Fundamentals of Bioinformatics

Dr. Upeksha Ganegoda
Department of Computational Mathematics

Course Objective

On successful completion of this module, the students will be able to define concepts in bioinformatics, different usage of biological databases, and explain how bioinformatics used in telemedicine and decision making. Furthermore, identify different types of sequence alignment methods and biological algorithm techniques.

Outline Syllabus

- Introduction to bioinformatics and computational biology
- Biological terms and usage
- Biological databases
- Telemedicine
- Clinical decision support
- Sequence alignments
- Biological algorithm techniques

} IT Application

} Concepts & algorithms

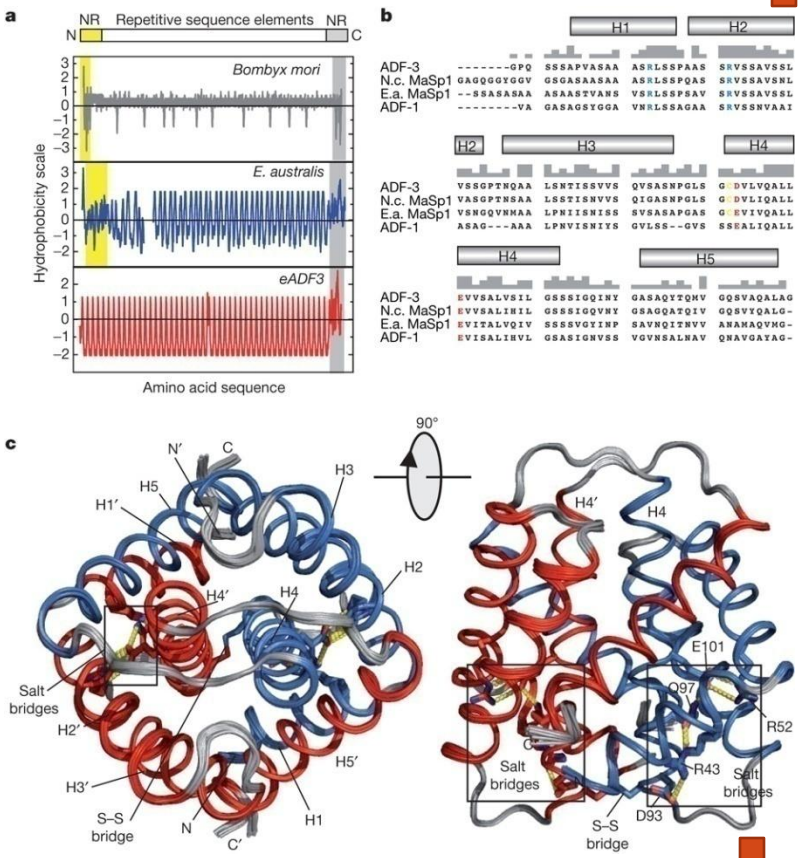
Recommended Text

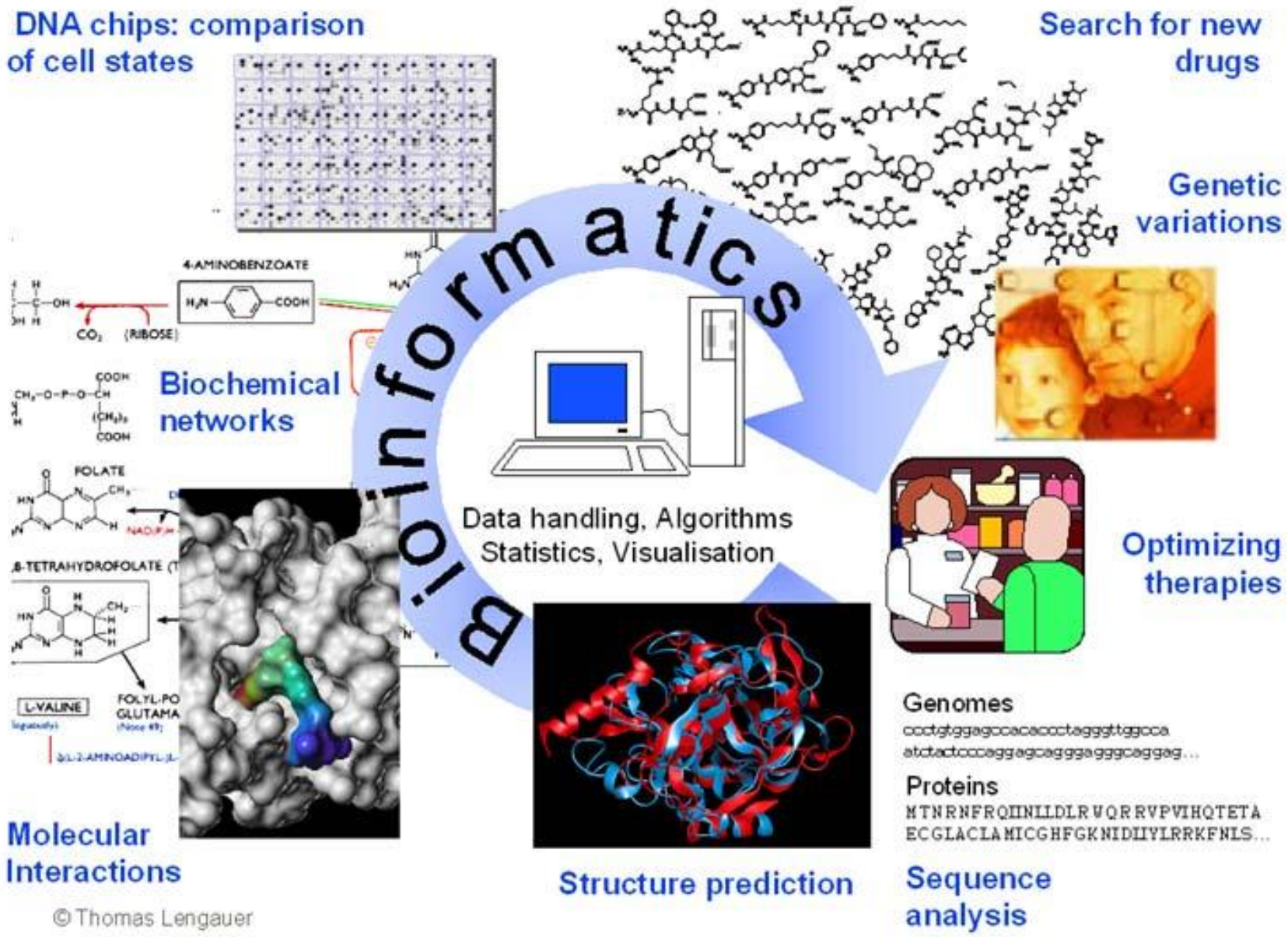
- Essential Bioinformatics by Jin Xiong
- An Introduction to Bioinformatics Algorithms by Neil C. Jones and Pavel A. Pevzner

Course Evaluation

- One Assignment
- Final Examination

What is Bioinformatics




DNA chips: comparison of cell states
Search for new drugs
Genetic variations
Bioinformatics
Biochemical networks
Data handling, Algorithms
Statistics, Visualisation
Optimizing therapies
Molecular Interactions
Structure prediction
Genomes
ccctgtggagccacaccctagggttggcca
atctactcccaggagcagggagggcaggag...
Proteins
MTNRNFRQIINLLDLRVQRRVPVIHQTETA
ECGLACLAMICGHFGKNIDIIYLRRKFNLS...
Sequence analysis
4-AMINOBENZOATE
FOLATE
L-VALINE
© Thomas Lengauer

Definition of bioinformatics and computational biology

- Bioinformatics involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA, and proteins. It develop applications of computational tools to manage all kinds of biological data.

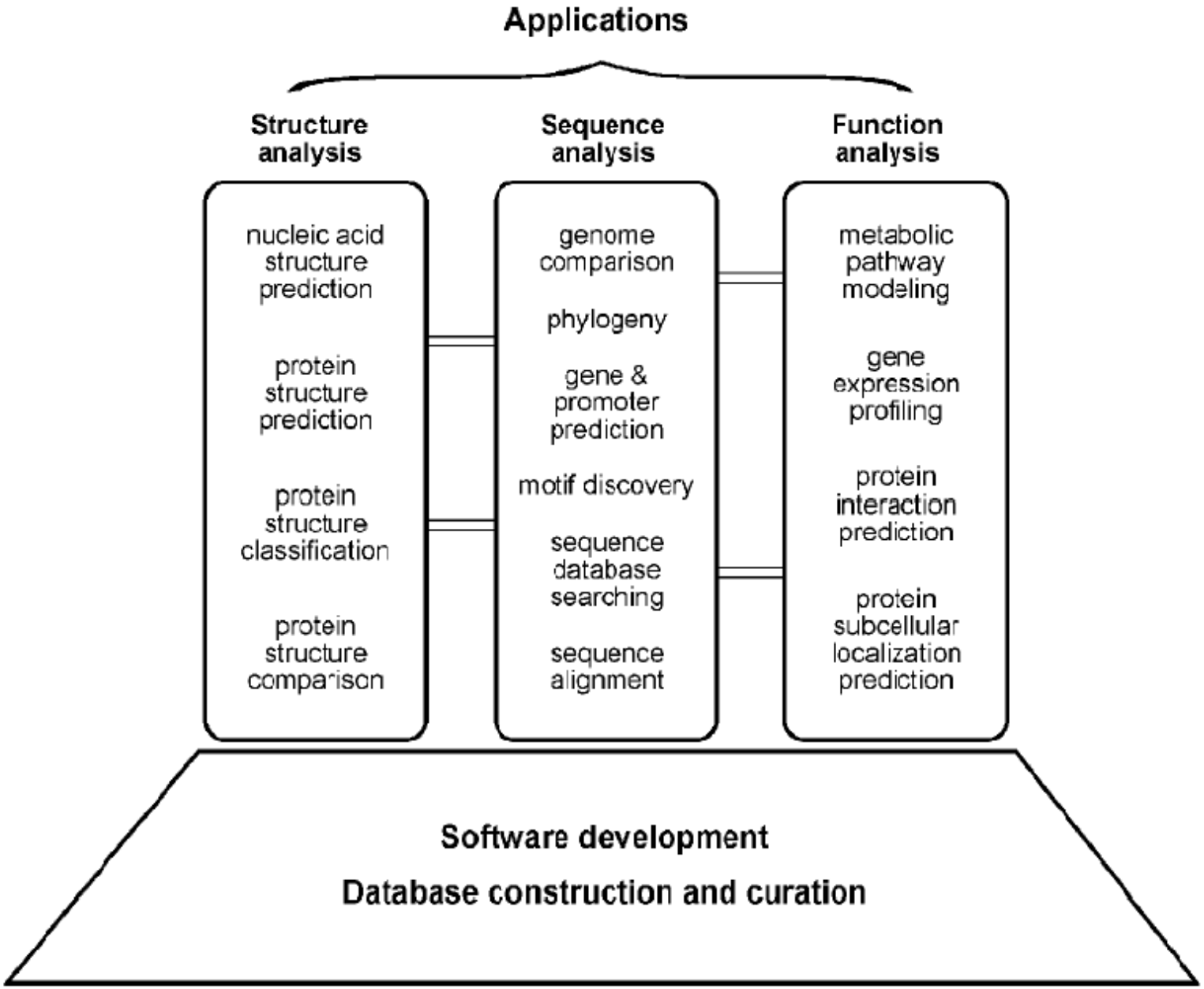
- Bioinformatics = the creation of tools (algorithms, databases) that solve problems. The goal is to build useful tools that work on biological data. It is about engineering.

Definition of computational biology

- Computational biology concern about developing theoretical algorithms to be used in bioinformatics.
- Computational biology = the study of biology using computational techniques. The goal is to learn new biology, knowledge about living systems. It is about science.
- Computational Biology, is the science of using biological data to develop algorithms and relations among various biological systems.

Sub field of Bioinformatics

- Development of computational tools and databases
  - Ex. writing software for sequence, structural, and functional analysis, as well as the construction and curating of biological databases.
- Application of these tools and databases in generating biological knowledge to better understand living systems.
  - Ex. knowledge-based drug design, forensic DNA analysis, and agricultural biotechnology


Applications
Structure analysis
Sequence analysis
Function analysis
nucleic acid structure prediction
protein structure prediction
protein structure classification
protein structure comparison
genome comparison
phylogeny
gene & promoter prediction
motif discovery
sequence database searching
sequence alignment
metabolic pathway modeling
gene expression profiling
protein interaction prediction
protein subcellular localization prediction
Software development
Database construction and curation

Limitations

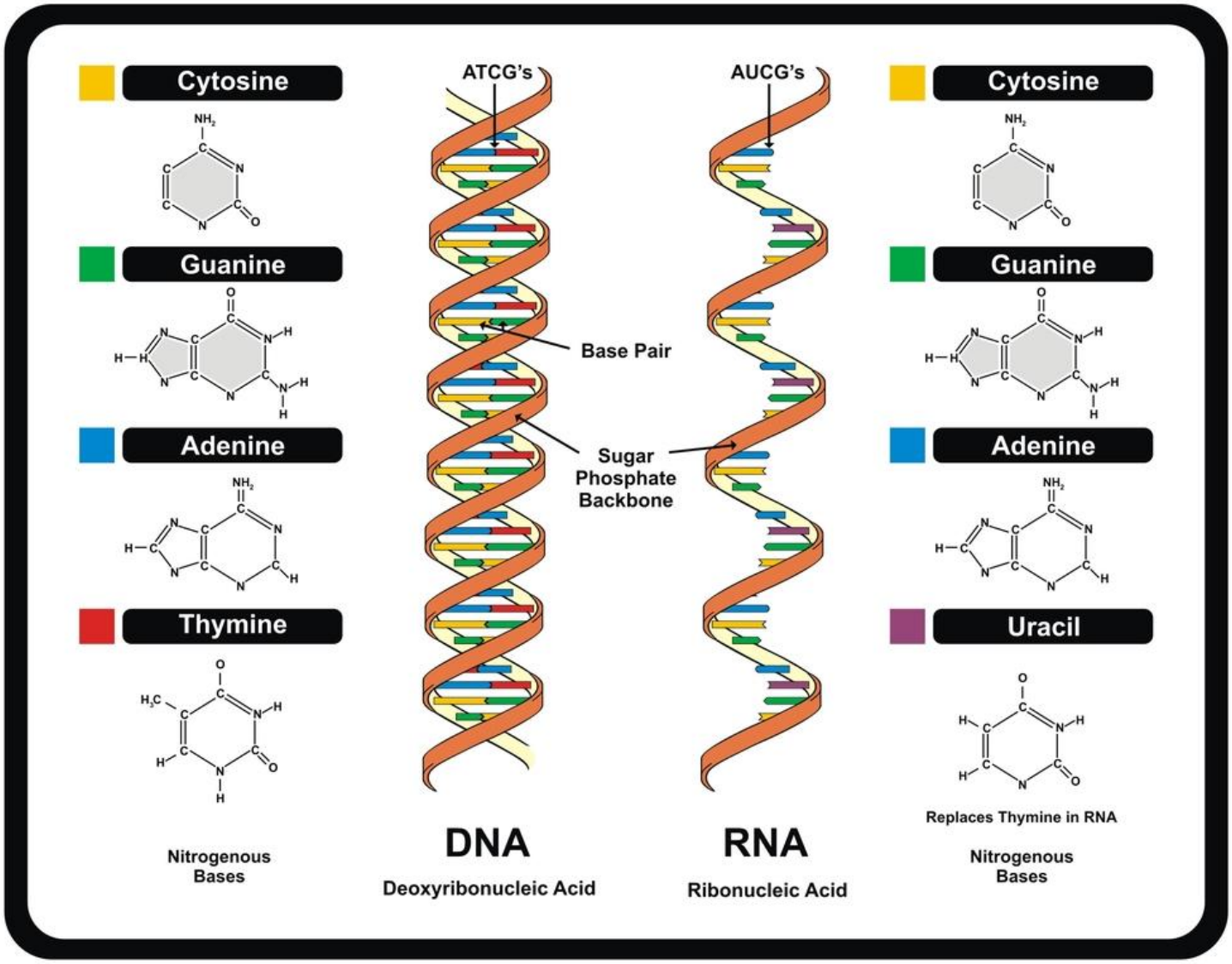
- Bioinformatics predictions are not formal proofs of any concepts. They do not replace the traditional experimental research methods of actually testing hypotheses.

- The quality of bioinformatics predictions depends on the quality of data and the sophistication of the algorithms being used.

Biological Terms

- DNA & RNA
- Chromosome
- Gene
- Protein
- Genotype and Phenotype
- Mutation
- Gene expression

DNA & RNA


Cytosine
Guanine
Adenine
Thymine
Nitrogenous Bases
ATCG's
AUCG's
Base Pair
Sugar Phosphate Backbone
DNA
Deoxyribonucleic Acid
RNA
Ribonucleic Acid
Cytosine
Guanine
Adenine
Uracil
Replaces Thymine in RNA
Nitrogenous Bases

DNA

- DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a **nucleotide**. Nucleotides are arranged in two long strands that form a spiral called a double helix.

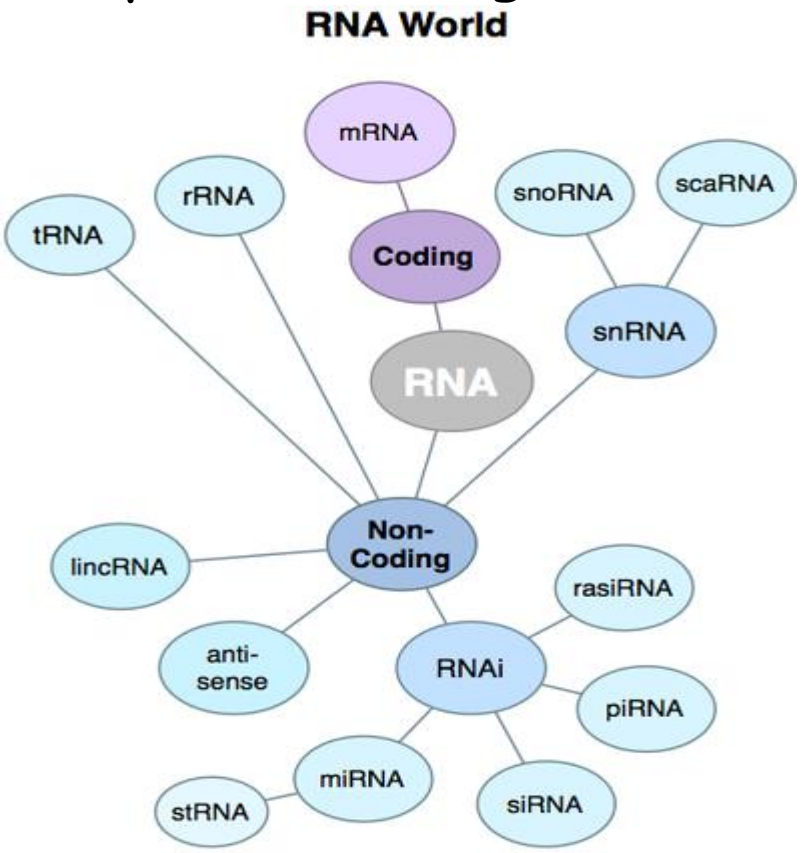
- **DNA can replicate**, or make copies of itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical when cells divide because each new cell needs to have an exact copy of the DNA present in the old cell.

- The complete set of information in an organism's DNA is called its **genome**, and it carries the information for all the proteins the organism will ever synthesize. (The term genome is also used to describe the DNA that carries this information.)

- The amount of information contained in genomes is staggering: For Example, a typical human cell contains 2 meters of DNA. Written out in the four-letter nucleotide alphabet, and complete sequence of nucleotides in the human genome would fill more than a thousand books the size of this one. In addition to other critical information, it carries the instructions for about 30,000 distinct proteins.

RNA

- RNA is a polymeric molecule which resides in all living cells and viruses. It contain in various types of biological roles such as coding, decoding, regulation and expression of genes.



- Coding RNA

Describe the RNA which carry the code or chemical blueprint of a specific protein. Hence it will describe the functionality of the protein.

Ex. MessengerRNA (mRNA)

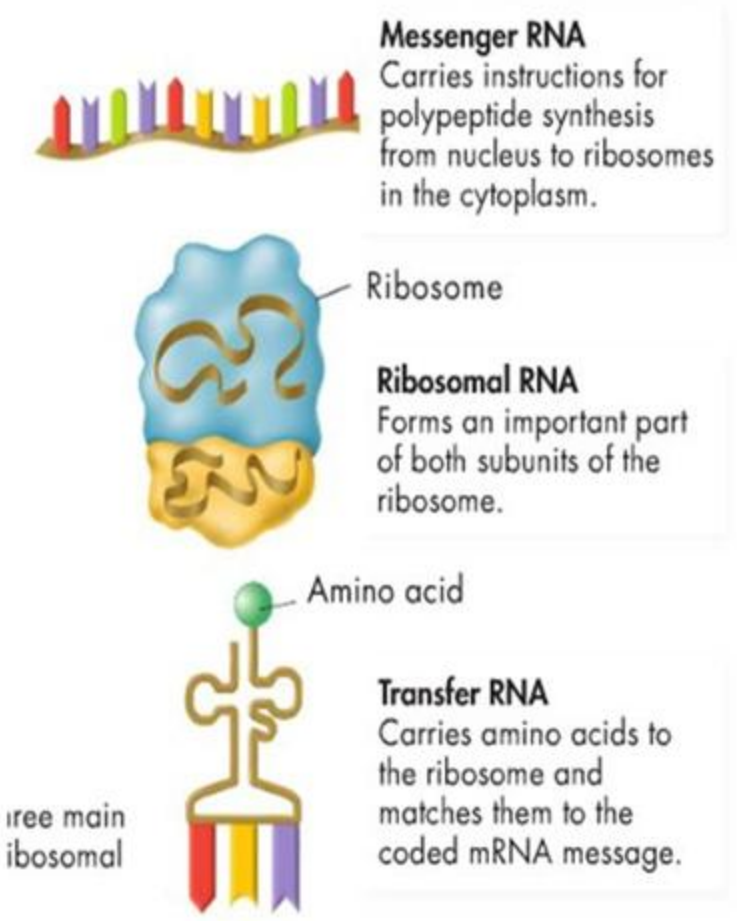
- Non-coding RNA (ncRNA)

ncRNA is an RNA molecule that is not translated into a protein. Ex. transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), snoRNAs, microRNAs, siRNAs, snRNAs, long ncRNAs, etc.

RNA Functions

Three different types of RNA:

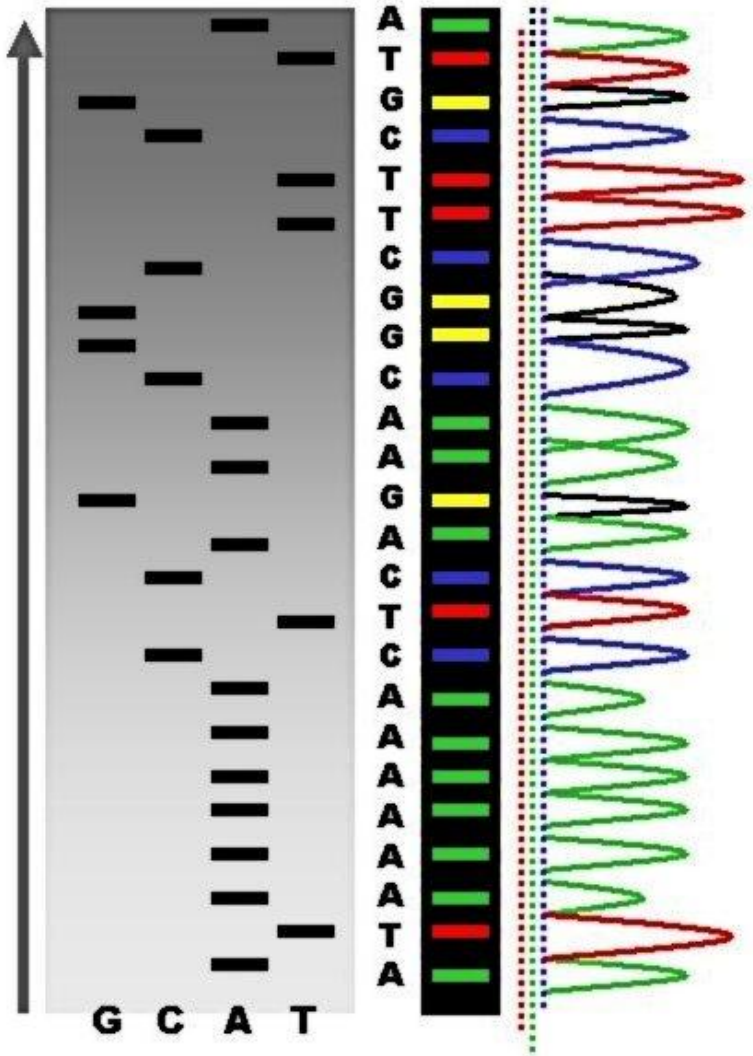
- mRNA(messenger): Used as template to make proteins
- rRNA (ribosomal): Makes up ribosome
- tRNA (transfer): Matches amino acids to mRNA to help make proteins



DNA & RNA sequencing

- **DNA sequencing** is the process of determining the precise order of nucleotides within a DNA molecule.

- RNA is less stable in the cell, and also more prone to nuclease attack experimentally. As RNA is generated by transcription from DNA, the information is already present in the cell's DNA. However, it is sometimes desirable to sequence RNA molecules. While sequencing DNA gives a genetic profile of an organism, sequencing RNA reflects only the sequences that are actively expressed in the cells.



Sequencing methods

- BLAST method
- Pairwise alignment method
- Similarity sequence method
- Dynamic programming

Advantages in DNA sequencing

Sequencing provides the order of individual nucleotides present in molecules of DNA or RNA isolated from animals, plants, bacteria, archaea, or virtually any other source of genetic information.

- **Forensics:** To identify particular individuals. As every individual has unique sequence of DNA. It is particularly used to identify the criminals by finding proof from the crime scene.

- **Medicine:** It can be used to detect the genes which are associated with some heredity or acquired diseases. Scientists use different techniques of genetic engineering like gene therapy to identify the defected genes and replace them with the health ones.

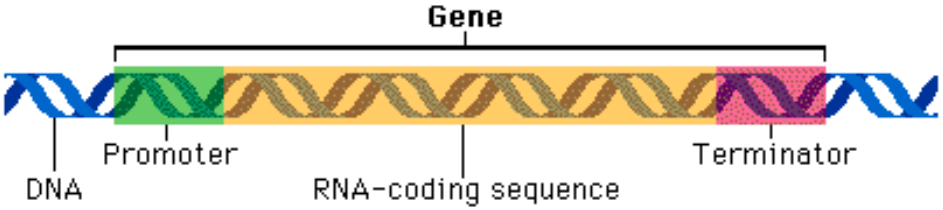
- **Agriculture:** mapping and sequencing of the whole genome of microorganisms has allowed the agriculturists to make them useful for the crops and food plants. Ex: some genes of bacteria have been used in some food plants to increase their resistance against insects and pests. Hence the productivity and nutritional value of the plants increase.

Gene

- Is a molecular unit of heredity of a living organism. It consist a region (or locus) of DNA that encodes a functional RNA or protein product. It helps to pass some specific variant of phenotypic characteristic to offspring by their parents. It consist some stretches (or region) of deoxyribonucleic acids (DNA) and ribonucleic acids (RNA) that code for a polypeptide or for an RNA chain that has a function in the organism.
- Nucleotide sequence of a very small human gene occupies a quarter of a page of text.

Different types of genes

Protein-coded genes



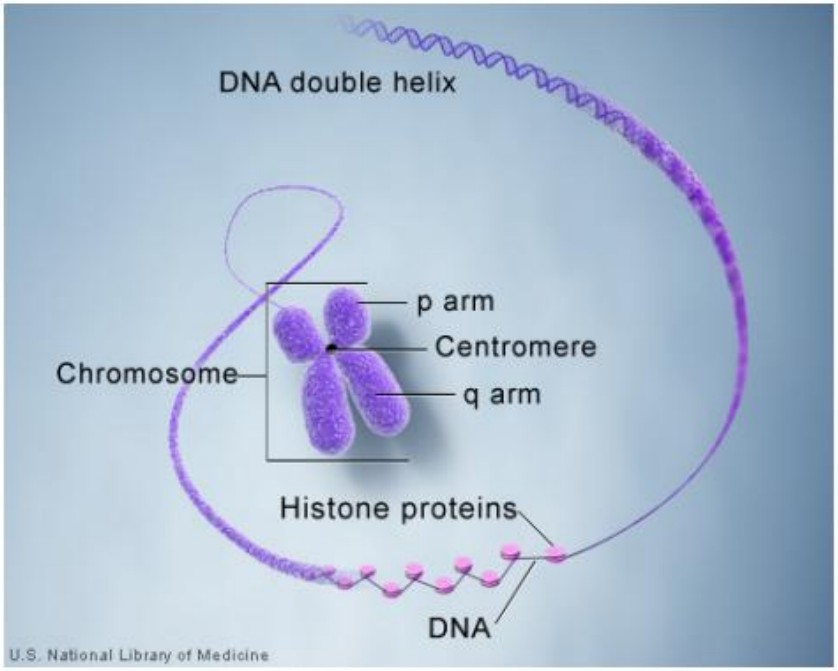
- The promoter is a base-pair sequence that specifies where transcription begins.
- The coding sequence is a base-pair sequence that includes coding information for the polypeptide chain specified by the gene. It contain information to generate a specific protein.
- The terminator is a sequence that specifies the end of the mRNA transcript.

Protein-non-coded genes

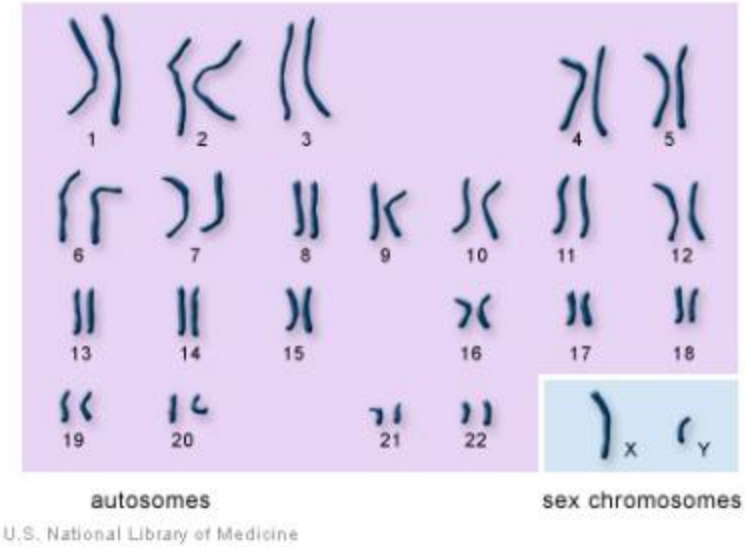
- These set of genes not contain information to generate any specific protein, but useful for different types of functionality in a living organism or cell.

Chromosome

The total complement of genes in an organism or cell is known as its genome, which may be stored on one or more chromosomes. A chromosome consists of a single, very long DNA helix on which thousands of genes are encoded.



DNA and histone proteins are packaged into structures called chromosomes.



Human Chromosome

- Chromosomes come in pairs. Normally, each cell in the human body has 23 pairs of chromosomes (46 total chromosomes). Half come from the mother; the other half come from the father.
- Two of the chromosomes (the X and the Y chromosome) determine if you are born a boy or a girl (your gender). They are called sex chromosomes:
  - Females have 2 X chromosomes.
  - Males have 1 X and 1 Y chromosome.
- The mother gives an X chromosome to the child. The father may contribute an X or a Y. The chromosome from the father determines if the baby is a girl or a boy.
- The remaining chromosomes are called autosomal chromosomes. They are known as chromosome pairs 1 through 22.

Protein

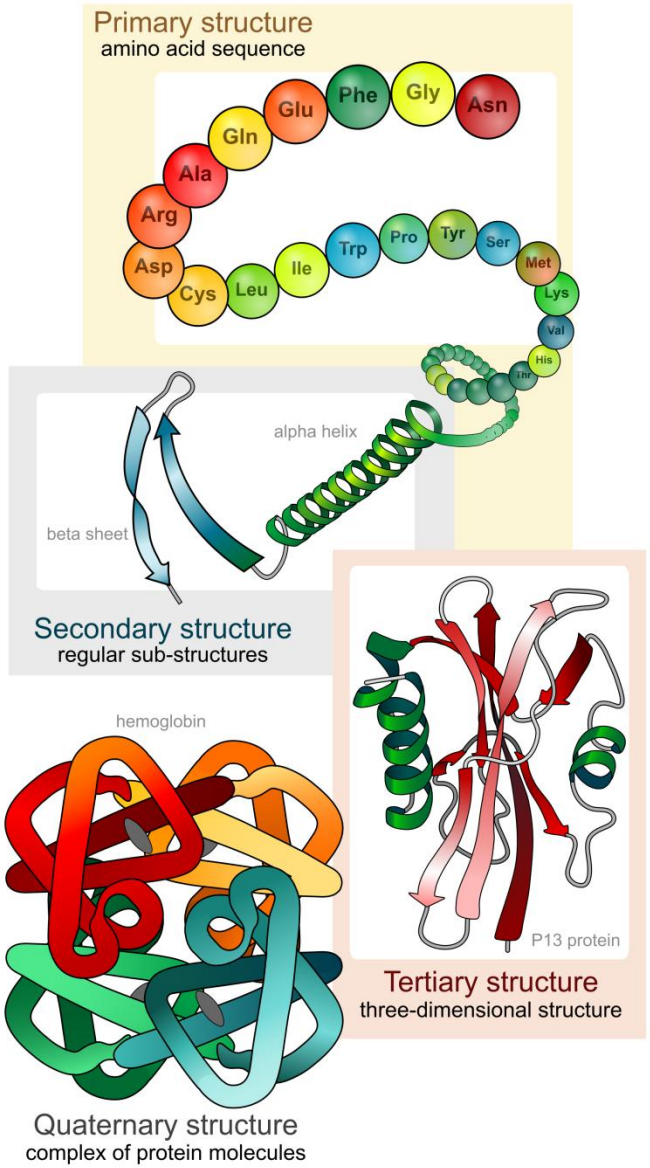
- Protein is a large molecule composed of one or more chains of amino acids in a specific order, which is determined by the base sequence of nucleotides in the DNA coding. Proteins differ from each other primarily in their sequence of amino acids, which is dictated by the nucleotide sequence of their genes.

Building block of protein

- Proteins are polymers of amino acids joined together by peptide bonds. Hence, the building blocks of protein are called amino acids. There are 20 different types of amino acids that the human body needs to function correctly. Eleven of these amino acids are produced naturally in the body, while the other nine need to be acquired by consuming food.

Different protein structures

- Four levels of protein structures can be specify,
  - **Primary structure:** the linear arrangement of amino acids in a protein and the location of covalent linkages such as disulfide bonds between amino acids.
  - **Secondary structure:** areas of folding or coiling within a protein; examples include alpha helices and pleated sheets, which are stabilized by hydrogen bonding.

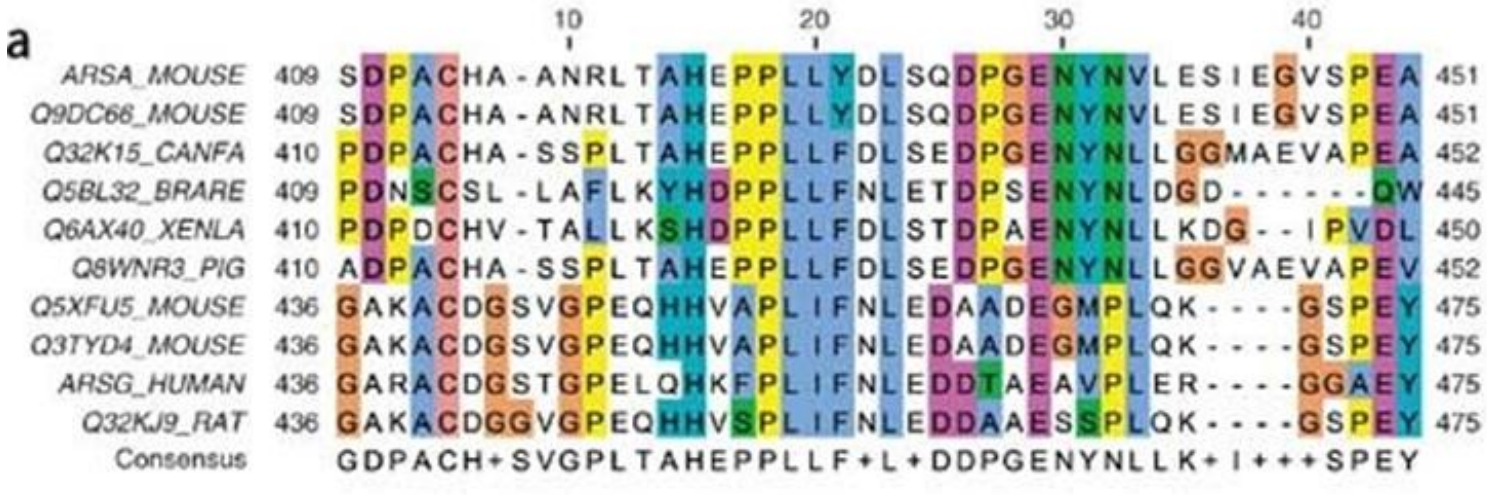


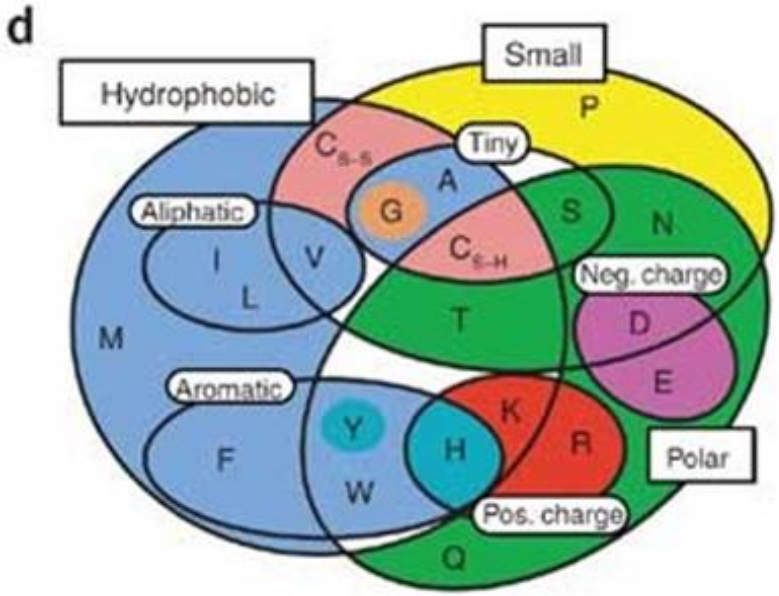
- **Tertiary structure:** the final three-dimensional structure of a protein, which results from a large number of non-covalent interactions between amino acids.

- **Quaternary structure:** non-covalent interactions that bind multiple polypeptides into a single, larger protein. Hemoglobin has quaternary structure due to association of two alpha globin and two beta globin polyproteins.

Software to check the structure: Cytoscape

Protein sequence





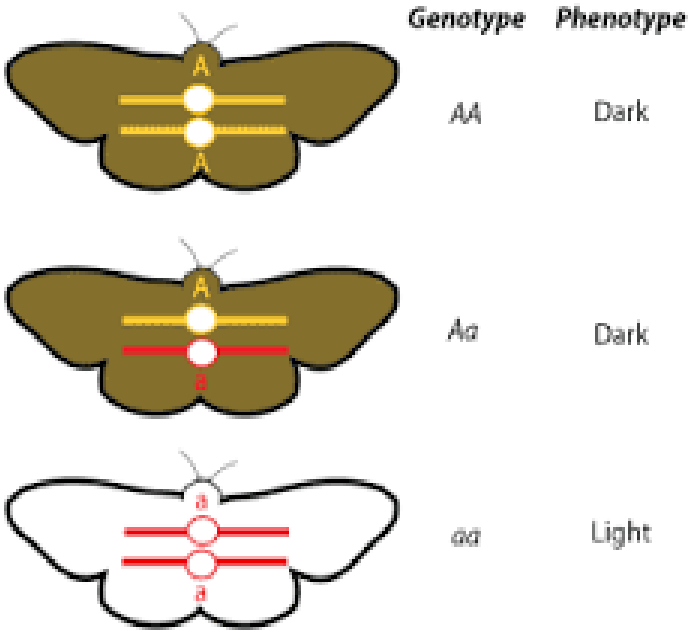
Protein complex

- **Protein complex** is a group of two or more associated polypeptide chains. The different polypeptide chains may have different functions. This is distinct from a multienzyme polypeptide, in which multiple catalytic domains are found in a single polypeptide chain.

- Protein complexes are a form of quaternary structure. Proteins in a protein complex are linked by non-covalent protein–protein interactions, and different protein complexes have different degrees of stability over time.

Genotype and Phenotype

- Genotype is the genetic pattern of a cell, an organism or an individual usually with reference to a specific characteristic under considered.
- Phenotype is the combination of an organism's observable characteristics such as its morphology, development, biochemical or physiological properties, behavior and products of behavior.



	Genotype	Phenotype
Example	DNA, susceptibility to diseases	Hair color, eye color, weight, the ability to roll one's tongue
Depends upon	The hereditary information that was given to an individual by their parents.	Genotype and the influence of the environment.
Inheritance	Partly inherited by offspring, as one of the two alleles is passed on during reproduction.	Cannot be inherited.
Contains	All the hereditary information of an individual, even if those genes are not expressed.	Expressed genes only.
Can be determined by	Genotyping – using a biological assay, such as PCR, to find out what genes are on an allele. (Inside the body)	Observation of the individual. (Outside the body)

Mutation

- A change in the nucleotide sequence of the gen ome of an organism or virus, sometimes resultin g in the
  appearance of a new character or trait not foun d in the parental type.
- The process by which such a change occurs, eit her through an alteration in the nucleotide sequ ence coding
  for a gene or through a change in the physical a rrangement of the genetic material.
- The nucleotide sequence, trait, or individual tha t results from such a change.

Types of Mutations

1) **Substitution**
   A substitution is a mutation that exchanges one base for another (i.e., a change in a single "chemical letter" such as switching an A to a G). Such a substitution could: change a codon to one that encodes a different amino acid and cause a small change in the protein produced. For example, sickle cell anemia is caused by a substitution in the beta-hemoglobin gene, which alters a single amino acid in the protein produced.

2) Insertion

An insertion changes the number of DNA bases in a gene by adding a piece of DNA. As a result, the protein made by the gene may not function properly.

3) Deletion

A deletion changes the number of DNA bases by removing a piece of DNA. Small deletions may remove one or a few base pairs within a gene, while larger deletions can remove an entire gene or several neighboring genes. The deleted DNA may alter the function of the resulting protein(s).

4) Duplication

A duplication consists of a piece of DNA that is abnormally copied one or more times. This type of mutation may alter the function of the resulting protein.

CTGGAG

CTCTGGAG

5) Frameshift

This type of mutation occurs when the addition or loss of DNA bases changes a gene's reading frame. A reading frame consists of groups of 3 bases that each code for one amino acid. A frameshift mutation shifts the grouping of these bases and changes the code for amino acids. The resulting protein is usually nonfunctional. Insertions, deletions, and duplications can all be frameshift mutations.

Reasons for mutation

1. **DNA fails to copy accurately**
   Most of the mutations that we think matter to evolution are "naturally-occurring." For example, when a cell divides, it makes a copy of its DNA - and sometimes the copy is not quite perfect. That small difference from the original DNA sequence is a mutation.
2. **External influences can create mutations**
   Mutations can also be caused by exposure to specific chemicals or radiation. These agents cause the DNA to break down. This is not necessarily unnatural - even in the most isolated and pristine environments, DNA breaks down. Nevertheless, when the cell repairs the DNA, it might not do a perfect job of the repair. So the cell would end up with DNA slightly different than the original DNA and hence, a mutation.

Effects of mutation

Somatic mutations

Some mutations cannot be passed on to offspring and do not matter for evolution. Somatic mutations occur in non-reproductive cells and won't be passed onto offspring.

Ex: the golden color on half of this Red Delicious apple was caused by a somatic mutation. Its seeds will not carry the mutation.

Germ line mutations

The only mutations that matter to large-scale evolution are those that can be passed on to offspring. These occur in reproductive cells.

Difference between mutation and variation

- Mutation is a change that occurs in the genome of an individual organism. Variation is something that occurs within a species. Therefore, variation is the differences between individuals of a species. Due to mutation occur in reproduction cells variation can occur.

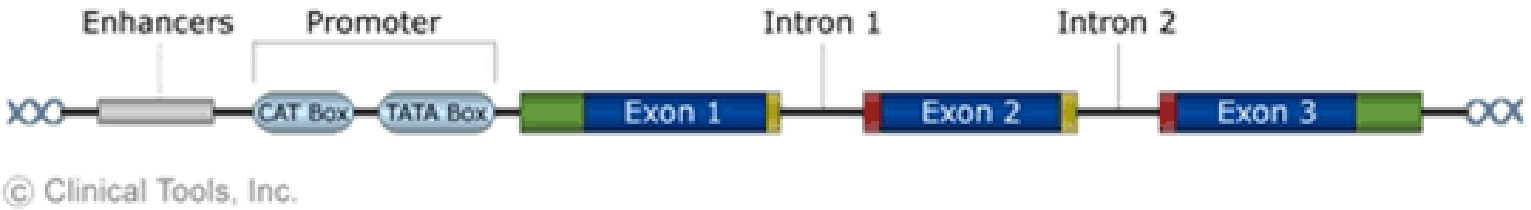
Gene expression

- Is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, the product is a functional RNA.

The process of gene expression involves two main stages:

- **Transcription:** the production of messenger RNA (mRNA) by the enzyme RNA polymerase, and the processing of the resulting mRNA molecule.
- **Translation**: the use of mRNA to direct protein synthesis, and the subsequent post-translational processing of the protein molecule.

- A structural gene involves a number of different components:



- Exons. Exons code for amino acids and collectively determine the amino acid sequence of the protein product. It is these portions of the gene that are represented in final mature mRNA molecule.
- Introns. Introns are portions of the gene that do not code for amino acids, and are removed (spliced) from the mRNA molecule before translation.

Gene control regions

- **Start site**. A start site for transcription.
- **A promoter**. A region a few hundred nucleotides 'upstream' of the gene (toward the 5' end). It is not transcribed into mRNA, but plays a role in controlling the transcription of the gene. Transcription factors bind to specific nucleotide sequences in the promoter region and assist in the binding of RNA polymerases.
- **Enhancers**. Some transcription factors (called activators) bind to regions called 'enhancers' that increase the rate of transcription. These sites may be thousands of nucleotides from the coding sequences or within an intron. Some enhancers are conditional and only work in the presence of other factors as well as transcription factors.
- **Silencers**. Some transcription factors (called repressors) bind to regions called 'silencers' that depress the rate of transcription.

What is the use.....

- Future jobs..
  - Forensic analysts
  - NGS providers
  - Research analysts (medical, agriculture)
  - Computational biology data managers
  - Academic
- Research areas...
  - Level 4 project????
  - Higher studies