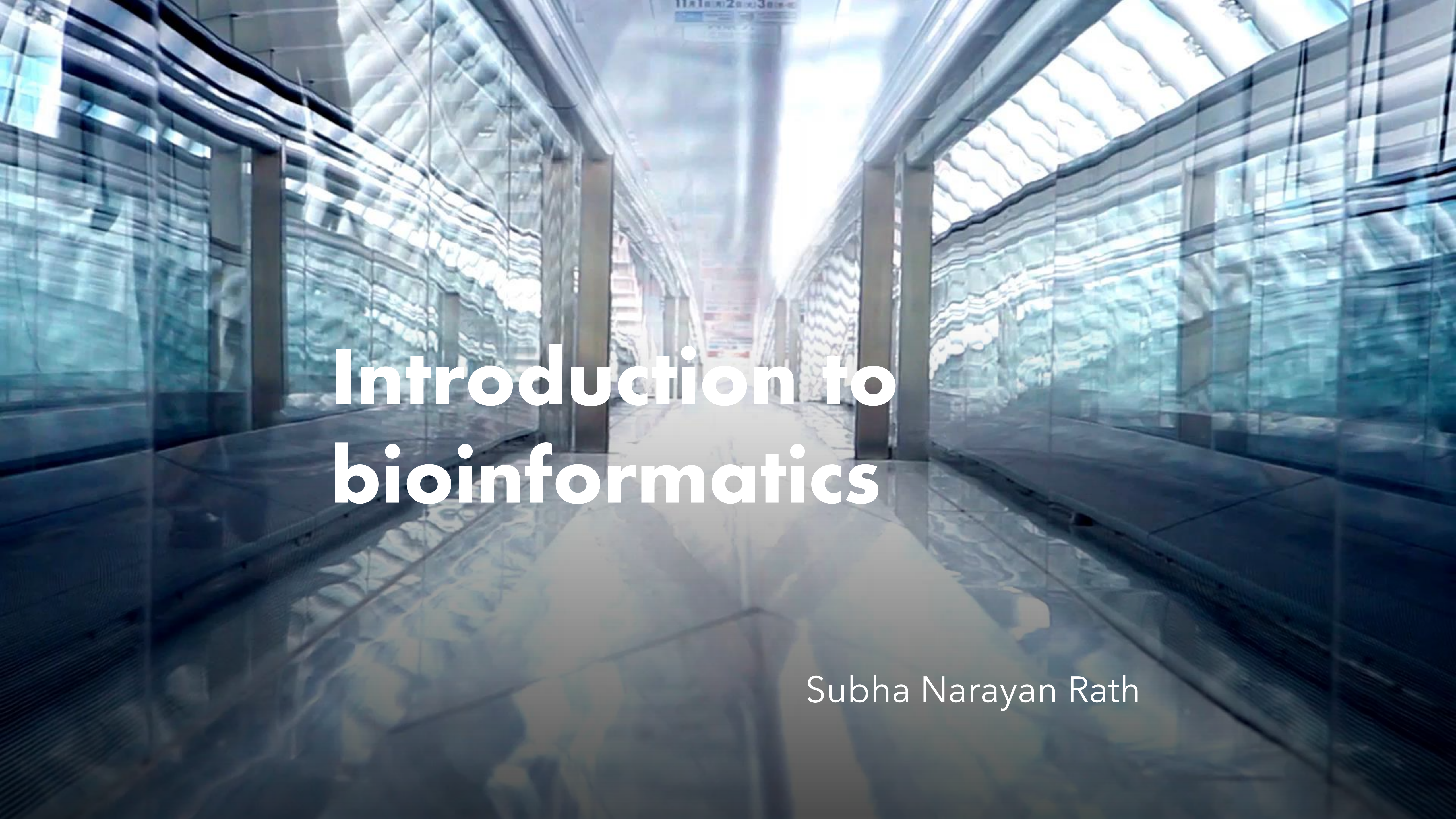


# Introduction to bioinformatics

Subha Narayan Rath

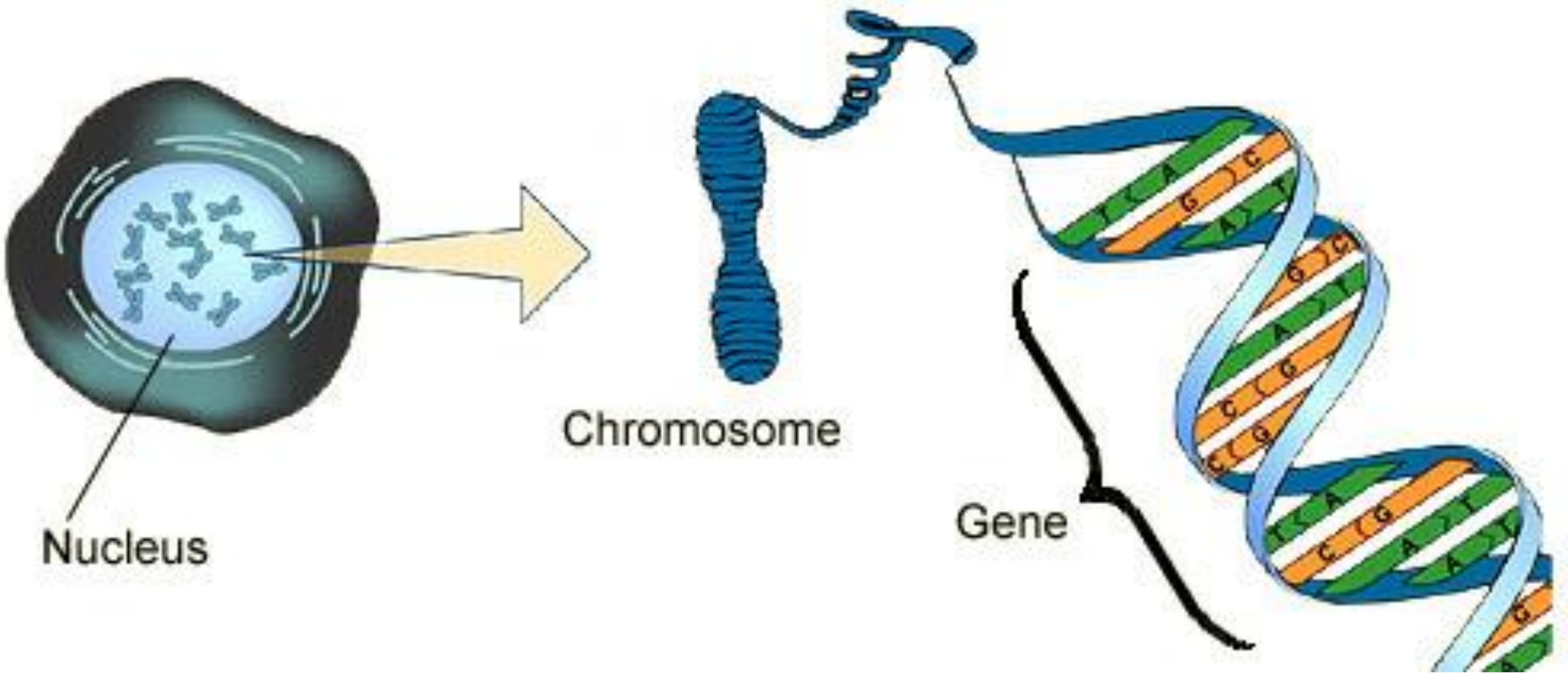
# The data of bioinformatics

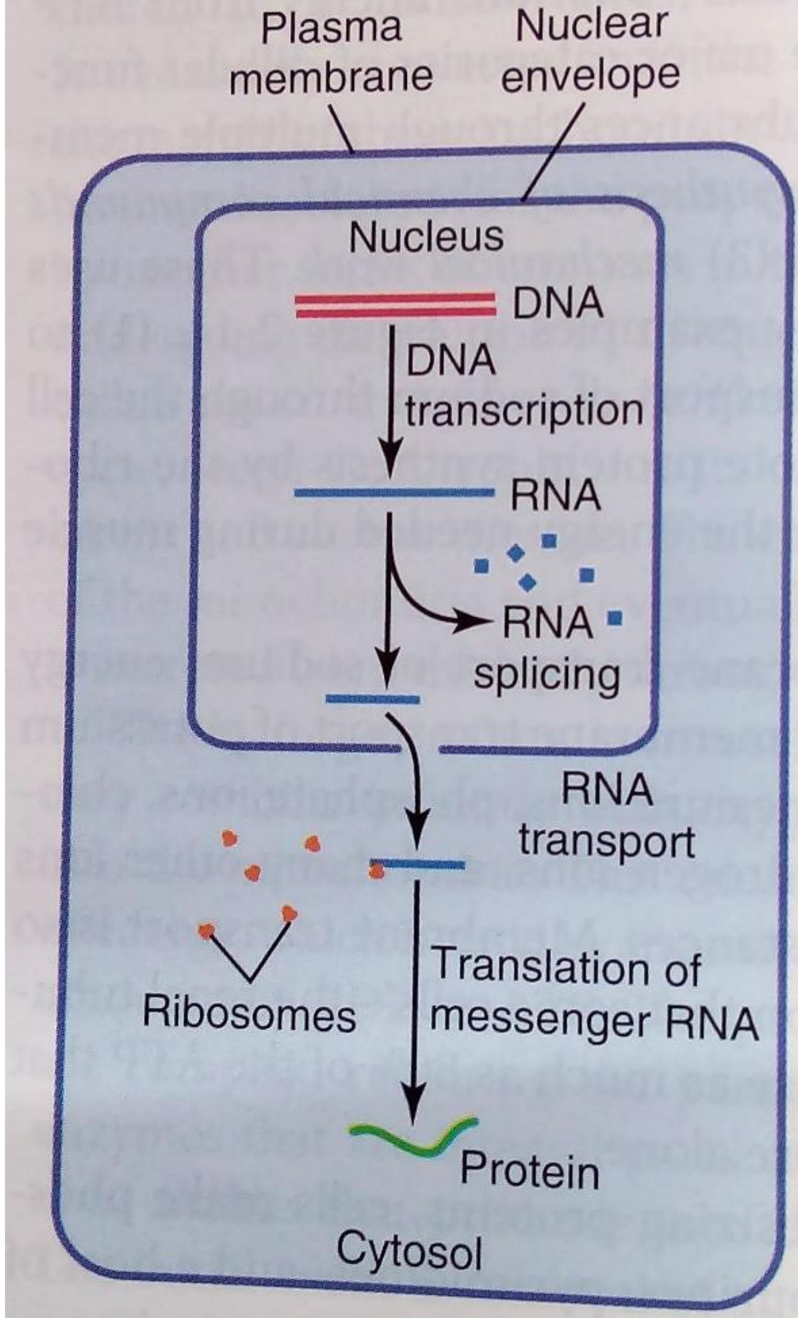
- Nucleotide sequence database: 6*10e11 bases (600 Gbp)
- Human genome: 3 * 10e9
- 200 human genome equivalent data are there.
- The databse of macromolecular structures: 100 000 entries with full 3d corordinates of proteins, amino acids etc.
- Phenotype= genotype+ environment+ life history+ epigenetics
- Alleles are different forms or sequences of the same gene
- Homozyogsity and heterozygosity…

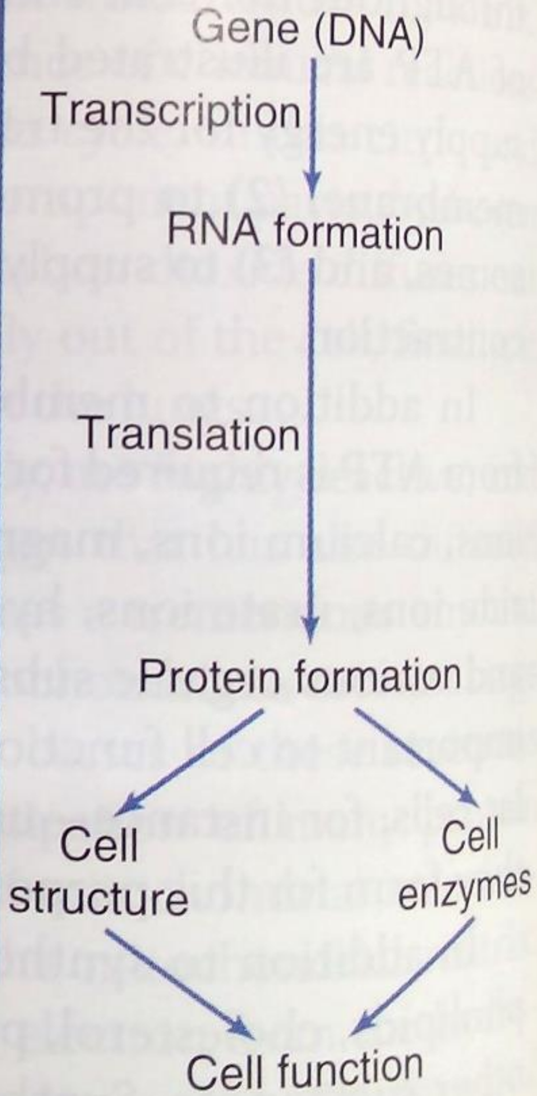
# Definition

- Bioinformatics is defined as the **application of tools of computation and analysis to the capture and interpretation of biological data**. It is an interdisciplinary field, which harnesses computer science, mathematics, physics, and biology

- (Molecular) Bioinformatics is conceptualizing biology in terms of
  - molecules (in the sense of Physical chemistry) 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.

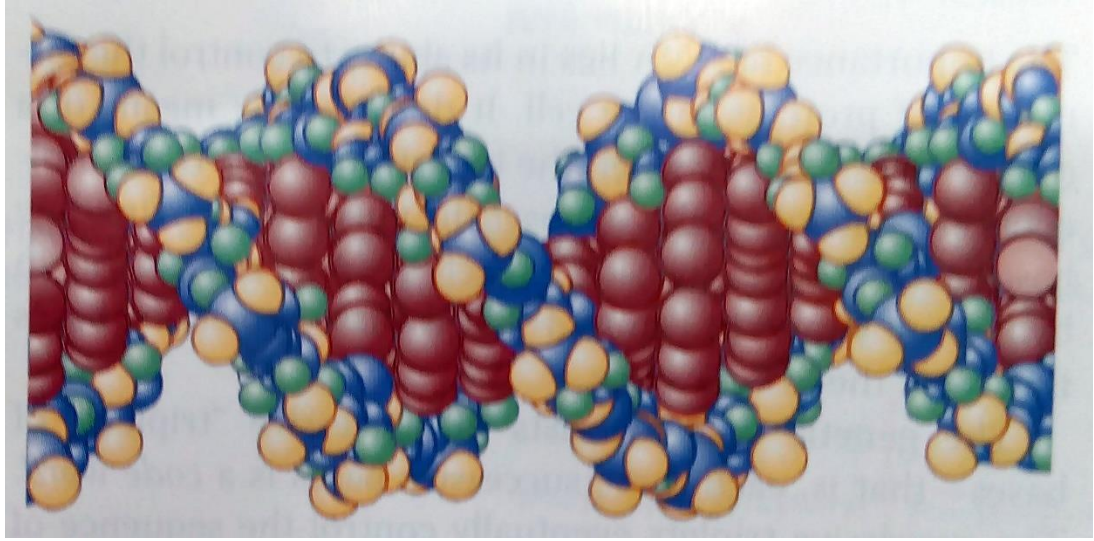

Nucleus
Chromosome
Gene

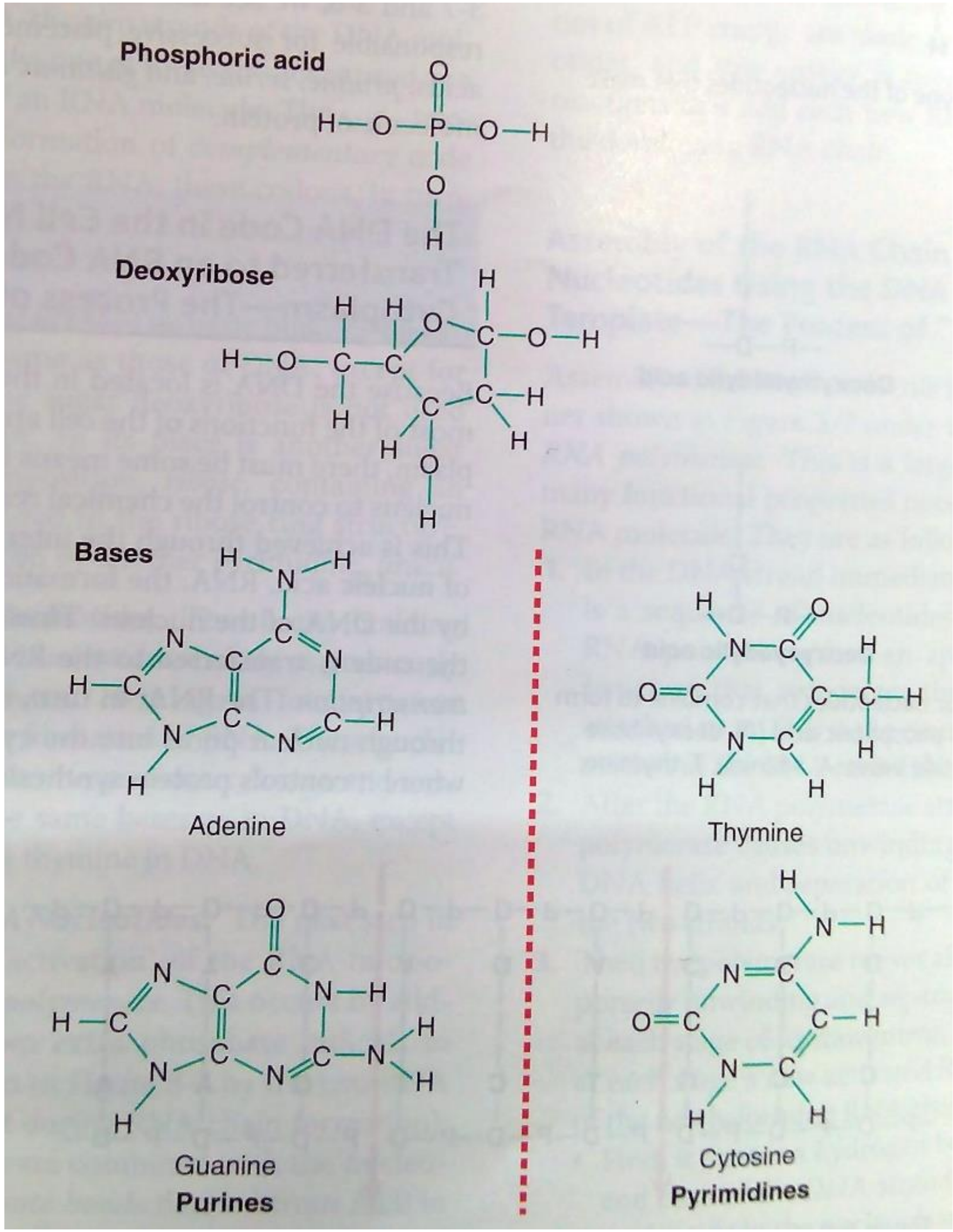

Plasma membrane
Nuclear envelope
Nucleus
DNA
DNA transcription
RNA
RNA splicing
RNA transport
Ribosomes
Translation of messenger RNA
Protein
Cytosol


Gene (DNA)
Transcription
RNA formation
Translation
Protein formation
Cell structure
Cell enzymes
Cell function

# Around 30,000 genes in human

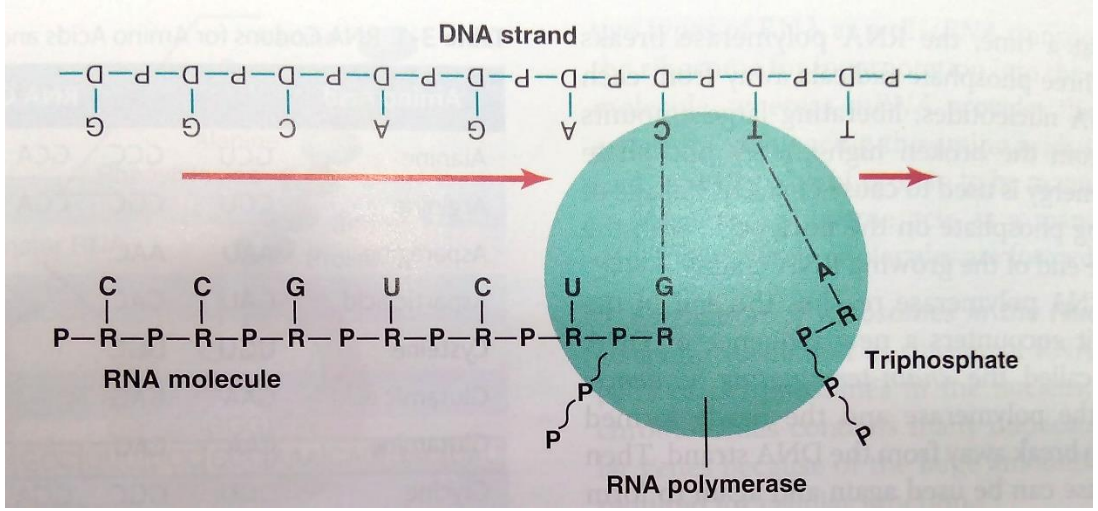
- Outside strands= phosphoric acid+ deoxyribose sugar
- Inside strands= 4 nitrogenous bases such as purines (AG) or pyrimidines (CT) bases determining the code of the genes
- Nucleotide= Phosphoric acid+deoxyribose+Base

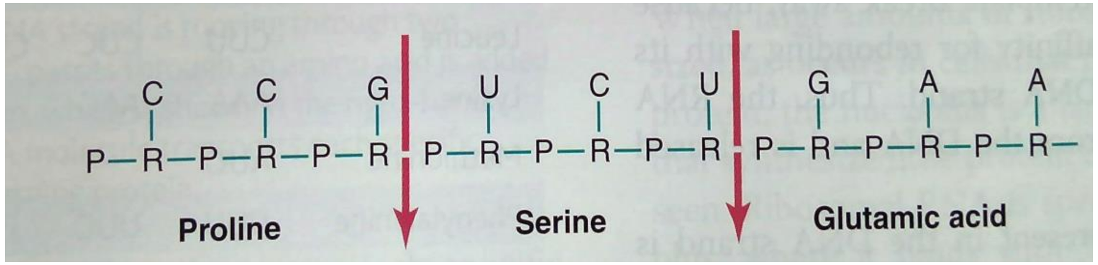



Phosphoric acid
Deoxyribose
Bases
Adenine
Thymine
Guanine
Cytosine
Purines
Pyrimidines

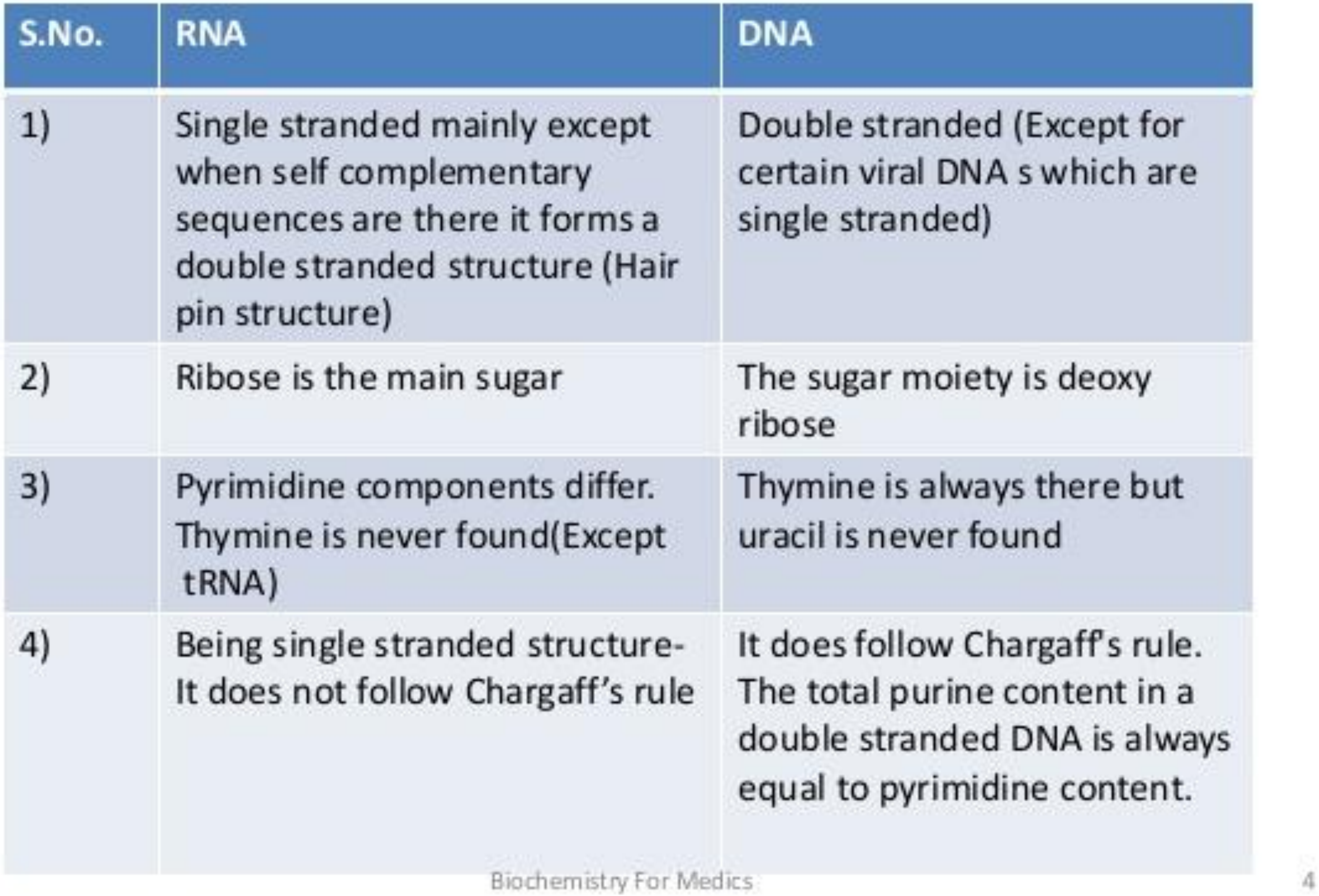
# Genetic code

- Three successive bases= a code word
- Transcription: DNA > RNA



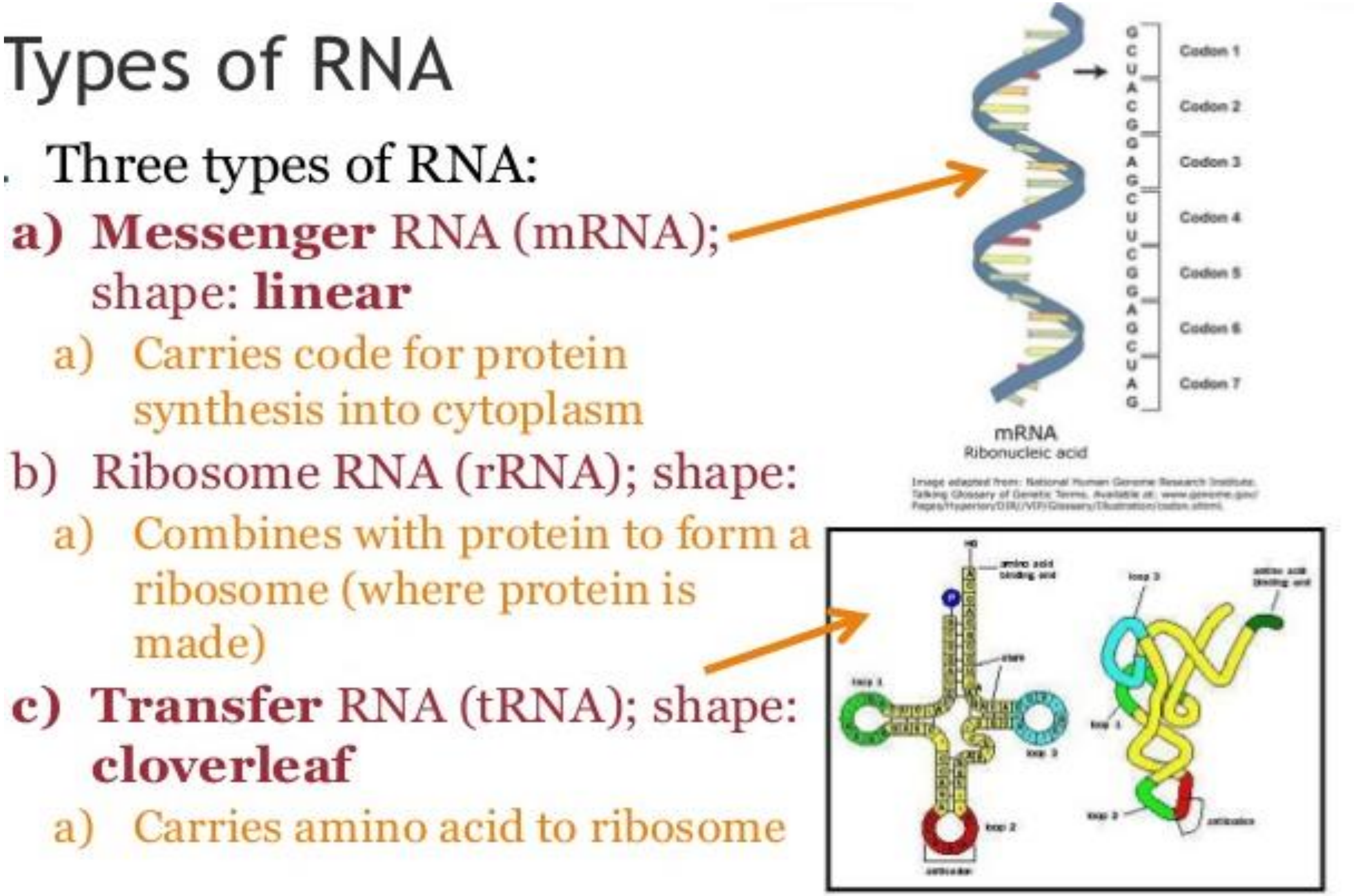


# Difference between DNA & RNA



| S.No. | RNA | DNA |
|---|---|---|
| 1) | Single stranded mainly except when self complementary sequences are there it forms a double stranded structure (Hair pin structure) | Double stranded (Except for certain viral DNA s which are single stranded) |
| 2) | Ribose is the main sugar | The sugar moiety is deoxy ribose |
| 3) | Pyrimidine components differ. Thymine is never found(Except tRNA) | Thymine is always there but uracil is never found |
| 4) | Being single stranded structure- It does not follow Chargaff's rule | It does follow Chargaff's rule. The total purine content in a double stranded DNA is always equal to pyrimidine content. |

# Types of RNA



## Types of RNA

- Three types of RNA:

a) **Messenger** RNA (mRNA); shape: **linear**

   a) Carries code for protein synthesis into cytoplasm

b) Ribosome RNA (rRNA); shape:

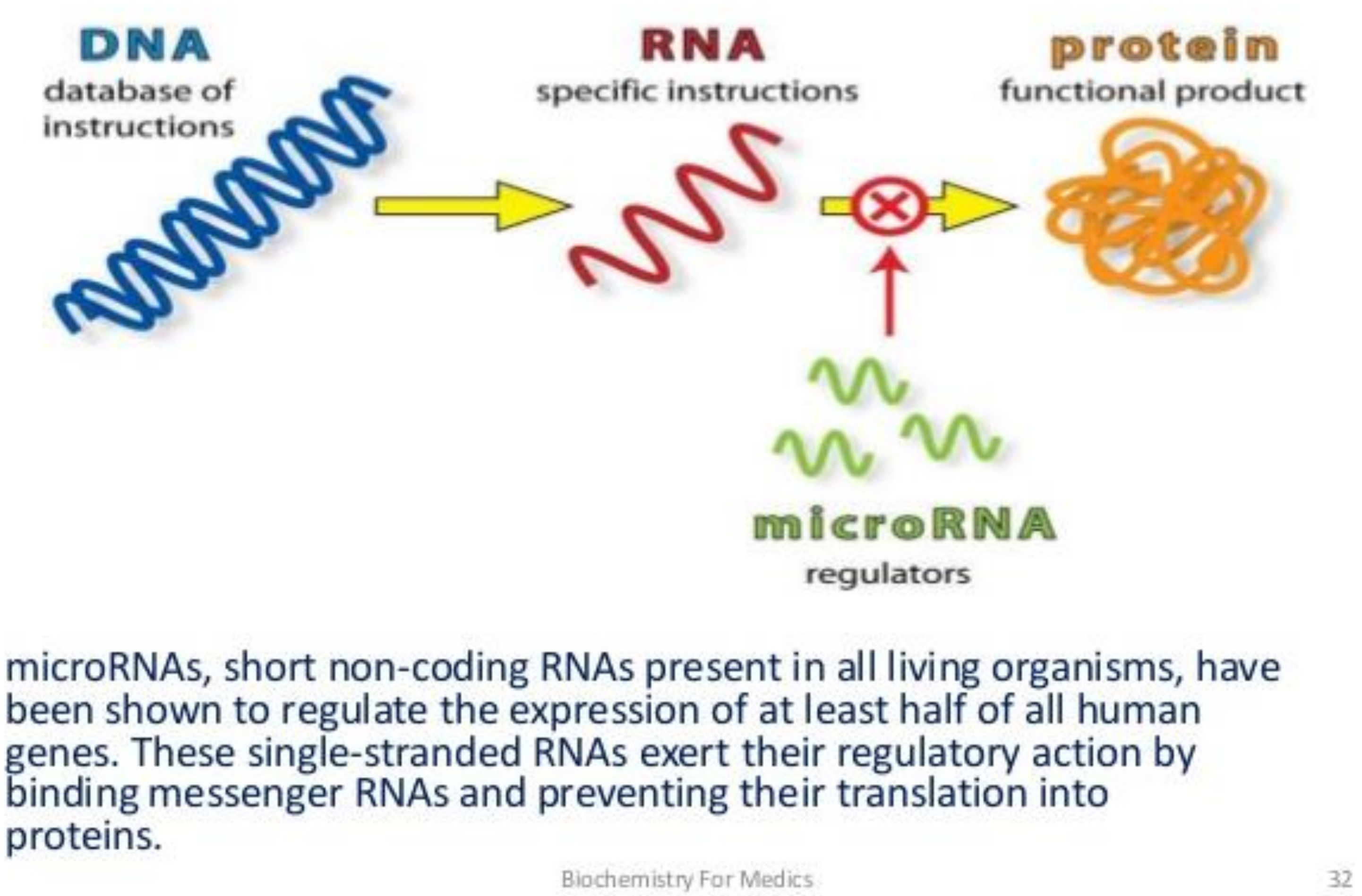
   a) Combines with protein to form a ribosome (where protein is made)

c) **Transfer** RNA (tRNA); shape: **cloverleaf**

   a) Carries amino acid to ribosome

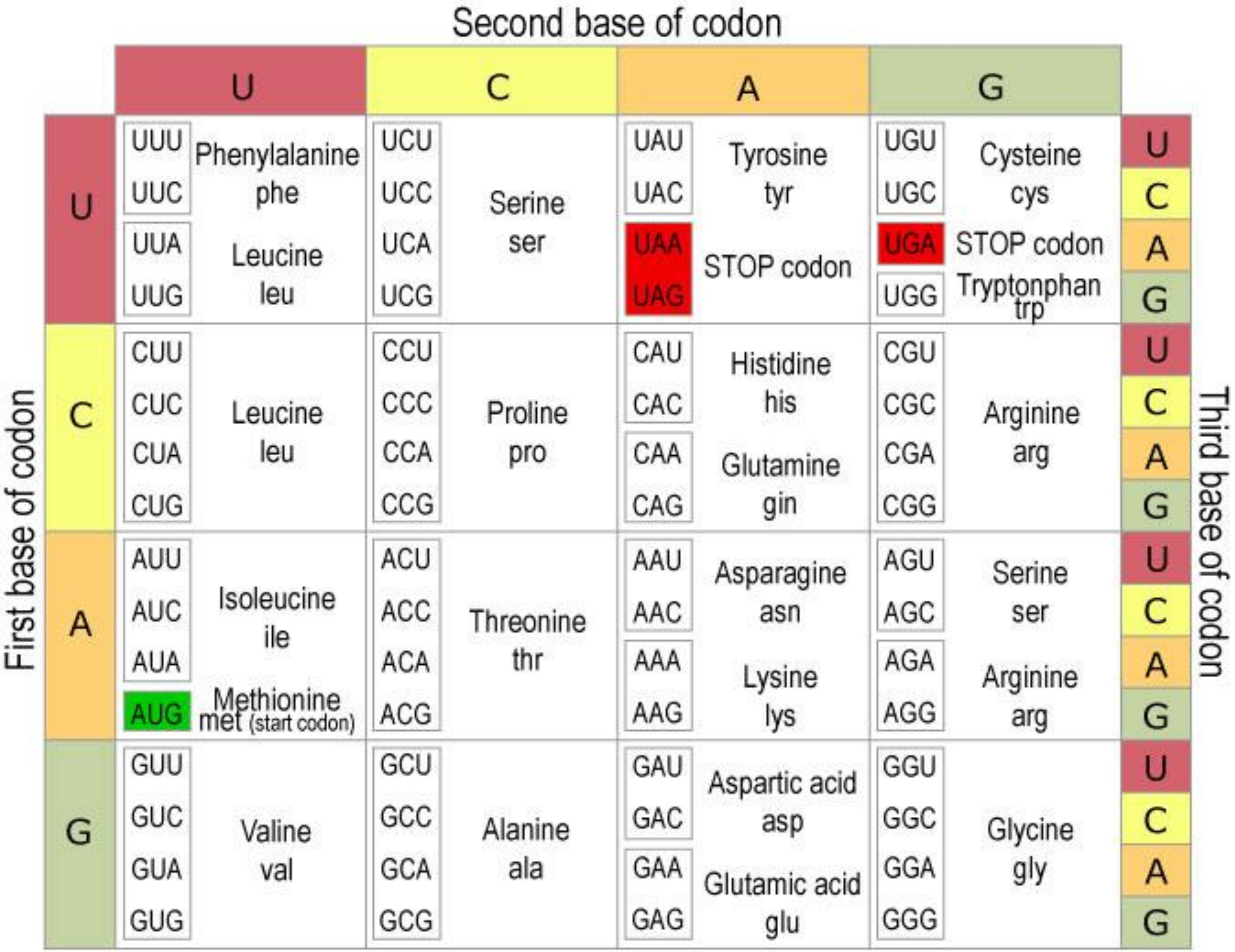
Image adapted from: National Human Genome Research Institute. Talking Glossary of Genetic Terms. Available at: www.genome.gov/Pages/Hyperion/DIR/VIP/Glossary/Illustration/codon.shtml.

# Micro RNA



microRNAs, short non-coding RNAs present in all living organisms, have been shown to regulate the expression of at least half of all human genes. These single-stranded RNAs exert their regulatory action by binding messenger RNAs and preventing their translation into proteins.

Biochemistry For Medics 32



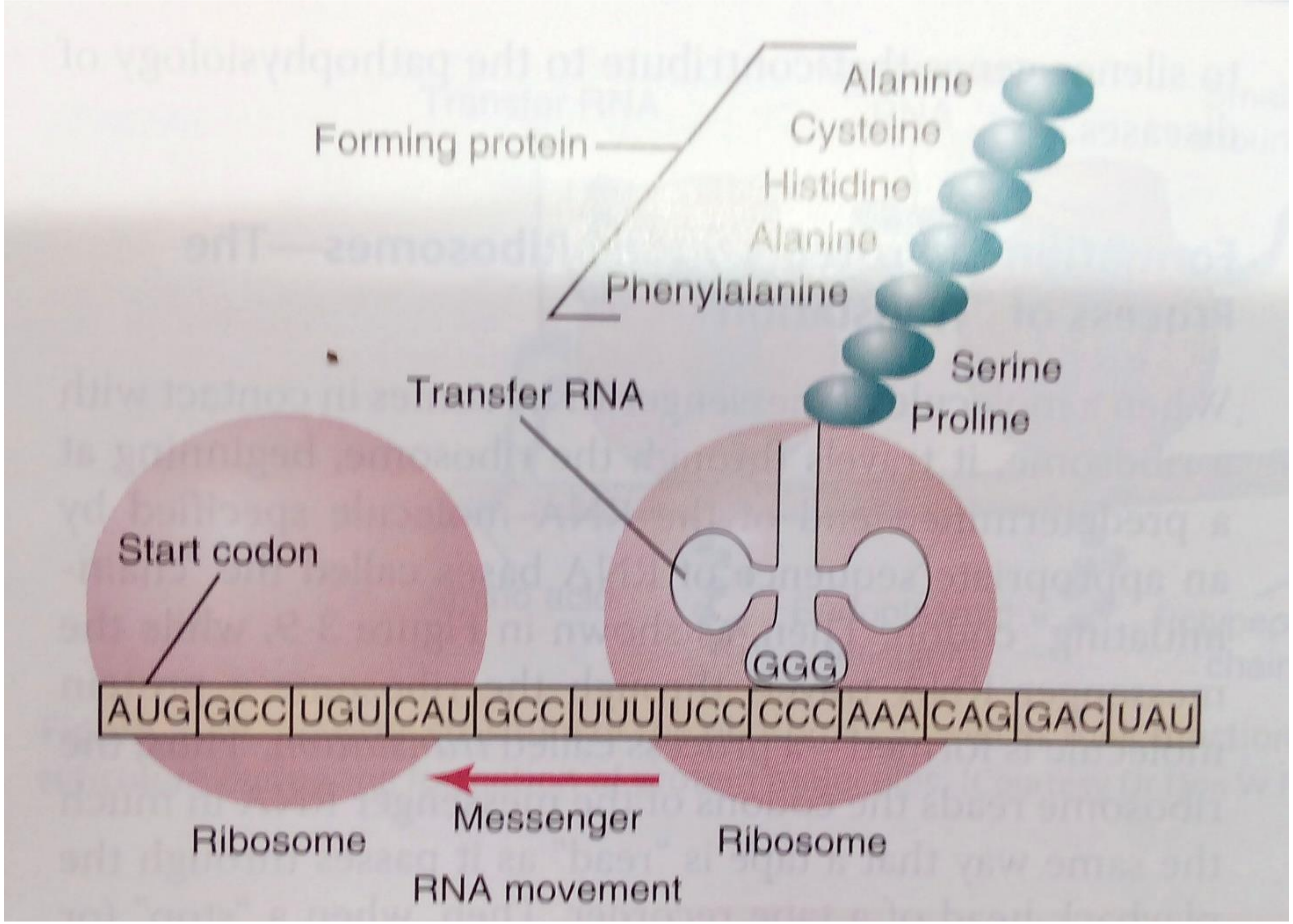
Second base of codon

First base of codon

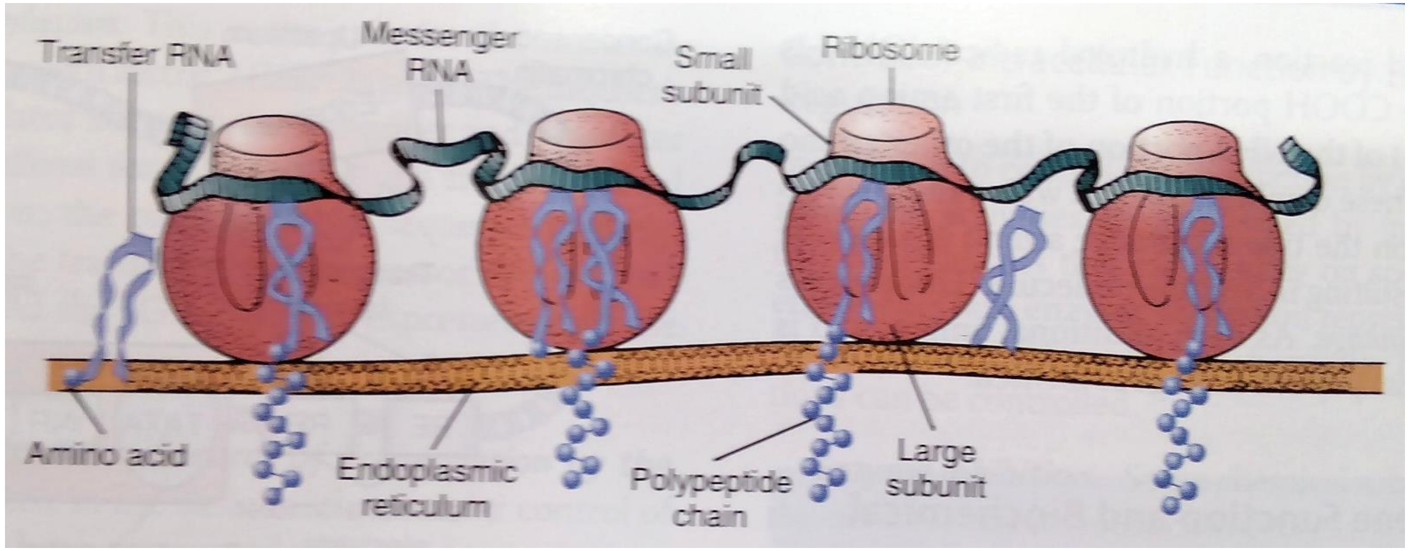
Third base of codon

| | U | | C | | A | | G | | |
|---|---|---|---|---|---|---|---|---|---|
| U | UUU | Phenylalanine<br>phe | UCU | Serine<br>ser | UAU | Tyrosine<br>tyr | UGU | Cysteine<br>cys | U |
| | UUC | | UCC | | UAC | | UGC | | C |
| | UUA | Leucine<br>leu | UCA | | UAA | STOP codon | UGA | STOP codon | A |
| | UUG | | UCG | | UAG | | UGG | Tryptonphan<br>trp | G |
| C | CUU | Leucine<br>leu | CCU | Proline<br>pro | CAU | Histidine<br>his | CGU | Arginine<br>arg | U |
| | CUC | | CCC | | CAC | | CGC | | C |
| | CUA | | CCA | | CAA | Glutamine<br>gin | CGA | | A |
| | CUG | | CCG | | CAG | | CGG | | G |
| A | AUU | Isoleucine<br>ile | ACU | Threonine<br>thr | AAU | Asparagine<br>asn | AGU | Serine<br>ser | U |
| | AUC | | ACC | | AAC | | AGC | | C |
| | AUA | | ACA | | AAA | Lysine<br>lys | AGA | Arginine<br>arg | A |
| | AUG | Methionine<br>met (start codon) | ACG | | AAG | | AGG | | G |
| G | GUU | Valine<br>val | GCU | Alanine<br>ala | GAU | Aspartic acid<br>asp | GGU | Glycine<br>gly | U |
| | GUC | | GCC | | GAC | | GGC | | C |
| | GUA | | GCA | | GAA | Glutamic acid<br>glu | GGA | | A |
| | GUG | | GCG | | GAG | | GGG | | G |

# mRNA=codons; tRNA= anti-codons



# Translation: formation of protein on the ribosomes by polyribosomes



# Control of gene function and biochemical activity of cells

- 1. GENETIC REGULATION
- Promoter (TATA box or TATAAAA)
- TF: + OR – transcription factors
- Enhancers
- Hormones: signals from outside the cells

- 2. ENZYME REGULATION
- Enzyme inhibition (negative feedback control)
- Enzyme activation (e.g. cAMP activates glycogen breakdown to form more ATPs)
- E.g. purine and pyrimidine formation uses both of the above mechanisms

# Dogmas: central and peripheral

- DNA > m RNA > Protein
- Strands in double helix are anti-parallel, direction either 3` or 5` (for deoxy-ribose ring)
- Transcription of DNA TO RNA and translation of m RNA: always read from 5` position
- Protein formation requires splicing or removal of non-coding regions
- Several proteins from same gene: by mixing and matching of exons
- Other types of RNA such as siRNA, microRNA, piwi-interacting RNAs: control translation
- Triplets of code from DNA act as cipher for protein code (as fig)

# Encode project

- To understand function of entire human genome: Encyclopaedia of DNA elements
- https://www.nature.com/collections/aghcdefffg/
- 80% of human genome can be ascribed to some function…compared to previously thought 23,000 protein coding genes which is 1.5% of genome
- The rest is called by some junk DNA
- Variable splicing means number of proteins are not limited to these genes
- TWO ways non-coding regions to have function
  - 1. involved in sequence dependent physical interaction: within chromatin that either expose it or block from protein ligands
  - 2. if transcribed to RNA: functions like regulation of transcription

# Results of ENCODE analysis

- 75% of human genome is transcribed
- Mapping and dictionary of regulatory sites: many to one i.e. many proteins can bind to same regulatory sites
- A sketch of structure of regulatory network
- Mapping of exposed sites of chromatin, which are unprotected from DNase1 cleavage: these are regulatory sites near genes for binding of regulators of expression

# Proteins as end results

- 200-400 Amino acids length
- Exons and introns: among introns some are for regulation and some considered junks
- Proteins and structural RNAs vary a lot in their 3d structure
- For each protein or peptide sequence: there is a stable native state adopted spontaneously
- The paradigm is: (and this is focus of bioinformatics)
    - DNA structure determines protein sequence
    - That determines protein structure
    - That determines protein function
    - Regulatory mechanisms like control of expression patterns

# Molecular biology dynamics

- Cell RNA content: Transcriptome
- DNA methylation patterns
- Splice variants and post-translational modifications of proteins in any cell
- Patterns in protein-protein interaction, DNA-Protein interaction with finding of exact regions binding for it
- Integration of individual regulatory steps into networks