Introduction to Bioinformatics

Goal

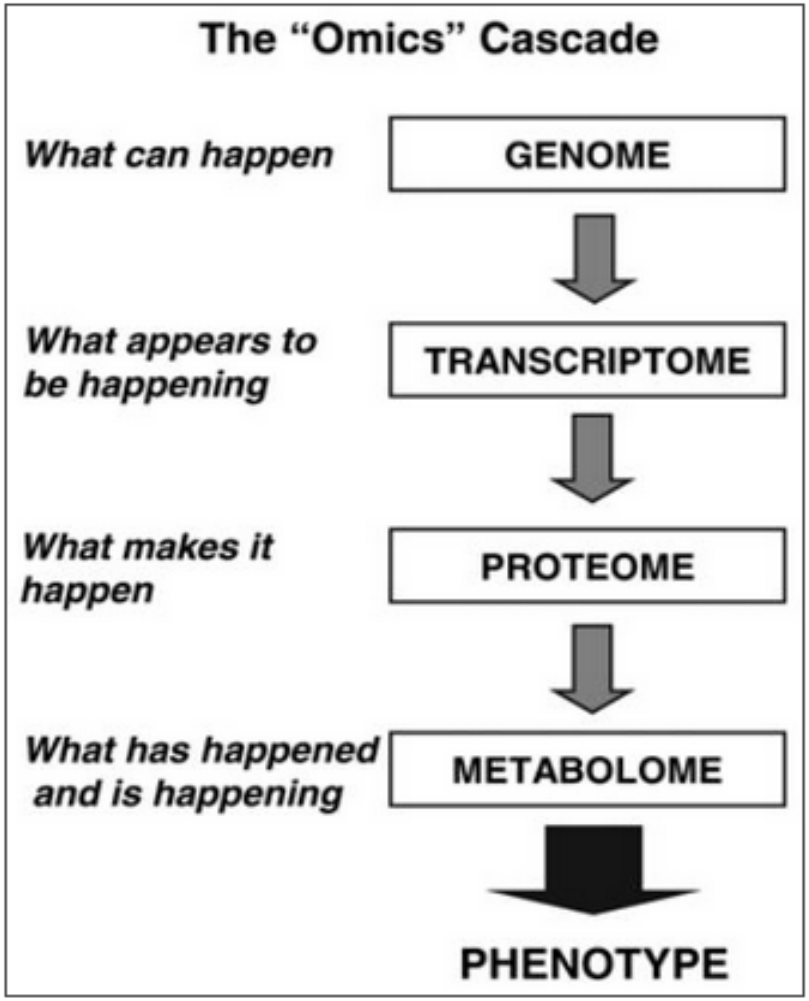
Goal of molecular cell biology - to understand the physiology of living cells in terms of the information that is encoded in the genome of the cell

How computer science can help in achieving this goal?

⇒ **To identify where are the genes in the genome, which gene is expressed when, where and how much, what factors affect its expression, what's its function, what happens in the disease state, etc.**

The name 'bioinformatics' was coined by Paulien Hogeweg in 1979, for the study of informatic processes in biological systems.

Various Omics studies, *viz.*, Genomics, Transcriptomics, Proteomics & Metabolomics are data-driven fields that aim to answer the question of how genomes code for living organisms.



Major inputs from CS – develop algorithms for mining meaningful information from biological data, and develop efficient data storage and data retrieval systems for managing large volumes of data

Biological Data: Levels of Organization

Central Dogma of MB

Central Dogma of Bioinformatics

Metabolic Pathways, Interacting Networks

Intercellular Interactions

Organs/Tissues Physiology

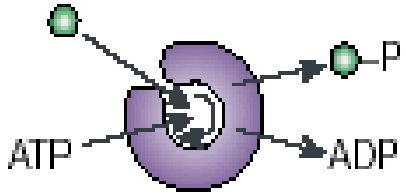
Ecological modeling
Inter-species interactions

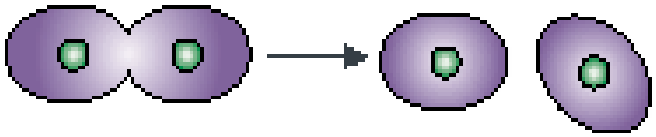
DNA → mRNA → Protein → Enzyme → Reaction network → Cell physiology

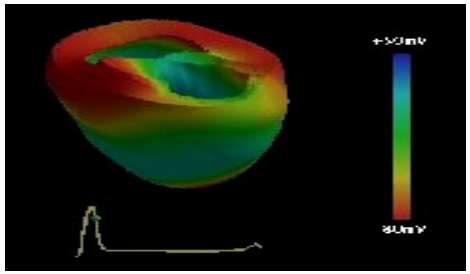
...TACCCGATGGCGAAATGC...

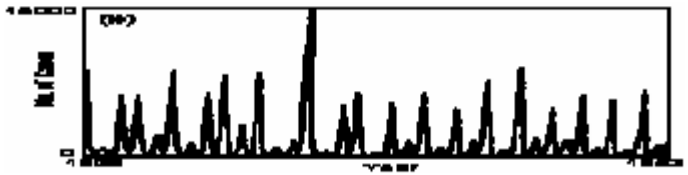
...AUGGGCUACCGCUUUACG...

...Met-Gly-Tyr-Arg-Phe-Thr...









Sequence/Structure Alignment, Pattern Recognition, Molecular Modeling

Network Modeling, Dynamical Systems Modeling, Differential Eqns, Graph Theory, Chaos Theory, Pattern Formation, Characterization, Time-series data analysis

If it were required, in a single model, to span all the scales from

molecular motions ⇒ **cell responses**

(nanometers/picosecs) **(micrometers/secs)**

- theoretical approach to cell physiology would be beyond grasp, both computationally & intellectually

Fortunately, considerable progress can be made – at any given level of hierarchy – independent of the successes or failures at levels above/below

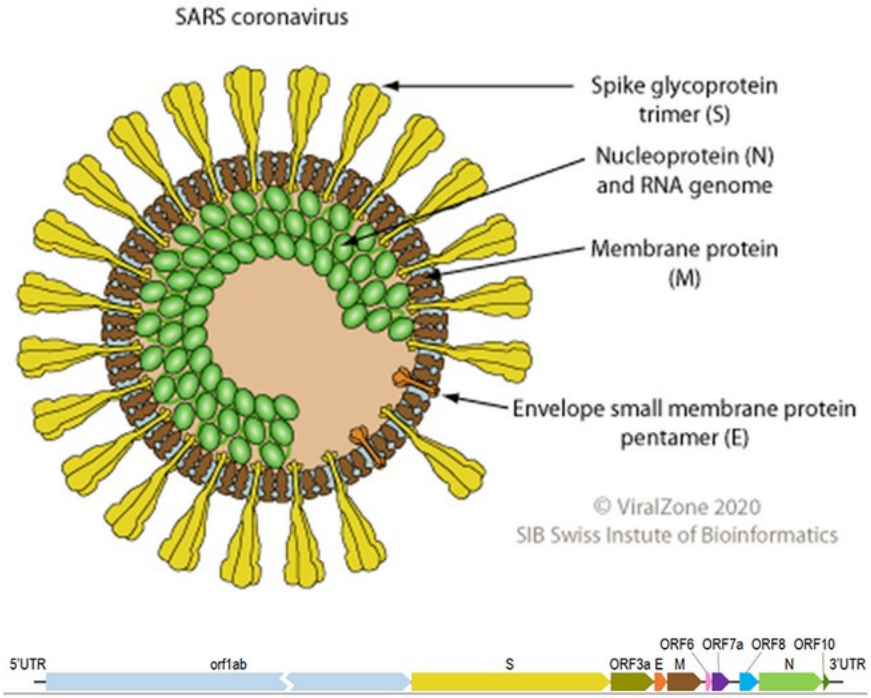
Systems biology is aiming towards achieving this goal of understanding the functional behaviour of a cell/tissue as a whole.

Disease - COVID19

- When a new virus strikes the population, such as we saw in the case of COVID-19, no specific treatment is available

What kind of sequence analysis can help in combating the disease?

SARS-CoV-2



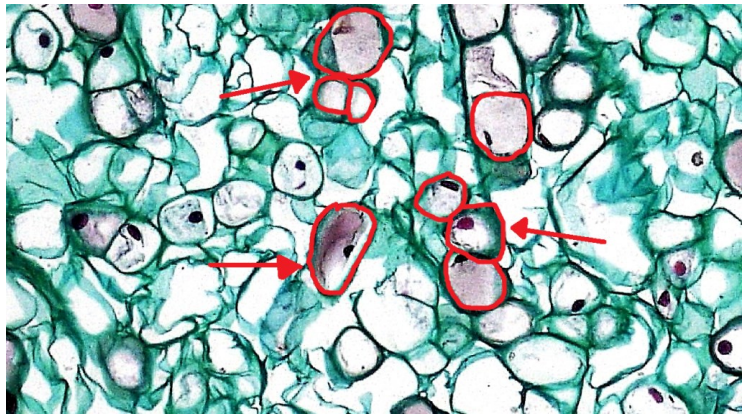
What kind of Bioinformatics analysis can we carry out to know about the virus causing COVID-19?

- How to identify if a person is infected with SAR-COV-2?
- Is it the only known human coronavirus?
- Comparing its genome with other viral genomes – to identify its closest relative
- What proteins aid in its transmission and infection?
- Identifying drug targets and develop vaccines
- What organs/tissues are affected by its infection?
- Its rate of propagation
- Is it mutating and becoming more virulent, or milder with time
- etc.

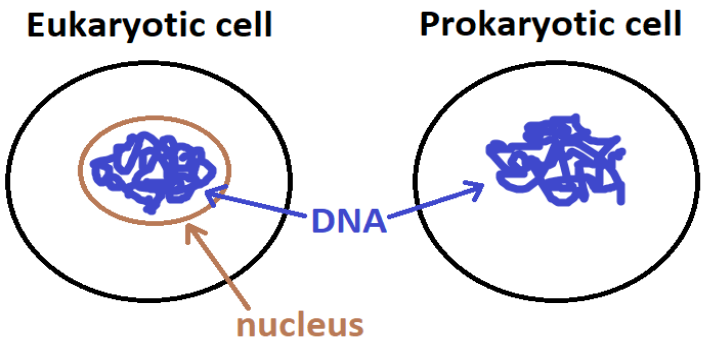
The Cell

Cell - the basic building block of all living creatures.

All forms of "omics" measure large quantities of something found inside cells.



Cells are of two types:

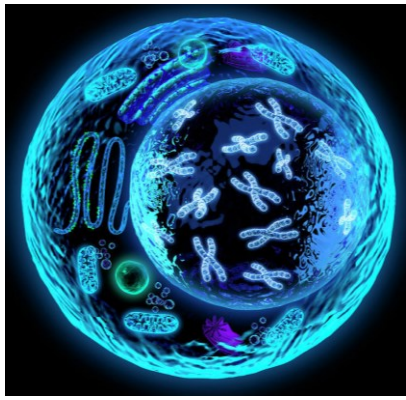


Original image

Apart from prokaryotes and eukaryotes, there is a third category called Viruses, acellular entities. May contain DNA or RNA as their nuclear material.

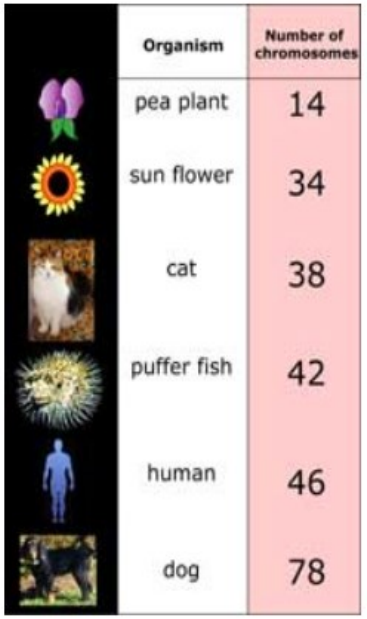
Cells and Chromosomes

E. B. Wilson: "the key to every biological problem must be sought in the cell; for every living organism is, or at some time has been, a cell."



~ 10^{13} cells that form a human body, the whole organism has been generated by cell divisions from a single cell

Cells are the fundamental units of life - the vehicle for all the hereditary information that defines each species.



Organism	Number of chromosomes
pea plant	14
sun flower	34
cat	38
puffer fish	42
human	46
dog	78

Genome – the total DNA content of an organism

Chromosomes – are physically separate molecules that range in length from ~ 50 - 250Mbp

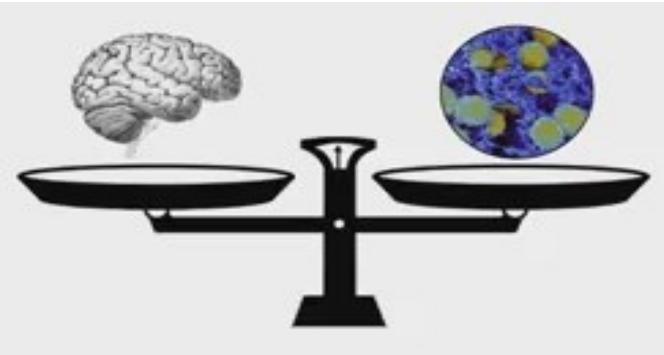
In mammals and many other eukaryotes, the chromosomes occur in homologous pairs, called diploids, except for sex chromosomes.

Cells and Genomes

Do we carry cells of any other organism within us, apart from human cells?

How human are we?

- We have 10 trillion human cells and 100 trillion microbial cells: with respect to cell count we are just 10% human
- Our genome has 20-30K genes, our microbiome has 2-20M genes: with respect to genes we are 0.1-1% human
- Our microbiome weighs ~ 3 pounds, about the same weight as our brain, and maybe as important to our well being, if not more!
- We share 99.9% of our genome with other individuals, but we share only 10% of our microbiome
- Microbiota include bacteria, archaea, viruses, eukaryota



DNA and Genes

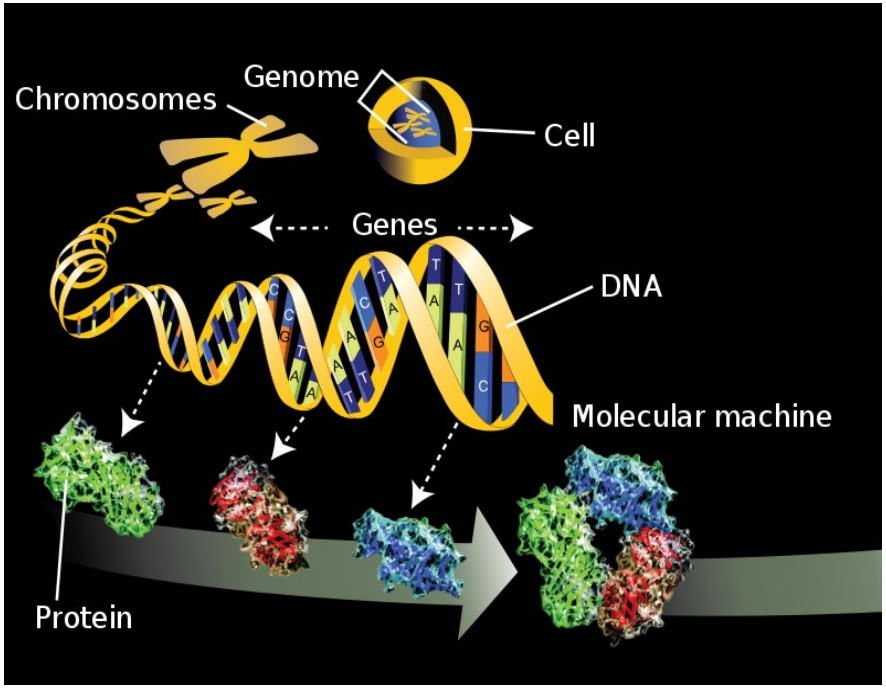
Human DNA - a long sequence of 3 billion letters, inside every cell in our body. There are ~ 37.2 trillion cells in our body.

A "gene" is a particular segment of DNA that encodes instructions for making a protein, hence genes are referred to as the "coding" part of the genome.

~ 25,000 genes in the human genome covering ~ 2-5% of the genome.

Function of remaining 98% of the "non-coding" part of the genome?

DNA: Deoxyribonucleic acid



DNA

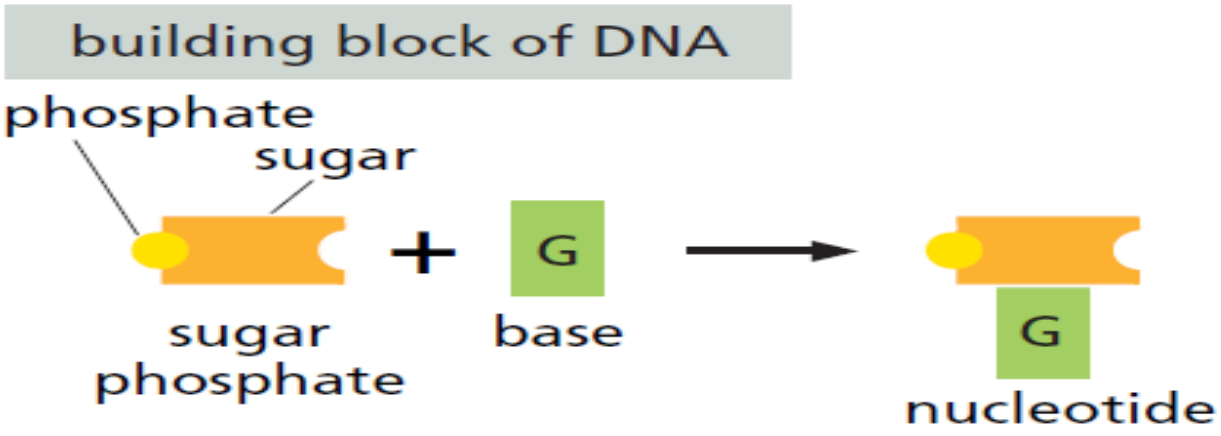
DNA (Deoxyribonucleic acid):

Composed of four basic units - called nucleotides

Each nucleotide contains - a sugar, a phosphate and one of the four bases:

Adenine (A), Thymine (T),

Guanine (G), Cytosine (C).

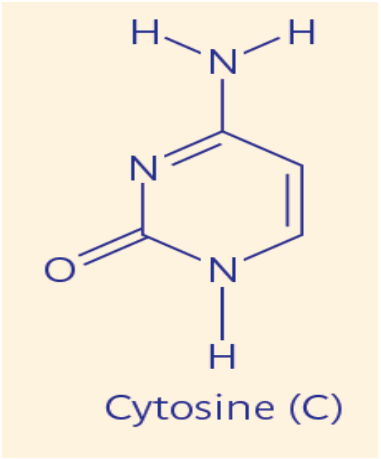


DNA

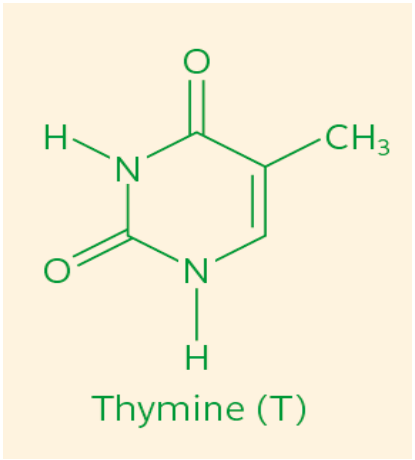
Bases: are ring-shaped and come in four types which fit together in pairs - this pairing forms the basis of information carrying capacity of DNA.

These are categorized as:

Pyrimidines

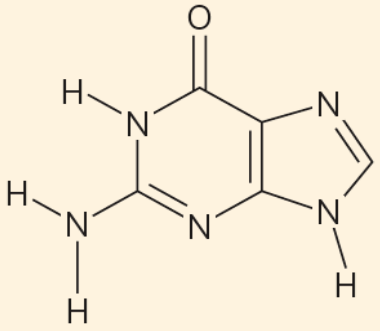


Cytosine (C)

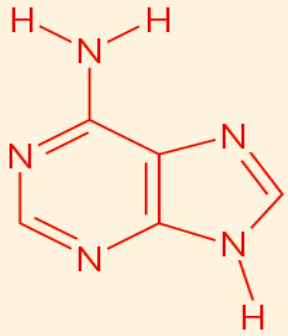


Thymine (T)

Purines



Guanine (G)



Adenine (A)

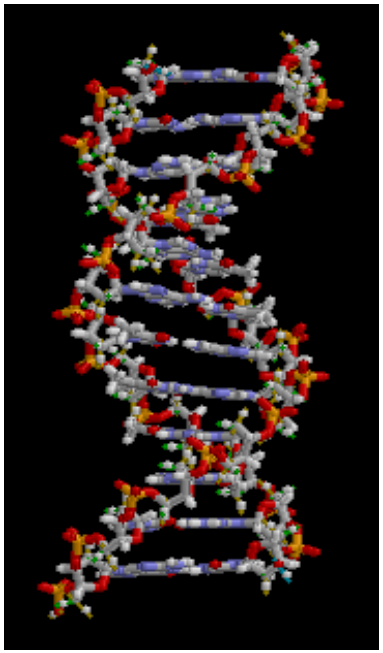
Which of these form base-pairs?

DNA

DNA is double-stranded - the two strands of DNA wind around each other to form a double helix.

Information in one strand is a "mirror copy" of the information in the other strand, achieved by base-pairing:

A ⇔ T, G ⇔ C



So, if the sequence on one strand is GATTACA, what is the sequence of the other strand?

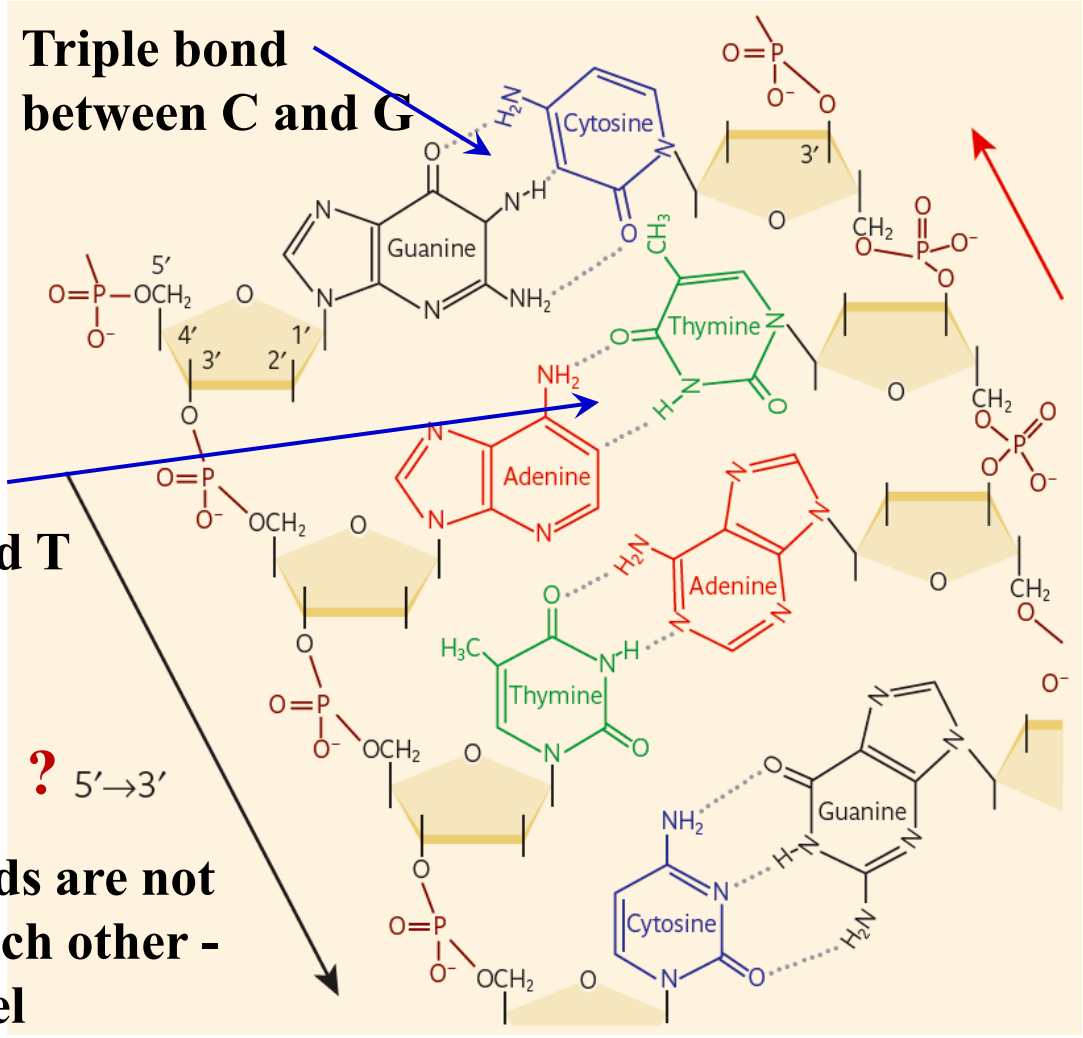
Importance of double-stranded nature of DNA:

- **Facilitates DNA replication**
- **Error-correct the genome**
- **Provides stability**

From a computational perspective, the sequence of only one strand is needed.

DNA

Base Pairing: If two polynucleotide strands face each other, sugar-phosphate backbone runs down each side, and complementary pairs of bases in the middle spontaneously form hydrogen bonds:



Note: the two strands are not mirror images of each other - they run antiparallel

Double-Stranded DNA: **If the sequence in the forward strand in 5′ to 3′ direction is:**

5′ CATTGCCAGT 3′

Then what is the sequence on the reverse strand when read in 5′ to 3′ orientation?

3' GTAACGGTCA 5'

5' ACTGGCAATG 3'

DNA

If the sequence in the forward strand in 5′ to 3′ direction is:

5′ CATTGCCAGT 3′

Then what is the sequence on the reverse strand when read in 5′ to 3′ orientation?

First write its complement:

5′ CATTGCCAGT 3′

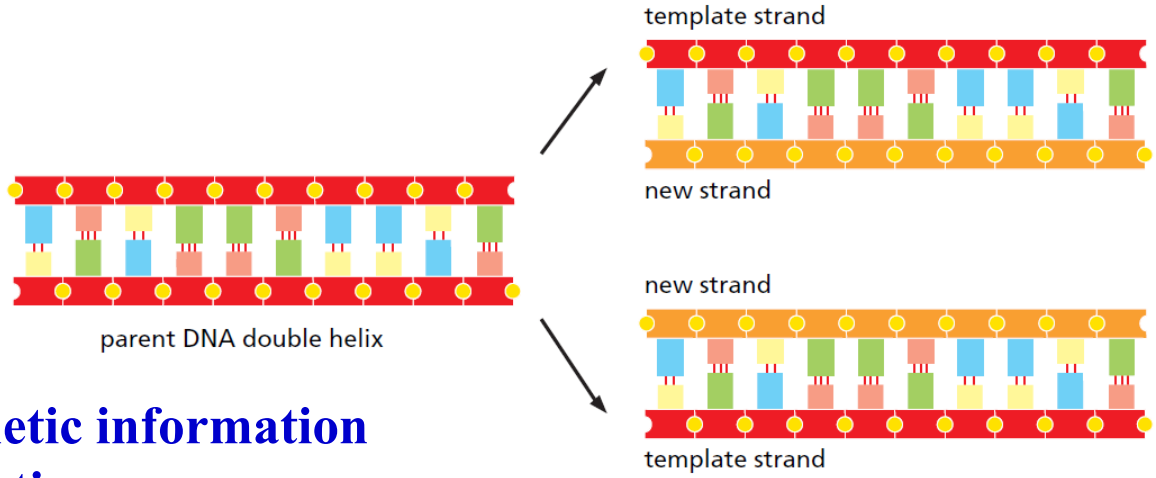
3′ GTAACGGTCA 5′

When read in 5′ to 3′ orientation, the sequence on the reverse strand is:

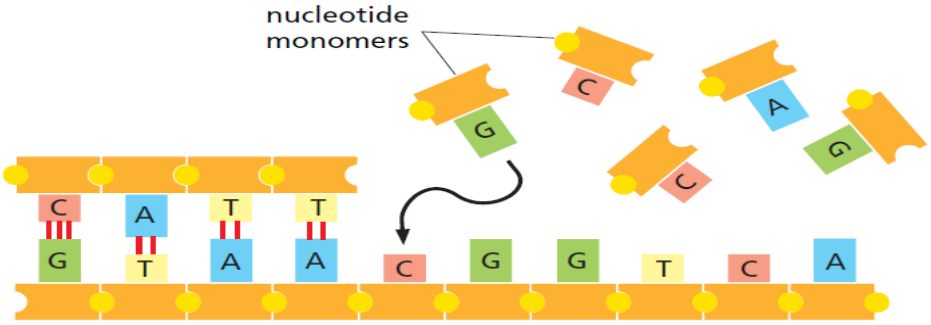
5′ ACTGGCAATG 3′

DNA Replication

In living cells DNA is not synthesized as a free strand in isolation, but on a template formed by a pre-existing DNA strand.

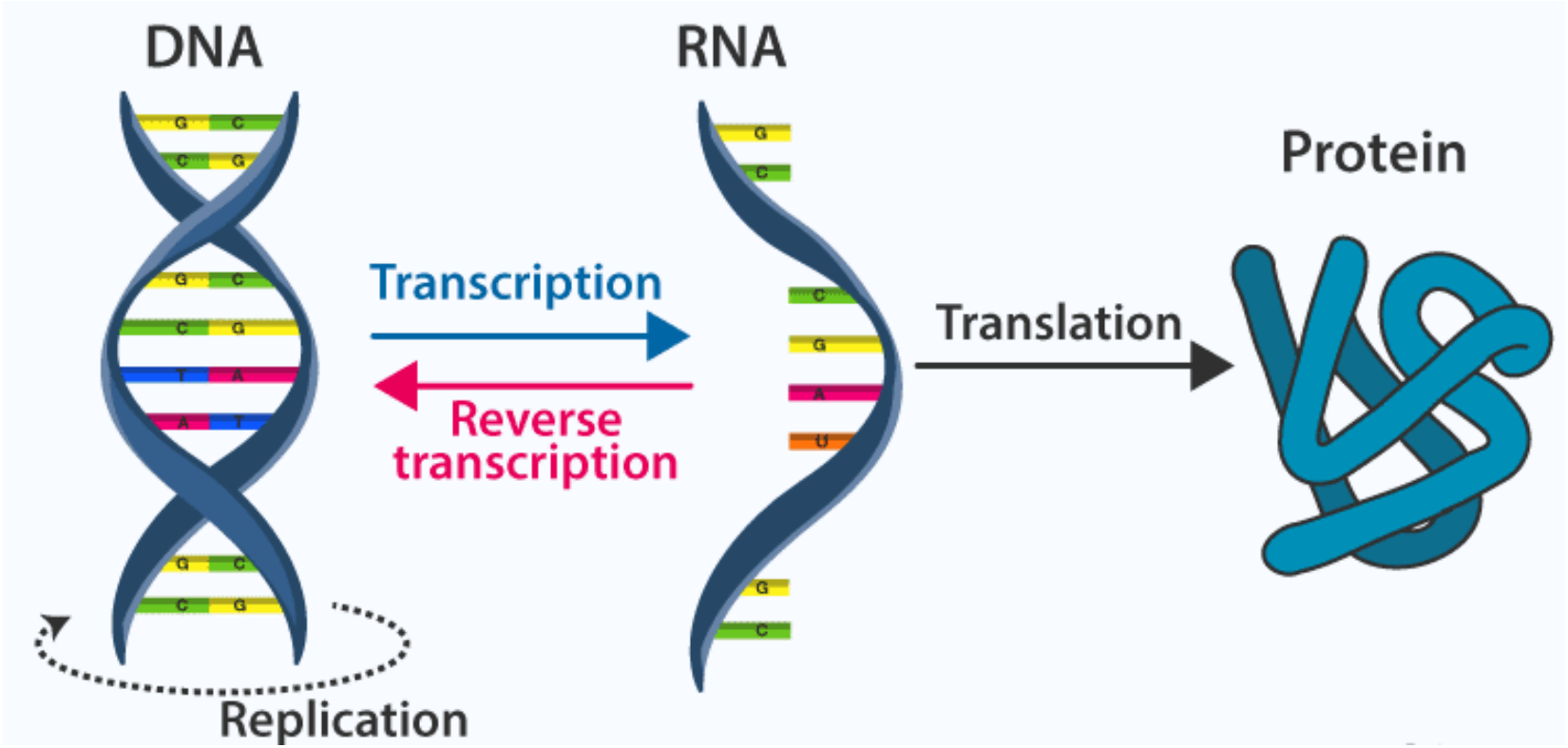


Copying of genetic information by DNA replication



templated polymerization

Central Dogma of Molecular Biology



Ribonucleic Acid (RNA):

It is single-stranded molecule

Composed of four basic units - called nucleotides:

Each nucleotide contains - a sugar (ribose), a phosphate and one of the four bases: Adenine (A), Uracil (U), Guanine (G), Cytosine (C)

RNA polynucleotide strand is built by creating a phosphodiester bond between nucleotides.

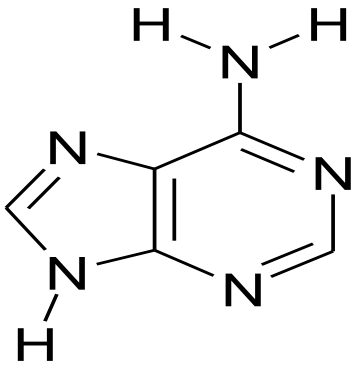
Intra-strand base pairing is a characteristic feature of RNA

Base Pairing – formed by weak H-bonds and follows the following complementarity rule:

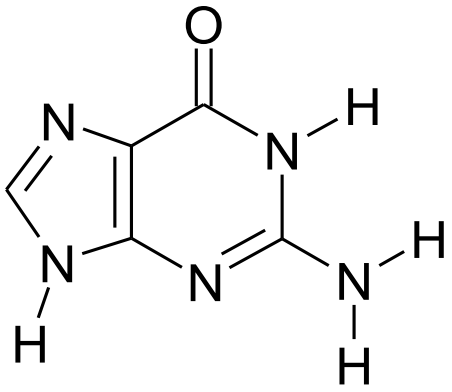
A ⟷ U, G ⟷ C, and G ⟷ U

Ring Structure of Nucleic Acid bases

Adenine (A)

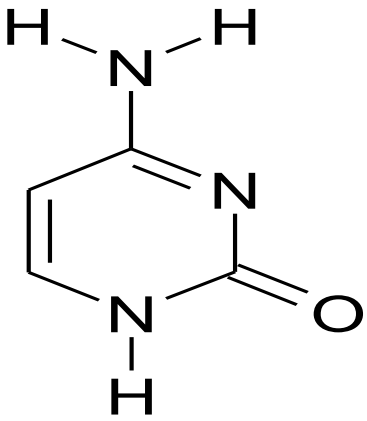


Guanine (G)

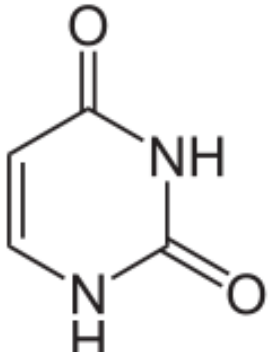


Purines

Cytosine (C)



Uracil (U)

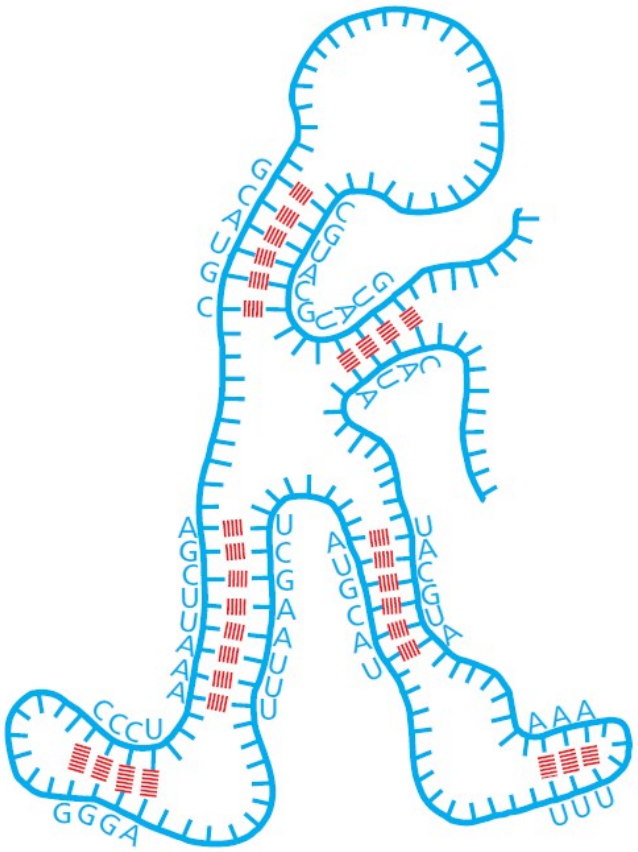


Pyridimines

Note: No CH3 in Uracil as in Thymine

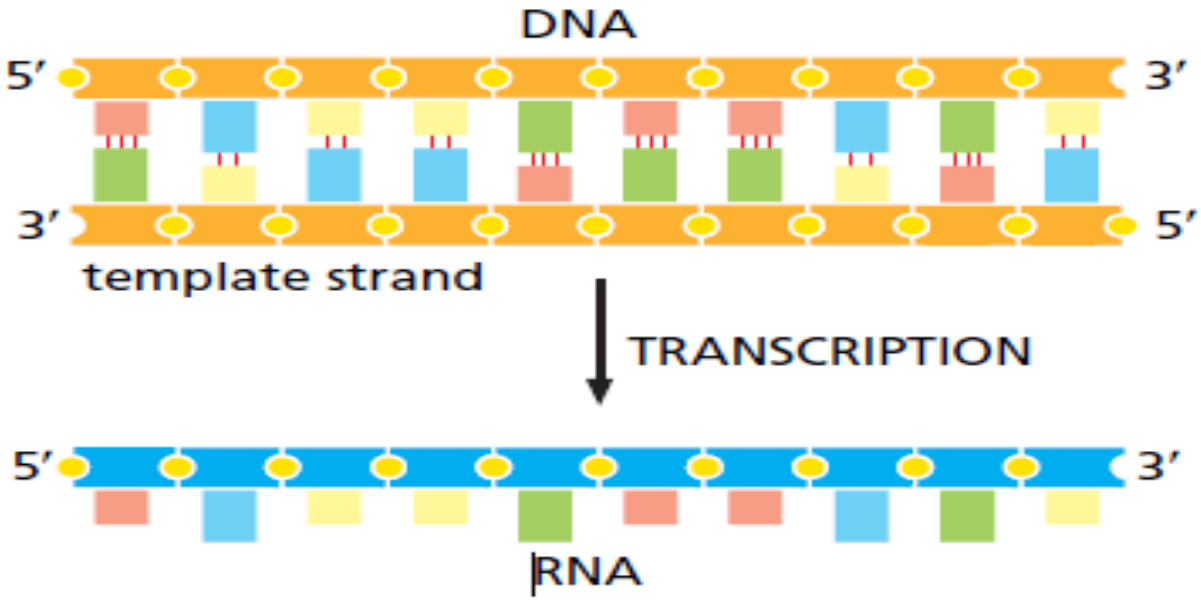
RNA

Nucleotide pairing between different regions of the RNA polymer chain causes the molecule to adopt a distinctive shape - enables it to recognize other molecules by selective binding, or, catalyze chemical changes in the molecules that are bound.



RNA Synthesis:

RNA is also read in the 5′ to 3′ orientation.



RNA molecules that are copied from the genes (which ultimately direct the synthesis of proteins) are called messenger RNA (mRNA) molecules.

RNA Synthesis

1. If the following DNA sequence is the forward strand:

5′ CATTGCCAGT 3′

What will be the sequence of the RNA strand synthesized?

2. If the following DNA sequence is used as template for RNA synthesis:

5′ CATTGCCAGT 3′

Give the sequence of the RNA strand read in 5′ to 3′ orientation.

RNA Synthesis

1. If the DNA sequence in the forward strand is given:

5′ CATTGCCAGT 3′

Template sequence used for RNA synthesis is its complement:

5′ CATTGCCAGT 3′

3′ GTAACGGTCA 5′ **template**

The synthesized RNA sequence is the reverse complement of the template:

3′ GTAACGGTCA 5′ **template**

5′ CAUUGCCAGU 3′ **RNA**

- i.e., synthesized RNA sequence is basically the DNA sequence in the forward strand with T replaced by U

RNA Synthesis

If the following DNA sequence is used as template for RNA synthesis:

5′ CATTGCCAGT 3′

First write its complement:

5′ CATTGCCAGT 3′

3′ GUAACGGUCA 5′ complement

Then the synthesized RNA sequence in 5′ to 3′ orientation is:

5′ ACUGGCAAUG 3′ RNA

- i.e., synthesized RNA sequence is basically the complement of the template DNA sequence with T replaced by U, when read in the 5′ to 3′ orientation

RNA Synthesis:

There are other RNA molecules also obtained from genes. The final product in such cases is RNA.

These RNAs are known as noncoding RNAs because they do not code for protein.

e.g., in yeast *Saccharomyces cerevisiae*, over 1200 genes (~15%) produce RNA as their final product; Humans may produce on the order of 10,000 noncoding RNAs.

These RNAs, like proteins, serve as enzymatic, structural, and regulatory components for a wide variety of processes in the cell.

TABLE 6–1 Principal Types of RNAs Produced in Cells	
Type of RNA	Function
mRNAs	Messenger RNAs, code for proteins
rRNAs	Ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis
tRNAs	Transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids
snRNAs	Small nuclear RNAs, function in a variety of nuclear processes, including the splicing of pre-mRNA
snoRNAs	Small nucleolar RNAs, help to process and chemically modify rRNAs
miRNAs	MicroRNAs, regulate gene expression by blocking translation of specific mRNAs and cause their degradation
siRNAs	Small interfering RNAs, turn off gene expression by directing the degradation of selective mRNAs and the establishment of compact chromatin structures
piRNAs	Piwi-interacting RNAs, bind to piwi proteins and protect the germ line from transposable elements
lncRNAs	Long noncoding RNAs, many of which serve as scaffolds; they regulate diverse cell processes, including X-chromosome inactivation

Note: rRNA, tRNA and snRNA play an important role in protein synthesis

Protein Synthesis

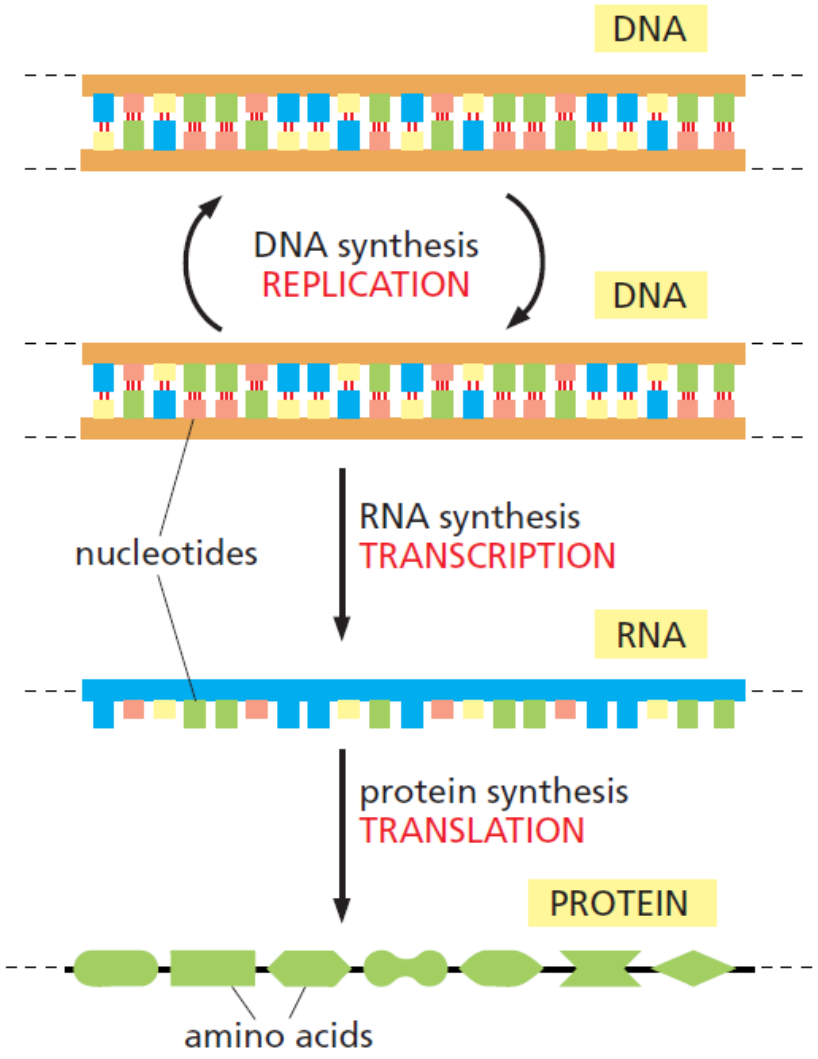
Proteins are synthesized from DNA in a two-step process:

Each chromosome has several genes that code for various traits in the body.

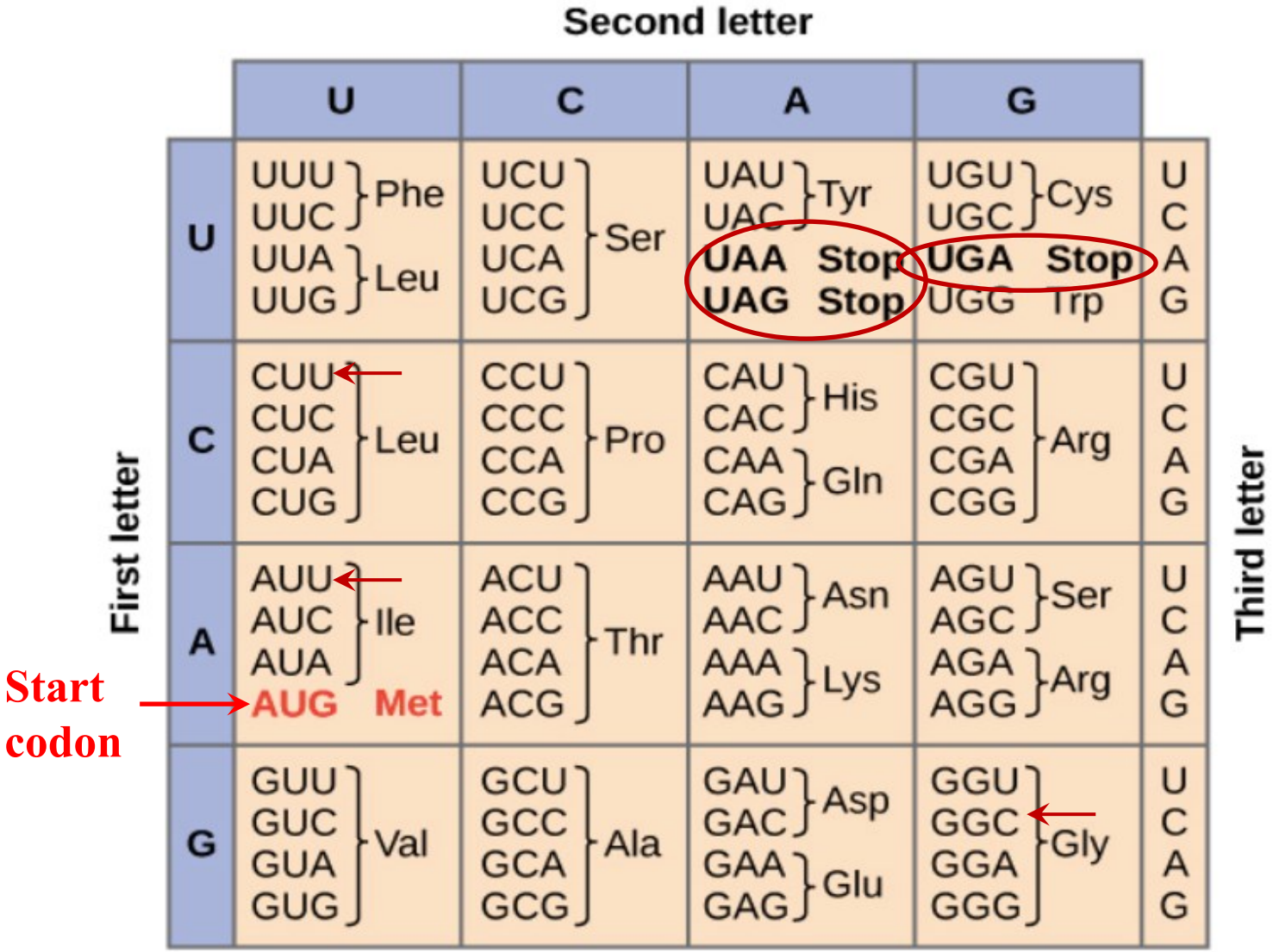
- from enzymes to the color of eye

RNA molecules direct synthesis of proteins in a complex process called translation.

- information in mRNA is read out in groups of three nucleotides, called codons.



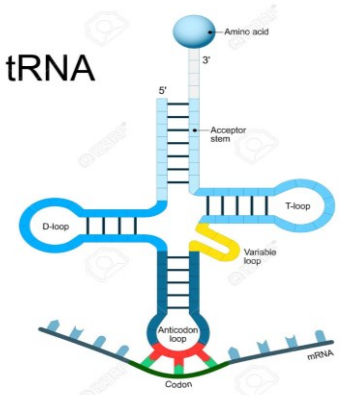
The Genetic Code



The genetic code is degenerate

Protein Synthesis

Using the genetic code, obtain the amino acid sequence synthesized from the following mRNA sequence:

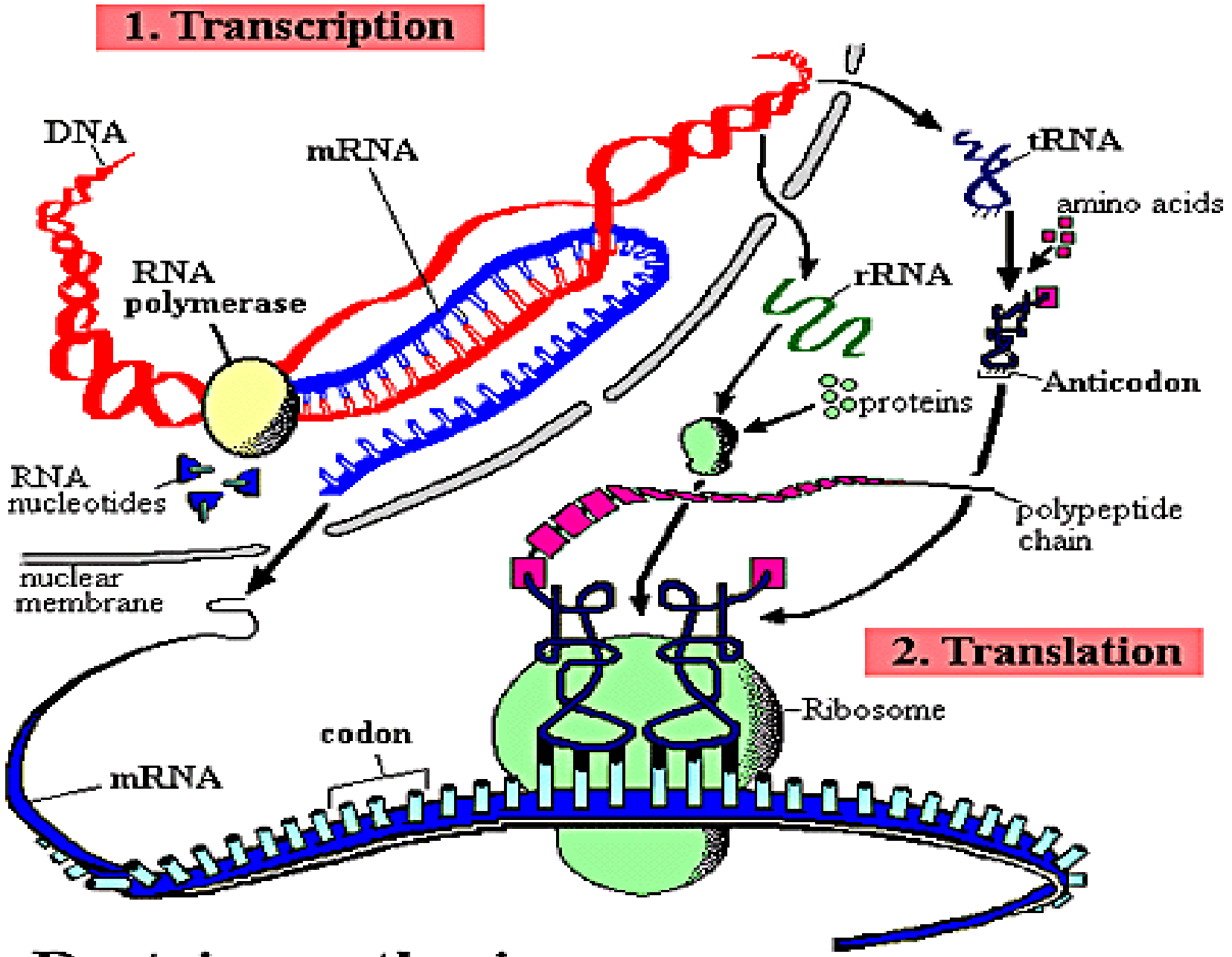


5′ ACU GGC AAU 3′

Thr Gly Asn

This genetic code is read out by a class of small RNA molecules, called transfer RNAs (tRNAs).

- **each type of tRNA attaches at one end a specific amino acid and at its other end has a specific sequence of 3 nucleotides, an anticodon that enables it to recognize, through base-pairing, a particular codon in the mRNA sequence.**
- **This process occurs on ribosome, a large multi-molecular machine composed of both proteins and ribosomal RNA.**



Protein synthesis

Proteins

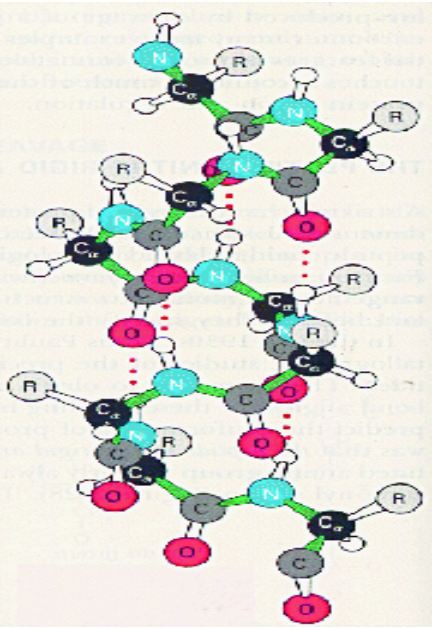
Like DNA and RNA, Proteins carry information in linear sequence on a 20-letter alphabet, called amino acids:

ATRVGTCWPRA

Protein structure is divided in 4 hierarchical levels:

- **Primary structure - represented by AA sequences**
- **Secondary structure - α-helices & β-sheets**
- **Tertiary and Quaternary structures - represented by 3D structures**

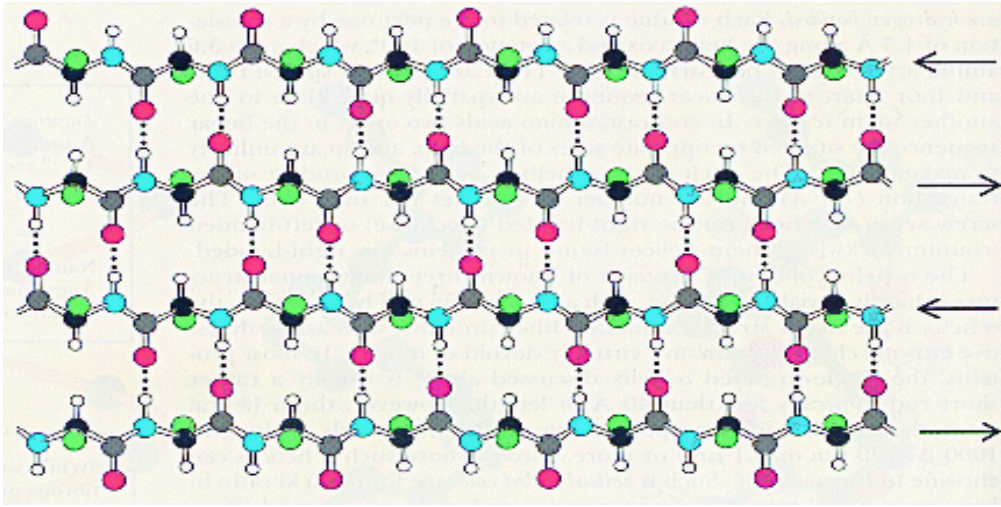
Primary Structure: ATRVGTCWPRA



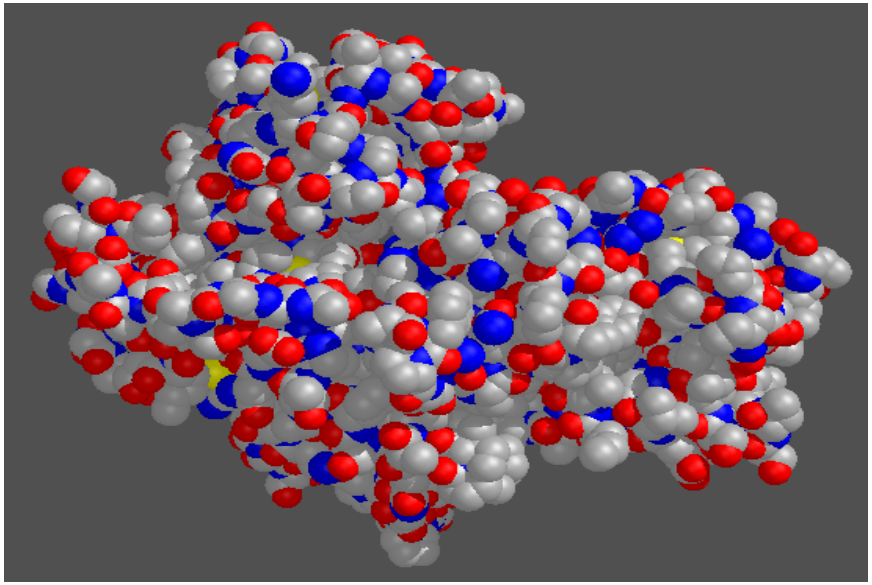
α-helix

Secondary Structures

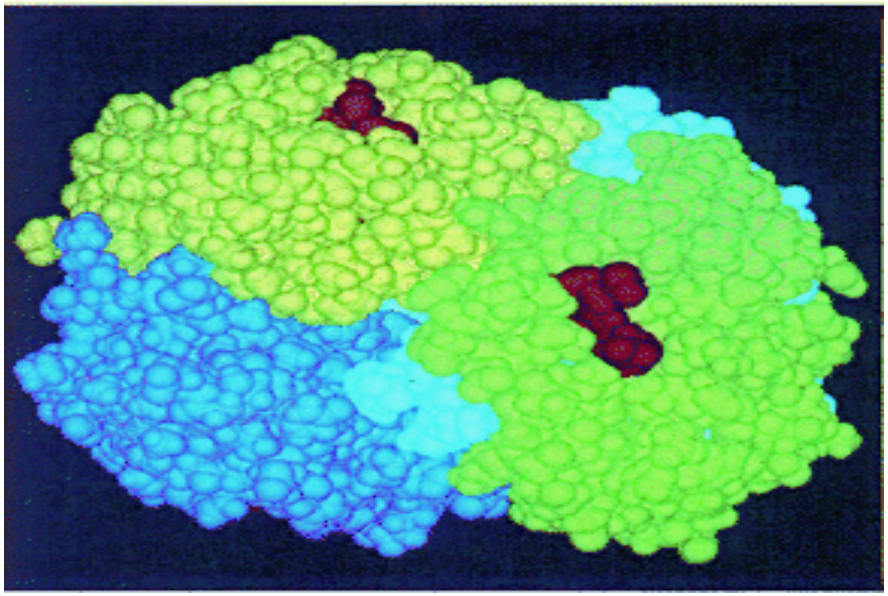
β-sheets



Tertiary Structure



Quaternary Structure



Function of Proteins

- **Proteins make up much of the cellular structure – hair, skin, fingernails, etc.**
- **Enzymes – catalyze chemical reactions within the cell**
- **Transcription factors – regulate the manner in which genes direct production of other proteins**
- **Receptors – proteins on the surface of cells act as receptors for hormones and other signaling molecules**
- **Recognize and bind to Nucleic acids (DNA, RNA) and Proteins – to carry out their functions in the cell**

Genes

Special sequences in the DNA code for genes:

- **Protein-coding genes, for which the final product is a protein.**
  - **same gene may give rise to more than one protein (~ 6 per gene in humans).**
- **Non-coding RNA genes - for which the final product is RNA**

Genotype – An organism's genotype is the set of genes that it carries.

Phenotype – An organism's phenotype is all of its observable characteristics which are influenced both by its genotype and by the environment e.g., height, hair colour, levels of hormones, etc.

Differences in the genotypes can produce different phenotypes

Genes for ear form are different, causing one of the cats to have normal ears and the other to have curled ears

Change in environment can also affect the phenotype. Pinkness is <u>not</u> encoded in the genotype of flamingos - the food they eat makes their phenotype white or pink - a natural pink dye, canthaxanthin, obtained from their diet of brine shrimp and blue-green algae

Genes

The biological function of a gene is to preserve and express the genetic information encoded within it

Genes are normally very stable entities

Genetic stability is not absolute, however.

Genes may occasionally become altered; these changes called mutations create new alleles.

Mutant genes are also stable entities and are inherited in the same way as normal, wild-type genes.

Genes

Normal diploid cells such as somatic cells of humans contain two sets of genes – one set inherited from each parent.

- corresponding genes derived from each parent are called alleles.

Together the two alleles govern the phenotype of an organism.

What is the percentage of genes in a genome?

Genes

Gene-fraction varies from 70% in prokaryotes to ~ 2 - 3% in humans

- **does this imply prokaryotes have more gene content than eukaryotes?**

- **Size of a prokaryotic genome? Eukaryotic genome?**

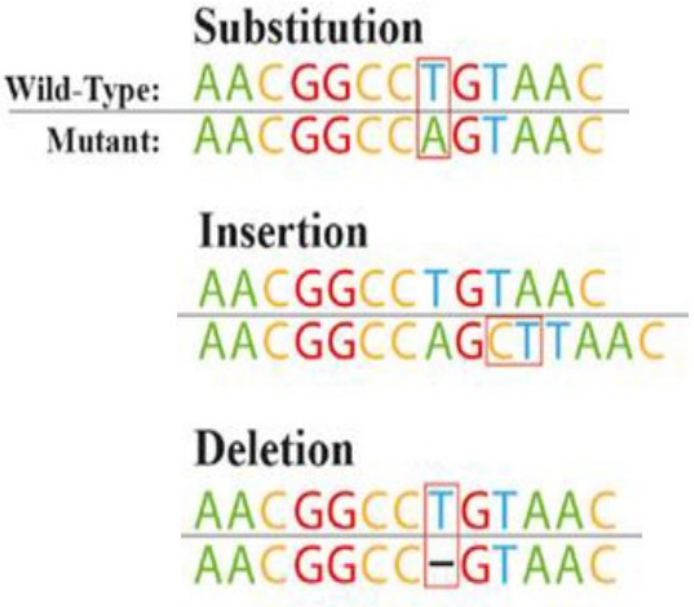
What's the function of remaining ~97-98% of human genome?

The remaining part of the genome consists of noncoding regions, whose functions may include providing chromosomal structural integrity and regulating where, when, and in what quantity proteins are made.

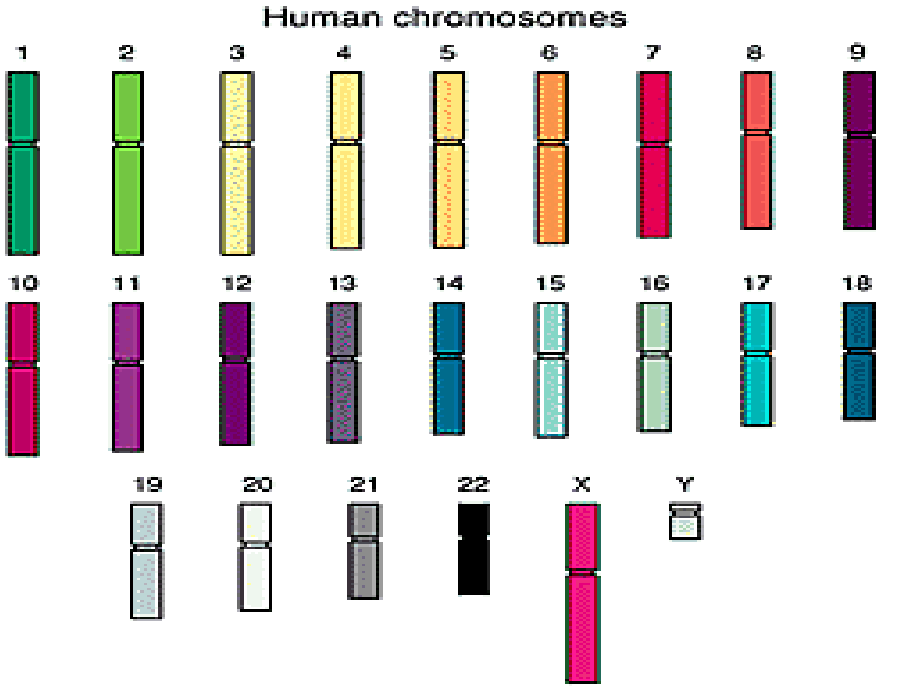
Mutations

Mutations - are local changes in the DNA content, caused by inexact replication and are of various kinds:

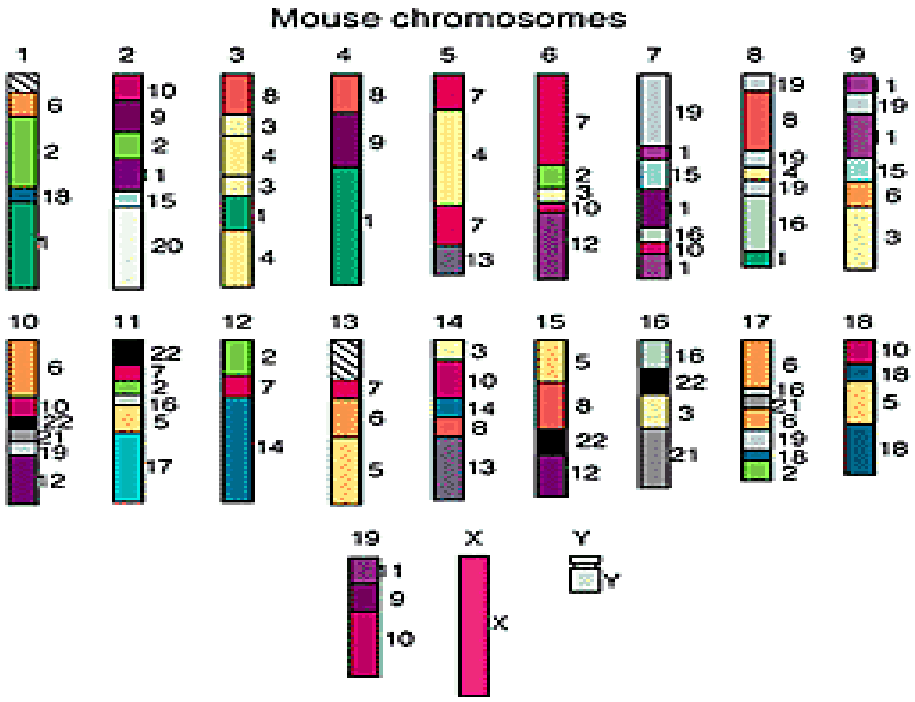
- **Substitution - a base is replaced by another - may or may not alter the protein sequence depending on the place it occurs.**
- **Insertion/Deletion – addition/removal of one or more bases – results in a frame-shift in coding regions.**
- **Rearrangement - a change in the order of complete segments along a chromosome.**



Chromosomal rearrangements occur both within and between chromosomes during evolution



The colors on the mouse chromosomes and the numbers alongside indicate the human chromosomes containing homologous segments.



Mutations

Role of Mutations:

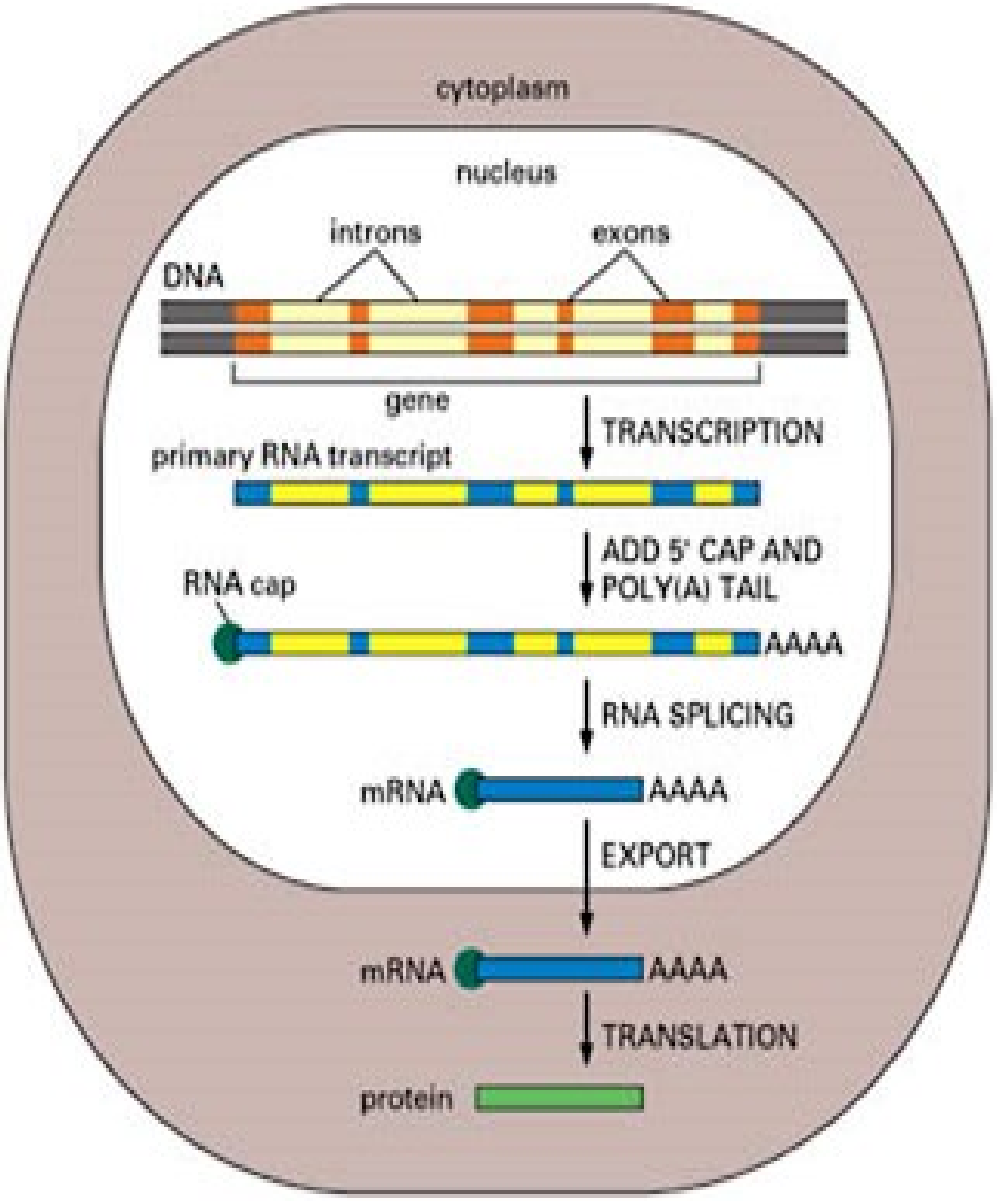
- **Mutations are the source of phenotypic variation on which natural selection acts, creating species & changing them.**

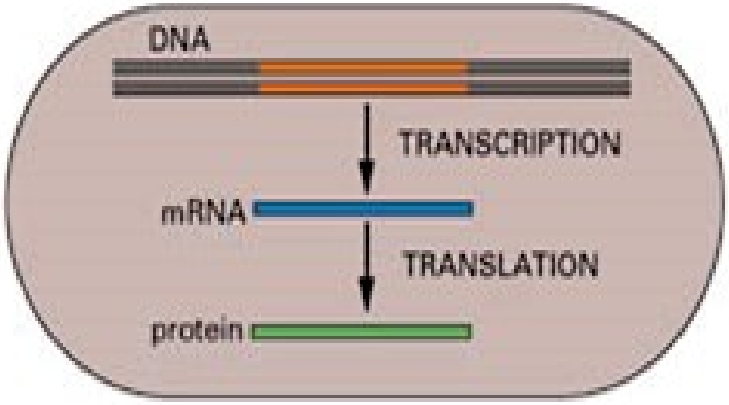
  e.g., the human and mouse genome are very similar – major difference being the internal order of DNA segments.

  Without mutations there wouldn't be any evolution!

- **They are responsible for inherited disorders and diseases, which involve alterations in gene.**

Steps Leading from Gene to Protein

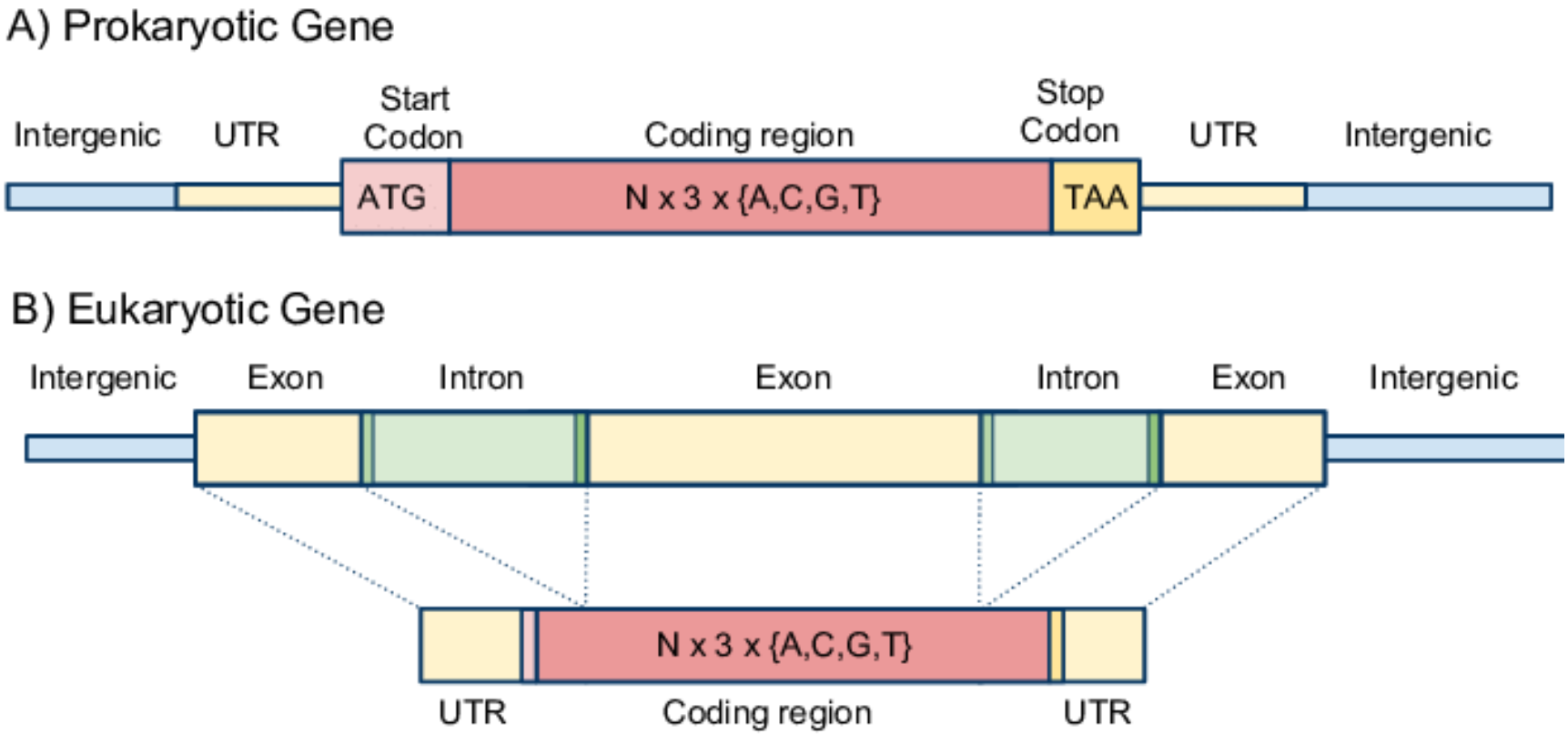




Prokaryotes

Eukaryotes

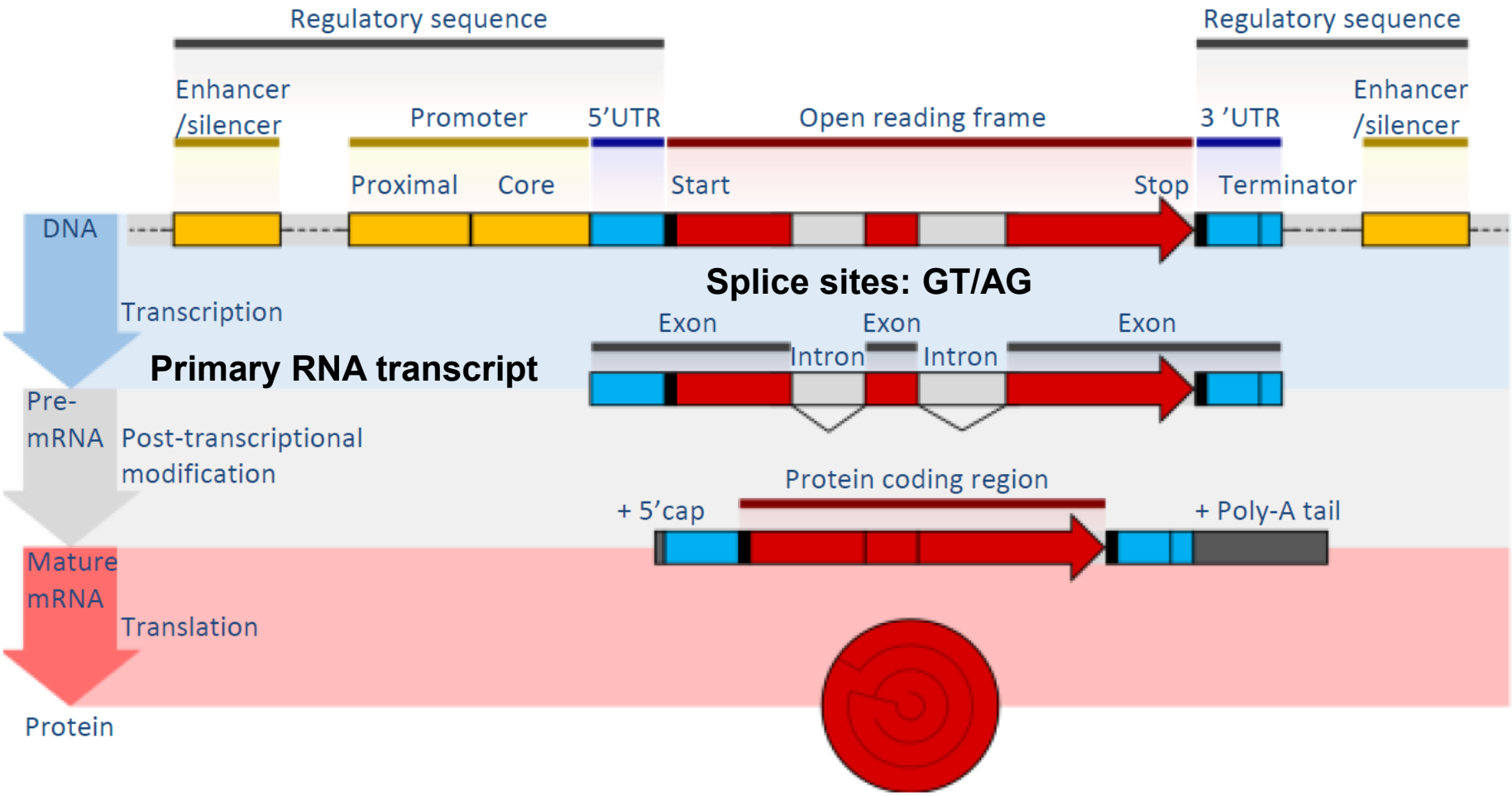
Gene Structure



UTRs – Untranslated Regions – are transcribed, but not translated

Eukaryote Gene Structure

Start codon: ATG **Stop Codon: TAA/TAG/TGA**



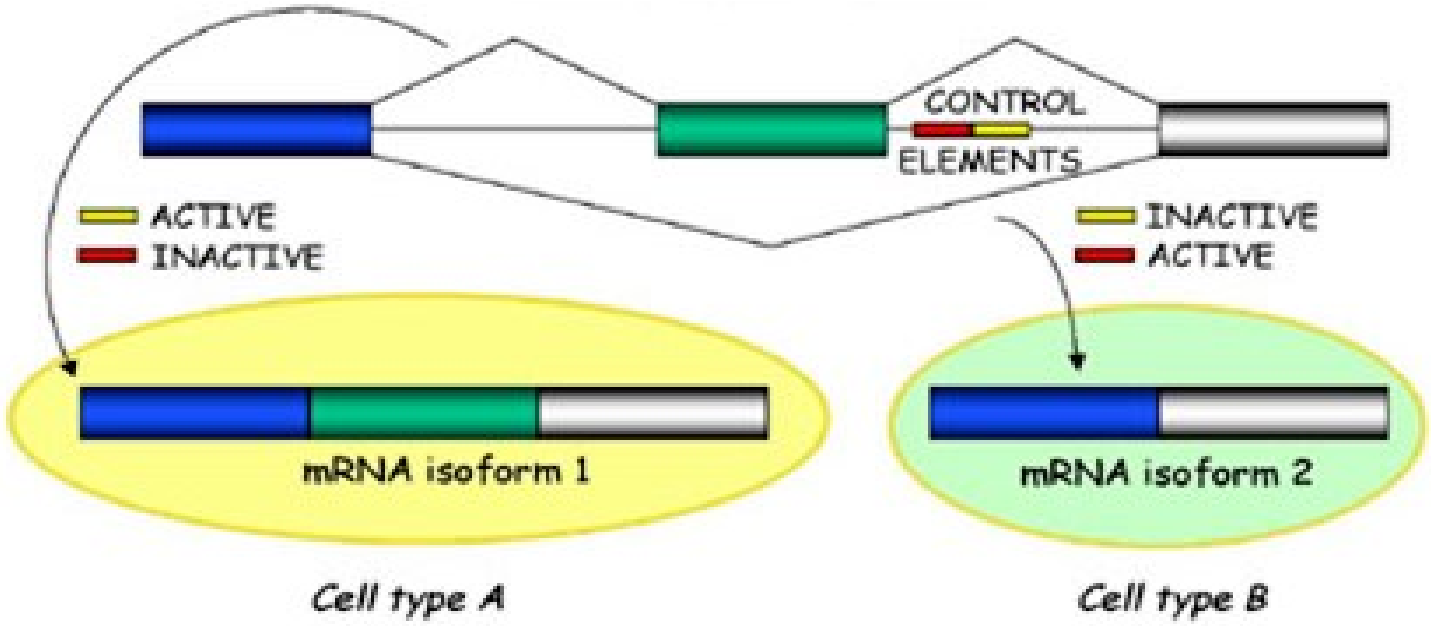
Transcription is initiated only at certain specific positions in the sequence, signaling the beginning of genes, called promoters.

Alternative Splicing

- **In many cases, the pattern of splicing can vary depending on the tissue in which the transcription occurs.**

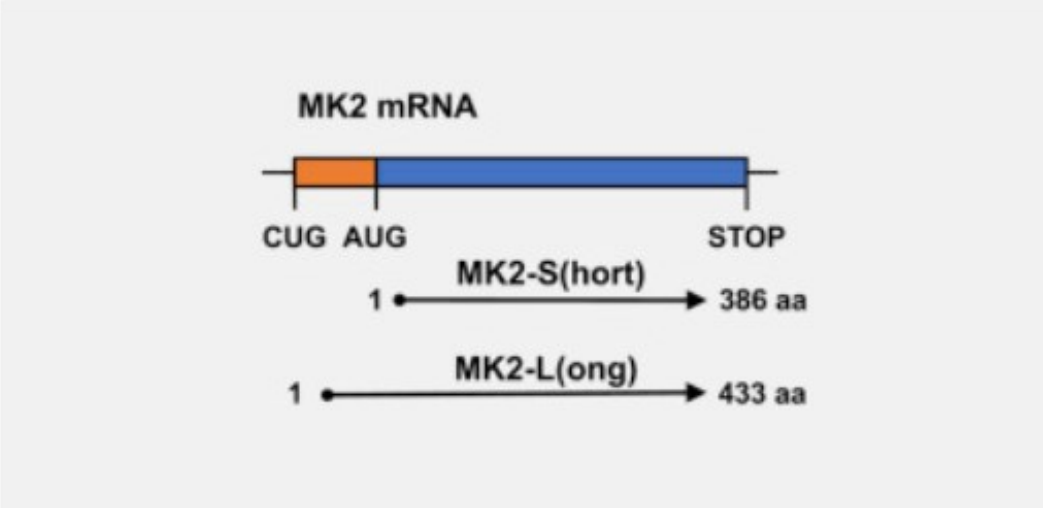
  e.g., an exon maybe spliced in the gene transcribed in liver but retained when transcribed in the brain.

- **This variation called alternative splicing, contributes to the overall protein diversity in the organism**



Alternative Initiation

- Another type of variation that contributes to protein diversity is alternative initiation



- Alternative translation is an important mechanism of post-transcriptional gene regulation leading to the expression of different protein isoforms originating from the same mRNA

Data Representation

DNA - a complex, dynamic, three-dimensional molecule represented as a string of alphabets

- a perfect representation for computer analysis

Aim: to find grammar & syntax rules of DNA language based on this 4-letter alphabet

- similar to English Grammar to form meaningful sentences

Biological Sequence Analysis

Pattern Recognition:

Assumption in biological sequence analysis:

- strings carrying information will be different from random strings

If a hidden pattern can be identified in a string, it must be carrying some functional information

Biological Sequence Analysis

Order of occurrence of bases:

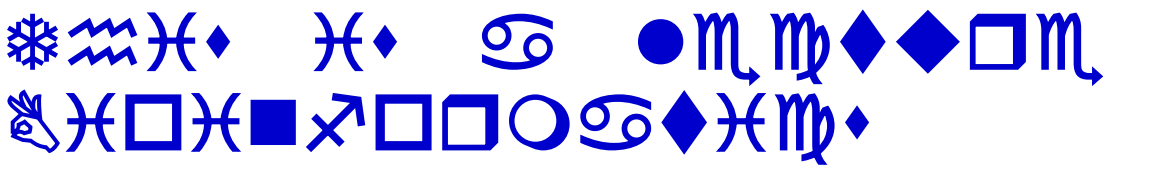
not completely random

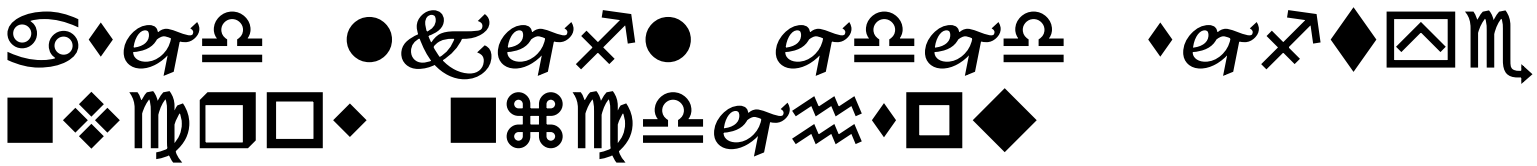
- Different regions of the genome exhibit different patterns of the four bases, A, T, G, C

e.g., protein coding regions, regulatory regions, intron/exon boundaries, repeat regions, etc.

Aim: Identify various patterns to infer their functional roles

Example





This is a lecture on bioinformatics

asjd lkjfl jdjd sjftye nvcrow nzcdjhspu

Frequency of letters

A.	7.3%	N.	7.8%
B.	0.9%	O.	7.4%
C.	3.0%	P.	2.7%
D.	4.4%	Q.	0.3%
E.	13.0%	R.	7.7%
F.	2.8%	S.	6.3%
G.	1.6%	T.	9.3%
H.	3.5%	U.	2.7%
I.	7.4%	V.	1.3%
J.	0.2%	W.	1.6%
K.	0.3%	X.	0.5%
L.	3.5%	Y.	1.9%
M.	2.5%	Z.	0.1%

Other statistics

Frequencies of the most common first letter of a word, last letter of a word, doublets, triplets, etc.

20 most used words in written English

- **the of to in and a for was is that on at he with by be it an as his**

20 most used words in spoken English

- **the and I to of a you that in it is yes was this but on well he have for**

Parallels in DNA language

ATGGTGGTCATGGCGCCCCGAACCCTCTTCCTGCTG
CTCTCGGGGGCCCTGACCCTGACCGAGACCTGGGCG
GGTGAGTGCGGGGTCAGGAGGGAAACAGCCCCTGC
GCGGAGGAGGGAGGGGCCGGCCCGGCGGG

GTCTCAACCCCTCCTCGCCCCCAGGCTCCCACTCCA
TGAGGTATTTCAGCGCCGCCGTGTCCCGGCCCGGCC
GCGGGGAGCCCCGCTTCATCGCCATGGGCTACGTGG
ACGACACGCAGTTCGTGCGGTTC

Parallels in DNA language

ATG GTG GTC ATG GCG CCC CGA ACC CTC TTC
CTG CTG CTC TCG GGG GCC CTG ACC CTG ACC
GAG ACC TGG GCG GGT GAG TGC GGG GTC AGG
AGG GAA ACA GCC CCT GCG CGG AGG AGG GAG
GGG CCG GCC CGG CGG…

GTC TCA ACC CCT CCT CGC CCC CAG GCT CCC ACT
CCA TGA GGT ATT TCA GCG CCG CCG TGT CCC
GGC CCG GCC GCG GGG AGC CCC GCT TCA TCG
CCA TGG GCT ACG TGG ACG ACA CGC AGT TCG
TGC GGT TC…

1st exon and 1st intron of Human HLA gene

This task needs to be automated because of the large genome sizes:

Smallest genome:

Mycoplasma genitalium 0.5×10^6 bp

Human genome: 3×10^9 bp – not the largest!

~ 10-100 times the Britannica Encyclopedia

Plant genomes are even larger.

DNA Sequence Analysis

- **Evolution has operated on every sequence that we see today - genes and sequences involved in gene regulation are conserved.**
- **these are transferred, like code modules, from one organism to another. Because of evolution, similar sequences have similar functions.**
- **Algorithms for comparing sequences and finding similar regions are at the heart of computational biology.**