BIF-101

FOR MID TERM (LEC-01-to-95)

Muhammad Faryad

0333-4590795

Student

BS-zoology 2nd Semester

Note=In this handouts Lec 76-85 missing

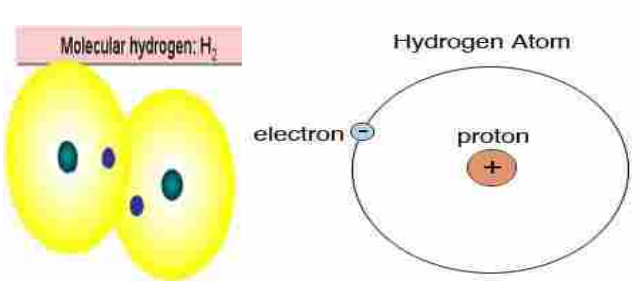
رب زدنی علما

1-Unit of Life

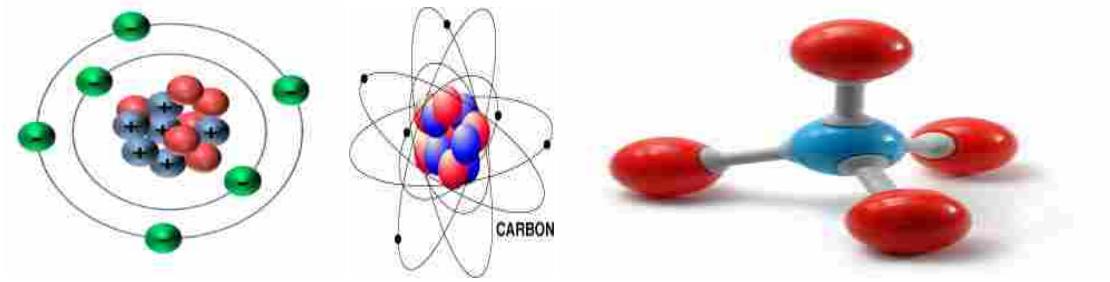
All organisms are composed of cells the basic unit of life and all cells come from preexisting cells

Life=collection of macromolecules that can perform unique functions because the are enclosed in structural compartment that provides consistency (homeostasis).

2-Composition of Matter =

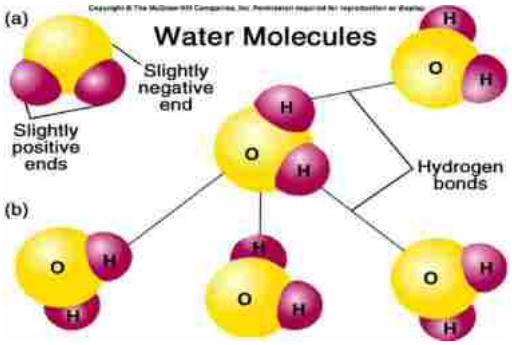


Carbon Atom

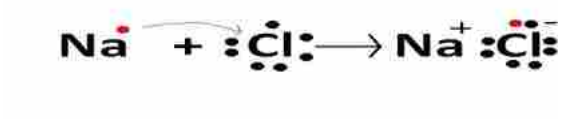


Electronegativity =The tendency of an atom to attract electrons towards itself.

Polarity of H2O

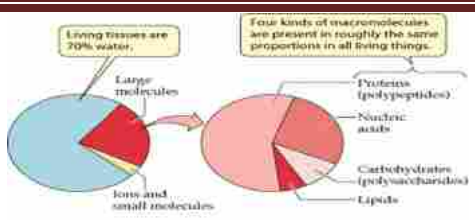


Ionic Bond=



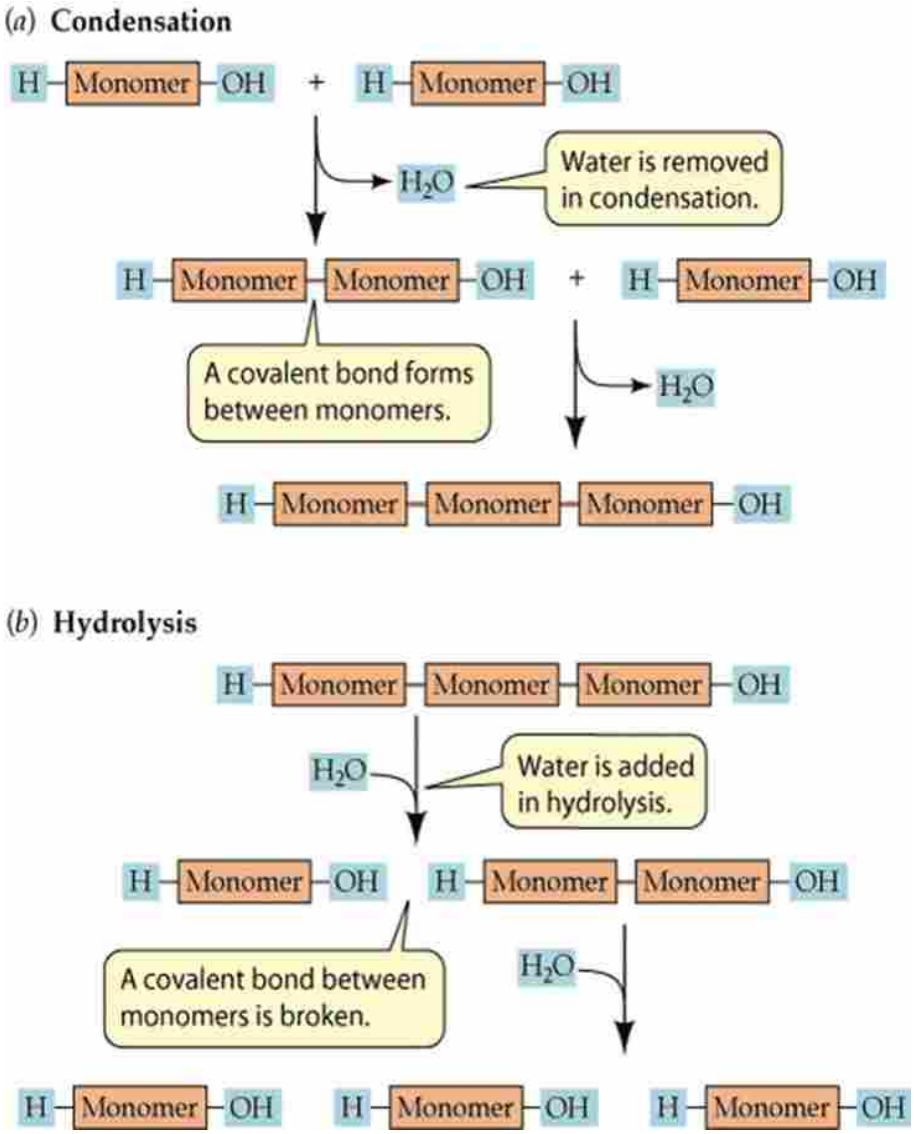
3-Molecules of Life

Composition of Life



Polymers & Monomers=Macromolecules are polymers of smaller monomers molecules
Condensation reactions: Monomers join H_2O Removal
Hydrolysis reactions: Use water to break polymers into monomers.

Condensation & Hydrolysis



4-Journey into the Cell

View animation

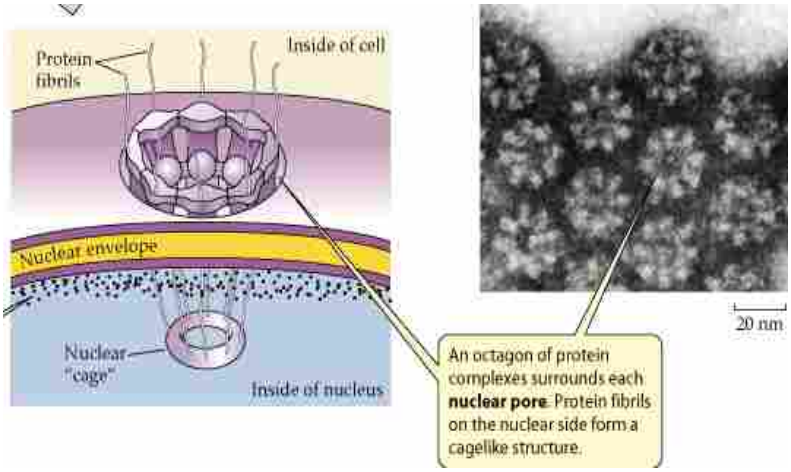
5-Size Matters

Cell Structure=Cells are small to maintain large surface area to volume ratio
Prokaryotic Cells=No membrane enclosed internal compartments.
Plasma membrane regulates traffic (barrier).

Nucleoid region contains DNA.
Most have cell wall.
Special Features=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

6-The Nucleus

Nuclear EMs



Chromatin=Euchromatin , Hetrochromatin , Chromosome

7- Introduction to Molecular Biology

- Molecular Biology
- (BIO-302)
- In the broader terms, definition of molecular biology includes all aspects of the study of life from a molecular perspective.
- More precisely, the term "Molecular Biology" refers to the biology of the molecules related to genes, gene products and heredity.
- In other words, the term molecular biology is often substituted for a more appropriate term, Molecular Genetics. So molecular biology grew out of the disciplines of genetics and biochemistry.
- 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.

8- History of Molecular Biology

- Molecular Biology takes its roots mainly from the the disciplines of Genetics and Biochemistry.

Some important discoveries which led to the emergence of Molecular Biology as a discipline include:-
Biochemistry

- Synthesis of Urea by Friedrich Wohler (1828).
- The first enzyme diastase was discovered by Anselme Payen in 1833.

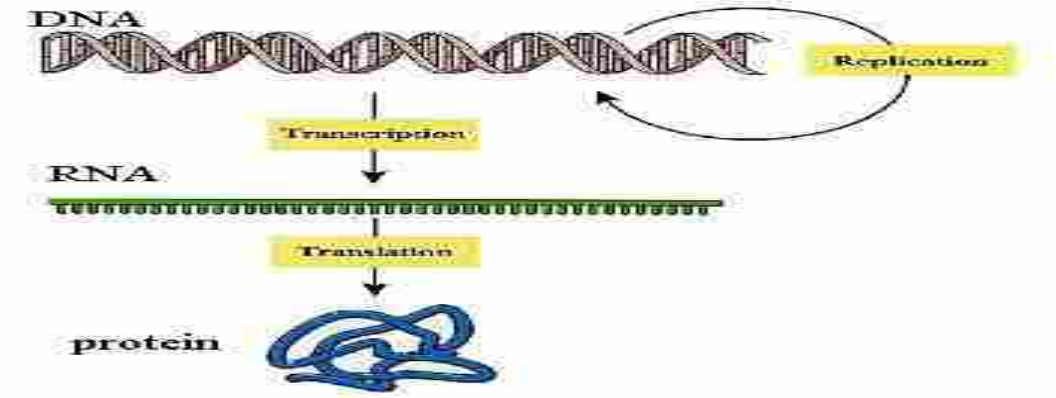
- The Discovery of DNA by Friedrich Miescher in 1869.
- Eduard Buchner discovered cell-free fermentation in1897.
- James B. Sumner crystallized enzyme Urease in 1926.
- James Watson and Francis Crick discovered the double helical structure of DNA in 1953.

Genetics

- Mendel's Experiments on garden pea by Gregor Mendel in 1865.
- Chromosome theory of inheritance by Thomas Hunt Morgan in 1910.
- Barbara McClintock and Harriet Creighton provided physical evidence of recombination in 1931.
- One-gene/one-enzyme hypothesis proposed by George Beadle and Edward Tatum in 1941.
- Oswald Avery and his colleagues demonstrated in 1944 that DNA is the hereditary material.
- It was further confirmed by Alfred Hershey and Martha Chase in 1952
- At this point, the discoveries in the two fields led to the foundation of a new discipline called Molecular Biology.

9- Achievements of Molecular Biology

In 1957, Francis Crick laid out the "Central dogma of moleular biology" which foretold the relationship between DNA, RNA, and proteins



- In 1958, Mathew Meselson & Franklin Stahl proved that DNA replication was semi-conservative.
- Marshall Nirenberg and Gobind Khorana working independently cracked the code in the early 1960s.
- They found that 3 bases constitute a code word, called a codon, that stands for one amino acid.

Gene Cloning

- Since 1970s, scientists have learned to isolate genes, place them in new organisms and reproduce them by a set of techniques collectively known as gene cloning

Genetically Modified

- Organisms (GMOs)
- The technique of Gene cloning led to the creation of a large number of genetically modified organisms with desirable characters

Human Genome

Project

The Human Genome Project (HGP) was launched in 1990 and completed in 2003.

10- Nucleic Acids

- 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.

Nucleic acids

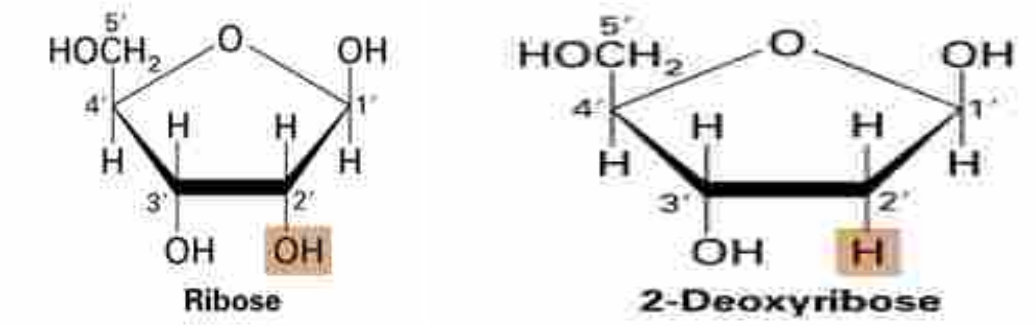
Ribonucleic acids (RNA)

Deoxyribonucleic acids (DNA)

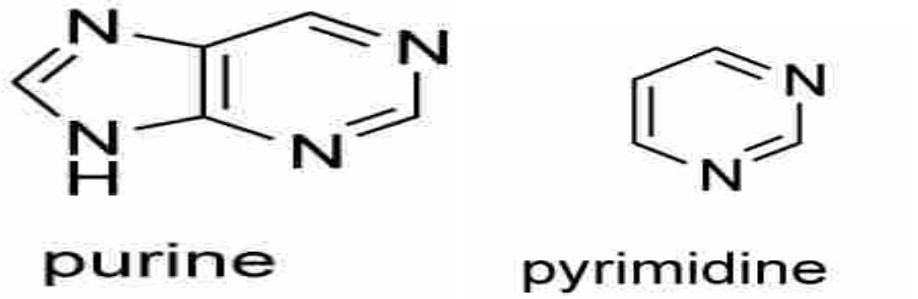
11-Chemical composition of DNA

- DNA is a polymer of Deoxyribonucleotides.
- Deoxyribonucleotide is composed of three components:
  - Deoxyribose
  - Nitrogenous Base
  - Phosphoric acid

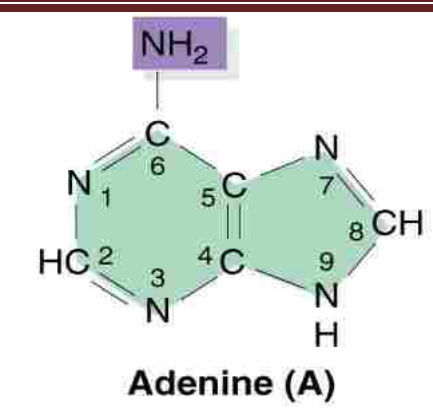
Deoxyribose (a pentose sugar derivative)

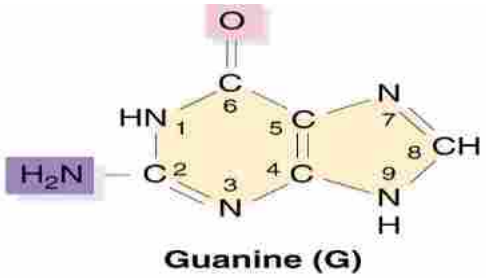


Nitrogenous Bases

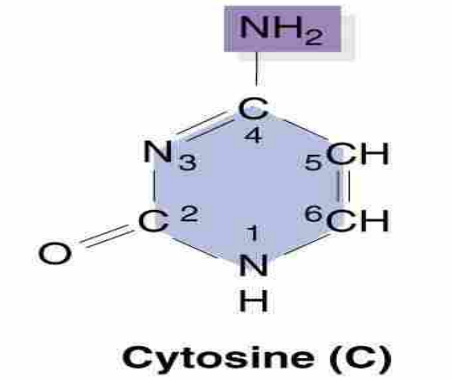


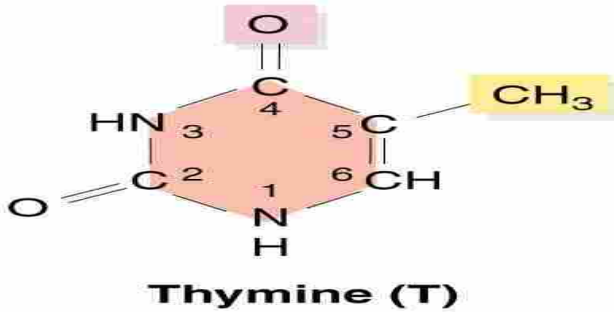
Purines



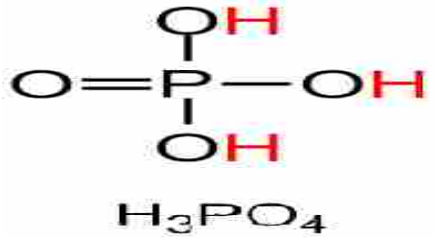


Pyrimidines





Phosphoric acid

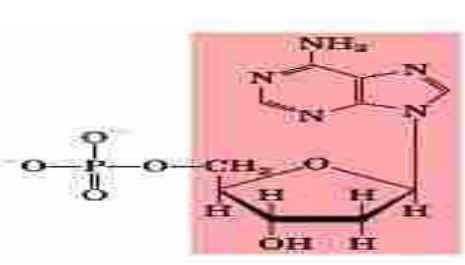


12-Nucleoside & Nucleotide

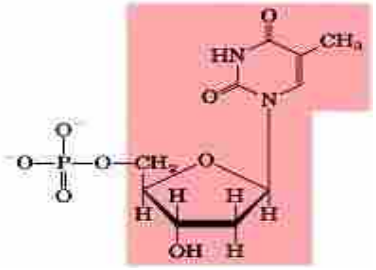
- A molecule containing all these three components is called a nucleotide.
- While a molecule without the phosphate group is called a nucleoside.
- Nucleotide = Nucleoside + Phophoric acid

Nucleoside = Nucleotide – Phoshoric acid

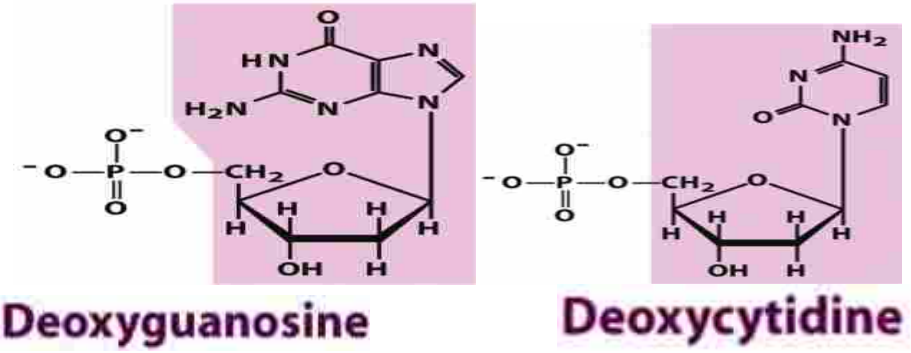
13-Types of Deoxyribonucleotides



Deoxyadenylate



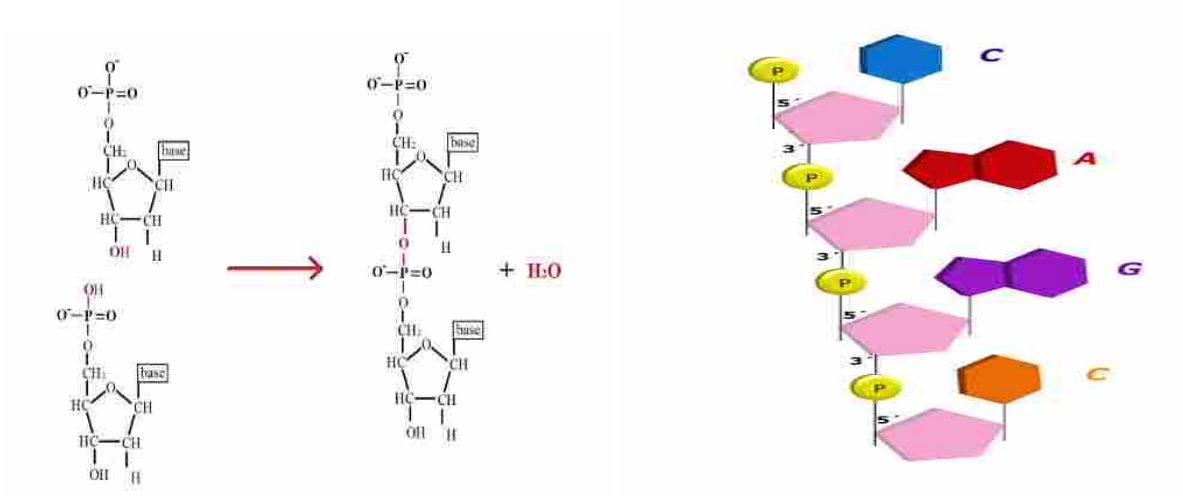
Deoxythymidylate



- These four deoxy-ribonucleotides make the structural units of DNA.

14-How do Deoxyribonucleotides Join?

- The successive nucleotides of DNA are joined together through phospho-diester linkages.



15-Structure of DNA

Work of Chargaff (Late 1940s)

- The discovery of the structure of DNA is one of the greatest events in the history of science.
- Erwin Chargaff and his colleagues provided a most important clue to the structure of DNA.
- The work of Chargaff led him to following conclusions, also called “Chargaff Rules”:-
- Base composition of DNA varies from one species to another.
- 2. The DNA isolated from different tissues of the same species have the same base composition.
- 3. The base composition of DNA in a given species does not change with an organism’s age, nutritional state, or changing environment.
- 4. In DNA, the number of adenosine residues is equal to the number of thymidine (A=T) and the number of guanosine residues is equal to the number of cytidine (G=C).
- It means that the sum of the purine residues equals the sum of the pyrimidine residues (AG=TC).

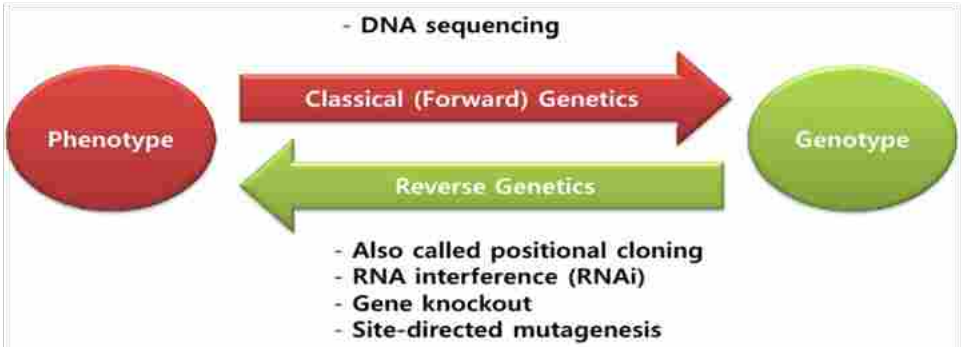
16-What is genetics

Definition

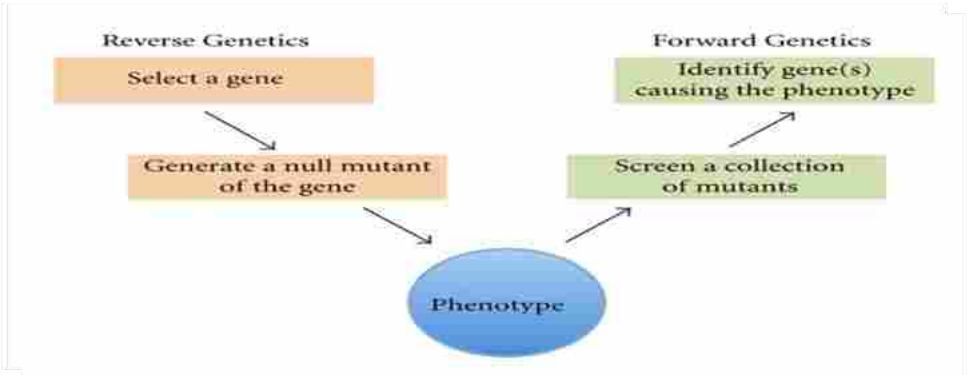
- Study of genes, heredity, variation.
- A field of Biology.
- The principles of heredity.
- Mendel unaware chromosomes, gene.

What is Genetics

GENETICS



What is Genetics



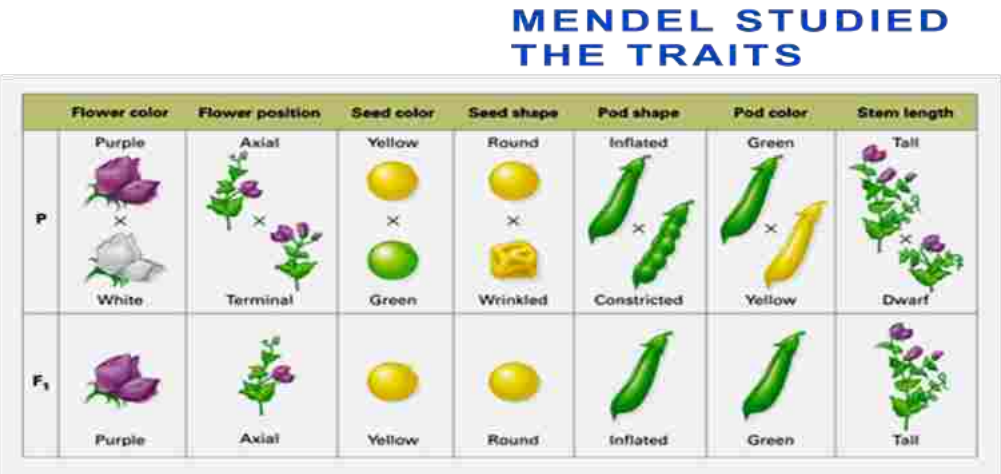
Garden Pea

- Seed in a variety of shapes and colors.
- Self, cross pollinate.
- Takes up little space.
- Short generation time
- Produces many offspring.

Mendel was fortunate

- Peas - in many varieties.
- Strict control over which plants mated.
- The pea traits are distinct, contrasting.

What is Genetics



Genetics

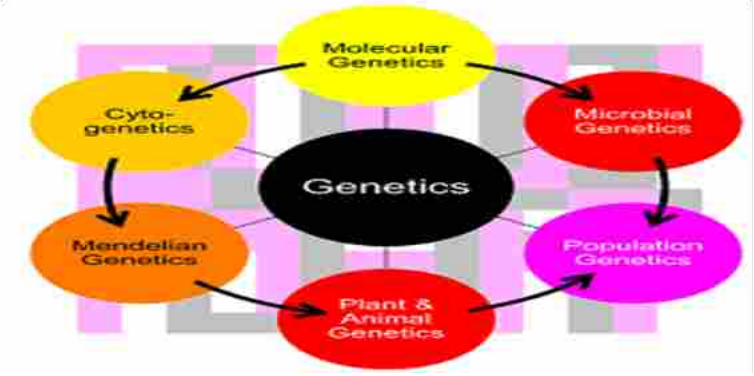
- Study of genes, chromosomes and heredity is called as Genetics

17-Sub disciplines of Genetics

four sub disciplines

- Transmission (Classical) Genetics
- Population Genetics
- Quantitative Genetics
- Molecular Genetics

Sub Disciplines of Genetics



developments in genetics

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

Transmission or classical genetics

- Deals with movement of genes and genetic traits from parents to offspring. Mendel s' Laws
- Deals with genetic recombinations.

population genetics

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

Quantitative genetics

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

Molecular genetics

- Deals with the molecular structure and function of genes.

18-Genetics terminologies

Genetics terminologies

- Character: heritable feature (skin color, height etc).
- Trait: variant for a character (i.e. brown, black, white etc).
- True-breed: all offspring of same variety.

different generations of a cross

- P generation (parents).
- F1 generation (1st filial generation).
- F2 generation (2nd filial generation).

Pure Cross and Hybrid Cross

- Pure Cross

True breeding X True breeding

WW X ww

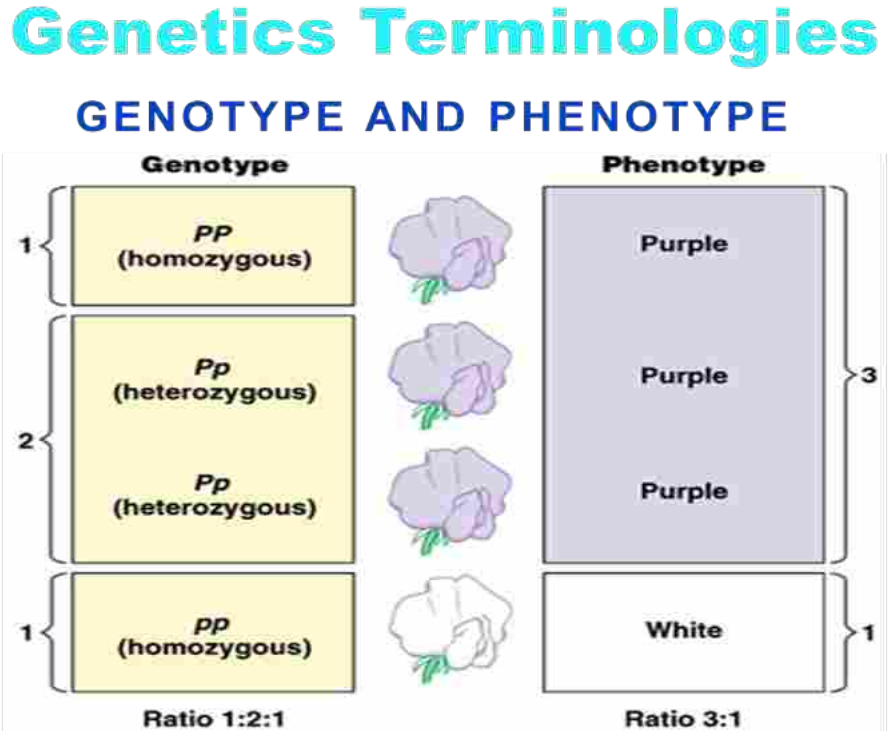
- Hybrid Cross

F1 generation X F1 generation

Ww X Ww

Genotype and Phenotype

- Genetic make-up of an organism is called Genotype.
- Physical appearance of an organism is called Phenotype.



Dominant and Recessive

- Dominant: when one characteristic expresses itself over the other i.e. round over wrinkled.
- Recessive: The trait that does not show through in the first generation i.e. wrinkled.

19-Genome Informatics

Introduction

- Genome sequencing provides the sequences of all the genes of an organism
- A major application of Bioinformatics is analysis of full genomes that have been sequenced
- Challenge is to identify those genes that are predicted to have a particular biological function

Genomics

- Study of all of a person's genes (the Genome), including interactions of those genes with each other and with the person's environment (NHGRI)

Genome Informatics

- Genome informatics is the field in which computational and statistical techniques are applied to derive biological information from genome sequences
- Genome informatics includes methods to analyze DNA sequence information and to predict protein sequence and structure

Genome Sequences

Availability of genome sequences facilitates;

- The discovery and utilization of sequence polymorphisms
- Opportunity to explore genetic variability both between and within the organisms

Genome Analysis

Following tasks;

- Sequencing
- Assembly
- Repeat identification and masking out
- Gene prediction
- Looking for EST and cDNA sequences
- Genome annotation
- Expression analysis
- Metabolic pathways and regulation studies
- Functional genomics
- Gene location/gene map identification
- Comparative genomics
- Identify clusters of functionally related genes
- Evolutionary modeling
- Self-comparison of proteome

Model organisms

- E. coli – bacteria
- S. cerevisiae – yeast
- C. elegans – worm
- D.melanogaster – fly
- Danio rerio – zebrafish
- Mus musculus - mouse
- Homo sapiens – you and me

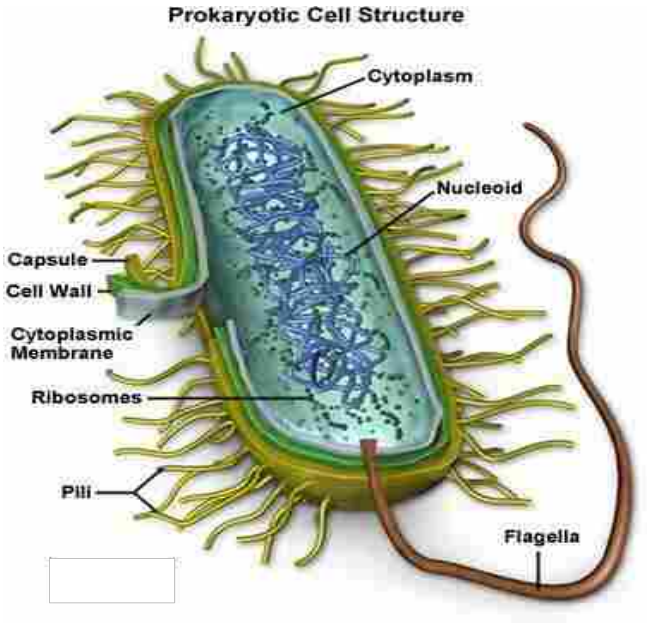
Arabidopsis – plant

Conclusions

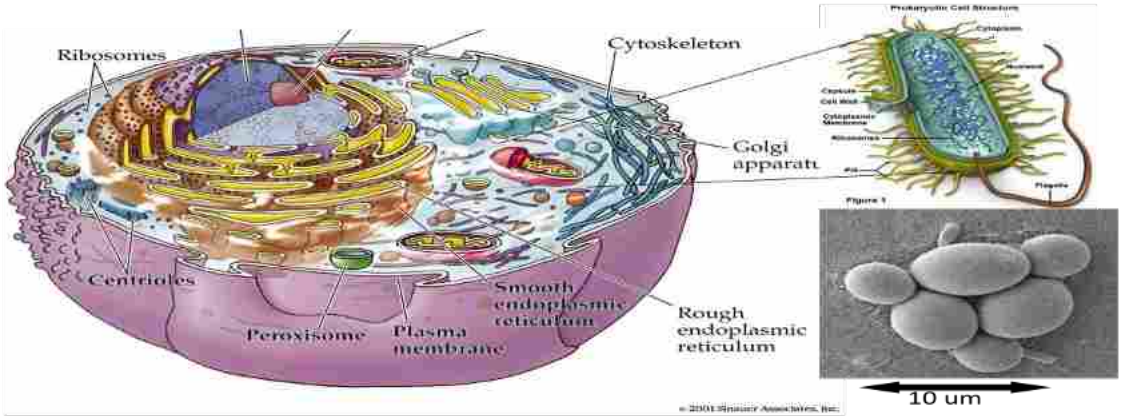
- Sequencing and analysis of full genomes paves the way for future discoveries
- Different model organisms can help explore our Genome and what maters most for us

20-Prokaryotic Genome

- Prokaryotes are the organisms whose Genetic material (DNA) is not enclosed in a nuclear membrane
- No membrane bound organelles



Prokaryotic Genome



Mitochondria evolved from a bacterial endosymbiont

- First prokaryotic Genome sequenced was that of *Hemophilus influenzae*
- Paved the way for sequencing of many other organisms

Selection criteria

Following are the criteria for organisms selected for sequencing;

- They had been subjected to a detailed biological analysis and thus were model organisms
- important human pathogens
- They were of phylogenetic interest
- Sequences were annotated as they were sequenced

Conclusions

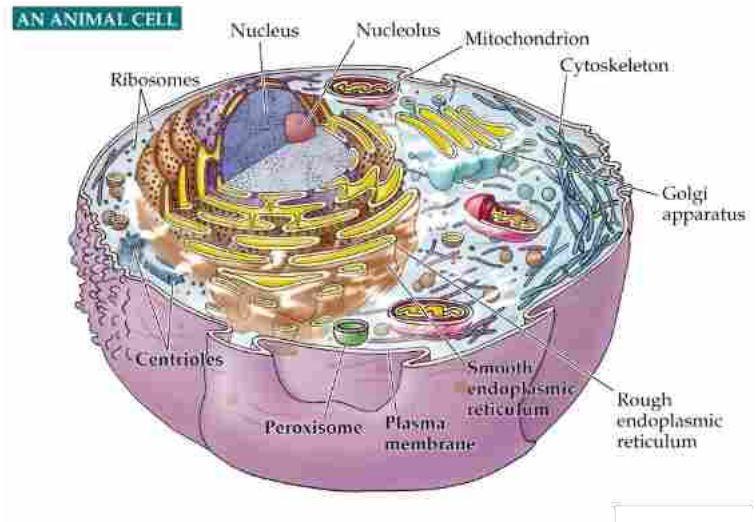
- Prokaryotes are simple Genomes
- Easy models to study Biochemistry and Molecular biology of life processes
- Sequencing is done on economically important organisms

21-Eukaryotic Genome

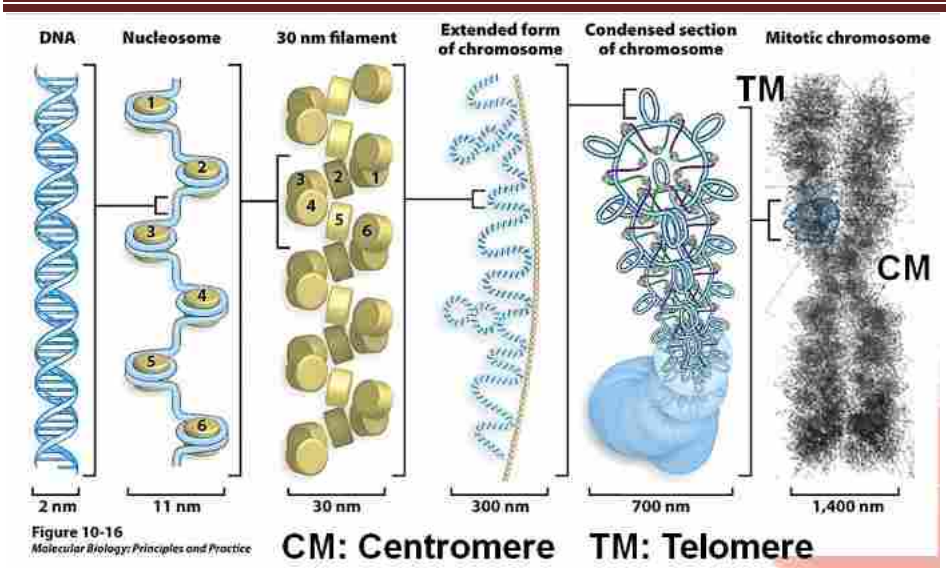
Introduction:

We have seen that prokaryotes are simple genomes in comparison to eukaryotes which are relatively complicated. So distinct properties of eukaryotes are mentioned below:

- Eukaryotes have larger genomes
- Have tandem repeats
- Have introns in their protein-coding genes (introns are between the exons which are the protein coding regions within the genes).
- Heterochromatin and euchromatin region (eukaryotes have complicated genome, so the chromosome is grouped as heterochromatin; densely packed region and euchromatin; lightly packed region).



Here, in this diagram we can see a typical eukaryotic cell which is pretty stuffed as compare to prokaryotic one. We have nucleus in the middle, channels coming out of the nucleus known as *endoplasmic reticulum* (helps in transportation), ribosomes (for protein synthesis), mitochondria (energy synthesis), we can also see the cytoskeleton that makes the structure of this cell intact and Golgi apparatus (are concerned with the secretion). So complicated membrane bounded organelles are present in the eukaryotes



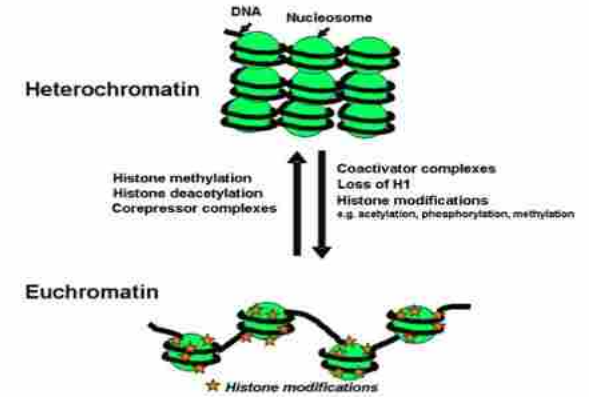
Here, in this diagram we see the connection between the DNA and the chromosomes.

On the left-hand side, we see a DNA strand that is a 2nm wide strand. So, the DNA wraps over the protein complex molecules (histones - labelled as 1, 2, 3...), and this structure is known as *nucleosome*. Then these histones, turn around and makes a wider structure and it makes a 3nm filament (in third section). Then these nucleosome structures supercoil on their selves to make those further bigger fibres and until they reach the chromosome the width is 1,400 nm. So, if we look into the chromosome, we can recognize that there are different arms in it, which are known as sister *chromatids* (remember this is just one chromosome but we have two chromatids), somewhere in the middle we see a constricted part known as *centromere,* whereas the terminals are known as *telomeres,* (remember these nomenclature while we are discussing the heterochromatin and euchromatin parts).

Staining with dyes:
So chromosomes if stained with the dyes, they give different coloring patterns, we can come up with the following:

- Dense heterochromatin (dense regions obviously take more color)
- Light Euchromatin (light regions take less color)

If we look into the gene expression, the *heterochromatic regions* are packed so the enzymatic machinery cannot reach there; hence these regions are poorly transcribed (expressed).
Whereas, the *euchromatic regions* are highly expressed because they are loosely packed and the enzymatic machinery can easily reach to them.



Here, is the diagram in which we can see the relationship between heterochromatin and euchromatin. You can look into these nucleosomes (combination of DNA and histones), which are quite jammed packed with one another, so definetly the enzymes cannot access the DNA which is embedded inbetween. There are different modifications on the DNA or histones that bring about those structures (we can see on the top), so for example there are histone methylations (in which methyl groups are added to those histones) in that case the system moves towards downside (as shown in the diagram) so it becomes *euchromatin* and

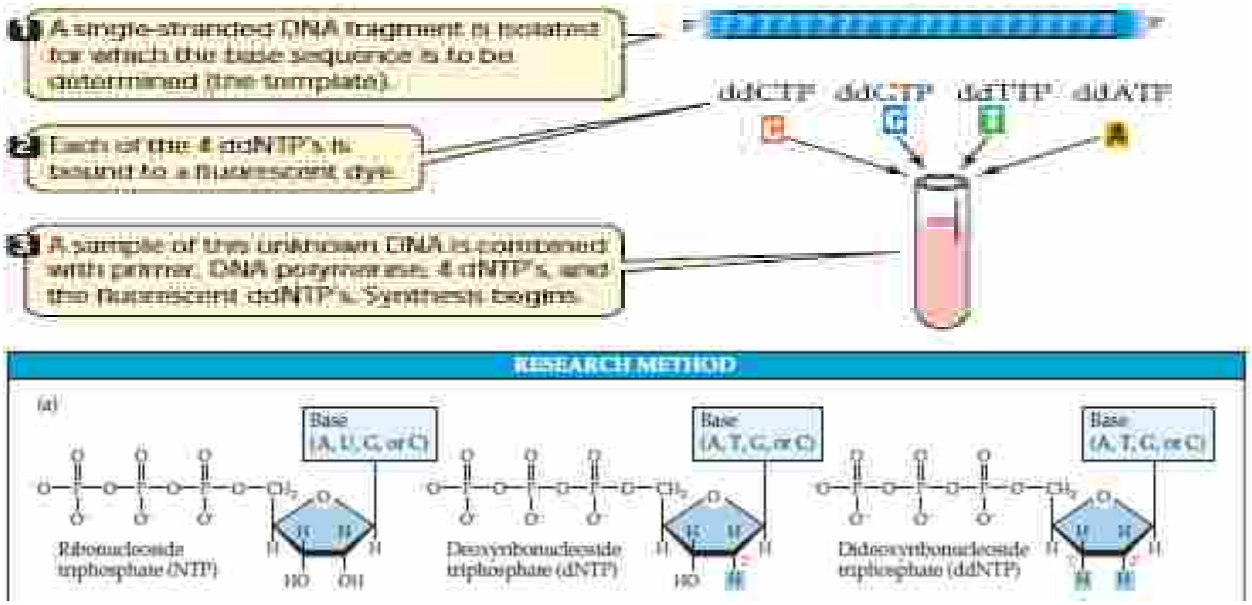
similarly, we see that there are some other methylations on some other aminoacids that can move back into the opposite direction also. So, histone methylation, histone deacetylation and there are some other complex proteins which gets attached and give us this *heterochromatin region* and in the reverse process, we get the *euchromatin region*. In the Euchromatin region, the histones are quite spaced and DNA can be acessible. So, this is the reason why the euchromatin region is expressed more as compared to the heterochromatin region.

Conclusions:

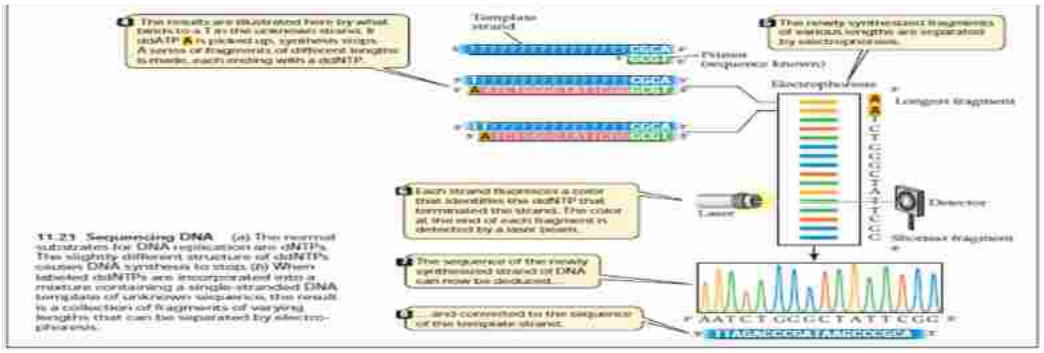
We conclude that:

- Eukaryotes are distinguished by the presence of prominent nuclei
- Eukaryotes have larger genomes, tandem repeats and introns in their protein-coding genes (i.e. they are complicated).

22- DNA Sequencing



DNA Sequencing



END

23- Genetic Information is Converted into Protein

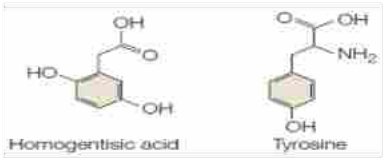
Archibald Garrod 1908

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.

24- The Dogma

The Dogma

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

25- Background of Bioinformatics

Modern Biology

- Humans have over 25,000 genes
- The genes may produce > 250,000 different proteins
- Each protein may further take up multiple modifications
- Humans have over 25,000 genes
- The genes may produce > 250,000 different proteins
- Each protein may further take up multiple modifications
- But then, there are numerous organelles & intermediary molecules as well

The scale and speed of biomolecular interactions, reactions and transport is simply bewildering
Experiments in Biology

Brisk enhancements in instrumentation:

- Next Generation Sequencers
- High Resolution Mass Spectrometers
- Nuclear Magnetic Resonance Spectroscopy

Digitalization of Biology

- Modern biological experiments produce data which is stored on computers
- The data includes text, numbers, symbols and images

Speed of Data Growth

- Data is being accumulated at an exponentially increasing rate
- E.g. Genome sequences in genome databases are doubling every few years!

Conclusion

- Humans are limited in recalling information from their memory
- So, we need the help of computers to store and recall this data!
- Computers can also quickly process this information and help in its manipulation.

26- Introduction to Bioinformatics

Motivation

1. An interdisciplinary field
2. New but rapidly developing field
3. Low requirement on infrastructure and research equipment
4. Vast opportunities for scientific discovery

Scope of Bioinformatics
Use of computational algorithms and techniques for:

1. Storage,

2. Organization,
3. Analysis, and
4. Representation of biological information

Activities

1. Developing algorithms
2. Writing software
3. Data processing pipelines
4. Statistical evaluation of data
5. Data visualization

Moving ahead…

- Learn the available bioinformatics concepts, algorithms and tools
- Advance these concepts by developing novel ideas
- Apply these ideas to forward the understanding of biology and life

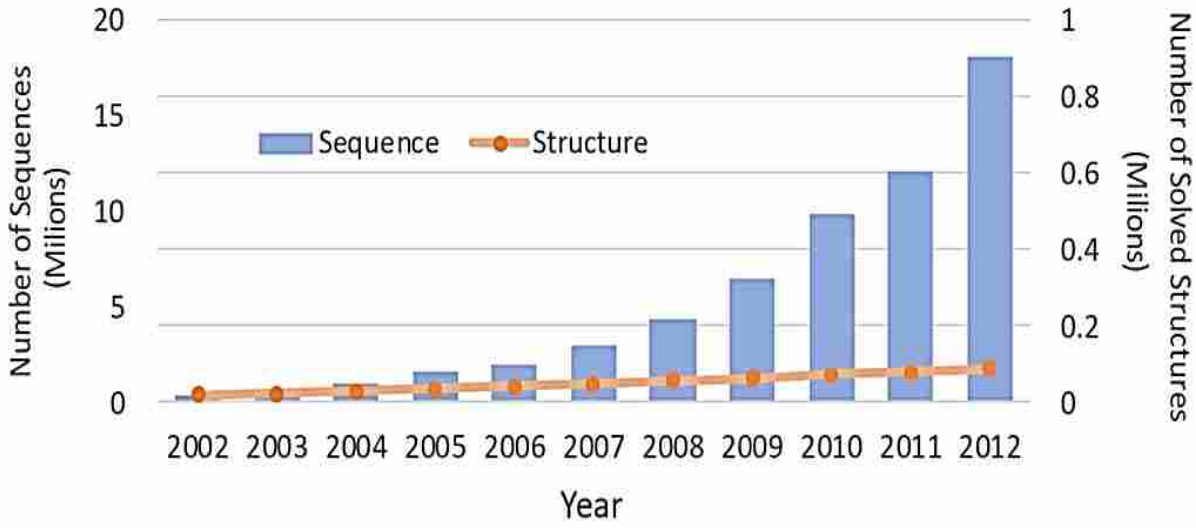
Conclusion

- The benefits of studying bioinformatics in Pakistan
- New degree programs in several universities
- Opportunity for cutting edge discoveries in a resource-limited setting

27- Need for Bioinformatics – I

Motivation

1. An interdisciplinary field
2. New but rapidly developing field
3. Low requirement on infrastructure and research equipment
4. Vast opportunities for scientific discovery



Conclusion

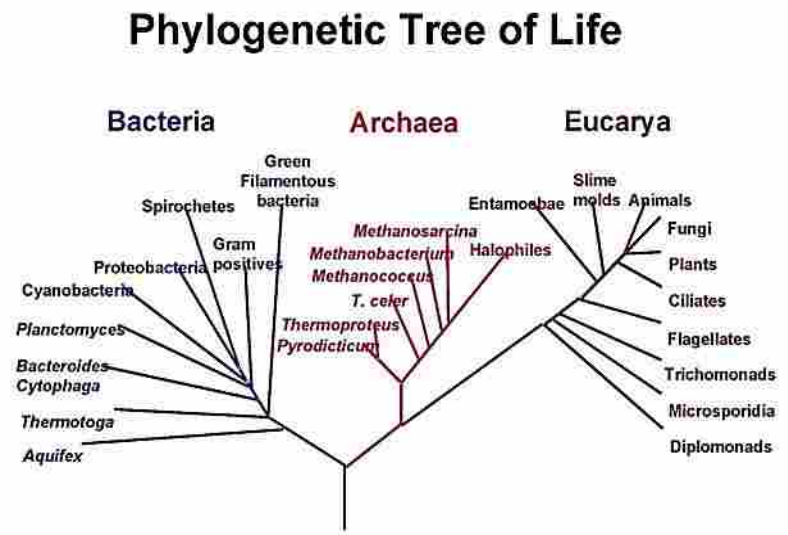
- Since genome and proteome information is publically available, you can process this information after downloading it
- If you carefully search this online data, with a little bit of biology background, it is a gold mine!

28- Need for Bioinformatics - II

Quote

Biology easily has 500 years of exciting problems to work on.

Donald Knuth, Professor, Stanford University



What can bioinformatics deliver?

- Life, evolution and disease can be better understood
- Drugs can be better developed after understanding molecular basis of life

Possible contributions

- Solve important problems related to life
- Handle mammoth quantities of experimental data
- Develop efficient solutions towards these goals

Conclusion

- From gene sequences to protein sequences Bioinformatics is the way forward
- Protein structure, protein-protein interactions and systems biology are other research areas in bioinformatics

29- Applications of Bioinformatics - I

Introduction

- Where can bioinformatics be applied specifically?
- What are the avenues in biology that can benefit from this science?
- What benefits can it deliver to the society?

1. Genomics
   - DNA Sequencing
   - Gene Finding
   - Genome Assembly
   - Variation in Genomes
   - Transcription Data
   - Databases
2. Evolutionary Studies
   - Evolutionary relationships
   - Evolutionary distances

- Phylogenetics
- Tree of life

3. Proteomics

- Protein Sequencing
- Protein Structures
- Post-translational Modifications
- Protein-Protein Interactions
- Database Development

4. Systems Biology

- Model protein and gene interactions
- Dynamical analysis of such models
- Understand system properties
- Predict system level behaviors

Conclusion

- Genomics
- Transcriptomics
- Proteomics
- Metabolomics
- Structural Proteomics
- Drug Design
- Systems Biology
- Personalized Medicine

30- Applications of Bioinformatics - II

Introduction

- Genomics
- Transcriptomics
- Proteomics
- Metabolomics
- Structural Proteomics
- Drug Design
- Systems Biology
- Personalized Medicine
- Bioinformatics techniques have enabled us to generate 'omic' data and its utilization!
- Step by step application of bioinformatics from genome level to entire biosystems

Small to Big

- Gene finding, function prediction
- Structure prediction, modeling
- Predicting Molecular (Protein) interaction
- Biological network building & modeling
- Cell level modeling & simulation
  Cell signaling modelling & simulation
- Tissue & morphology modelling

- Biological network building & modeling
- Models integrate biological data
- Simulations help validate testable hypotheses
- Moreover, these simulations can also predict disease progression & outcomes

Conclusion

- Bioinformatics not only organizes, stores and analyzes biological data, but can also validate novel hypotheses
- Modern day bioinformatics also helps predict disease outcomes as well as drugs to treat them!

31- Frontiers in Bioinformatics - I

Introduction

- Bioinformatics is a relative young scientific area
- It is rapidly expanding to cater for the variety and scale of biological data
- Several unresolved challenges exist in this expansion!

Frontiers in Genomics

- Next generation sequencing of whole genomes
- Massive amounts of data (Tera byte files)
- Handling & analysis of this data
- Genome assembly & prediction

Frontiers in Transcriptomics

- Identification of hitherto unknown types as well as roles of RNA
- Understanding the regulatory dynamics of these RNA molecules
- Interactions of the RNA molecules

Frontiers in Proteomics

- Identification of low abundance proteins in patient tissue samples
- Large scale protein identification
- Identification of known as well as novel post-translational modifications

Conclusion

- Bioinformatics is full of challenges and opportunities
- Amongst other frontiers in bioinformatics, there are protein structures, systems biology and personalized medicine!

32- Frontiers in Bioinformatics - II

Introduction

Frontiers in bioinformatics included:

- Next generation genomics
- Transcriptomics
- High throughput proteomics

Frontiers in Protein Structure

- Accurate solution of protein structures is one of the toughest problems
- How to determine protein structures?
- How to predict protein structures?
- How to simulate the protein folding process?
- How to predict the protein – protein interaction?

How to model protein-drug interaction

Frontiers in Systems Biology

- Cells function as a whole!
- How to integrate genes, proteins and metabolites into a unified system?
- How to simulate these models in real-time?

Frontiers in Personalized Medicine

- Not all medicines work all the time!
- Some medicines have side effects on certain patients
- How to evaluate drugs using Bioinformatics tools?

Conclusion

- 21st century is the century of Bioinformatics
- What will it hold for us, we can already guess that!
- An era of personalized medicines and eradication of several common diseases!

33- 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.
- 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.

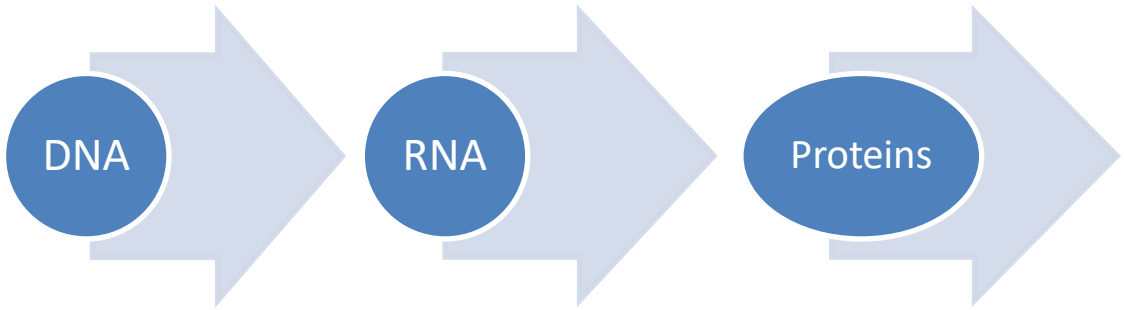
34- Gene, mRNA and Protein Sequences

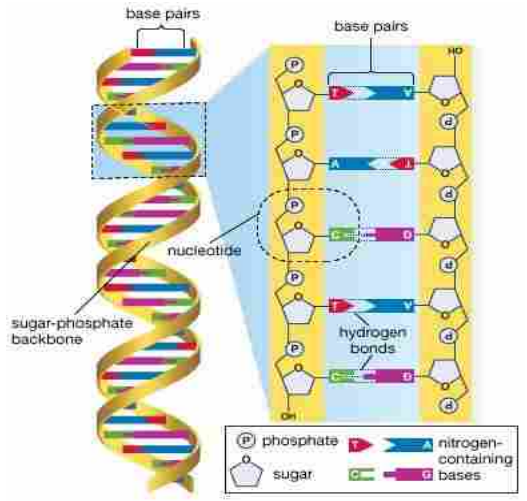
Introduction

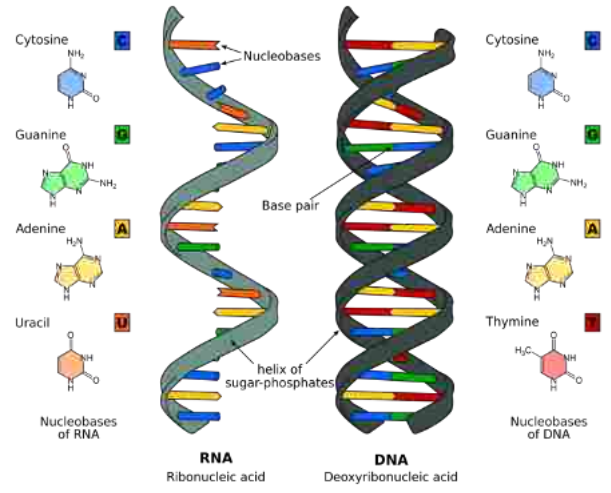
- All living things are composed of cells
- How are cells made?
- DNA provides the blueprint for building cells
- DNA dictates the production of cell's proteins, carbohydrates & vitamins

- The transformation of information in DNA into these molecules is termed the Central Dogma
- Central dogma:
  DNA -> RNA -> Proteins

- Proteins are then used in constructing cells







Conclusion

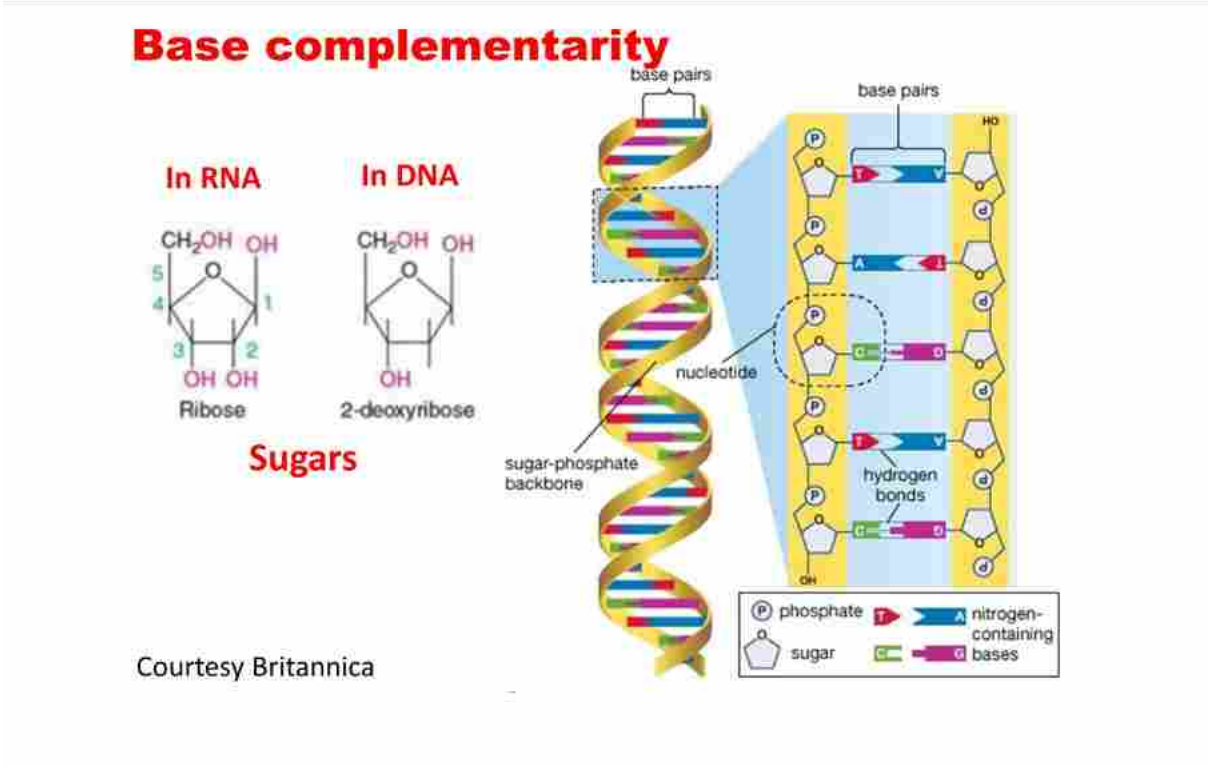
- DNA codes for an RNA
- RNA in turn helps produce Proteins
- Proteins along with some other molecules form cells including their organelles and membrane

35- Nucleotides

Background

- DNA and RNA molecules are built up from four types of molecules, each
- These molecules are called nucleotides
- These are Adenine (A), Cytosine (C), Thymine (T), Uracil (U) and Guanine (G)

Nucleotides



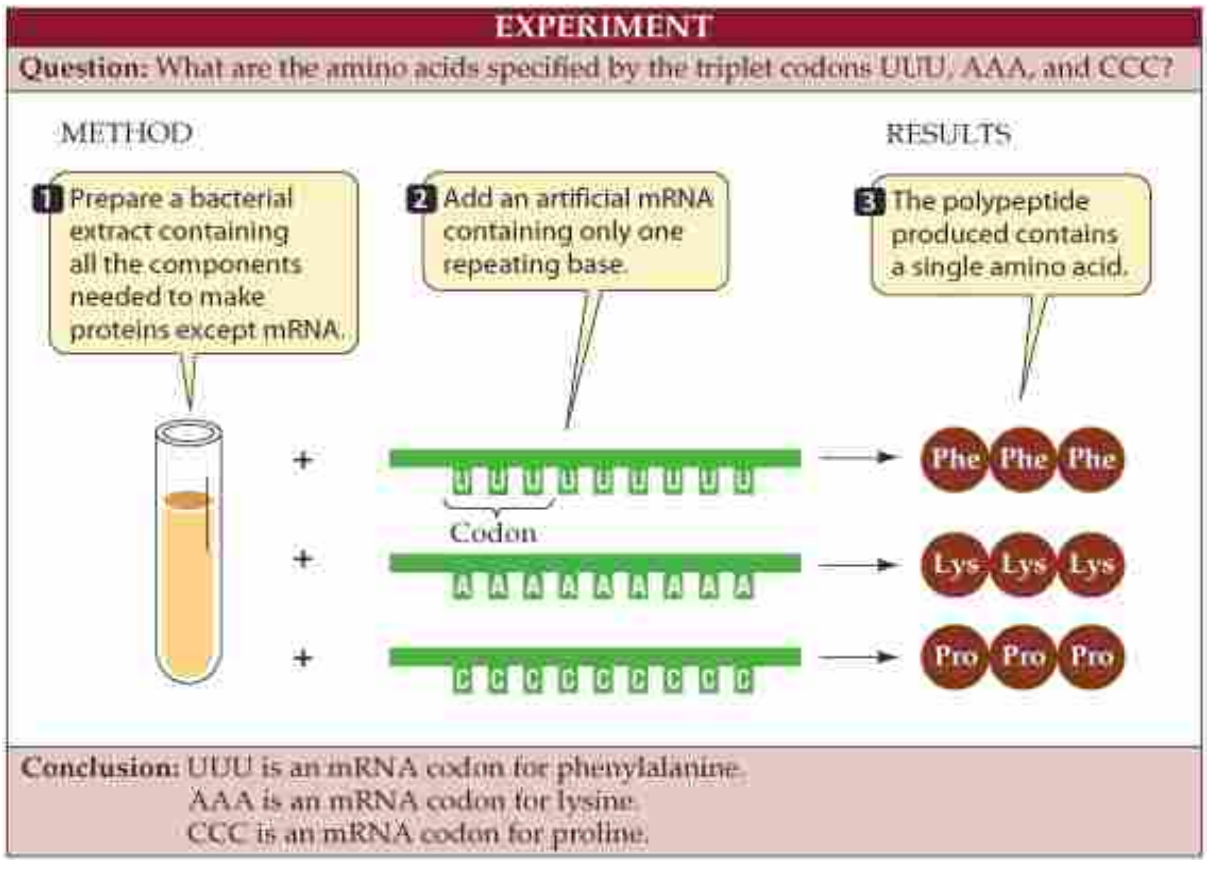
Conclusion

- DNA gives rise to RNA
- DNA has A, T, C & G while RNA has A, U, G & C
- DNA is double stranded while RNA is single stranded
- RNA helps produce proteins

36- Genetic Code

Three mRNA nucleotides form a codon
which is complementary & antiparallel to corresponding DNA.
Genetic code relates codons to their specific AA.
~ Universal & Redundant but not ambiguous some exceptions.
Start & Stop

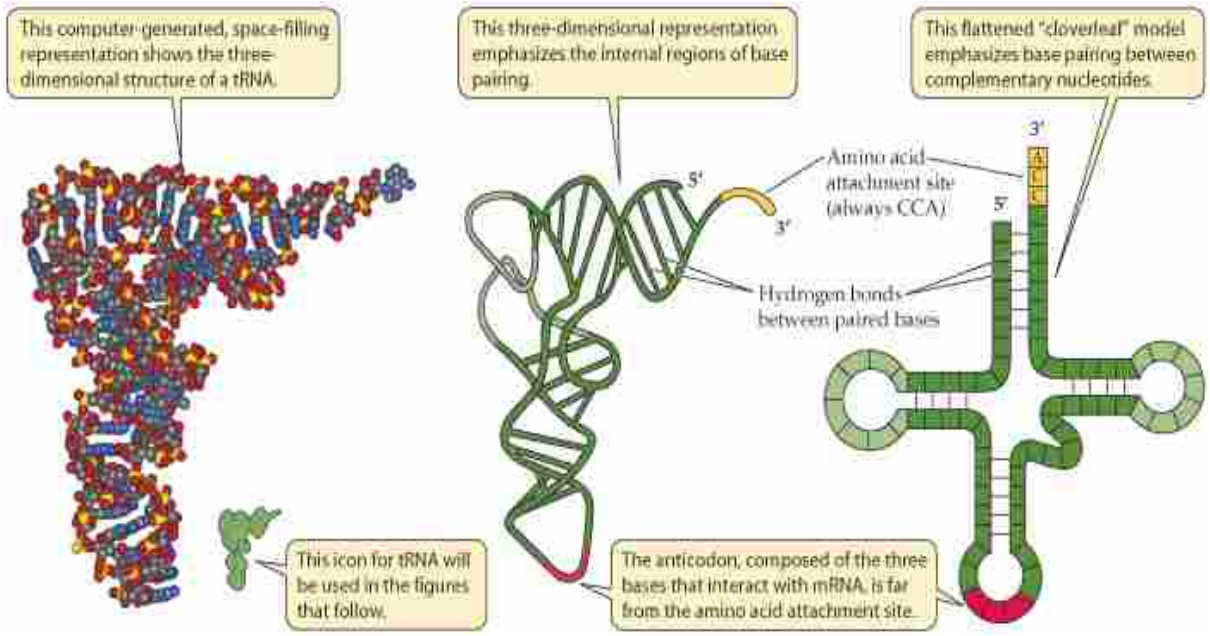
First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC: Phenylalanine UUA, UUG: Leucine	UCU, UCC, UCA, UCG: Serine	UAU, UAC: Tyrosine UAA: Stop codon UAG: Stop codon	UGU, UGC: Cysteine UGA: Stop codon UGG: Tryptophan	U C A G
C	CUU, CUC, CUA, CUG: Leucine	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



tRNA Translates Nucleotide Language into AA Language

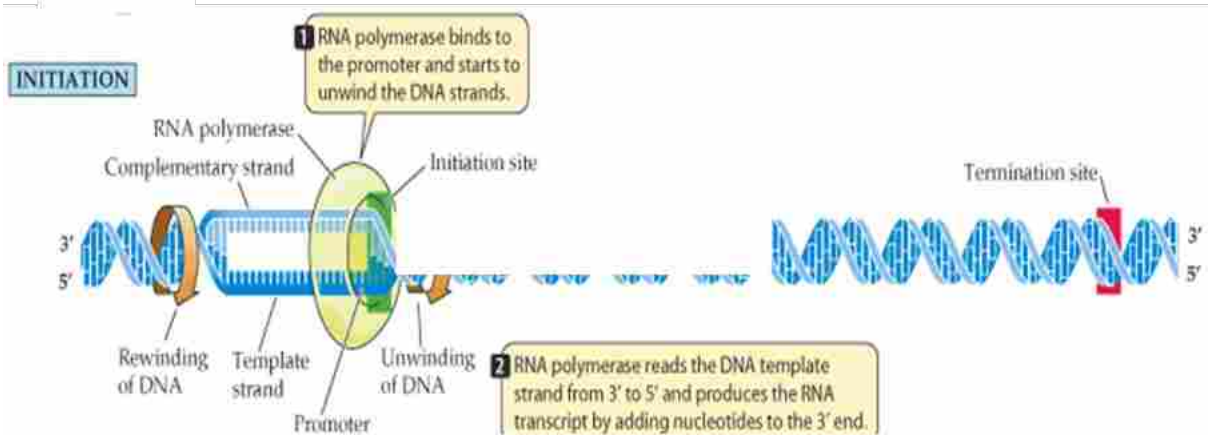
1) Carries AA
2) Associates with mRNA
3) Binds ribosome

tRNA has 75-80 nucleotides with internal H-bonding.
Conformation of tRNA allows it to bind ribosome.
3' end of tRNA (CCA) is attached to AA & mid point has anticodon (antiparallel).



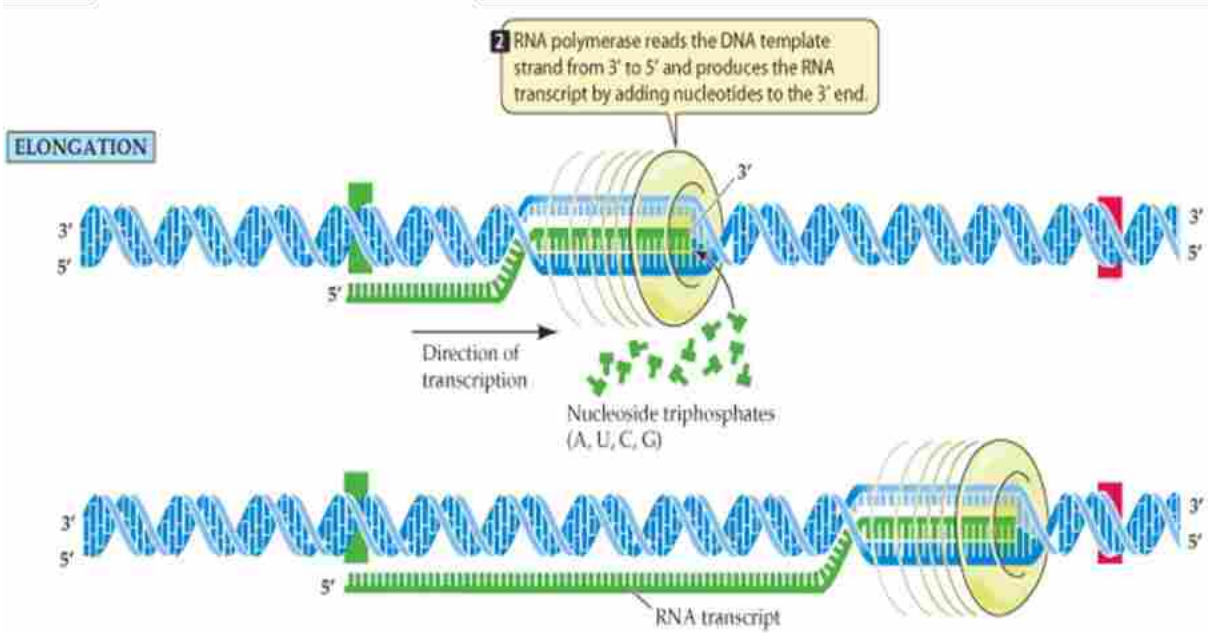
37- Transcription-I

Initiation begins at promoter -tightly binds RNA Pol

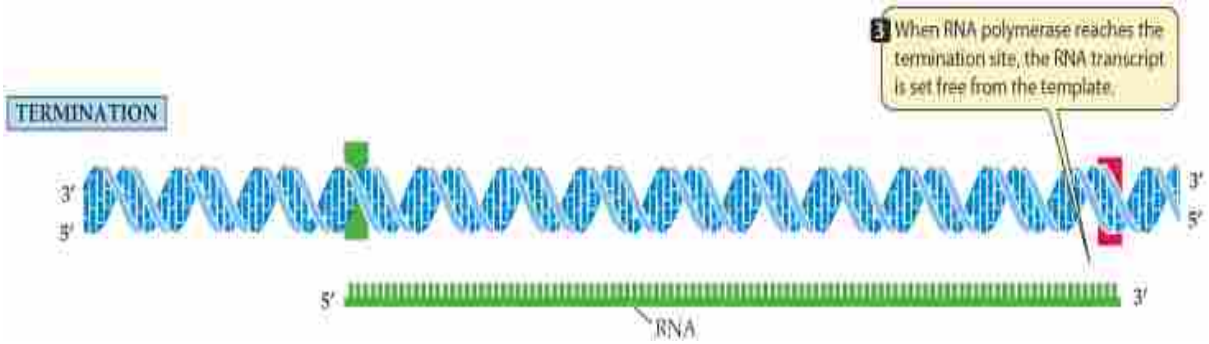


1) Where to start transcription.
2) Which strand to read.
3) The direction to take start from.

Part of promoter is initiation site (transcription start).
RNA Pol moves 3' to 5', synthesis 5' to 3', primer
Unwinds 20 bp at a time.



Transcript antiparallel to DNA template strand.
No proof reading error rate 1 in 10,000-100,000
Translation starts before termination in prokaryotes.
Pre-mRNA 1st product of transcription in eukaryotes.



38- Transcription -II

Background

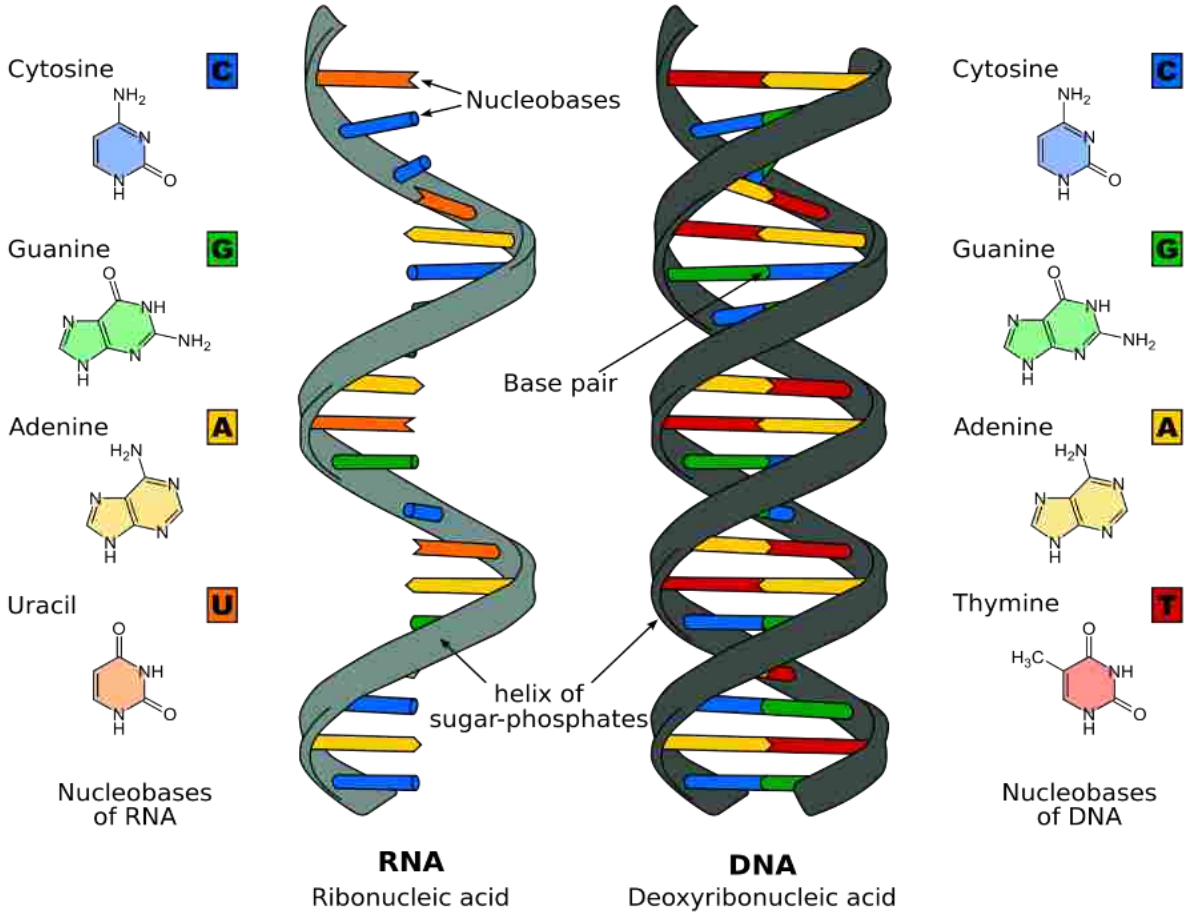
- Cells are built up of proteins and carbohydrates
- These molecules are produced after transformation of DNA into proteins
- What are the salient steps underlining this transformation?

Transcription

- DNA is constituted by four bases
- These are Adenine (A), Cytosine (C), Thymine (T), and Guanine (G)
- DNA molecule may contain thousands of A, C, T and G.
- A pairs with T only and C pairs with G

Transcription

- Process of transcription decodes these bases in the DNA
- Based on the information received after decoding DNA, an RNA molecule is encoded



Transcription

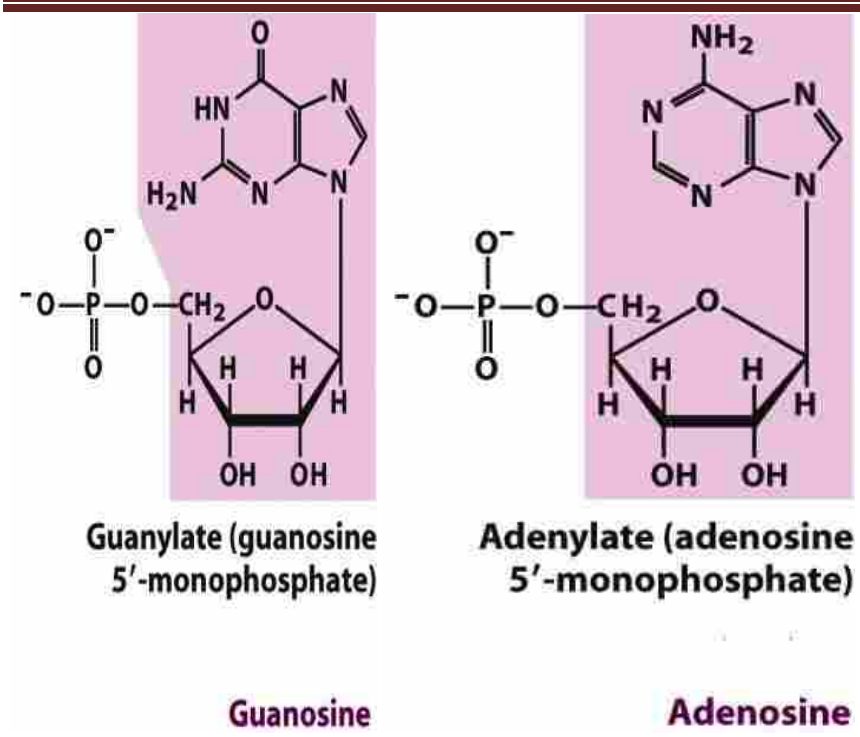
- RNA is constituted by four bases
- These are Adenine (A), Cytosine (C), Uracil (U), and Guanine (G)
- RNA molecule may contain thousands of A, C, U and G.
- A pairs with U only and C pairs with G

Conclusion

- DNA gives rise to RNA
- DNA has A, T, C & G while RNA has A, U, G & C
- DNA is double stranded while RNA is single stranded
- RNA helps produce proteins

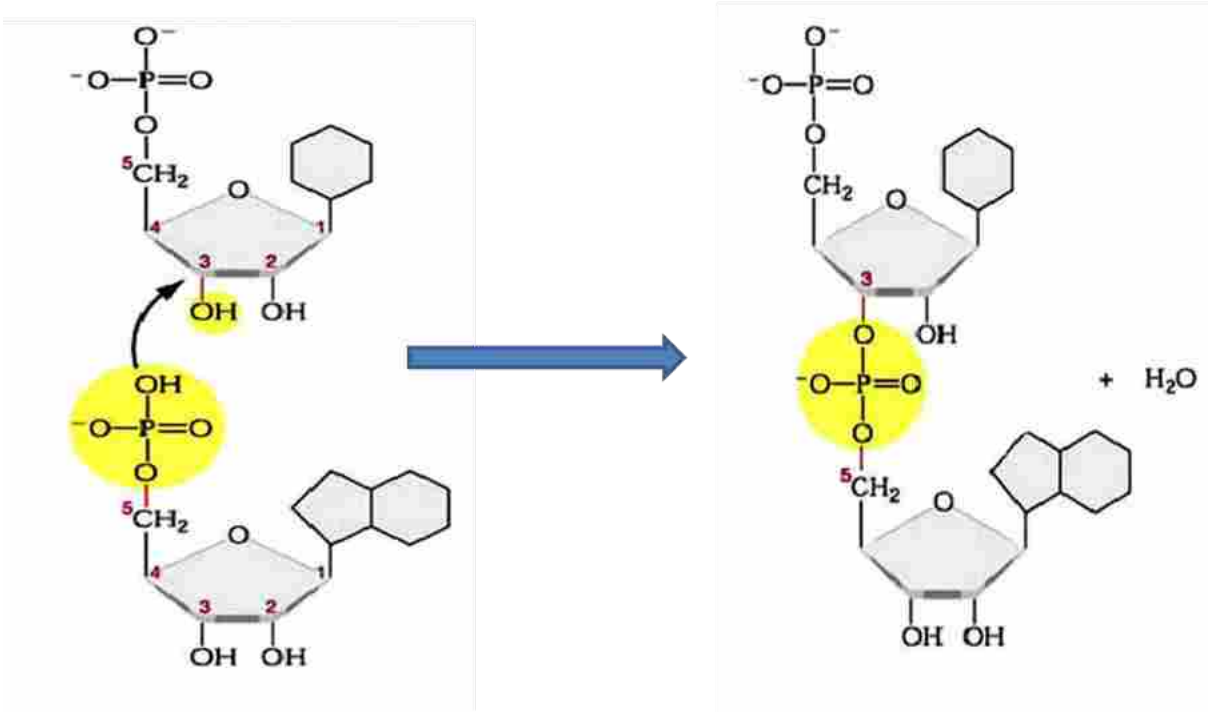
39- Types of Ribonucleotides

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



Types of Ribonucleotides

How do Ribonucleotides Join?



40- Types of RNAs

- 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.

Messenger RNA (mRNA)

- 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.

Transfer RNA (tRNA)

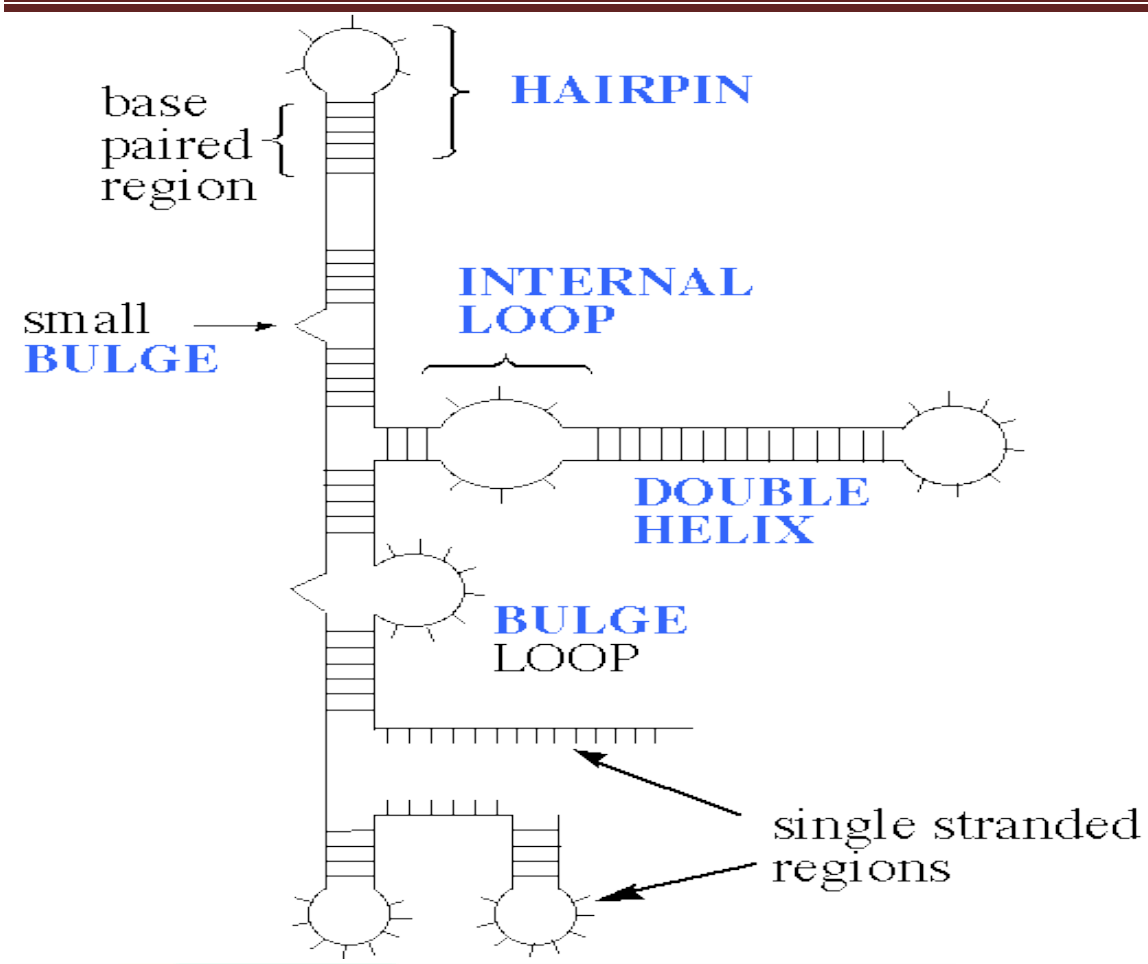
- They pair with the mRNA in such a way that amino acids are joined to a growing polypeptide in the correct sequence.

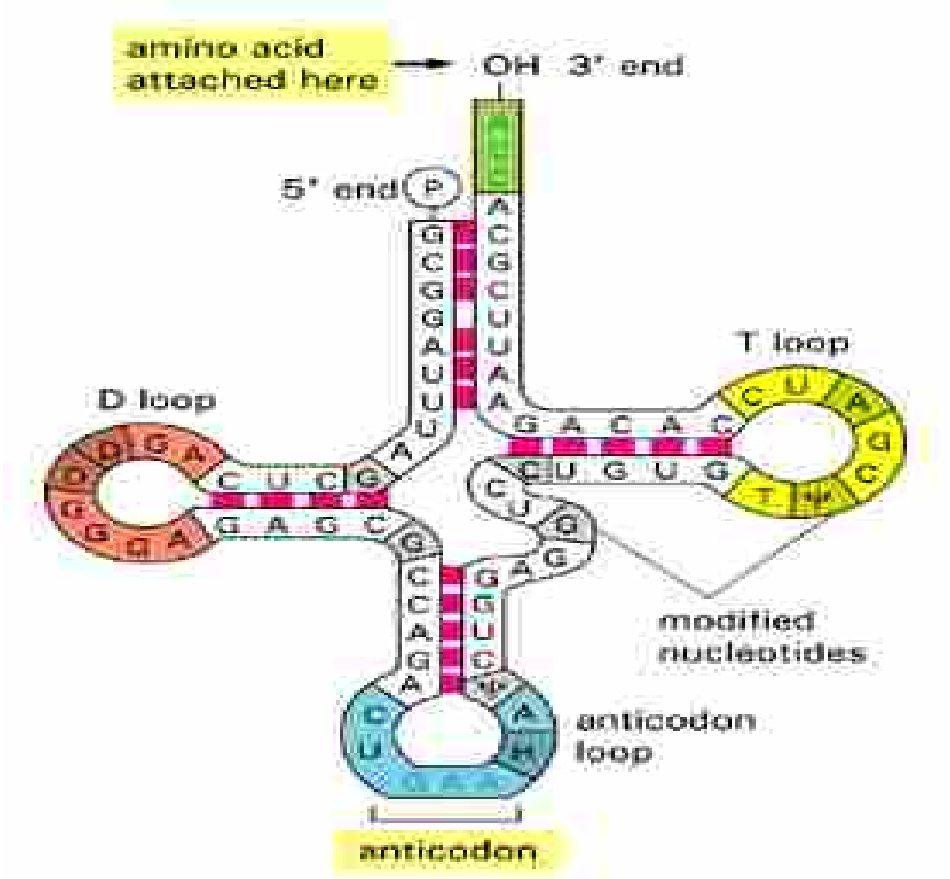
Ribosomal RNA (rRNA)

- 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.

41- Structures of RNAs

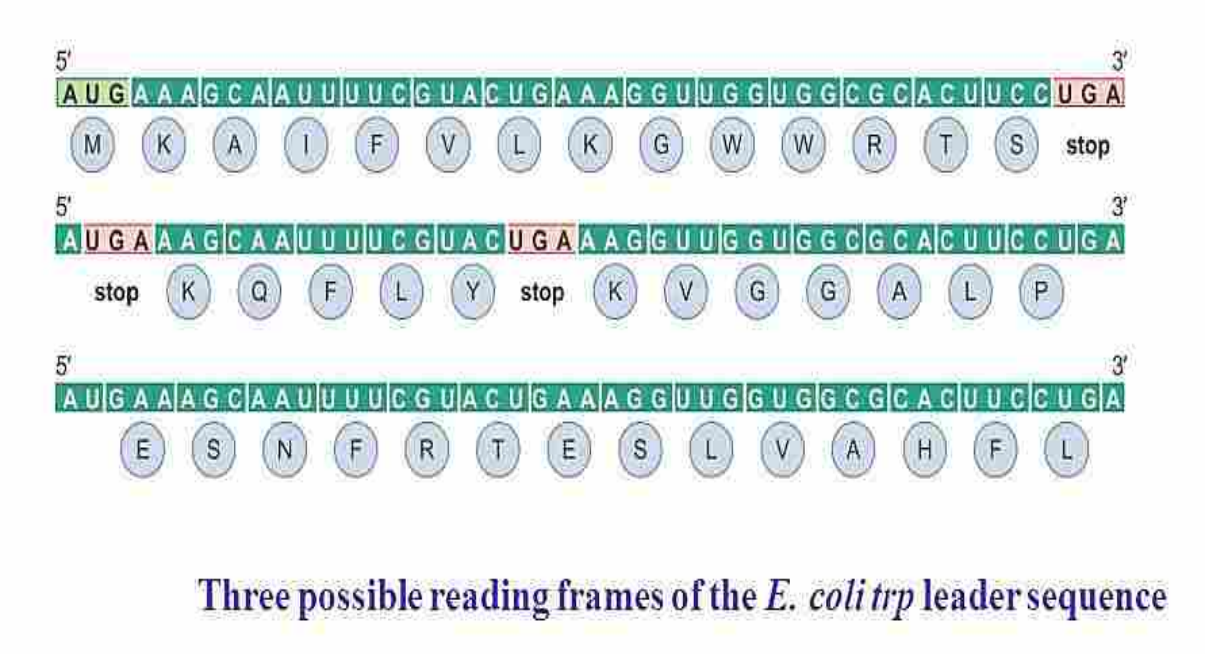
- mRNA is always single stranded when it is formed from DNA.
- But this single strand assumes a double helical conformation soon after its formation.
- This confirmation is achieved mainly due to base stacking interactions.
- Self-complementary sequences may occur in the RNA molecules which produce more 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.





42- Messenger RNA

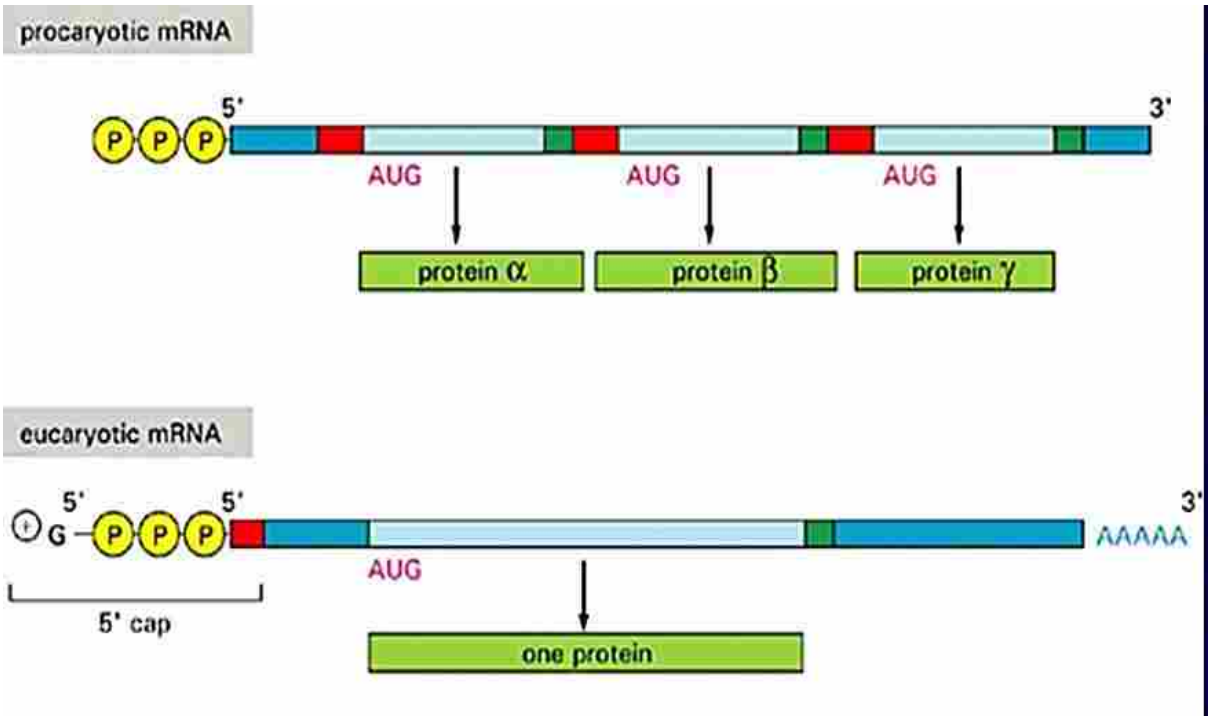
- The protein-coding region(s) of each mRNA is composed of a contiguous, non-overlapping string of codons called an open reading frame (commonly known as an ORF).
- Each ORF specifies a single protein and starts and ends at internal sites within the mRNA.
- That is, the ends of an ORF are distinct from the ends of the mRNA.
- 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.
- In bacteria, the start codon is usually 5'-AUG-3', but 5'-GUG-3' and sometimes even 5'-UUG-3' are also used.
- Eukaryotic cells always use 5'-AUG-3' as the start codon.
- The start codon has two important functions.
- First, it specifies the first amino acid to be incorporated into the growing polypeptide chain.
- Second, it defines the reading frame for all subsequent codons.
- Because each codon is immediately adjacent to the next codon, and because codons are three nucleotides long, any stretch of mRNA could be translated in three different reading frames.



Three possible reading frames of the *E. coli trp* leader sequence

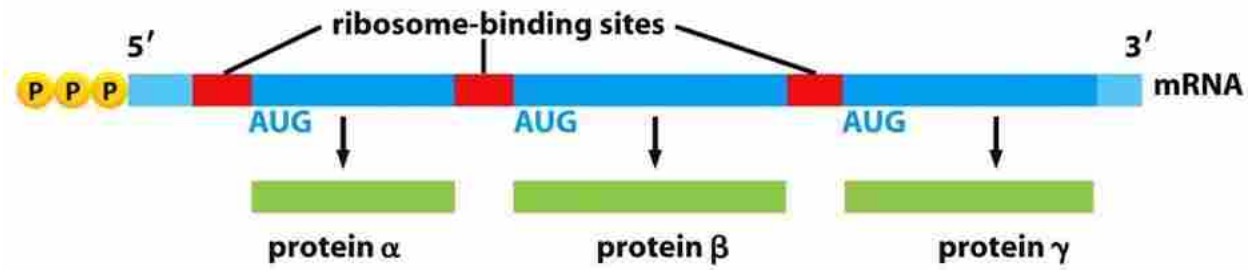
- Once translation starts, however, the reading frame is determined.
- Thus, by setting the location of the first codon, the start codon determines the location of all following codons.
- Stop codons, of which there are three (5'-UAG-3', 5'-UGA-3', and 5'-UAA-3'), define the end of the ORF and signal termination of polypeptide synthesis.
- 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.
- 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.

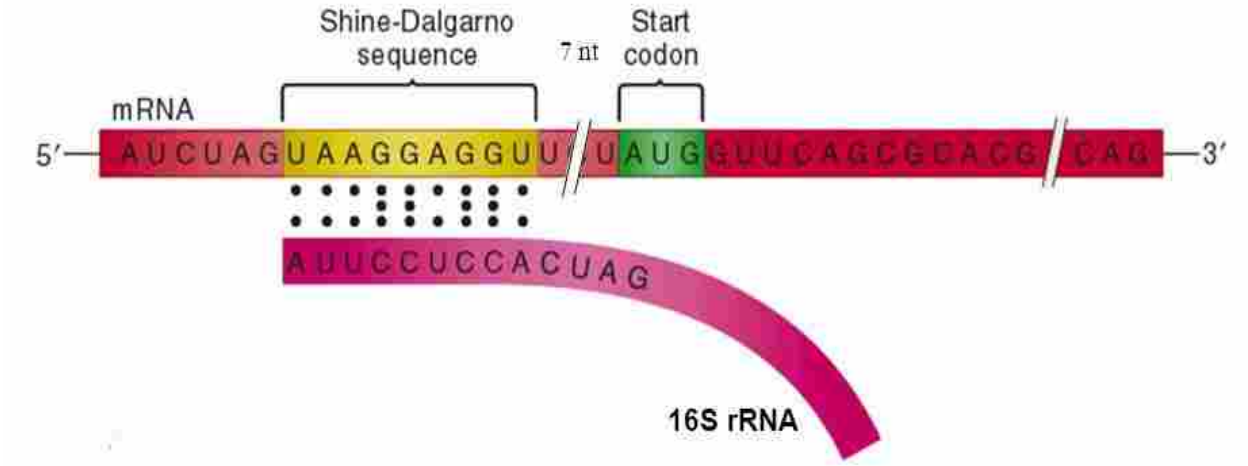
- In contrast, prokaryotic mRNAs frequently contain two or more ORFs.
- mRNAs containing multiple ORFs are known as polycistronic RNAs and those encoding a single ORF are known as monocistronic RNAs.
- The polycistronic mRNAs found in bacteria often encode proteins that perform related functions, such as different steps in the biosynthesis of an amino acid or nucleotide.



43- Prokaryotic mRNAs

- For translation to occur, the ribosome must be recruited to the mRNA.
- Prokaryotic mRNAs have a ribosome-binding site that recruits the translational machinery.
- 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).
- This element is also referred to as a Shine Őalgarno sequence after the scientists who discovered it by comparing the sequences of multiple mRNAs.
- The RBS, typically located 3–9 bp on the 5’ side of the start codon, is complementary to a sequence located near the 3’ end of one of the ribosomal RNA components, the 16S ribosomal RNA (rRNA).





- The RBS base-pairs with this RNA, thereby aligning the ribosome with the beginning of the ORF. The core of this region of the 16S rRNA has the sequence 5'-CCUCCU-3'.
- Not surprisingly, prokaryotic RBS are most often a subset of the sequence 5'-AGGAGG-3'.
- 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.
- High complementarity and proper spacing promote active translation, whereas limited complementarity and/or poor spacing generally support lower levels of translation.
- 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.
- In these cases, the start codon of the downstream ORF often overlaps the 3' end of the upstream ORF.
- Thus, a ribosome that has just completed translating the upstream ORF is positioned to begin translating from the start codon for the downstream ORF.
- This 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.
- Indeed, with two translationally coupled genes, a mutation that leads to a premature stop codon in the upstream ORF also prevents translation of the downstream ORF.

44- Eukaryotic mRNAs

- 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.

- 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.
- Created in three steps, the guanine nucleotide of the 5’ cap is connected to the 5’ end of the mRNA through three phosphate groups.
- The resulting 5’ cap is required to recruit the ribosome to the mRNA. Once bound to the mRNA, the ribosome moves in a 5’ 3’ direction until it encounters a 5’-AUG-3’ start codon, a process called scanning.
- 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 (5’-G/ANNAUGG-3’).
- This sequence was originally identified by Marilyn Kozak and is referred to as the Kozak sequence. Many eukaryotic mRNA slack these bases, but their presence increases the efficiency of translation.
- In contrast to the situation in prokaryotes, these bases are thought to interact with the initiator tRNA, not with an RNA component of the ribosome.
- 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.
- 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.
- Importantly, in addition to their roles in translation, these 5’- and 3’-end modifications also protect eukaryotic mRNAs from rapid degradation.

45-Transfer RNA

- The heart of protein synthesis is the “translation” of nucleotide sequence information (in the form of codons) into amino acids.
- This is accomplished by tRNA molecules, which act as adaptors between codons and the amino acids they specify.
- 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).
- tRNA molecules are between 75 and 95 ribonucleotides in length.
- Although the exact sequence varies, all tRNAs have certain features in common.
- First, all tRNAs end at the 3’ terminus with the sequence 5’-CCA-3’.
- Consistent with this absolute conservation, the 3’ end of this sequence is the site that is attached to the cognate amino acid.
- A second 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.
- For example, pseudouridine (ψU) is derived from uridine by an isomerization in which the site of attachment of the uracil base to the ribose is switched from the nitrogen at ring position 1 to the carbon at ring position 5.

- Likewise, dihydrouridine (D) is derived from uridine by enzymatic reduction of the double bond between the carbons at positions 5 and 6.
- Other 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.
- This observation suggests that the modified bases lead to improved tRNA
- function.
- For example, hypoxanthine plays an important role in the process of codon recognition by certain tRNAs.

46-Same as lec-45

47- Attachment of Amino Acids to tRNA

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.

- 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.
- This acyl linkage is a high-energy bond because its hydrolysis results in a large change in free energy.
- 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.
- All aminoacyl-tRNA synthetases attach an amino acid to a tRNA in two enzymatic steps:
    - Adenylation
    - tRNA charging
- Step one is adenylation in which the amino acid reacts with ATP to become adenylylated with the concomitant release of pyrophosphate.
- Adenylylation refers to transfer of AMP, as opposed to adenylation, which would indicate the transfer of adenine.
- The principal driving force for the adenylylation reaction is the subsequent hydrolysis of pyrophosphate by pyrophosphatase.
- 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.
- Step two is tRNA Charging in which the adenylylated amino acid, which remains tightly bound to the synthetase, reacts with tRNA.

- 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.
- There are two classes of tRNA synthetases:

- Class I enzymes attach the amino acid to the 20-OH of the tRNA and are generally monomeric.
- Class II enzymes attach the amino acid to the 30-OH of the tRNA and are typically dimeric or tetrameric.
- Although the initial coupling between the tRNA and the amino acid is different, once released from the synthetase, the amino acid rapidly equilibrates between attachment at the 3'-OH and the 2'-OH.
- Each of the 20 amino acids is attached to the appropriate tRNA by a single, dedicated tRNA synthetase.
- 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).
- Nevertheless, the same tRNA synthetase is responsible for charging all tRNAs for a particular amino acid.
- Thus, one and only one tRNA synthetase attaches each amino acid to all of the appropriate tRNAs.

48- The Ribosomes

- 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.

49- Formation of Peptide Bonds

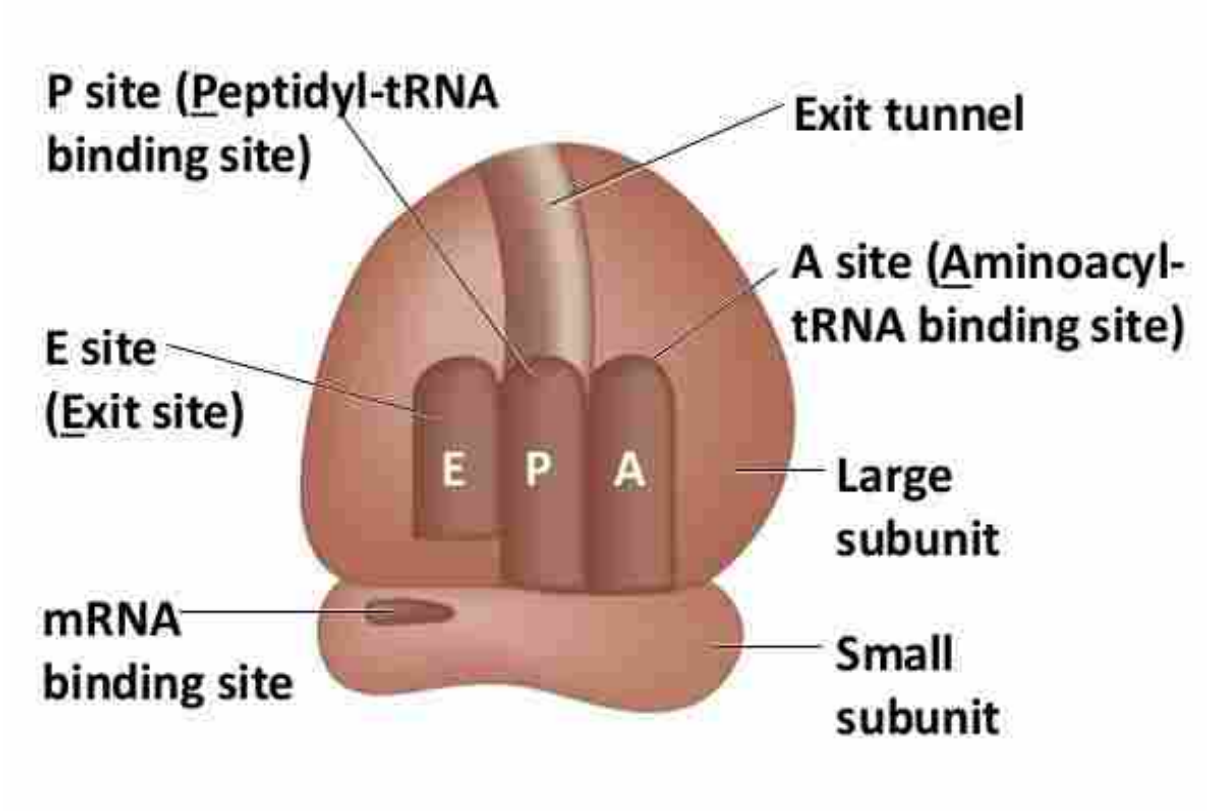
- Each new amino acid is added to the carboxyl terminus of the growing polypeptide chain (often referred to as synthesis in the amino- to carboxy-terminal direction).
- The ribosome catalyzes a single chemical reaction —the formation of a peptide bond.
- This reaction occurs between the amino acid residue at the carboxy-terminal end of the growing polypeptide and the incoming amino acid to be added to the chain.
- 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.
- The actual substrates for each round of amino acid addition are two charged species of tRNAs —an aminoacyl-tRNA and a peptidyl-tRNA.

- As you know the 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.
- The bond between the aminoacyl-tRNA and the amino acid is not broken during the formation of the next peptide bond.
- 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.
- To catalyze peptide-bond formation, the 3' ends of these two tRNAs are brought into close proximity by the ribosome.
- 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.
- 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. First, this mechanism of peptide-bond formation requires that the amino terminus of the protein be synthesized before the carboxyl terminus.
- 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.
- Interestingly, peptide-bond formation takes place without the simultaneous hydrolysis of a nucleoside triphosphate.
- This is because peptide-bond formation is driven by breaking the high-energy acyl bond that joins the growing polypeptide chain to the tRNA.
- Recall that this bond was created during the tRNA synthetase – catalyzed reaction that is responsible for charging tRNA.
- And the charging reaction involves the hydrolysis of a molecule of ATP.
- Thus, the energy for peptide-bond formation originates from the molecule of ATP that was hydrolyzed during the tRNA charging reaction.

50- Binding Sites on Ribosomes for tRNA

- The ribosome is composed of two subassemblies of RNA and protein known as the large and small subunits.
- The large subunit contains the peptidyl transferase center, which is responsible for the formation of peptide bonds.
- The small subunit contains the decoding center in which charged tRNAs read or "decode" the codon units of the mRNA.
- Both the decoding center and the peptidyl transferase center are buried within the intact ribosome.
- Yet, mRNA must be threaded through the decoding center during translation, and the nascent polypeptide chain must escape from the peptidyl transferase center.
- How do these polymers enter and exit the ribosome?

- The answer is provided by the structure of the ribosome, which reveals that there are "tunnels" in and out of the ribosome.
- To perform the peptidyl transferase reaction, the ribosome must be able to bind at least two tRNAs simultaneously.
- In fact, the ribosome contains three tRNA-binding sites, called the A-, P-, and E-sites.
- The A-site is the binding site for the aminoacylated-tRNA, the P-site is the binding site for the peptidyl-tRNA, and
- 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").



- Each tRNA binding site is formed at the interface between the large and the small subunits of the ribosome.
- 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.
- 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.
- The anticodon loops of the bound tRNAs are located adjacent to the small subunit.

51- Initiation of Translation

- For translation to be successfully initiated, three events must occur:-

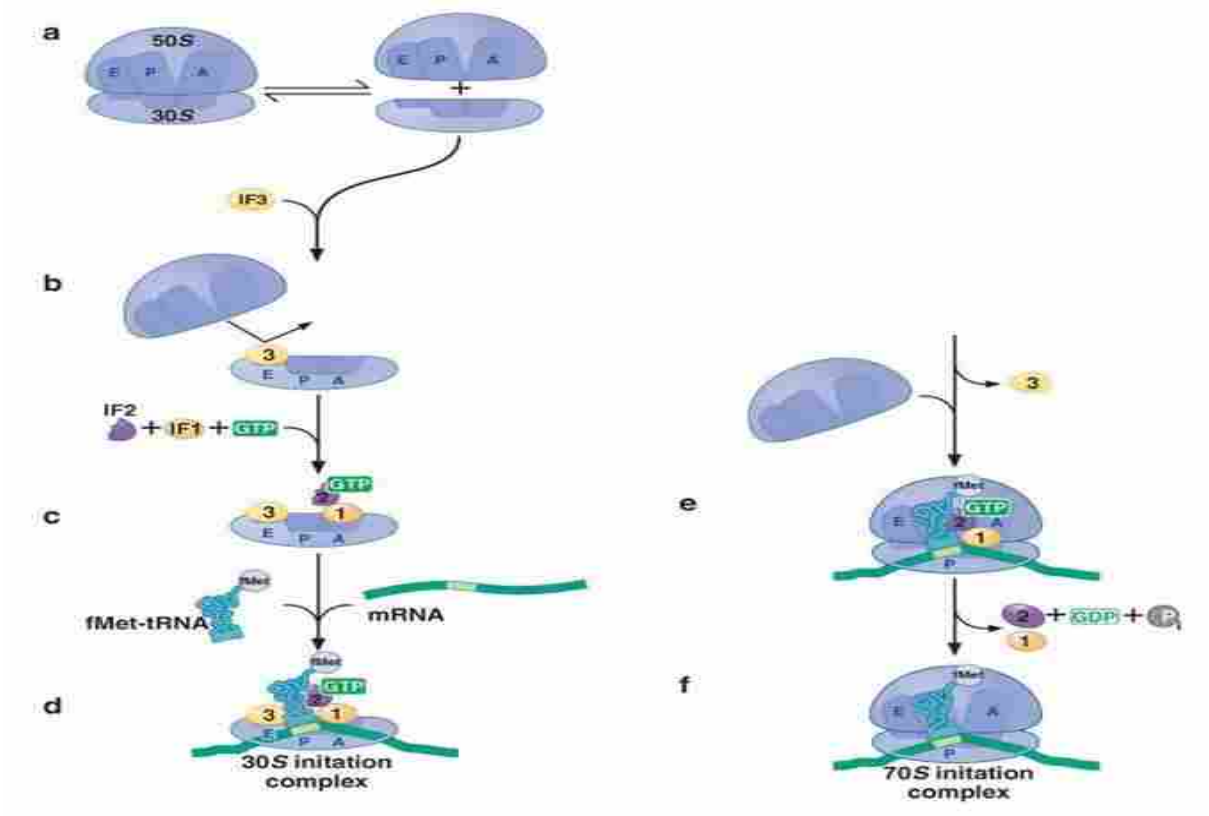
- i) the ribosome must be recruited to the mRNA.
- ii) a charged tRNA must be placed into the P-site of the ribosome.
- iii) the ribosome must be precisely positioned over the start codon.
- The correct positioning of the ribosome over the start codon is critical because this establishes the reading 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.
- In prokaryotes, the association of the small subunit with the mRNA is mediated by base-pairing interactions between the RBS and the 16S rRNA.
- 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.

52-The Initiation Factors

- 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.
- IF1:
- It prevents tRNAs from binding to the portion of the small subunit that will become part of the A-site.
- 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}).
- By interacting with these components, IF2 facilitates the association of fMet-tRNA ^{fMet} with the small subunit and prevents other charged tRNAs from associating with the small subunit.
- 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.
- IF3 becomes associated with the small subunit at the end of a previous round of translation when it helps to dissociate the 70S ribosome into its large and small subunits.
- Each of the initiation factors binds at, or near, one of the three tRNA binding sites on the small subunit.
- Consistent with its role in blocking the binding of charged tRNAs to the A-site, IF1 binds directly to the portion of the small subunit that will become the A-site.
- IF2 binds to IF1 and reaches over the A-site into the P-site to contact the fMet - tRNA ^{fMet}.
- Finally, IF3 occupies the part of the small subunit that will become the E-site.
- Thus, of 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.
- 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.



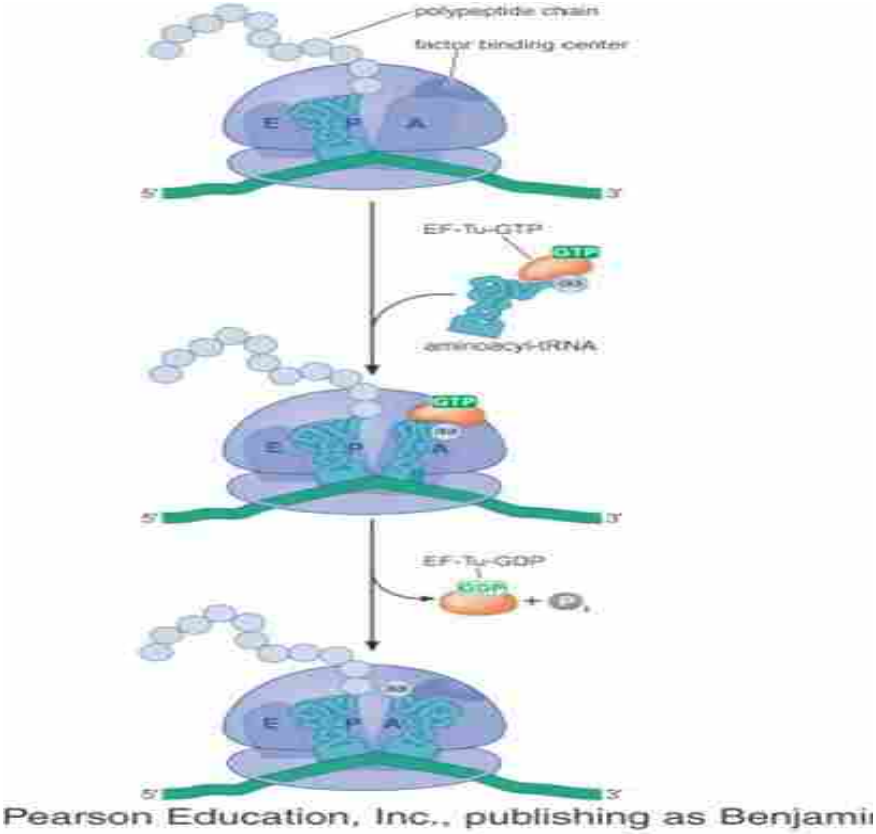
- Binding fMet-tRNA ^{fMet} to the small subunit is facilitated by its interactions with IF2 bound to GTP and base pairing between the anticodon and the start codon of the mRNA.
- Similarly, base pairing between the fMet-tRNA ^{fMet} and the mRNA serves to position the start codon in the P-site.
- This altered conformation results in the release of IF3.

- In the absence of IF3, the large subunit is free to bind to the small subunit with its cargo of IF1, IF2, mRNA, and fMet-tRNA^{fMet}.
- In particular, IF2 acts as an initial docking site of the large subunit, and this interaction subsequently stimulates the GTPase activity of IF2.GTP.
- 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.
- Thus, the net result of initiation is the formation of an intact (70S) ribosome assembled at the start site of the mRNA with fMet-tRNA^{fMet} in the P-site and an empty A-site.
- The ribosome – mRNA complex is now poised to accept a charged tRNA into the A-site and commence polypeptide synthesis.

53-Translation Elongation

- 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.
- First, the correct aminoacyl-tRNA is loaded into the A site of the ribosome as dictated by the A-site codon.
- 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.
- 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.
- As with the original positioning of the mRNA, this shift must occur precisely to maintain the correct reading frame of the message.
- Two auxiliary proteins known as elongation factors control these events.
- Both of these factors use the energy of GTP binding and hydrolysis to enhance the rate and accuracy of ribosome function.
- Unlike the initiation of translation, the mechanism of elongation is highly conserved between prokaryotic and eukaryotic cells.
- Unlike the initiation of translation, the mechanism of elongation is highly conserved between prokaryotic and eukaryotic cells.
- Aminoacyl-tRNAs do not bind to the ribosome on their own. Instead, they are "escorted " to the ribosome by the elongation factor EF-Tu.
- Once a tRNA is aminocylated, EF-Tu binds to the tRNA's 3 ' end, masking the coupled amino acid. This interaction prevents the bound aminoacyl-tRNA from participating in peptide bond formation until it is released from EF-Tu.
- Like the initiation factor IF2, the elongation factor EF-Tu binds and hydrolyzes GTP and the type of guanine nucleotide bound governs its function.
- EF-Tu can only bind to an aminoacyl-tRNA when it is associated with GTP. EF-Tu bound to GDP, or lacking any bound nucleotide, shows little affinity for aminoacyl-tRNAs.
- Thus, when EF-Tu hydrolyzes its bound GTP, any associated aminoacyl-tRNA is released.

- 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.
- 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.
- At this point, EF-Tu hydrolyzes its bound GTP and is released from the ribosome.
- The control of GTP hydrolysis by EF-Tu is critical to the specificity of translation.



- The error rate of translation is between 10^{-3} to 10^{-4}.
- 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.
- 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.

54-The Ribosome Is a Ribozyme

- 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.

55-Translocation in the Large Subunit

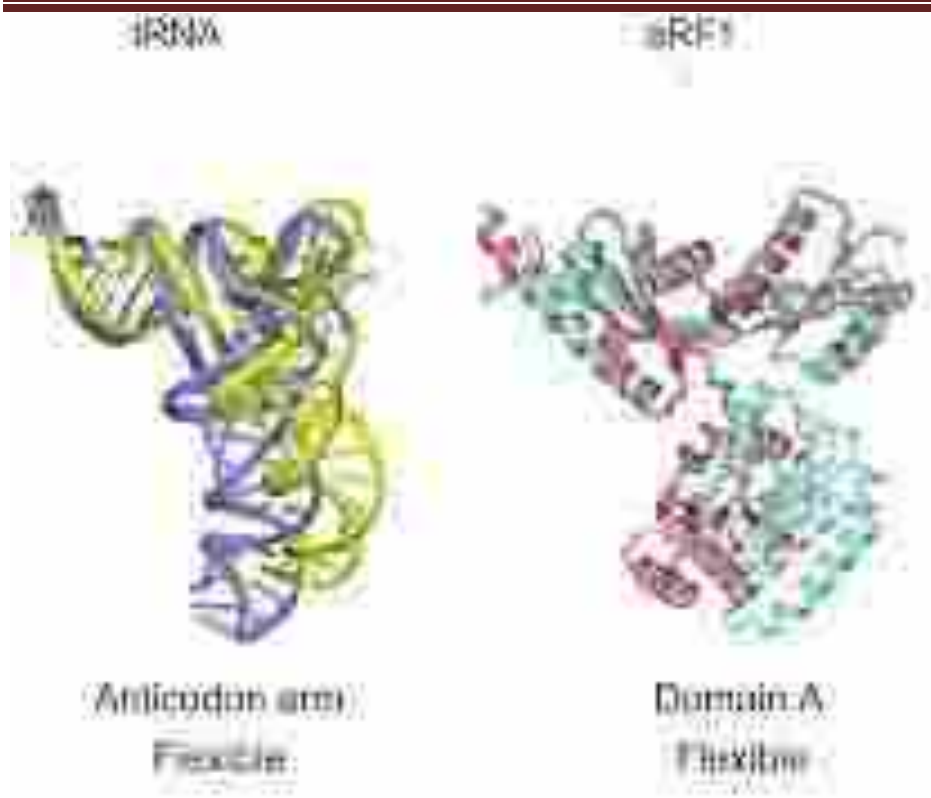
- Once the peptidyl transferase reaction has occurred, the tRNA in the P-site is deacetylated (no longer attached to an amino acid), and the growing polypeptide chain is linked to the tRNA in the A-site.
- For a new round of peptide chain elongation to occur, the P-site tRNA must move to the E-site and 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.
- These movements are coordinated within the ribosome and are collectively referred to as translocation.
- The initial steps of translocation are coupled to the peptidyl transferase reaction.
- 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.
- 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.
- 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.
- In contrast, at this time, the anticodons of these tRNAs remain in their initial location in the small subunit bound to the mRNA.
- Thus, translocation is initiated in the large subunit before the small subunit, and the tRNAs are said to be in “hybrid states.”
- Their 3’ ends have shifted into a new location, but their anticodon ends are still in their pre–peptidyl transfer position.

- Importantly, this change is associated with a counter clockwise rotation of the small subunit relative to the large subunit facilitating interaction of the tRNAs with distinct tRNA-binding sites in the different subunits.
- 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.
- After the peptidyl transferase reaction, EF-G–GTP binds to and stabilizes the ribosome in the rotated, hybrid state.
- 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.
- First, interactions between EF-G–GDP and the ribosome are thought to “unlock” the ribosome.
- Structural studies reveal that there are “gates” that separate the A-, P-, and E-sites and EF-G–GDP is said to unlock the ribosome by opening these gates.
- 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.
- Because the ribosome is unlocked, the formerly A-site tRNA can move into the P-site, allowing EF-G–GDP to bind the A-site.
- Completion of translocation is accompanied by a clockwise rotation of the small subunit back to its starting position.
- The resulting ribosome structure has dramatically reduced affinity for EF-G–GDP.
- 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.
- 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.

56-Termination of Translation

- 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.
- It was initially postulated that there would be one or more chain terminating tRNAs that would recognize these codons.
- However, this is not the case.
- Instead, 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.
- Class I release factors recognize the stop codons and trigger hydrolysis of the peptide chain from the tRNA in the P-site.
- Prokaryotes have two class I release factors called RF1 and RF2.
- RF1 recognizes the stop codon UAG and RF2 recognizes the stop codon UGA.
- The third stop codon, UAA, is recognized by both RF1 and RF2.

- In eukaryotic cells, there is a single class I release factor called eRF1 that recognizes all three stop codons.
- Class II release factors stimulate the dissociation of the class I factors from the ribosome after release of the polypeptide chain.
- Prokaryotes and eukaryotes have only one class II factor called RF3 and eRF3, respectively.
- Like EF-G, IF2, and EF-Tu, class II release factors are regulated by GTP binding and hydrolysis.
- How do release factors recognize stop codons?
- Because release factors are composed entirely of protein, protein–RNA interaction must mediate stop codon recognition.
- Experiments in which short coding regions were genetically swapped between RF1 and RF2 (having different stop-codon specificity) identified a three-amino-acid sequence that is critical for release factor specificity.
- Exchange of these three amino acids between RF1 and RF2 swaps the stop-codon specificity of the two complexes.
- For this reason, this three-amino-acid sequence is called a peptide anticodon and must interact with and recognize stop codons.
- A 3D structure of RF1 bound to the ribosome confirms that RF1 binds to the A-site of the ribosome.
- In this structure, the peptide anticodon is located very near the anticodon, but it is likely that there are additional protein regions that contribute to codon recognition.
- A region of class I release factors that stimulates polypeptide release has also been identified.
- All class I factors share a conserved three-amino-acid sequence (glycine, glycine, glutamine) that is essential for polypeptide release.
- Moreover, the structure of RF1 bound to the ribosome confirms that the GGQ motif is located in close proximity to the peptidyl transferase center.
- It remains unclear whether the GGQ motif is directly involved in the release of polypeptide from the peptidyl-tRNA or it induces a change in the peptidyl transferase center that allows the center itself to catalyze hydrolysis.
- Studies of the conserved bases found adjacent to the CCA ends in the peptidyl transferase center indicate that several of these residues are required for peptide hydrolysis.
- Indeed, these bases appear to play a more important role in peptide release than they do in peptide-bond formation.
- Together, these studies have led to the hypothesis that class I release factors functionally mimic a tRNA, having a peptide anticodon that interacts with the stop codon and a GGQ motif that reaches into the peptidyl transferase center.



- Once the class I release factor has triggered the hydrolysis of the peptidyl tRNA linkage, it must be removed from the ribosome.
- This step is stimulated by the class II release factor, RF3.
- RF3 is a GTP-binding protein but, unlike the other GTP-binding proteins involved in translation, this factor has a higher affinity for GDP than GTP.
- Thus, free RF3 is predominantly in the GDP-bound form.
- RF3-GDP binds to the ribosome in a manner that depends on the presence of a class I release factor.
- After the class I release factor stimulates polypeptide release, a change in the conformation of the ribosome and the class I release factor stimulates RF3 to exchange its bound GDP for a GTP.
- The binding of GTP to RF3 leads to the formation of a high-affinity interaction with the ribosome that favors the rotated hybrid state.
- This change in conformation displaces the class I factor from the ribosome.
- These changes also allow RF3 to associate with the factor-binding center of the large subunit. As with other GTP-binding proteins involved in translation, this interaction stimulates the hydrolysis of GTP.
- In the absence of a bound class I factor, the resulting RF3.GDP has a low affinity for the ribosome and is released.

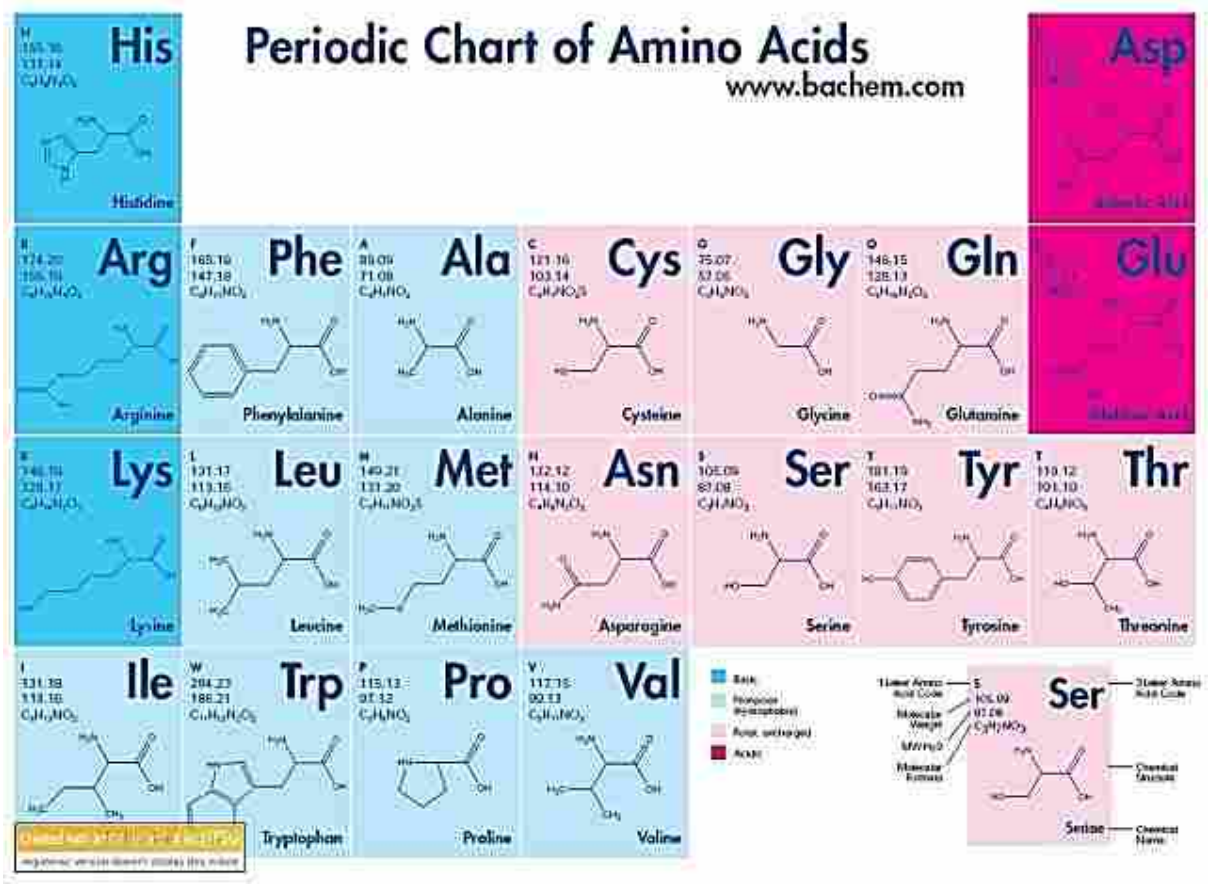
57-Missing
58-Amino Acids

Background

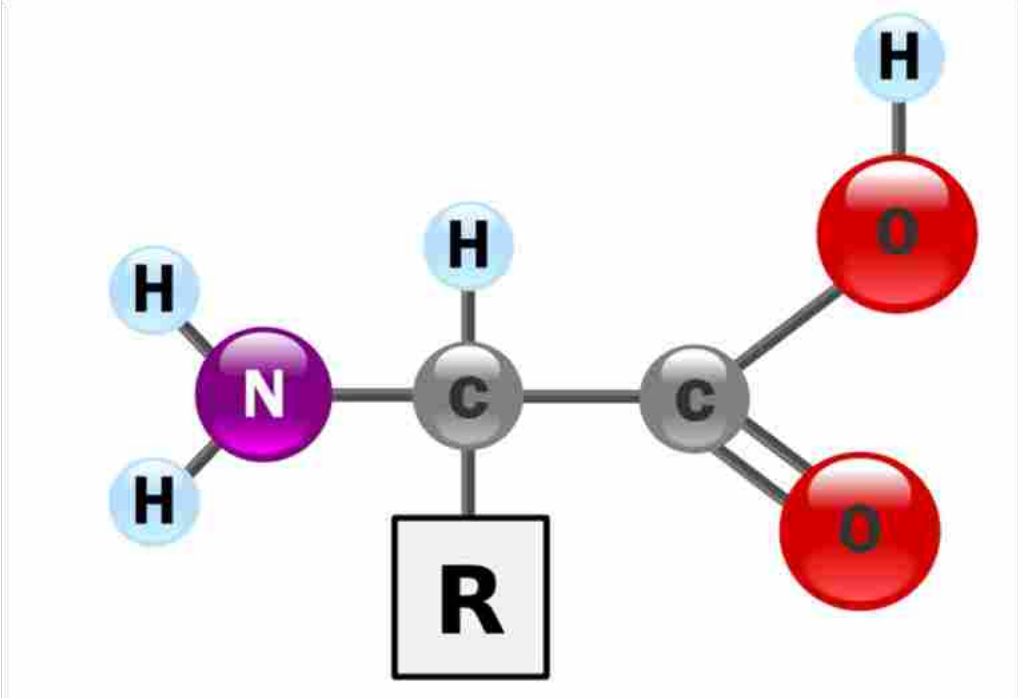
- Translation involves coding of proteins by RNAs at ribosomes
- Codons of three nucleotides from the RNA molecules code one amino acid at a time

Introduction

- In all, 20 different amino acids can be coded by such codons during translation
- The amino acids polymerize to form proteins
- Next, we will study amino acids and the respective codons!



Chemical Structure of an Amino Acid



Conclusion

- Proteins are formed after polymerization of amino acids
- Amino acids are chained together by peptide bonds
- Proteins thus formed fold to take 3D conformations!

59-Proteins: Polymers of Amino Acids

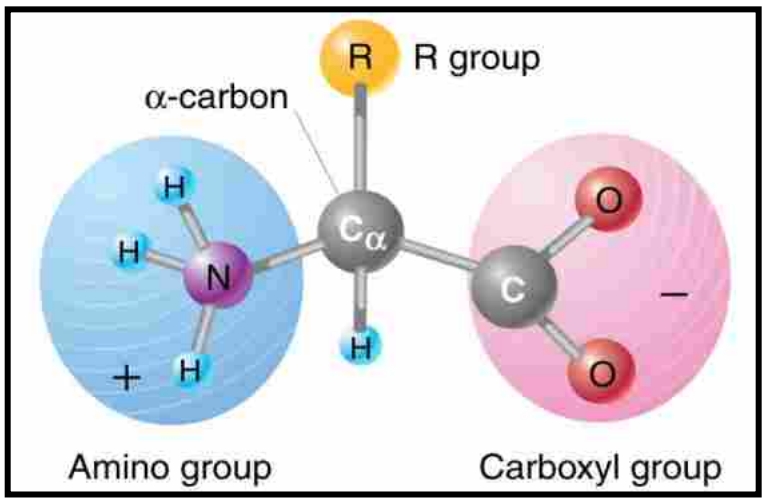
Protein Functions

Most Abundant Polymer in the cell

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

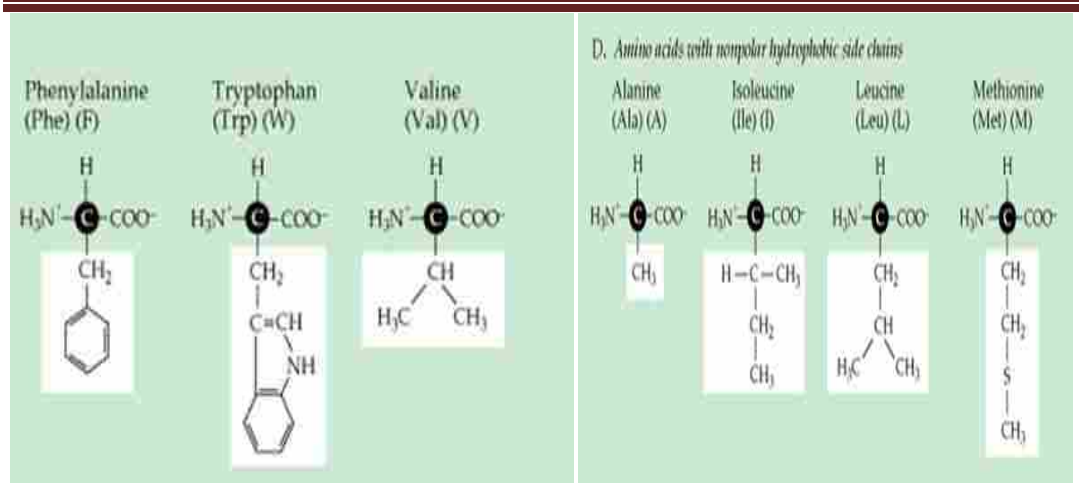
Do not store genetic information

Amino Acid Structure



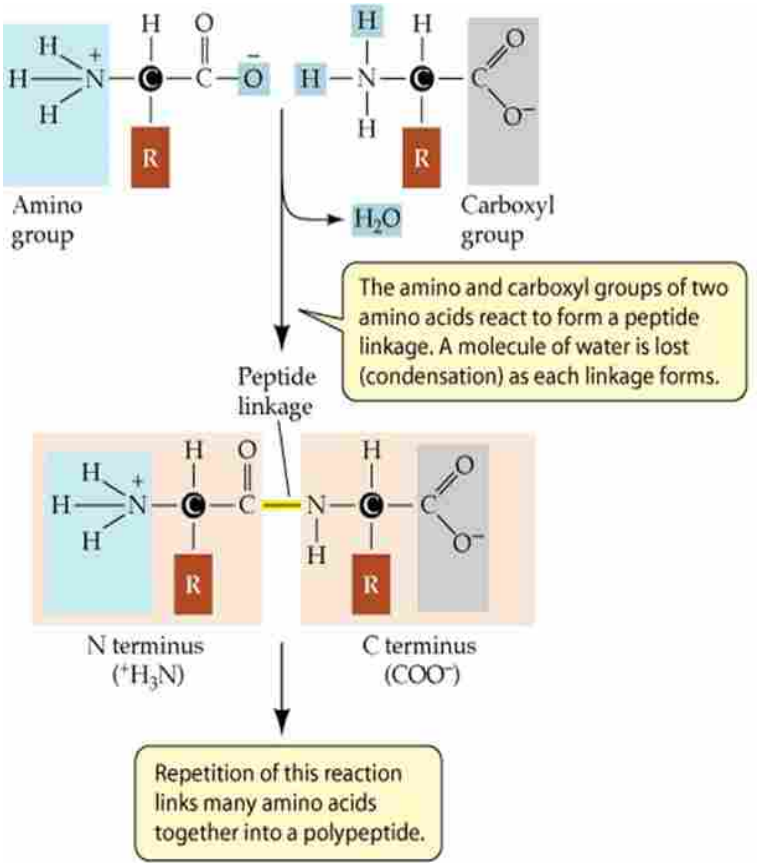
The side chains or R groups give different properties to each of the amino acids

Nonpolar R Groups



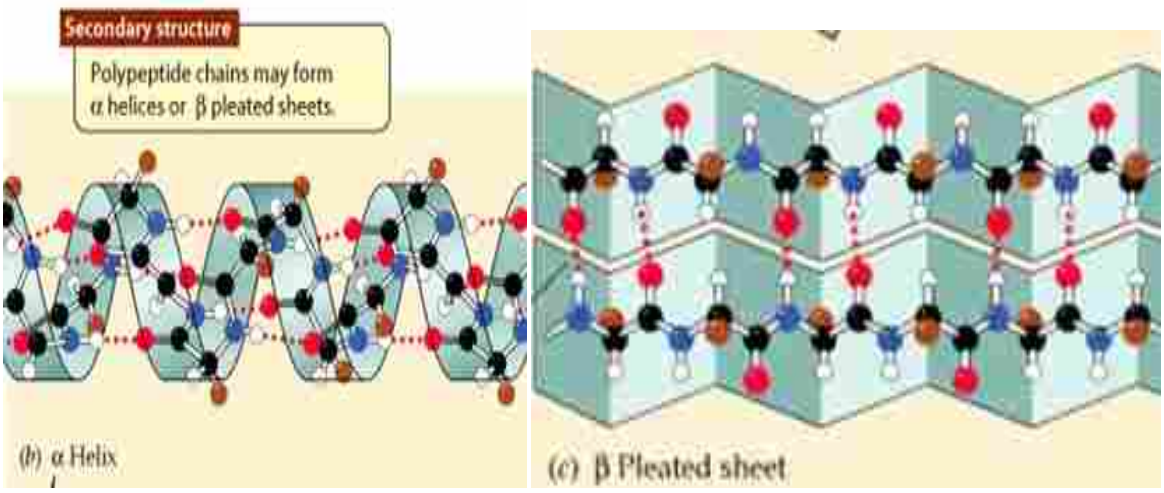
60-Proteins: Structure

The Peptide Bond



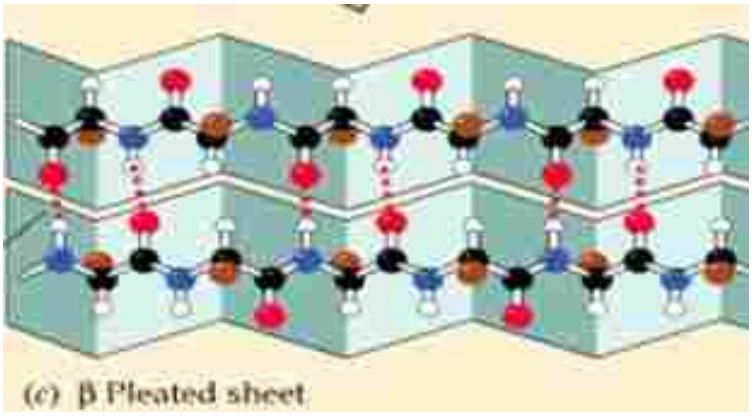
Secondary Structure

α helices (hydrogen bonds between amino acid residues)



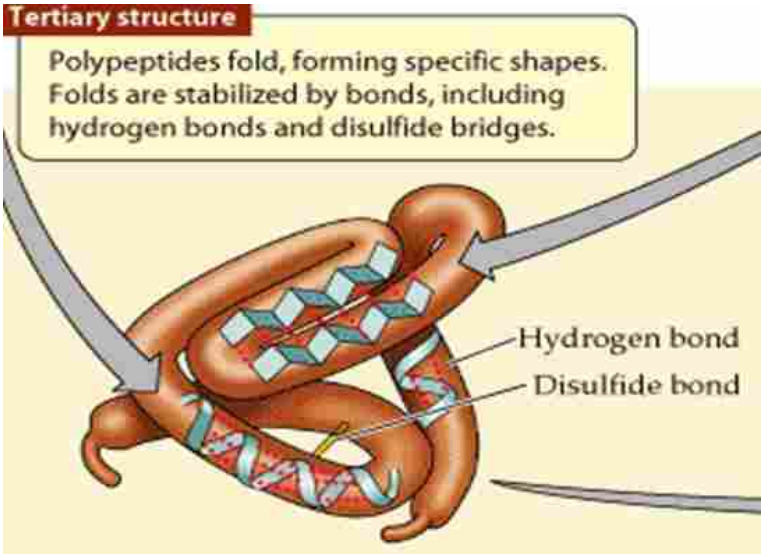
Secondary Structure

β pleated sheets (hydrogen bonds between amino acid residues)



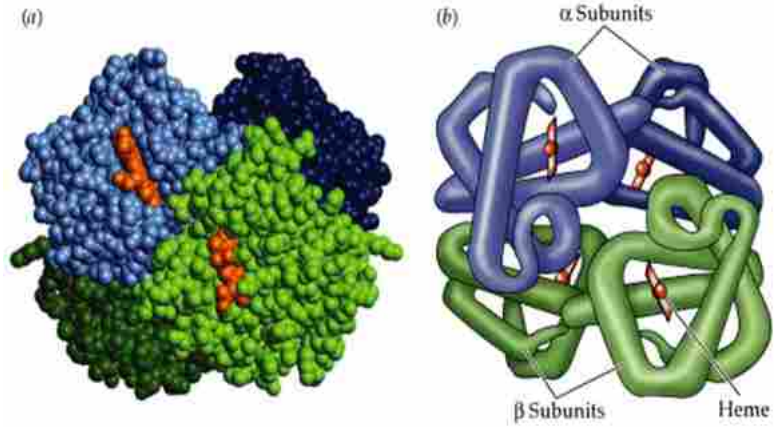
Tertiary Structure

Generated by bending and folding of the polypeptide chain



Quaternary structure

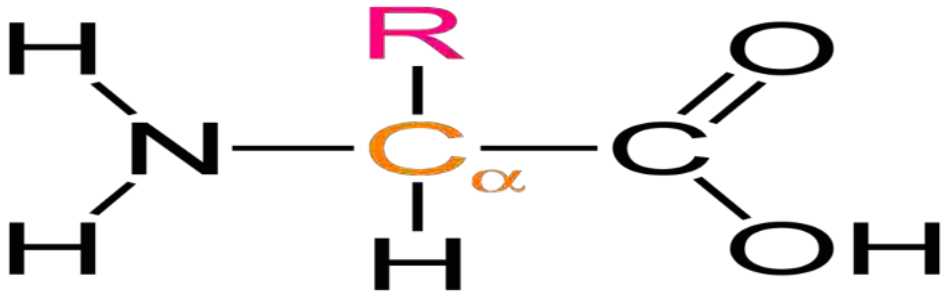
Arrangement of multiple folded protein molecules in a multi-subunit complex.



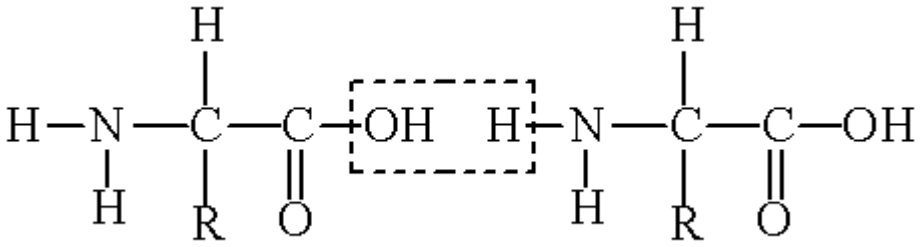
61-Missing

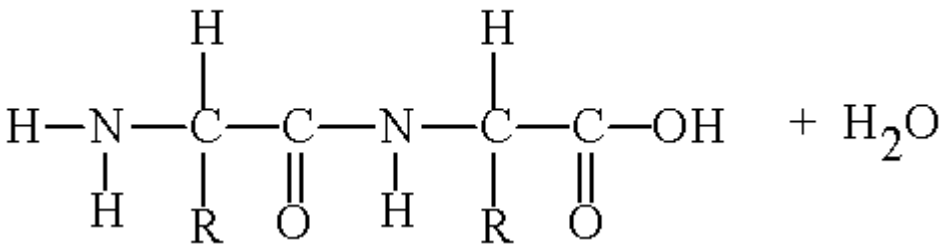
62-Chemical composition of proteins

- 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.





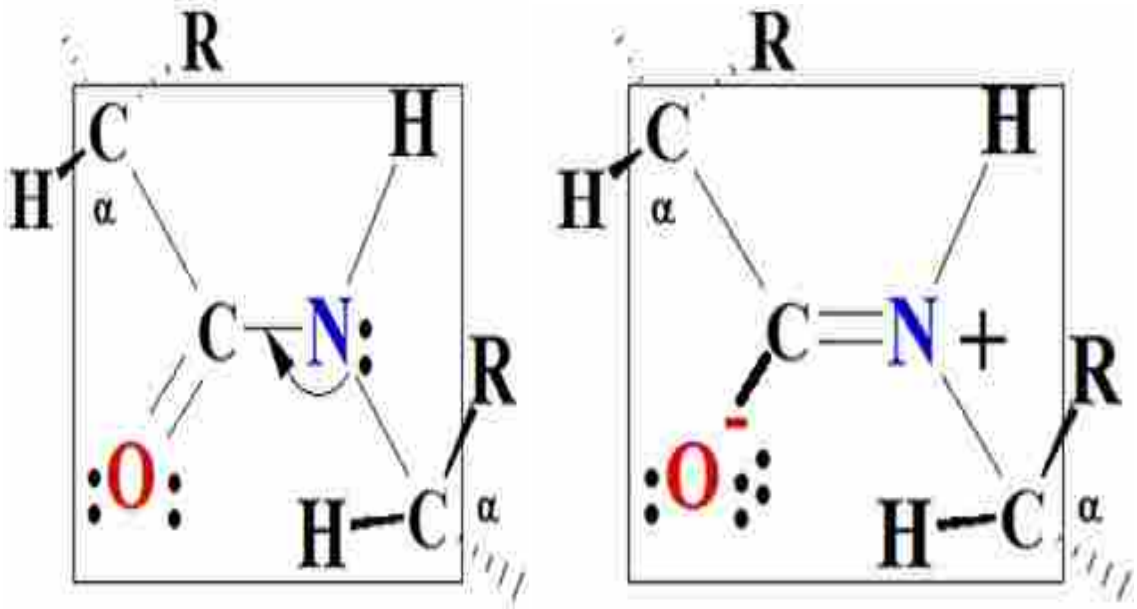
- Amino acids can successively join to form dipeptides, tripeptides, tetrapeptides, oligo peptides and polypeptides.

63- Primary structure of proteins

- 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

- Linus Pauling and Robert Corey carefully analyzed the peptide bond.
- Their findings laid the foundation for our present understanding of protein structure.
- 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.

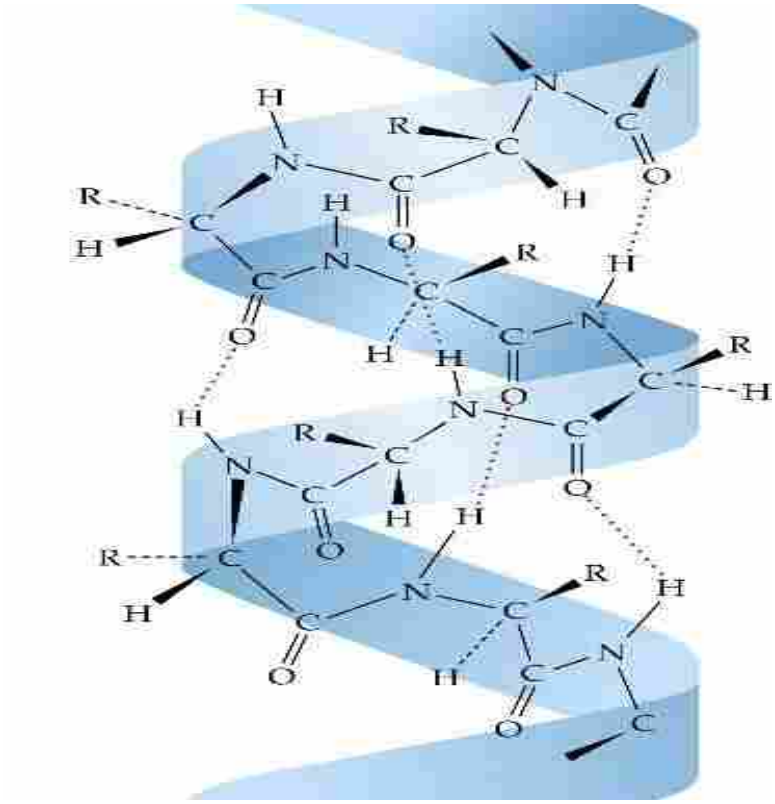
- In principle, ϕ and ψ can have any value between +180 & -180.

64- Secondary structure of proteins

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

65- α- Helix

- 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.

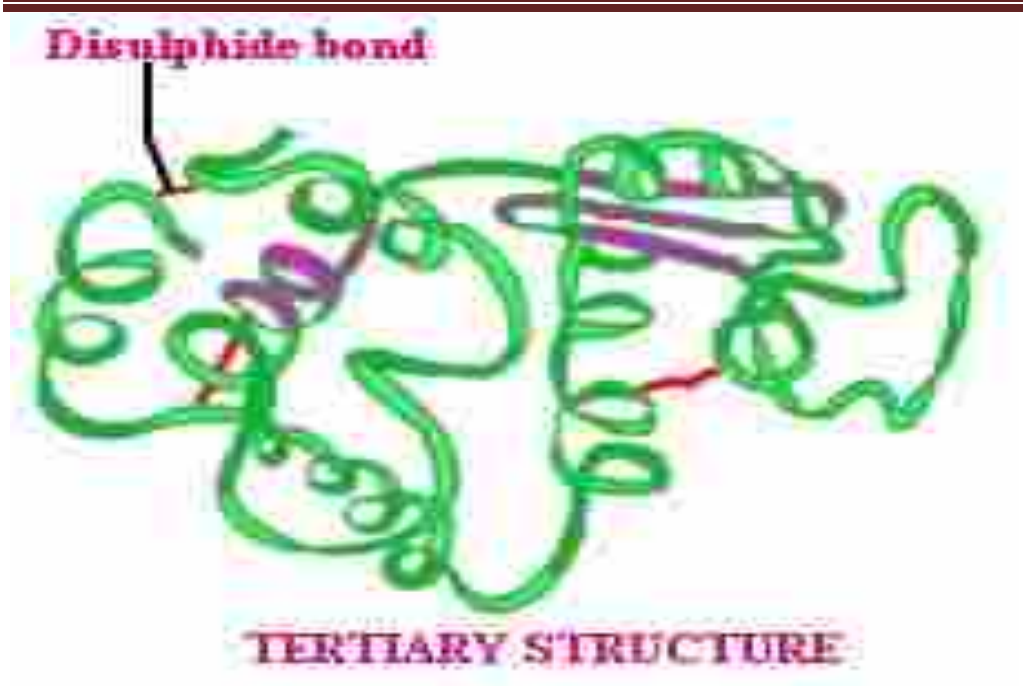


- The helical twist of the α-helix found in all proteins is right-handed.
- The repeating unit is a single turn of the helix, which extends about 5.4 Å (includes 3.6 amino acid residues) along the long axis.
- The amino acid residues in an helix have conformations with psi = –45 to –50 and phi = – 60.
- An helix makes optimal use of internal hydrogen bonds.
- About one-fourth of all amino acid residues in polypeptides are found in α-helices while in some proteins it is the predominant structure.

66-Missing

67- Tertiary Structure of Proteins

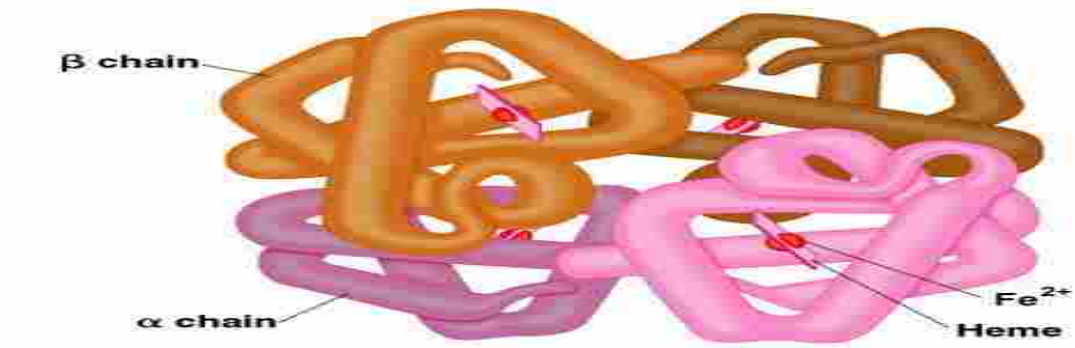
- The overall three-dimensional arrangement of all atoms in a protein is referred to as the protein's tertiary structure.



- 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.

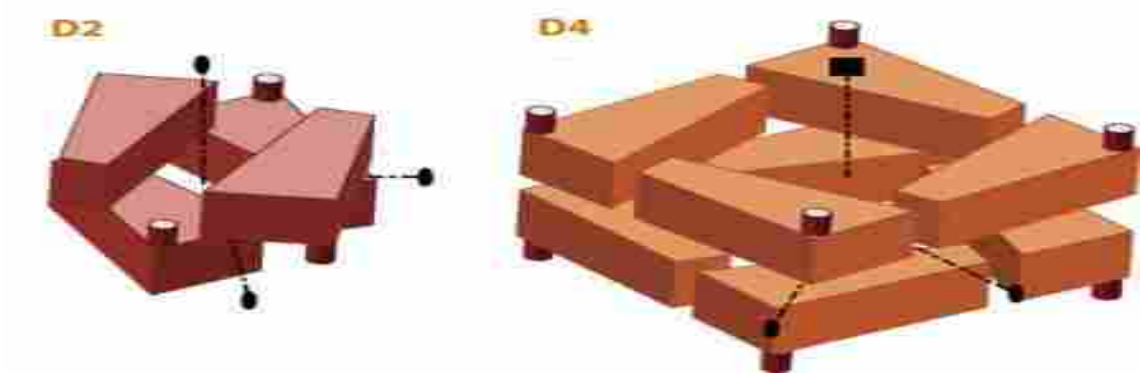
68- Quaternary Structure of Proteins

- Some proteins contain two or more separate polypeptide chains or subunits, which may be identical or different.
- The spatial arrangement of these subunits is known as a protein's quaternary structure.
- A multi-subunit protein is also referred to as a multimer.
- A multimer with just a few subunits is called as oligomer and a single subunit or a group of subunits, is called a protomer.



- Identical subunits of multimeric proteins are generally arranged in a symmetric patterns.
- Oligomers can have either rotational symmetry or helical symmetry.

- There are several forms of rotational symmetry. The simplest is cyclic symmetry, involving rotation about a single axis.
- A somewhat more complicated rotational symmetry is dihedral symmetry, in which a twofold rotational axis is present.



- More complex rotational symmetries include icosahedral symmetry.
- An icosahedron is a regular 12-cornered polyhedron having 20 triangular faces.
- The other major type of symmetry found in oligomers is helical symmetry.

69- Storage of Biological Sequence Information

Background

- DNA sequences contain A, C, T & G
- mRNA sequences contain A, C, U & G
- Protein sequences contain A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y & P
- 20 Amino Acids

Amino Acid	3-Letters	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Introduction

- DNA, mRNA and Proteins can be sequenced in the laboratory
- We can therefore obtain DNA, mRNA and Protein sequences which are simple lists of alphabets

Storing Sequences

- DNA sequences may contain hundreds of thousands of bases
- RNA sequences can be a little less than the DNA but they are in a much larger variety
- Proteins have long sequences and are large in number
- A problem then arises on how to store this information?

Solution: Databases

- The solution lies with public sequence databases
- DNA/RNA-> GenBank (by NIH)
- Proteins->UniProt (by UniProt consortium)

Conclusion

- Nucleotide and amino acid sequences are stored in online databases
- DNA & RNA information is available in GenBank
- Protein sequence information is available in UniProt

70- Comparing Sequences

Background

- A large number of sequences are available in GenBank and Uniprot, publically
- What if we try to compare them?
- What will be get from such a comparison?

Introduction

By comparing sequences, we can get:

- Similarity
- Specific Differences
- Relationship
- Evolutionary Insight

DNA

- ACTGGTCAGTCATGTCATGTCA

RNA

- CAGUACGUAGCUAGGAGUAGC

Protein

- MQVKLFTLPHLKIHQLFTLPHLKIHDH

Conclusion

Several conclusions can be drawn on a set of sequences in terms of:

- Similarity
- Specific Differences
- Relationship

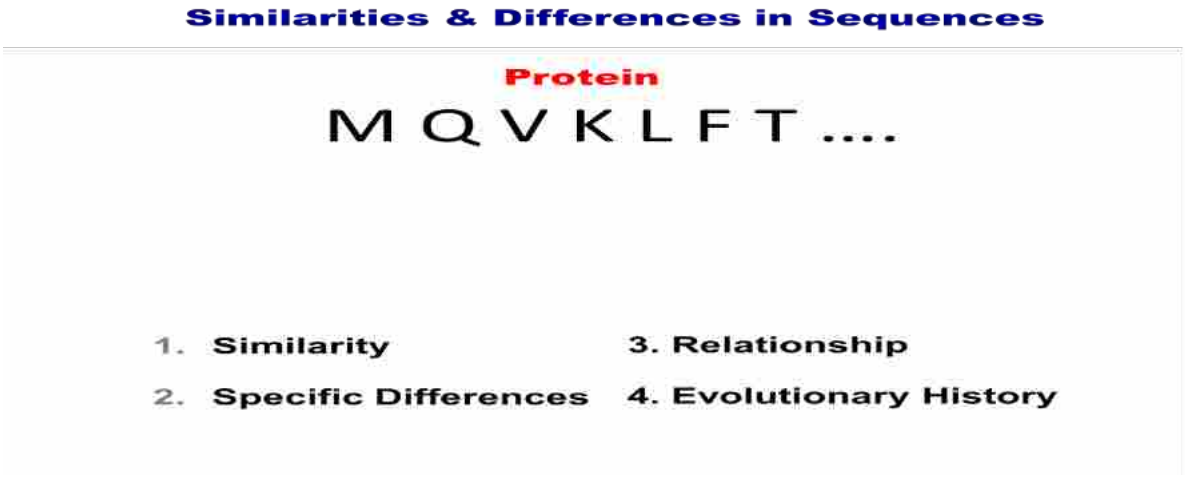
- Evolutionary Insight

71- Similarities & Differences in Sequences

Background

By comparing two sequences, one can study:

- Similarity
- Specific Differences
- Relationship
- Evolutionary History



Exact Matching

Not only should the two or more sequences being compared have the exact same number of each nucleotide (for DNA /RNA) or amino acid (for Proteins), but that they should be arranged in the exact same order!

Accommodating Differences – In exact Matching

- Regular Expression
- Signature Sequences
- PROSITE Patterns Database

- Hyphens separate the elements of the pattern,
- Letters refer to amino acids,
- X indicates any amino acid,
- Bracketed numbers denote the repeat length of a residue,
- If this repeat length varies, the range of this variation is quoted in the brackets.

Conclusion

- Exact Matches include complete matching in terms of residues and their order
- In exact matching includes partial matching in terms of residues and order with room for variations in both.

72- Pairwise Sequence Alignment - I

Background

- Exact matches include complete matching in terms of residues and their order

- In exact matching includes partial matching in terms of residues and order with room for variations in both.

Introduction

- Concept of conserved residues or nucleotides
- Before starting to matching nucleotide or amino acid sequences, we need to find location of conserved residues or nucleotides

Alignment

- The process of inexact matching while keeping in view the conserved residues is called sequence "Alignment"

Pairwise sequence alignment is therefore alignment of a pair of two sequences

- Two Examples of Pairwise Alignment

Conclusion

- Pairwise sequence alignment helps compare two sequences
- It takes into consideration inexact matching
- Matches are colored and missing nucleotides are denoted by '.'

73- Pairwise Sequence Alignment - II

Background

- Pairwise sequence alignment compares two amino acid or nucleotide sequences
- It employs inexact matching
- Matches are colored and missing nucleotides are denoted by '.'

Salient Points

- Sliding the sequences past each other
- Aim is to maximize matches
- Gaps ('.') could be inserted to account for insertions and deletions

Salient Points

- Gaps ('.') may carry a penalty for reducing scores from unreasonable alignments
- Without a gap penalty an exact match and match with gaps will get equal scores

- Types of pairwise alignments

Global	**Local**

• **Maximizing sequence matches over the entire length of two sequences, by introducing gaps.** • **Used to determine overall similarity, conservation.**	• **Find regions between two sequences that have the strongest similarity, excluding less similar regions.** • **Finding similar domains, motifs, detecting distant homology**

Global alignment - maximizes the number of matches between the query and source sequences along the entire length of both the sequences.
Local alignment - gives the highest scoring local match between both query and sequences.
Optimal alignment - one that exhibits the most correspondences between the query and the source sequences. It is the alignment with the highest score. Biologically meaningful?
Conclusion

- Gaps are inserted in a sequence being aligned to accommodate missing amino acids or nucleotides
- Global and Local alignment is used depending on the required type of pairwise alignment

74- Pairwise Sequence Alignment - III

Background

- Given a pair of sequences (strings of characters)
- Pairwise alignment tells the similarity between the sequences by maximizing the matches

- Sequences can differ from each other in three different ways:
- Substitutions ACGA → AGGA
- Insertions ACGA → ACCGA
- Deletions ACGA → AGA
- No change? *match*
- Insertions or deletions ("indels") result in *gaps* in alignments
- Substitutions result in *mismatches*
- Align 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…….

Conclusion

- Global and local pairwise alignments are two types of alignment and they can lead to different results
- Indels lead to Gaps in alignments
- Substitutions lead to mismatches

75- Identity vs. Similarity

Background

- How to compare two biological sequences?
- How to measure the degree of match between sequences?
- Two concepts:
  - Identity
  - Similarity

- Sequence Identity
- "Identity is the number of nucleotides or amino acids which match exactly between two biological sequences"
- 1: CATGCTT
  2: CATGC

- Calculate the identity between sequences 1 & 2.
- Number of matches = 5
- Smaller Length: Length (1) = 7, Length (2) = 5
- Identity = Num. of matches/Smaller Legth * 100% = 100%

Points to remember

- Gaps are not counted
- Identity measurement is made on the shorter of the two sequences

- Sequence Similarity
- "Result of matching and transforming one sequence to the other by finding the smallest set of edit operations (insertions, deletions, and substitutions)."
- Can be calculated by pairwise sequence alignment.
- For computing similarity between sequences, you will need to first align the two sequences using pairwise sequence alignment!

Conclusion

- Identity is the count of exact matches between two sequences
- Gaps are excluded
- Similarity is the comparison between sequences calculated by using alignment approach

76-85 IN LAST SOME HINTS
86-Introduction to FASTA - I

Background

- Comparison between two sequences can be performed by pairwise sequence alignment

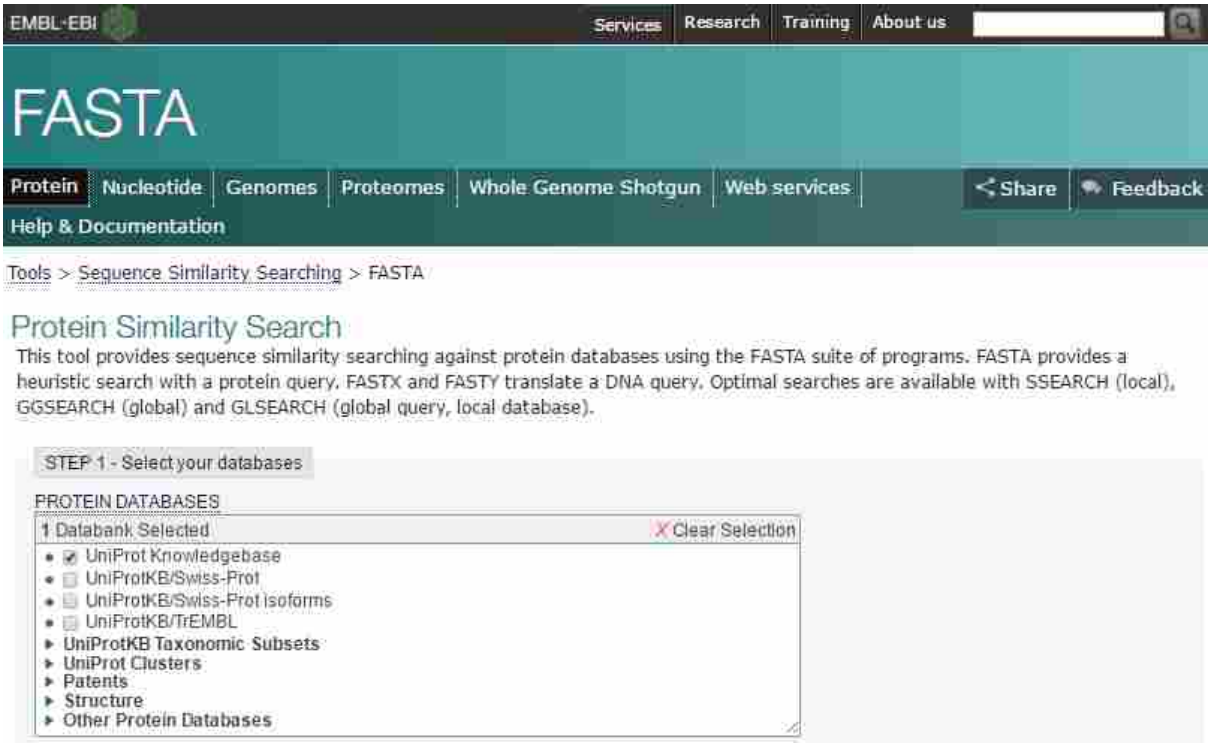
- For comparing multiple sequences, multiple sequence alignment (MSA) is used!
- MSA is a progressive pairwise alignment so as to handle multiple sequences
- Smith-Waterman performs local alignment
- Needleman-Wunsch calculates global alignment The local and global alignment algorithms use Dynamic Programming Approaches
- But, this takes time if there are several sequences that need comparison!

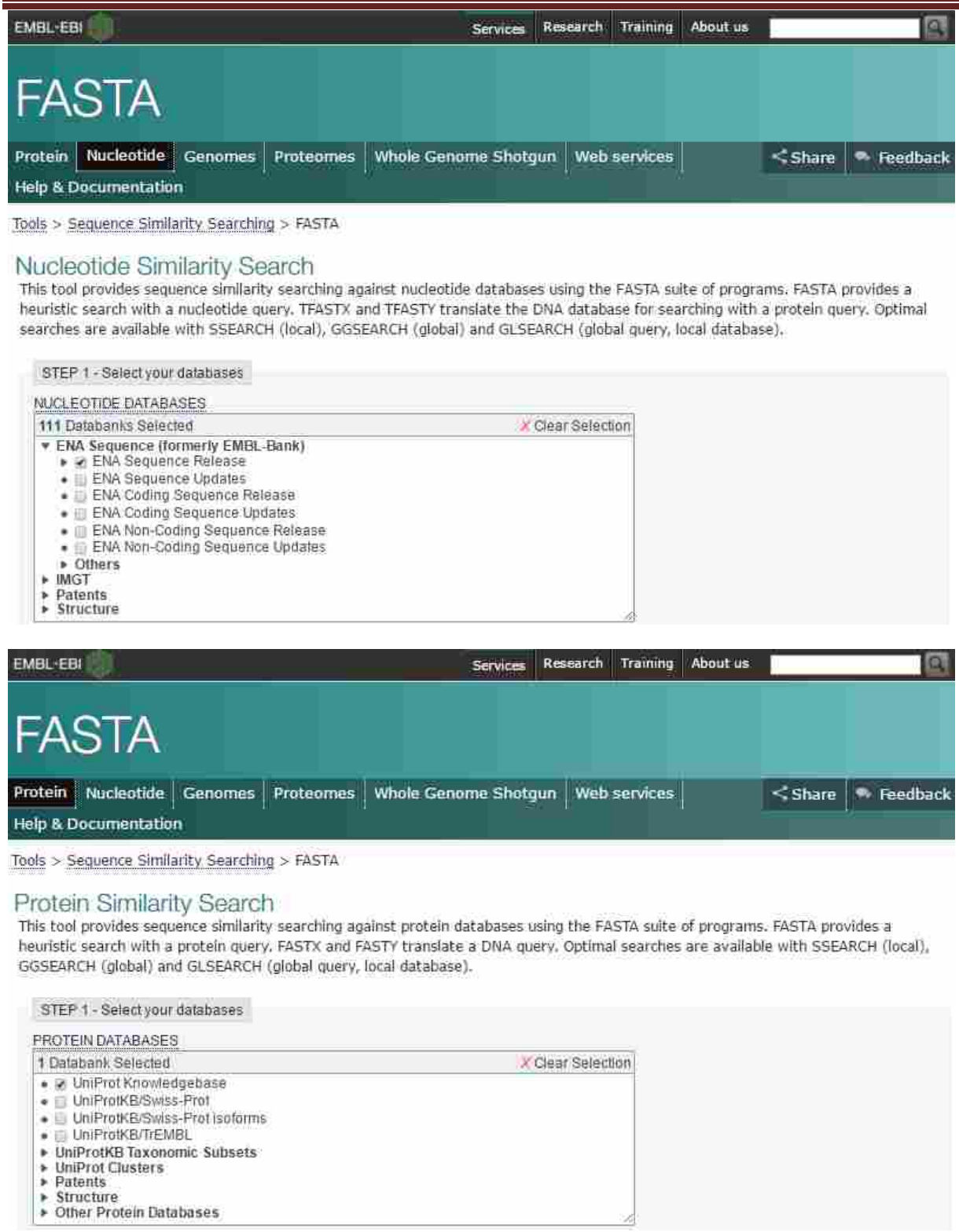
Solution?

- We used BLAST which is an approximate local alignment search tool
- BLAST compares a large number of sequences, quickly!
- FASTA took a similar approach!

FASTA

- Developed in 1988
- "Fast Alignment"
- Searches databases for query protein and nucleotide sequences
- Was later improved upon in BLAST





Conclusion

- FASTA can perform quick comparison of protein and nucleotide sequences
- Can also perform genome and proteome similarity search
- It is available online

87-Introduction to FASTA – II

Introduction

- FASTA – Fast Alignment Algorithm
- Classical global and local alignment algorithms are time consuming
- FASTA achieves alignment by using short lengths of exact matches

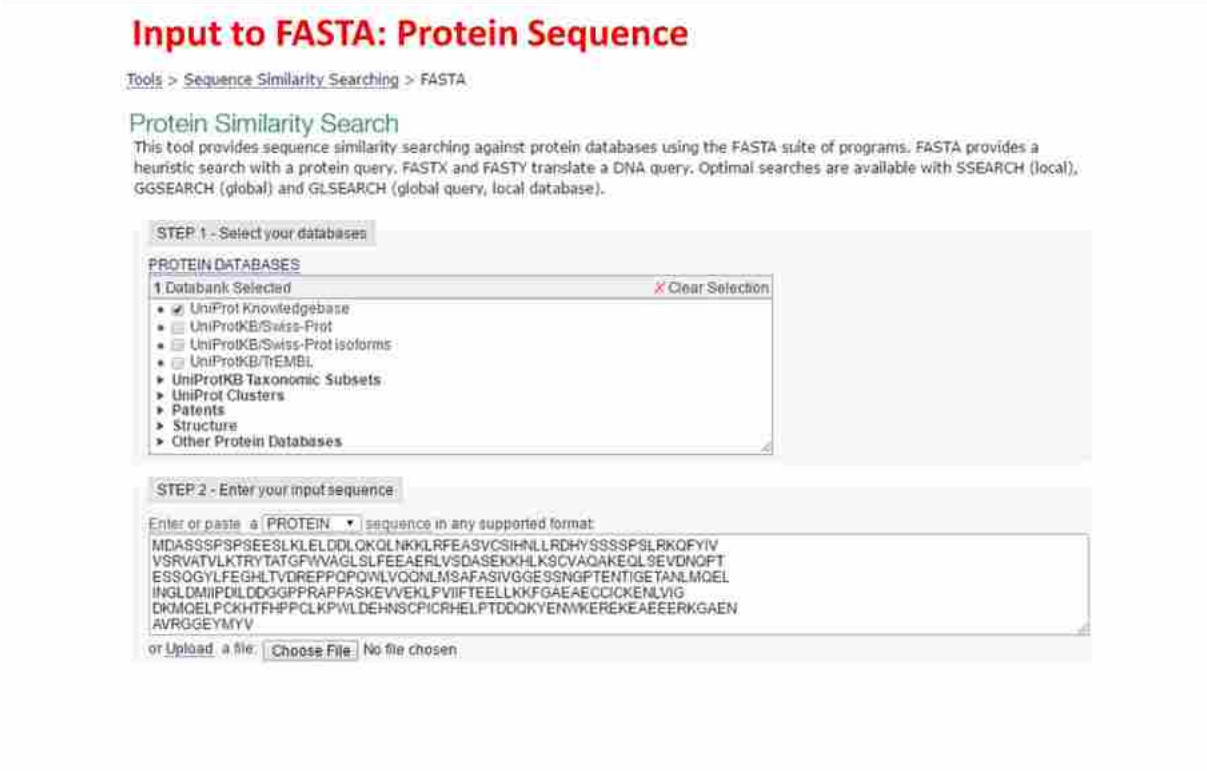
Use Cases of FASTA

- FASTA relies on aligning subsequences of absolute identity
- Input to FASTA search can be in FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP or UniProt formats

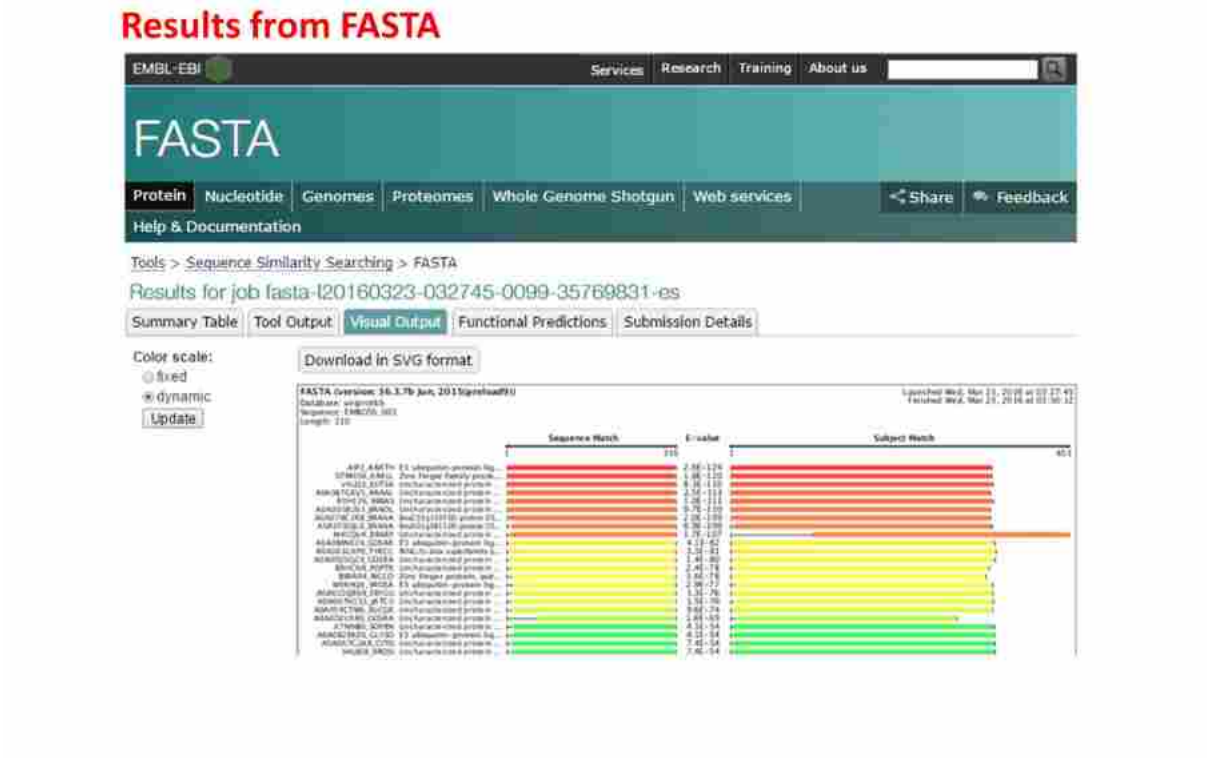
Output of BLAST

- Results are output in visual format along with functional prediction
- A table lists the sequence hits found along with scores
- Users can click on each reported match to look at the details

Introduction to FASTA- II



Introduction to FASTA- II



Conclusion

- Purpose of FASTA
- FASTA vs. Smith Waterman
- Inputs and outputs from FASTA
- Internal working of FASTA?

88-FASTA Algorithm

Introduction

- FASTA can search sequence databases and identify unknown sequences by comparing them to the known sequence databases

- This can help obtain information on the parent organism, function and evolutionary history

Conclusion

- FASTA performs preliminary alignments quickly followed by Smith-Waterman

- Results are output in visual format along with statistical scores

89-Types of FASTA

Introduction

- FASTA can search sequence databases and compare sequences to the known sequences in DNA/RNA/Protein Databases

- It performs search on matching words followed by application Smith-Waterman

Types of FASTA

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 (6ORFs) to a protein sequence database

fasts35

- Compare unordered peptides to a protein sequence database

fastm35

- Compare ordered peptides (or short DNA sequences) to a protein (DNA) sequence database

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

90-Summary of FASTA

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

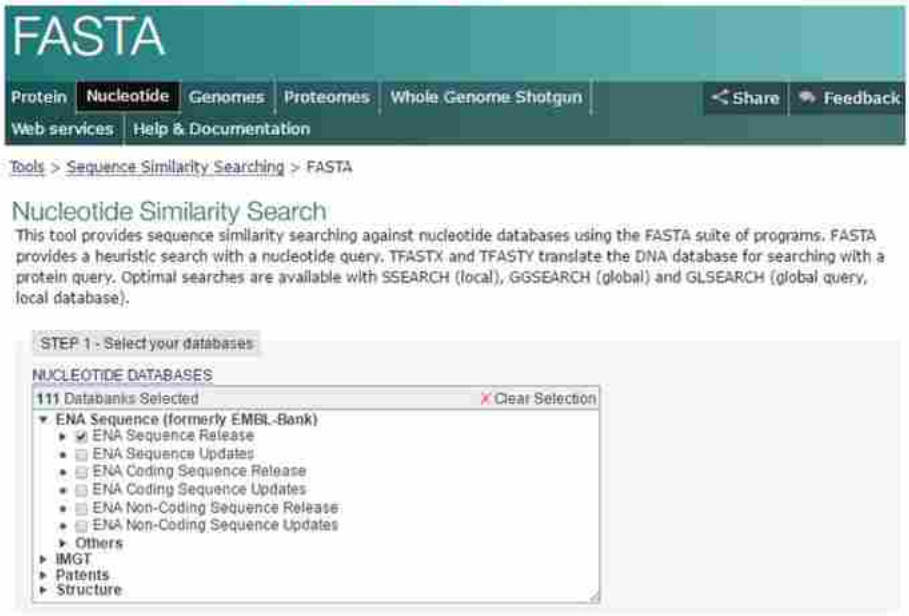
Summary of FASTA

Step 1: Obtain a query sequence

For known sequences: Use NCBI, UCSC etc..
For unknown sequences: Use NGS or Mass Spectrometry

Summary of FASTA

Step 2: Choose a type of FASTA



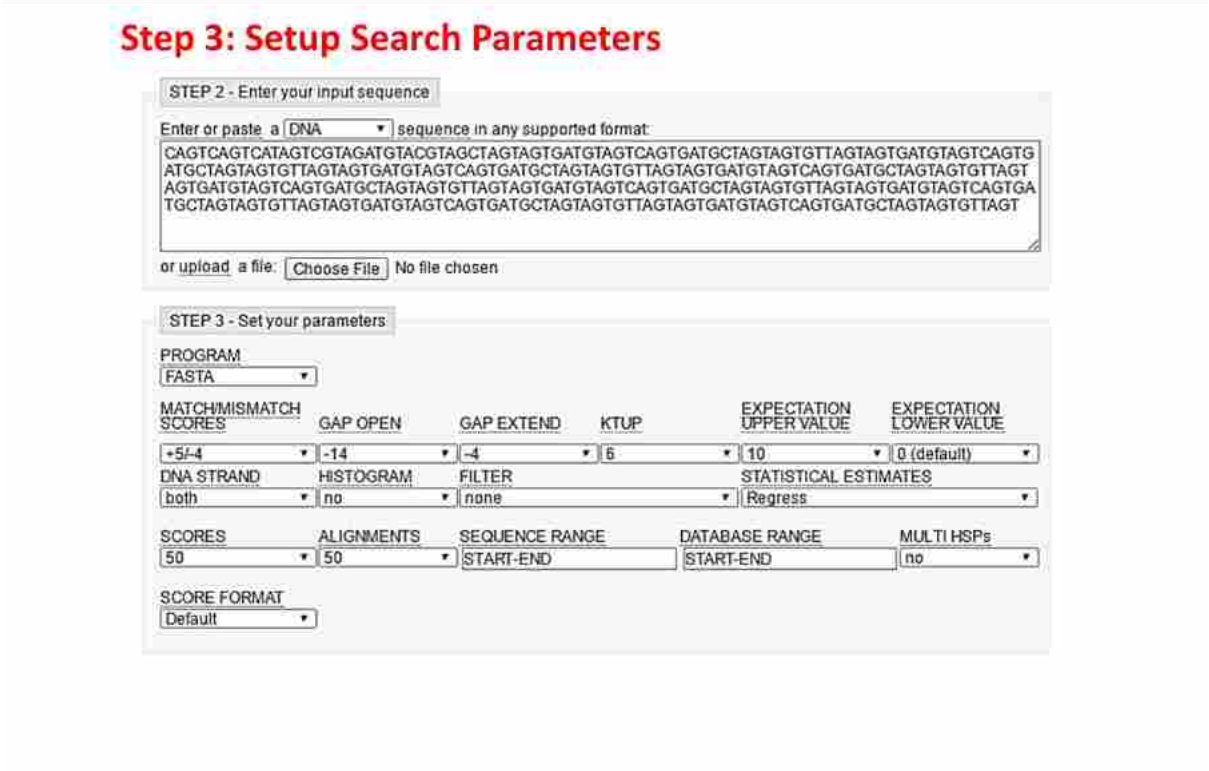
Summary of FASTA

Step 2: Type of FASTA

http://fasta.bioch.virginia.edu/fasta_docs/fasta35.shtml

`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.

Summary of FASTA



Conclusion

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

91- Database Management System

Reading Material

"Database Systems Principles, Design and Implementation" written by Catherine Ricardo, Maxwell Macmillan.	Chapter 1.

Overview of Lecture

- Introduction to the course
- Database definitions

- Importance of databases
- Introduction to File Processing Systems
- Advantages of the Database Approach

Intr oduction to thecourse

This course is first (fundamental) course on database management systems. The course discusses different topics of the databases. We will be covering both the theoretical and practical aspects of databases. As a student to have a better understanding of the subject, it is very necessary that you concentrate on the concepts discussed in the course.

Areas to be covered in this Course:

- Database design and application development: How do we represent a real-world system in the form of a database? This is one major topic covered in this course. It comprises of different stages, we will discuss all these stages one by one.

- Concurrency and robustness: How does a DBMS allow many users to access data concurrently, and how does it protect against failures?

- Efficiency and Scalability: How does the database cope with large amounts of data?

Study of tools to manipulate databases: In order to practically implement, that is, to perform different operations on databases some tools are required. The operations on databases include right from creating them to add, remove and modify data in the database and to access by different ways. The tools that we will be studying are a manipulation language (SQL) and a DBMS (SQL Server).

Database definitions:

Definitions are important, especially in technical subjects because definition describes very comprehensively the purpose and the core idea behind the thing. Databases have been defined differently in literature. We are discussing different definitions here, if we concentrate on these definitions, we find that they support each other and as a result of the understanding of these definitions, we establish a better understanding of use, working and to some extent the components of a database.

Def 1: A shared collection of logically related data, designed to meet the information needs of multiple users in an organization. The term database is often erroneously referred to as a synonym for a "database management system (DBMS)". They are not equivalent and it will be explained in the next section.

Def 2: A collection of data: part numbers, product codes, customer information, etc. It usually refers to data organized and stored on a computer that can be searched and retrieved by a computer program.

Def 3: A data structure that stores metadata, i.e. data about data. More generally we can say an organized collection of information.

Def 4: A collection of information organized and presented to serve a specific purpose. (A telephone book is a common database.) A computerized database is an updated, organized file of machine readable information that is rapidly searched and retrieved by computer.

Def 5: An organized collection of information in computerized format.

Def 6: A collection of related information about a subject organized in a useful manner that provides a base or foundation for procedures such as retrieving information, drawing conclusions, and making decisions.

Def 7: A Computerized representation of any organizations flow of information and storage of data.

Each of the above given definition is correct, and describe database from slightly variant perspectives. From exam point of view, anyone will do. However, within this course, we will be referring first of the above definitions more frequently, and concepts discussed in the definition like, logically related data, shared collection should be clear. Another important thing that you should be very clear about is the difference between database and the database management system (DBMS). See, the database is the collection of data about anything, could be anything. Like cricket teams, students, busses, movies, personalities, stars, seas, buildings, furniture, lab equipment, hobbies, hotels, pets, countries, and many more anything about which you

want to store data. What we mean by data; simply the facts or figures. Following table shows the things and the data that we may want to store about them:

Thing	Data (Facts or figures)
Cricket Player	Country, name, date of birth, specialty, matches played, runs etc.
Scholars	Name, data of birth, age, country, field, books published etc.
Movies	Name, director, language (Punjabi is default in case of Pakistan) etc.
Food	Name, ingredients, taste, preferred time, origin, etc.
Vehicle	Registration number, make, owner, type, price, etc.

There could be infinite examples, and please note that the data that is listed about different things in the above table is not the only data that can be defined or stored about these things. As has been explained in the definition one above, there could be so many facts about each thing that we are storing data about; what exactly we will store depends on the perspective of the person or organization who wants to store the data. For example, if you consider food, data required to be stored about the food from the perspective of a cook is different from that of a person eating it. Think of a food, like, Karhahi Ghost, the facts about Karhahi ghosht that a cook will like to store may be, quantity of salt, green and red chilies, garlic, water, time required to cook and like that. Where as the customer is interested in chicken or meat, then black or red chilies, then weight, then price and like that. Well, definitely there are some things common but some are different as well. The thing is that the perspective or point of view creates the difference in what we store; however, the main thing is that the database stores the data.

The database management system (DBMS), on the other hand is the software or tool that is used to manage the database and its users. A DBMS consist of different components or subsystem that we will study about later. Each subsystem or component of the DBMS performs different function(s), so a DBMS is collection of different programs but they all work jointly to manage the data stored in the database and its users. In many books and may be in this course sometimes database and database management system are used interchangeably but there is a clear difference and we should be clear about them. Sometimes another term is used, that is, the database system, again, this term has been used differently by different people, however in this course we use the term database system as a combination of database and the database management system. So database is collection of data, DBMS is tool to manage this data, and both jointly are called database system.

Importance of the Databases

Databases are important; why? Traditionally computer applications are divided into commercial and scientific (or engineering) ones. Scientific applications involve more computations, that is, different type of calculations that vary from simple to very complex. Today such applications exist, like in the fields of space, nuclear, medicine that take hours or days of computations on even computers of the modern age. On the other hand, the applications that are termed as commercial or business applications do not involve much computations, rather minor computation but mainly they perform the input/output operations. That is, these applications mainly store the data in the computer storage, then access and present it to the users in different formats (also termed as data processing) for example, banks, shopping, production, utilities billing,

customer services and many others. As is clear from the example systems mentioned, the commercial applications exist in the day to day life and are related directly with the lives of common people. In order to manage the commercial applications more efficiently databases are the ultimate choice because efficient management of data is the sole objective of the databases. So such applications are being managed by databases even in a developing country like Pakistan, yet to talk about the developed countries. This way databases are related directly or indirectly almost every person in society.

Databases are not only being used in the commercial applications rather today many of the scientific/engineering application are also using databases less or more. databases are concern of the effectively latter form of applications are more Commercial applications involve The goal of this course is to present an in-depth introduction to databases, with an emphasis on how to organize information in the database and to maintain it and retrieve it efficiently, that is, how to design a database and use it effectively.

Databases and Traditional File Processing Systems

Traditional file processing system or simple file processing system refers to the first computer-based approach of handling the commercial or business applications. That is why it is also called a replacement of the manual file system. Before the use computers, the data in the offices or business was maintained in the files (well in that perspective some offices may still be considered in the pre-computer age). Obviously, it was laborious, time consuming, inefficient, especially in case of large organizations. Computers, initially designed for the engineering purposes were though of as blessing, since they helped efficient management but file processing environment simply transformed manual file work to computers. So processing became very fast and efficient, but as file processing systems were used, their problems were also realized and some of them were very severe as discussed later.

It is not necessary that we understand the working of the file processing environment for the understanding of the database and its working. However, a comparison between the characteristics of the two definitely helps to understand the advantages of the databases and their working approach. That is why the characteristics of the traditional file processing system environment have been discussed briefly here.

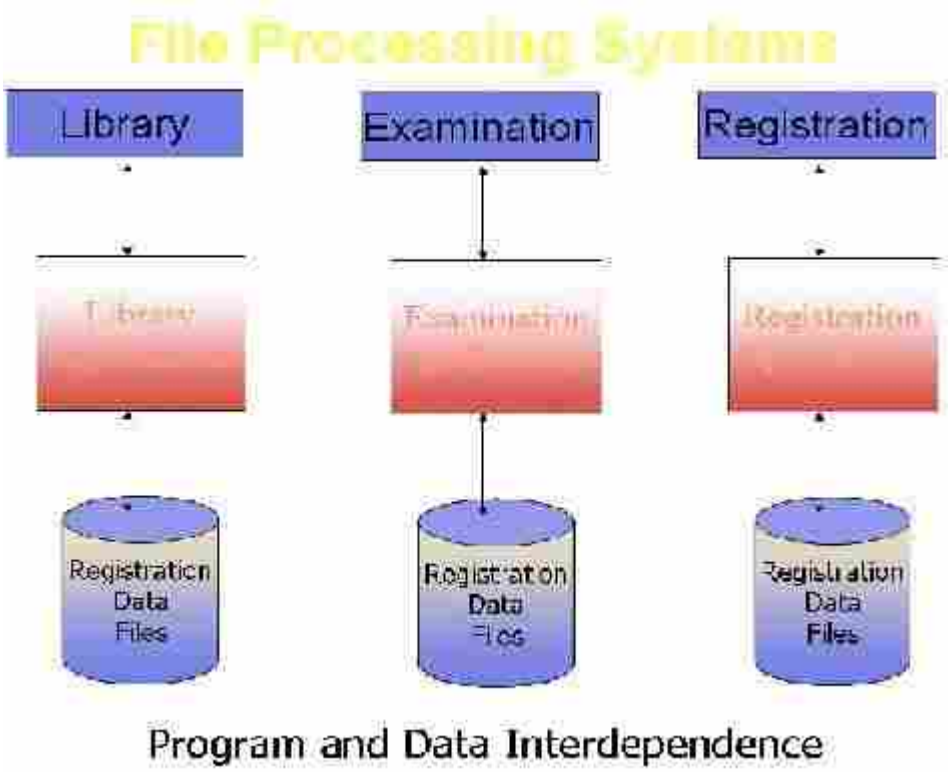


Fig. 1: A typical file processing environment

The diagram presents a typical traditional file processing environment. The main point being highlighted is the program and data interdependence, that is, program and data depend on each other, well they depend too much on each other. As a result any change in one affects the other as well. This is something that makes a change very painful or problematic for the designers or developers of the system. What do we mean by change and why do we need to change the system at all. These things are explained in the following.

The systems (even the file processing systems) are created after a very detailed analysis of the requirements of the organizations. But it is not possible to develop a system that does not need a change afterwards. There could be many reasons, mainly being that the users get the real taste of the system when it is established. That is, users tell the analysts or designers their requirements, the designers design and later develop the system based on those requirements, but when system is developed and presented to the users, it is only then they realize the outcome of the effort. Now it could be slightly and (unfortunately) sometimes very different from what they expected or wanted it to be. So the users ask changes, minor or major. Another reason for the change is the change in the requirements. For example, previously the billing was performed in an organization on the monthly basis, now company has decided to bill the customers after every ten days. Since the bills are being generated from the computer (using file processing system), this change has to be incorporated in the system. Yet another example is that, initially bills did not contain the address of the customer, now the company wants the address to be placed on the bill, so here is change. There could be many more examples, and it is so common that we can say that almost all systems need changes, so system development is always an on-going process.

So we need changes in the system, but due to program-data interdependence these changes in the systems were very hard to make. A change in one will affect the other whether related or not. For example, suppose data about the customer bills is stored in the file, and different programs use this file for different purposes, like adding data into the bills file, to compute the bill and to print the bill. Now the company asks to add the customers' address in the bills, for this we have to change the structure of the bill file and also the program that prints the bill. Well, this was necessary, but the painful thing is that the other programs that are using these bills files but are not concerned with the printing of the bills or the change in the bill will also have to be changed, well; this is needless and causes extra, unnecessary effort.

Another major drawback in the traditional file system environment is the non-sharing of data. It means if different systems of an organization are using some common data then rather than storing it once and sharing it, each system stores data in separate files. This creates the problem of redundancy or wastage of storage and on the other hand the problem on inconsistency. The change in the data in one system sometimes is not reflected in the same data stored in other system. So different systems in organization;

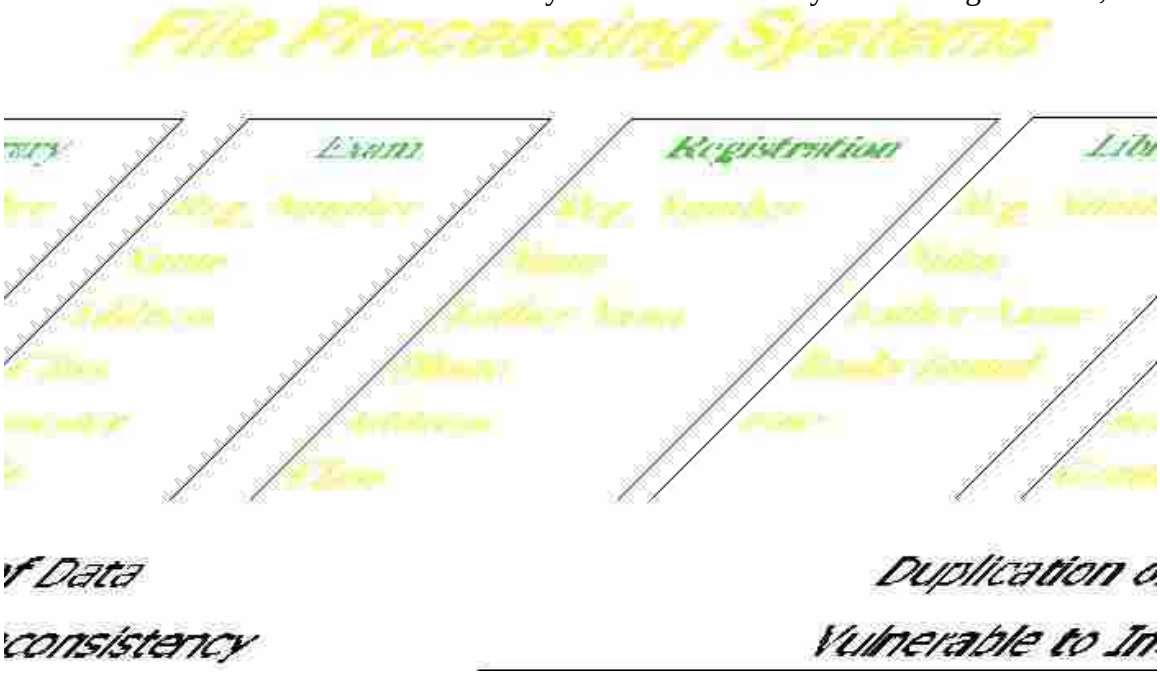


Fig. 2: Some more problems in File System Environment

Previous section highlighted the file processing system environment and major problems found there. The following section presents the benefits of the database systems.

Advantages of Databases

It will be helpful to reiterate our database definition here, that is, database is a shared collection of logically related data, designed to meet the information needs of multiple users in an organization. A typical database system environment is shown in the figure 3 below:

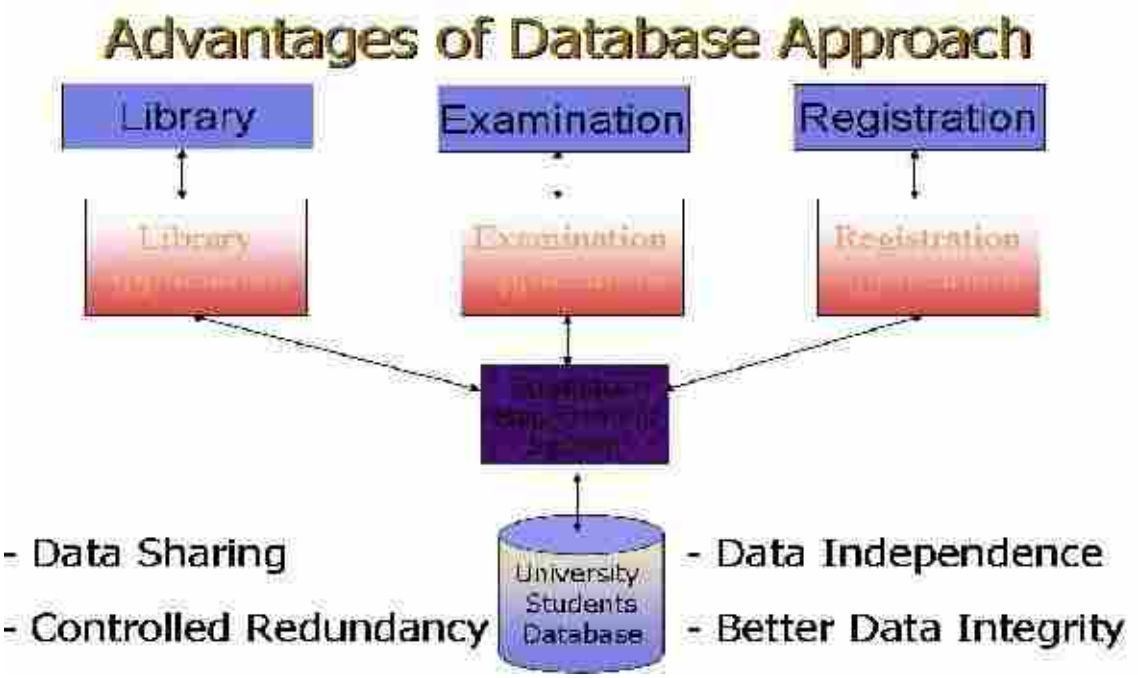


Fig. 3: A typical Database System environment

The figure shows different subsystem or applications in an educational institution, like library system, examination system, and registration system. There are separate, different application programs for every application or subsystem. However, the data for all applications is stored at the same place in the database and all application programs, relevant data and users are being managed by the DBMS. This is a typical database system environment and it introduces the following advantages:

- Data Sharing
  The data for different applications or subsystems is placed at the same place. This introduces the major benefit of data sharing. That is, data that is common among different applications need not to be stored repeatedly, as was the case in the file processing environment. For example, all three systems of an educational institution shown in figure 3 need to store the data about students. The example data can be seen from figure 2. Now the data like registration number, name, address, father name that is common among different applications is being stored repeatedly in the file processing system environment, where as it is being stored just once in database system environment and is being shared by all applications. The interesting thing is that the individual applications do not know that the data is being shared and they do not need to. Each application gets the impression as if the data is being for stored for it. This brings the advantage of saving the storage along with others discussed later.

- Data Independence
  Data and programs are independent of each other, so change is once has no or minimum effect on other. Data and its structure is stored in the database where as application programs manipulating this data are stored separately, the change in one does not unnecessarily effect other.

- Controlled Redundancy
  Means that we do not need to duplicate data unnecessarily; we do duplicate data in the databases, however, this duplication is deliberate and controlled.

- Better Data Integrity
  Very important feature; means the validity of the data being entered in the database. Since the data is being placed at a central place and being managed by the DBMS, so it provides a very conducive to check or ensure that the data being entered into the database is actually valid. Integrity of data is very important, since all the processing and the information produced in return are based on the data. Now if the data entered is not valid, how can we be sure that the processing in the database is correct and the results or the information produced is valid? The businesses make decisions on the basis of information produced from the database and the wrong information leads to wrong decisions, and business collapse. In the database system environment, DBMS provides many features to ensure the data integrity, hence provides more reliable data processing environment.

Dear students, that is all for this lecture. Today we got the introduction of the course, importance of the databases. Then we saw different definitions of database and studied what is data processing then studied different features of the traditional file processing environment and database (DB) system environment. At the end of lecture we were discussing the advantages of the DB approach. There some others to be studied in the next lecture. Suggestions are welcome.

Exercises

- Think about the data that you may want to store about different things around you
- List the changes that may arise during the working of any system, lets say Railway Reservation System

Thanks and good luck

92- Database Management Systems

Overview of lecture-1

- **Database definitions**
- **File processing systems**
- **Advantages of database**
- **More advantages**
- **Some costs**
- **Levels of data**
- **Database users**

Data & Information

Company: Super Soft Dept: Sales

Emp Name	Age	Salary
Malik Sharif	23	55
Sh. M. Akmal	24	55
M. A. Butt	20	40
Malik Junaid	19	20

- **Schema**
- **Database Applications**
- **Database Management System (DBMS)**
- **Data consistency**
- **Better data security**
- **Faster development of new applications**
- **Economy of scale**
- **Better concurrency control**
- **Better backup and recovery procedures**
- **BUT**

Its not always just the

SUGAR

Disadvantages

- **Higher costs**
- **Conversion cost**
- **More difficult recovery**

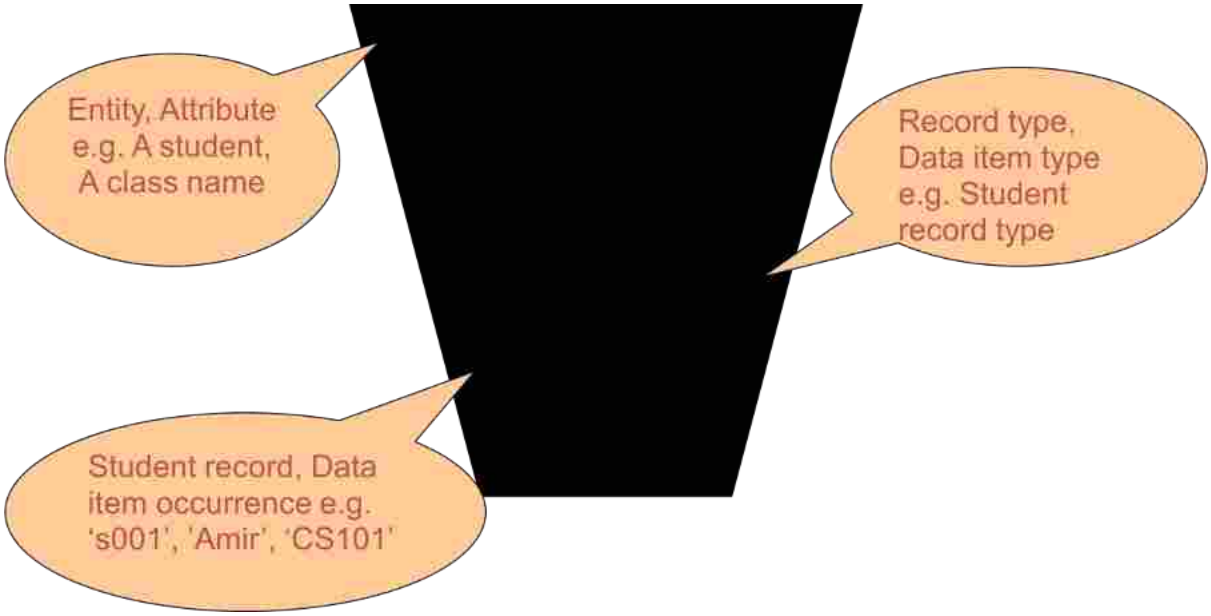
Data As Resource

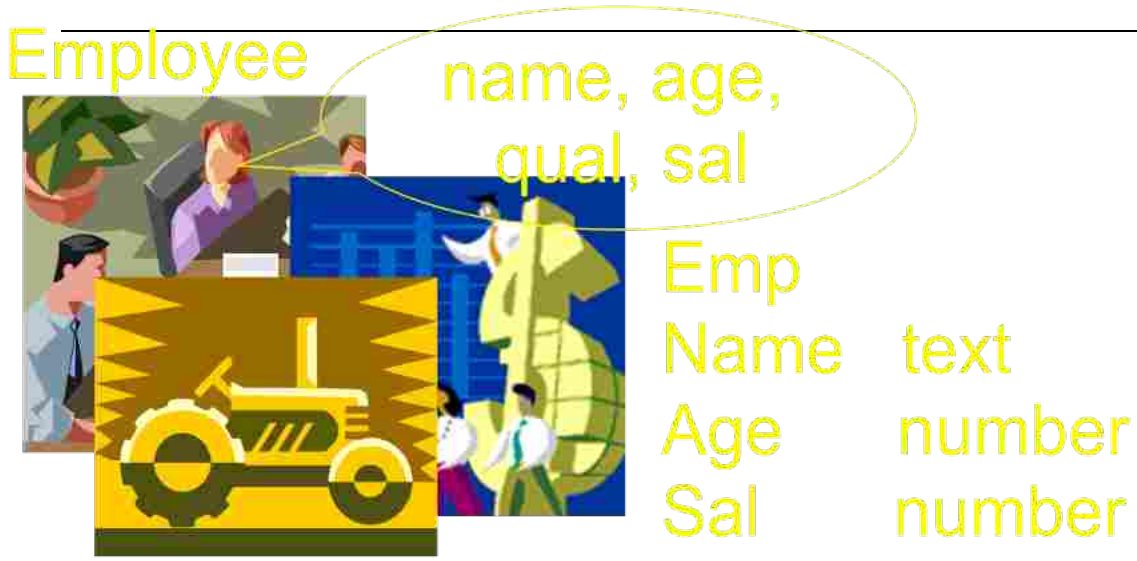
Resource

Any asset that is of value to an organization and that incurs cost

Is data a resource ? YES

YES





Malik Sharif	23	55
Sh. M. Akmal	24	55
M. A. Butt	20	40
Malik Junaid	19	20

Levels of Data

- **Real-world data**
- **Metadata**
- **Data Occurrence**

Database Users

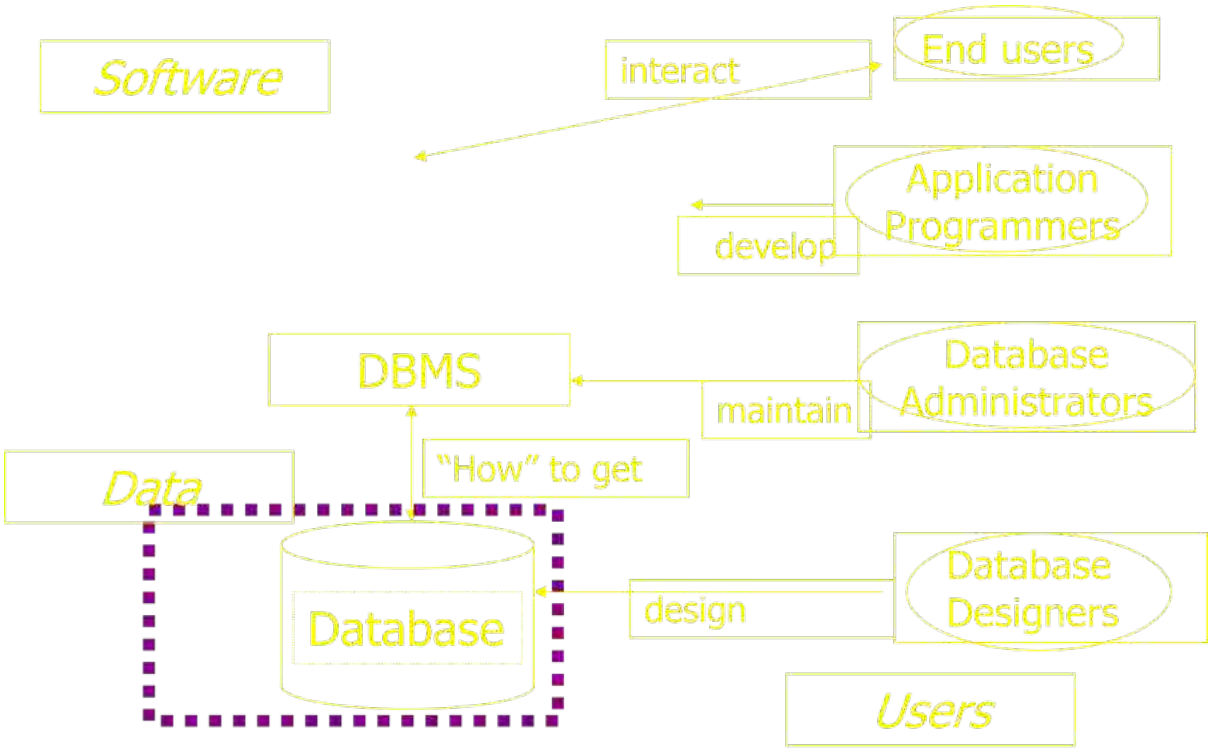
- **Application Programmers**
- **End Users**
  - **Naïve**
  - **Sophisticated**
- **Database Administrator (DBA)**

  A person who has central control over data and programs that access this data

Functions of DBA

- **Schema definition**
- **Granting data access**
- **Routine Maintenance**
  - **Backups**
  - **Monitoring disk space**
  - **Monitoring jobs running**

Typical Components



Database

Architecture

History

- **Initiates in 1960s; IMS, IDS**
- **Conference on DAta Systems Languages (CODASYL)**
- **Data Base Task Group (DBTG)**

- **General architecture in 1971**
- **DBTG's two layered architecture having**
  - **System View- Schema**
  - **User View- Sub rschemas**
- **American National Standards Institute (ANSI)**
- **Standards Planning and Requirements Committee (SPARC)**
- **ANSI-SPARC's 3-layered Architecture**

Summary

- **Basic Terminology**
- **Pros n Cons of DB Env**
- **Levels of Data**
- **Types of Users**
- **History of 3-L Architecture**

CS101 Introduction to Computing

During the last lecture …

(Intelligent Systems)

- We looked at the distinguishing features of intelligent systems w.r.t. other software systems

- We looked at the role of intelligent systems in scientific, business, consumer and other applications

nt systems

(Artificial) Intelligent Systems

- SW programs or SW/HW systems designed to perform *complex* tasks employing strategies that mimic some aspect of human thought

Not a Suitable Hammer for All Nails!

if the nature of computations required in a task is not well understood

or there are too many exceptions to the rules

or known algorithms are too complex or inefficient

then AI has the potential of offering an acceptable solution

SelectedApplications

- Games: Chess, SimCity
- Image recognition
- Medical diagnosis
- Robots

Neural Networks (1)

- Original inspiration was the human brain; emphasis now on usefulness as a computational tool

GeneticAlgorithms (1)

- Based on Darwin's evolutionary principle of ‘survival of the fittest’
- GAs require the ability to recognize a good solution, but not how to get to that solution

Rulebased Systems (1)

- Based on the principles of the logical reasoning ability of humans

Fuzzy Logic (1)

- Based on the principles of the approximate reasoning faculty that humans use when facedwith linguistic ambiguity

The Right Technique

- Selection of the right AI technique requires intimate knowledge about the problem as wellas the techniques under consideration

- Real problems may require a combination of techniques (AI and/or nonAI) for an optimal solution

Three exciting areas areas ofAI applications

Robotics

- Automatic machines that perform various tasksthat were previously done by humans

Autonomous WebAgents (1)

- Computer program that performs various actions continuously, autonomously on behalf oftheir principal!

Decision Support Systems

- Interactive software designed to improve the decision-making capability of their users

- The do not make decisions - just assist in the process

Today'sGoals:

(Data Management)

- First of a two-lecture sequence
- Today we will become familiar with the issuesand problems related to data-intensive computing
- We will find out about flat-files, the simpleast databases
- Next time, in our 4th lecture on productivity software, we will discuss relational databaase

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

BholiBooks.com (1)

- Consider the situation of a large, online bookstore

- They have an inventory of millions of books, with new titles constantly arriving, and old onesbeing phased out on a regular basis

- The price for a book is not a static feature; it varies every once in a while

BholiBooks.com(2)

- Thousands of books are shipped each day, changing the inventory constantly
- Some are returned, again changing the inventory situation constantly
- The cost of each shipped order depends on:
  - Prices of individual books
  - Size of the order
  - Location of the customer

- For each order, the customer's particulars –_ name, address, phone number, credit card number – are required

- Generally, that data is not deleted after the completion of the transaction; instead, it is keptfor future reference

- All the transaction activity and the inventory changes result in:
  - Thousands of data items changing every day
  - Thousands of additional data items being added everyday

- Keeping track & taking care (i.e. management) of all that constantly changing and expanding data is not a trivial task and requires disciplinedattention

Issues in Data Management

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

Data Entry

- New titles are added every day
- New customers are being added every day

- Some of the above *may* require manual entry ofnew data into the computer systems
- That new data needs to be added accurately

Data Updates (1)

- Old titles are deleted on a regular basis
- Inventory changes every instant
- Book prices change
- Shipping costs change
- Customers' personal data change
- Various discount schemes are alwaysc

Data Updates (2)

- All those actions require updates to existingdata
- Those changes need to be entered accurately
- That can also be achieved by user-interfacesthat prevent the input of invalid data

Data Security (1)

- All the data that BholiBooks has in its computer systems is quite critical to its operation
- The security of the customers' personal data is of utmost importance. Hackers are always looking for that type of data, especially for creditcard numbers
- Enough leaks of that type, and customers will stop doing business with BholiBooks

Data Security (2)

- This problem can be managed by using appropriate security mechanisms that provide access to authorized persons/computers only

- Security can also be improved through:
  - Encryption
  - Private or virtual-private networks
  - Firewalls
  - Intrusion detectorss

Data Integrity

- Integrity refers to maintaining the correctnessand consistency of the data
  - Correctness: Free from errors
  - Consistency: No conflict among related data items
- Integrity can be compromised in many ways:
  - Typing errors
  - Transmission errors
  - Hardware malfunctions
  - Program bugs

- Viruses
- Fire, fd

Ensuring Data Integrity (1)

- *Type Integrity* is implemented by specifying the type of a data item:
  - Example: A credit card number consists of 12 digits.An update attempting to assign a value with more orfewer digits or one including a non-numeral should be rejected

- *Limit Integrity* is enforced by limiting the valuesof data items to specified ranges to prevent illegal valueAge of person should not be negative

Ensuring Data Integrity (2)

- *Referential Integrity* requires that an item referenced by the data for some other itemmust itself exist in the database
  - Example: If an airline reservation is requested for a particular flight, then the corresponding flight number must actually exist

- *Physical Integrity* is ensured through hardware redundancy, backups, et

DataAccessibility (1)

- If the transaction and inventory data is placed ina disorganized fashion on a hard disk, it becomes very difficult to later search for a stored data item

- What is required is that:
  - Data be stored in an organized manner
  - Additional info about the data be stored

so that the data access times are minimized

DataAccessibility (2)

- What if two customers check on the availabilityof a certain title simultaneously?
- On seeing its availability, they both order thetitle – for which, unfortunately, only a single copy is available
- Same is the case when two airline customerstry booking the only available s

DataAccessibility (3)

- A solution to this *concurrency control* problem: Lock access to data while someone is using it

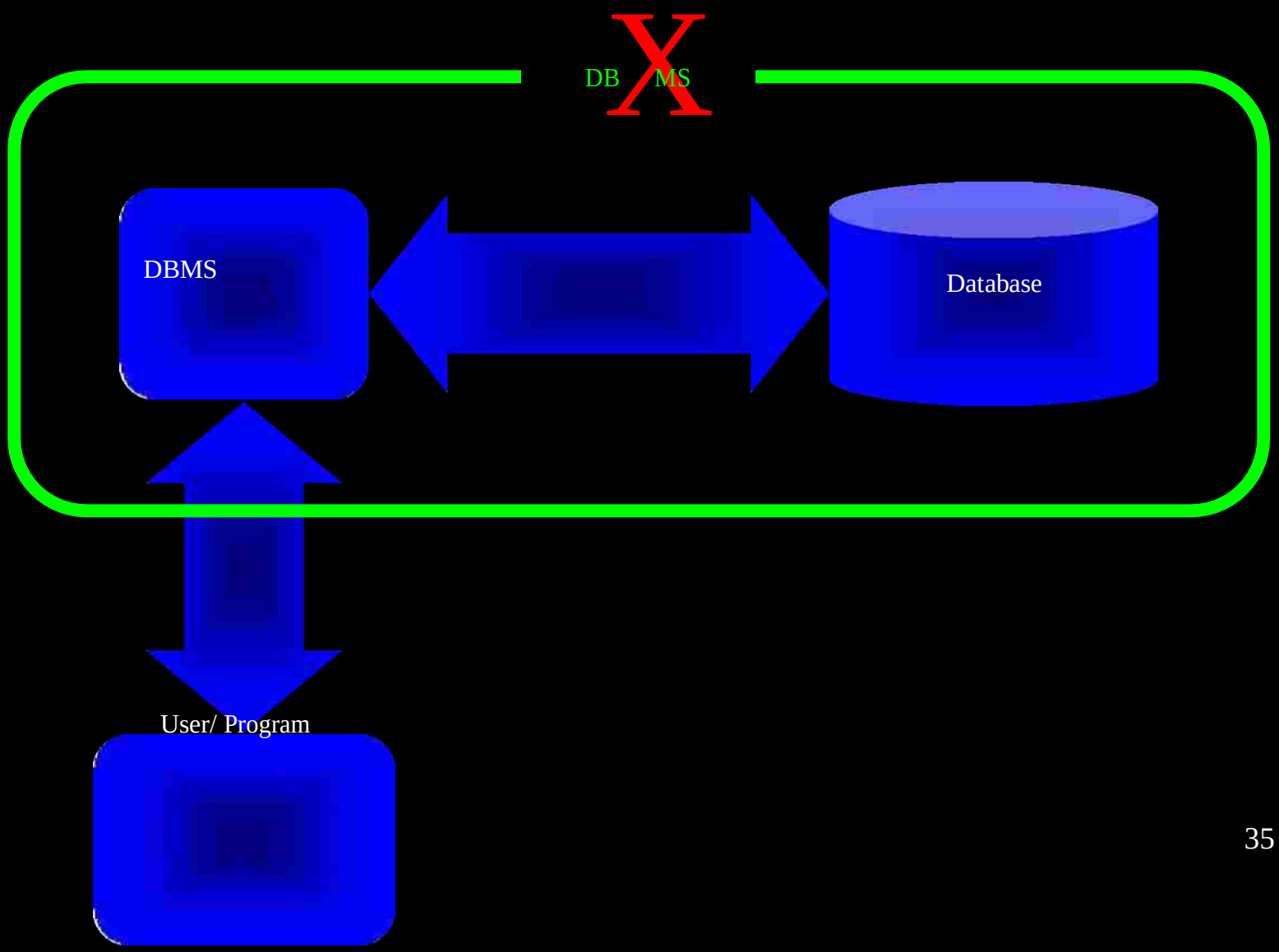
We can write our own SW that can take care of all the issues that we just discussed

OR

We can save ourselves lots of time, cost, and effort by buying ourselves a Database Management System (DBMS) that takes care of most, if not all, of the

DBMS (1)

- 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 bea component of the DBMS, and not an entity outside the DBMS


DB MS
DBMS
Database
User/ Program

DBMS (2)

- A DBMS takes care of the storage, retrieval,and management of large data sets on a database
- It provides SW tools needed to organize & manipulate that data in a flexible manner
- It includes facilities for:
  - Adding, deleting, and modifying data
  - Making queries about the stored data
  - Producing reports summarizing the required

Database (1)

- A collection of data organized in such a fashionthat the computer can quickly search for a desired data item

- All data items in it are generally related to each other and share a single domain

Database (2)

- They allow for easy manipulation of the data
- They are designed for easy modification & reorganization of the information they contain
- They generally consist of a collection of interrelated computer files

Example: VU Student Database

- Student's name
- Student’s photograph
- Father’s name
- Phone number
- Street address
- eMail address
- Courses being taken

Example: BholiBooks' Customer DB

- Name, address, phone & fax, eMail
- Credit card type, number, expiration date
- Shipping preference
- Books on order
- All books that were ever shipped to the customer

Example: BholiBooks' Inventory DB

- Book title, author, publisher, binding, date of publication, price
- Book summary, table of contents
- Customers', editors', newspaper reviews
- Number in stock
- Number on order

OS Independence (1)

- DBMS stores data in a database, which is a collection of interrelated files
- Storage of files on the computer is managed bythe computer OS's file system
- Intimate knowledge of the OS & its file systemis required to provide rapid access to the data

OS Independence (2)

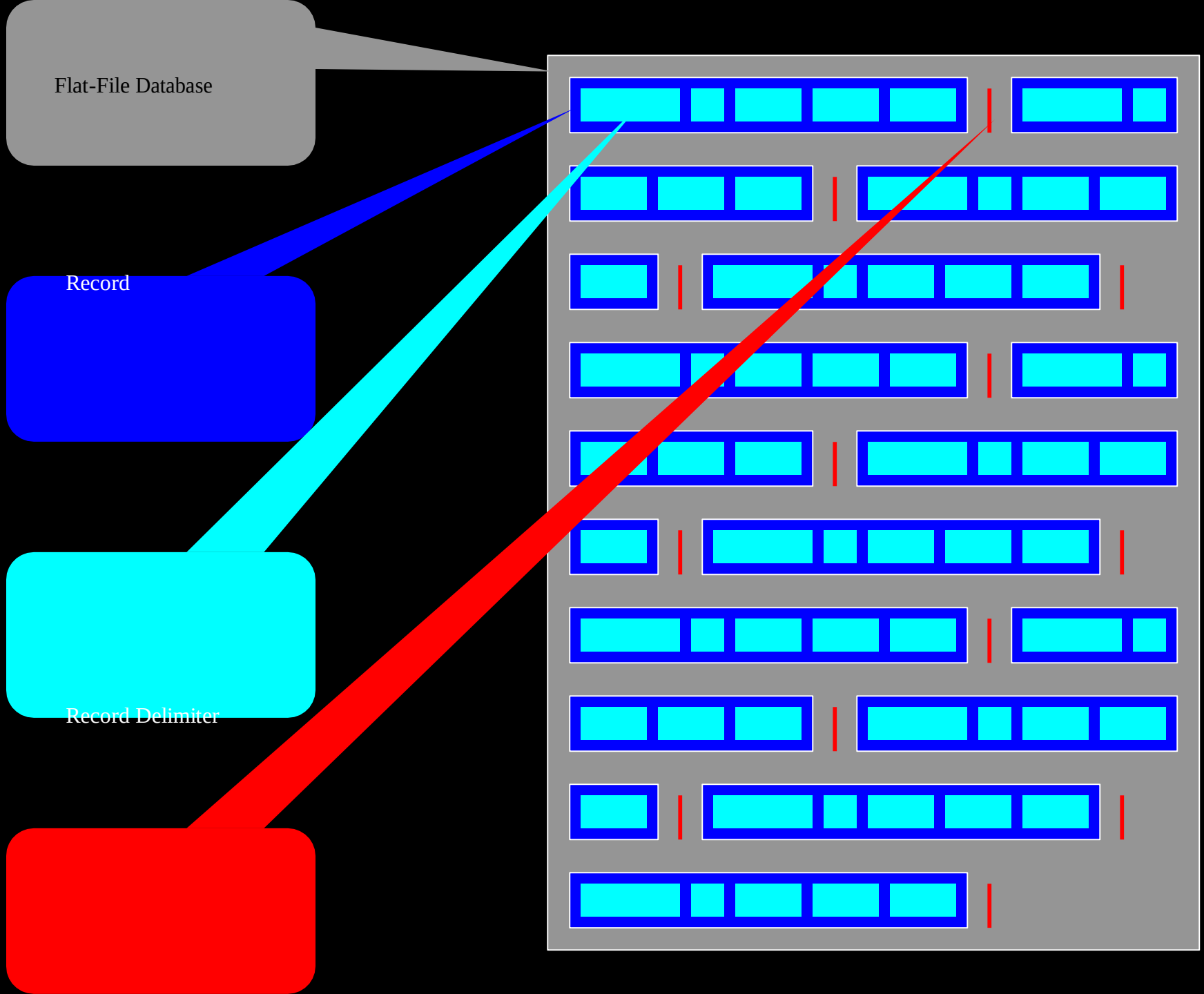
- The DBMS takes care of those details
- It hides the actual storage details of data filesfrom the user
- It provides an OS-independent view of the datato the user, making data manipulation and management much more convenient

What can be stored in a database?

- In the old days, databases were limited to numbers, Booleans, and text
- These days, anything goes
- As long as it is digital data, it can be stored:
  - Numbers, Booleans, text
  - Sounds

In the very, very old days ...

- Even large amounts of data was stored in textfiles, known as *flat-file databases*
- All related info was stored in a single long, tab-or comma-delimited text file
- Each group of info – called a *record* - in that file was separated by a special character; vertical bar '|' was a popular opti


Flat-File Database
Record
Record Delimiter

Title, Author, Publisher, Price, InStock|Good Bye Mr. Bhola, Altaf Khan, BholiBooks, 1000, Y|The Terrible Twins, Bhola Champion, BholiBooks, 199, Y|Calculus & Analytical Geometry, SmithSahib, Good Publishers, 325, N|Accounting Secrets, Zamin Geoffry, Sangg-e-Kilometer Publishers, 29, Y|

The Trouble with Flat-File Databases

- The text file format makes it hard to search for specific information or to create reports that include only certain fields from each record

- Reason: One has to search sequentially through the entire file to gather desired info,such as ‘all books by a certain author’

- However, for small sets of data – say, consi

Consider this tabular approach ...

(same records, same fields, but in a different format)

Title	Author	Publisher	Price	InStock
Good Bye Mr. Bhola	Altaf Khan	BholiBooks	1000	Y
The Terrible Twins	Bhola Champion	BholiBooks	199	Y
Calculus & Analytical Geometry	Smith Sahib	Good Publishers	325	N
Accounting	Zamin	Sung-e-		

Secrets	Geoffry	Kilometer Publishers	29	Y 49

Tabular Storage: Features & Possibilities

1. Similar items of data form a column

2. *Fields* placed in a particular row – same as a flat-file *record* – are strongly interrelated

3. One can sort the table w.r.t. any column

4. That makes searching – e.g., for all the books written by a certain author – straight forwar

5. Similarly, searching for the 10 cheapest/most expensive books can be easily accomplished through a sort

6. Effort required for adding a new field to all the records of a flat-file is much greater thanadding a new column to the table

CONCLUSION: Tabular storage is better than flat-file storage

We will continue on this theme next time

Today's Summary: (Data Management)

- First of a two-lecture sequence
- Today we became familiar with the issues and problems related to data-intensive computing
- We also found out about flat-file and tabular storage

Next Lecture:

(Database SW)

- Next time, in our 4th lecture on productivity SW,we will continue our discussion on data management

- We will find out about relational databases

CS101 Introduction to Computing

Lecture 37

Database Software

1

Focus of the last Lecture was on Data Management

- First of a two-lecture sequence
- We became familiar with the issues and problemsrelated to data-intensive computing
- We also found out about flat-file and tabular storage

2

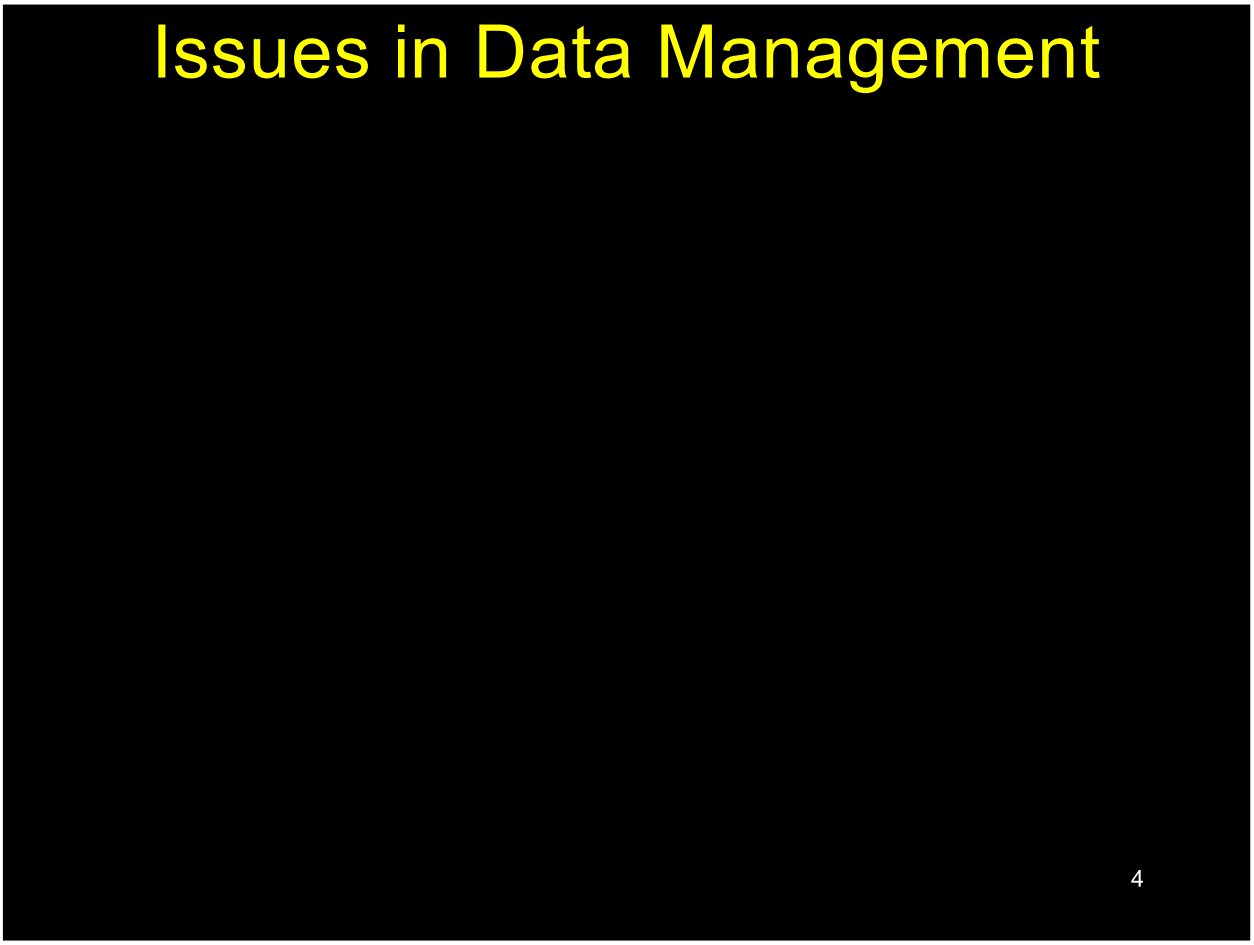
Data Management

Data Management

- Keeping track of a few dozen data itemsis 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

3

Issues in Data Management



Data Entry

Data Entry

- New titlesare added every day
- New customersare being added every day
- That new data needs to be added accurately

5

Data Updates (2)

Data Updates (2)

- All those actions require updatesto existing data
- Those changes need to be entered accurately

6

Data Security (1)

- All the data that BholiBooks has in its computer systems is quite critical to its operation
- The security of the customers' personal datais of utmost importance. Hackersare always looking for that type of data, especially for credit card numbers

7

Data Security (2)

Data Security (2)

- This problem can be managed by using appropriate security mechanismsthat provide access to authorized persons/computersonly
- Security can also be improved through:
  - Encryption
  - Privateor virtual-private networks
  - Firewalls
  - Intrusiondetectors
  - Virusdetectors

8

Data Integrity

Data Integrity

- Integrity refers to maintaining the correctness and consistency of the data
  - Correctness: Free from errors
  - Consistency: No conflict among related data items
- Integrity can be compromisedin many ways:
  - Typingerrors
  - Transmissionerrors
  - Hardwaremalfunctions
  - Program bugs
  - Viruses
  - Fire, flood, etc.

9

Ensuring Data Integrity (1)

Ensuring Data Integrity (1)

- *Type Integrity*
- *Limit Integrity*

10

Ensuring Data Integrity (2)

Ensuring Data Integrity (2)

- *Referential Integrity*
- *Physical Integrity*

11

Data Accessibility (1)

Data Accessibility (1)

- What is required is that:
  - Data be stored in an organized manner
  - Additional infoabout the data be stored

  so that the data access timesare minimized

12

Data Accessibility (3)

Data Accessibility (3)

- A solution to this *concurrency control* problem: Lock access to data while someone is using it

13

DBMS (2)

- A DBMS takes care of the storage, retrieval, and managementof large data setson a database
- It provides SW toolsneeded to organize & manipulate that data in a flexible manner
- It includes facilities for:
  - Adding, deleting, and modifyingdata
  - Making queriesabout the stored data
  - Producing reportssummarizing the required contents

14

Database (1)

Database (1)

- A collection of data organizedin such a fashion that the computer can quickly searchfor a desired data item

15

OS Independence (2)

OS Independence (2)

- It provides an OS-independent view of the data to the user, making data manipulation and management much more convenient

16

What can be stored in a database?

What can be stored in a database?

- As long as it is digital data, it can be stored:
  - Numbers, Booleans, text
  - Sounds
  - Images
  - Video

17

In the very, very old days …

In the very, very old days …

- Even large amounts of data was stored in text files, known as *flat-filedatabases*
- All related info was stored in a single long, tab- or comma-delimitedtext file
- Each group of info– called a *record*- in that file was separated by a special character; vertical bar‘|’ was a popular option
- Each record consisted of a group of *fields*, each field containing some distinct data item

The Trouble with Flat-File Databases

The Trouble with Flat-File Databases

- The text file format makes it hardto search for specific infoor to create reports that include only certain fieldsfrom each record
- Reason: One has to search sequentially through the entire file to gather desired info, such as 'all books by a certain author'
- However, for small sets of data– say, consisting of several tens of kB– they can provide reasonable performance

19

Tabular Storage: Features & Possibilities

Tabular Storage: Features & Possibilities

1. Similar itemsof data form a column

2. *Fields*placed in a particular row – same as a flat-file *record*– are strongly interrelated

3. One can sort the table w.r.t. any column

4. That makes searching– e.g., for all the books written by a certain author – straight forward

20

Tabular Storage: Features & Possibilities

Tabular Storage: Features & Possibilities

5. Similarly, searching for the 10 cheapest/most expensive bookscan be easily accomplished through a sort

6. Effortrequired for adding a new fieldto all the records of a flat-file is much greater than adding a new column to the table

21

CONCLUSION: Tabular storage is better than flat-file storageWe will continue on with tables' theme today

CONCLUSION: Tabular storage is betterthan flat-file storage

We will continueon with tables' theme today

22

Today's Lecture:Database SW

Today's Lecture: Database SW

- In our 4th& final lecture on productivity software, we will continue our discussionfrom last week on data management
- We will find out about relational databases
- We will also implementa simple relational database

23

Let's continue on with the tabular approachWe stored data in a table last time, and liked itLet's revisit that table and then put together another one

Let’s continue on with the tabular approach

We stored data in a table last time, and liked it

Let’s revisit that table and then put together another one

24

Table from the Last Lecture

Table from the Last Lecture

Title	Author	Publisher	Price	InStock
Good Bye Mr. Bhola	Altaf Khan	BholiBooks	1000	Y
The Terrible Twins	Bhola Champion	BholiBooks	199	Y
Calculus & Analytical Geometry	Smith Sahib	Good Publishers	325	N
Accounting Secrets	Zamin Geoffry	Sung-e-Kilometer Publishers	29	Y

25

Another table …

Another table …

Customer	Title	Shipment	Type
Aadil Ali	Good Bye Mr. Bhola	2002.12.26	Air
Aadil Ali	The Terrible Twins	2002.12.26	Air
Miftah Muslim	Calculus & Analytical Geometry	2002.12.25	Surface
Karen Kaur	Good Bye Mr. Bhola	2002.12.24	Air

26

This & the previous table are related

This & the previous table are related

- They sharea column, & are related through it
- A program can match infofrom a field in one table with info in a corresponding fieldof another table to *generate*a 3rd table that combines requested datafrom both tables
- That is, a program can use matching values in 2 tablesto *relate*info in one to infoin the other

27

Q: Who is BholiBooks' best customer?

Q:Who is BholiBooks' best customer?

- That is, who has spent the most moneyon the online bookstore?
- To answer that question, one can process the inventoryand the shipmenttables to generate a third tablelisting the customer namesand the prices of the books that they have ordered

28

The generated table

The *generated* table

Customer	Price
Aadil Ali	1000
Aadil Ali	199
Miftah Muslim	325
Karen Kaur	1000

Can you now process this table to find the answer to our question?

29

Relational Databases (1)

- Databases consisting of two or more related tablesare called *relational databases*
- A typicalrelational database may have anywhere from 10 to over a thousandtables
- Each columnof those tables can contain only a single type of data(contrast this with spreadsheetcolumns!)
- Table rows are called records; row elements are called fields

30

Relational Databases (2)

Relational Databases (2)

- A relational database stores all its data inside tables, and nowhere else
- All operations on dataare done on those tables or those that are generatedby table operations
- Tables, tables, and nothing but tables!

31

RDBMS

RDBMS

- Relational DBMS software
- Contains facilities for creating, populating, modifying, and querying relationaldatabases
- Examples:
  - Access
  - FileMaker Pro
  - SQL Server
  - Oracle
  - DB2
  - Objectivity/DB
  - MySQL
  - Postgres

32

The Trouble with Relational DBs (1)

The Trouble with Relational DBs (1)

- Much of current SW development is done using the object-oriented methodology
- When we want to store the object-oriented data into an RDBMS, it needs to be translatedinto a form suitable for RDBMS

33

The Trouble with Relational DBs (2)

The Trouble with Relational DBs (2)

- Then when we need to read the data backfrom the RDBMS, the data needs to be translated back into an object-oriented form before use
- These two processing delays, the associated processing, and time spent in writing and maintainingthe translation code are the key disadvantages of the current RDBMSes

34

Solution?

Solution?

- Don’t have time to discuss that, but try searching the Web on the following terms:
  - Object-oriented databases
  - Object-relational databases

35

Classification of DBMS w.r.t. Size

Classification of DBMS w.r.t. Size

- Personal/Desktop/Single-user (MB-GB)
  - Examples: Tech. papers' list; Methai shop inventory
  - Typical DMBS: Access
- Server-based/Multi-user/Enterprise (GB-TB)
  - Examples: HBL; Amazon.com
  - Typical DMBS: Oracle, DB2
- Seriously-huge databases (TB-PB-XB)
  - Examples: 2002 – BaBar experiment at Stanford (500TB); 2005 – LHC database at CERN (1XB)
  - Typical DMBS: Objectivity/DB

36

Some Terminology (1)

Some Terminology (1)

- *Primary Key* is a field that uniquely identifies each record stored in a table
- *Queries* are used to view, change, and analyze data. They can be used to:
  - Combine data from different tables, efficiently
  - Extract the exact data that is desired
- *Forms* can be used for entering, editing, or viewing data, one record at a time

37

Some Terminology (2)

Some Terminology (2)

- *Reports* are an effective, user-friendly way of presenting data. All DBMSes provide tools for producing custom reports.

- *Data normalization* is the process of efficiently organizing data in a database. There are two goals of the normalization process:
  - Eliminate redundant data
  - Storing only related data in a table

38

Before we do a demo, let me just mention my favorite database application: Data Mining

Before we do a demo, let me just mention my favorite database application: Data Mining

39

Data Mining

Data Mining

- The process of analyzinglarge databases to identify patterns
- Example: Mining the sales recordsfrom a BholiBooks could identify interesting shopping patterns like “53% of customers who bought book A also bought book B”. This pattern can be put to good use!
- Data mining often utilizes intelligent systems’ techniques

40

Let’s now demonstrate the use of a desktop RDBMS

Let's now demonstrate the use of a desktop RDBMS

- We will create a new relational database
- It will consist of two tables
- We will populate those tables
- We will generate a report after combining the data from the two tables

41

Assignment # 13

Assignment # 13

Develop a database by designing two tables, populate them, and then generate a report

Further information on this assignment will be provided to you on the CS101 Web site

42

Access Tutorial

http://www.microsoft.com/education/DOWNLOADS/tutorials/classroom/office2k/acc2000.doc

43

Today's Lecture:

Today's Lecture:

- In this final lecture on productivity software, we continued our discussion from last week on data management
- We found out about relational databases
- We also implementeda simple relational database

44

Next Lecture' Goals(Cyber Crime)

Next Lecture' Goals (Cyber Crime)

- To know the different typesof computer crimes that occur over cyber space
- To familiarize ourselves with with several methodsthat can be used to minimizethe effect of these crimes
- To get familiar with a few policiesand legislationdesigned to tackle cyber crime

45

95- Nucleotide Sequence Databases

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
- Secondary databases
- Specialized Databases

Classification

Biological databases can also be classified on the bases of types of data as

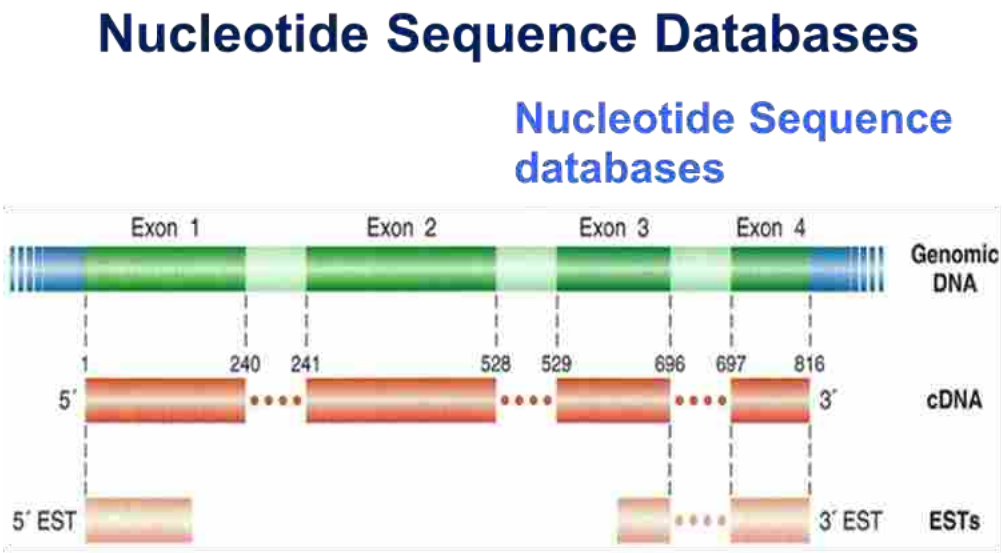
- Nucleotide databases
- Protein databases
- RNA databases
- Genome databases

- Expression databases

Issues

Due to limited Q/A

- Redundancy
- Inconsistency
- Incompatibility (format, terminology, data types, etc.)
- Nucleotide Sequence databases
- Nucleotide Sequence Databases store DNA and cDNA or EST sequences.
- Nucleotide Sequence databases

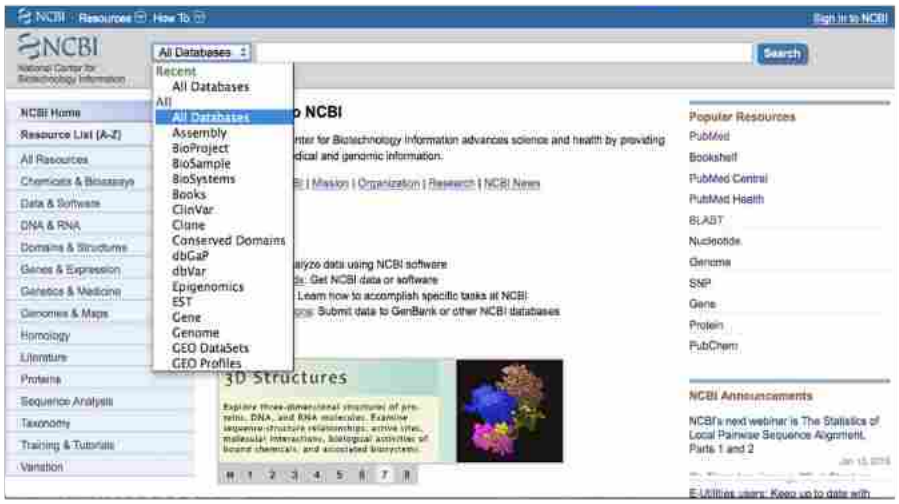


Origin

- First assembled into Genebank (1982) at Los Alamos National Laboratory (LANL), New Maxico by Walter Goad et al
- **GeneBank** is now under NCBI

Nucleotide Sequence Databases

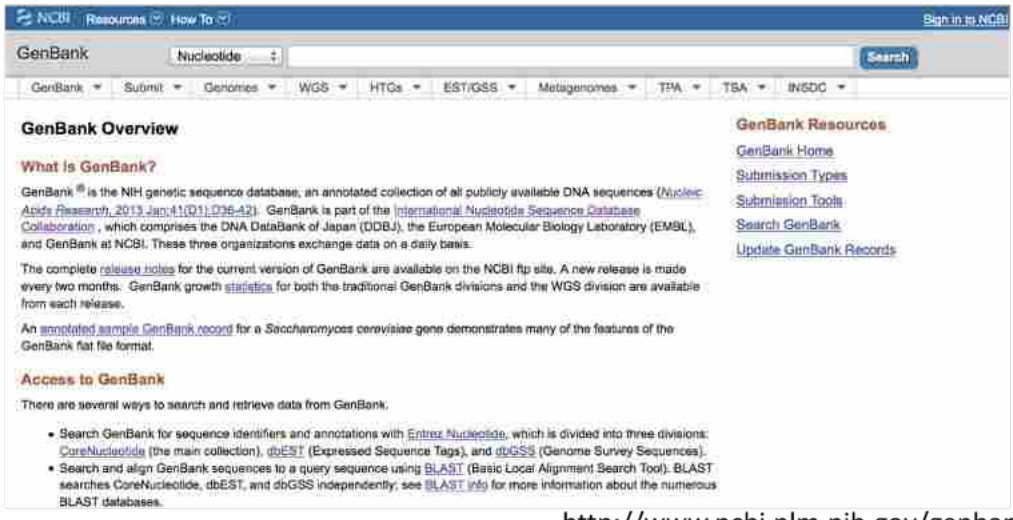
National Center for Biotechnology Information



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

Nucleotide Sequence Databases

GeneBank



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

EMBL and DDBJ

- European Molecular biology Lab (EMBL established1980)
- DNA databank of Japan (DDBJ) established in Mishima japan (1984)

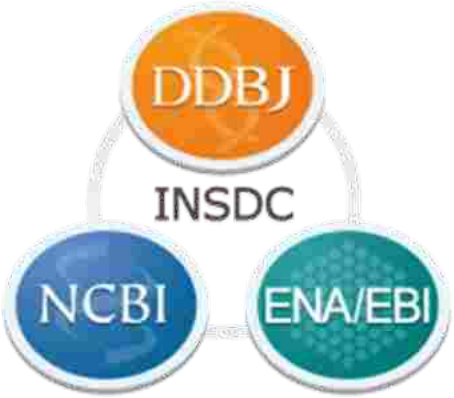
EMBL and DDBJ

- European Molecular biology Lab (EMBL established1980)

- DNA databank of Japan (DDBJ) established in Mishima japan (1984)

Nucleotide Sequence Databases

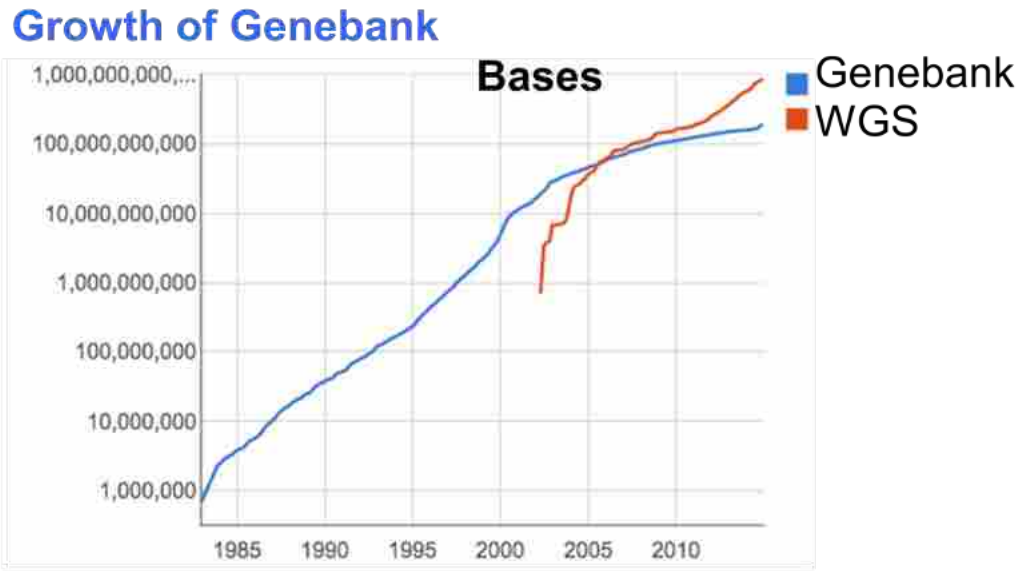
INSDC



Nucleotide Sequence Databases

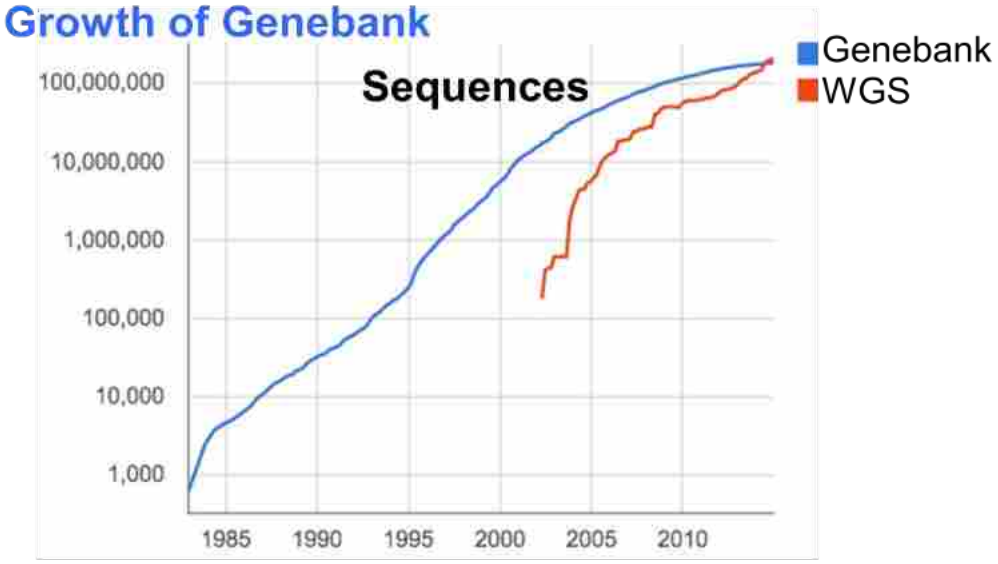
http://www.insdc.org/

Nucleotide Sequence Databases



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

Nucleotide Sequence Databases



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

Conclusions

- Biological databases store biological data
- INSDC is joint venture of NCBI, EMBL and DDBJ
- Growth of bases in Genebank is exponential, doubbling every 18 months

Lesson 69-90

STORAGE OF BIOLOGICAL SEQUENCE INFORMATION

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 contain 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.

When both DNA and RNA or mRNA are sequenced in lab their sequences contains larger number of nucleotides with variety

And when we talk about protein its sequences contain large number of bases as they are complex in nature.

SOLUTIONS DATABASES

This large number of sequence or bases cannot be stored in a single computer that's why solution lies in public sequence data bases for DNA & RNA the public database is 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.

COMPARING SEQUENCES

There are millions sequences on GenBank and UniProt what will happen if we will compare them? 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 relationship.

As there nucleotides can be similar or differ from each other

By comparison of nucleotides and Amino acids of any DNA, RNA and protein sequence we canfind many evolutionary facts and relations among species.

SIMILARITIES & DIFFERENCES IN SEQUENCES

When we compare the sequences of DNA and RNA we can get the similarity and differences or relationship in evolution. And same case is with amino acids of proteins.

In compression not only they have the same number of nucleotides but they have same orderor arrangements. If some sequence is exactly similar to each other it means;

- They might have some regular expression in cell or system.
- Or they indicate some specific presence like signature of any protein or gene.
- Or they might have similar nucleotide just one or two between them are different from rest.

CONCLUSION

If there is exact match in sequences it means their order or arrangement and maximum numbers of nucleotides match to each other not all of those.

While the genome of each created kind is unique, many animal kinds share some specific types of genes that are generally similar in DNA sequence. When comparing DNA sequences between animal taxa, evolutionary scientists often hand-select the genes that are commonly shared and more similar (conserved), while giving less attention to categories of DNA sequence that are dissimilar. One result of this approach is that comparing the more conserved sequences allows the scientists to include more animal taxa in their analysis, giving a broader data set so they can propose a larger evolutionary tree. Although these types of genes can be easily aligned and compared, the overall approach is biased towards evolution. It also avoids the majority of genes and sequences that would give a better understanding of DNA similarity concepts.

PAIR WISE ALIGNMENT –II

In pair wise alignment of nucleotides the nucleotides comes in pairs and matching are colored while missing amino acids are indicated with “” and this empty space is called as gap.

And these Gaps are inserted for deletion or insertion of any nucleotide. Increase in Gaps can increase the chance of plenty in sequencing and less number of Gaps can increase the similarity rate of sequences.

There are two types of pair alignments.

1. Global

2. Local

In Global ways of sequence pair alignment we introduce the Gaps in all sequence to know over all matching. While in Local type of sequence pair alignment we find those regions where nucleotides are maximum matching with each other it is used to find the similarity or some mutation.

Most important the Gaps are introduced so that we may add the missing nucleotides. 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).

PAIR WISE SEQUENCE ALIGNMENT –III

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

- Substitutions ACGA → AGGA
- Insertions ACGA → ACCGA
- Deletions ACGA → AGA

By applying all above ways to any sequence the matching and mismatching can be increased or decreased between to different comparing sequencing. Both local and Global ways of alignments give us different results. But among above Substitution increases mismatch of sequence.

IDENTY VS SIMILARITY

When we talk about the comparison of two sequences than question arise that how we can compare the biological sequences and after comparison what will be the degree of comparison.

There are two concepts for sequence analysis

1. Identity

2. Similarity

Identity means the counting number of nucleotides or amino acids which exactly match when two biological sequences are matched.

For example:

Number of match = 5

Smaller length = 5

Sequence (1) = 7

Sequence (2) = 5

Formula for Identity:

Identity = No. of Matches / smaller length × 100

= 5/5*100=100 percent

And Similarity means the comparison between two different sequences calculated by alignment approach. In both identity and similarity the dots are not counted.

MULTIPLE SEQUENCE ALIGNMENT

In pair wise sequence alignments we use pairs of sequence to compare them. And scoring matrices were used to score the sequence ranks. In Multiple sequence Alignments we compare multiple number of protein and DNA sequences to identify the matches and mismatches.

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.

MORE ON MULTIPLE SEQUENCE ALIGNMENT

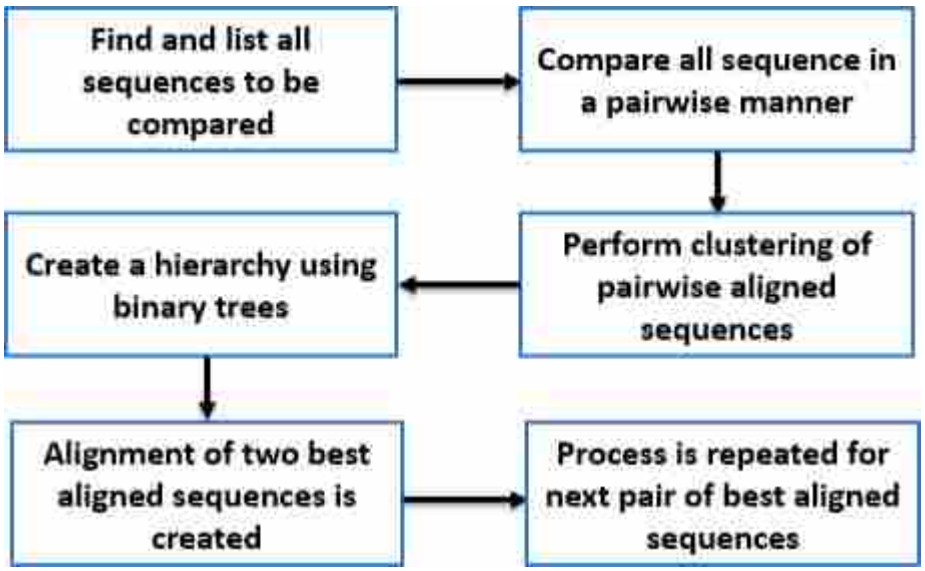
MSA helps compare several sequences by aligning them. MSA can extract consensus sequences from several aligned sequences. Characterize protein families based on homologous regions.

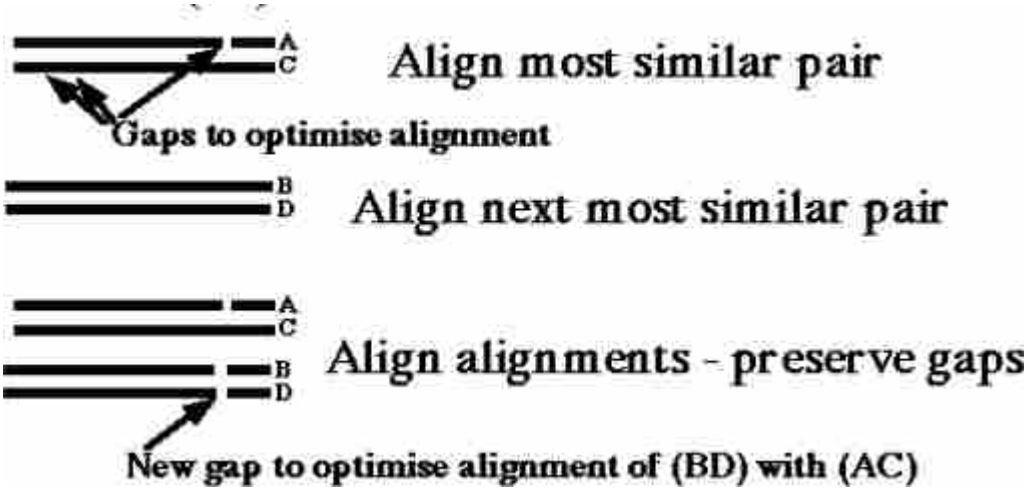
APPLICATION OF MSA

- Predict secondary and tertiary structures of new protein sequences
- Evaluate evolutionary order of species or “Phylogeny”

METHODOLOGY

- Pairwise alignment is the alignment of two sequences
- MSA can be performed by repeated application of pairwise alignment





CONCLUSION

MSA can help align multiple sequences. Progressive alignment can help perform MSA. Need to remove sequences with >80% similarity.

PROGRESSIVE ALIGNMENT FOR MSA

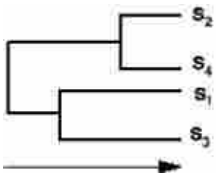
MSA involves progressive alignment of sequences. Doing so many progressive alignments can be slow.

STEPS:

Step 1: Pairwise Alignment of all sequences

Example: S1, S2, S3, S4, so that is 6 pairwise comparisons.

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



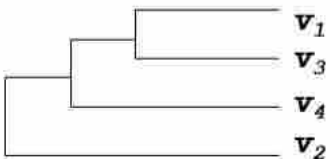
$$Distance = 1 - \frac{Matches}{Comparisons}$$

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

Step 3: Progressive alignment following branching order in tree.

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

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



SHORTCOMING OF THIS APPROACH

- Dependence upon initial alignments
- If sequences are dissimilar, errors in alignment are propagated
- Solution: Begin by using an initial alignment, and refine it repeatedly

Progressive alignments are used in aligning multiple sequences. Iterative approaches can help refine results from progressive alignments

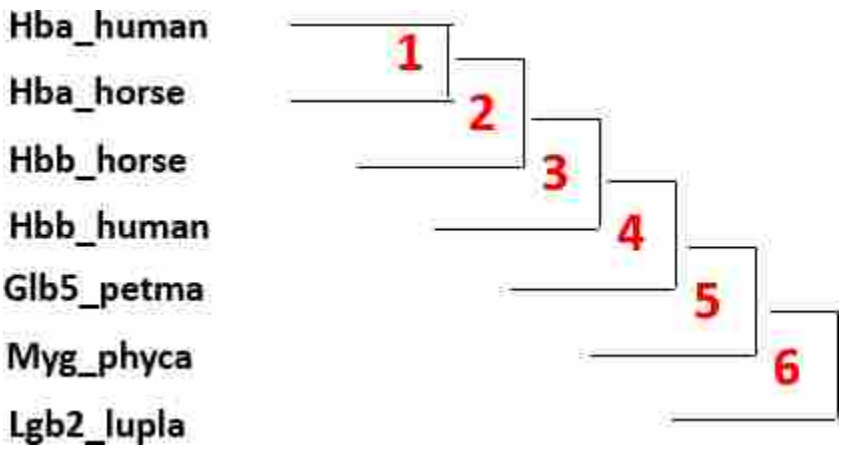
MSA-EXAMPLE

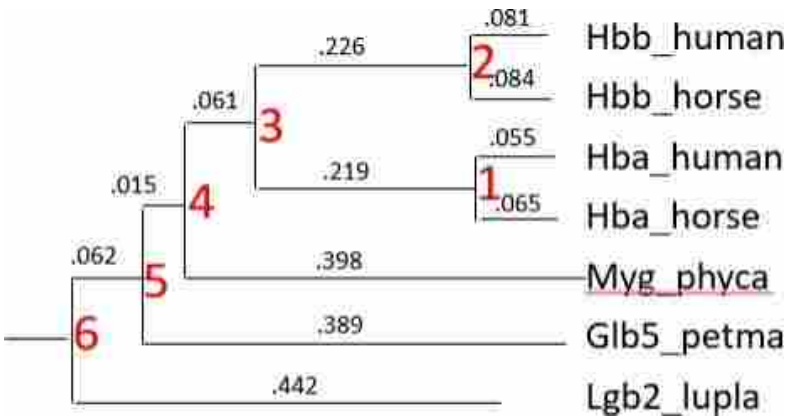
MSA involves progressive alignment of sequences. Doing so many progressive alignments can be slow.

For example:

$$\text{distance between 2 sequences} = 1 - \frac{\text{No. identical residues}}{\text{No. aligned residues}}$$

		Hbb_human	Hbb_horse	Hba_human	Hba_horse	Myg_phyca	Glb5_petma	Lgb2_lupla
Hbb_human	1	-						
Hbb_horse	2	.17	-					
Hba_human	3	.59	.60	-				
Hba_horse	4	.59	.59	.13	-			
Myg_phyca	5	.77	.77	.75	.75	-		
Glb5_petma	6	.81	.82	.73	.74	.80	-	
Lgb2_lupla	7	.87	.86	.86	.88	.93	.90	-





MSA can be better performed using clustering strategies followed by alignment of the alignments later. CLUSTAL is a free online tool that does all of this for us

CLUSTALW

MSA involves progressive alignment of sequences. Doing so many progressive alignments canbe slow. CLUSTALW is an online tool to perform MSA.Developed by European Molecular Biology Laboratory & European Bioinformatics Institute.

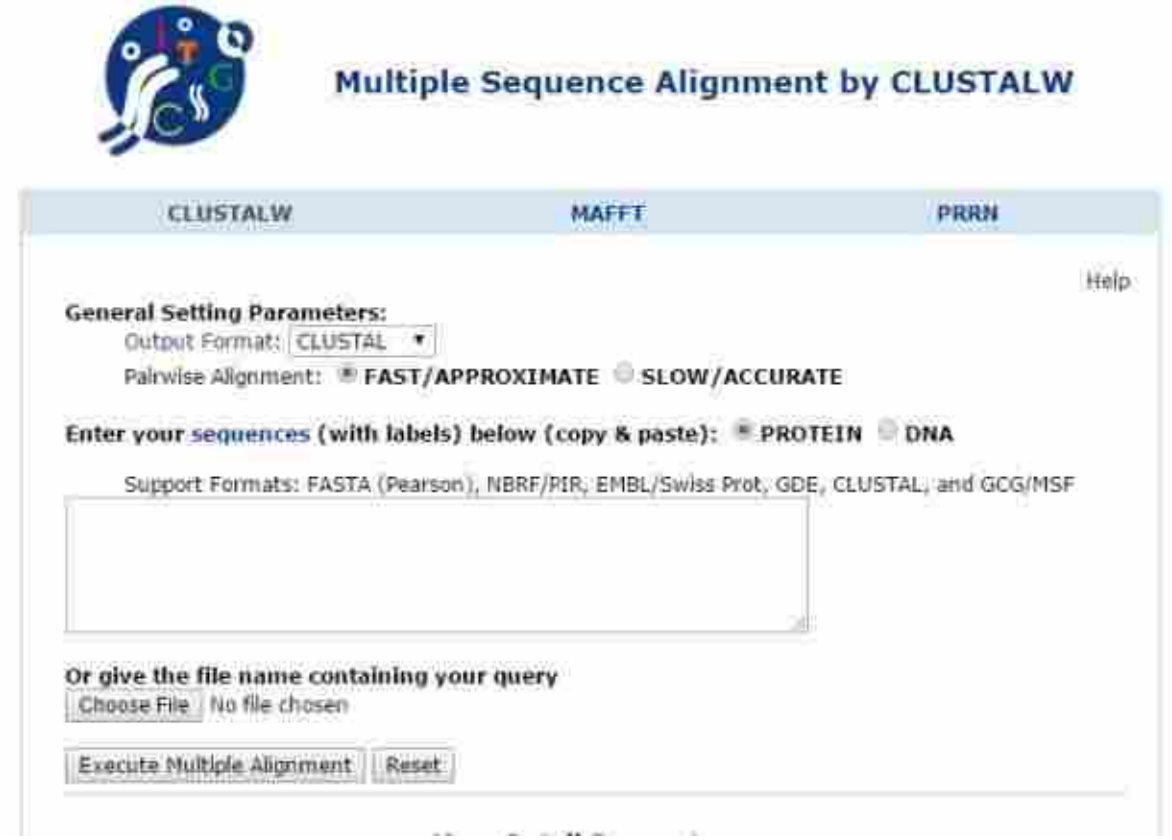
Performs alignment in:

- slow/accurate
- fast/approximate

SCOPE

- create multiple alignments,
- optimize existing alignments,

- profile analysis &
- create phylogenetic trees



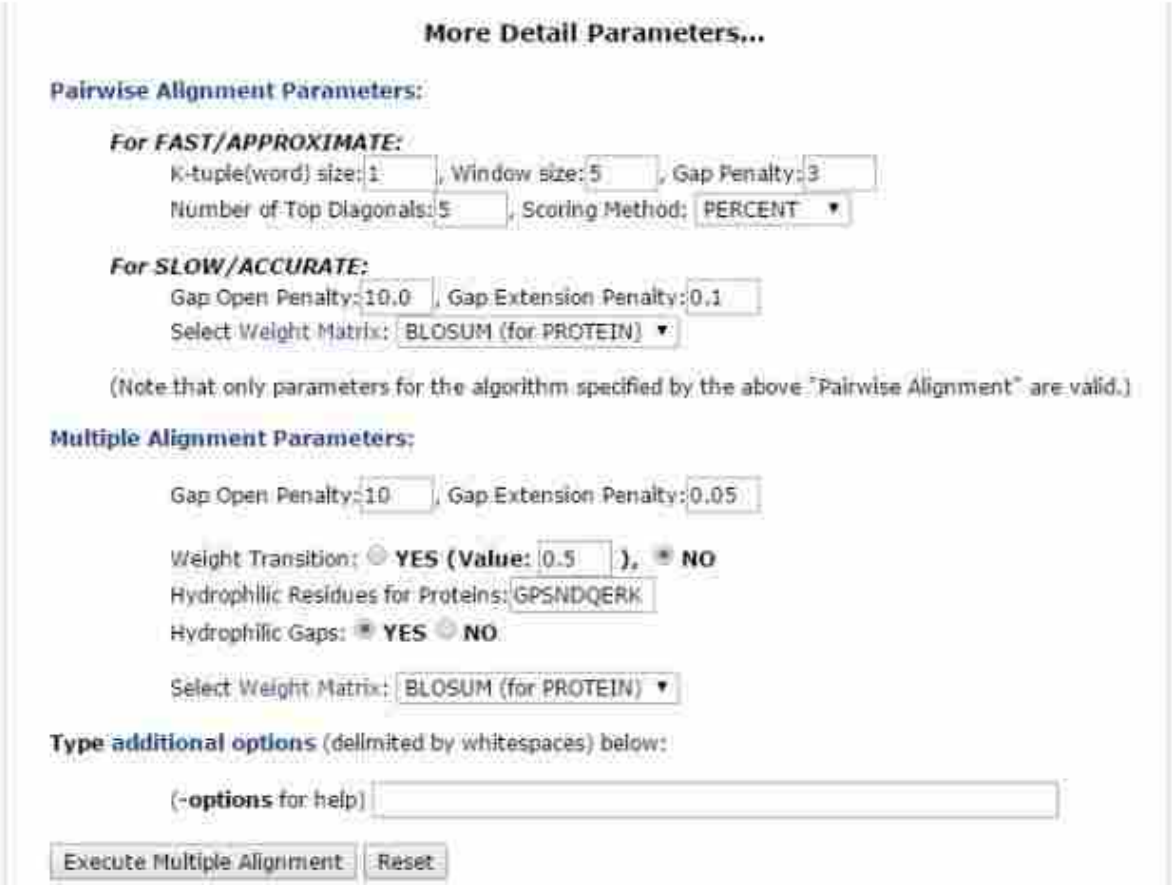
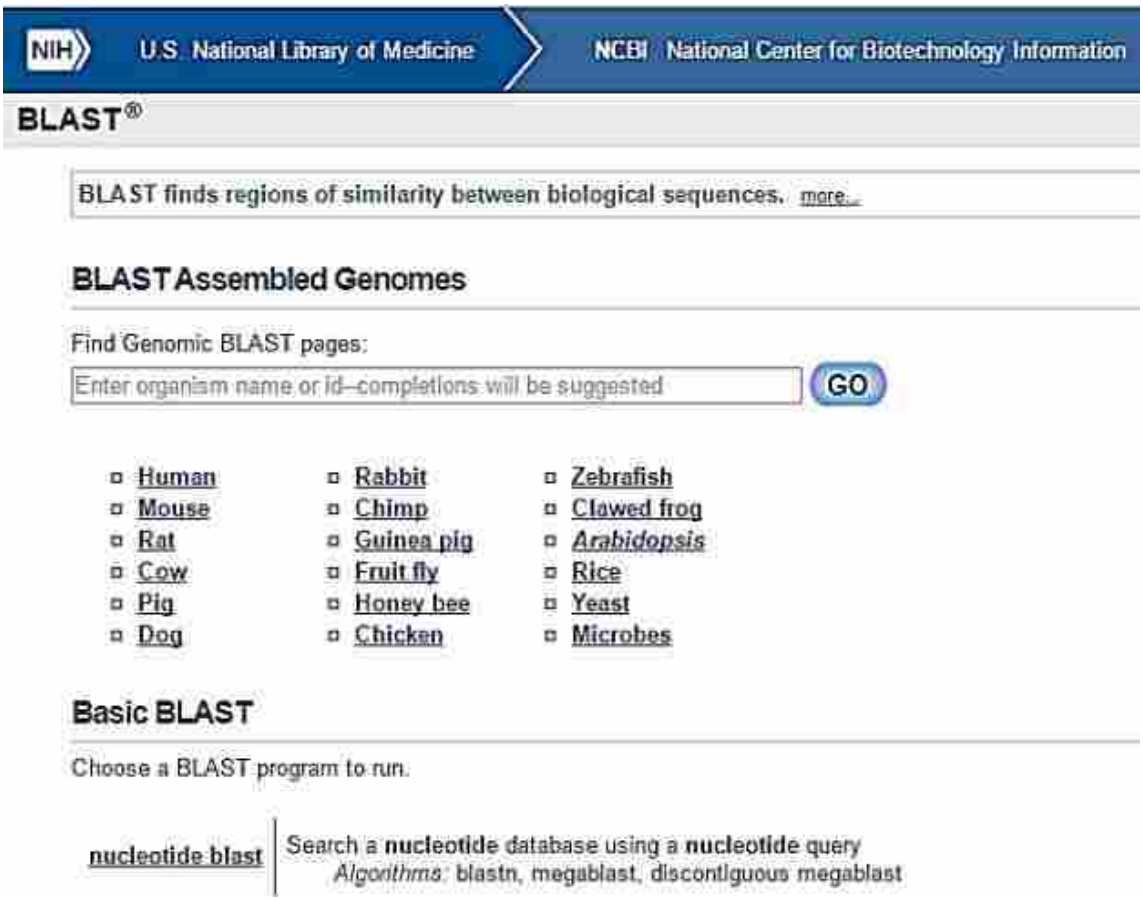


Table 1. Some recent and less recent available methods for MSAs.

Name	Algorithm	URL
MSA	Exact	http://www.ibc.wustl.edu/ibc/msa.html
DCA	Exact (requires MSA)	http://bibiserv.techfak.uni-bielefeld.de/dca
OMA	Iterative DCA	http://bibiserv.techfak.uni-bielefeld.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
ConAlign	Consistency-based	http://www.daimi.au.dk/~ 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

INTRODUCTION TO BLAST-I

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



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

INTRODUCTION TO BLAST-II

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

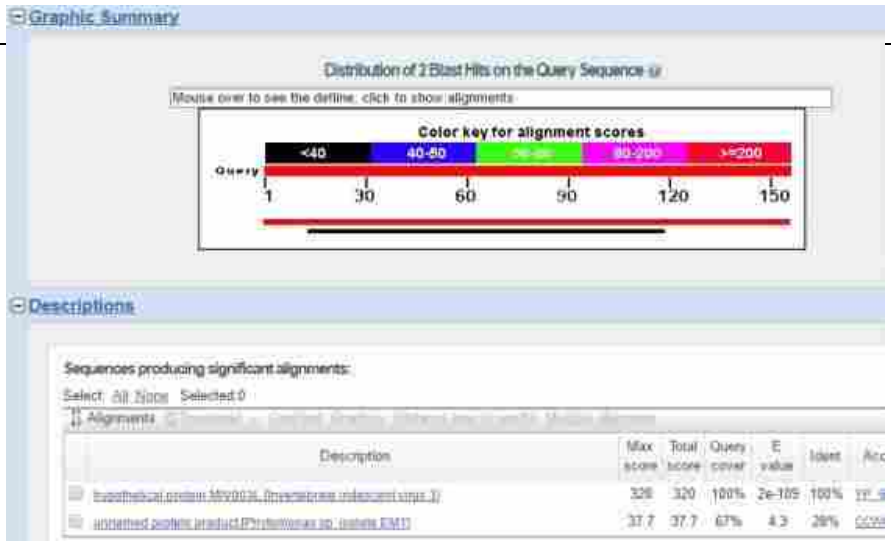
www.blast.ncbi.nlm.nih.gov/Blast.cgi

Smith Waterman can align complete sequences. BLAST does it in an approximate way. Hence,

BLAST is faster BUT does not ensure optimal alignment. BLAST provides for approximatesequence matching. Input to BLAST is a FASTA formatted sequence and a set of search parameters

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



BLAST ALGORITHM

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.

For example:

Query sequence: PQGELV

Make list of all possible worlds (length 3 for proteins)

PQG (score 15)

QGE (score 9)

GEL (score 12)

ELV (score 10)

Assign scores from Blosum62, use those with score> 11: PQG & GEL

Mutate words such that score still > 11

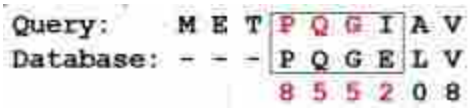
PQG (score 15) similar to PEG (score

13)At the end, we get: 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):



High Scoring Pair: PQGI (score 8+5+5+2)

If 2 HSP in query sequence are < 40 positions away

Full dynamic alignment on query and hit sequences

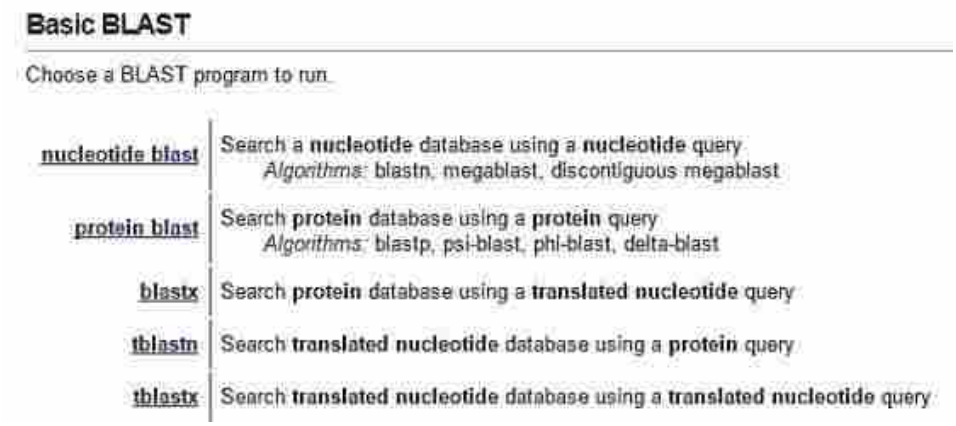
BLAST performs quick alignments on sequences. The results are tabulated with alignment regions overlapping each other. Statistical evaluation is also provided alongside

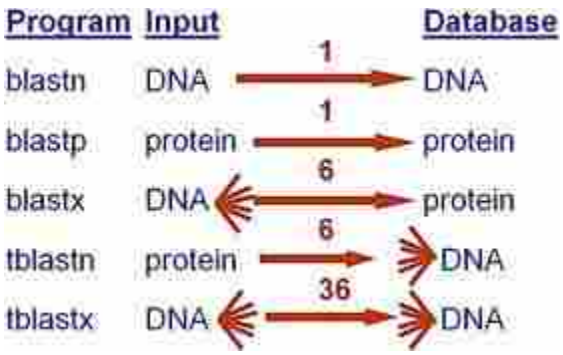
SUMMERY OF BLAST

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.

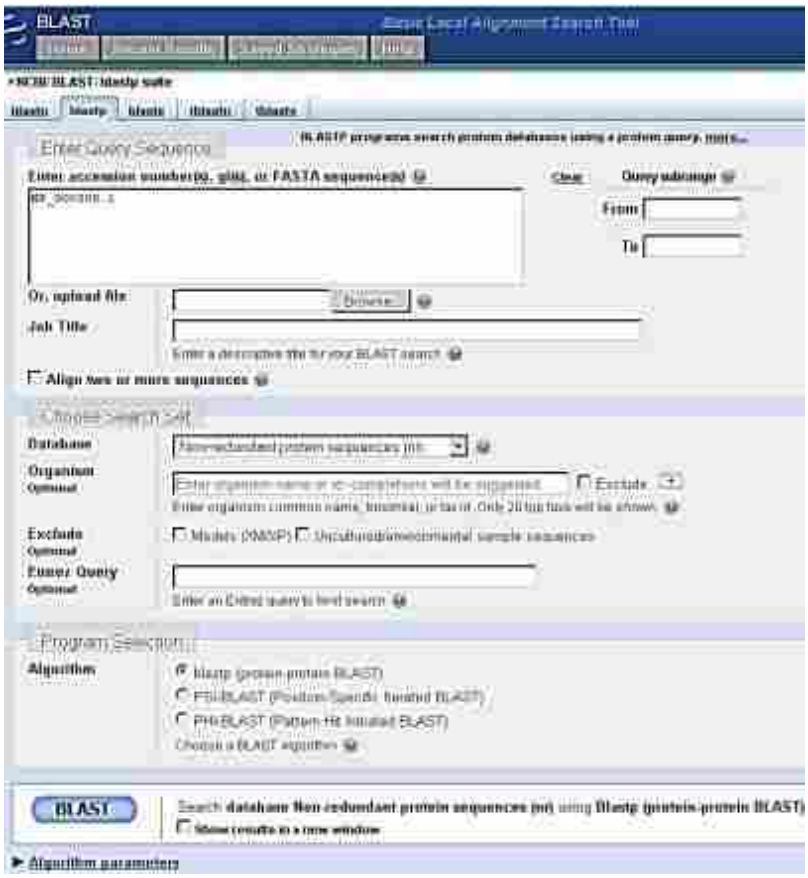
Step1: obtain a query of sequence

Step2: choose a type of BLAST

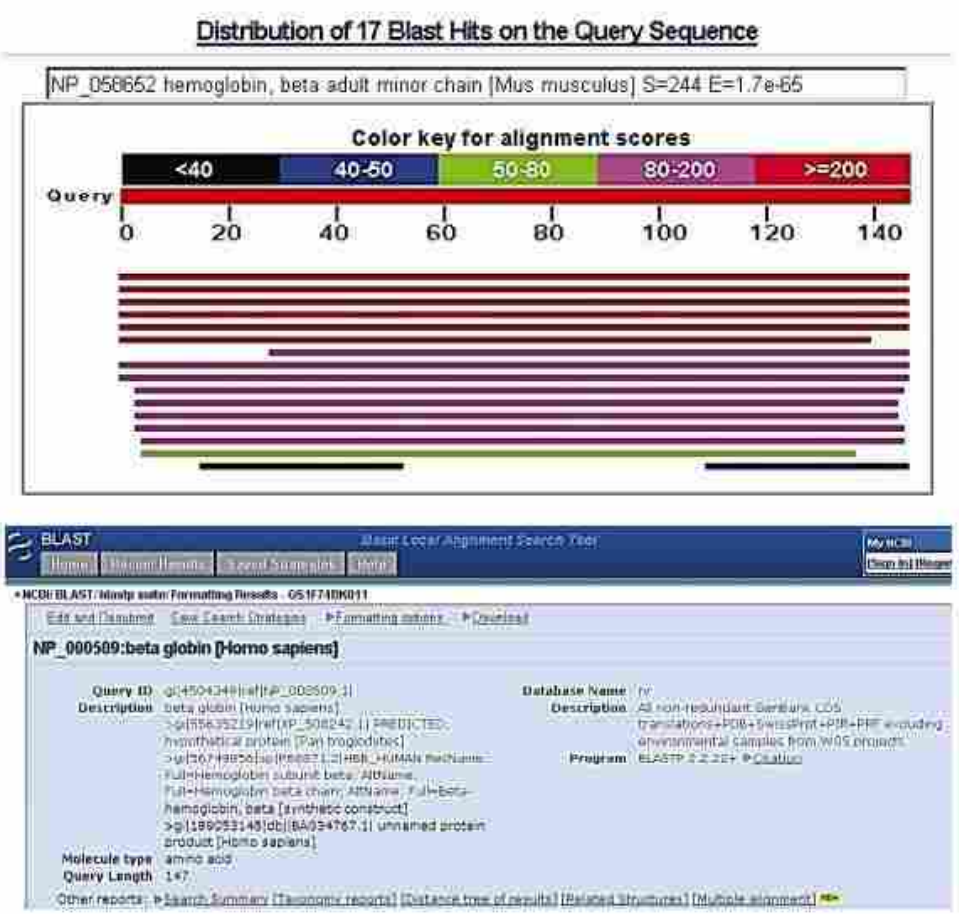




Step3: search parameter



Step4: tabulated search results



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_486069.1|	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_705842.2|	5' nucleotidase, cytosolic II like 1 protein [M	28.9	1.5	U G

INTRODUCTION TO FASTA

For comparing two sequences we use pair wise sequencing and for the comparison of many sequences we use multiple sequence alignment. To handle the multiple alignments we perform alignment through smith-waterman algorithm for local one. 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. Developed in 1988.it does Fast Alignment .Searches databases for query protein and nucleotide sequences. It was later improved upon in BLAST.

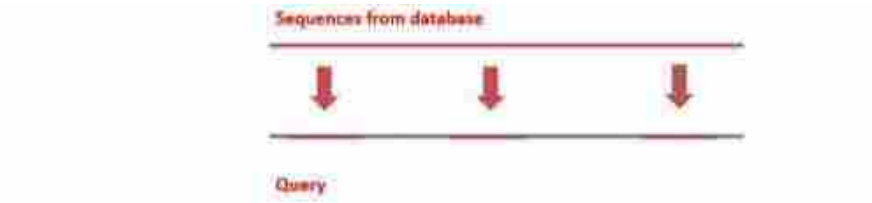
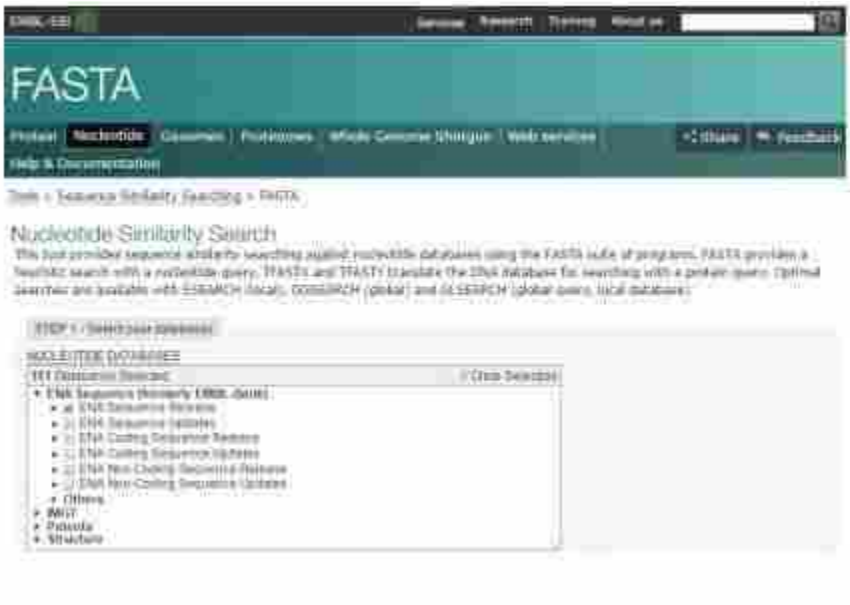


Figure 0.29 Regions of absolute identity

http://www.ebi.ac.uk/Tools/sss/fasta/


FASTA
Protein
Nucleotide
Genomes
Proteomes
Whole Genome Shotgun
Web services
Share
Feedback
Help & Documentation
Tools > Sequence Similarity Searching > FASTA
Nucleotide Similarity Search


FASTA
Protein
Nucleotide
Genomes
Proteomes
Whole Genome Shotgun
Web services
Share
Feedback
Help & Documentation
Tools > Sequence Similarity Searching > FASTA
Protein Similarity Search

INTRODUCTION TO FASTA-II

FASTA – Fast Alignment Algorithm. Classical global and local alignment algorithms are timeconsuming. FASTA achieves alignment by using short lengths of exact matches.

USES OF FASTA

FASTA relies on aligning subsequences of absolute identity. Input to FASTA search can be in

FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP or UniProt formats

OUTPUT OF BLAST

Results are output in visual format along with functional prediction. Makes table lists the sequence hits found along with scores. Users can click on each reported match to look at the details

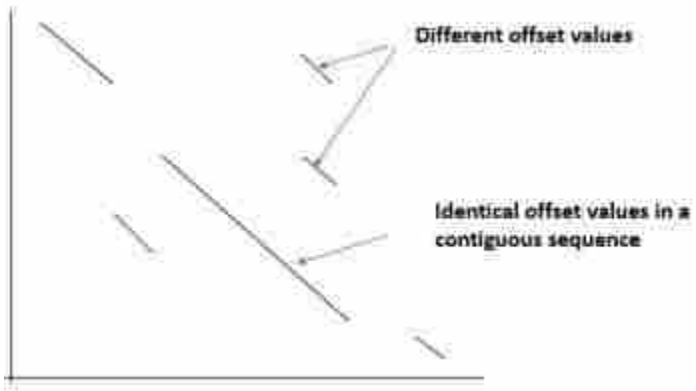

Tools > Sequence Similarity Searching > FASTA
Protein Similarity Search
This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).
STEP 1 - Select your databases
PROTEIN DATABASES
0 Databank Selected
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
MDASSSPSFSEESLKLELDDLGKGLNKKLRFEASVCSIHNLLRDHYSSSSPSLRKDFYIV
VIRVATVLKTRYTATDFWVAGLSLPEEAERLVTIDASEKKHLKSCVAQAKEGLSEVDNQPT
ESSQIGYLFEGHLTVDREFFQPQWLVQGNLMSAFASNVGGESSNGFTENTIGETAINLMQEL
INGLDMIIFDILDDGGPPRAPPASKEVSEKLPVIIFTEELLKKFGAEAECCICKENLVG
DKMQELPCKHTFHPPCLKPVLDEHNSCPICRHELPTDDGKYEMWKEREKAEEERKGAEN
AVRQGEYMYV
or Upload a file: Choose File No file chosen


EMBL-EBI
Services Research Training About us
FASTA
Protein Nucleotide Genomes Proteomes Whole Genome Shotgun Web services
Share Feedback
Help & Documentation
Tools > Sequence Similarity Searching > FASTA
Results for job fasta-20160323-032745-0099-35769831-es
Summary Table Tool Output Visual Output Functional Predictions Submission Details
Color scale:
fixed
dynamic
Update
Download in SVG format

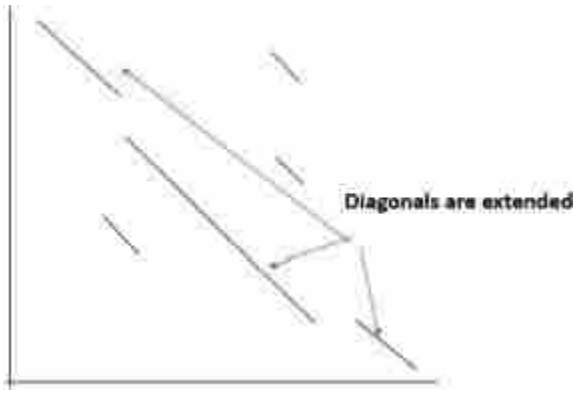
FASTA ALGORITHM

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

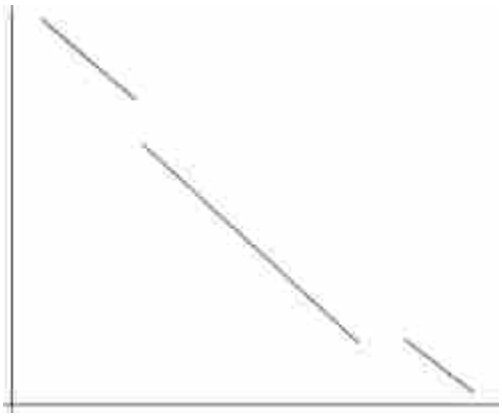
STEP1: Local regions of identity are found



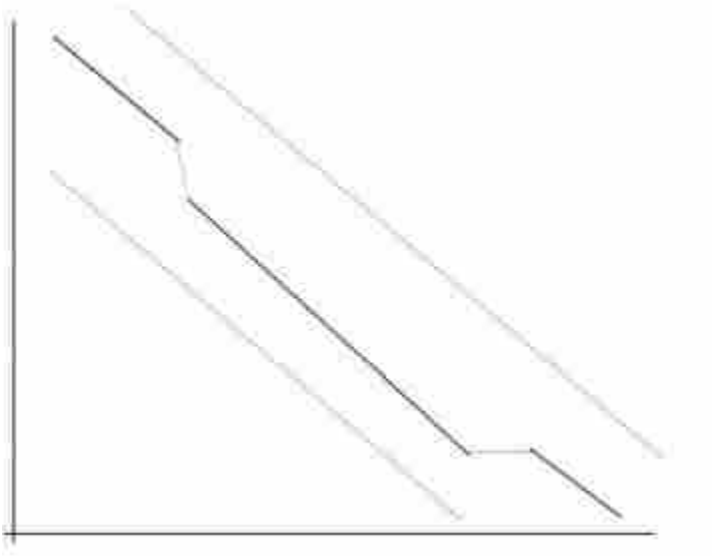
STEP2: Rescore the local regions using PAM or BLOSUM matrix



STEP3: Eliminate short diagonals below a cutoff score



STEP4: Create a gapped alignment in a narrow segment and then perform Smith Watermann Alignment



TYPES OF FASTA

There are six types of FASTS:

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

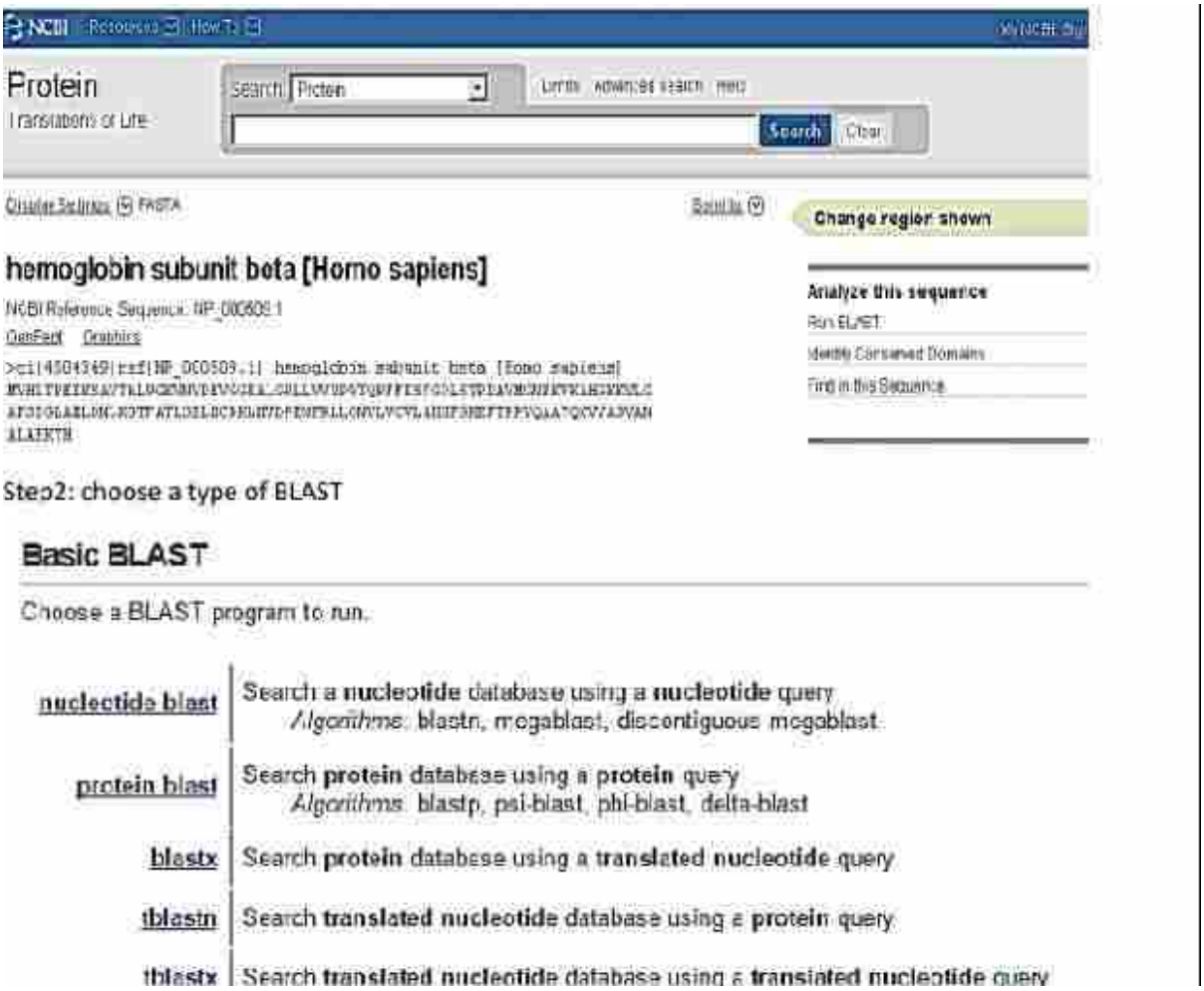
Compare a DNA sequence (6ORFs) to a protein sequence database

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

SUMMERY OF BLAST

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.

Step1: obtain a query of sequence

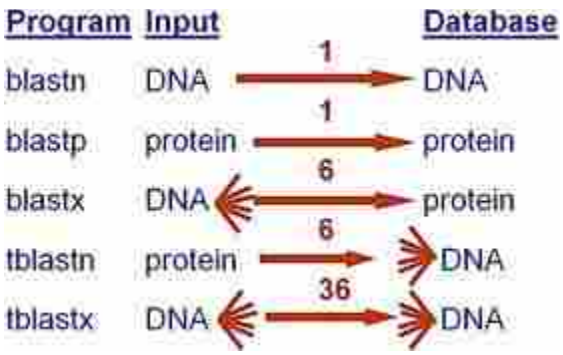


Step2: choose a type of BLAST

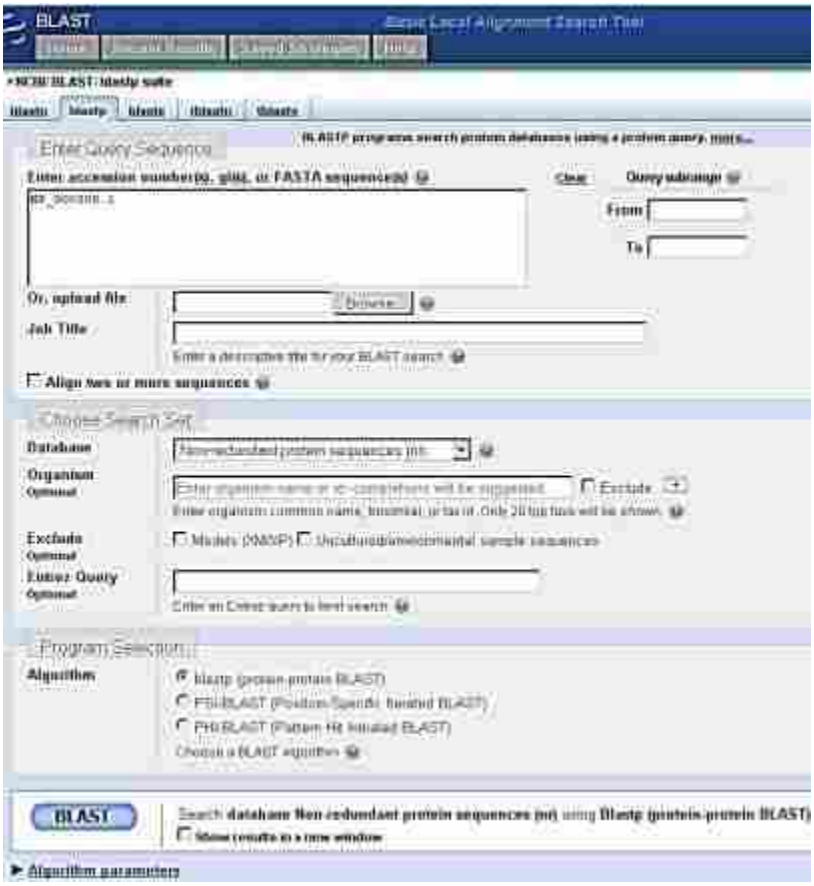
Basic BLAST

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

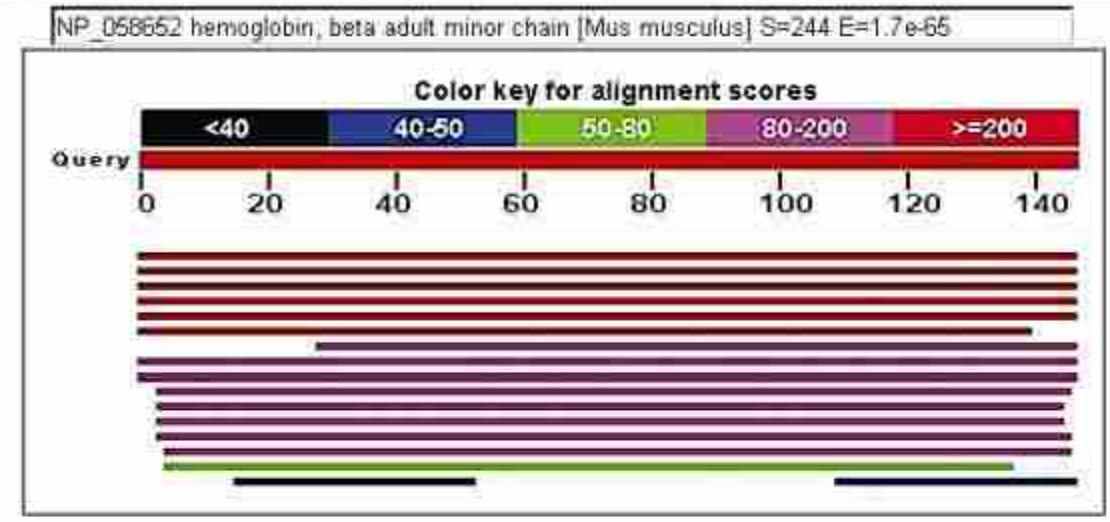


Step3: search parameter



Step4: tabulated search results

Distribution of 17 Blast Hits on the Query Sequence



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_486069.1|	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