Mansoura University
Faculty of Computers and Information
Department of Computer Science
First Semester: 2020-2021

[MED121] Bioinformatics: An Introduction

Grade: Third Year (Medical Informatics Program)

Sara El-Metwally, Ph.D.

Faculty of Computers and Information,

Mansoura University,

Egypt.

AGENDA

- What is Bioinformatics.
- Bioinformatics Vs. Medical Informatics.
- Central Dogma of Life.
- DNA, Genomes.
- RNA, Transcriptomes.
- Proteins, Proteomics .
- Transcriptions and Translation.
- Omics era.

COURSE OUTLINES

- **Course Meeting Time**: Wednesday, 2 :15 pm – 4.00 pm
- **Course Instructor**: Sara El-Metwally, PhD
- **Course TAs**: Eng. Aya Ayad, Eng. Mohamed Ashraf,
- **Course Labs**: Python
- **Course Grading**:
  - Midterm: 10%
  - Oral: 10%
  - Practical: 20%
  - Final: 60%
  - Any Projects are welcomed ! (up to 40%)
  - Quizzes, attendance , etc..

COURSE OBJECTIVES

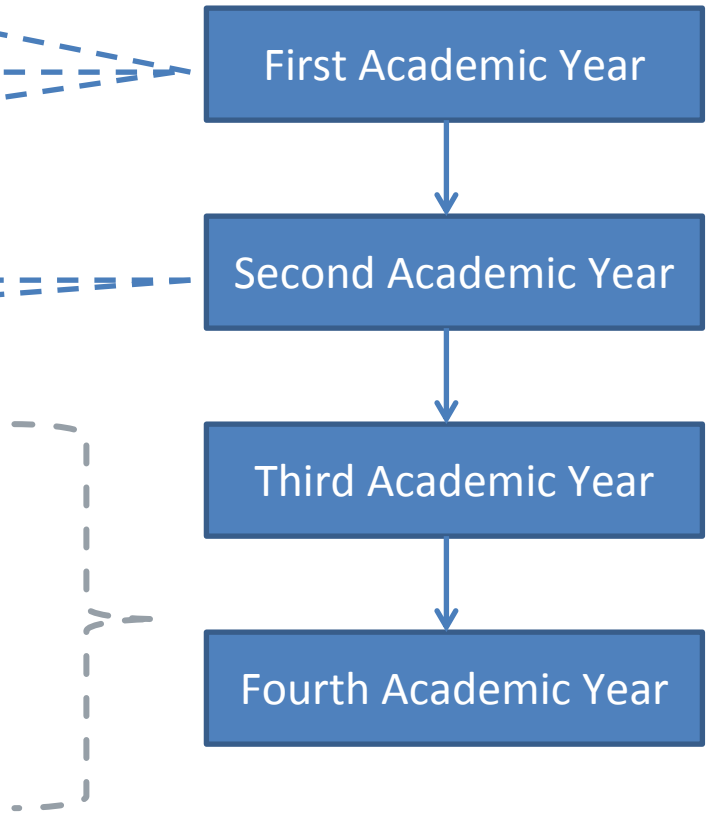
- Bioinformatics course works as an interface between the computer science and biology and opens new opportunity to deal with different algorithms and tools for sequence comparisons, assembly, pattern matching, and efficient data structure to process a biological text.
- Understand the basic computer science terminologies regarding algorithm analysis and design, data analysis, and computer programming.
- Define the challenges that are facing the large scale biological data and think how to solve them through developing algorithms and bioinformatics tools.

MANSOURA FCIS COURSE DEPENDENCY

Basics of Programming Languages
Basics of Biology
Statistics & probabilities

Data Structures
Genetics

Bioinformatics, Genomics, Medical Informatics
Artificial Intelligence,
Machine Learning,
Algorithms analysis and design,
Big Data and Data mining techniques,
Distributed Systems,
Cloud Computing, etc..



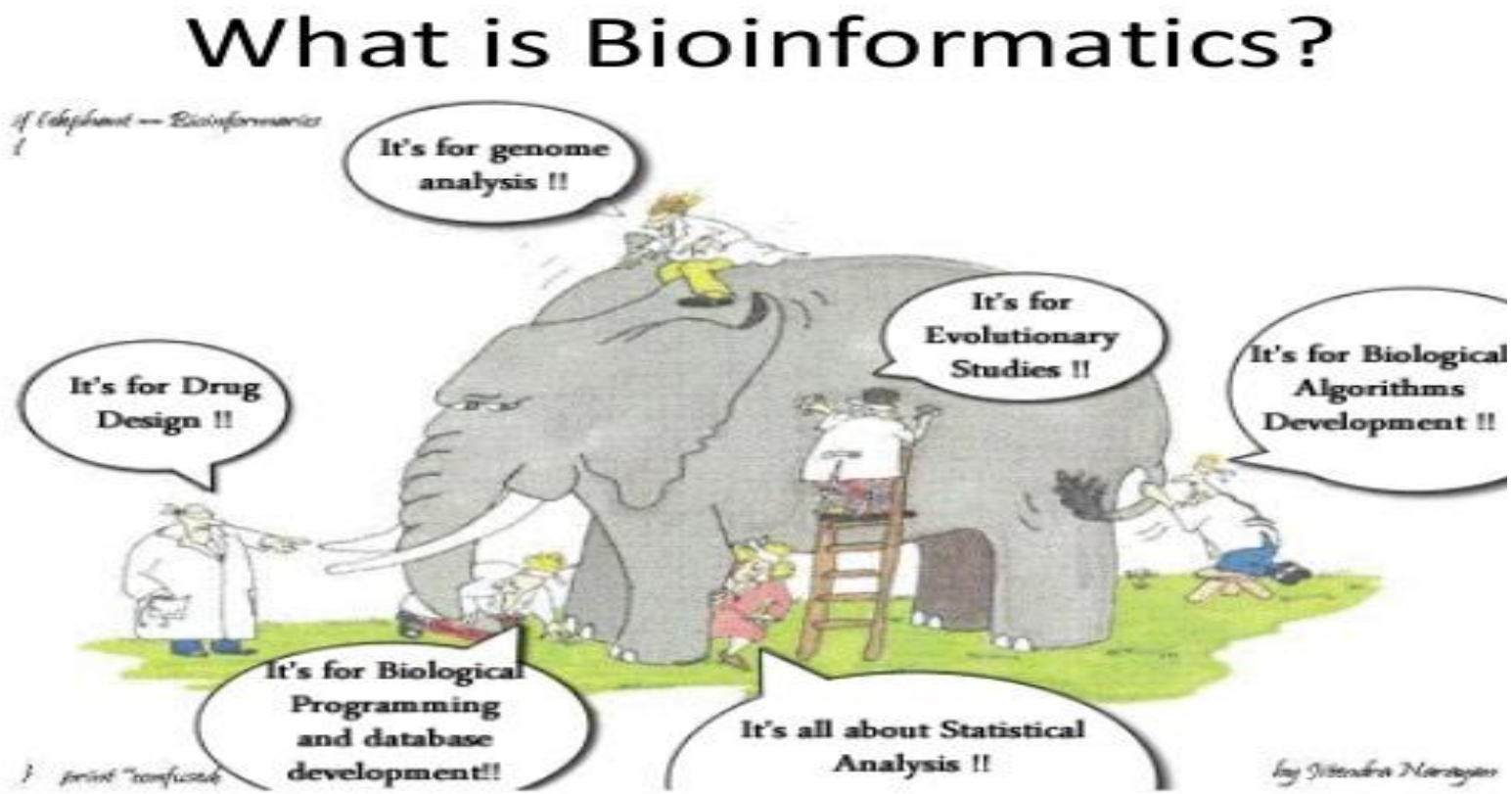
COURSE OUTPUT

http://phdcomics.com/

LET'S START!

- What do you think " Bioinformatics" is? and why?

LET'S START!



- a professional in the pharmaceutical industry
- a policeman worrying about DNA testing
- a computer scientist developing bio-databases
- a consumer concerned about GMOs (Genetically Modified Organisms)
- … …

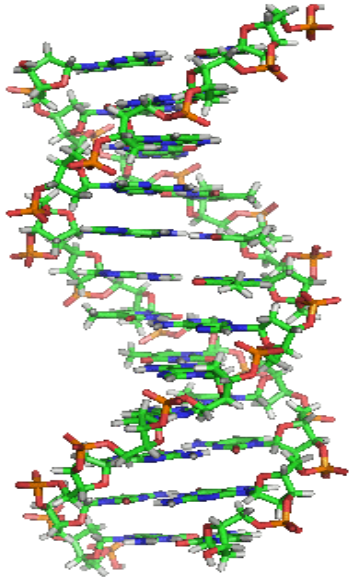
Photo credit: https://bioinformaticsonline.com/file/view/4482/bioinformatics-definitions-and-applications

WHAT IS BIOINFORMATICS?

- **Bioinformatics** is conceptualizing biology in terms of molecules and applying "informatics techniques" (derived from disciplines such as applied maths, computer science and statistics) to understand and organize the information associated with these molecules, on a large scale.

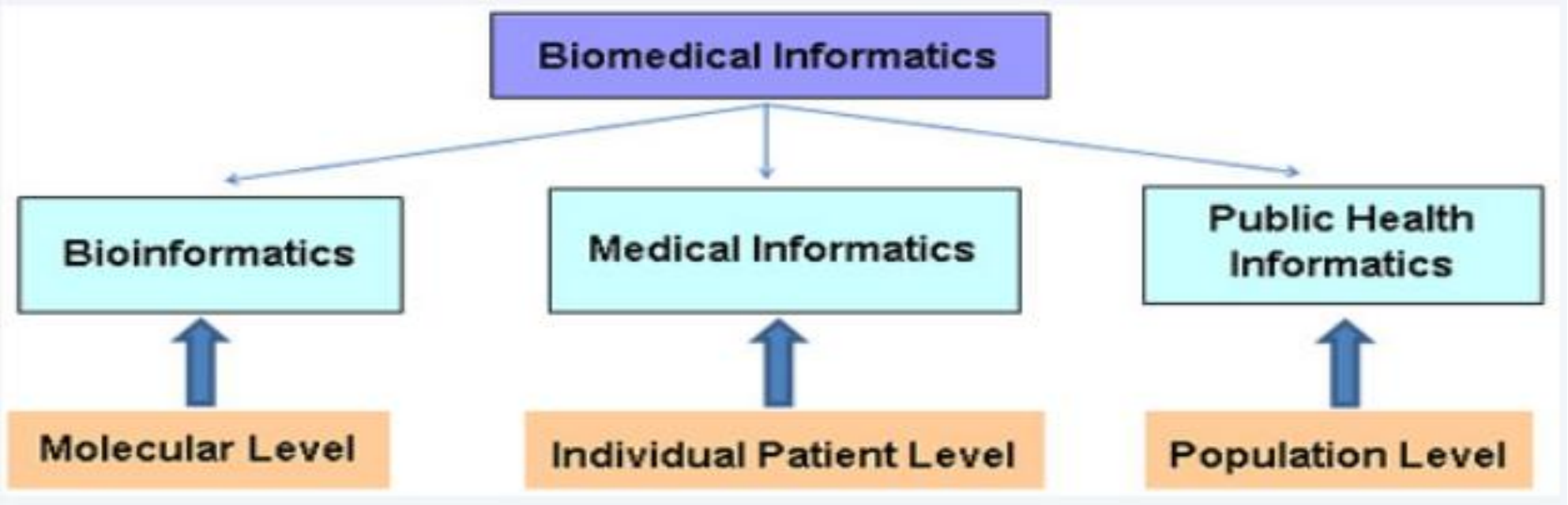
- **Bioinformatics** is computer aided biology! Also called computational biology.

BIOINFORMATICS: A SIMPLE VIEW



Biological Data + **Computer Calculations**

BIOINFORMATICS, MEDICAL INFORMATICS &BIOMEDICAL INFORMATICS



https://osteopathic.nova.edu/msbi/evolution.html

BIOINFORMATICS IS FUN!

CENTRAL DOGMA OF LIFE

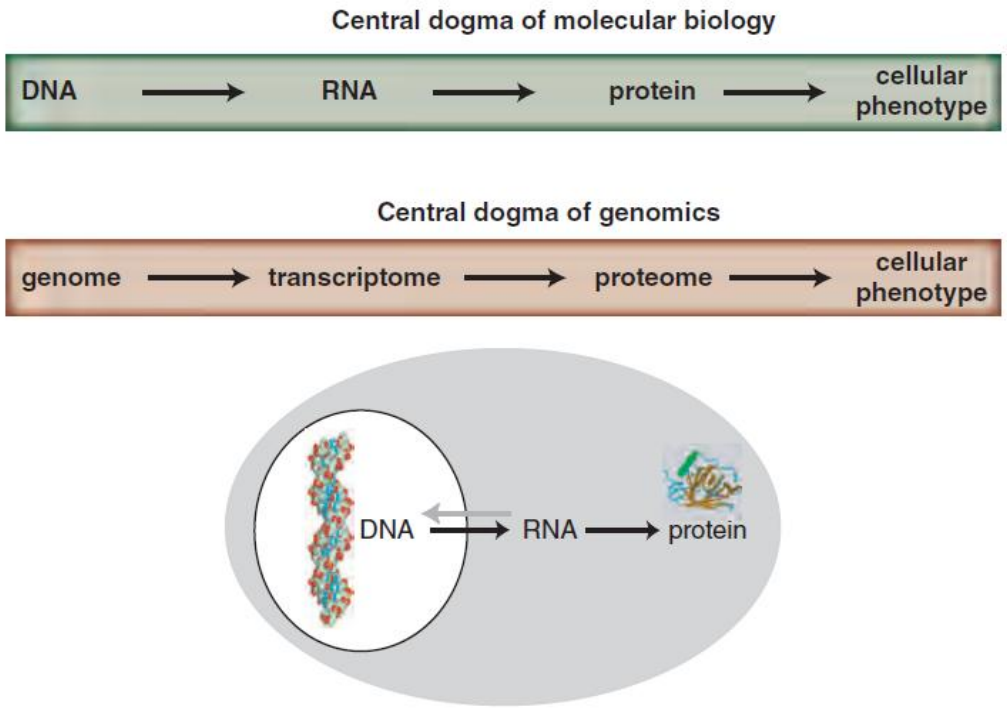


Photo credit :Functional Genomics Book by Jonathan Pevsner

Photo credit : https://edu.t-bio.info/course/transcriptomics-1/

DNA

- DNA stands for Deoxyribo Nucleic Acid that carries the genetic instructions required for the development, functioning and reproduction of all known living organisms.
- In eukaryotic organisms (like animals, plants, and fungi), DNA is present in the nucleus of each cell.
- In prokaryotic organisms (single-celled organisms like bacteria and mitochondria), DNA is present in the cell's cytoplasm.

DNA

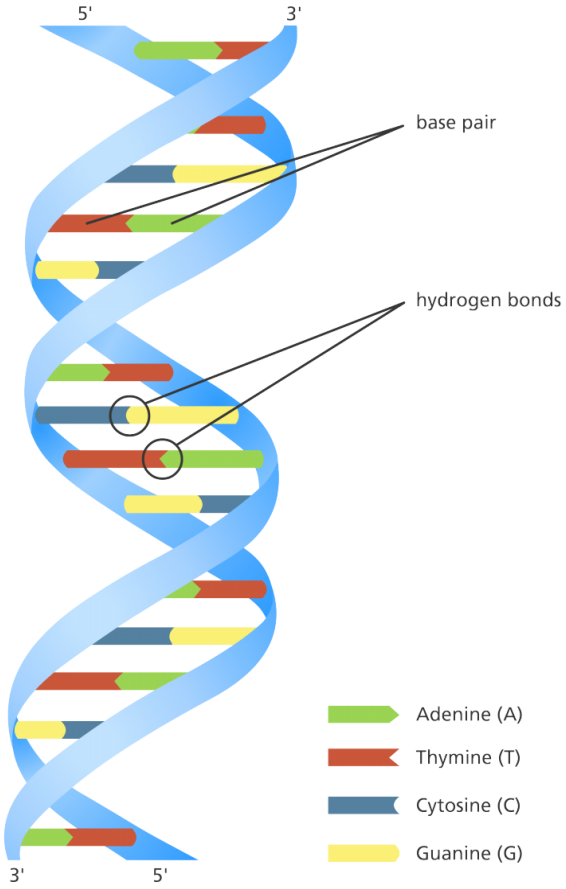


Photo Credit: https://www.yourgenome.org/facts/what-is-dna

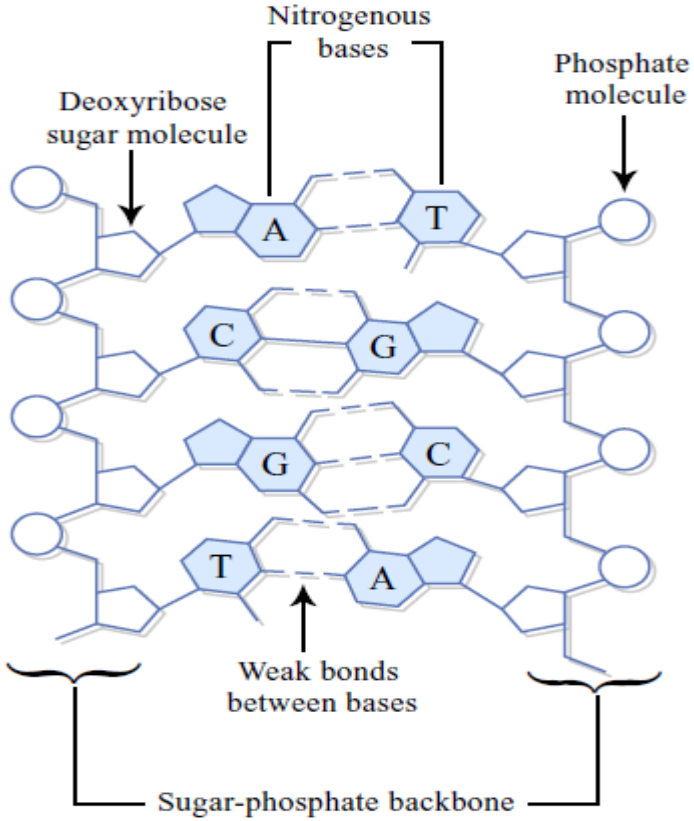


Photo Credit: https://www.yourgenome.org/facts/what-is-dna

DNA

TACTGTGACACTGT

```
Y     Pyrimidine (C or T)
R     Purine (A or G)
W     Weak (A or T)
S     Strong (G or C)
K     Keto (T or G)
M     Amino (C or A)
D     A, G, T (not C - remember as after C)
V     A, C, G (not T - remember as after T/U - We'll get to "U" soon)
H     A, C, T (not G - remember as after G)
B     C, G, T (not A - remember as after A)
N     Any base
-     Gap
```

Photo credit : BIOSTAR Handbook

DNA

sense/forward/+/watson/code strand

5' ATGACACTGTGACA 3'

3' TACTGTGACACTGT 5'

anti-sense/reverse/-/crick/template strand

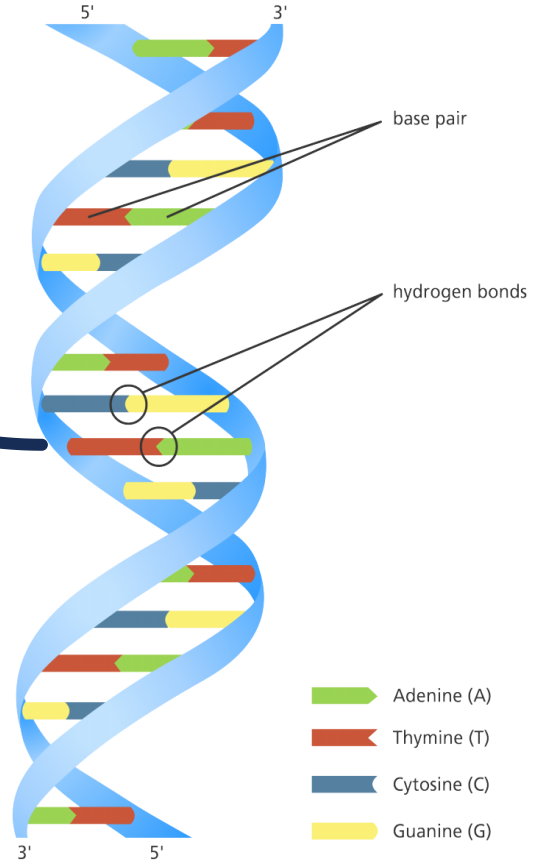


Photo Credit: https://www.yourgenome.org/facts/what-is-dna

GENOMES

- Genome refers to the complete set of genetic information in an organism, it is your DNA!
- Human genome contains approximately 3 billion base pairs, which reside in the 23 chromosomes.
- Mosquitoes have 3 pairs, i.e. tiger mosquito has 1,967, 000, 000 bp = 1,967 Mbp.

GENOME SIZES

Comparison of Genome Size in Different Organisms

Species	*T2 phage*	*Escherichia coli*	*Drosophila melanogaster*	*Homo sapiens*	*Paris japonica*
Genome Size	170,000 bp	4.6 million bp	130 million bp	3.2 billion bp	150 billion bp
Common Name	 Virus	 Bacteria	 Fruit fly	 Human	 Canopy Plant

Photo Credit: https://ib.bioninja.com.au/standard-level/topic-3-genetics/32-chromosomes/genome-size.html

Organism	Completion date	Size	Description
phage phiX174	1978	5,368 bp	1st viral genome
human mtDNA	1980	16,571 bp	1st organelle genome
lambda phage	1982	48,502 bp	important virus model
HIV	1985	9,193 bp	AIDS retrovirus
H. influenzae	1995	1,830 Kb	1st bacterial genome
M. genitalium	1995	580 Kb	smallest bacterial genome
S. cerevisiae	1996	12.5 Mb	1st eukaryotic genome
E. coli K12	1997	4.6 Mb	bacterial model organism
C. trachomatis	1998	1,042 Kb	internal parasite of eukaryotes
D. melanogaster	2000	180 Mb	fruit fly, model insect
A. thaliana	2000	125 Mb	thale cress, model plant
H. sapiens	2001	3,000 Mb	human
SARS	2003	29,751 bp	coronavirus

Photo Credit: http://book.bionumbers.org/how-many-genes-are-in-a-genome/

GENES

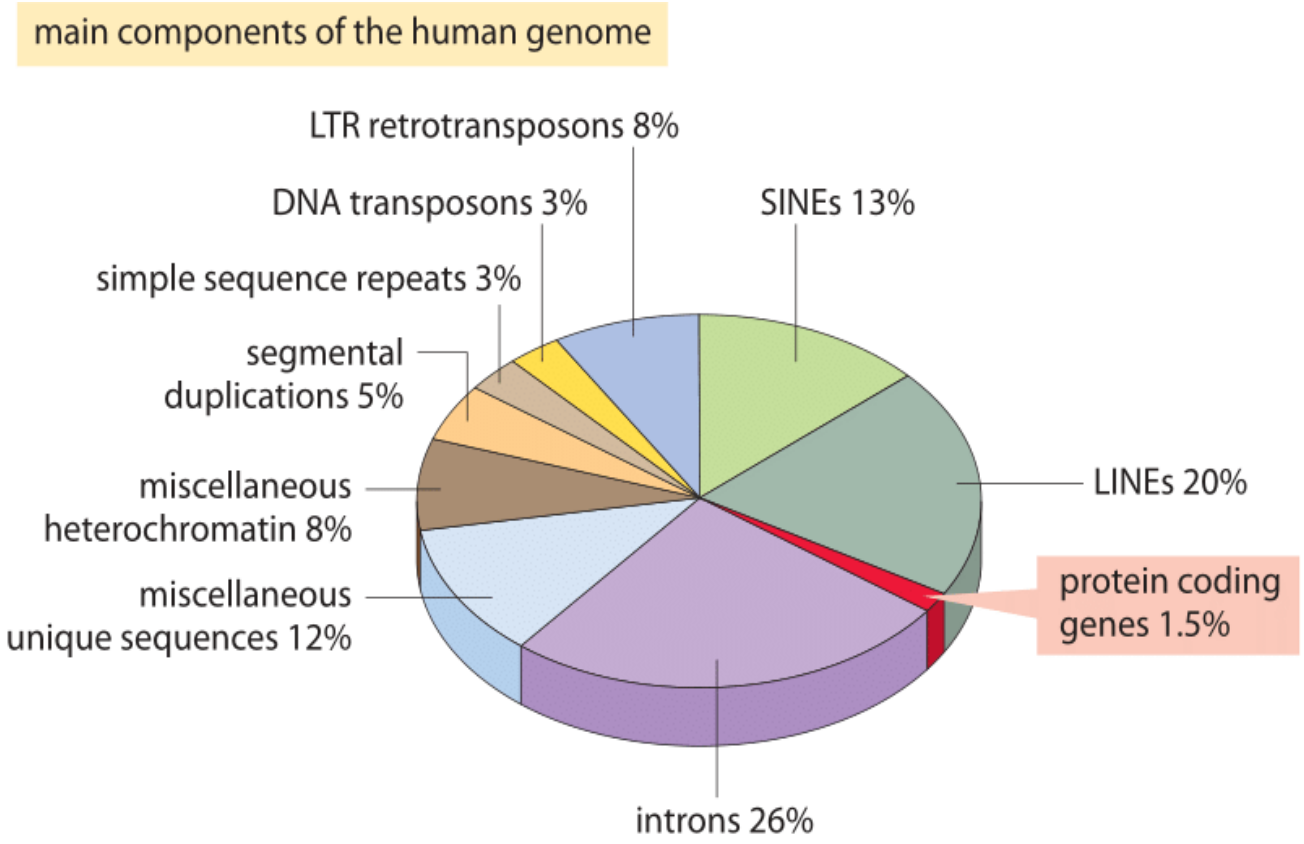


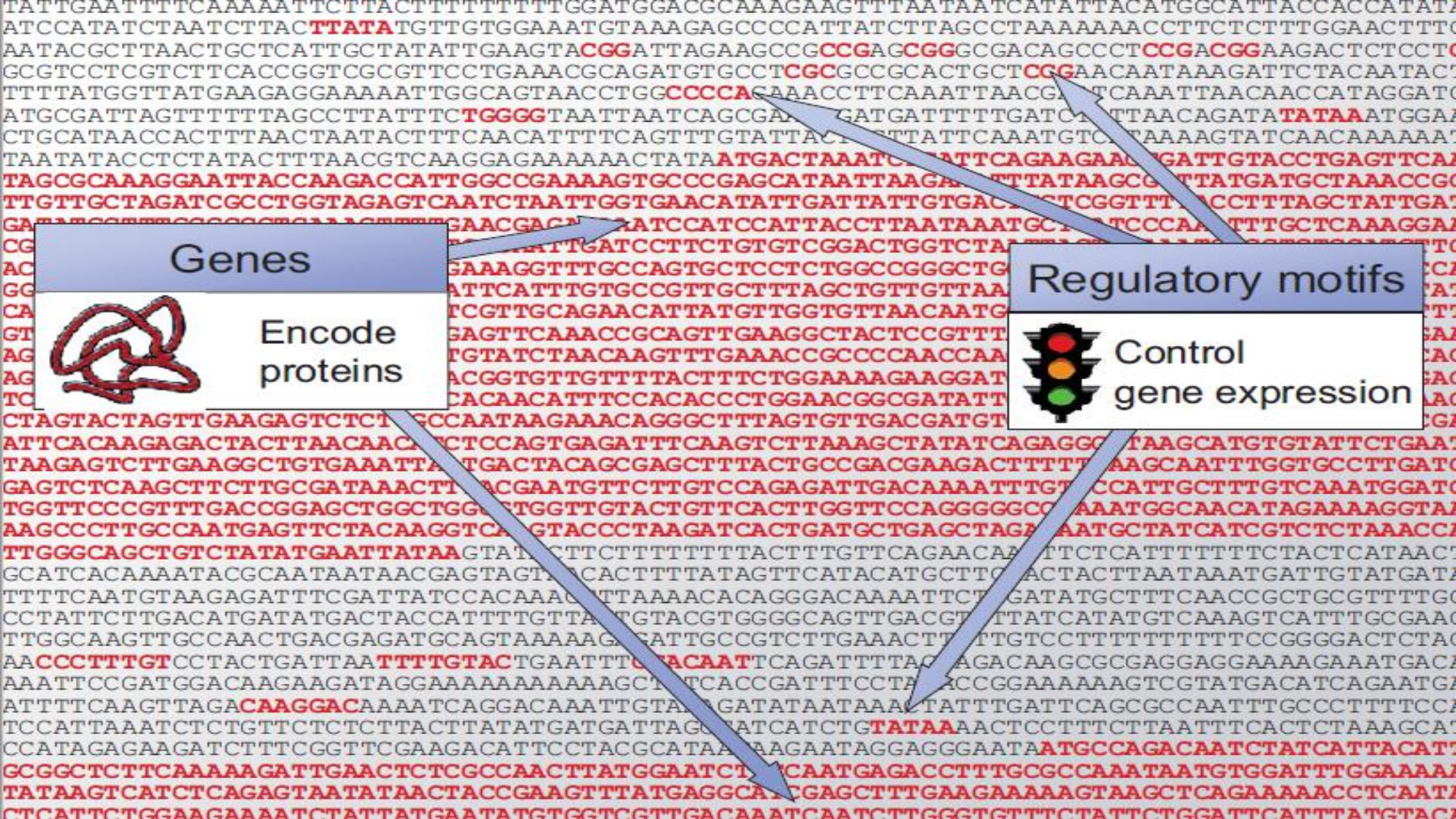
Photo Credit: https://ib.bioninja.com.au/standard-level/topic-3-genetics/32-chromosomes/genome-size.html

	Organism	# of protein-coding genes	# of genes naïve estimate: (genome size /1000)	BNID
viruses	HIV 1	9	10	105769
	Influenza A virus	10-11	14	105767
	Bacteriophage λ	66	49	105770
	Epstein Barr virus	80	170	103246
prokaryotes	*Buchnera sp.*	610	640	105757
	T. maritima	1,900	1,900	105766
	S. aureus	2,700	2,900	105500
	V. cholerae	3,900	4,000	105760
	B. subtilis	4,400	4,200	111448
	E. coli	4,300	4,600	105443
eukaryotes	*S. cerevisiae*	6,600	12,000	105444
	C. elegans	20,000	100,000	101364
	A. thaliana	27,000	140,000	111380
	D. melanogaster	14,000	140,000	111379
	F. rubripes	19,000	400,000	111375
	Z. mays	33,000	2,300,000	110565
	M. musculus	20,000	2,800,000	100308
	H. sapiens	21,000	3,200,000	100399, 111378
	T. aestivum (hexaploid)	95,000	16,800,000	105448, 102713

GENES

- Gene is a basic unit of heredity, it is a sequence of bases that carries the instructions to make a particular protein.
- There are an estimated 19,000 - 20,000 human protein-coding genes.
- Not all of the DNA in a genome encodes protein, i.e. ~ 1.5 % of human genome are protein coding regions.

TATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
ATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
AATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACT
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATC
ATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGAA
CGATTTGCCGTTGGACGGTTCTTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTTC
ACTCTTTTCTAAAGAAACTTGCACCGGAAAGGTTTGCCAGTGCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACCA
GGCAGTGGATTGTCTTCTTCGGCCGCATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATAT
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTTGGTGTTAACAATGGCGGTATGGATCAGGCTGCCTCTGTTT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTAAAAAACCATGAA
AGCTTTGTTATTGCGAACACCCTTGTTGTATCTAACAAGTTTGAAACCGCCCCAACCAACTATAATTTAAGAGTGGTAGAAGTCAC
AGCTGCAAATGTTTTAGCTGCCACGTACGGTGTTGTTTTACTTTCTGGAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGAC
TCATGAACGTTTATTATGCCAGATATCACAACATTTCCACACCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAAAC
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTTGACGATGTCGCACAATCCTTGAATTGTTCTCGCGA
ATTCACAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATC
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATC
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTATAAGTATACTTCTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAACT
GCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATA
TTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGC
CCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAC
TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTGTCCTTTTTTTTTTCCGGGGACTCTAC
AACCCTTTGTCCTACTGATTAATTTTGTACTGAATTTGGACAATTCAGATTTTAGTAGACAAGCGCGAGGAGGAAAAGAAATGACA


Genes
Encode proteins
Regulatory motifs
Control gene expression

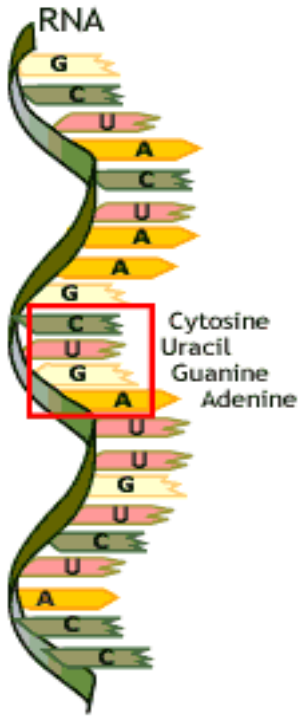
GENOTYPE VS. PHENOTYPE

- **Genotype** is the set of genes in an organism.
- **Phenotype** is the set of all observable characteristics — which are influenced both by its genotype and by the environment.




Sources: http://evolution.berkeley.edu/evolibrary/article/genovspheno_01

RNA



- RNA is like DNA except that:
  1. Backbone is a little different.
  2. Often single stranded.
  3. The base Uracil (U) is used in place of Thymine (T).

- RNA as a string of four alphabets:

AUGACACUGUGACA

Photo Credit: https://www.micropia.nl/en/discover/microbiology/rna/

DNA vs. RNA

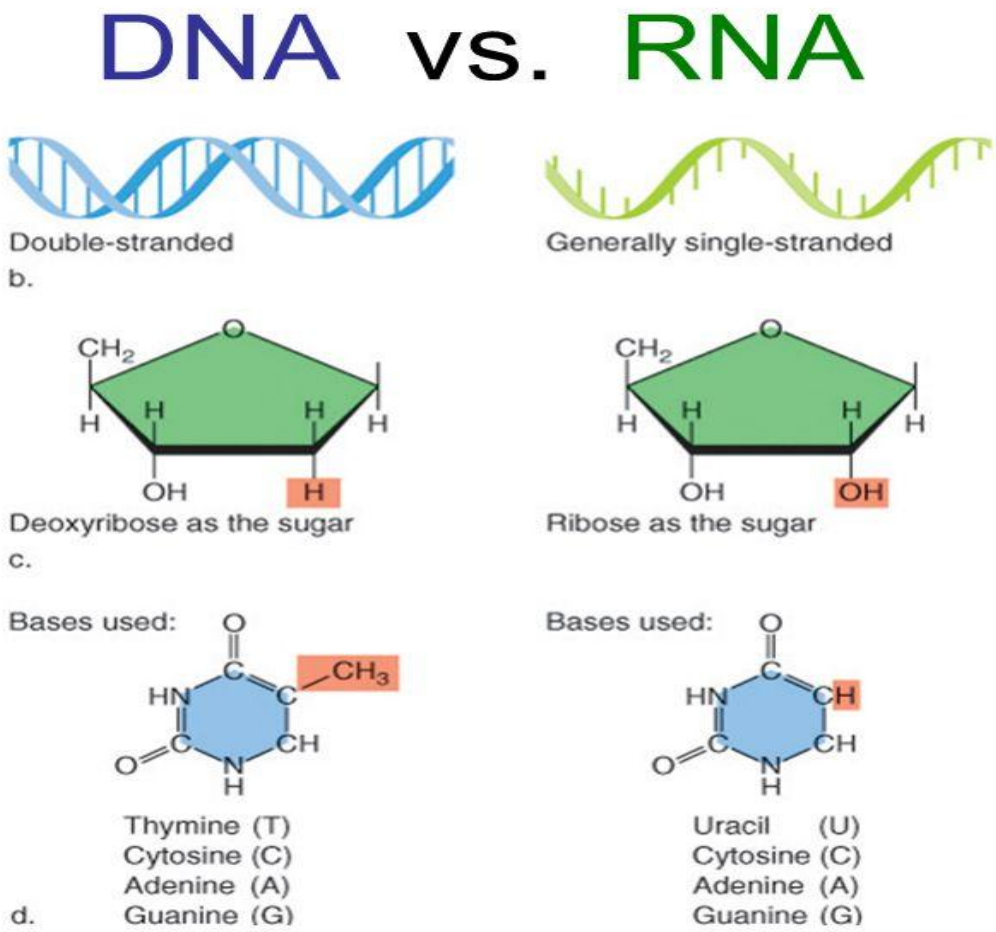
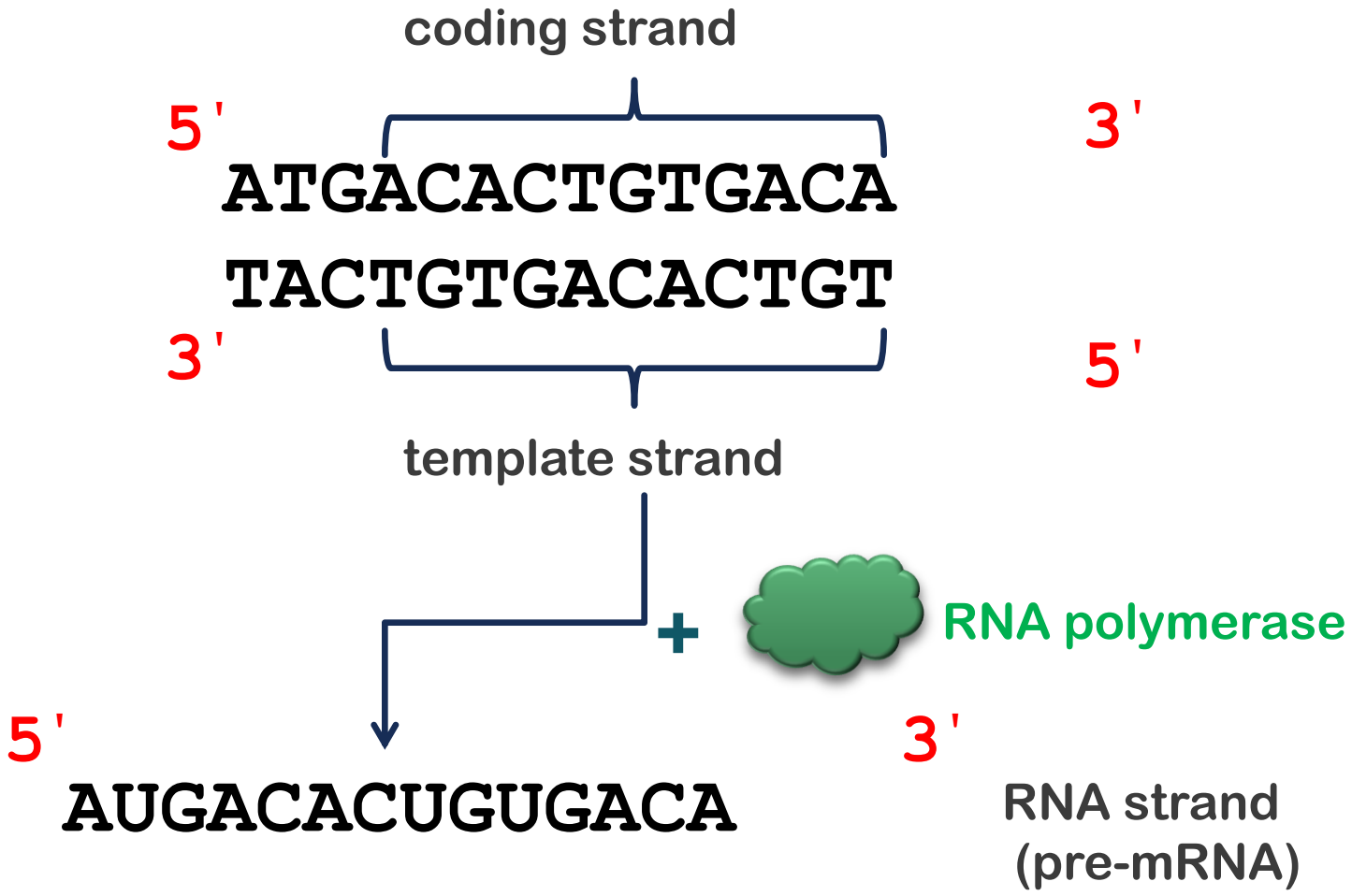


Photo credit: http://genes.com.pk/home/detail/why-dna-is-double-stranded-and-rna-is-single-stranded-

TRANSCRIPTION

- Transcription is the process of transforming information in a gene in a DNA strand to an RNA strand, i.e. making RNA from a DNA template.
- Genetic information is carried on only one of the two strands of the DNA, which known as a coding strand.

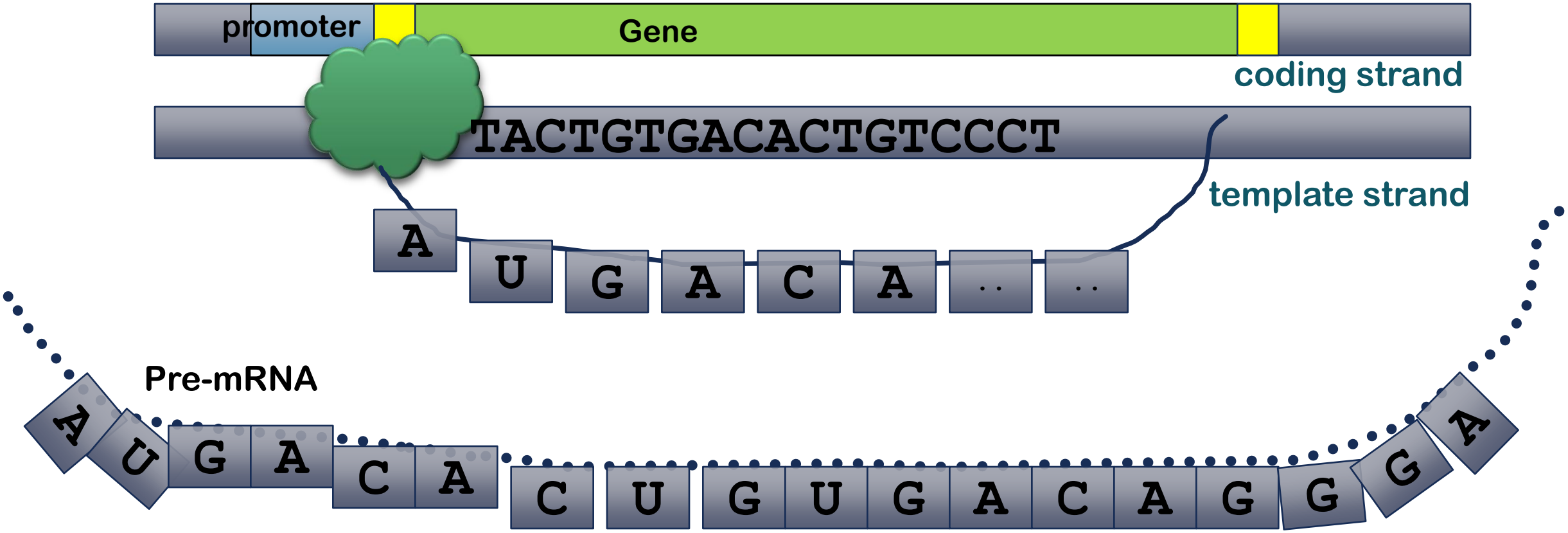
TRANSCRIPTION



TRANSCRIPTION

- RNA polymerase recognizes short sequence of bases called promoter sequence and bind to it.

TRANSCRIPTION



RNA PROCESSING

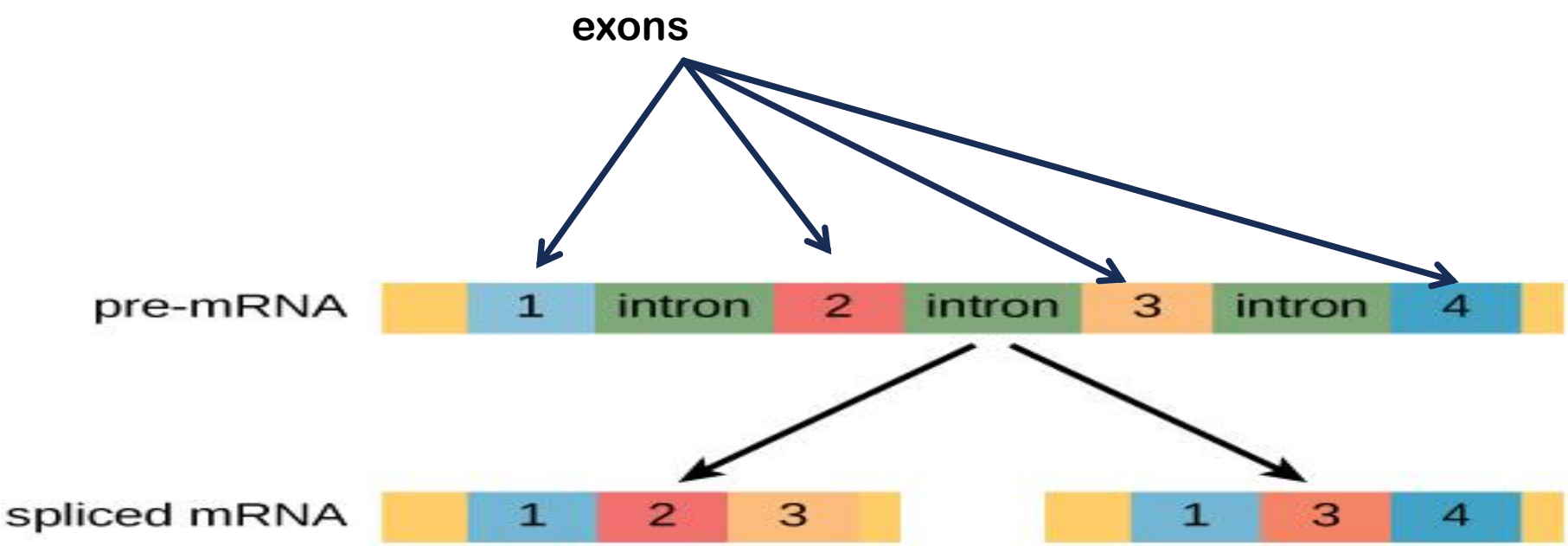


Photo credit: https://en.wikipedia.org/wiki/Alternative_splicing

ALTERNATIVE SPLICING

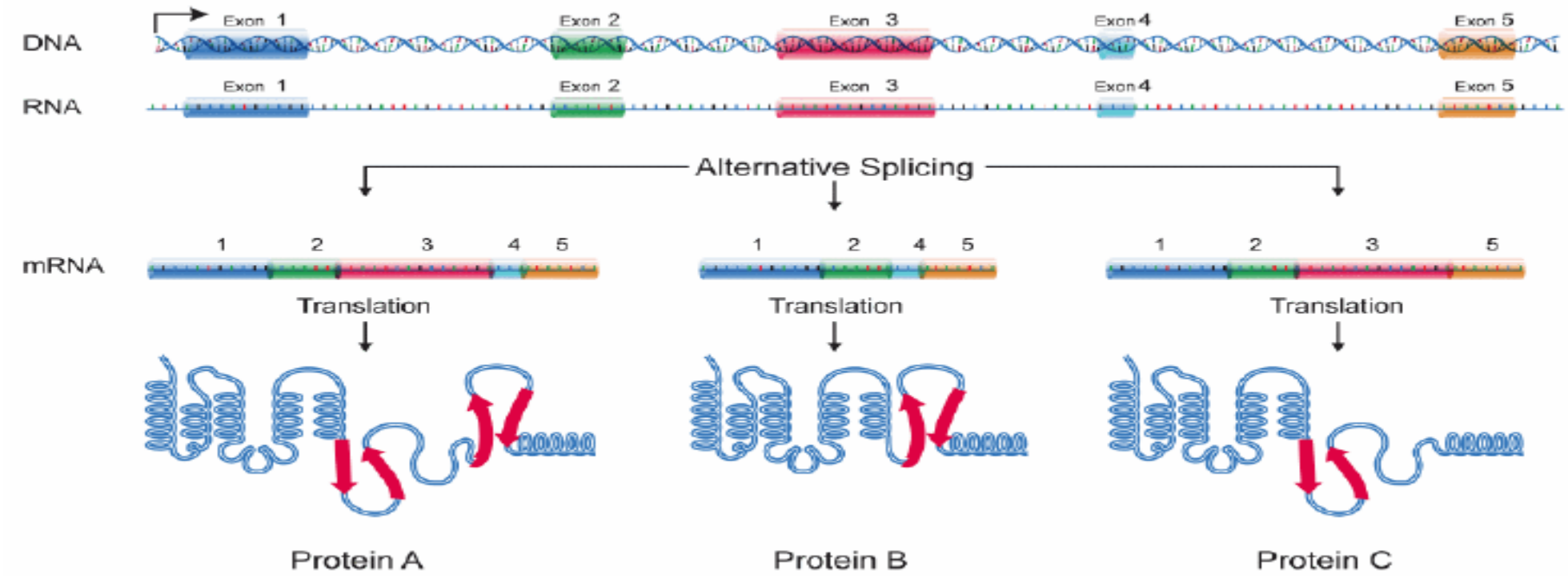


Photo credit: https://en.wikipedia.org/wiki/Alternative_splicing

NOTES

- **Other DNA products:**
  1. ribosomal RNA (**rRNA**), which includes major constituents of ribosomes.
  2. transfer RNAs (**tRNAs**), which carry amino acids to ribosomes.
  3. micro RNAs (**miRNAs**), which play an important regulatory role in various plants and animals.

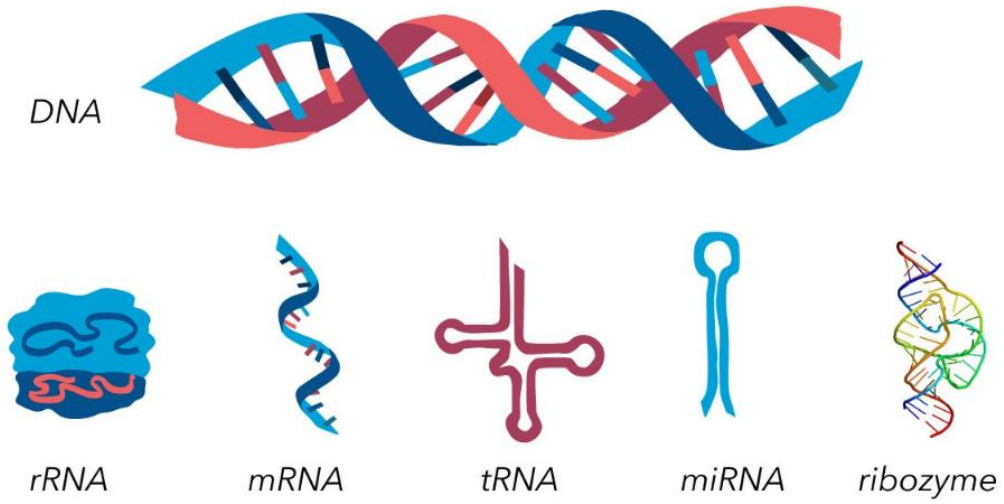


Photo credit : https://edu.t-bio.info/course/transcriptomics-1/

TRANSLATION

- Ribosome is the machine that synthesizes proteins from mRNA.
- Proteins are molecules composed of one or more polypeptides.
- a polypeptide is a polymer composed of amino acids.
- Cells build their proteins from 20 different amino acids.
- A polypeptide can be thought of as a string composed from a 20-character alphabet.

TRANSLATION

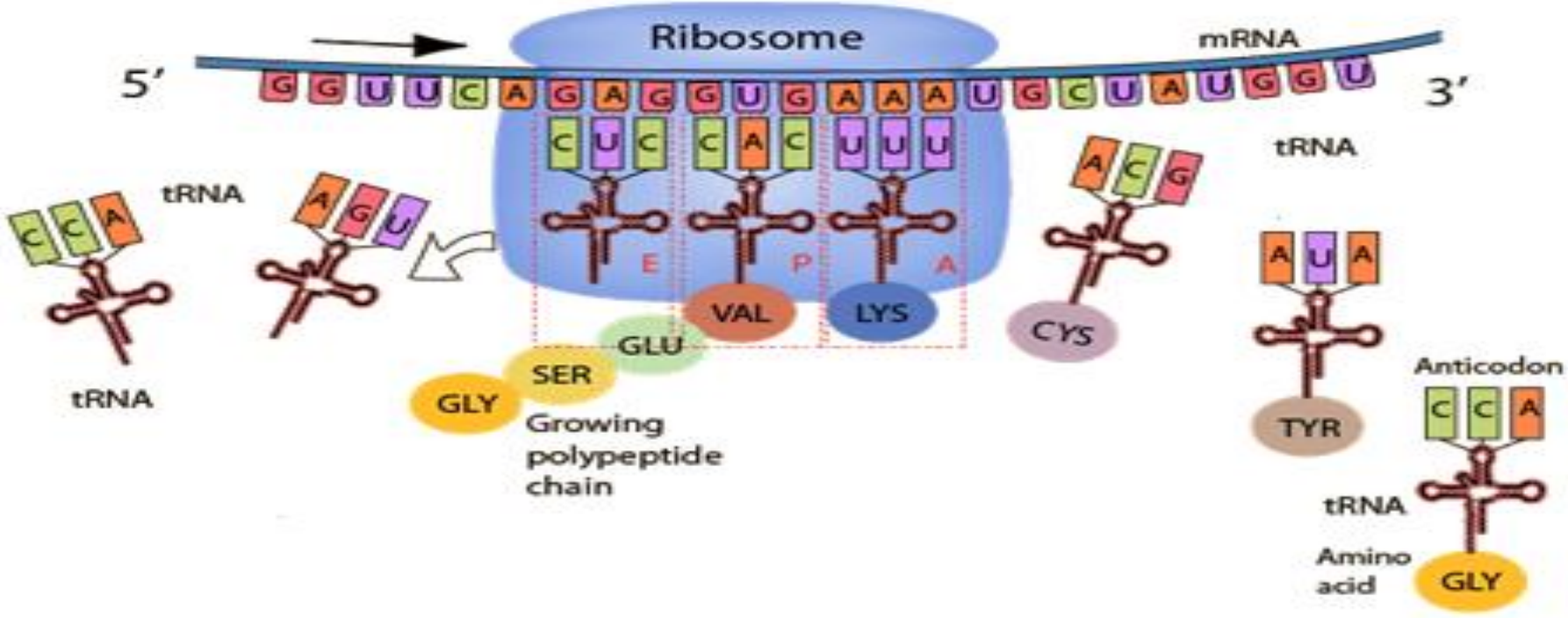


Photo credit: http://hyperphysics.phy-astr.gsu.edu/hbase/Organic/translation.html

AMINO ACIDS

Second base

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

Photo credit: http://lyl4science.blogspot.com/2013/02/amino-acid-chart.html

PROTEIN SEQUENCE

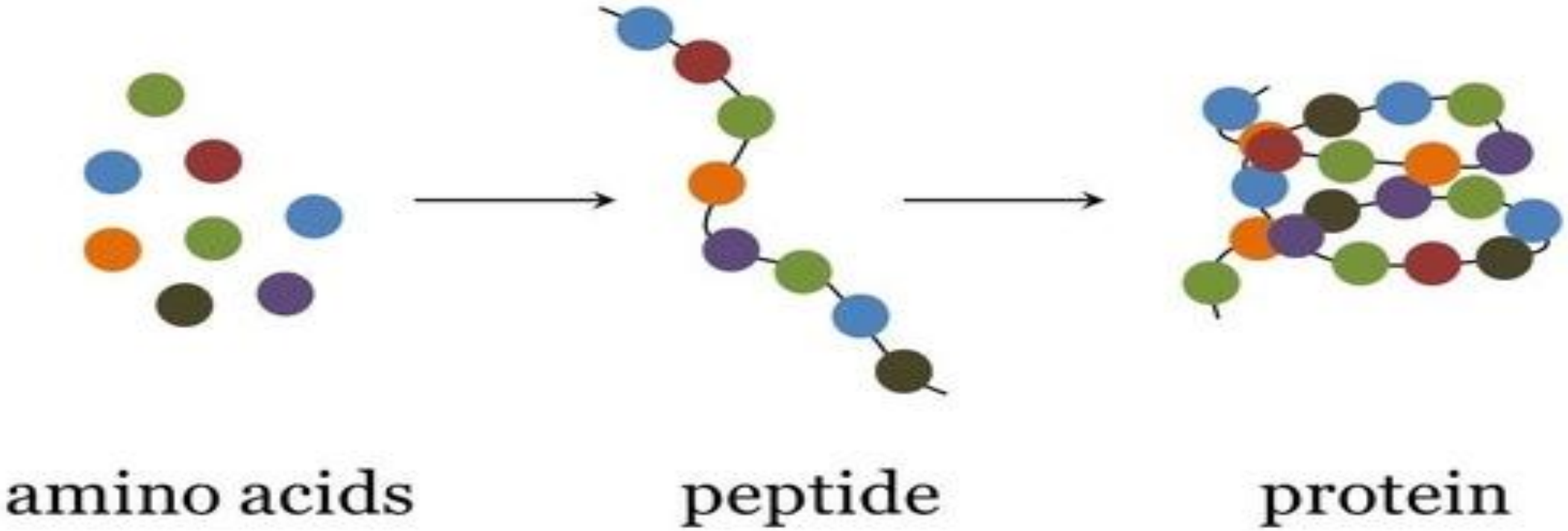
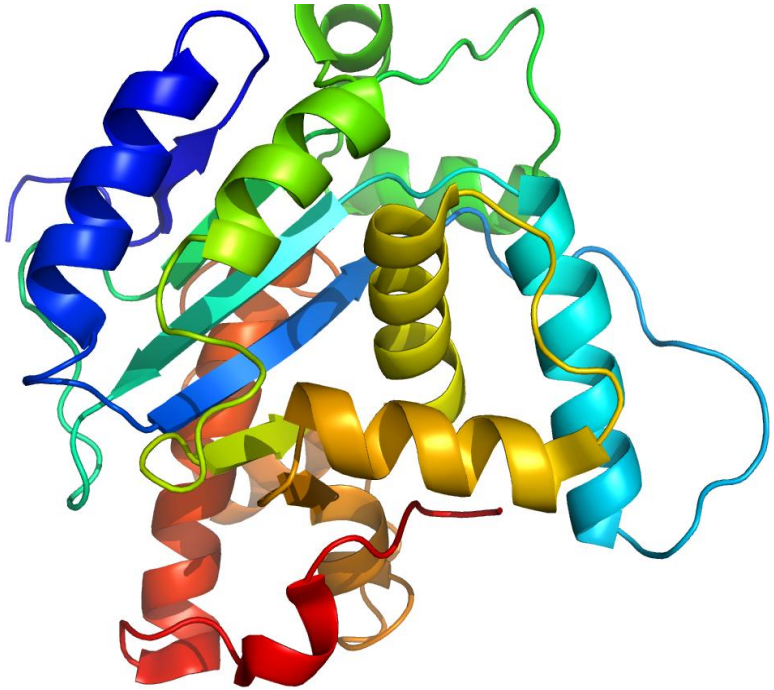


Photo Credit: https://www.peptidesciences.com/information/peptides-vs-proteins/

PROTEIN SEQUENCE



→ **FLIVQNATPSVIE**

Photo credit: https://www.zmescience.com/medicine/designer-protein-upon-us-25072019/

VIDEO

- https://www.youtube.com/watch?v=gG7uCskUOrA

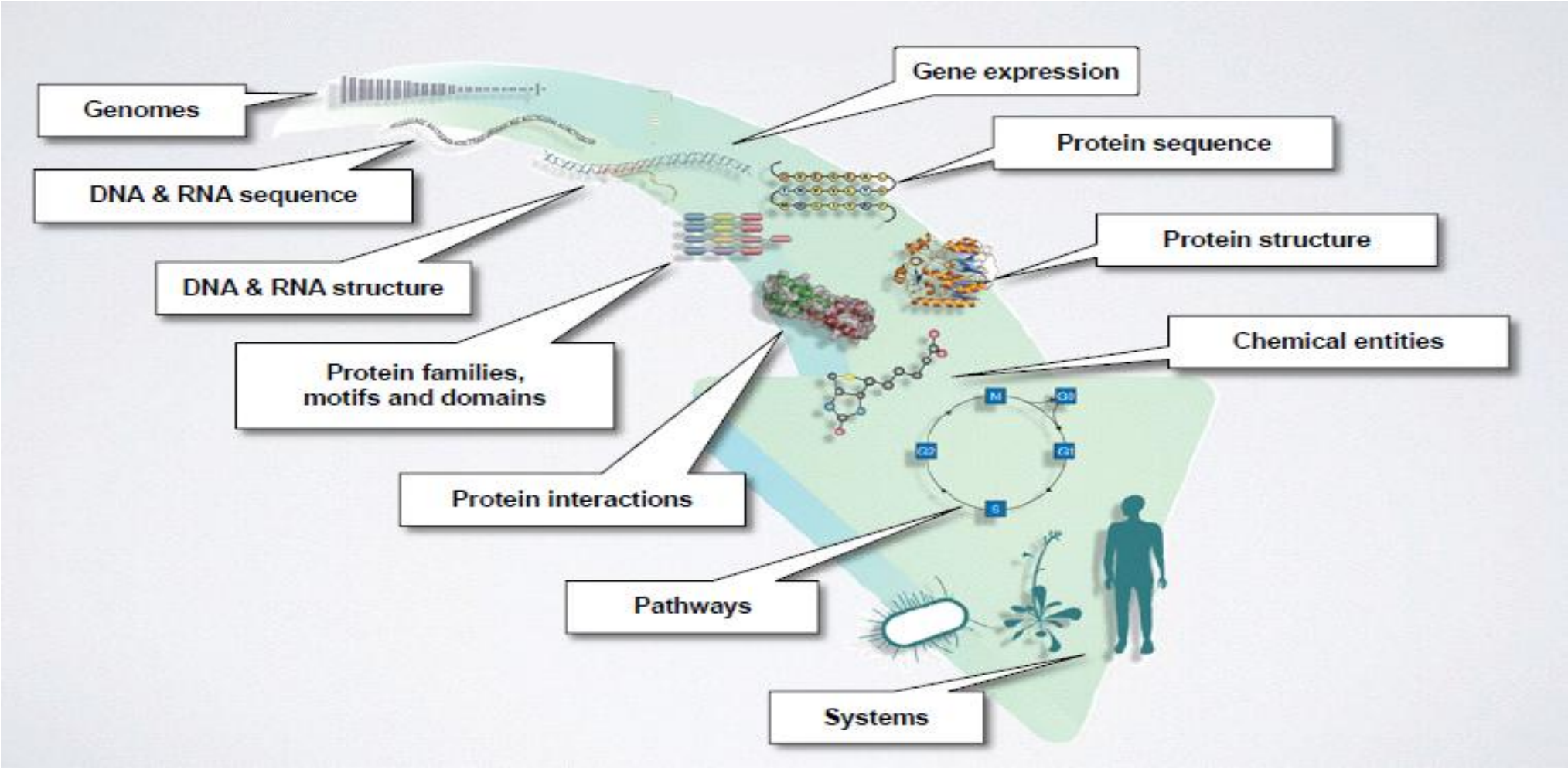
OMICS

- Genomics: DNA.
- Transcriptomics: RNA.
- Proteomics: Proteins.
- Metagenomics: is a new research area focused on the analysis of mixture of DNA sequences extracted from different organisms i.e. viral, bacterial, or eukaryotic that are living together in a symbiotic community.
- Epigenomics is the study of the complete set of epigenetic modifications on the genetic material of a cell, known as the epigenome. The epigenome is made up of chemical compounds and proteins that can attach to DNA and direct such actions as turning genes on or off, controlling the production of proteins in particular cells.

OMICS ERA (LARGE SCALE BIOLOGY)

Name	Study of
Genomics	entire genome of an organism
Transcriptomics	expressed genes
Exomics	coding sequences
Proteomics	proteins within an organism
Metagenomics	mixture of DNA sequences from different organisms
Metabolomics	metabolites within an organism
Interactomics	interactions between nucleotides, proteins and metabolites
Connectomics	neural pathways in the brain
Pharmocogenomics	application of genomics to pharmacology
Phenomics	observable phenotypes
Physiomics	functional behaviour of an organism
Exposomics	organism's environment

TYPES OF BIOINFORMATICS DATA (BIOLOGICAL MOLECULES)



Thank you!