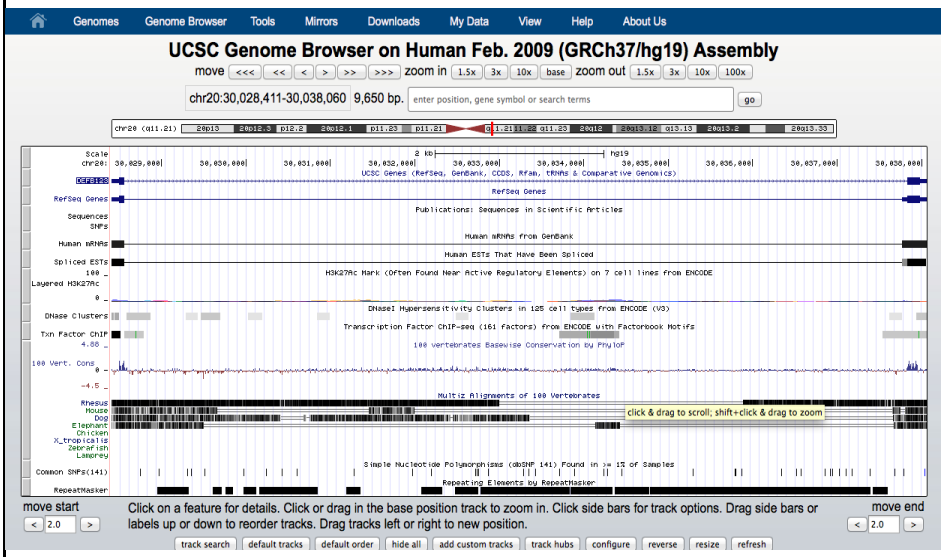
Human Genome Project:

Human Genome Project started initially as a Pilot Project which begun by **D**epartment **O**f**E**nergy (**DOE**) of USA in 1986. Two organizations, one is **N**ational **H**uman **G**enome **R**esearch **I**nitiative (**NHGRI**), which was federally funded organization through NIH (**N**ational **I**nstitute of **H**ealth) that started in 1988 by **Francis Collin** which was joined to Commercial organization named ***Celera***

(Celera Genomics)in 1998, a commercial under the leadership of **Craig Venter.**



While we have those genomes available, we want to see their graphical views where we can get the reports, get the idea about where different genes are located, so in order to do that we needed to make something which we call it as genome browsers- are the webpages where we can look into the different features within our genomes so UCSC is one of the example (shown on the left) which is University of California Santa Cruz which is the biggest genome browser.



The figure of UCSC Genome Browser, where we can have information, so on the top we see a chromosome and down below we see various lines which are known as different tracks (for snips, genes, EST's etc.) so we can look or zoom into different regions of the genome by using those genome browsers.

Conclusion:

In the end, we conclude the following:

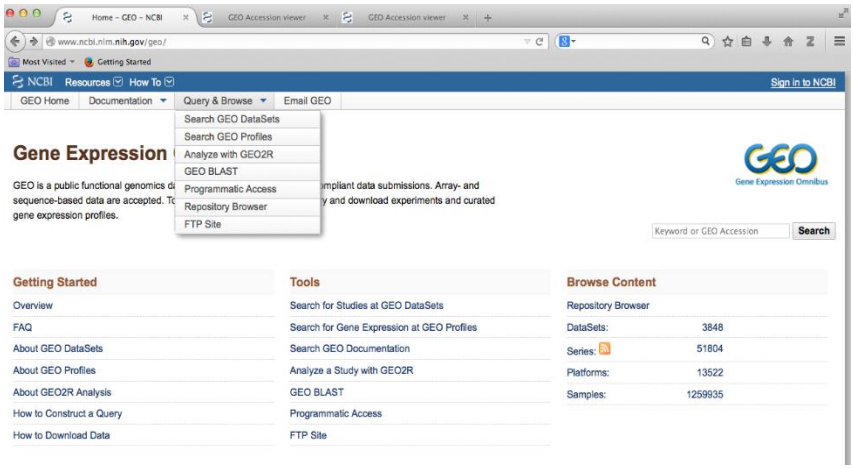
- Success of ***Haemophilusinfluenzae***paved the way for other genome sequencing projects
- Human Genome Project was accomplished by NHGRI and Celera (they were working independently from one another).

Module 98: Gene expression data bases

Text (9:00)

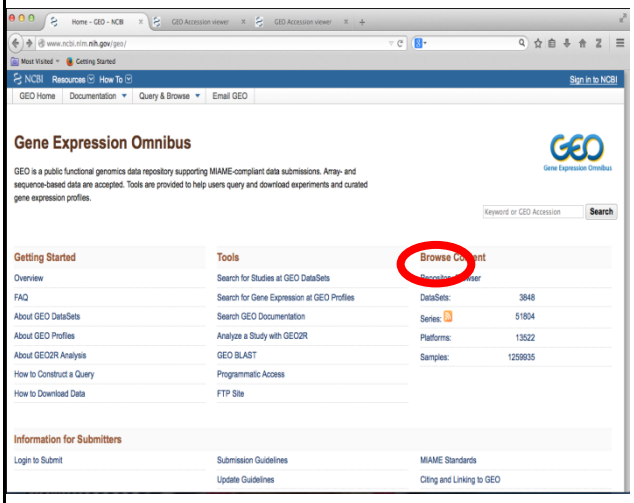
Gene Expression Omnibus (GEO):

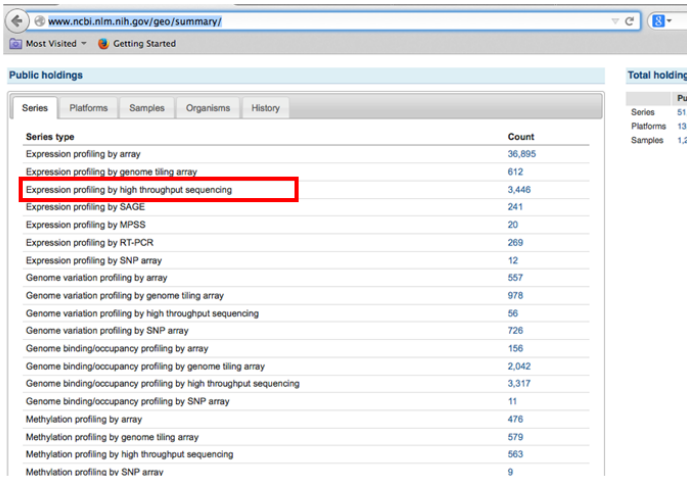
Genes are expressed into mRNA, and whenever we talk about gene expression, we generally mean the mRNA sequences so we can normally get those mRNA from techniques like microarray and another famous technique nowadays which is being established is known as RNAseq. And microarray data and RNAseq can be classified into Gene Expression Data which is stored in Gene Expression Databases.Gene Expression Omnibus is convenient for deposition of gene expression data, as required by funding agencies and journals and it's also a curated resource for gene expression data where we can do Browsing, querying, analysis and retrieval of the data.



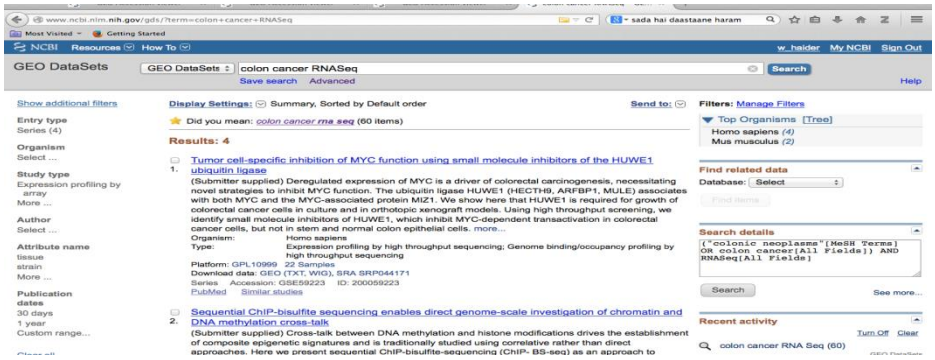
Here, is the webpage of GEO which is Gene Expression Omnibus running under NCBI (you can visit NCBI where you can get to the GEO Database) which are having different datasets, has expression profiles where we can see the change in expression of genes across different treatments and we can also analyze this expression data.

There is a tool called as GEO2R, we can use BLAST in it.

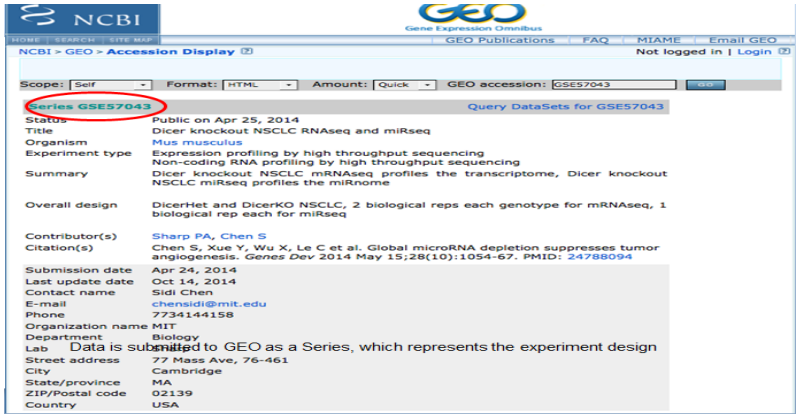




Here, is the Gene Expression Omnibus page and if we look into the different types of datasets it have, we can have *Series* (on the top left side of right figure), different records for the *Platform, Samples.* If you look into the types of series, you can see there are expression profiling by array, expression profiling by high throughput sequencing (in our course we'll be getting some RNAseq data which is under the expression profiling by high throughout sequencing), similarly there are other various techniques for getting the expression which are listed below in the *Series* section as can be seen and number of datasets available are also present in the column called as *count.*



If you want to look into some dataset, you can simply type into search bar say for example, you write colon cancer RNAseq data which leads us to the sets of records it gets and when we click onto one of them the page appears (shown below).

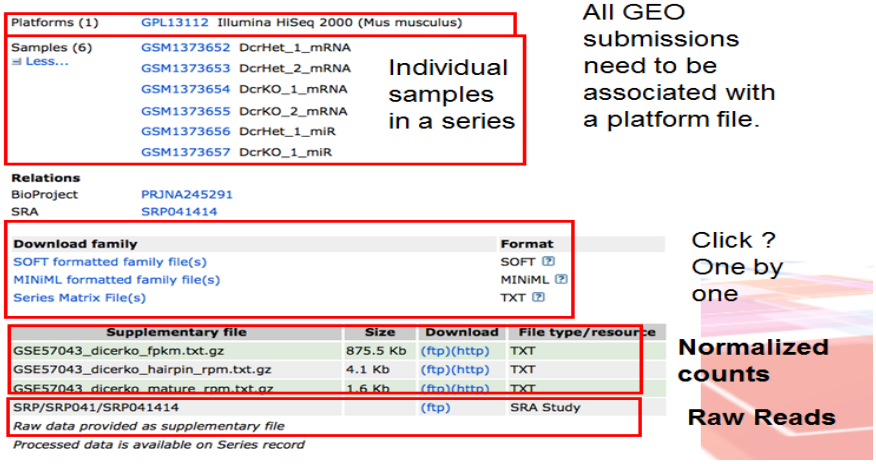


And we get this file, so here is the information of the particular dataset.

You can see on the top, that each dataset is submitted as a series with a unique number, say for example here the number is GSE57043 which is basically an id number for this dataset.

So, if we look into this page, we can have the idea about the experiment, the organism from which it's coming, type of experiment, a little bit summary of the experiment, we can also see the contributors name, their publications and addresses.

It is a huge page which is portioned and the other side of it is shown in the figure below



Conclusion:

So we sum-up that **GEO** is a public repository for the archiving and distribution of gene expression data and is the Best resource to get microarray and Next Generation Sequencing (RNASeq) data.

Module 99: Medical database

Text (9:00)

Introduction:

Informatics in health care may be called as health informatics. It deals with the resources, devices, and methods required in optimizing the acquisition, storage, retrieval, and use of information in health and biomedicine.

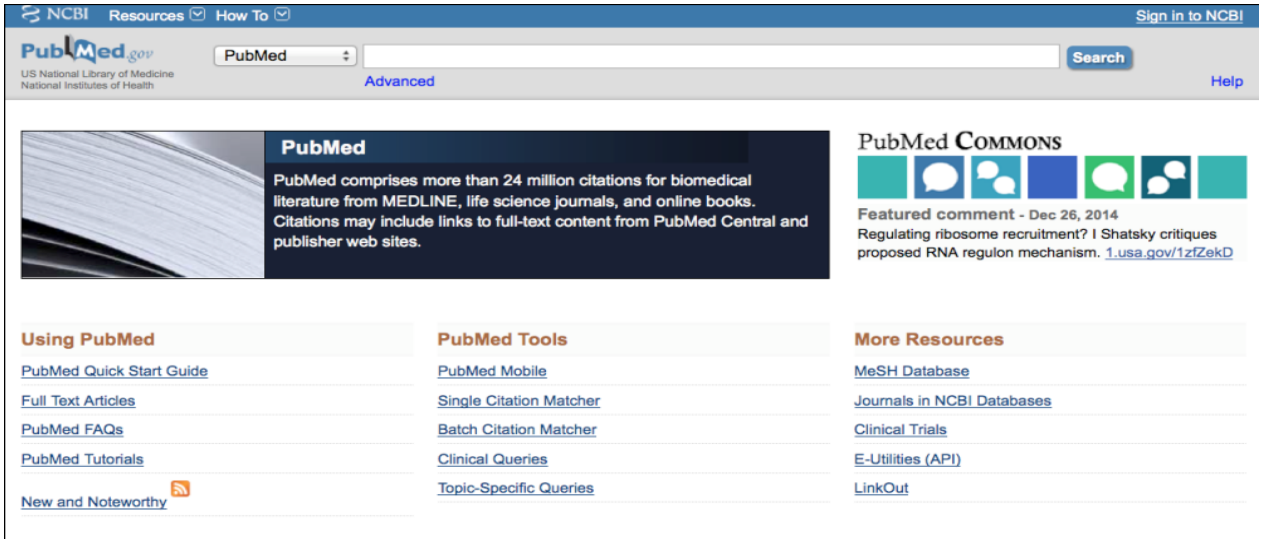
Medical databases store and provide medical information. The premier database for biomedical literature is the National Library of Medicine (NLM)'s MEDLINE, which is accessible through PubMed.

There are other databases where we can have medical information in addition to MEDLINE and are as follows:

- AcademicOneFile
- CINAHL (Cumulated Index of Nursing and Allied Health Literature)
- PsycINFO
- Web of Knowledge

PUBMED

PUBMED comprises of more than 24 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

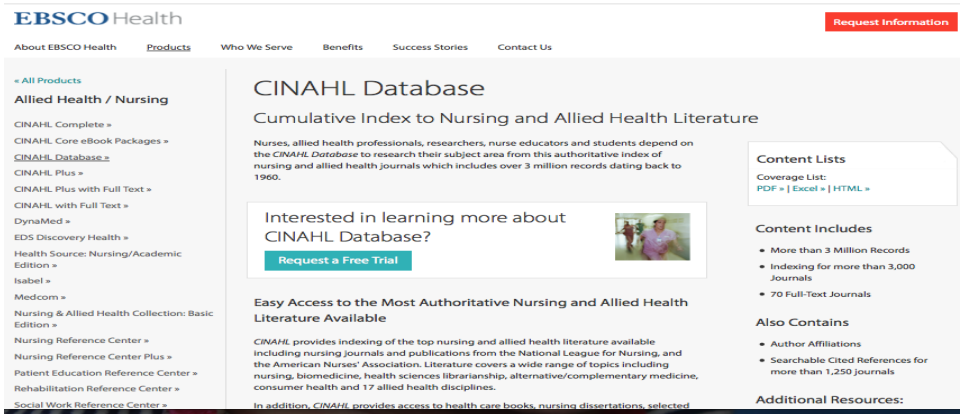


MEDLINE:

MEDLINE is the primary resource for biomedical journal articles and millions of citations to articles in biomedical journals can be found here.

Academic OneFile:

Academic OneFile lists articles from journals covering a broad range of subjects where you can also have medical data in it. It does not primarily focus on the medical topics but useful articles related to medical can still be found here in this database.



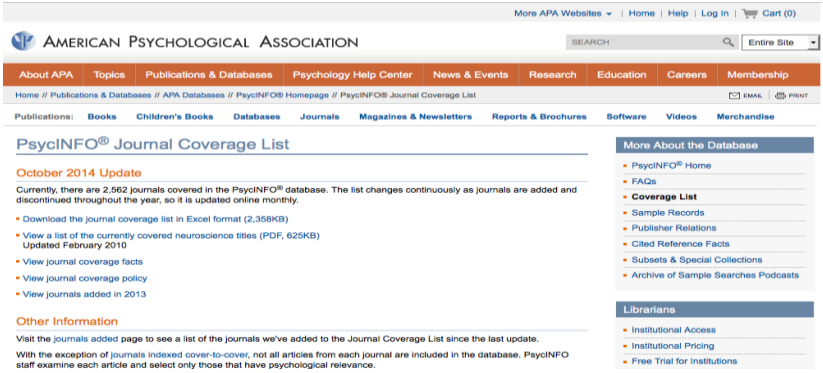
CINAHL:

Here, is a webpage of CINAHL Database (Cumulated Index of Nursing and Allied Health Literature), so you can have health literature from this database.

PsycINFO:

It is actually working under the ***American Psychological Association-APA*** (which also has literature referencing style).

PsycINFO searches the psychological literature while it does not primarily focus on medical topics, useful articles related to medical literatures can still be found here



Here, is the webpage for American Psychological Association (APA), where you can see PsycINFO- some *Journal* and its *Coverage List.* .

Web of Science:

It is a major source for articles in a wide range of fields, including the sciences, social sciences, and humanities and it is an outstanding place to find articles from scientific journals that may not be included in MEDLINE.

Conclusions:

In the end, we conclude the following:1. Informatics in health care may be called as health informatics.2. Medical databases deal with the acquisition, storage, retrieval, and use of information in health and biomedicine.

Module 100: Sequence Submission

Text (9:00)

Introduction

Sequences are submitted to the databases in order to share them with the scientific community

(sometimes they are also required by the Publication and funding agencies to submit them). Generally sequences are submitted at the time of publication and are reviewed by peers.

Caution

It is important to ensure that sequence files do not contain any special characters because sometimes the control characters can also be incorporated into or normal sequences, which can then mess-up the down-stream analysis or data-transfer.

Table 2.1. *Base–nucleic acid codes*

Symbol	Meaning	Explanation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	A or G	puRine
Y	C or T	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	C or G	Strong interactions 3 h bonds
W	A or T	Weak interactions 2 h bonds
H	A, C or T not G	H follows G in alphabet
B	C, G or T not A	B follows A in alphabet
V	A, C or G not T (not U)	V follows U in alphabet
D	A, G or T not C	D follows C in alphabet
N	A,C,G or T	Any base

Adapted from NC-IUB (1984).

Mount, pg 28

So, there is an issue of how to put the ambiguous nucleotides or amino acids in the sequences (because at some places you are not sure whether it is 'A' or 'T' or 'G' or 'C' and you are restricted to put a single letter). So, there is an organization known as International Union of Biochemistry (IUB), it has established some standard codes to represent those ambiguous bases or amino acids.

For example, here we see that G, A, T or C are just Guanine, Adenine, Thymine and Cytosine respectively. If we see R, it can be either A or G and the word is derived from the group they are coming from i.e. the puRines. We see Y that is the pYrimidine, it can be C or T.
M stands for if they are having some amine group / amino group in them, K is if they have Keto group i.e. G or T.
S is if they have strong interactions (3 hydrogen bonds) like C and G, who forms triple bonds.
W is for weak interactions, A or TSince H follows G in Alphabet so it's everything except G, it can be A, C or T and similar procedure is followed for B,V, and D whereas N can be any base.

Table 2.2. *Table of standard amino acid code letters*

1-letter code	3-letter code	Amino acid
A[a]	Ala	alanine
C	Cys	cysteine
D	Asp	aspartic acid
E	Glu	glutamic acid
F	Phe	phenylalanine
G	Gly	glycine
H	His	histidine
I	Ile	isoleucine
K	Lys	lysine
L	Leu	leucine
M	Met	methionine
N	Asn	asparagine
P	Pro	proline
Q	Gln	glutamine
R	Arg	arginine
S	Ser	serine
T	Thr	threonine
V	Val	valine
W	Trp	tryptophan
X	Xxx	undetermined amino acid
Y	Tyr	tyrosine
Z[b]	Glx	either glutamic acid or glutamine

Adapted from IUPAC-IUB (1969, 1972, 1983).
[a] Letters not shown are not commonly used.
[b] Note that sometimes when computer programs translate DNA sequences, they will put a "Z" at the end to indicate the termination codon. This character should be deleted from the sequence.

Mount, pg 28

Similarly, for amino acids, we have single letter codes i.e. from A to Z. And we can see in the figure on the left that some letters are missing.

There are four amino acids that are starting with G, but we gave that G letter to Glycine and for rest of them, we might use some other letters like Glutamic acid is represented

as E.

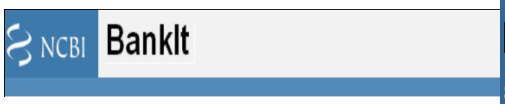
Y stands for Tyrosine (down below) and X can be any amino acid like N (in case of the nucleotide sequences).

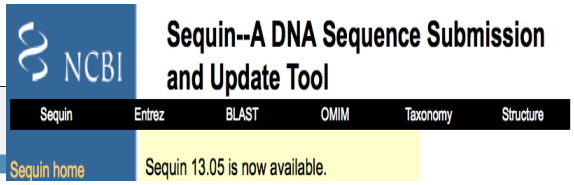
NCBI:

NCBI has two options for sequence submission

BANKIt - for simple sequences (not related with down-stream analysis) and annotations and can be submitted through web (if the datasets are small) which does not requires any advanced tools.

Sequin - For Complex sequences and annotations and is also good if we want to do some off-line submissions normally where we have our datasets which are huge ones and can be used in future with some advanced tools (for analysis) and graphical reports.

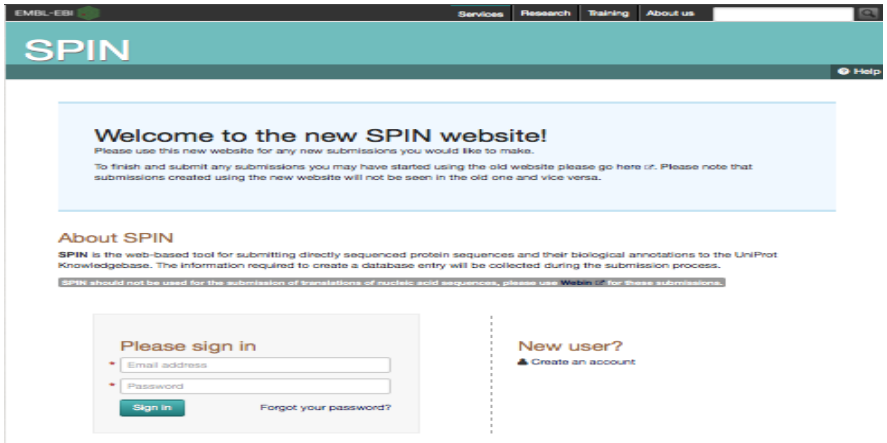




http://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank

In the figures above, are the glances BankIt and Sequin webpages.

UniProt:



For protein sequences, just like NCBI tools, we have UniProt and the similar tool is called as**SPIN** which is a web-based tool for submitting directly sequenced protein sequences and biological annotations to the knowledgebase.

Shown in this figure, is the webpage of SPIN.

Conclusion:

We conclude that sequences are stored in databases in specific format and when we want to submit them into a database then we need to follow the guidelines provided by those databases.

Module101: DNA Sequence Retrieval

Text (12:00)

Databases not merely collect and organize data (i.e. not only stores it) but allow intelligent data retrieval (we can do some down-stream analysis on those data sets). Let's see how we can get the

Gene | p53

Save search Advanced

Display Settings: Tabular, 20 per page, Sorted by Relevance Send to:

Did you mean p53 as a gene symbol?
Search Gene for p53 as a symbol.

Results: 1 to 20 of 9031 << First < Prev Page 1 of 452 Next > Last >>

Filters activated: Current only. Clear all to show 9271 items.

Name/Gene ID	Description	Location	Aliases	MIM
p53 ID: 2768677	CG33336 gene product from transcript CG33336-RB [*Drosophila melanogaster* (fruit fly)]	Chromosome 3R, NT_033777.3 (23049657..23054082, complement)	Dmel_CG33336, CG10873, CG31325, CG33336, D- DMP53, Dm-P53, DmP53, Dmel\CG33336, Dmp53, Dp53, dmp53, dp53, prac	
TP53 ID: 7157	tumor protein p53 [*Homo sapiens* (human)]	Chromosome 17, NC_000017.11 (7668402..7687550, complement)	BCC7, LFS1, P53, TRP53	191170

DNA data from the NCBI.

So, here is the webpage of NCBI, for example you want to search for say p53 gene; tumour suppressor gene. We write p53 on the search bar, then we get then results, so here we can find many ID entries like 9000 entries are there, we are just looking into the first page in this we choose the first two. So let's click the first one, the p53 where the ID is 2768677, there is a description that what sort of gene is it, and its actually coming from *Drosophila melanogaster*, the location is Chromosome number 3 and we see some Aliases; the alternative names of this gene. The link to NCBI is

p53 [*Drosophila melanogaster* (fruit fly)]

Gene ID: 2768677, updated on 4-Jan-2015

Summary

Official Symbol	p53 provided by FlyBase
Primary source	FLYBASE:FBgn0039044
Locus tag	Dmel_CG33336
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Drosophila melanogaster (old-lineage: Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora)
Lineage	Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora
Also known as	CG10873; CG31325; CG33336; D-p53; Dm-P53; Dmel\CG33336; dmp53; Dmp53; DmP53; DMP53; dp53; Dp53; prac

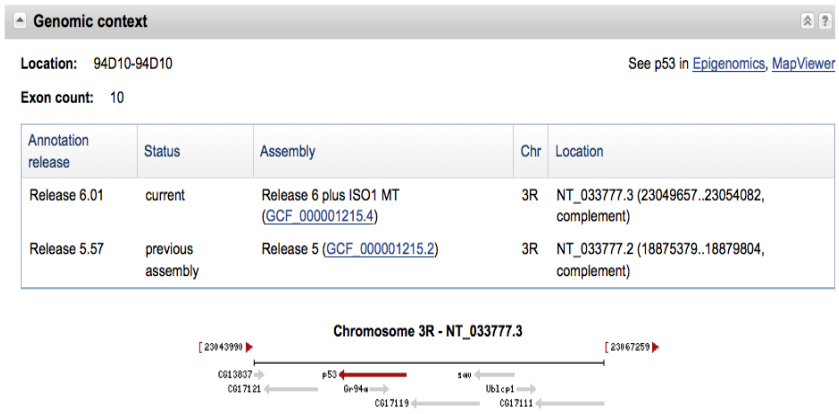
http://www.ncbi.nlm.nih.gov/.

When we clicked on the first gene as shown in the figure above, we now come to this webpage

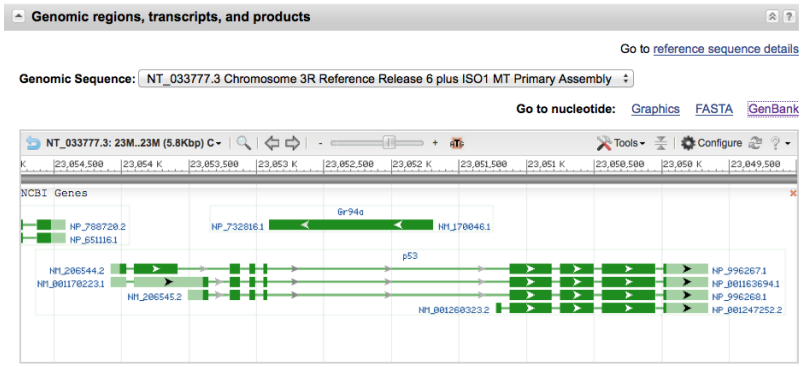
which is a huge page that is portioned into different figures.

In this figure (on the left), we can see the **summary** of this gene.

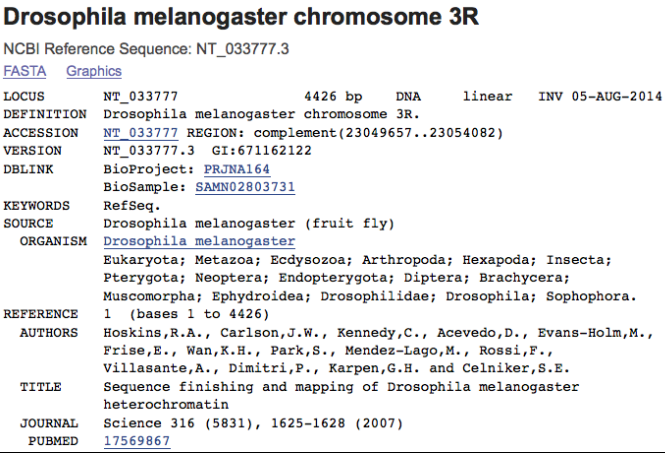
The official symbol is p53 provided by FlyBase which is also written in the *Primary source* (FlyBase is the databases that stores the genome of this fruit-fly *Drosophila*), then the *locus tag*, *gene type* is protein coding, RefSeq says reviewed (sometimes the genes are submitted and reviewed by some other scientist so it means that this gene has been REVIEWED). In the organism section, we see the classification of that organism and the Aliases are written beneath it.



In this figure, we can look into the structure of this gene and its coordinates (genomic coordinates), where we can see the location from where it is coming from, we can also see the orientations- the directions in which it is going (down below).



In this figure, we can see the genomic region, the transcripts and products tabs, we can look into the products of this gene (the gene when is expressed, the DNA is converted into the RNA). Since it's a eukaryotic genome where there is alternative splicing, so we can find different alternative splice variants of this gene.



Drosophila melanogaster chromosome 3R

NCBI Reference Sequence: NT_033777.3

FASTA Graphics

```
LOCUS       NT_033777               4426 bp    DNA     linear   INV 05-AUG-2014
DEFINITION  Drosophila melanogaster chromosome 3R.
ACCESSION   NT_033777 REGION: complement(23049657..23054082)
VERSION     NT_033777.3  GI:671162122
DBLINK      BioProject: PRJNA164
            BioSample: SAMN02803731
KEYWORDS    RefSeq.
SOURCE      Drosophila melanogaster (fruit fly)
  ORGANISM  Drosophila melanogaster
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Endopterygota; Diptera; Brachycera;
            Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE   1  (bases 1 to 4426)
  AUTHORS   Hoskins,R.A., Carlson,J.W., Kennedy,C., Acevedo,D., Evans-Holm,M.,
            Frise,E., Wan,K.H., Park,S., Mendez-Lago,M., Rossi,F.,
            Villasante,A., Dimitri,P., Karpen,G.H. and Celniker,S.E.
  TITLE     Sequence finishing and mapping of Drosophila melanogaster
            heterochromatin
  JOURNAL   Science 316 (5831), 1625-1628 (2007)
   PUBMED   17569867
```

On the upper right side of the figure, it is written as *Go to nucleotide, Graphics, FASTA and GeneBank* , so these are the different views with which we can get access to data files associated with this gene. When we click GeneBank, we are guided to another page, shown in the next figure.

We can see the entry in GeneBank and how

does it look.

Here, again we see the *name* of the gene, *locus* (where it's ID is written), length of the gene (it is 4426 BP), DNA, it is a linear type of DNA then we have the submission date.

Then the *definition* which is describing the organism's name, chromosome from which it is coming, then it has *accession* (the regions of the genome from which it is coming from), then we have the *version* (which is NT_033777.3, so there should have been .1 and .2 and since this is the third review, we can see .3 version here), we also see the reference (down below) and the authors from which this gene is coming and then their publications (it was seen to be published in Science).

```
FEATURES             Location/Qualifiers
     source          1..4426
                     /organism="Drosophila melanogaster"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:7227"
                     /chromosome="3R"
                     /genotype="y[1]; Gr22b[1] Gr22d[1] cn[1] CG33964[R4.2]
                     bw[1] sp[1]; LysC[1] MstProx[1] GstD5[1] Rh6[1]"
     gene            1..4426
                     /gene="p53"
                     /locus_tag="Dmel_CG33336"
                     /gene_synonym="CG10873; CG31325; CG33336; D-p53; Dm-P53;
                     Dmel\CG33336; dmp53; Dmp53; DmP53; DMP53; dp53; Dp53;
                     prac"
                     /map="94D10-94D10"
                     /db_xref="FLYBASE:FBgn0039044"
                     /db_xref="GeneID:2768677"
     mRNA            join(1..118,178..501,884..964,1035..1071,1135..1161,
                     2959..3268,3333..3579,3642..4036,4096..4426)
                     /gene="p53"
                     /locus_tag="Dmel_CG33336"
                     /gene_synonym="CG10873; CG31325; CG33336; D-p53; Dm-P53;
                     Dmel\CG33336; dmp53; Dmp53; DmP53; DMP53; dp53; Dp53;
                     prac"
                     /product="p53, transcript variant B"
                     /note="p53-RB; Dmel\p53-RB; CG33336-RB; Dmel\CG33336-RB"
                     /transcript_id="NM_206544.2"
                     /db_xref="GI:281362333"
                     /db_xref="FLYBASE:FBtr0084360"
                     /db_xref="FLYBASE:FBgn0039044"
                     /db_xref="GeneID:2768677"
```

When we scrolled down, we can see in the figure on left, that there are features of this gene so the total length of the gene is 4426.

We can see mRNA (down below), and since it's a eukaryotic gene, so mRNA is coming from the exons and the regions from which it came are shown below with the word *join*.

```
CDS                  join(75..118,178..501,884..964,1035..1071,1135..1161,
                     2959..3268,3333..3579,3642..4036,4096..4118)
                     /gene="p53"
                     /locus_tag="Dmel_CG33336"
                     /gene_synonym="CG10873; CG31325; CG33336; D-p53; Dm-P53;
                     Dmel\CG33336; dmp53; Dmp53; DmP53; DMP53; dp53; Dp53;
                     prac"
                     /note="CG33336 gene product from transcript CG33336-RB;
                     CG33336-PB; p53-PB; p53-like regulator of apoptosis and
                     cell cycle; Dmp53; protein 53; drosophila p53"
                     /codon_start=1
                     /product="p53, isoform B"
                     /protein_id="NP_996267.1"
                     /db_xref="GI:45553461"
                     /db_xref="FLYBASE:FBpp0083753"
                     /db_xref="FLYBASE:FBgn0039044"
                     /db_xref="GeneID:2768677"
                     /translation="MSLHKSASFSLTFNQNTSIVSRSNSRTIFEAFKEFLDFWDIGNE
                     VSAESAVRVSSNGAFNLPQSFGNESNEYAHLATPVDPAYGGNNTNNMMQFTNNLEILA
                     NNNSDGNNKINACNKFVCHKGTDSEDDSTEVDIKEDIPKTVEVSGSELTTEPMAFLQG
                     LNSGNLMQFSQQSVLREMMLQDIQIQANTLPKLENHNIGGYCFSMVLDEPPKSLWMYS
                     IPLNKLYIRMNKAFNVDVQFKSKMPIQPLNLRVFLCFSNDVSAPVVRCQNHLSVEPLT
                     ANNAKMRESLLRSENPNSVYCGNAQGKGISERFSVVVPLNMSRSVTRSGLTRQTLAFK
                     FVCQNSCIGRKETSLVFCLEKACGDIVGQHVIHVKICTCPKRDRIQDERQLNSKKRKS
                     VPEAAEEDEPSKVRRCIAIKTEDTESNDSRDCDDSAAEWNVSRTPDGDYRLAITCPNK
                     EWLLQSIEGMIKEAAAEVLRNPNQENLRRHANKLLSLKKRAYELP"
```

Then, within this mRNA we find the coding sequences (shown in the figure on the left), where coding sequences are the parts of the mRNA which are translated into the proteins so there are further sub-sets within those mRNA regions.

Down below, we see the translated version where we can see the word written as *translation,* and here we see the amino acid sequences coming from this gene.

```
ORIGIN
        1 cctggagcac ggaagattct tgcggacaca aatcgcaact gctaaataaa atttatttat
       61 ttgagtgcac agccatgagt cttcacaagt ccgcgtcgtt tagcttgact tttaaccagt
      121 gagcggagat attttattcg gtcttaccca acaaaataat gttgcgcctt tttgcagaaa
      181 cacttcgatt gtttcgcgta gcaatagtcg cacaattttt gaagctttca aggagttcct
      241 ggatttttgg gatatcggca acgaagtttc tgcagagtca gcagttcggg tctccagcaa
      301 cggagctttc aacttgccgc agagttttgg caacgaatcc aacgaatatg cccacctggc
      361 tacgcctgtg gatccagcct acggaggcaa caacacgaac aacatgatgc agttcacgaa
      421 caatctggaa attttggcca acaataattc cgatggcaat aacaaaatta atgcatgcaa
      481 caaattcgtc tgccacaagg ggtgagcaaa ttcaaaacac gcgctccaat cgataaacat
      541 tggctacggc gattgttcgc gctgcgtggc gaatggcaaa atccaaatag tcggtggcca
      601 ctacgattct gtagtttttt gttagcgaat ttttaatatt tagcctcctt ccccaacaag
      661 atcgcttgat cagatatagc cgactaagat gtatatatca cagccaatgt cgtggcacaa
      721 agaaaggtac agtgcggcaa caaattgatg atcgaacagt agaaaccttg catgtagcaa
       ------------------------------------
     4261 ggcatgttcg atggccgaaa agaaaacatt tttatatttt tgatagtata ctgttgttaa
     4321 ctgcagttct atgtgactac gtaacttttg tctaccacaa caaacatact ctgtacaaaa
     4381 aagccaaaag tgaatttatt aaagagttgt catattttgc aaacat
//
```

In the end, till we reach the word called as *origin,* and here we can see the actual nucleotide sequences which are present starting from 1 until the last nucleotide and the sequence ends with a double slash sign (//).

Conclusions:

So, we conclude that DNA Sequences are stored in DNA sequence databases in specified formats and Genebank format is a standard format.

Module102: Protein Sequence Retrieval

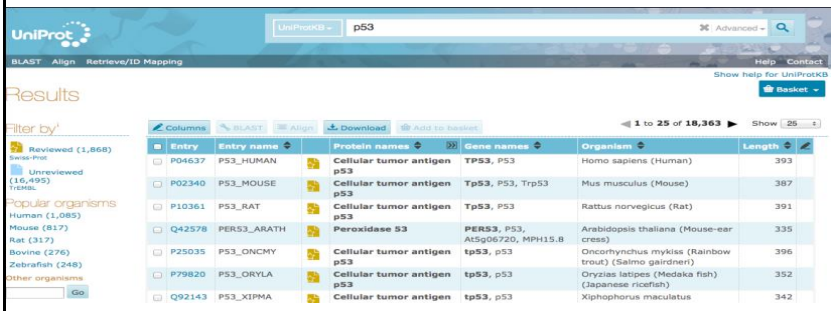
Text (15:00)

Protein Sequences:

Now we talk about the data retrieval and first we'll talk about the protein sequence retrievals and structures. Protein data is of the following types:

- Actual sequences (from the proteomic data or some other experimental techniques) or translated sequences (sometimes, we go to nucleotides databases, we get those nucleotides and then we translate them by using some softwares, so these are kind of predicted protein sequences) .
- Structures (we can also make structures from those proteins that maybe predicted or the real structures coming from various X-ray Crystallography Techniques).
- Annotations (sometimes, we are interested in the functions of the proteins so those are stored as annotations).

UniProt (*It is an international partnership between PIR, EBI and SIB*):

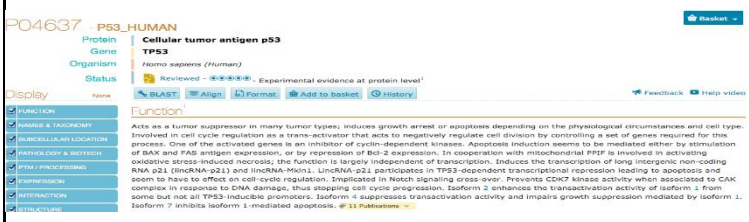


Now as far as the resources are concerned, we have multiple resources for protein sequences but **UniProt**claims to be the biggest and integrated resource whereas for the structures **PDB** seems like a good resource.

As shown in his figure, is the webpage for data retrieval from UniProt, so we want to search a protein, say p53, where we put it into the search box and press enter which gives us the output. And we see that there are 18,000 different records and it is showing us the first 25 out of them.

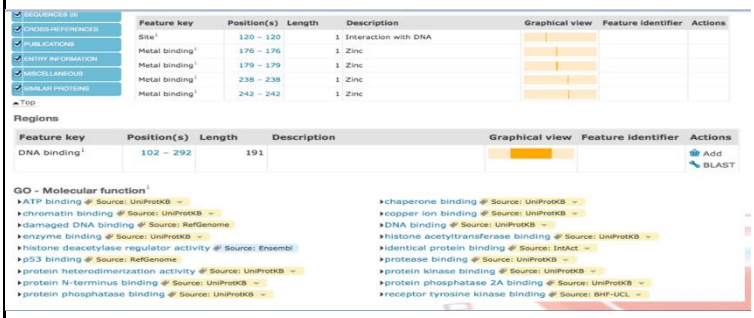
We can have different columns for the output on this webpage so we can have *entry; it's ID, entry name* (the Suffix Human is written so it's coming from Human, it can be from mouse, rat and Arabidopsis), the *protein name* is Cellular tumour antigen, then *gene name* which is TP53 (where TP stands for Tumour Protein), the *organism* is obviously the human (here) and in the end we have it's length i.e. 393 bp (base pairs).

The link to this webpage is **http://www.uniprot.org/uniprot/.**



So, let's check the first one and here we reach on the record for this protein (shown in the figure on the left) which is Cellular Tumour Antigen p53 protein, commonly known as TP53.

We can have different tabs, showing us the outputs. We can look into the *functions*, its *taxonomy*, and lot many other characteristics so if we look into the function so it gives us some description about what it's doing.



After scrolling the same webpage (shown in the figure on the left), we can see the *feature key* and in some *site* written (there are unique sites in different proteins which are having some specific properties in them so this is just one amino-acid present in this protein that *interacts with the DNA*). Similarly, there are different *metal binding sites* and we can see that it's mainly binding to the *Zinc*metal.The

number of amino-acids is shown here so these are the regions where it interacts with the metal.

Down below, we can also see the *DNA binding* region, for example here, the amino acids are from 102 to 292 and that is also shown in the *Graphical view* as well.

GO-Molecular function or GO-Gene Ontologies, so gene ontologies are the different functional annotation term, there they define different functions, so amongst them we have molecular functions, biological processes, and we have cellular components. So here we just see a *Molecular function,* so it tells us that it performs the functions as shown in the figure , mainly it's a *ATP binding, it's p53 binding* with various other functions like *DNA binding.* So all those functions related to these proteins are present in the heading of GO-Molecular Function.

Keywords - Molecular function
Activator
Keywords - Biological process
Apoptosis, Cell cycle, Host-virus interaction, Necrosis, Transcription, Transcription regulation
Keywords - Ligand
DNA-binding, Metal-binding, Zinc
Enzyme and pathway databases

Reactome	REACT_118568. Pre-NOTCH Transcription and Translation. REACT_1194. Activation of NOXA and translocation to mitochondria. REACT_121. Activation of PUMA and translocation to mitochondria. REACT_169121. Formation of Senescence-Associated Heterochromatin Foci (SAHF). REACT_169185. DNA Damage/Telomere Stress Induced Senescence. REACT_169325. Oncogene Induced Senescence. REACT_169436. Oxidative Stress Induced Senescence. REACT_20549. Autodegradation of the E3 ubiquitin ligase COP1. REACT_24970. Factors involved in megakaryocyte development and platelet production. REACT_309. Stabilization of p53.
SignaLink	P04637.

Next, we move on to some other functions, in the *Biological process* category (shown in the figure) we see that it is related to *Apoptosis* (which is a cell death and it is related to cell-cycle and some other components).

In the below section, we see some *enzymes and pathway databases,* and*Reactome*is a database in which we have a group of reactions which are categorized so these are the list of those reactions with which it is related.

Protein family/group databases

TCDB	1.C.110.1.1. the pore-forming pnc-27 peptide of 32 aas from the p53 tumor suppressor protein (pnc-27) family.

Names & Taxonomy

Protein names	*Recommended name:* **Cellular tumor antigen p53** *Alternative name(s):* • Antigen NY-CO-13 • Phosphoprotein p53 • Tumor suppressor p53
Gene names	*Name:*TP53 Synonyms:P53
Organism	Homo sapiens (Human)
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo
Proteomes	UP000005640: Chromosome 17

When we move further (as shown in the figure on the left) till we reach its *Taxonomy.*

On the top, we can see something written as *Protein family* or *group databases* whichis TCDB. Basically, there is another classification in which the proteins are classified on the basis of being as transporter proteins so it is associated with the transportation across the membranes and there is 5-digit number, so there is a specific classification code which is given to each protein, and this protein has the specific code as shown in the figure.

So then we have the *names and taxonomies,* where there are *protein names,* and thetaxonomyof the individual can be seen in the *Taxonomic lineage* row. Let's see how we reach to its sequence and is shown in the figure below:

Isoform 1 (identifier: **P04637-1**) [UniParc] FASTA Add to Basket

Also known as: p53, p53alpha

This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it.
This is also the sequence that appears in the downloadable versions of the entry.

« Hide

```
        10         20         30         40         50
MEEPQSDPSV EPPLSQETFS DLWKLLPENN VLSPLPSQAM DDLMLSPDDI
        60         70         80         90        100
EQWFTEDPGP DEAPRMPEAA PPVAPAPAAP TPAAPAPAPS WPLSSSVPSQ
       110        120        130        140        150
KTYQGSYGFR LGFLHSGTAK SVTCTYSPAL NKMFCQLAKT CPVQLWVDST
       160        170        180        190        200
PPPGTRVRAM AIYKQSQHMT EVVRRCPHHE RCSDSDGLAP PQHLIRVEGN
       210        220        230        240        250
LRVEYLDDRN TFRHSVVVPY EPPEVGSDCT TIHYNYMCNS SCMGGMNRRP
       260        270        280        290        300
ILTIITLEDS SGNLLGRNSF EVRVCACPGR DRRTEEENLR KKGEPHHELP
       310        320        330        340        350
PGSTKRALPN NTSSSPQPKK KPLDGEYFTL QIRGRERFEM FRELNEALEL
       360        370        380        390
KDAQAGKEPG GSRAHSSHLK SKKGQSTSRH KKLMFKTEGP DSD
```

In this figure, we can see the sequence of the protein which is found to be at the end of the page.

Here, it says *Isoform 1,* so different proteins have different isoforms, different alternative splice variants so this is Isoform 1 as exhibited by its name which is P04637-1, and is the kind of first isoform. We can see the sequence of the protein and starts with a methionine (always a first amino acid in those proteins) and ending at 390TH amino acid. So, it's a 393 aa long protein and the sequence is right here. You can click on the FASTA button on the top and then you can get this output in FASTA format (we'll discuss it later).

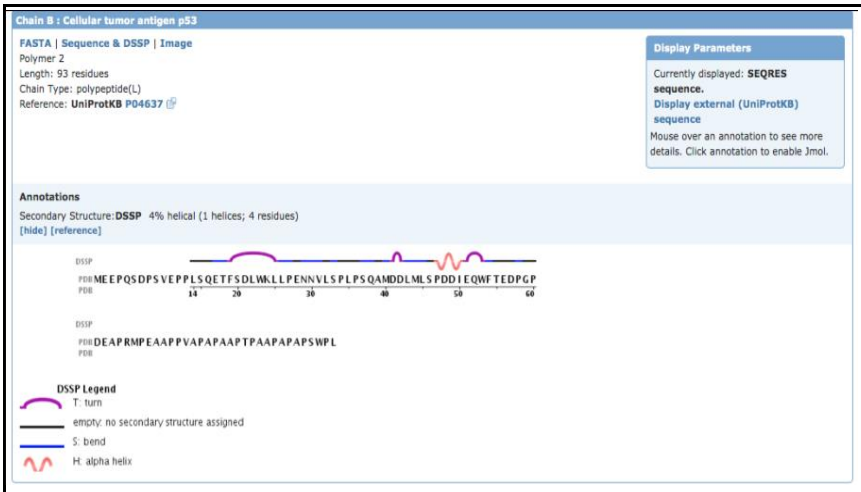
NCBI:

```
ORIGIN
        1 meepqsdpsv epplsqetfs dlwkllpenn vlsplpsqam ddlmlspddi eqwftedpgp
       61 deaprmpeaa ppvapapaap tpaapapaps wplsssvpsq ktyqgsygfr lgflhsgtak
      121 svtctyspal nkmfcqlakt cpvqlwvdst pppgtrvram aiykqsqhmt evvrrcphhe
      181 rcsdsdglap pqhlirvegn lrveylddrn tfrhsvvvpy eppevgsdct tihynymcns
      241 scmggmnrrp iltiitleds sgnllgrnsf evrvcacpgr drrteeenlr kkgephhelp
      301 pgstkralpn ntssspqpkk kpldgeyftl qirgrerfem frelnealel kdaqagkepg
      361 gsrahsshlk skkgqstsrh kklmfktegp dsd
//
```

We can also get the same protein from NCBI (as shown in the figure on the left)

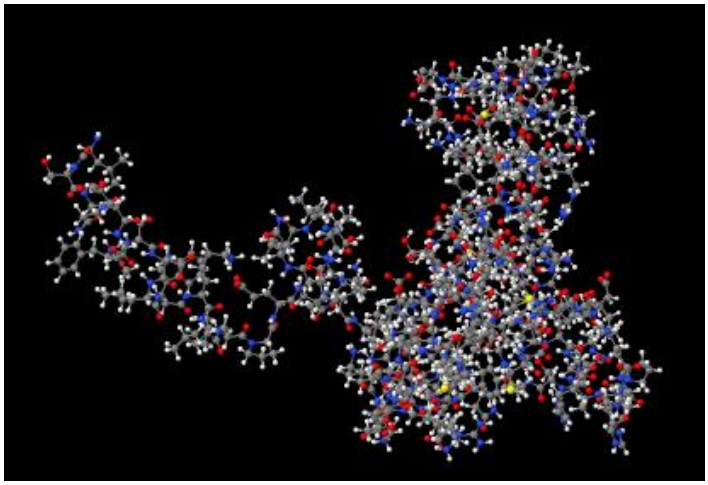
In NCBI, obviously the sequence is pretty similar and the arrangement is slightly different so it is *ORIGIN*, where the sequence starts and sequence ends at those two slashes (//). So, we can get the protein sequence from NCBI as well and the link to this website is http://www.ncbi.nlm.nih.gov/.

PDB:



PDB gives us the structures, so we can go to PDB webpage (as shown in the figure on the left) and search for the same ID i.e. P04637 and it gives us the sections or the regions from where it can make up some specific structures.

You can see the *turns* in Annotations section, the black ones are the empty lines where no secondary structure can be formed, blue ones show those bends and the orange ones are designated as alpha helices regions. So in PDB, we can have structures in this format as well as the 3D-Structures as shown in the figure below:



Conclusions:

We conclude that:

- UniProt is the integrated resource between PIR, EBI and SIB and
- PDB is a good resource to get the protein structure.

Module103: Sequence Formats

Text (07:00)

Sequences are stored in different formats in databases and since software requires those sequences in specific format so it's good to have an idea about what major formats are, we'll look into few of them.

FASTA Sequence Format

FASTA is the most recognized and well distributed format to present DNA and Protein sequences.

The sequence starts with a '*greater than*' sign (>), whereas the actual sequence is always on the next

line. It is recommended that all lines of text should be shorter than 80 characters in length (generally we have 60 characters).

Example:

>gi|568815581:c7687550-7668402 Homo sapiens chromosome 17, GRCh38 Primary Assembly
GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCATCTAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCT
CTCCACGACGGTGACACGC--------
>gi|120407068|ref|NP_000537.3| cellular tumor antigen p53 isoform a [Homo sapiens]
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLAPPVLGFLHSGTAKSVTCTYSPALNKMFCQLAKT--*
This sequence is of DNA in the *fasta* format (shown above), which starts with the '*greater than*' sign (>), same as in the case of the protein sequence in the *fasta* format (shown below) that also starts with the same symbol.

Then we have '*gi'* written which stands for 'gene identification' and the numbers shown are the 'ID' in both of the sequences.

In the DNA sequence, we have 'c' followed by the 'ID', this 'c' basically means the sequence is of the complementary strand and the regions from where it is coming are designated here; the base positions in between them, this gene is located. Then we have a short description of this gene that it belongs to '*Homo sapiens* ', 'chromosome 17' and the 'Primary Assembly' (assembly is where we get short sequence reads or small sequences and we put them together into a gene, known as assembly). Then finally, we have the actual sequence which is around 60 characters long in each line (as the sequence was quite long, we have used dashes to represent further characters).

In the protein sequence, we have 'ref' followed by the 'ID', which gives us an idea that it is a reference sequence (reference sequences are the curated sequence, there is a sub-section in NCBI called as *ref seq*, so they have reference sequences ; a kind of standard sequences to avoid any kind of redundancy. So, we can say these are the primary or the main sequences and we might have other alternative splice variants but references are the kind of true representative of the class). Followed by ref, we have another ID, which is the 'protein ID'. Then we have its brief description that it's a 'cellular tumor antigen' protein 'p53 isoform' and is also from '*Homo sapiens'*. Finally, we have the actual sequence of this protein and in the end we have dashes that represents it is an incomplete sequence and steric (*) is shown (sometimes the steric (*) is found to be seen in *fasta*files but sometime it don't, so the software must know what does this specific steric (*) stands for).

GeneBank Sequence Format:

GeneBank sequence format is found to be in the GeneBank Database which is a kind of standard format and other formats are pretty similar to it.
A sequence file in Gene Bank format can contain several sequences. One sequence 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 ("//").

EMBL Format:
This format is similar to that of GeneBank Format. An example sequence in EMBL format is:

ID AA03518 standard; DNA; FUN; 237 BP.

```
XX
AC   U03518;
XX
DE   Aspergillusawamori internal transcribed spacer 1 (ITS1) and 18S
DE   rRNA and 5.8S rRNA genes, partial sequence.
XX
SQ   Sequence 237 BP; 41 A; 77 C; 67 G; 52 T; 0 other;
aacctgcggaaggatcattaccgagtgcgggtcctttgggcccaacctcccatccgtgtc        60
tattgtaccctgttgcttcggcgggcccgccgcttgtcggccgccgggggggcgcctctg       120
ccccccgggcccgtgcccgccggagaccccaacacgaacactgtctgaaagcgtgcagtc  180
tgagttgattgaatgcaatcagttaaaactttcaacaatggatctcttggttccggc  237
//
```

Here, we have ID, accession number (AC), descriptions (DE), and the sequence actually starts from where the word 'SQ' is there, and we can observe that we have pretty similar lines as seen in the previous example. Finally, the sequence ends with doubles slashes same as in GeneBank format.

SwissProt Format:

SwissProt protein sequence format is similar to EMBL format but there is considerably more information about physical and biochemical properties of a protein (as you can see below there is more description).

ID - Identification.
AC - Accession number(s).
DT - Date.
DE - Description.
GN - Gene name(s).
OS - Organism species.
OG - Organelle.
OC - Organism classification.
RN - Reference number.
RP - Reference position.
RC - Reference comments.
RX - Reference cross-references.
RA - Reference authors.
RL - Reference location.
CC - Comments or notes.
DR - Database cross-references.
KW - Keywords.
FT - Feature table data.
SQ - Sequence header.
// - Termination line.

XML Format:

It is a modern practice in which we try to put those sequences in kind of a machine language. So, XML stands for Extensible Markup Language. The format is similar to HTML (language for Web

programming).
The good part is that this language is in between machine and man readable so it's kind of easy to code over this.
And it is becoming standard data format for transferring genome data.

GCG FORMAT:

GCG stands for Genetics Computer Group (basically it was a group of scientists who were helping the biological community to develop different software and training programs to help with the biological sequence analysis problems, so they also came up with the sequence formats). This format is kind of similar to the NBRF format (we have checksum but we don't have greater than (>) sign as in fasta, we have length of the sequence). There can be multiple sequences in one file.

Sequence converters:

Sometimes, we need to convert between sequences so you can come up with your own script or you can come up with your own codes and there are also some programs meant for this purpose alone such as READSEQ is a useful sequence converter (developed by D.G.Gilbert at Indiana University, USA) basically it recognizes DNA or Protein sequence file and interconvert them between different formats.

Conclusions:

What we conclude in the end of this lecture is the following:

- Databases store sequences in specified formats
- Genebank, DDBJ and EMBL has similar formats
- Different software need sequences in different formats

We might convert the sequences into other formats on our own or we can also simply use one of the programs available for converting like READSEQ

Module104: Data Retrieval

Text (9:00)

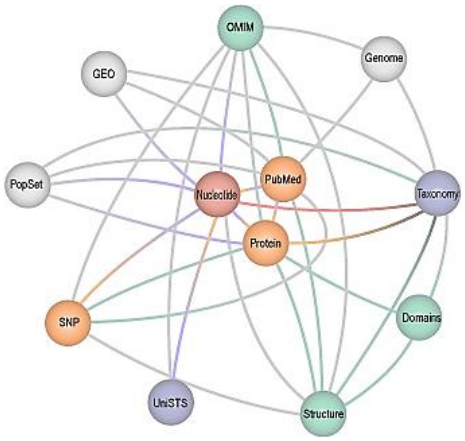
Data Retrieval:

Nearly all biological databases are available for download as simple text (flat) files. Sometimes we are interested to download the database and do the analysis locally in our own machines which might save our time as the local version of the database allows one greater freedom in processing the data.

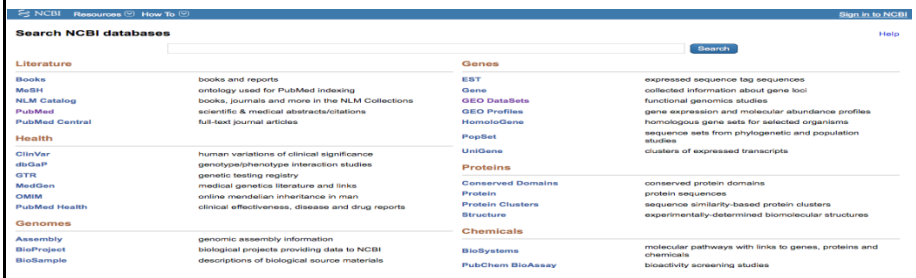
ENTREZ:

It is an integrated search engine that works behind NCBI, so you can do lot of researches and can look for variety of data using it (It can be accessed from the site www.ncbi.nlm.nih.gov/Entrez/). It integrates PubMed and 39 other scientific literatures, nucleotide and protein databases. For

example, it can be **protein domain data**, **population studies**, **expression data**, **pathways**, **genome details** and **taxonomic information.**



Here, we can see it integrates between GEO (gene expression sets), OMIM (Online Mendelian Inheritance in Man), Genome Databases, taxonomy Databases, etc. And we can see that in the middle we have Nucleotide, PubMed and Protein. So it is an integrated system which operates between different databases, so you can simply search for whatever you are looking after and ENTREZ will search it for you.



Here, is the page of ENTREZ that allows you to search anything by the help of a search bar at the top. It has different connections like we have Literature resources, we have Health Databases, Genomes, different Genes Databases, Proteins and Chemicals.

Bulk Data Retrieval:

Sometimes, we need to obtain data in bulk amount and for this purpose normally we use Linux but for Windows users, there are some packages or programs available and are known as FTP clients so the best option is to use FTP (File transfer protocol). The File Transfer Protocol (**FTP**) is a standard network protocol used to transfer files Via command line or application programs like FTP clients (we'll be using it).

Once, we get the data which is mostly not in a proper format and every other software require different specific formats so we might want to use some programming languages to help convert the data into the required format. The programming languages like PERL and Python are good for processing Biological data in Bioinformatics.

Conclusions:

We have learned that :

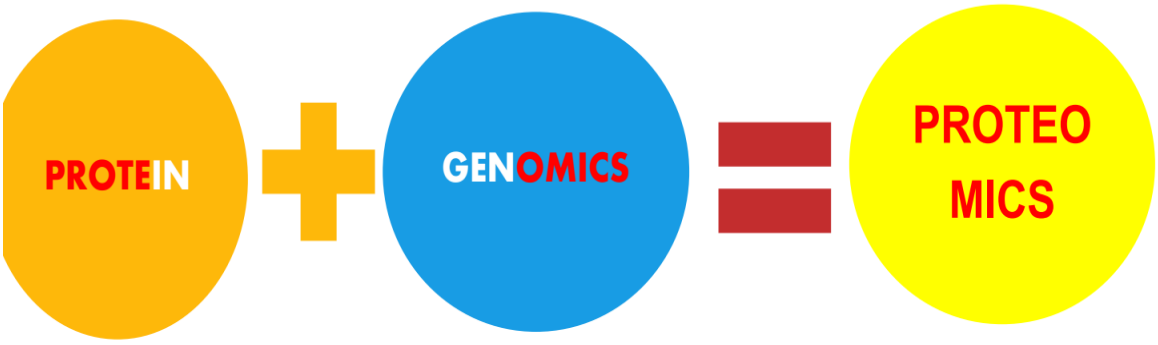
- Data is transferred over the internet.

Data needs to be transformed or processed before handing it over to any software.

Module105: Why Proteomics

Text (9:00)

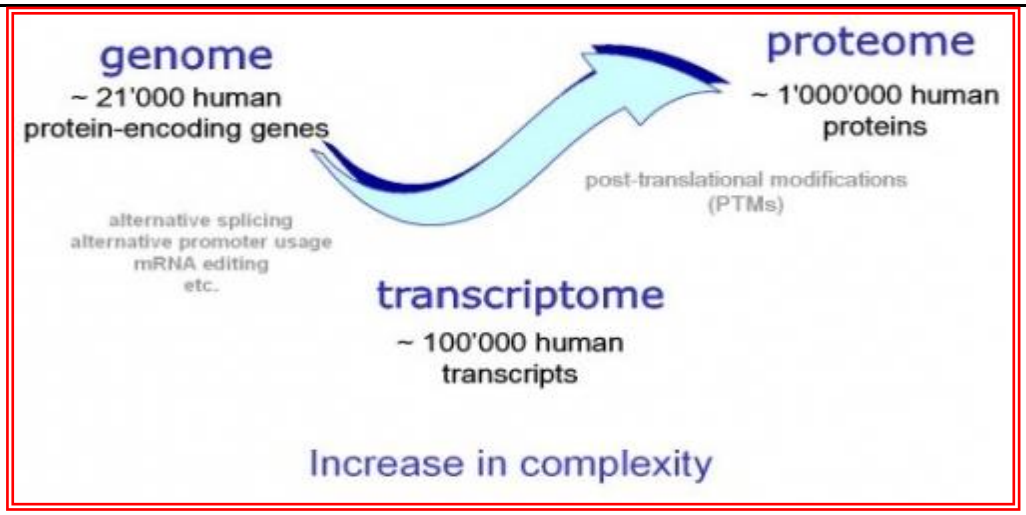
- Proteomics is the large-scale study of proteins, usually by biochemical methods. The word proteomics has been associated
- traditionally with displaying a large number of proteins from a given
- cell line or organism on two -dimensional polyacrylamide gels



- Many types of information cannot be obtained from the study of genes alone. For example, proteins, not genes, are responsible for the phenotypes of cells. It is impossible to elucidate mechanisms of disease, aging, and effects of the environment solely by studying the genome.

Aims of Proteomics

- Genomics integrated strategies
- Study of post-translational modifications
- Identification of novel protein targets for drugs
- Analysis of tumor tissues
- Comparison between normal and diseased tissues
- Comparison between diseased and pharmacologically treated tissues

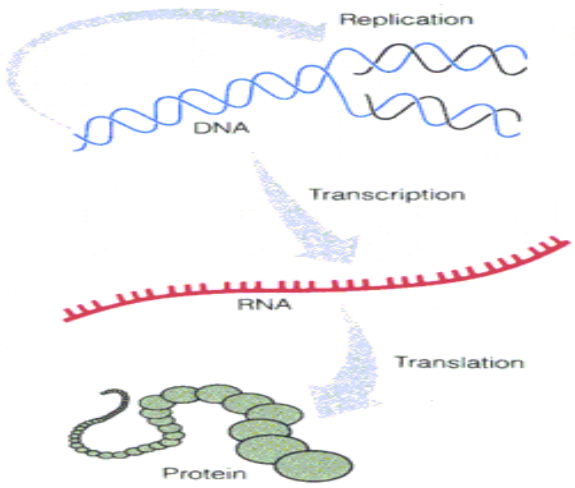


Module106: Central dogma of molecular biology

Text (8:00)

The central dogma:

The central dogma states that information in nucleic acid can be perpetuated or transferred, but the transfer of information form into protein is irreversible.

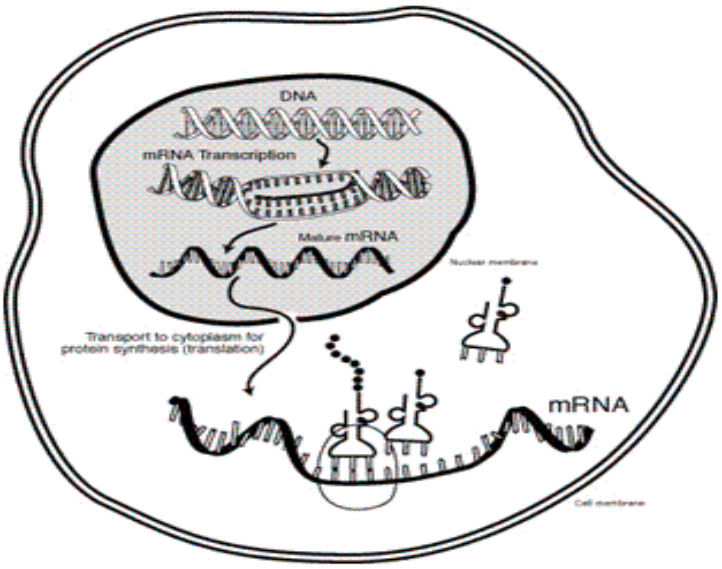


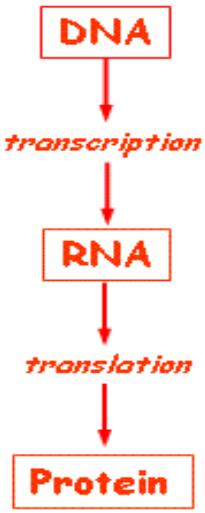
Central Dogma of Biology:

- DNA -> RNA -> Protein Synthesis
- Transcription:
    - Process of DNA serving as a template for RNA synthesis
- Translation:
    - Process of RNA serving as a template for protein synthesis

How do DNA and genes relate to proteins?

- DNA provides the genes, or genetic code, for protein synthesis
- Genes are expressed because DNA codes for RNA which then codes for ALL of our proteins





Background: 3 Types of RNA

- mRNA: Messenger RNA
  - 1st RNA's made DIRECTLY from DNA template
  - Travel from nucleus to ribosome
- rRNA: Ribosomal RNA
  - Helps form ribosomes in cytoplasm
- tRNA: Transfer RNA
  - Brings amino acids from cytoplasm to ribosome so proteins can be made

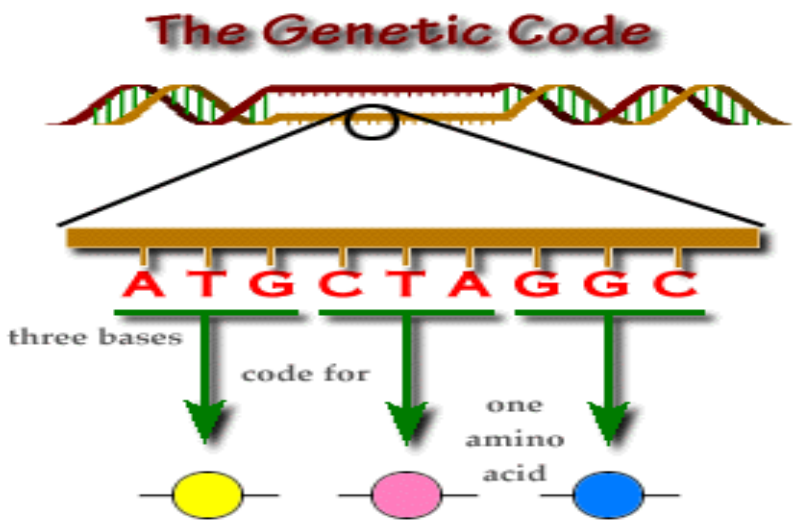
Step 1: Transcription

- INSIDE of the nucleus DNA is used to make mRNA
- DNA is unzipped then RNA polymerase makes an mRNA strand from the DNA template
- New mRNA strand then leaves the nucleus and travels into the cytoplasm
- DNA is ALWAYS left protected in the nucleus
- DNA: 5' AAA TTT GGG CCC ATC GCA 3'
- mRNA: 3' UUU AAA CCC GGG UAG CGU 5'
- DNA: CTA GTT CCC TAA AAG GAG
- mRNA: GAU CAA GGG AUU UUC CUC
- DNA: TAC CGA GGT TTA ACT

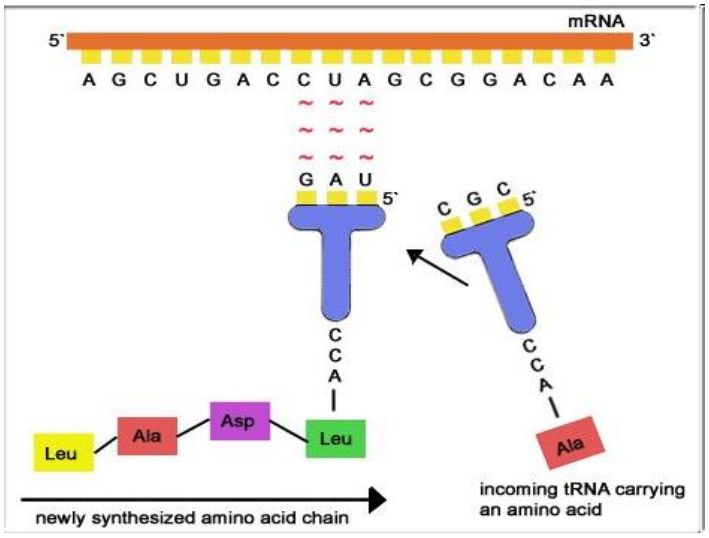
- mRNA: AUG UGA CCA AAU UGA

Step 2: Translation

- Each nucleotide sequence serves as a code for what amino acid will be added to the protein being made
- Nucleotides read in triplets, or codons



- mRNA is now connected to the ribosome
- tRNA has a corresponding anti-codon and brings over the corresponding amino acid



The end result…

- An amino acid sequence that makes a protein
- GENES code for proteins/enzymes
- We NEED proteins to function

- The shape of the protein determines its function

Module107: Types of proteomics

Text (7:00)

Scope of Proteomics:

- Expression proteomics
- Structural proteomics
- Functional proteomics

EXPRESSION PROTEOMICS:

- Expression proteomics is used to study the qualitative and quantitative expression of total proteins under two different conditions.
  ¢ Normal and diseased state.
  ¢ E.g. :tumor or normal cell.
  ¢ It studied that protein is over expressed or under expressed.
  ¢ 2-D electrophorasis.

STRUCTURAL PROTEOMICS:

- Structural proteomics helps to understand three dimensional shape and structural complexities of functional proteins.
- It determine either by amino acid sequence in protein or from a gene this process is known as **homology modeling**.
- It identify all the protein present in complex system or protein-protein interaction.
- Mass spectroscopy is used for structure determination.

FUNCTIONAL PROTEOMICS:

- Functional proteomics explains understanding the protein functions as well as unrevealing molecular mechanisms within the cell that depend on the identification of the interacting protein partners. So that detailed description of the cellular signaling pathways might greatly benefit from the elucidation of protein- protein interactions

Limitations of Genomics Challenge of Proteomics:

- **co-translational modifications**

differential mRNA splicing

- **post-translational modifications (PTMs)**

C-terminal GPI anchor

phosphorylation

sulfation

glycosylation

N-myristoylation

hydroxylation

N-methylation

carboxymethylation

signal peptidase site.....

Module108: STRUCTURAL PROTEOMICS

Text (8:00)

What is structural proteomics/genomics?

- High-throughput determination of the 3D structure of proteins
- Goal: to be able to determine or predict the structure of every protein.

Direct determination - X-ray crystallography and nuclear magnetic resonance (NMR).

Prediction

Comparative modeling -

Threading/Fold recognition

Ab initio

Why structural proteomics?

- To study proteins in their active conformation.

Study protein:drug interactions

Protein engineering

- Proteins that show little or no similarity at the primary sequence level can have strikingly similar structures.

Module109: Expressional Proteomics

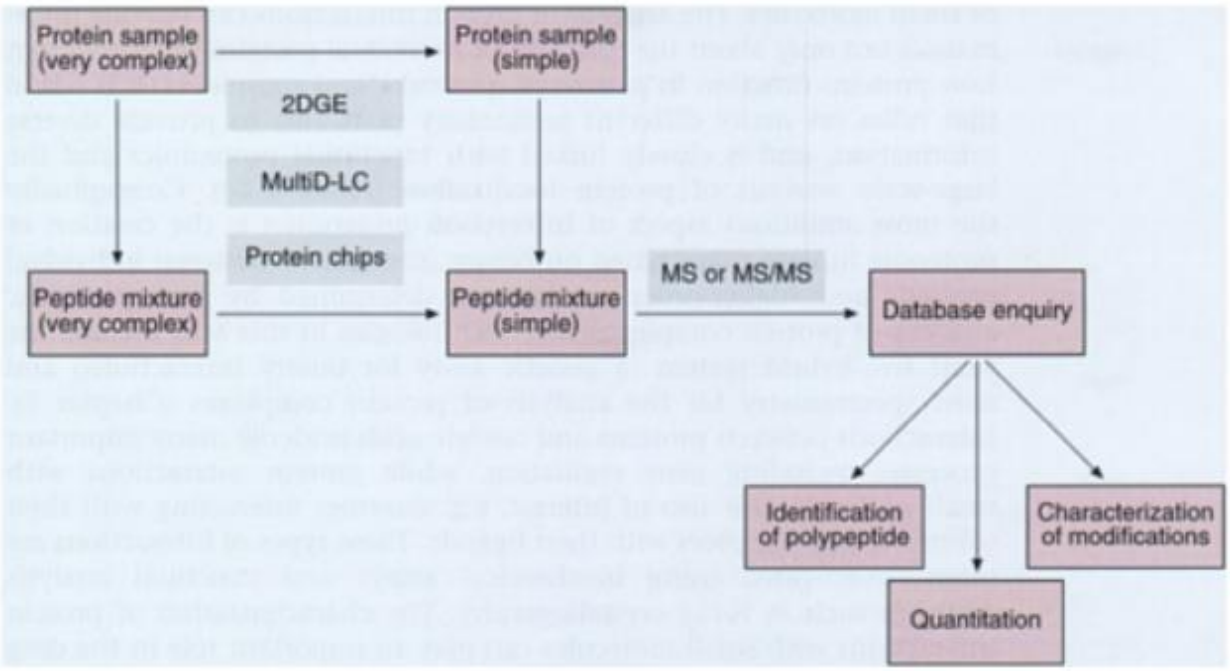
Text (8:00)

Expression Proteomics

Expression proteomics includes the analysis of protein expression at larger scale. It helps identify main proteins in a particular sample, and those proteins differentially expressed in related samples—such as diseased vs. healthy tissue.

Expression proteomics is devoted to the analysis of protein abundance and involves the separation of complex protein mixtures, the identification of individual components and their systematic quantitative analysis. Methods for the separation of protein mixtures based on two dimensional gel

electrophoresis (2DGE) were first developed in the 1970s and even at this time it was envisaged that databases could be created to catalog the proteins in different cells and look for differences representing alternative states, such as health and disease. Many of the statistical analysis methods which are usually associated with microarray analysis, such as clustering algorithms and multivariate statistics, were developed originally in the context of 2DGE protein analysis.

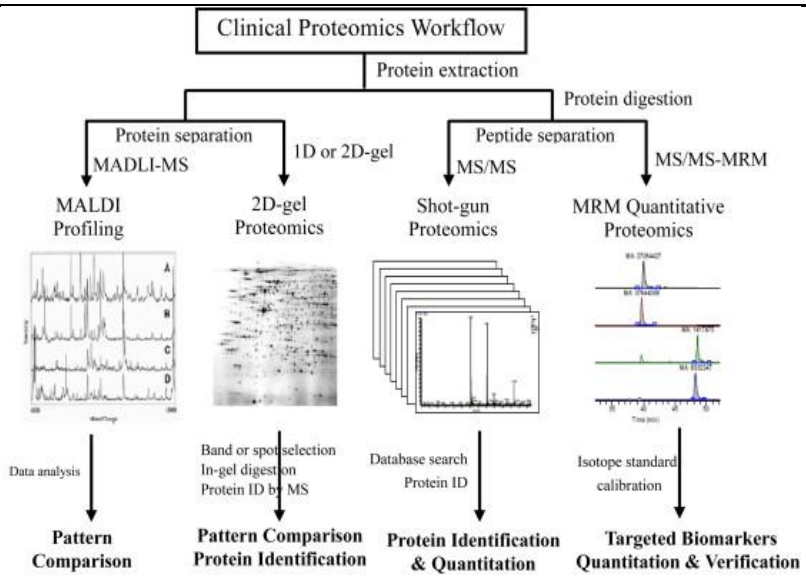


Expression proteomics is concerned with protein Identification and qualitative analysis. This figure shows the aims of expression proteomics and major technology platforms used.

Module110: Clinical Proteomics

Text (08:00)

Proteins are highly evolved nanomachines that carry out the work of the cell. Biologic information is transmitted both by and through proteins. Just as proteins are the functional elements of the organism, the field of clinical proteomics is an effector arm of the upcoming revolution in molecular medicine. Recent advances in proteomic technology have yielded biologic discoveries and pathologic insights at a rapid pace. Proteomics has opened a treasure chest of candidate biomarkers that were never before known to exist in the blood. Protein–protein interactions, posttranslational modifications, and entire proteomic circuits have become the new scaffolding for drug target discovery. Investigators have graduated from tissue-culture cell lines, and are now routinely applying proteomics to human tissue samples. The mission of Clinical Proteomics is to provide a scholarly forum for novel scientific research in the field of translational proteomics. The special emphasis of Clinical Proteomics is the application of proteomic technology to clinical research.



Areas of emphasis will include the following:

• Clinical sample collection and handling to preserve proteins and posttranslational modifications.

• New technology for protein-based clinical bioassays and clinical chemistry assays.

• Translational pathology related to proteomics.

• Bioinformatic tools and protein circuit building.

• Biomarker discovery and validation from clinical samples.

• Signal transduction pathway profiling in clinical tissue samples.

• Discovery of new drug targets from clinical samples.

• Use of proteomic technologies in the drug development pipeline (hit to lead screening and lead optimization and preclinical screening).

• Use of proteomic technologies to monitor prognosis, therapeutic end points, toxicity, and efficacy.

• Clinical trials using proteomic monitoring

• Clinical trials using proteomics to individualize therapy.

This inaugural issue of Clinical Proteomics is a showcase of scientific research spanning discovery, functional analysis and biomarker profiling.

Module111: Origins of Proteomics

Text (9:00)

- In 1975, the introduction of the 2D gel by O'Farrell who began mapping proteins from *E. coli.*
- Although many proteins could be separated and visualized, they could not be identified.

- Despite these limitations, shortly thereafter a large-scale analysis of all human proteins was proposed.
- The goal of this project, termed the human protein index, was to use two-dimensional protein electrophoresis (2-DE) and other methods to catalog all human proteins.
- However, lack of funding and technical limitations prevented this project from continuing.

Proteomics Origins

- The first major technology to emerge for the identification of proteins was the sequencing of proteins by Edman degradation.
- A major breakthrough was the development of microsequencing techniques for electroblotted proteins.
- Microsequencing technique was used for the identification of proteins from 2-D gels to create the first 2-D databases.
- Improvements in microsequencing technology resulted in increased sensitivity of Edman sequencing in the 1990s to high picomole amounts.
- One of the most important developments in protein identification has been the development of MS technology.
- The sensitivity of analysis and accuracy of results for protein identification by MS have increased by several orders of magnitude.
- It is now estimated that proteins in the femtomolar range can be identified in gels.

Because MS is more sensitive, can tolerate protein mixtures, and is amenable to high-throughput operations

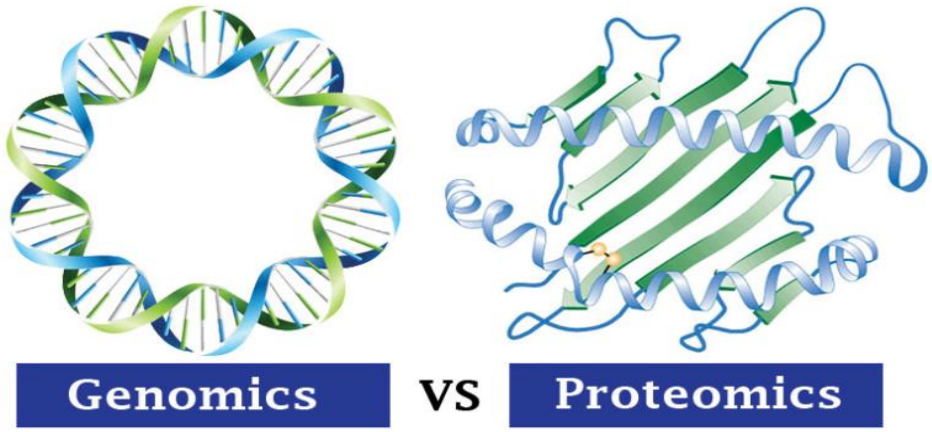
Module112: Genomics vs Proteomics

Text (7:00)

S.N.	Character	Genomics	Proteomics

1.	**Definition**	Genomics is the study of genomes which refers to the complete set of genes or genetic material present in a cell or organism.	Proteomics is the branch of molecular biology that studies the set of proteins expressed by the genome of an organism.
2.	**Study of**	Genomics is the study of the genes in an organism.	Proteomics is the study of the all the proteins in a cell.
3.	**Unit under Study**	The study of the function of genomes	The study of the function of proteomes
4.	**Nature of Study Material**	The genome is constant. Every cell of an organism has the same set of genes.	Proteome is dynamic and varies. The set of proteins produced in different tissues varies according to the gene expression.
5.	**Use of High throughput techniques**	High throughput techniques are used in the genomics to map, sequence, and analyze genomes.	In proteomics, characterization of the 3D structure and the function of proteins is carried out by the use of high throughput methods.
6.	**Techniques involved**	The techniques involved in genomics include gene sequencing strategies such as directed gene sequencing, whole genome shotgun sequencing, construction of expressed sequence tags (ESTs), identification of single nucleotide polymorphisms (SNPs), and the analysis and interpretation of sequenced data using different software and databases.	Techniques involved in proteomics include extraction and electrophoretic separation of proteins, digestion of proteins with the use of trypsin into small fragments, determination of the amino acid sequence by mass spectrometry, and identification of proteins using the information in the protein databases. Moreover, the 3D structure of the protein can be predicted using software-based methods. The expression of proteins can be studied by protein microarrays. Protein-network maps can be developed to determine

			protein-protein interactions.
7.	**Types**	The two types of genomics are structural genomics and functional genomics.	The three types of proteomics are structural, functional, and expression proteomics.
8.	**Important Areas**	Genome sequencing projects such as the Human Genome Project are the important areas of genomics.	Proteome database developments such as SWISS-2DPAGE and software development for computer-aided drug design are the important areas of proteomics.
9.	**Importance**	Genomic studies are important to understand the structure, function, location, regulation of the genes of an organism.	The study of the entire set of proteins produced by a cell type is done in order to understand its structure and function.
10.	**Significance**	Genes in the nucleus may not accurately portray conditions in the cell due to regulation at the RNA and protein level that cannot be viewed in Genomics studies.	Proteomics studies are more beneficial because proteins are the functional molecules in cells and represent actual conditions.



Module113: Life and death of protein

Text (8:00)

Life and death of protein:

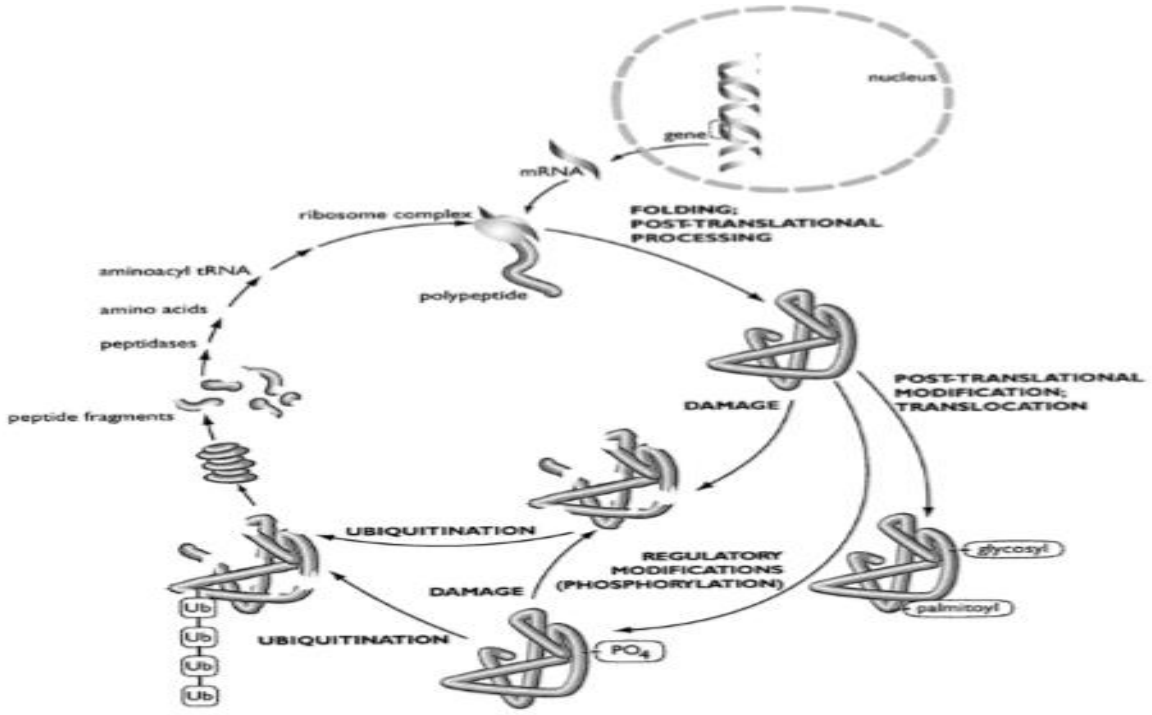
Proteins are synthesized by the translation of mRNAs into polypeptides on ribosomes.

In most cases, the initial polypeptide-translation product undergoes some type of modification before it assumes its functional role in a living system.

These changes are broadly termed "posttranslational modifications" and encompass a wide variety of reversible and irreversible chemical reactions.

Approximately 200 different types of posttranslational modifications have been reported. Some of these are summarized in Fig.

Life Cycle of the Cell



Modifications during Protein Cycle:

Modifications those occur early in the life of the protein

- Carboxylation of glutamate residues
- Removal of the N-terminal methionine
- Glycosylation
- Addition of Prosthetic groups
- Formation of multisubunit complexes
- Prenylation of cysteine residues assists anchoring of proteins in or on membranes.

These more or less "permanent" modifications and transport ultimately result in the delivery of functional proteins to specific locations in cells.

- The activities of many proteins are then controlled by posttranslational modifications.
- The most prominent and best-understood of these is phosphorylation of serine, threonine, or tyrosine residues.
- Phosphorylation may activate or inactivate enzymes, alter proteinprotein interactions and associations, change protein structures, and target proteins for degradation.
- Protein phosphorylation regulates protein function in diverse contexts and appears to be a key switch for rapid on-off control of signaling cascades, cell-cycle control, and other key cellular functions.

Degradation of Proteins:

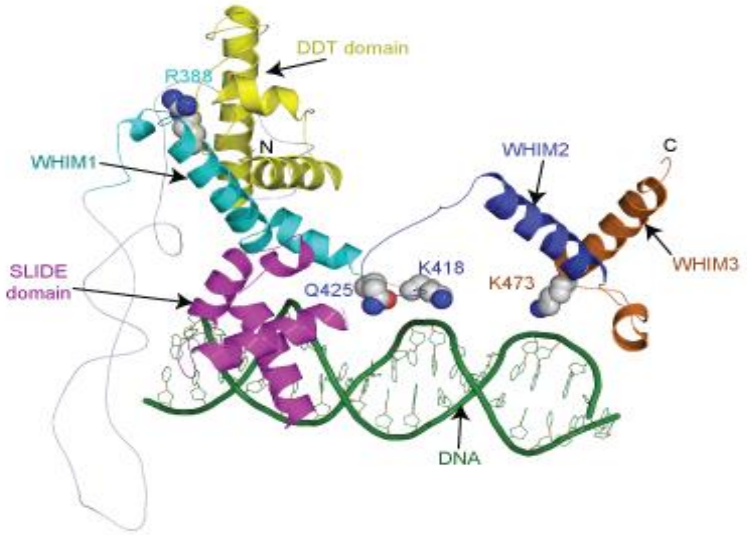
- Protein modifications appear to be critical to initiating processes that ultimately degrade proteins.
- Phosphorylation of some proteins is rapidly followed by conjugation with ubiquitin, which leads to degradation by the 26S proteasomal complex.
- There evidently are other stimuli for protein ubiquitination and turnover, including oxidative damage and other protein modifications.
- Proteins also undergo degradation by lysosomal enzymes.
- Any protein may be present in many forms at any one time in a cell.
- Collectively, the proteome of a cell comprises all of these many forms of all expressed proteins. This certainly makes the proteome bewilderingly complex.

Module114: Proteins as Modular Structure

Text (9:00)

Segments of amino acid sequences can be considered as functional building blocks or modules. The modular units in proteins that confer specific properties and functions are referred to as "motifs" or "domains". Motifs and domains are recognizable sequences that confer similar properties or functions when they occur in a variety of proteins. In some cases, amino acid sequences within

motifs and domains are highly conserved and do not vary from protein to protein. In other cases, some key amino acids occur in a reproducible relationship to each other in a sequence, even though various substitutions in other amino acids occur.

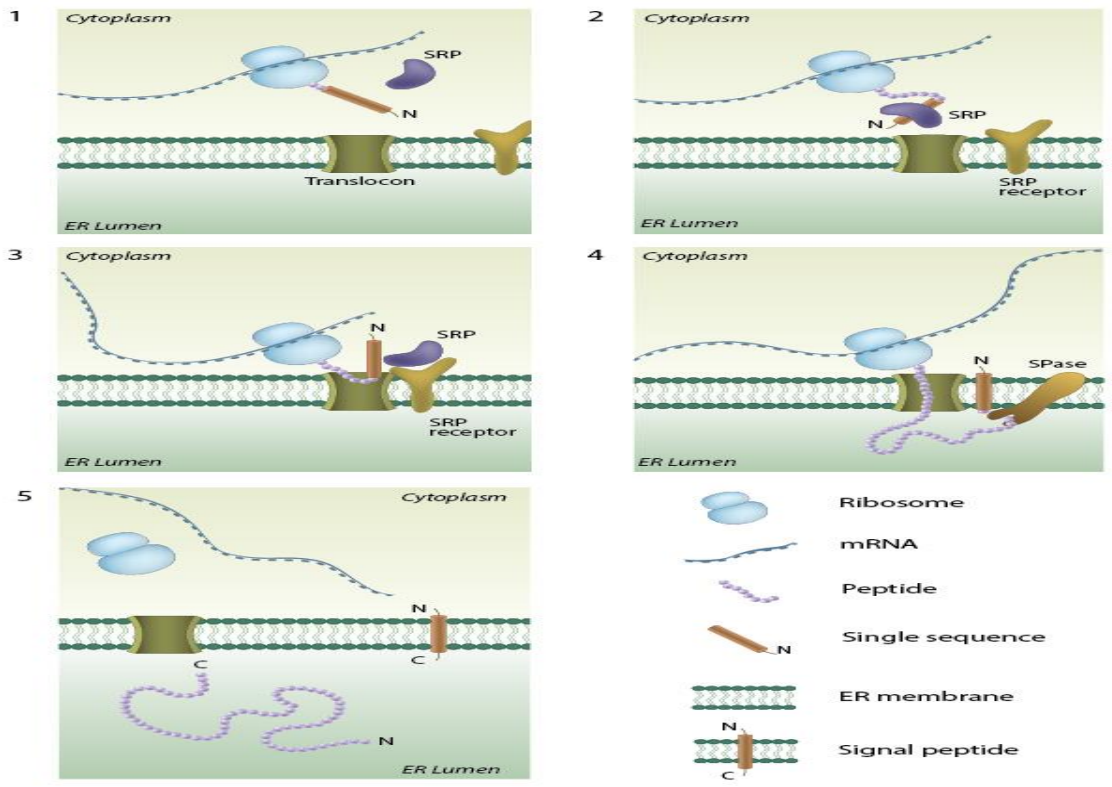


Longer amino acid sequences often form domains, which confer specific properties or functions on a protein. Some domain structures refer simply to sequences that confer a bulk physical property to a segment of the polypeptide, such as transmembrane domains, which simply form helices that span a lipid bilayer membrane. Other domain structures provide hydrogen bonding or other contacts for key enzyme substrates or prosthetic groups. In many cases, domains are made up of combinations of units of secondary structure, such as helix-loop-helix domains.

Module115: Localization of Proteins

Text (8:00)

In order for subcellular processes to be carried out within defined compartments or cellular regions, mechanisms must exist to ensure the required protein components are present at the sites and at an adequate concentration. The accumulation of a protein at a given site is known as protein localization.



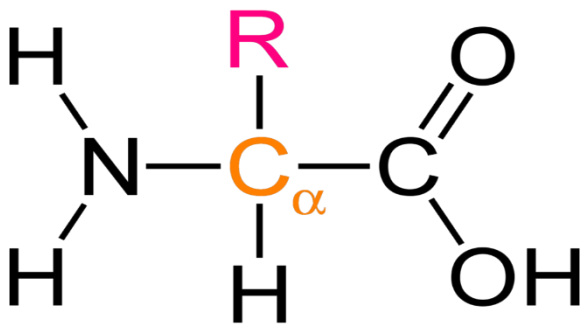
Protein localization can result from the recognition of passively diffusing soluble proteins or protein complexes; however, this may not guarantee a sufficient concentration of components to maintain a given process. This can impede its completion, particularly when carried out in regions with a limited cytoplasmic volume, such as the tip of a filopodia, or when components are rapidly turned over.

A more efficient way of maintaining the concentration of protein components is by their directed delivery via the cytoskeletal network. The cytoskeleton, which is comprised of actin filaments and microtubules, spans the entire cell and connects the plasma membrane to the nucleus and other organelles. These filaments perform many purposes, from providing structural support to the cell, to generating the forces required for cell translocation. They may also serve as 'tracks' on which motor proteins can translocate as they carry cargo from one location to another; analogous to a freight train transporting cargo along a network of railway tracks.

Module116: Chemical Composition of Proteins

Text (11:00)

Proteins are polymers of amino acids. They range in size from small to very large. All the proteins are made up of Twenty different types of amino acids. So these amino acids are called standard amino acids.



Amino acids are tiny molecules with a common structure. They have a central carbon atom attached to a hydrogen atom, an amino and a carboxyl group, and a fourth functional group (R), which is variable. Amino acids attach to one another through bonds called peptide bonds between the amino nitrogen and the carboxyl carbon.

When the bond is formed, a water molecule is released. Using these peptide bonds, amino acids can join together in chains of nearly any sequence, which are known as polypeptides. When a polypeptide is of an appropriate size, structure and sequence, it functionally becomes a protein.

Peptide bond is produced when carboxyl radical $(-\overset{\overset{O}{||}}{C}-OH)$ of one amino acid reacts with the amino ($-NH_2$) group of the other amino acid.

Module117: Introduction to homology modelling

Text (9:00)

1. **BACKGROUND**

   Proteins have 3-D structure. Each protein is unique in structure. And structure of protein determines its functionality. Proteins are classified as 1', 2', 3' and 4' structures. 1' structure is the simpler ones having linear structure of amino acids. Helices, beta sheets, loops and coils formed 2' structures. When 2' structures combined with the help of interaction then, 3' structure formed. 4' structure is the most complex structure.

   We can determine the structure of proteins with the help of X-ray crystallography and NMR spectroscopy. But these methods are expensive so, we used alternative approach in which we predict the structures of proteins.

2. **INTRODUCTION**

   Protein sequence determines its structures. So, if we have two proteins; we know the structure and sequence of first protein and we know only the sequence of other protein (unknown) both proteins are similar according to its structures so, we can determine the structure of unknown protein. We can identify the unknown protein structure by homologous protein sequence.

3. **CONCLUSION**

Homologous protein helps us to identify unknown proteins. Sequence alignment and identity help us to determine homology.

Module118: Homology, Paralogy and Orthology

Text (7:00)

1. **BACKGROUND**

   In homology modelling, prediction depends upon which type of protein we are using. Because 1' protein predicts 1' protein and similarly 3' protein will predict 3' protein.

2. **INTRODUCTION**

   According to evolutionary theory it predicts that related organisms have similarities of the structure, physiology etc. between two species which reflects that these two-species having common ancestor.

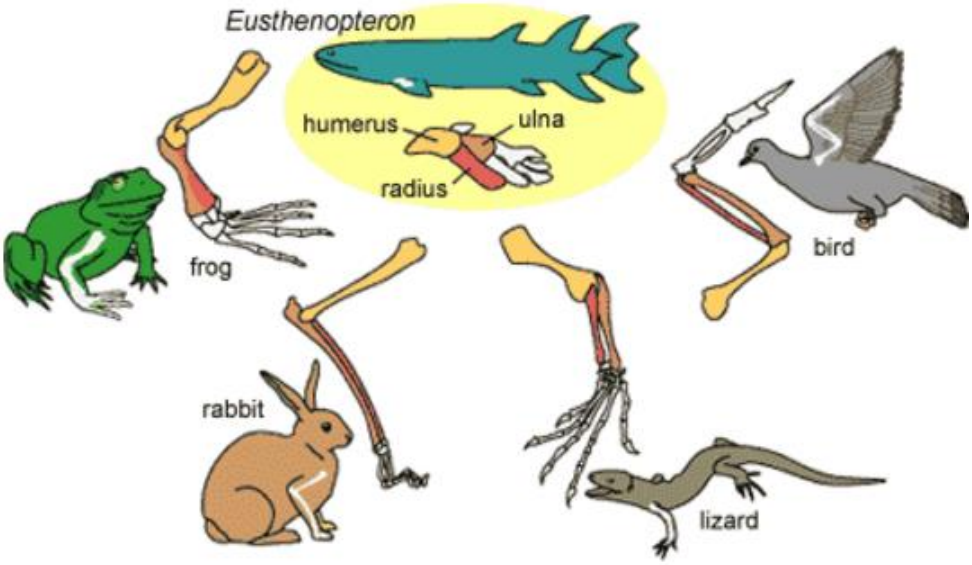
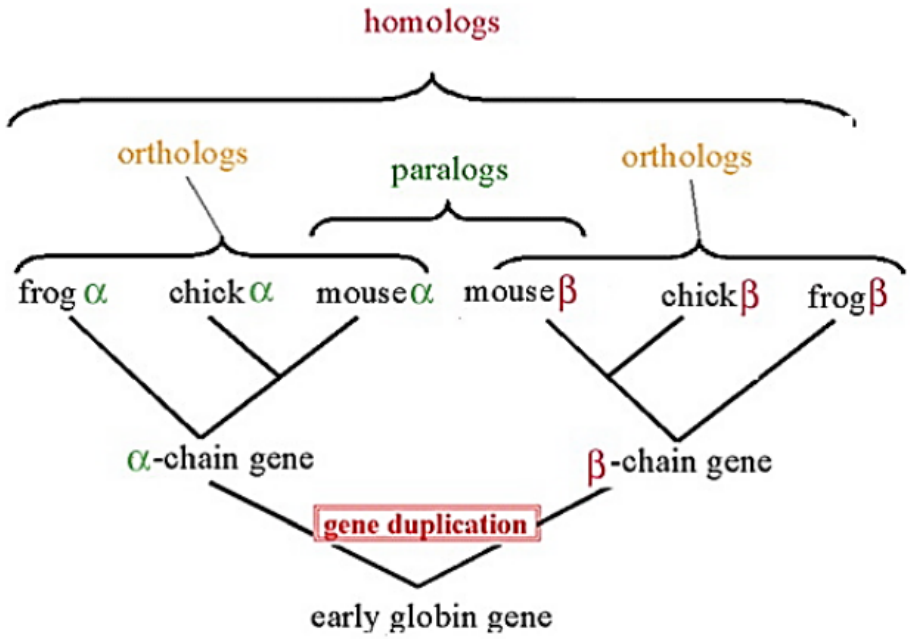


Figure 40.6.1: Homology behavior between different species

Table 40.6.1: Ortholog vs. paralog

Ortholog	Paralog
Gene from different species which evolved by common ancestral gene	Genes related by duplication within genome
Ortholog gene retain same functionality after evolution	Paralog gene evolve in a new function



http://bioweb.uwlax.edu/GenWeb/Molecular/Bioinformatics/Unit_4/Lab_4-2/lab_4-2.htm

Figure 40.6.2: Flow diagram of ortholog & paralog

Given below graphs gives the idea about homology and relationship between sequence identity and alignment length. If point lies above and before the curve then, we observe homology otherwise not.

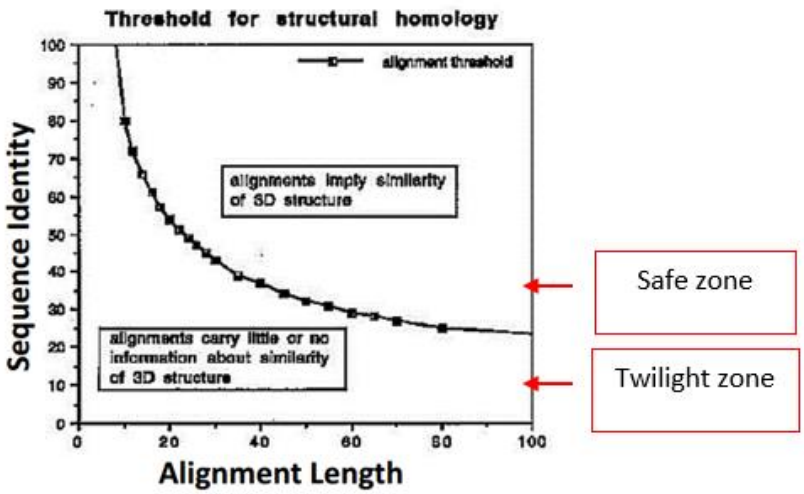


Figure 40.6.3: Graphical representation of sequence identity and alignment length

3. CONCLUSION

We can acquire accurate results if two species have good sequence alignment and they are also identical.

Module119: Workflow for Structural Modelling

Text (9:00)

1. BACKGROUND

Homology modelling is used for prediction of proteins.

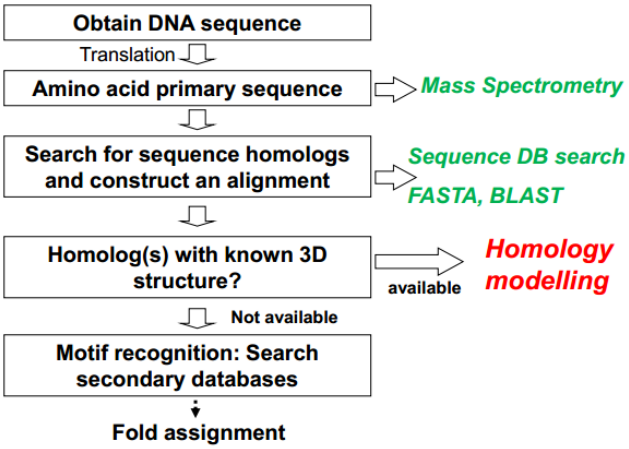


Figure 40.7.1: Initial steps before structural modelling

2. INTRODUCTION

There are three types of structural prediction of proteins. But here we will discuss only three types.

- Homology Modelling
- Thread Fold Recognition
- Ab Initio Modelling

3. CONCLUSION

We can only use homology modelling if we have high identity and high alignment score. If unknown protein lies in "twilight zone" then, we use other techniques.

Module120: Seven Steps to Homology Modelling-I

Text (9:00)

1. BACKGROUND

There are three types of structural prediction of proteins. But here we will discuss only three types.

- Homology Modelling
- Thread Fold Recognition
- Ab Initio Modelling

2. INTRODUCTION

There are seven steps for homology modelling for structural prediction of proteins. **Template (known):** all parameters are known of that protein. **Target (unknown):** some parameters are unknown of that protein and we are willing to determine it.

3. HOW IT WORKS?

Homology modeling having 7 steps for prediction of proteins.

- **Step 1:** Template recognition and initial alignment
- **Step 2:** Alignment correction

- **Step 3:** Backbone generation
- **Step 4:** Loop modeling
- **Step 5:** Side-chain modeling
- **Step 6:** Model optimization
- **Step 7:** Model validation

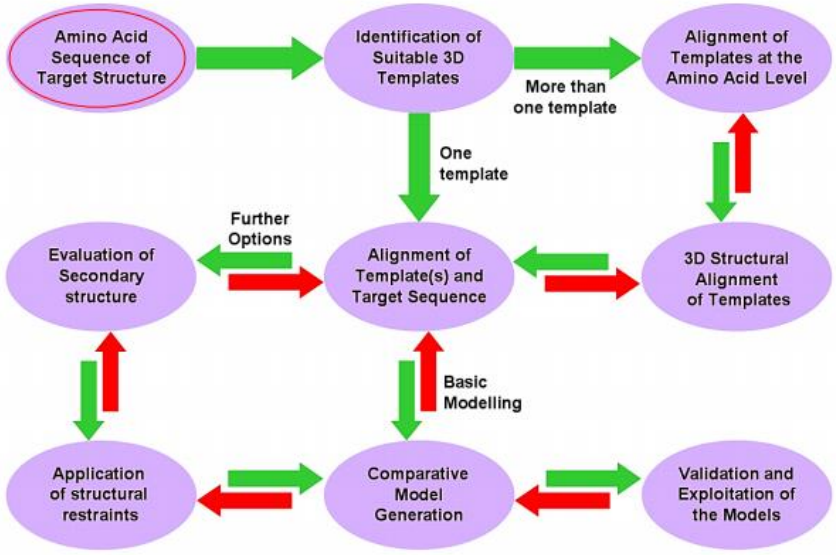


Figure 41.1.1: Workflow of homology modelling in 7 steps

4. CONCLUSION

Homology modelling having seven steps and its repetitive process.

Module121: Seven Steps to Homology Modelling-II

Text (14:00)

1. BACKGROUND

Homology modelling operates in 7 steps.

- **Step 1: Template recognition and initial alignment**
- **Step 2:** Alignment correction
- **Step 3:** Backbone generation
- **Step 4:** Loop modeling
- **Step 5:** Side-chain modeling
- **Step 6:** Model optimization
- **Step 7:** Model validation

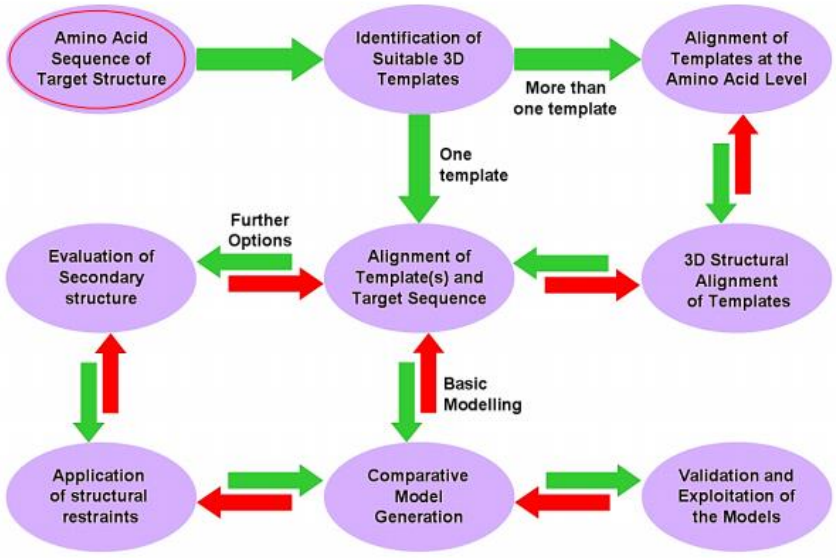


Figure 41.2.1: Workflow of homology modelling in 7 steps

2. HOW IT WORKS?

Given below flow diagram gives detail about first step which is Template reorganization and initial alignment.

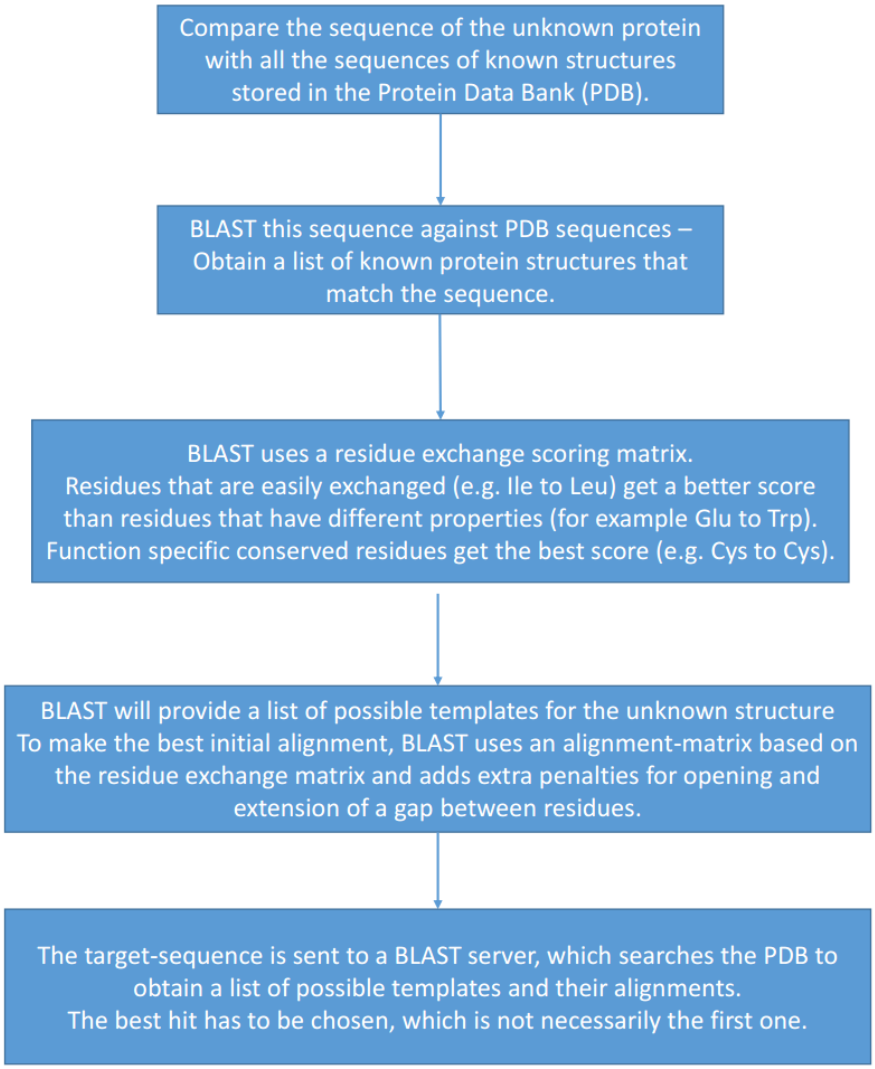


Figure 41.2.2: Working on selection of template & target

3. CONCLUSION

We successfully select the template and targets which is our first step in homologous modelling i.e. template reorganization and initial alignment.

Module122: Seven Steps to Homology Modelling-III

Text (08:00)

1. BACKGROUND

Homology modelling operates in 7 steps.

- **Step 1:** Template recognition and initial alignment
- **Step 2: Alignment correction**
- **Step 3:** Backbone generation
- **Step 4:** Loop modeling
- **Step 5:** Side-chain modeling
- **Step 6:** Model optimization
- **Step 7:** Model validation

2. HOW IT WORKS?

Given below flow diagram gives detail about second step which is alignment correction.

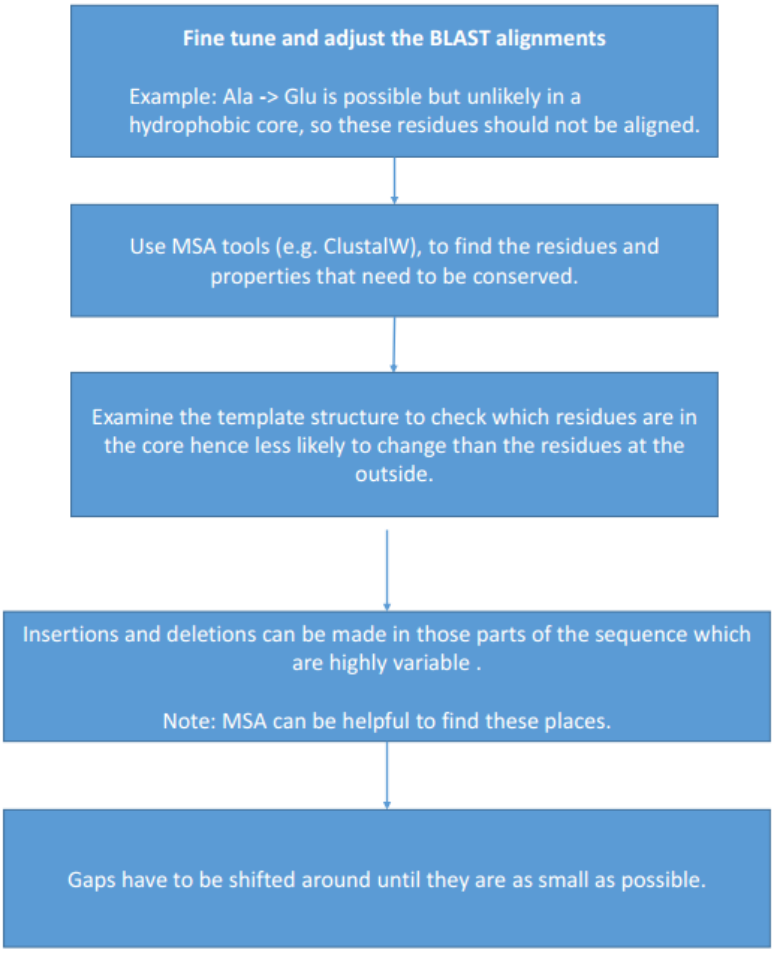


Figure 41.3.1: Working on alignment correction

During alignment correction residues are deleted so that we acquire best alignment (highest score). After deletion gap generated so, shifting of several residues decreases the gap.

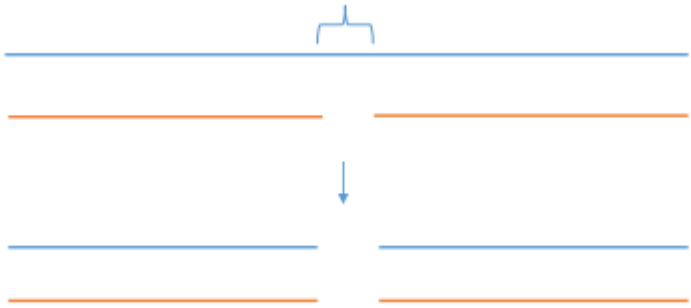


Figure 41.3.2: Deletion of residues in first sequence (blue line)

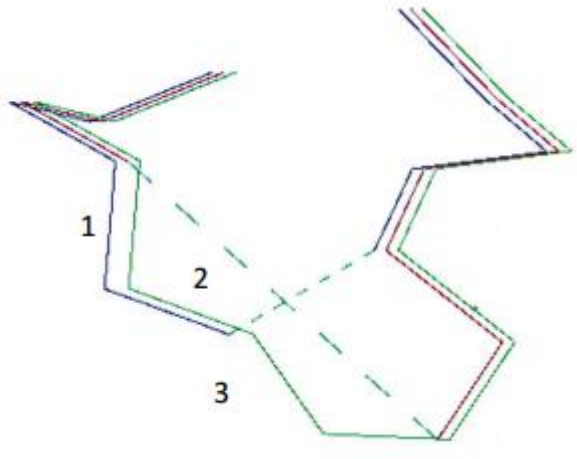


Figure 41.3.3: Shifting of residues in sequence (blue line)

3. CONCLUSION

We determined the mis matches and gaps which have adjusted. So, now we have alignment which is not only fine-tuned but also corrected.

Module123: Seven Steps to Homology Modelling-IV

Text (10:00)

1. BACKGROUND

Homology modelling operates in 7 steps.

- **Step 1:** Template recognition and initial alignment
- **Step 2:** Alignment correction
- **Step 3: Backbone generation**
- **Step 4: Loop modeling**
- **Step 5:** Side-chain modeling
- **Step 6:** Model optimization
- **Step 7:** Model validation

2. HOW IT WORKS?

Given below flow diagrams give detail about third and fourth steps which are backbone generation and loop modelling.

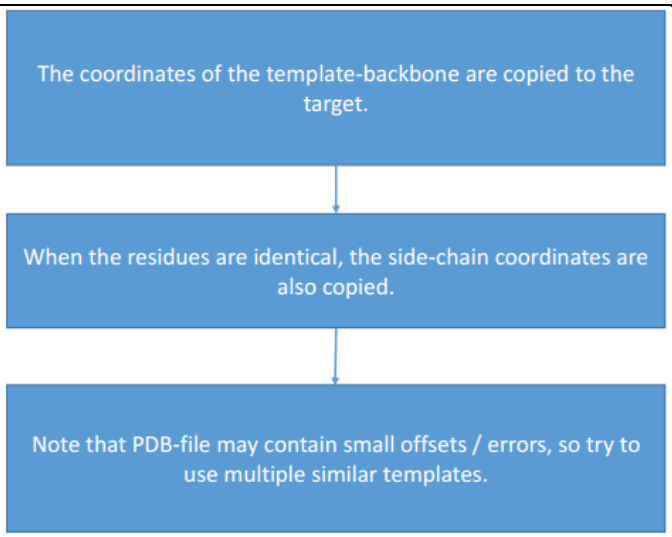


Figure 41.4.1: Working on backbone generation

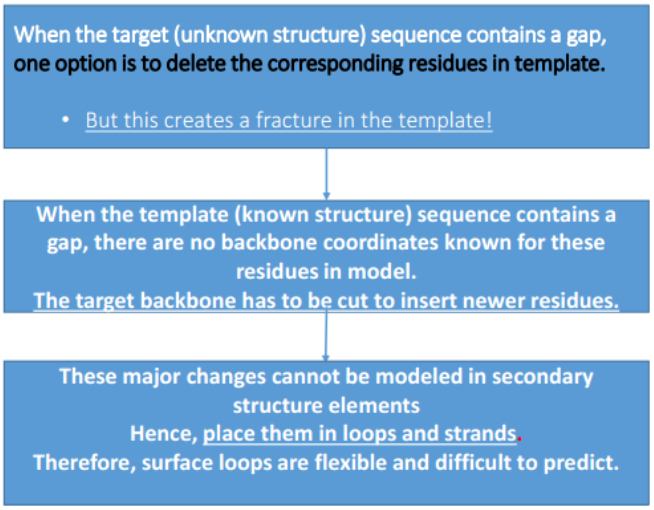


Figure 41.4.2: Working on loop modelling

Sometimes loops are attached as an anchor residue so, we searched it in PDB for loops having similar anchor-residues. And then, best loop copies in the model.

3. CONCLUSION

Now, backbone of protein is ready. In next step we will move towards side chains.

Module124: Seven Steps to Homology Modelling-V

Text (08:00)

1. BACKGROUND

Homology modelling operates in 7 steps.

- **Step 1:** Template recognition and initial alignment
- **Step 2:** Alignment correction

- **Step 3:** Backbone generation
- **Step 4:** Loop modeling
- **Step 5: Side-chain modeling**
- **Step 6: Model optimization**
- **Step 7:** Model validation

2. HOW IT WORKS?

Given below flow diagrams give detail about fifth and sixth steps which are side-chain modelling and model optimization.

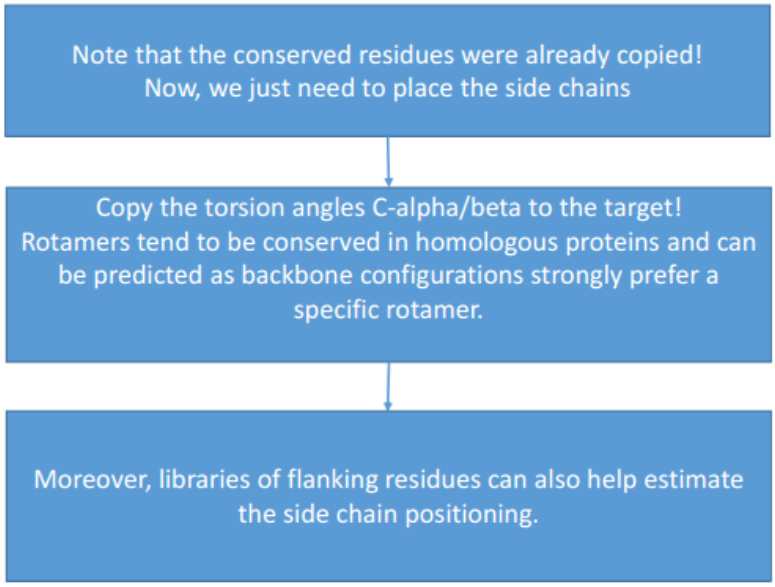


Figure 41.5.1: Working on side-chain modelling

Exceptional: Sometime amino acid has rotamers (simply said: rotational ability). So they can attach with backbone where they are fit properly. E.g. tyrosine has two rotamers.

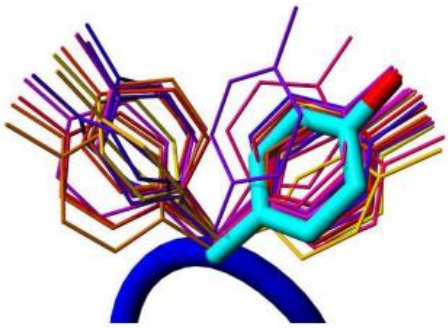


Figure 41.5.2: Rotamers in tyrosine (amino acid)

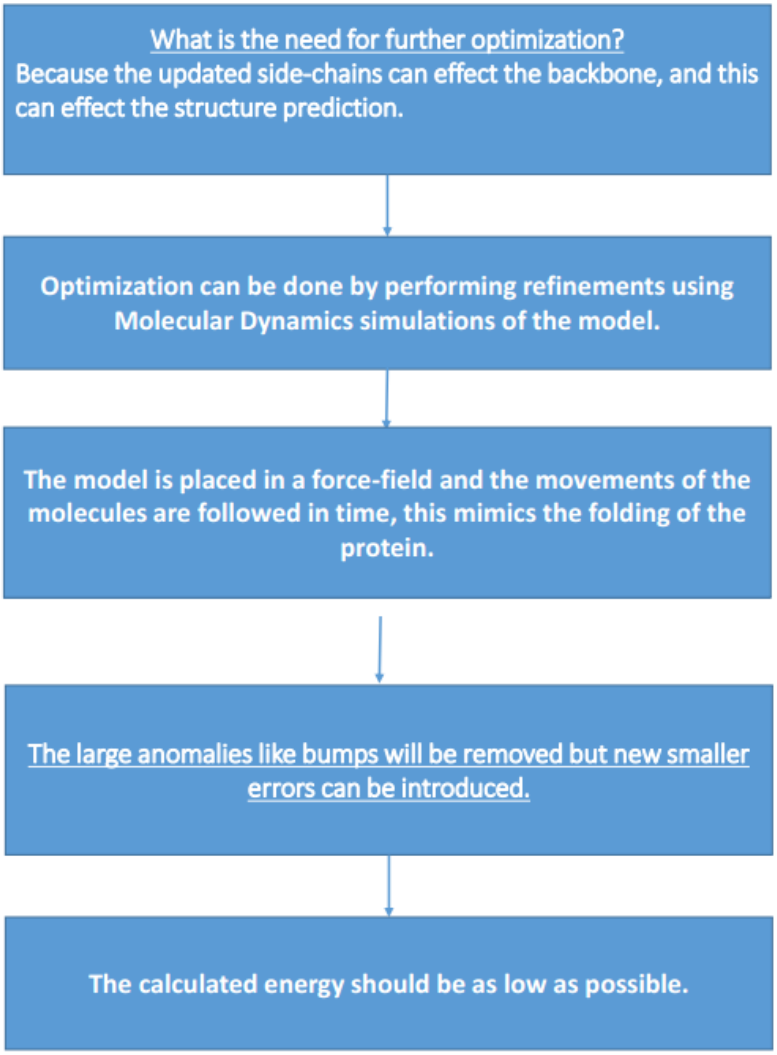


Figure 41.5.3: Working on model optimization

3. EXAMPLE

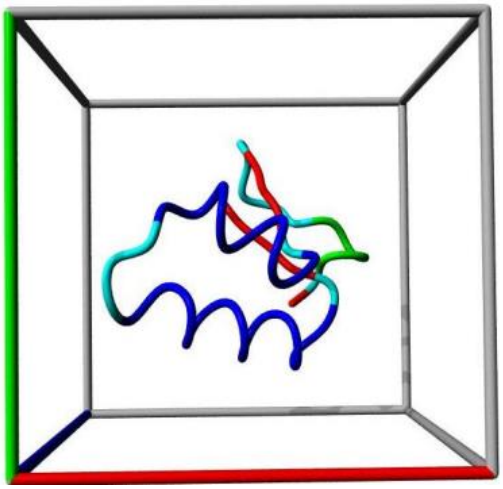


Figure 41.5.4: Modelling of Crambin (Ethiopian cabbage protein)

4. CONCLUSION

We reduced large errors, but smaller ones may still exist.

Module125: Seven Steps to Homology Modelling-VI

Text (9:00)

1. BACKGROUND

Homology modelling operates in 7 steps.

- **Step 1:** Template recognition and initial alignment
- **Step 2:** Alignment correction
- **Step 3:** Backbone generation
- **Step 4:** Loop modeling
- **Step 5:** Side-chain modeling
- **Step 6:** Model optimization
- **Step 7: Model validation**

2. HOW IT WORKS?

Given below flow diagram give detail about seventh (last) step which is model validation.

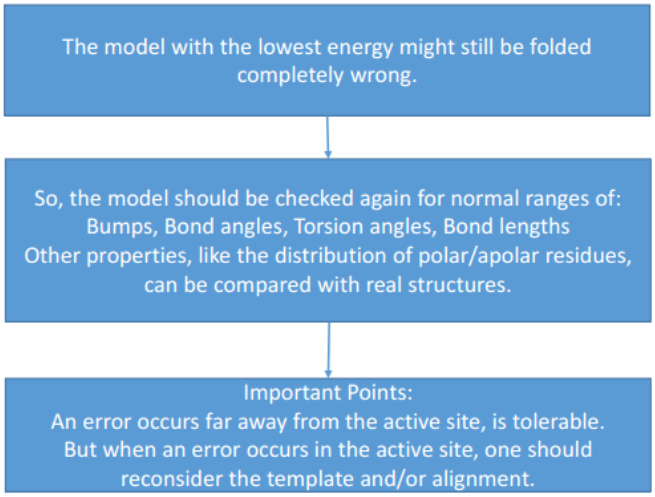


Figure 41.6.1: Working on model validation

3. LIMITATION OF HOMOLOGY MODELLING

In homology modelling three limitations are also exist which are:

- Large Bias towards structure of template
- Cannot study conformational (shape) changes
- Cannot elicit new catalytic/binding sites

4. CONCLUSION

Homology modelling having limitations so, we have some other strategies to avoid these limitations. These approaches are: Threading and Ab Initio Modelling.

Module126: Modeller for Homology Modelling

Text (7:00)

1. BACKGROUND

Homology modelling operates in 7 steps.

- **Step 1:** Template recognition and initial alignment
- **Step 2:** Alignment correction
- **Step 3:** Backbone generation
- **Step 4:** Loop modeling
- **Step 5:** Side-chain modeling
- **Step 6:** Model optimization
- **Step 7:** Model validation

2. INTRODUCTION

Modeller is a software for homology modelling. This link provides “MODELLER” for protein modling i.e. salilab.org/modeller. It takes input as python script file, sequence alignment and as a template (PDB).

```
from modeller import *
from modeller.automodel import *
log.verbose()
env = environ()
env.io.atom_files_directory = './'

a = automodel(
  env,
  alnfile = 'herg.ali',
  knowns = '1q5o',
  sequence = 'herg'
)

a.starting_model= 1
a.ending_model = 1
a.make()
```

Input Python Script (*.py)

```
>P1;1q5o
structureX: 1q5o : 443 : A : 644 : A ::::
DSSRRQYQEKYKQVEQYMSFHKLPADFRQKIHDYYEHRYQ-GKMFDEDSILGELNGPLRE
EIVNFNCRKLVASMPLFANADPNFVTAMLTKLKFEVFQPGDYIIREGTIGKKMYFIQHGV
VSVLTKGNKEMKLSDGSYFGEICLL--TRGRRTASVRADTYCRLYSLSVDNFNEVLEEYP
MMRRAFETVAIDRLDRIGKKNSIL.*

>P1;herg
sequence: herg : 1 :::::::
YSGTARYHTQMLRVREFIRFHQIPNPLRQRLEEYFQHAWSYTNGIDMNAVLKGFPECLQA
DICLHLNRSLLQHCKPFRGATKGCLRALAMKFKTTHAPPGDTLVHAGDLLTALYFISRGS
IEILRGDVVVAILGKNDIFGEPLNLYARPGKSNGDVRALTYCDLHKIHRDDLLEVLDMYP
EFSDHFWSSLEITFNLRDTN-MIP.*
```

Sequence Alignment (*.ali)

```
ATOM  1  N   ASP A 443   -15.943  41.425  44.702  1.00 44.68
ATOM  2  CA  ASP A 443   -15.424  42.618  45.447  1.00 43.15
ATOM  3  C   ASP A 443   -14.310  43.306  44.686  1.00 41.81
ATOM  4  O   ASP A 443   -14.298  44.528  44.539  1.00 42.61
                            etc...
```

Template Structure (*.pdb)

Figure 42.1.1: Different input schemes for MODELLER

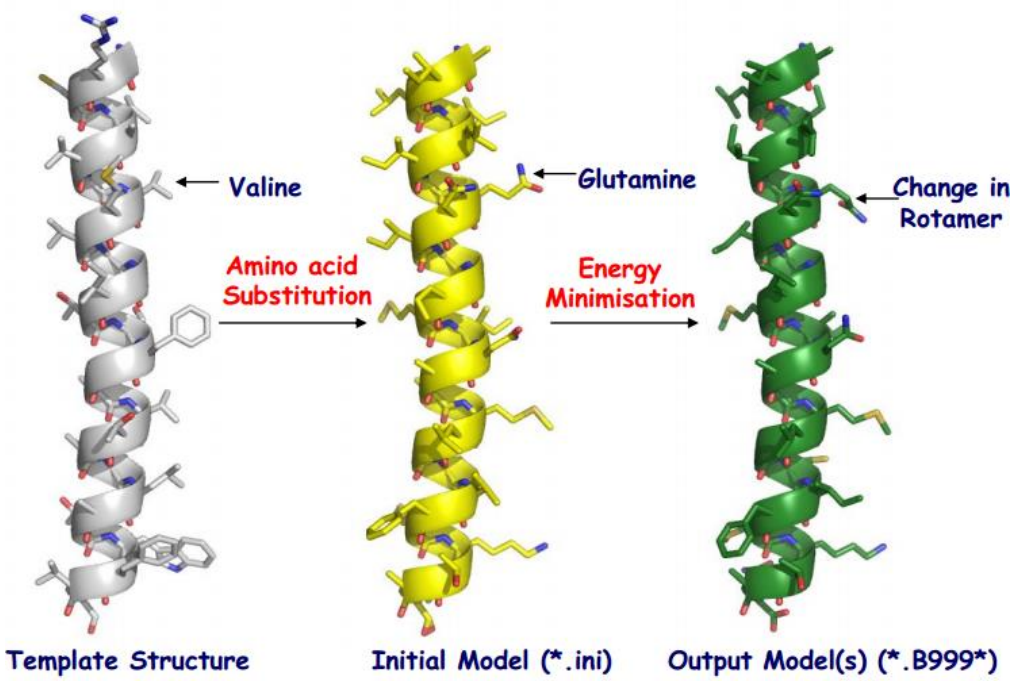


Figure 42.1.2: Output view in MODELLER

3. FILE EXTENSIONS

Given below are the file extensions which relates to MODELLER.

- **.log:** log output from the run
- **.B*:** model generated in the PDB format
- **.D*:** progress of optimization

- **.V*:** violation profile
- **.ini:** initial model that is generated
- **.rsr:** restraints in user format
- **.sch:** schedule file for the optimization process

4. AUTOMATED MODELLING SERVERS

Given below are the automated modelling servers with URL (Universal Resource Locator).

- **Swiss Model:** http://swissmodel.expasy.org//SWISSMODEL.html
- **Robetta:** http://robetta.bakerlab.org/
- **3D Jigsaw:** http://www.bmm.icnet.uk/servers/3djigsaw/
- **Phyre:** http://www.sbg.bio.ic.ac.uk/phyre/

5. CONCLUSION

Homology modelling helps us to predict protein structures by using prior structural information. There are several tools to perform this job either by programming or automated way.

Module127: Fold Recognition/Threading I

Text (8:00)

1. BACKGROUND

Homology modelling is used for prediction of proteins.

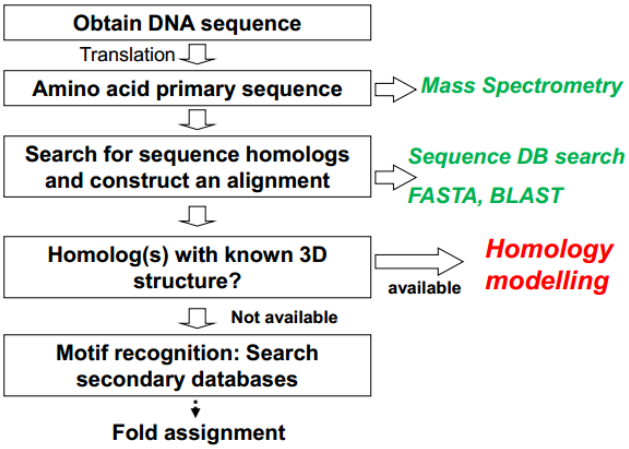


Figure 42.1.1: Initial steps before structural modelling

Homology modelling operates in 7 steps.

- **Step 1:** Template recognition and initial alignment
- **Step 2:** Alignment correction
- **Step 3:** Backbone generation
- **Step 4:** Loop modeling
- **Step 5:** Side-chain modeling
- **Step 6:** Model optimization

- **Step 7:** Model validation

Given below graphs gives the idea about homology and relationship between sequence identity and alignment length. If point lies above and before the curve then, we observe homology otherwise not.

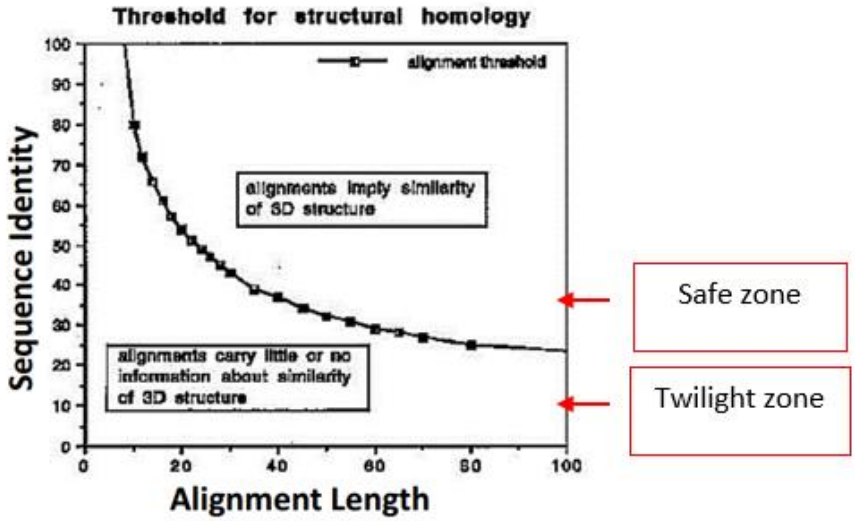


Figure 42.2.2: Graphical representation of sequence identity and alignment length

2. INTRODUCTION

In protein 2' structure protein elements are arranged in space (3-D) relative to the positions of each other. The common folds are 4-helix bundle and TIM barrel.

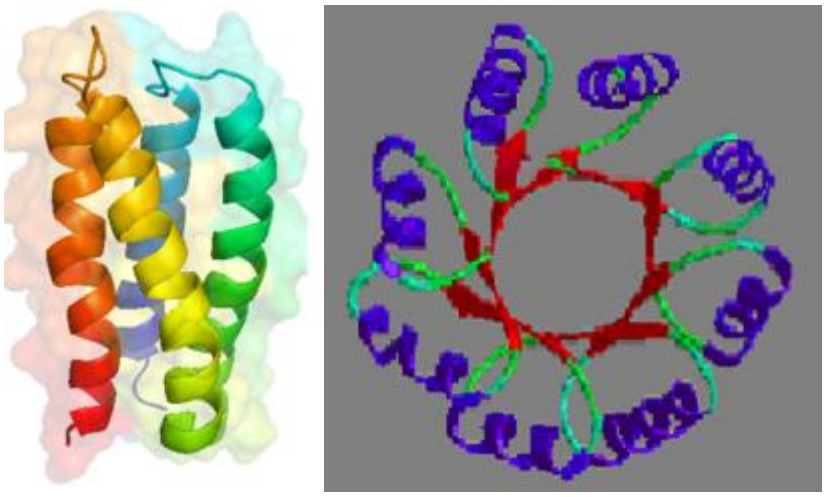


Figure 42.2.3: (a) Structure of 4 helix bundle & (b) Top cross-sectional view of TIM barrel

There are approximately 5,000 natural stable folds exists. And we choose best fold which fits according to our sequence this technique is known as threading or fold recognition.

3. **CONCLUSION**

Fold recognition or threading is a technique which is used for structural prediction of proteins. It is very useful where homology modelling fails. And it predicts quality structures.

Module128: Fold Recognition/Threading II

Text (9:00)

1. **BACKGROUND**

Fold recognition is also called Threading which is used for predicting protein structures. It helps us when homology modelling is not able to predict quality structures.

2. **HOW IT WORKS?**

- **Step 1:** In this technique, we mount the residue of unknown protein onto another known protein structure

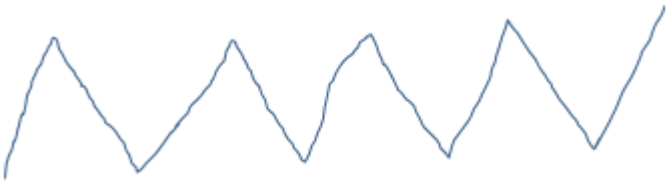


Figure 42.3.1: Residue of unknown protein

- **Step 2:** We drag unknown protein on template (known protein)
- **Step 3:** Then, we compute the fitness of sequence on it

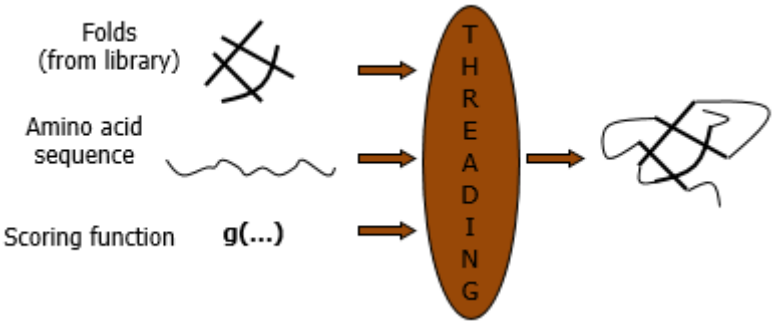


Figure 42.3.2: Inputs & outputs of threading

3. **CONCLUSION**

In threading, amino acid sequence passes through each fold in the database. And threading compute best matching using a scoring function.

Module129: Fold Recognition/Threading III

Text (8:00)

1. BACKGROUND

In threading, amino acid sequence passes through each fold in the database. And threading compute best matching using a scoring function.

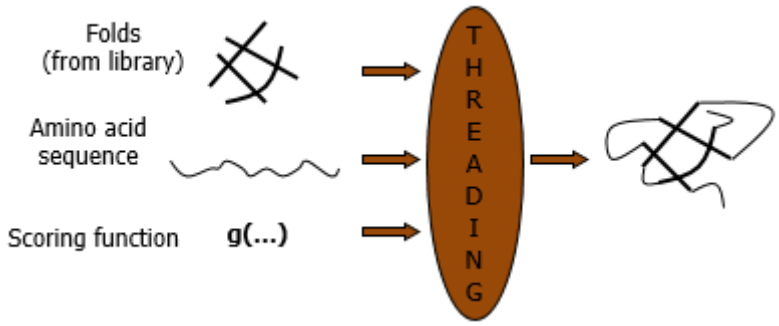


Figure 42.4.1: Inputs & outputs of threading

2. HOW IT WORKS?

Following flow chart shows that how threading works.

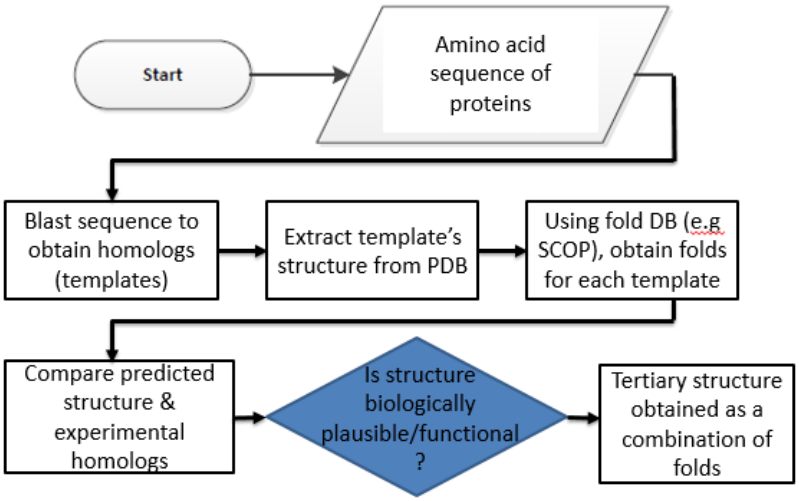


Figure 42.4.2: Flow chart of threading

3. EXAMPLE

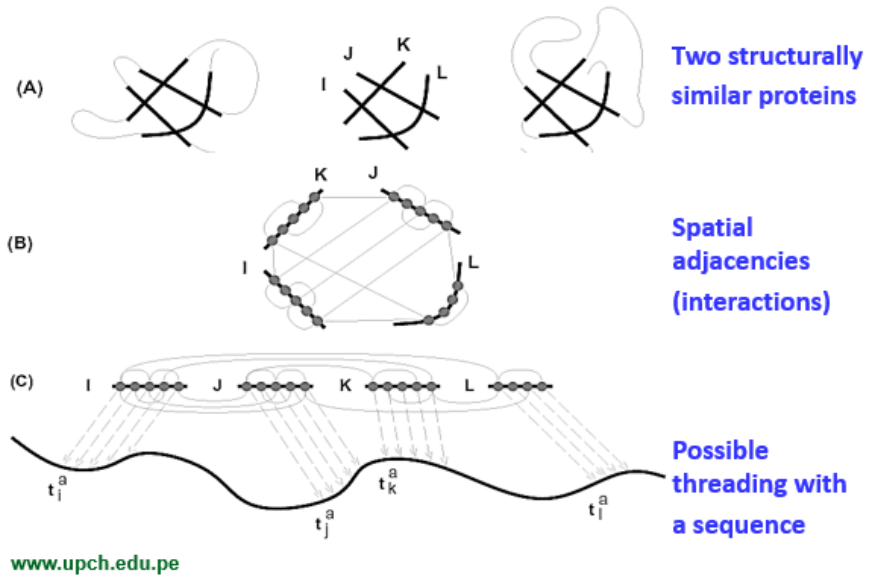


Figure 42.4.3: Example

4. CONCLUSION

In threading, different types of secondary structures make combination to form the best prediction. In this method scoring typically involves using a Z-Score (statistical term) function based on energy of a molecule.

Module130: Online Tools for Threading- iTasser

Text (11:00)

1. BACKGROUND

In threading, amino acid sequence passes through each fold in the database. And threading compute best matching using a scoring function.

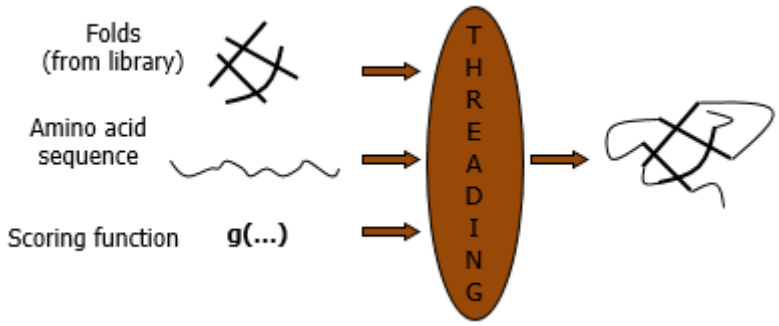


Figure 42.5.1: Inputs & outputs of threading

2. INTRODUCTION

ITASSER stands for "**I**terative **T**hreading **ASSE**mbly **R**efinement (I-TASSER) server. It's a software for automated protein structure and for functional prediction which is based on the **sequence-to-structure-to-function**.

3. **HOW IT WORKS?**
   - **Step 1:** Starts from amino acid sequence
   - **Step 2:** ITASSER first generates 3D atomic models from multiple threading alignments and iterative structural assembly simulations
   - **Step 3:** The function of the protein is then inferred by structurally matching the 3D models with other known proteins
   - **Step 4:** Outputs full-length secondary & tertiary structures and functional annotations on ligand-binding sites
   - **Step 5:** An estimate of accuracy of the predictions is provided based on the confidence score of the modeling

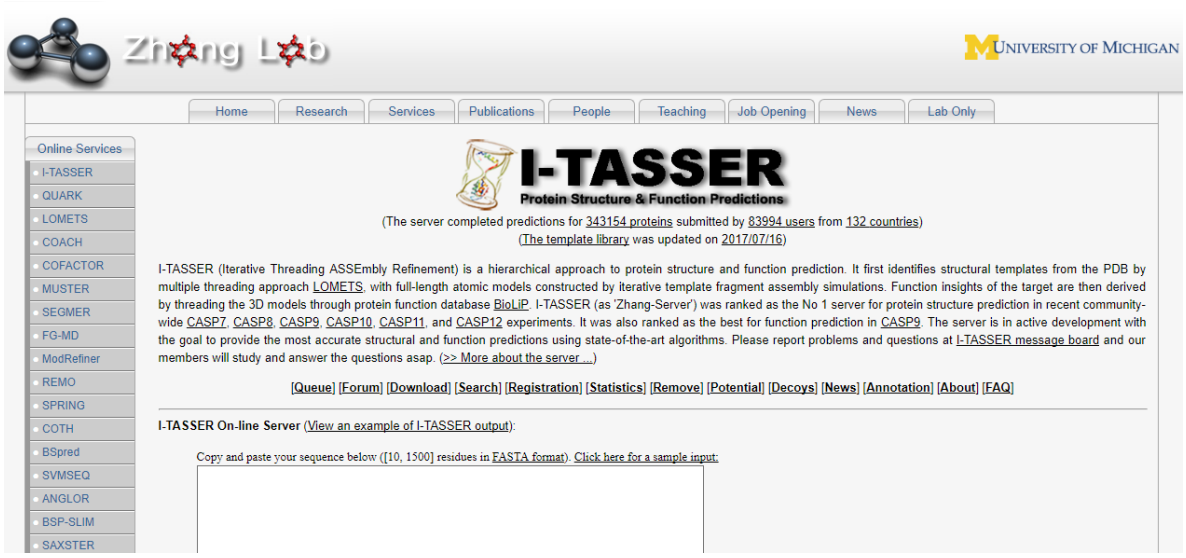


Figure 42.5.2: Homepage of ITASSER

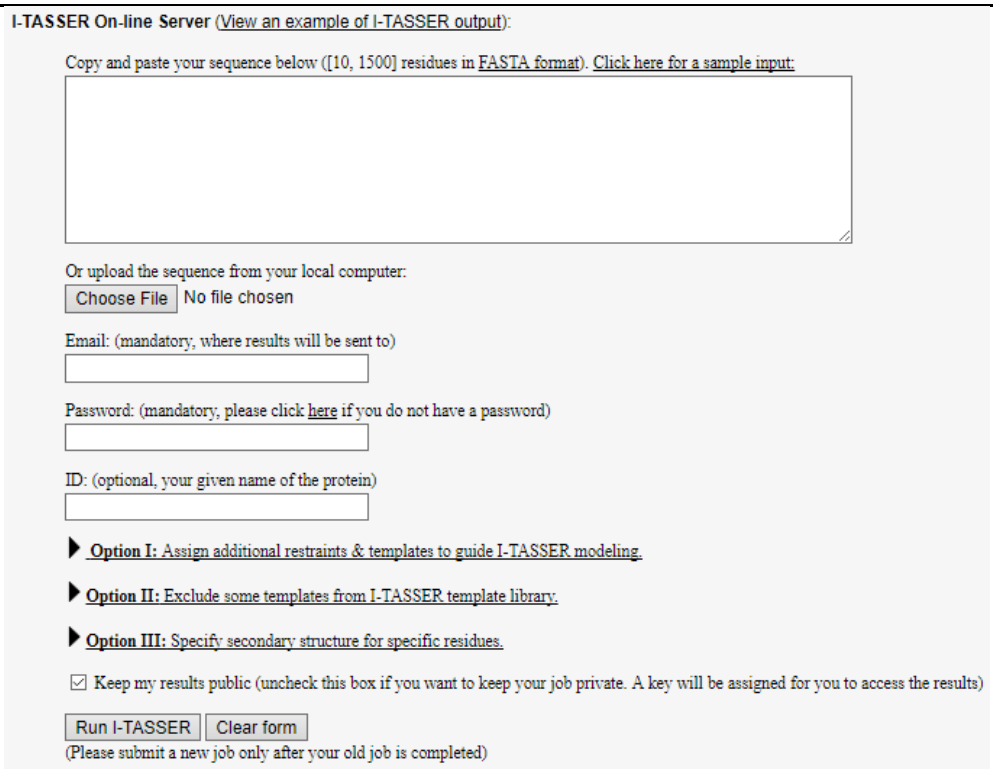


Figure 42.5.3: Input page of ITASSER

4. CONCLUSION

ITASSER helps us by predicting functions of structures.

Module131: Advantages and Disadvantages of Threading

Text (9:00)

1. BACKGROUND

Fold recognition or threading is a technique to predict protein structures. It is very useful technique which helps us when homology modelling fails to give the quality results.

EXAMPLE

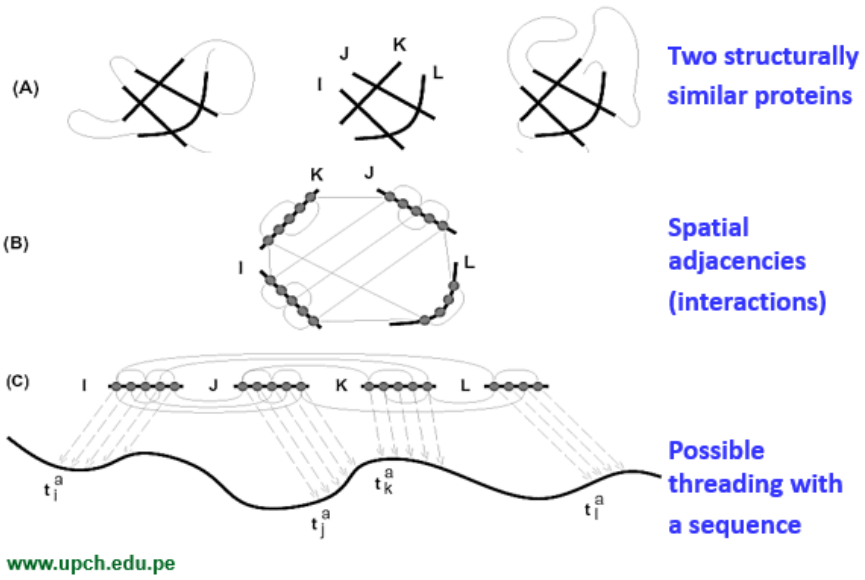


Figure 42.6.1: Example

2. ADVANTAGES

Threading helps us to predict the 2' structural protein to 3' structural protein. This method is also work "Twilight Zone" where homology modelling fails.

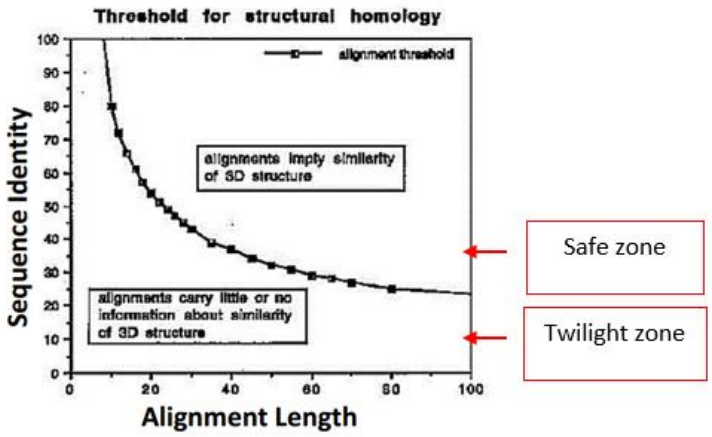


Figure 42.6.2: Graphical representation of sequence identity and alignment length

3. DISADVANTAGES

- Novel proteins cannot be predicted using threading
- Fewer than 30% of the predicted first hits are true remote homologues
- Validation of each result is necessary

Module132: 3D-1D Bowie Algorithm

Text (7:00)

1. BACKGROUND

Homology employed high alignment scores whereas threading work by creating combinations of 1' sequences and its corresponding 2' structures.

2. INTRODUCTION

Bowie Algorithm was purposed by Bowie in 1991. It converts all 3-D structures into 1-D string profiles. Based on 2' structure total 18 structural environments discussed in Bowie Algorithm e.g. solvent accessibility etc. Profiles of scores of each 20 amino acids computed. Then, it aligns with the target sequence to these profiles.

Identify amino acids based on: protein core, side chain positioning, solubility etc. (6 in all)

Part of secondary structure including α-helix, β-sheet etc (3 in all)

Total of 3 x 6 = 18 distinct states

$P_{a:j}$= Probability of finding amino acid (a) in environment (j)

P_a= Probability of finding (a) anywhere

Maximize sum of scores for the fold:

$$s_{aj} = \log\left(\frac{P_{a:j}}{P_a}\right)$$

Figure 43.1.1: Scoring formula, Bowie Algorithm

3. CONCLUSION

3D-1D method convert all information into "profiles". So, then we compute score for each amino acid for each profile.

Module133: Introduction to Ab-Initio Modelling

Text (9:00)

1. INTRODUCTION

Ab initio method is based on Anfinsen's dogma (or thermodynamic hypothesis). This method helps us to determine the structure with minimum free energy.

2. NEED FOR AB INITIO MODELLING

Ab Initio method is applicable for all sequences. But biologically it's not very accurate. Its accuracy and applicability are limited based on our requirements.

3. LIMITATION

Ab Initio method is computationally expensive and it is only suitable for those proteins who have less than 100 residues.

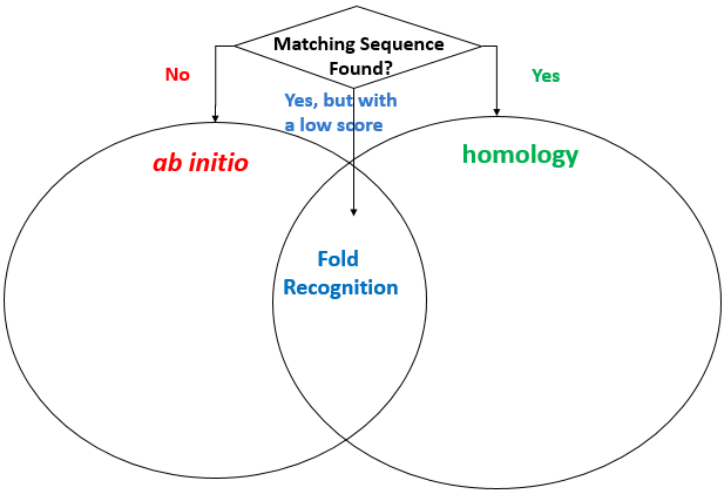


Figure 43.2.1: Comparative analysis of three different modelling techniques

4. CONCLUSION

Ab Initio method depends upon the energies of folded proteins. The protein structures with the lowest energy are inclined as plausible predictions.

Module134: Rationale of Ab-Initio Modelling

Text (9:00)

1. BACKGROUND

Ab Initio method depends upon the energies of folded proteins. The protein structures with the lowest energy are inclined as plausible predictions.

2. INTRODUCTION

Ab initio method is based on Anfinsen's dogma (or thermodynamic hypothesis). This method helps us to determine the structure with minimum free energy.

3. RATIONALE

Sometimes it happens that a protein with slightly homology does not available which renders the homology modelling and threading as futile. It is useful for the discovery of novel proteins. This method is independent to that method which uses matching with available structures. Other schemes included homology and fold recognition does not use physical and chemical properties for prediction of proteins.

4. CONCLUSION

Ab Initio method predicts the structure of proteins based on physical models. Amount of energy which is released during folding also computed for prediction of structure.

Module135: Strategies for Ab-Initio Modelling

Text (14:00)

1. BACKGROUND

Ab Initio method predicts the structure of proteins on the basis of physical models. Amount of energy which is released during folding also computed for prediction of structure.

2. HOW IT WORKS?

Whole ab Initio method works in two levels which are:

- **Level 1:** Energy optimization in Ab Initio modelling
  - **Step 1:** Start with a rough initial model
  - **Step 2:** Define an energy function mapping structures to energy values
  - **Step 3:** Solve the computational problem of finding the global minimum

- **Level 2:** Simulation of the folding process
  - **Step 1:** Build an accurate initial model (including energy and forces)
  - **Step 2:** Accurately simulate the dynamics of the protein folding process
  - **Step 3:** The native structure will steadily emerge

3. CONCLUSION

Ab Initio compute energy then, it formed structure of protein. That structure has minimum energy so, it has maximum stability.

Module136: Energy States of Folded Proteins

Text (08:00)

1. BACKGROUND

Ab Initio method predicts the structure of proteins based on physical properties.

- **Level 1:** Energy optimization in Ab Initio modelling
  - **Step 1:** Start with a rough initial model
  - **Step 2:** Define an energy function mapping structures to energy values
  - **Step 3:** Solve the computational problem of finding the global minimum

2. INTRODUCTION

Total energy calculated of the whole molecule by force field energy.

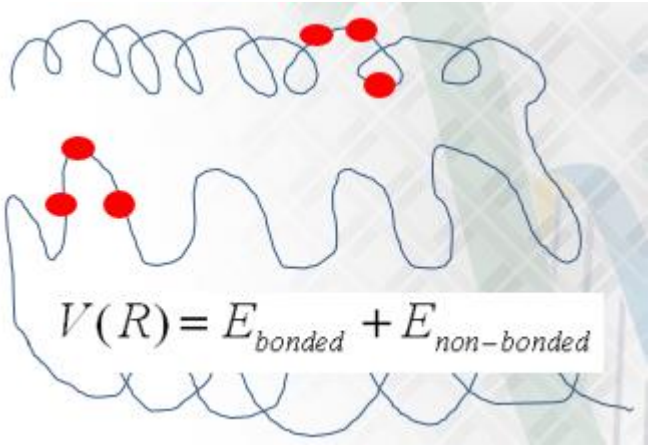


Figure 43.5.1: Energies of bonded atoms vs. non-bonded atoms

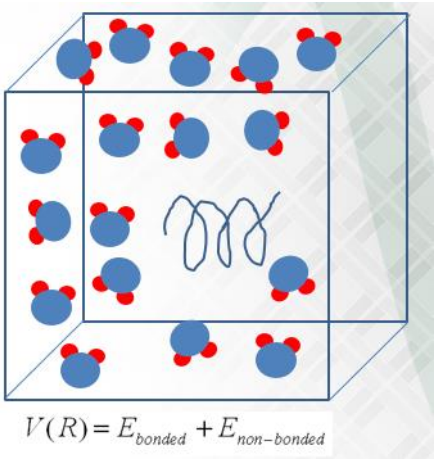


Figure 43.5.2: Force field energy calculation (starting)

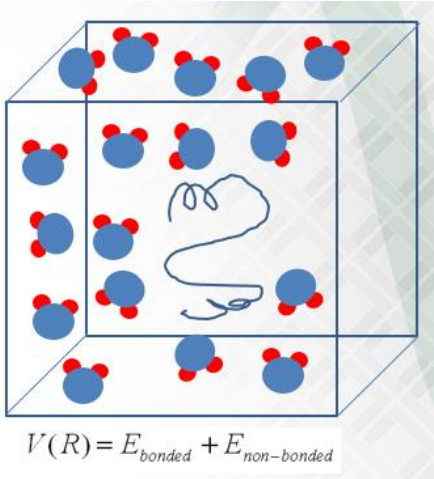


Figure 43.5.2: Force field energy calculation (during)

3. CONCLUSION

Lowest energy protein structure selected.

Module137: Local versus Global Minima

Text (10:00)

1. BACKGROUND

Lowest energy protein structure selected.

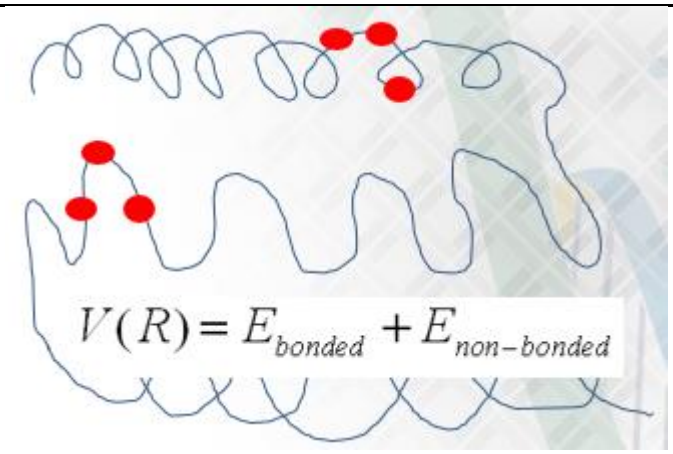


Figure 43.6.1: Energies of bonded atoms vs. non-bonded atoms

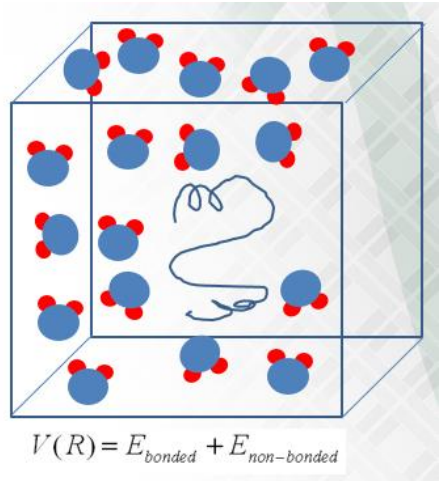


Figure 43.6.2: Force field energy calculation (during process)

2. BEST CASE ENERGY FUNCTION

First, Ab Initio method computes the global energy. Global energy helps us to find global minimum. Global minimum energy reflects the stability. So, that after computing global minimum we will be able to determine the most stable structure of protein.

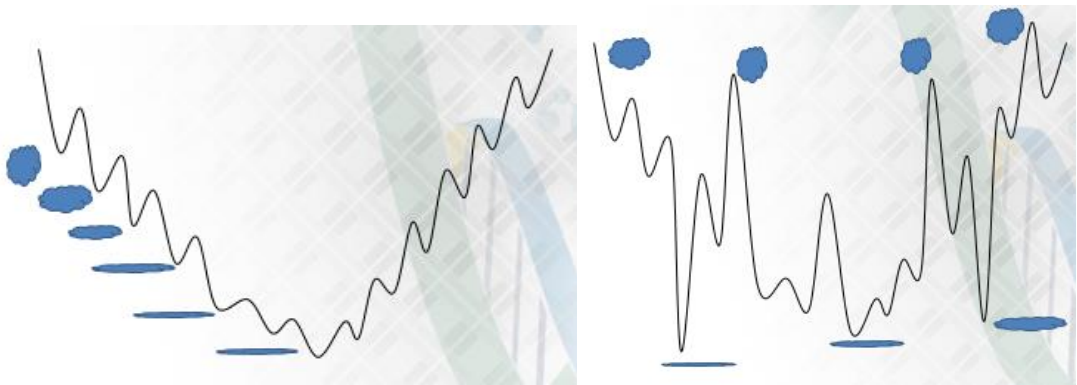


Figure 43.6.3: Determining all energy levels

3. **OPTIMAL ENERGY FUNCTION**

This function is easy to design. And we should remember that native structure of protein not always found at the global minimum. So, we have not clear way to generate alternative structure.

Module138: Pros and Cons of Ab Initio Modelling

Text (08:00)

1. **BACKGROUND**

The native structure of protein not always found at the global minimum. So, we have not clear way to generate alternative structure.

2. **ADVANTAGES**

Ab Initio method only fold any target protein based on physical atomic properties. And these predictions of proteins are mostly accurate and correct which describe the process of natural folding.

3. **DISADVANTAGES**

Ab Initio method is very difficult to design (energy function). And this method is also very slow because of large number of possibilities. E.g. 10^{12} steps are needed to simulate protein folding for medium sized protein structures.

4. **CHALLENGES IN AB INITIO MODELLING**

It's very hard to accurately describe energy functions that can reliably differentiate native and non-native structures. It has large number of calculations.

Module139: Summary of Structural Modelling - I

Text (9:00)

1. **STRATEGIES OF STRUCTURAL MODELLING**

There are many types of structural modelling. But here we will discuss only three types.

- **Homology Modelling**
- Fold Recognition
- Ab Initio Modelling

1.1. HOMOLOGY MODELLING

First, we determine the homologous sequence if available in database then, we will follow other 7 steps of homology modelling. (shown below).

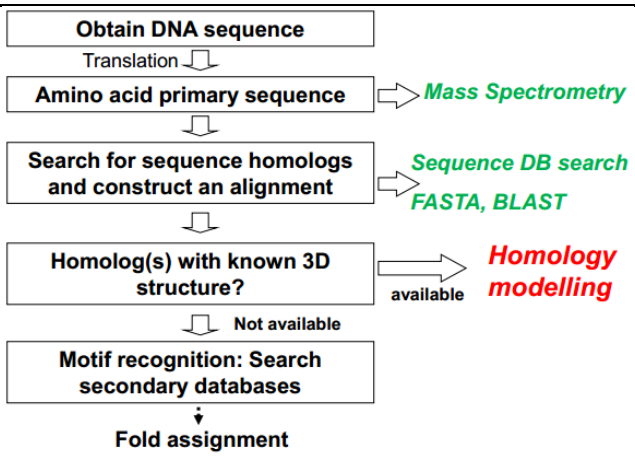


Figure 44.1.1: Initial steps for structural modelling

2. 7 STEPS

After determination of template sequence, we flow next steps. In homology modeling, there are 7 steps for prediction of proteins.

- **Step 1:** Template recognition and initial alignment
- **Step 2:** Alignment correction
- **Step 3:** Backbone generation
- **Step 4:** Loop modeling
- **Step 5:** Side-chain modeling
- **Step 6:** Model optimization
- **Step 7:** Model validation

3. CONCLUSION

We can only use homology modelling if we have high identity and high alignment score. If unknown protein lies in "twilight zone" then, we use other techniques.

Module140: Summary of Structural Modelling - II

Text (7:00)

1. STRATEGIES OF STRUCTURAL MODELLING

There are three types of structural modelling. But here we will discuss only three types.

- Homology Modelling
- **Fold Recognition**
- Ab Initio Modelling

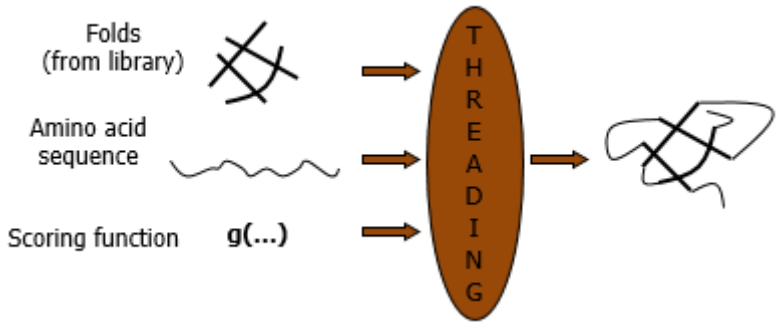


Figure 44.2.1: General input & output in fold recognition technique

1.1. FOLD RECOGNITION

First, we determine the homologous sequence if its available or not we will move towards next steps (shown below).

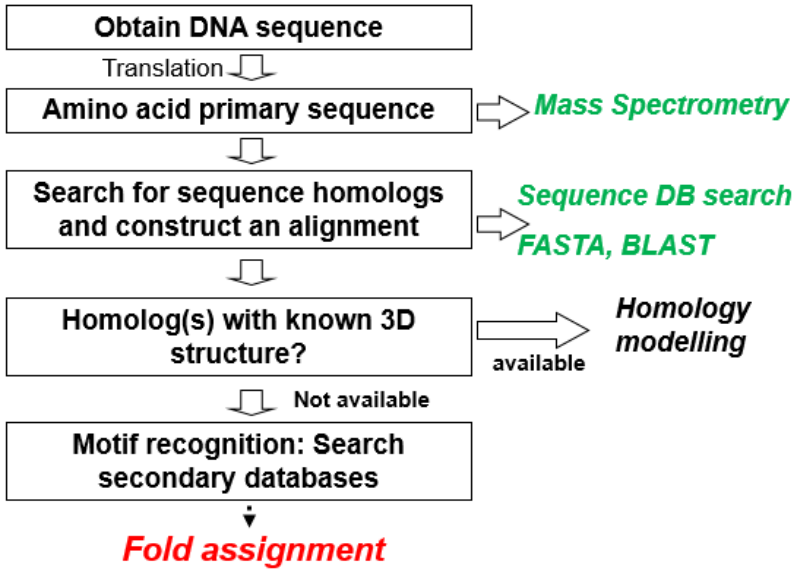


Figure 44.2.2: Initial steps before structural modelling

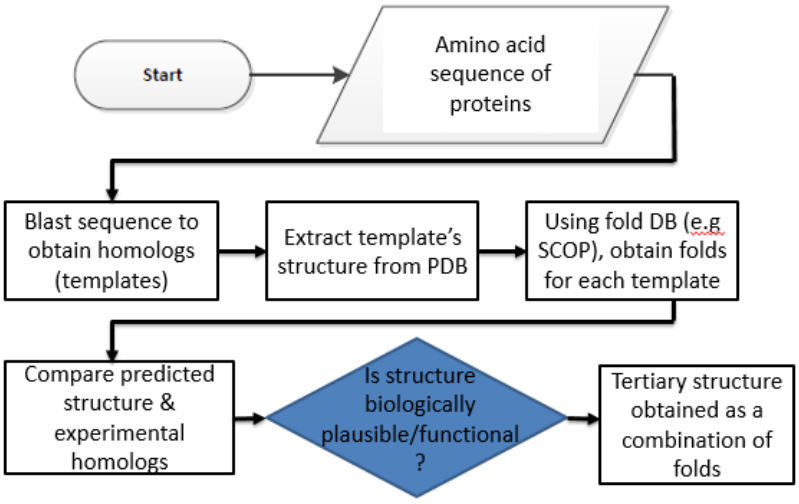


Figure 44.2.3: Flow chart of threading

2. CONCLUSION

In threading, amino acid sequence passes through each fold in the database. And threading compute best matching using a scoring function.

Module141: Summary of Structural Modelling - III

Text (8:00)

1. STRATEGIES OF STRUCTURAL MODELLING

There are three types of structural modelling. But here we will discuss only three types.

- Homology Modelling
- Fold Recognition
- **Ab Initio Modelling**

1.1. AB INITIO MODELLING

Whole ab Initio method works in two levels which are:

- **Level 1:** Energy optimization in Ab Initio modelling
  - **Step 1:** Start with a rough initial model
  - **Step 2:** Define an energy function mapping structures to energy values
  - **Step 3:** Solve the computational problem of finding the global minimum

- **Level 2:** Simulation of the folding process
  - **Step 1:** Build an accurate initial model (including energy and forces)
  - **Step 2:** Accurately simulate the dynamics of the protein folding process
  - **Step 3:** The native structure will steadily emerge

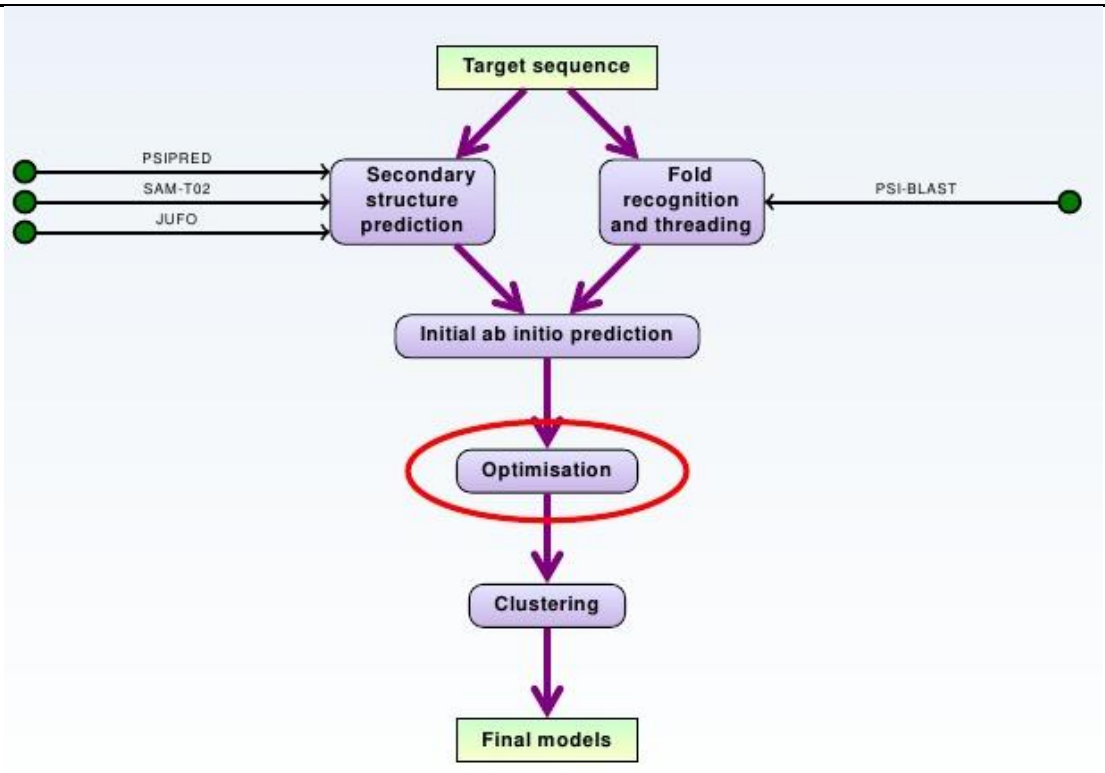


Figure 44.3.1: Workflow of Ab Initio modelling

2. CONCLUSION

Ab Initio method predicts the structure of proteins on the basis of physical models. Amount of energy which is released during folding also computed for prediction of structure. Ab Initio method is very good for prediction of proteins but still there are some limitations.

Module142: Review of Sequence Analysis

Text (8:00)

1. HOW DO WE SEQUENCE?

Generally, we classify sequences into two ways.

- **Genomes:** In this type of sequence, we sequenced at genetic level in other meanings, in the form of nucleotides. E.g. AAAACCCGGGTTT etc.
- **Proteomes:** In this type of sequence, we sequenced at protein level in other meanings, in the form of amino acids e.g. AEPEVLEGGI etc.

2. HOW DO WE COMPARE SEQUENCE?

We compared sequences into two ways.

- **Pair-wise Sequence Alignment:** In this type of comparison, we compared two sequences with each other.
- **Multiple Sequence Alignment:** In this type of comparison, we compared more than two sequences with each other.

3. **TYPES OF ALIGNMENT**

Mainly there are two types of alignment.

- **Global Alignment:** In this type of alignment, we align whole sequences. In this alignment, we usually used Needle-Wunsch Algorithm.
- **Local Alignment:** In this type of alignment, we align some specific region/part of sequence. In this alignment, we usually used Smith-Waterman Algorithm.

4. **ADVANCED TOOLS**

There are many types of tools for alignment. But two tools of alignments are most popular.

- **Fast Alignment (FASTA):** In this type of tool, we align the sequences. But it is not guaranteed that FASTA can find best alignment between query and alignment because it prefers speed.
- **Basic Local Alignment Search Tool (BLAST):** In this type of tool, we align the sequences. BLAST can search sequence databases and identify unknown sequences by comparing them to the known sequences. This can help identify the parent organism, function and evolutionary history. Updated version of FASTA is BLAST.

5. **ONLINE DATABASES**

There are many online databases. But two databases are most popular.

- **GenBank:** In this type of database, we have genetic sequences which is regularly updated Internationally.
- **UniProt:** In this type of database, we have protein sequences which is regularly updated Internationally.

6. **ONLINE PORTALS**

There are many online portals. But here we discuss only three portals.

- **Ensemble:** It is genome search engine which is used to search the genome of every recorded species. And it is regularly updated.
- **ExPASy:** It provides access to a variety of online databases and tools. Depending upon your requirement, you can find sequence information from ExPASy.

UniprotKB: It is the central hub for the collection of functional information's on protein. And this is a part of UniProt.

Module143: Review of Phylogenetics

Text (9:00)

1. **MOLECULAR EVOLUTION**

We divide molecular evolution into three types.

- **Insertions:** In this type of evolution, nucleotide or amino acid inserted into the sequence and affects the overall functionality.
- **Deletions:** In this type of evolution, nucleotide or amino acid deleted into the sequence and affects the overall functionality.

- **Substitutions:** In this type of evolution, nucleotide or amino acid substituted with another nucleotide or amino acid, respectively and affects the overall functionality.

2. **PHYLOGENETIC TREES: BRANCH LENGTH**

   There are many types of trees according to the area of classification. But here we will discuss two types of trees which are based on branch length.
   - **Scaled Trees:** Branch lengths are equal to the magnitude of variance in the nodes. Nodes represent the common ancestors between two species.
   - **Unscaled Trees:** Only representing the relationship between sequences.

3. **RATE OF EVOLUTION**

   As we know rate of evolution is usually different in different species.
   - **With clock:** We study evolution by considering time.
   - **Without clock:** We does not consider time when we are studying the evolution

4. **PHYLOGENETIC TREES: ANCESTORS & DIRECTION OF EVOLUTION**

   There are many types of trees based on the area of classification. But here we will discuss two types of trees which are based on direction of evolution.
   - **Rooted Trees:** It gives the idea about common ancestors and tell us the direction of evolution.
   - **Unrooted Trees:** It neither gives the idea about common ancestors and nor tell us the direction of evolution.

5. **UPGMA**

   It stands for Unweighted Pair- Group Method using Arithmetic Average. In this method we study the distance between species and their common ancestors.

6. **CLUSTERING VS. NON-CLUSTERING METHOD**

   UPGMA is a clustering method whereas maximum Parsimony etc. are non-clustering methods. Later discussed method is beyond the scope of this book.

Module144: Review of Protein Sequencing

Text (8:00)

1. **TECHNIQUES OF PROTEIN SEQUENCING**

   There are many types of techniques for protein sequencing. But we will discuss only two of them.
   - **Edman Degradation:** It is basically two step method. First is, labeling of amino terminal residues. Second is, removing the labeled residues.
   - **Mass Spectrometry:** Mainly it consists of two steps which are MS^1 and MS^2. And further it is dividing on many steps. It generates spectrum against proteins or peptides. It is more reliable method than Edman degradation.

2. **IMPORTANT TERMINLOGIES IN MASS SPECTROMETRY**

Whole process of mass spectrometry divides into small functions.

- **Protein ionization:** Protein ionizes in mass spectrometer by addition or removal of ions.
- **Mass analysis:** We analyzed the masses of proteins by the spectrum of mass spectrometry.
- **Protein Fragmentation:** We converts protein into small fragments by using three types of fragmenters (but one used at a time) which are ETD, CID, ECD.
- **MS^1:** Mass spectrometry have two types of level but having same types of steps. In MS^1 we perform mass spectrometry one times and then, get results.
- **MS^2:** Mass spectrometry have two types of level but having same types of steps. In MS^2 we perform mass spectrometry two times and then, get results.
- **Estimating and scoring whole protein mass:** We estimate and score the whole protein by mass spectrometry.
- **Extracting and scoring Peptide Sequence Tags:** By mass spectrometry we break down peptides into peptide sequence tags and then, extract and score it down.
- **Searching post-translational modification:** Some tools and techniques are help us to study the post-translational modifications. This is not encoded by the original genome. Therefore, these modifications tend to malfunction of protein etc.

3. **COMPOSITE SCORING SCHEMES: ONLINE TOOLS**

There are some online tools which are used for either Bottom Up proteomics (BUP) and/or Top Down Proteomics (TDP) which are:

- Mascot
- Sequest

Prosight PC

Module145: Review of RNA Structure Prediction

Text (11:00)

RNA is a hereditary material in many organisms like plants and viruses etc. We use different approaches for determining the structure of RNA, Atomic Force Microscopy is one of them. It's very important to predict the structure of RNA because RNA involves in transferring the information of DNA and it also have some other vital functions. Therefore, structure of RNA reflects its functionality.

1. **RNA SECONDARY STRUCTURE**

Mainly RNA has four types secondary structures which are:

- Hairpin loop
- Bulges
- Helices
- Junction or Intersection

2. **CONCEPTUAL BASIS OF STRUCTURAL PREDICTION**

RNA releases energy when nucleotides formed bond together. And lower the energy

increases the stability of the RNA.

3. **ALGORITHM FOR PREDICTING RNA STRUCTURE**

There are many different types of algorithms designed by different approaches so, that we predict the structure of RNA more accurately. Some algorithms are:

- Dot plot
- Zuker's Algorithm
- Martinez Algorithm
- Nussinov Jacob Algorithm

There are many online RNA structural databases which are readily available and up to date, provides information regarding different parameters. And there are many online tools available which predicts the structure of RNA by inputting sequence.

Module146: Review of Protein Structures

Text (9:00)

1. **TYPES OF PROTEIN STRUCUTRES**

Generally, we divide structures of proteins into four types.

- **Primary structure (1' structure):** A structure having linear sequence of amino acids.
- **Secondary structure (2' structure):** A structure which is formed by 1' structure. This type of structure is more complex than 1' structure.
- **Tertiary structure (3' structure):** A structure which is formed by 2' structure. This type of structure is more complex than 2' structure.
- **Quaternary structure (4' structure):** A structure which is formed by 3' structure. This type of structure is most complex than all other protein structures.

2. **TECHNIQUES FOR DETERMINING THE PROTEIN STRUCTURES**

There are many techniques which are used for determining protein structures some are:

- X-ray crystallography
- NMR spectroscopy

3. **PROTEINS: SEQUENCE VS. STRUCTURE**

We know more number of sequence of proteins than structure of proteins. Because determination of sequence is easy method as compared to the determination of structure. Protein mostly exist in 3-D complex conformation so, it's practically difficult to determine its structure.

4. **TYPES OF SECINDARY STRUCUTRE PROTEINS**

Generally, we divide structures of proteins into four types.

- Helices
- Beta Sheets
- Coils
- Loops

Based on DSSP, 2' structure also divided into 3 and 8 types, DSSP-3 and DSSP-8 respectively.

5. **STRUCTURAL PREDICTION OF PROTEINS**

   Amino acids have propensities to form specific 2' structures that's the foundation on which algorithm works to predict the structure of proteins. **Chou Fasman** is very famous algorithm used for the prediction of amino acid.

   We can acquire data of proteins regarding its structure and predict its structures by using different approaches which are:

   - Protein Data Bank (PDB)

Online tools for prediction of proteins

Module147: Review of Homology Modelling

Text (7:00)

1. **TYPES OF PROTEIN STRUCUTRES**

   Generally, we divide structures of proteins into four types.

   - **Primary structure (1' structure):** A structure having linear sequence of amino acids.
   - **Secondary structure (2' structure):** A structure which is formed by 1' structure. This type of structure is more complex than 1' structure.
   - **Tertiary structure (3' structure):** A structure which is formed by 2' structure. This type of structure is more complex than 2' structure.
   - **Quaternary structure (4' structure):** A structure which is formed by 3' structure. This type of structure is most complex than all other protein structures.

2. **JUSTIFICATION FOR HOMOLOGY MODELLING**

   We know a greater number of sequence of proteins than structure of proteins. Because determination of sequence is easy method as compared to the determination of structure. Protein mostly exist in 3-D complex conformation so, it's practically difficult to determine its structure.

3. **STRATEGIES FOR STRUCTURAL PREDICTION**

   There are many strategies for structural prediction but here we will discuss only three of them.

- **Homology Modelling:** In this strategy, we compare two proteins if they have good matching score; known protein (all parameters known) and unknown protein (some parameters are unknown) then, find then, unknown parameters of unknown proteins.

- **Fold Recognition:** If we have not significant matches then, we use this method. In this method, we mount the residues of unknown proteins onto the known protein when it fits on it. Then, we conclude it.

- **Ab Initio Modelling:** If above two methods failed then, we use this method. It computes

energy then, it formed structure of protein. That structure has minimum energy so, it has maximum stability. And at the end we predict the structure of proteins.

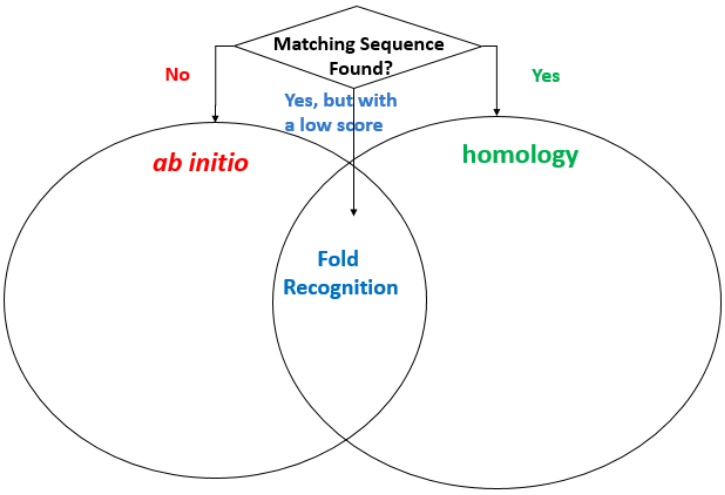


Figure 45.5.1: Comparative analysis of three different modelling techniques

Below graph show that if known vs. unknown protein exist in "safe zone" then, we use **Homology Modelling**, if they are in "twilight zone" then, we use **Ab Initio Modelling**. And in between we use **Fold recognition**.

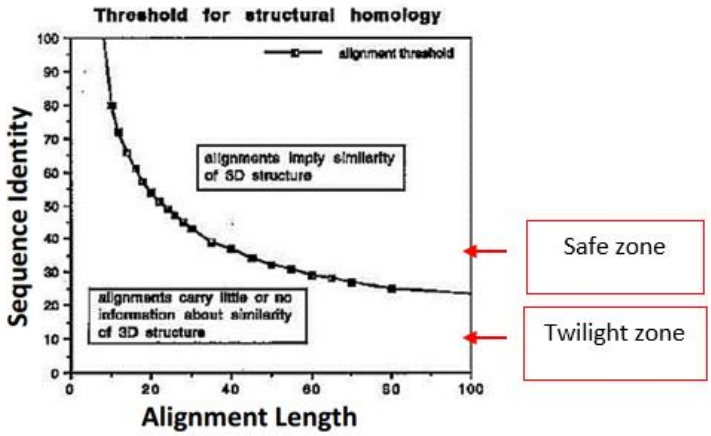


Figure 45.5.2: Graphical representation of sequence identity and alignment length

4. **STRUCTURAL PREDICTION OF PROTEINS**

   We can acquire data of proteins regarding its structure and predict its structures by using different approaches which are:

   - Protein Data Bank (PDB)

   Online tools for prediction of proteins e.g. **"ITASSER"**

Module148: Conclusions from this Course

Text (9:00)

1. **DEFINE BIOINFORMATICS**
   Bioinformatics is an interdisciplinary science which uses computational methods to acquire knowledge from biological data.

2. **WHY IT NEEDED?**
   We need Bioinformatics to address the biological problems with the help of computational algorithms.

3. **AREAS WITHIN BIOINFORMATICS**
   Bioinformatics is an interdisciplinary science at the cross-roads of biology, mathematics, computer science, chemistry and physics.

4. **ANALYSIS OF BIOLOGICAL DATA**
   We store, process and analyze the data regarding living organisms. For this purpose, we used different types of algorithms to perform various types of jobs.

5. **SPECIFIC AREAS**
   Mainly in bioinformatics we usually focus on three areas.
   - Comparing sequences
   - Comparing structures
   - Predicting structures

6. **APPROACHES**
   To study specific areas, we use different approaches.
   - Algorithms
   - Databases
   - Online tools

   We study the algorithms to study specific areas (discussed above) so, that we can extract more information. And we also use different new algorithms which gives better and detailed results than, older ones.

Module149: Advanced Follow-up Courses

Text (14:00)

Foundation of bioinformatics contains a fined mixture of different fields. Its interdisciplinary field. All topics which are already discussed in the whole book has undergone a lot of development and they are still growing with the help of new approaches.

1. **TOPICS FOR COMPUTATIONAL GENOMICS**
   For advanced level of study in Genomics, you may take "**Computational Genomics**" course. You will study given below topics and much more:
   - Gene Assembly
   - Gene Finding

- Annotation
- GWAS etc.

2. **TOPICS FOR COMPUTATIONAL PROTEOMICS**
   For advanced level of study in Proteomics, you may take "**Computational Proteomics**" course. You will study given below topics:
   - Protein Sequencing
   - PTM search
   - Structure Modelling
   - PPI Studies

3. **TOPICS FOR SYSTEMS BIOLOGY**
   For advanced level of study in Integrative Biology, you may take "**Systems Biology**" course. You will study given below topics and much more:
   - Metabolomics
   - Transcriptomics
   - Network Biology etc.

4. **OTHER COURSES**
   Now, there are also another cutting-edge course on:
   - Nano-Bio-IT
   - Computational Drug Design

Personalized Medicine

Module150: Careers in Bioinformatics

Text (08:00)

1. **BACKGROUND**
   Pakistan faces a problem of limited infrastructure. And its onset of digital revolution. In the field of Bioinformatics emergence of data is most precious commodity all over the world specifically in the form of health data. Health and diseases are the big challenges of mankind all over the world. Therefore, they also did very appreciative work by combatting with disease. And this work is still going on.

2. **UNIQUE OPPORTUINITES IN PAKISTAN**
   In Pakistan, for working in the field of Bioinformatics we require only two things:
   - Smart mind
   - Internet connected computer

3. **ONE MAN COMPANY**
   You can take public databases (freely available) and use it in drug designing.

   One man vs. Roche?

4. **BIGDATA**
   You can establish your company which manages and process health bigdata. You only need basic software development skills which are coupled with Bioinformatics.

5. **NEXT DISRUPTION**
   Now-a-days, multinational companies like **GOOGLE**, **FACEBOOK** and **UBER** working onto emerge the Health and Bioinformatics. E.g. Google and Facebook are specially working separately on the interface of **THE HUMAN BRAIN** with **DIGITAL WORLD.**

   Many Pharmaceutical companies are investing into Bioinformatics human resource development.

6. **JOBS MARKET**

- Job market for Bioinformatician is very vast. You can join:
- Pharmaceutical Giants
- Research Centers & Universities
- Hospital & Diagnostic IT departments

Your own startup company

Module151: Lesson-2- Special topics in bioinformatics

Text (10:00)

- Sequencing Techniques
  - Alignment
  - Assembly
  - Gene Regulation
  - Gene Annotation
  - Tools for Next Generation Sequencing
- Linux an operating system
- Python for Bioinformatics
  - Biopython
- R for Bioinformatics
- Metagenomics
- Advance techniques in Bioinformatics

Recent Techniques In Sequencing

- High throughput sequencing also called next generation sequencing (NGS) have the capacity to sequence full genomes.
- These technologies includes Roche's 454 GS FLX, Illumina's Solexa Technology, ABI's SOLiD Technology and Ion Torrent Technology.

Module152:Lesson-3 Next Generation Techniques

Text (08:00)

Sequencing:

Sequencing is the process to determine the precise order of nucleotides or amino acids in DNA or RNA molecule respectively.

- DNA Sequencing
- RNA Sequencing
- Single molecule detection/sequencing

History of Sequencing

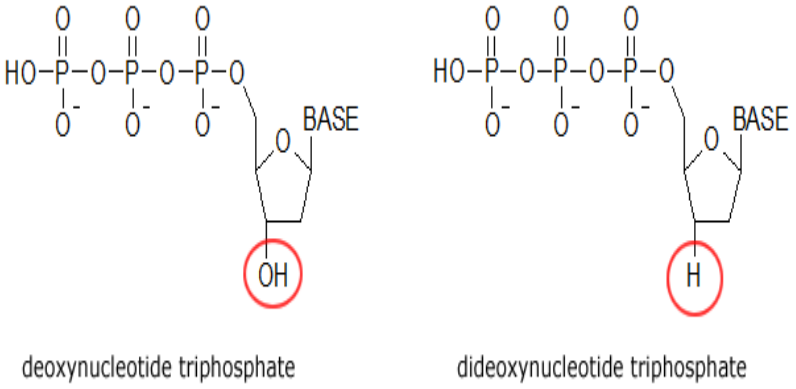
Early Sequencing was performed with transfer RNA (tRNA) via a technique developed by Richard Holley, a published in 1964.

Technique involves in breaking down RNA molecule and then baffling the fragments back together.

This technique is Time consuming due to its large molecular size

In 1988 Fredrick Sanger prospered a method that allowed sequencing of up to 50 nucleotide in length.

In 1975 Sanger developed "the plus and minus method" based on the principal of chain termination during polymerization and sequenced a complete genome of ϕX 174 bacteriophage



Locat addition of ddNTPs will terminate chain elongationion of ddNTP insertion within a nucleotide chain can be determined using gel separation.

Three innovations came about that greatly expedited the sequencing process:

Shot-gun sequencing

PCR (Polymerization Chain Reaction)

Automation of sequencing

Module153: Lesson-4 Current and Emerging Sequencing Techniques

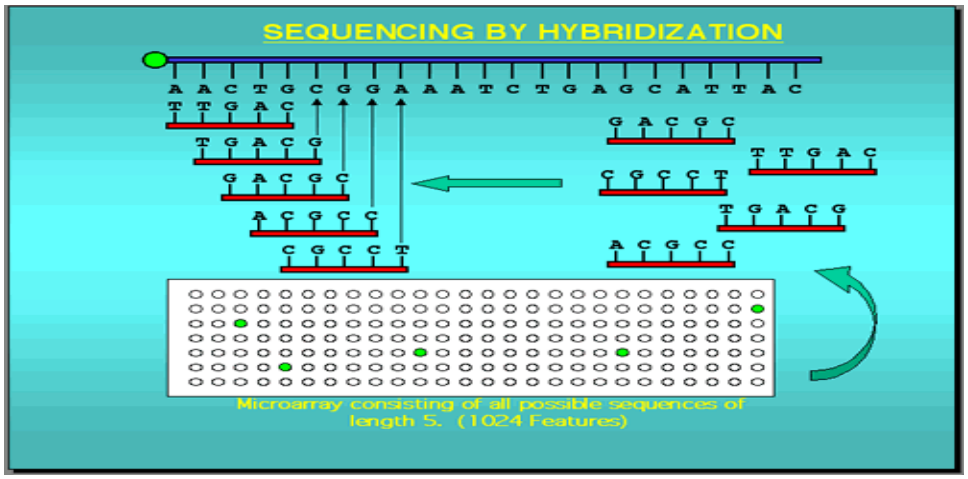
Text (9:00)

Current and Emerging Sequencing Techniques:

Sequencing By Hybridization (SBH)

The array contains all possible oligonucleotide sequences of a given length.

DNA of unknown sequence is incubated with the array.



- The target hybridizes to the array wherever there is complementation to a portion of the target.
- Hybridization of oligos are detected by fluorescence.
- The probes are organized by overlaps with one another to reconstruct the target sequence.

Limitations:

Difficult to reconstruct long sequences.

Very large libraries are required.

The normal approach to SBH is also sensitive to errors.

Latest improvement and advantages

Universal bases are used instead of normal oligonucleotides.

By acting as spacers the universal bases make consecutive probes less dependent on one another.

These are less sensitive to errors.

Does not require larger libraries.

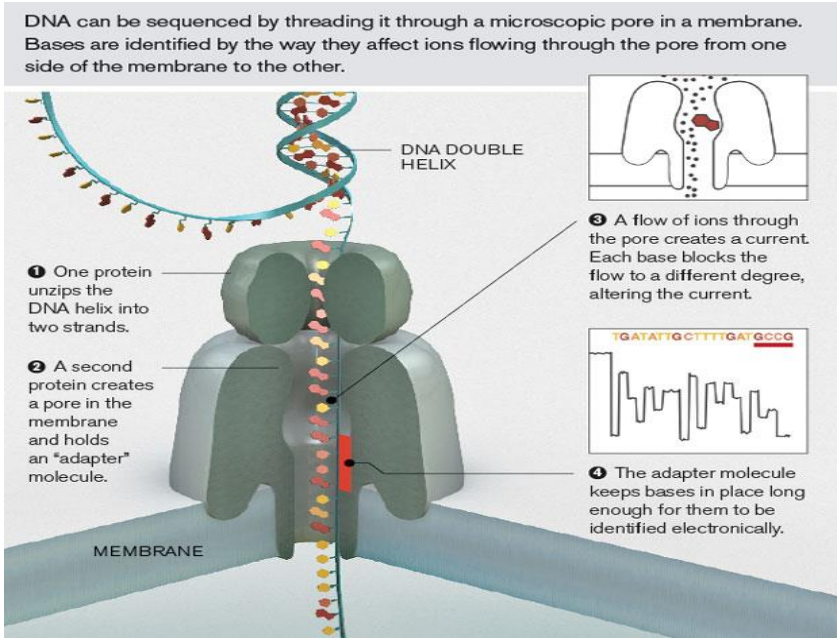
Module154: Lesson-5 Nanopore Sequencing

Text (7:00)

Nanopore Sequencing

- Determine the sequence of DNA fragments by passing DNA through a protein (or other) pore in a membrane.

  Transmembrane porins



  Oxford Nanopore became the first company to provide a commercially available nanopore sequencer in 2015 (available to community in 2012)

- Nanopore is a disruptive technology:
- Sequencer Size
- Read Length 10,000 bp to 15,000 bp average
- Potential direct RNA sequencing
- Biology Problem with Data Velocity Issues
  - Currently ~400GB/24 hours needs to be processed

  .
- ~50 (250-400 with new R9 pore) base pairs per second pass through a pore
- Need to segment signal into individual events representing base pairs and determine to which base each event corresponds

Module155:Lesson-6 Sequencing-By-Synthesis (SBS)

Text (8:00)

Sequencing-by-synthesis (SBS)

In NGS, a huge number of short length reads are sequenced parallel to each other in a single run.

The input sample is first cleaved into short length fragments. 100-150 bp.

These fragments are ligated to the generic adaptors and annealed to slide via adaptors where sequencing takes place.

Pcr amplifies each read, thus creating a spot with many copies of a same read. The sequences are then separated into single strands that are ultimately to be sequenced.

- The slide is occupied by a large number of nucleotides and DNA polymerases
- The nucleotides are fluorescently labeled with the corresponding color of the bases they are related to.
- A terminator also exists, to make sure that only one base is added at a time.
- Each slide is captured, and in each location, there will be a fluorescent signal that would be indicating the base that has been added.

The slide is then prepared for the following cycle.

- The terminators are removed, this allows the next base to be added, and the fluorescent signal is removed so that it does not happen to contaminate the next image.
- The process is repeated, ensuring addition of one nucleotide at a time and imaging side by side.
- Computational forces are then used to detect the base at each location in each image and then these are used to construct a complete sequence.

Advantages

Allows parallel sequencing.

Use of photons requires no additional chemical reagents.

Clean products with no need of subsequent purification.

Module156: Lesson-7a- Other Sequencing techniques

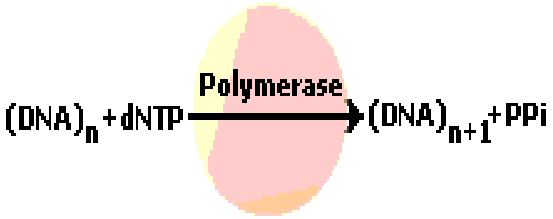
Text (9:00)

Single-nucleotide addition (SNA)

Pyrosequencing

- Non-sanger nonfluorescence technique that quantitatively measures released PPi

- Pyrogram corresponds to complementary base



Applications and advantages

- SNP analysis
- Ideal for rapidly mutating organisms
- Quantifications provide additional data

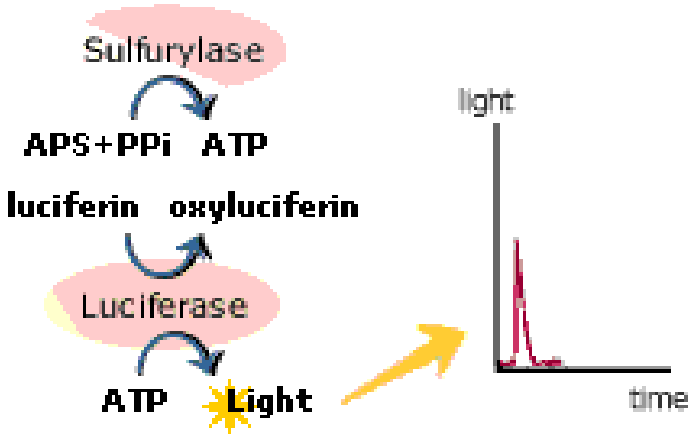
Single-nucleotide addition (SNA)

Pyrosequencing

- Non-sanger nonfluorescence technique that quantitatively measures released PPi
- Pyrogram corresponds to complementary base

Applications and advantages

- SNP analysis
- Ideal for rapidly mutating organisms
- Quantifications provide additional data



nucleotide incorporation generates light seen as a peak in the pyrogram

Single-nucleotide addition (SNA)

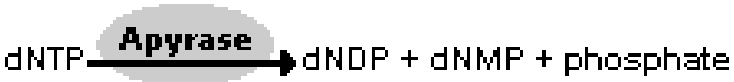
Pyrosequencing

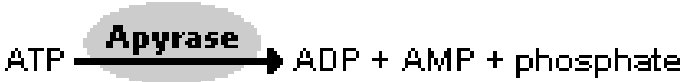
- Non-sanger nonfluorescence technique that quantitatively measures released PPi

- Pyrogram corresponds to complementary base

Applications and advantages

- SNP analysis
- Ideal for rapidly mutating organisms
- Quantifications provide additional data





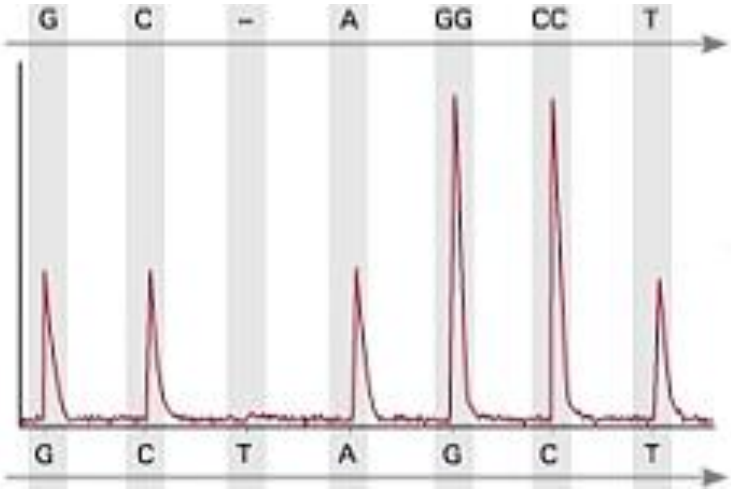
Single-nucleotide addition (SNA)

Pyrosequencing

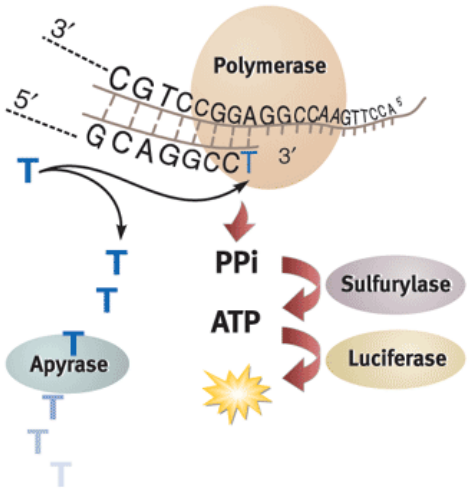
- Non-sanger nonfluorescence technique that quantitatively measures released PPi
- Pyrogram corresponds to complementary base

Applications and advantages

- SNP analysis
- Ideal for rapidly mutating organisms
- Quantifications provide additional data



deoxyadenosine alfa-thio triphosphate (dATPαS) is used as a substitute for the natural deoxyadenosine triphosphate (dATP) since it is efficiently used by the DNA polymerase, but not recognized by the luciferase



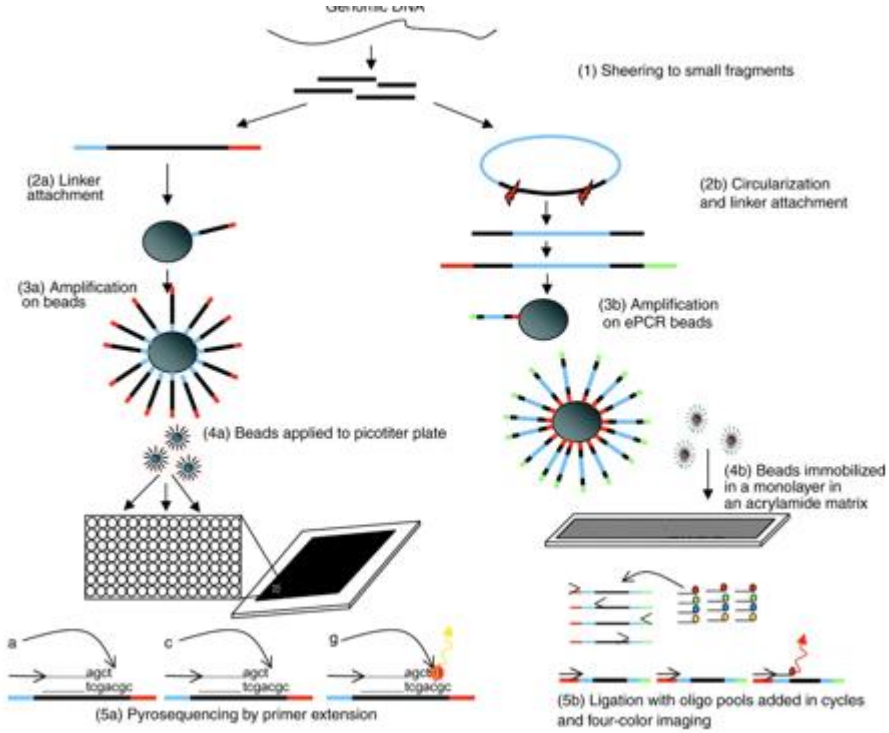
Limitations

- Short sequence reads
- Homopolymer repeat problems

Module157: Lesson-7b- Other Sequencing techniques

Text (8:00)

Polony Technology



Polony technology

Polony (polymerase colony) is amplified product from single DNA molecule in acrylamide gel.

Sequencing done by the incorporation of cleavable fluorescent labeled nucleotide.

Advantage

Scalability is easy by using 1µm magnetic beads.

Disadvantage

Failure in cleaving dye moiety.

Module158: Lesson-7c- Other Sequencing techniques

Text (11:00)

Comparative genome sequencing

Test DNA is hybridized with reference DNA to identify regions of genomic differences. Genomic different regions are sequenced to identify SNPs.

Advantages

Fast, accurate sequencing of the regions of interest

Comparative genome sequencing

Test DNA is hybridized with reference DNA to identify regions of genomic differences. Genomic different regions are sequenced to identify SNPs.

Advantages

Fast, accurate sequencing of the regions of interest

Table 1. Comparative genome sizes of humans and other model organisms

Organism	Estimated size (base pairs)	Chromosome number	Estimated gene number
Human (*Homo sapiens*)	3 billion	46	~25,000
Mouse (*Mus musculus*)	2.9 billion	40	~25,000
Fruit fly (*Drosophila melanogaster*)	165 million	8	13,000
Plant (*Arabidopsis thaliana*)	157 million	10	25,000
Roundworm (*Caenorhabditis elegans*)	97 million	12	19,000
Yeast (*Saccharomyces cerevisiae*)	12 million	32	6,000
Bacteria (*Escherichia coli*)	4.6 million	1	3,200

Cyclic reversible terminator (CRT)

Sequencing by CRT consists of three steps; incorporation, imaging and deprotection. The reversible terminator must be cleaved efficiently with photocleaving groups like 2-nitrobenzyl group.

Advantages

Avoids gel electrophoresis, functions in highly parallel fashion, high throughput, speed and accuracy.

Module159: Lesson-8a-NGS data formats

Text (9:00)

File Types

Many software packages have been developed for the analysis of DNA and protein sequences.

A variety of different file formats have been developed to store or analyse DNA and protein sequence information

The various software packages will usually only accept a specific file format.

The situation is made worse by the fact that different databases hold the information in different file formats

An essential skill is be able to recognize the different formats and to be able to interconvert files between formats

Main file formats used in Bioinformatics

ASN.1

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
//
```

EMBL, SwissProt

FASTA, FASTq

GCG

GeneBank

Phylip/PIR

Main file formats used in Bioinformatics

ASN.1

EMBL, SwissProt

FASTA, FASTq

GCG

GeneBank

Phylip/PIR

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

ASN.1

EMBL, SwissProt

FASTA**, FASTq**

GCG

GeneBank

Phylip/PIR

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

ASN.1

EMBL, SwissProt

FASTA, FASTq

GCG

GeneBank

Phylip/PIR

```
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
```

ASN.1

EMBL, SwissProt

FASTA, FASTq

GCG

GeneBank

Phylip/PIR

```
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
//
```

ASN.1

EMBL, SwissProt

FASTA, FASTq

GCG

GeneBank

Phylip/PIR

SRA format

The SRA is a "raw data" archive, and requires per-base quality scores for all submitted data. Therefore, FASTA and other sequence-only formats are not sufficient for submission! FASTA can, however, be submitted as a reference sequence(s) for BAM files or as part of a FASTA/QUAL pair.

Text formats, such as FASTQ, are supported, but are not the preferred file format for SRA submission.

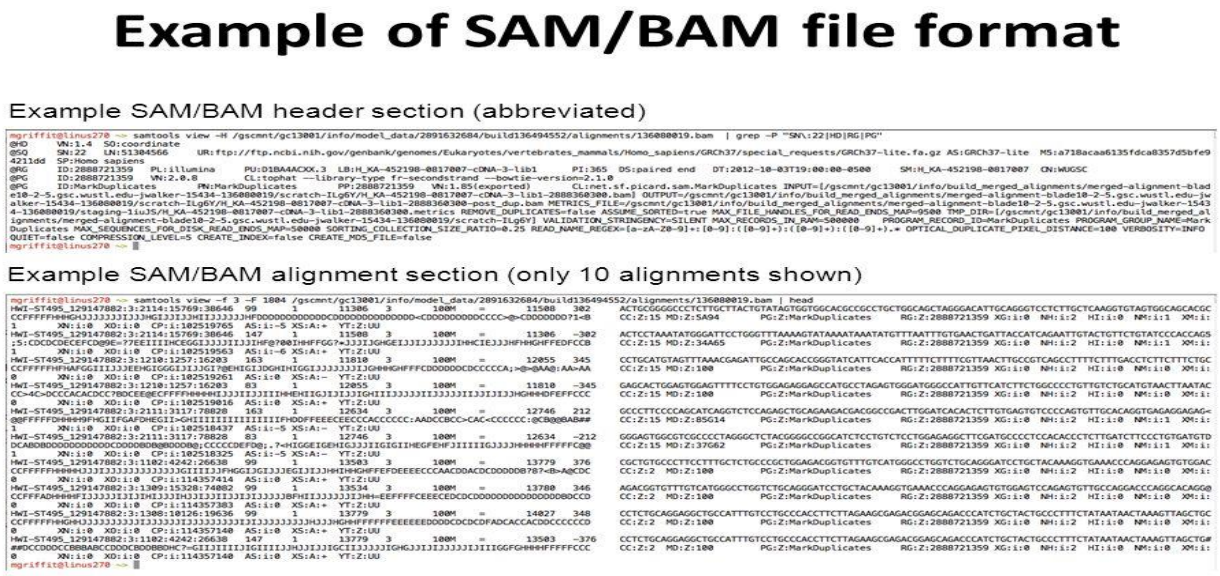
SRA prefers files such as BAM, SFF, and HDF5 formats

Module160: Lesson-8b- NGS data formats

Text (7:00)

BAM files

Binary alignment/map files (BAM) are the preferred SRA submission format. BAM is a compressed version of the sequence alignment/map (SAM) format BAM files can be decompressed to a human-readable text format (SAM) using SAM/bam-specific utilities and can contain unaligned sequences as well.



Module 2 – RNA-seq alignment and visualization bioinformatics.ca

SAM Files

Sequence Alignment Map format. A TAB-delimited text format consisting of a header section and alignment body/section. The Each line of header sections start with @ sign and alignment section don't. 11 compulsory fields having alignment information will be present in each alignment line such as aligner specific information and mapping position etc.

11 columns of SAM File

1. Read ID
2. The SAM flag
3. Chromosome/contig read aligned to
4. PosiCon which read aligned to
5. Mapping quality score
6. Cigar string
7. Chromosome/contig which read pair aligned to
8. PosiCon which read pair aligned to
9. Insert Size
10. Sequence in bases
11. Quality score for each base

11 columns of SAM File

1. ERR001268.25 Read ID
2. 147 The SAM flag
3. chr22 Chromosome/conCg read aligned to
4. 44549174 PosiCon which read aligned to
5. 60 Mapping quality score
6. 36M Cigar string
7. = Chromosome/contig which read pair aligned to
8. 44548985 PosiCon which read pair aligned to
9. -225 Insert Size
10. GGTTGGATGTGTATTTTT Sequence in bases
11.)(1)+.5+<.@9A%<;=0IIIHCII?III;IIIII Quality score for each base

SAM/BAM Format

SAM (Sequence Alignment Map) format is a text based format that stores alignment data.

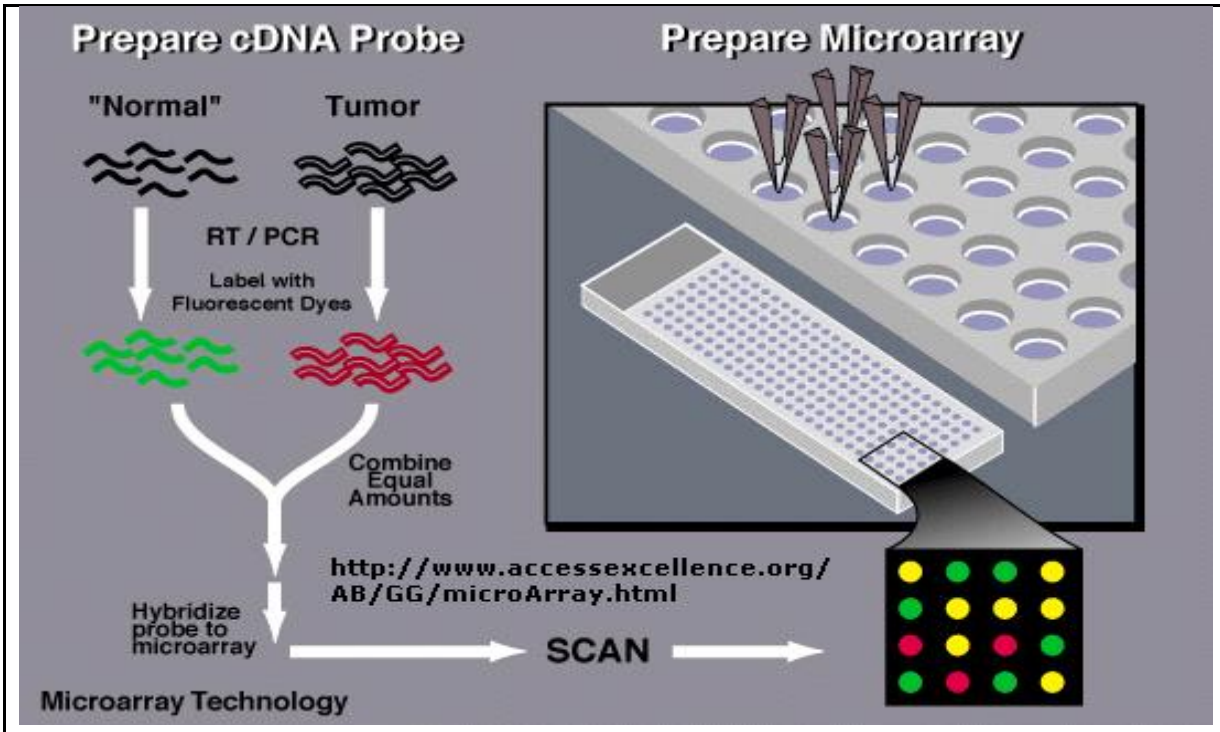
BAM (Binary Alignment Map) format is the binary version of SAM.

This is the generally accepted file format for aligned sequence data. SAM tools is a utility that can be used to convert between SAM and BAM format. http://samtools.sourceforge.net/

Module161: Lesson-9- Gene Expression Omnibus (GEO)

Text (9:00)

Microarray in general



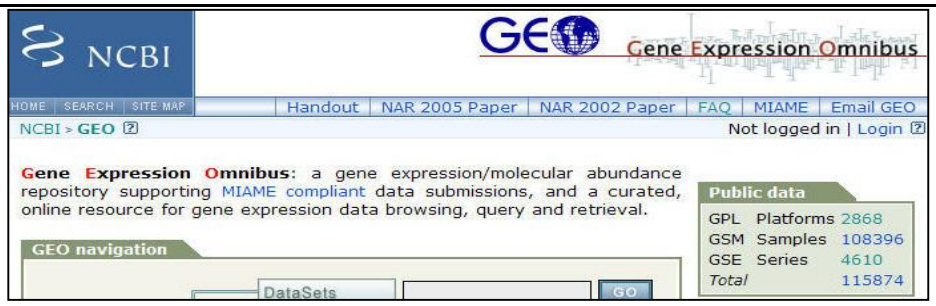
Gene Expression and Molecular abundance Data Repository

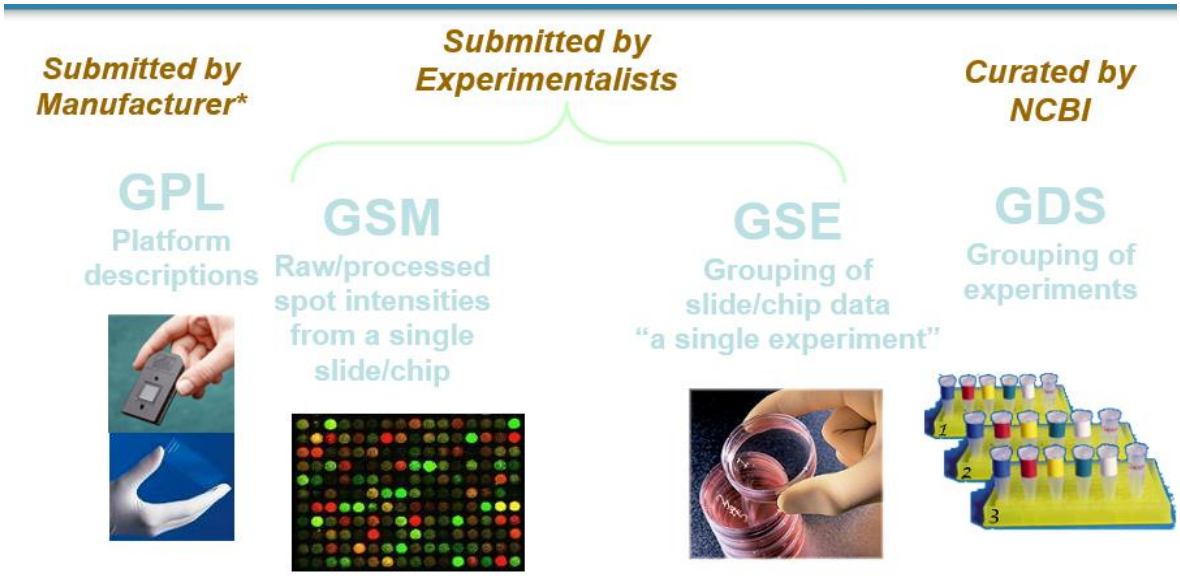
- A public repository for the archiving and distribution of gene expression data submitted by the scientific community.
- Miame compliant data.
  - Minimum information about a microarray experiment http://www.Mged.Org/workgroups/MIAME/miame.Html
- Convenient for deposition of gene expression data, as required by funding agencies and journals.
- Curated, online resource for gene expression data browsing, query, analysis and retrieval.

GEO Architecture

GEO has four kinds of data records

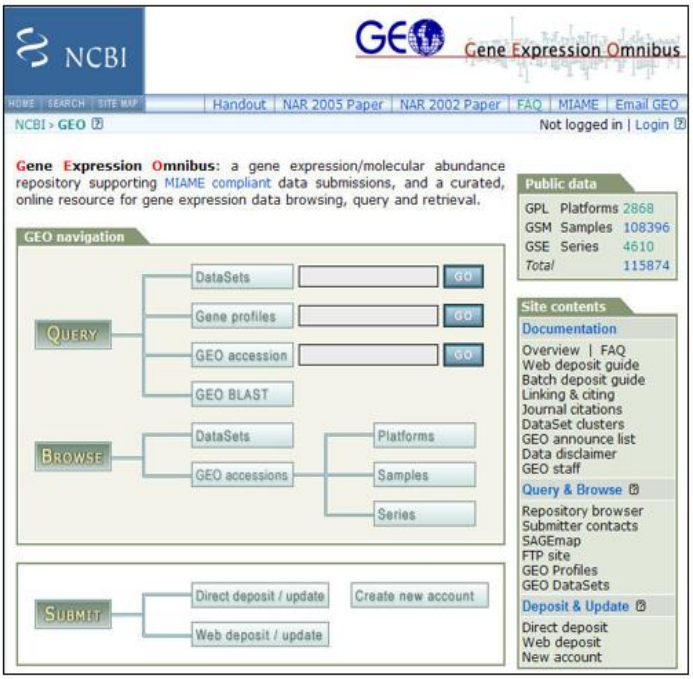
- Platform (GPL) = the technology used and the features detected.
- Sample (GSM) = preparation and description of the sample.
- Series (GSE) defines a set of samples and how they are related.
- Datasets (GDS) sample data collections assembled by geo staff.
- Submitters may provide raw data
- Original microarray scans
- Raw quantification data





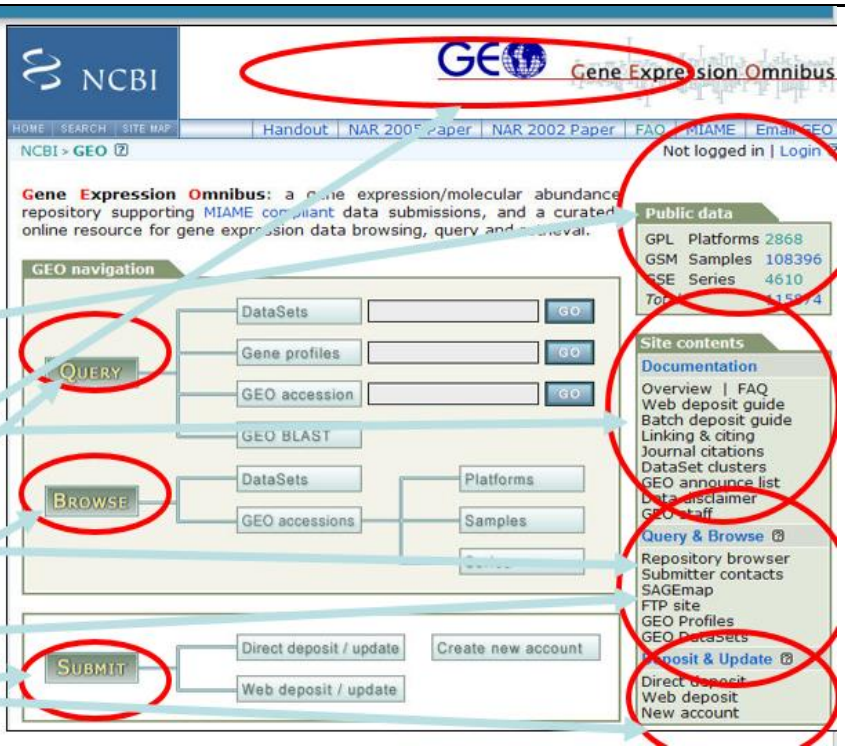
Simple interface to:

- show status
- find documentation
- query data
- browse data
- submit data

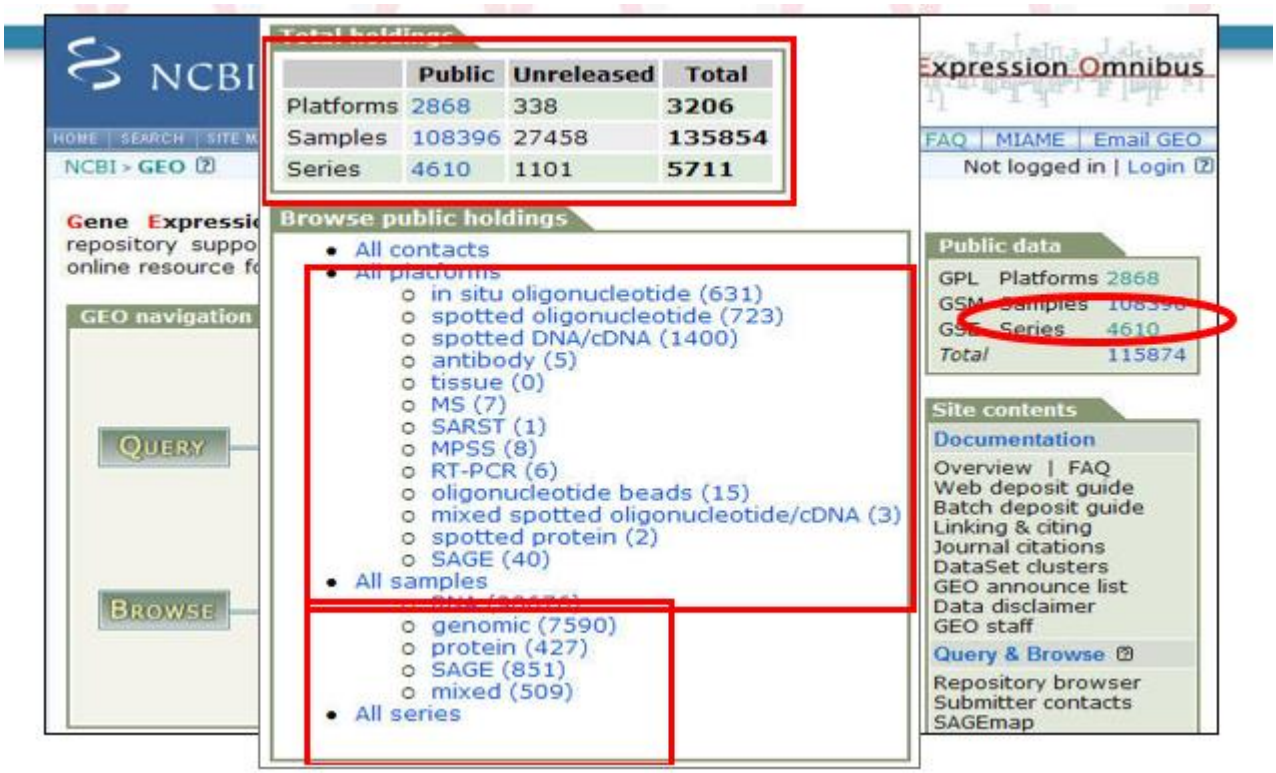


Simple interface to:

- show status
- find documentation
- query data
- browse data
- submit data



Basic Search: Repository Browser

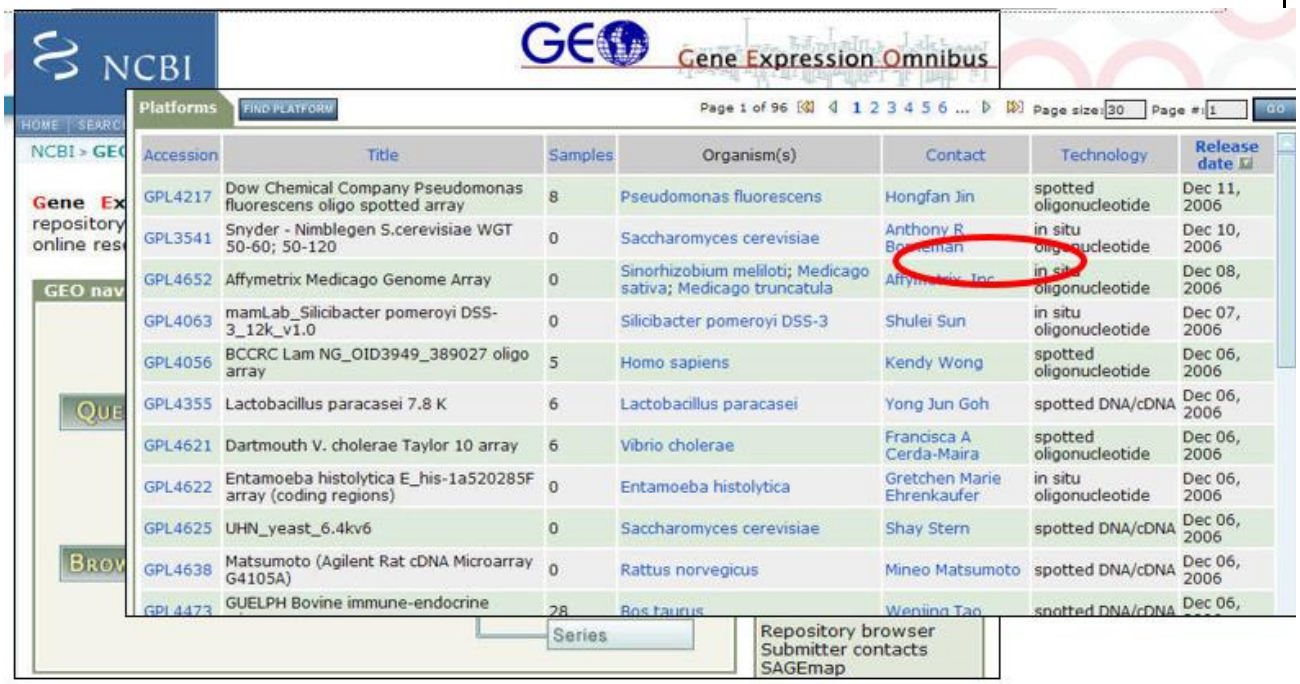


Selecting the total public data or repository browser links on the GEO home page, takes you to the repository browser, listing:

- Number of each type of submitted file, both public and unreleased

- The total number of each technology type under platforms
- The total number of each sample type

Basic Search: Browser Platforms



Accession	Title	Samples	Organism(s)	Contact	Technology	Release date
GPL4217	Dow Chemical Company Pseudomonas fluorescens oligo spotted array	8	Pseudomonas fluorescens	Hongfan Jin	spotted oligonucleotide	Dec 11, 2006
GPL3541	Snyder - Nimblegen S.cerevisiae WGT 50-60; 50-120	0	Saccharomyces cerevisiae	Anthony R Bo[illegible]man	in situ oligonucleotide	Dec 10, 2006
GPL4652	Affymetrix Medicago Genome Array	0	Sinorhizobium meliloti; Medicago sativa; Medicago truncatula	Affy[illegible] Inc	in situ oligonucleotide	Dec 08, 2006
GPL4063	mamLab_Silicibacter pomeroyi DSS-3_12k_v1.0	0	Silicibacter pomeroyi DSS-3	Shulei Sun	in situ oligonucleotide	Dec 07, 2006
GPL4056	BCCRC Lam NG_OID3949_389027 oligo array	5	Homo sapiens	Kendy Wong	spotted oligonucleotide	Dec 06, 2006
GPL4355	Lactobacillus paracasei 7.8 K	6	Lactobacillus paracasei	Yong Jun Goh	spotted DNA/cDNA	Dec 06, 2006
GPL4621	Dartmouth V. cholerae Taylor 10 array	6	Vibrio cholerae	Francisca A Cerda-Maira	spotted oligonucleotide	Dec 06, 2006
GPL4622	Entamoeba histolytica E_his-1a520285F array (coding regions)	0	Entamoeba histolytica	Gretchen Marie Ehrenkaufer	in situ oligonucleotide	Dec 06, 2006
GPL4625	UHN_yeast_6.4kv6	0	Saccharomyces cerevisiae	Shay Stern	spotted DNA/cDNA	Dec 06, 2006
GPL4638	Matsumoto (Agilent Rat cDNA Microarray G4105A)	0	Rattus norvegicus	Mineo Matsumoto	spotted DNA/cDNA	Dec 06, 2006
GPL4473	GUELPH Bovine immune-endocrine	28	Bos taurus	Wenjing Tao	spotted DNA/cDNA	Dec 06,

- All GEO submissions need to be associated with a platform file. These describe the features on a given platform, required to understand the data.
- A platform file must be submitted if one is not already present in geo.
- Commercial array platform files are submitted to geo by the manufacturer

Data Retrieval: Series Accession Page

Scope: Self Format: HTML Amount: Quick GEO accession: GSE3494

Series GSE3494

Status	Public on Oct 21, 2005
Title	An expression signature for p53 in breast cancer predicts mutation status, transcriptional effects, and patient survival
Organism(s)	Homo sapiens
Type	Tumor sample comparisons
Summary	The biological tumor samples (ie, breast tumor specimens) consisted of freshly frozen breast tumors from a population-based cohort of 315 women representing 65% of all breast cancers resected in Uppsala County, Sweden, from January 1, 1987 to December 31, 1989. Estrogen receptor status was determined by biochemical assay as part of the routine clinical procedure. An experienced pathologist determined the Elston-Ellis grades of the tumors, classifying the tumors into low, medium and high-grade tumors. The clinico-pathological characteristics accompanying each tumor include p53 status, ER status, tumor grade, lymph node status and patient age.
Overall design	All tumor specimens were assessed on U133 A and B arrays.
Contributor(s)	Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J
PubMed ID	16141321
Submission date	Oct 21, 2005
Contact name	Lance David Miller
E-mail(s)	millerl@gis.a-star.edu.sg
Phone	65 6478 8100
Fax	65 6478 9060
URL	http://www.gis.a-star.edu.sg/internet/site/investigators.php?f=cv&user_id=7
Organization name	Genome Institute of Singapore
Department	Microarray and Expression Genomics
Street address	60 Bioplois Street, #02-01 Genome
City	Singapore
ZIP/Postal code	138672
Country	Singapore
Platforms (2)	GPL96 Affymetrix GeneChip Human Genome U133 Array Set HG-U133A GPL97 Affymetrix GeneChip Human Genome U133 Array Set HG-U133B
Samples (502)	GSM79114, GSM79115, GSM79116, GSM79117, GSM79118, GSM79119

GEO Accession Results Display Options

Scope: Self Format: SOFT Amount: Quick GEO accession: GSE3494 GO

All GEO accession results pages have the same header that allows different views and formats for the data to be displayed

Scope: Self
Self
Platform
Samples
Series
Family

Scope controls what information is displayed
- Self
- Platform, Samples or Series
- Family

Format: HTML
HTML
SOFT
MINiML

Format controls how information is displayed:
- Html
- SOFT (simple omnibus format in text)
- Miniml (MIAME notation in markup language

Amount: Quick
Brief
Quick
Full
Data

Data Retrieval: Series Accession Page

Scope: Self Format: HTML Amount: Quick GEO accession: GSE3494

Series GSE3494

Status	Public on Oct 21, 2005
Title	An expression signature for p53 in breast cancer predicts mutation status, transcriptional effects, and patient survival
Organism(s)	Homo sapiens
Type	Tumor sample comparisons
Summary	The biological tumor samples (ie, breast tumor specimens) consisted of freshly frozen breast tumors from a population-based cohort of 315 women representing 65% of all breast cancers resected in Uppsala County, Sweden, from January 1, 1987 to December 31, 1989. Estrogen receptor status was determined by biochemical assay as part of the routine clinical procedure. An experienced pathologist determined the Elston-Ellis grades of the tumors, classifying the tumors into low, medium and high-grade tumors. The clinico-pathological characteristics accompanying each tumor include p53 status, ER status, tumor grade, lymph node status and patient age.
Overall design	All tumor specimens were assessed on U133 A and B arrays.
Contributor(s)	Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J
PubMed ID	16141321
Submission date	Oct 21, 2005
Contact name	Lance David Miller
E-mail(s)	millerl@gis.a-star.edu.sg
Phone	65 6478 8100
Fax	65 6478 9060
URL	http://www.gis.a-star.edu.sg/internet/site/investigators.php?f=cv&user_id=7
Organization name	Genome Institute of Singapore
Department	Microarray and Expression Genomics
Street address	60 Bioplois Street, #02-01 Genome
City	Singapore
ZIP/Postal code	138672
Country	Singapore
Platforms (2)	GPL96 Affymetrix GeneChip Human Genome U133 Array Set HG-U133A GPL97 Affymetrix GeneChip Human Genome U133 Array Set HG-U133B
Samples (502)	GSM79114, GSM79115, GSM79116, GSM79117, GSM79118, GSM79119

Biological sample summary

Design summary

Publication information

Platform (total)

Samples (total)

Data Retrieval: Sample File Summary

Scope: Self | Format: HTML | Amount: Quick | GEO accession: GSM79114 | GO

Sample GSM79114

Status	Public on Oct 21, 2005
Title	X100B08 (HG-U133A)
Sample type	RNA

Sample preparation

Source Name	Breast Tumor
Organism(s)	Homo sapiens
Characteristics	Breast Tumor tissue
Extracted molecule	total RNA
Extraction protocol	Qiagen RNeasy Mini Kit
Label	biotin
Label protocol	Approximately 10ug of total RNA was processed to produce biotynylated cRNA targets

Hybridization and data processing

Hybridization protocol	Standard Affymetrix procedure
Scan protocol	Standard Affymetrix procedure
Description	Series of 251 tumours
Data processing	Affymetrix Microarray Suite Version 5.0

Submission date	Oct 21, 2005
Contact name	Lance David Miller
E-mail(s)	millerl@gis.a-star.edu.sg
Phone	65 6478 8100
Fax	65 6478 9060
URL	http://www.gis.a-star.edu.sg/internet/site/investigators.php?f=cv&user_id=7
Organization name	Genome Institute of Singapore
Department	Microarray and Expression Genomics
Street address	60 Bioplois Street, #02-01 Genome
City	Singapore
ZIP/Postal code	138672
Country	Singapore

Platform Series

Platform ID	GPL96
Series (1)	GSE3494 An expression signature for p53 in breast cancer predicts mutation status, transcriptional effects, and patient survival

Querying GEO with IDs from Papers

PNAS

An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, a

Lance D. Miller*†, Joh
Yudi Pawitan§, Per Ha

*Genome Institute of Singa
Karolinska Institute and Ho
17177 Stockholm, Sweden

GENETICS. For the article "An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival," by Lance D. Miller, Johanna Smeds, Joshy George, Vinsensius B. Vega, Liza Vergara, Alexander Ploner, Yudi Pawitan, Per Hall, Sigrid Klaar, Edison T. Liu, and Jonas Bergh, which appeared in issue 38, September 20, 2005, of *Proc. Natl. Acad. Sci. USA* (**102,** 13550-13555; first published September 2, 2005; 10.1073/pnas.0506230102), the breast cancer microarray data discussed in this publication have been deposited in the National Center for Biotechnology Information's Gene Expression Omnibus database (GEO, www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO Series accession no. GSE3494 [NCBI GEO] .

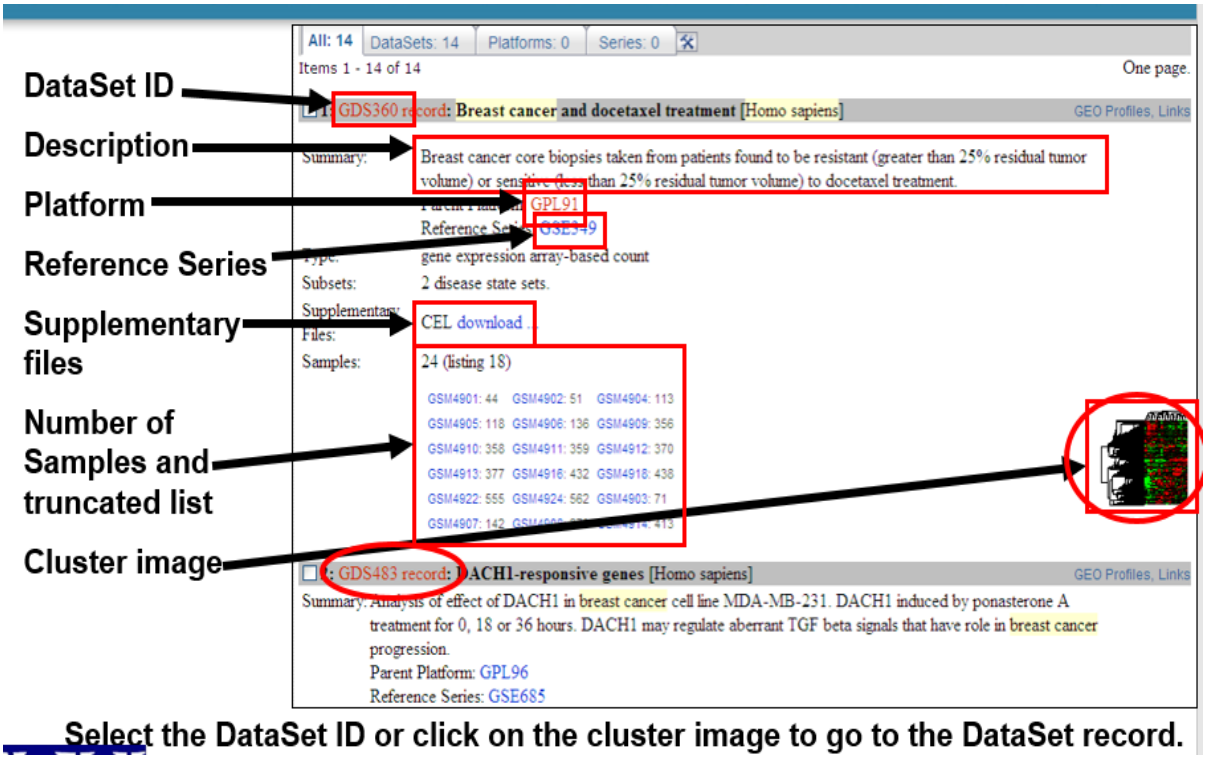
GEO navigation

QUERY

DataSets		GO
Gene profiles		GO
GEO accession	GSE3494	GO
GEO BLAST		

- A common way to access GEO data is through accessions from papers.
- Online journals include hyperlinks to the geo accession page.
- Or, at the geo home page enter the accession into the query>geo accession text box

DataSet Search Result



Module162: Lesson-10- SRA Files and SRA File Handling & SAM File

Text (9:00)

Sequence Read Archive (SRA) files are a common format used by the NCBI, EBI, and others for storing reads and read alignments. While the format provides good compression and data accessibility, some work is often needed to transform SRA-formatted files into a form suitable for visualization and other analytical processing needs

The SRA is one of the International Nucleotide Sequence Databases and this Collaboration (INSDC) sets policies and goals for the partner databases.

Goals

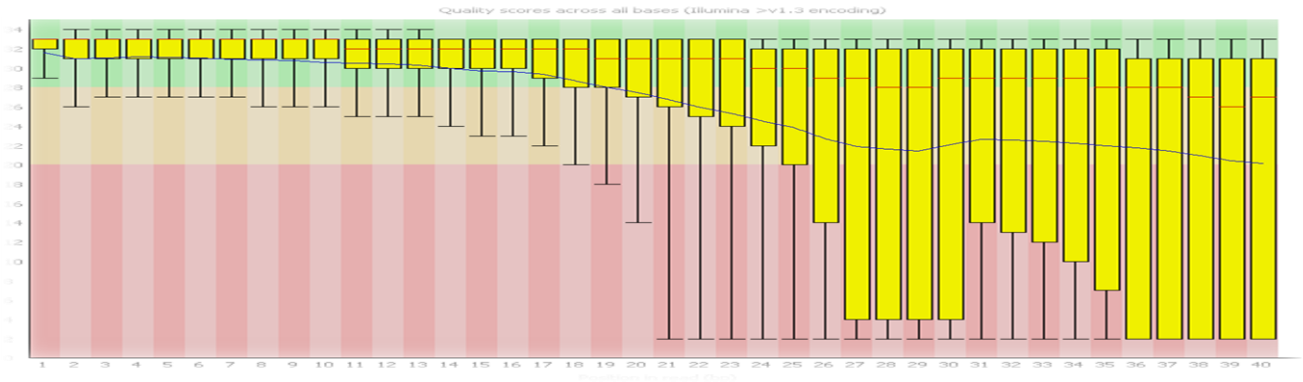
Guide for submitters of sequencing data in order to:

- ✓ Specify which data formats are currently supported by SRA.
- ✓ Enable submitters to validate and convert data prior submission to avoid unnecessary data transfers.

- ✓ Improve the speed of submission processing.
- ✓ Reduce the probability of failed submissions.
- ✓ Improve other services provided by SRA by freeing up time previously spend to correct and transform data.

The SRA is a "raw data" archive, and requires per-base quality scores for all submitted data. Thus, unlike GenBank and some other NCBI repositories, FASTA and other sequence-only formats are not sufficient for submission. FASTA can, however, be submitted as a reference sequence(s) for BAM file.

The SRA data model has transitioned from "dumps" of whole flowcell lanes or production runs into a semi-curated database of sample-specific sequencing libraries.



- The SRA generally prefers to obtain "container files". Container in this context means an unambiguous binary file. These are objects that contain both the data and a description or specification of the data. Examples include BAM, SFF, and PacBio HDF5 formats. Containers have the following advantages:
- All data for a given library is contained in one file.
- Data are indexed for random access.
- Data are compressed so *gzip* and other compression utilities are discouraged.
- Data are streamable (can be read from one input handle).
- Data are self-identifying (file type can be interrogated with *file*).

Data come with run-time configuration and execution parameters, including run date, instrument name, flow cell name, processing program and version, etc.

Installation *SRA toolkit*

For Window users

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

For Linux users

wget*http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-centos_linux64.tar.gz*

Unpacking the Toolkit

For Window users

use an archiving and compression utility (e.g., WinZip, 7-Zip, etc.), or simply double-click on the .zip file and drag the ‘ sratoolkit ...' folder to the preferred install location.

For Linux users

tar -xzf sratoolkit.current-centos_linux64.tar.gz

~/[user_name]/sra-toolkit/fastq-dump

Root->directory->sra-toolkit folder->fastq-dump folder

For linux

fastq-dump

Open a terminal or command prompt and "cd" into the directory containing the toolkit executables (e.g., [download_location]/sratoolkit[version]/bin/).

Linux/Mac OSX:

./fastq-dump -X 5 -Z SRR390728

Windows:

fastq-dump.exe -X 5 -Z SRR390728

SRA Toolkit to convert data into different format

- fastq-dump: Converts data to FASTq and FASTA format.
- sam-dump: Converts data to SAM (human-readable bam). Data submitted as aligned bam are output as aligned SAM, while other formats are output as unaligned SAM.
- sff-dump: Converts data to SFF format. Note that only data submitted as SFF can be converted back to this format.
- abi-dump: Converts data to csFASTA / csqual format. Note that data submitted in base-space can be represented in color-space.
- illumina-dump: Converts data to Illumina native and qseq formats.
- vdb-dump: Exports the vdb-formatted data of the .sra file.

Module163: Lesson-11-Important Terminologies in NGS analysis-copy

Text (14:00)

Template

A DNA/RNA sequence part of which is sequenced on a sequencing machine or assembled

from raw sequences.

Segment

A contiguous sequence or subsequence.

Read

A raw sequence that comes off a sequencing machine. A read may consist of multiple segments. For sequencing data, reads are indexed by the order in which they are sequenced. Range of a read varies from 90 bp to 180 bp even till 1000 bp depending upon the sequencing machine used.

Chimeric alignment

An alignment of a read that cannot be represented as a linear alignment. A chimeric alignment is represented as a set of linear alignments that do not have large overlaps. Typically, one of the linear alignments in a chimeric alignment is considered the "representative" alignment, and the others are called "supplementary" and are distinguished by the supplementary alignment flag. All the SAM records in a chimeric alignment have the same QNAME and the same values for 0x40 and 0x80 flags the decision regarding which linear alignment is representative is arbitrary.

Read alignment

A linear alignment or a chimeric alignment that is the complete representation of the alignment of the read.

Multiple mapping

The correct placement of a read may be ambiguous, e.g., Due to repeats. In this case, there may be multiple read alignments for the same read. One of these alignments is considered primary. All the other alignments have the secondary alignment flag set in the SAM records that represent them. All the SAM records have the same QNAME and the same values for 0x40 and 0x80 flags. Typically the alignment designated primary is the best alignment, but the decision may be arbitrary