# Molecular Biology Primer

Figures and slides taken from various sources on the internet including, www.bioalgorithms.info (sources mostly acknowledged)

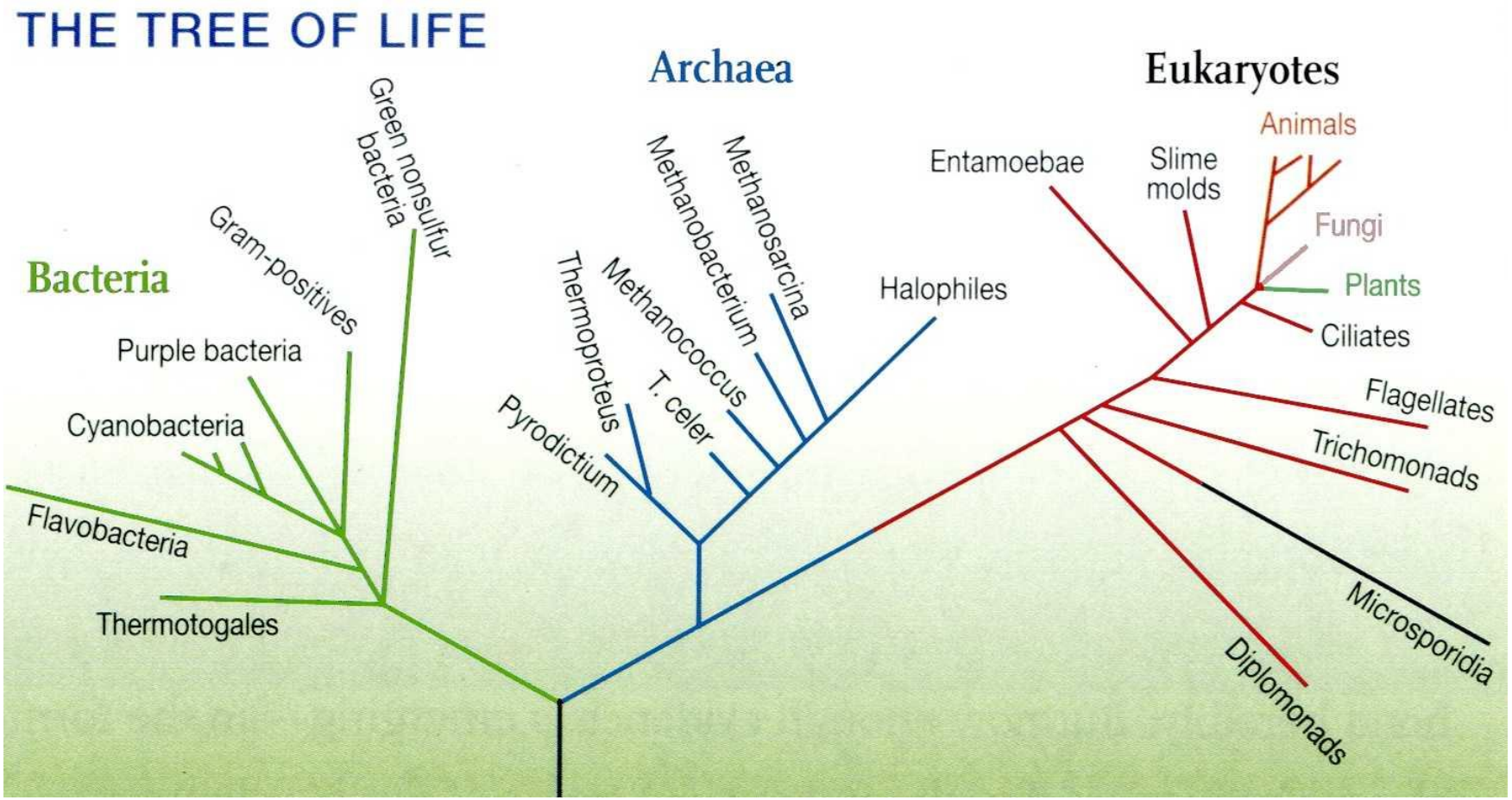
# Cells

- **Fundamental working units** of every living system.
- Every organism is composed of one of two radically different types of cells:

  **prokaryotic** cells or

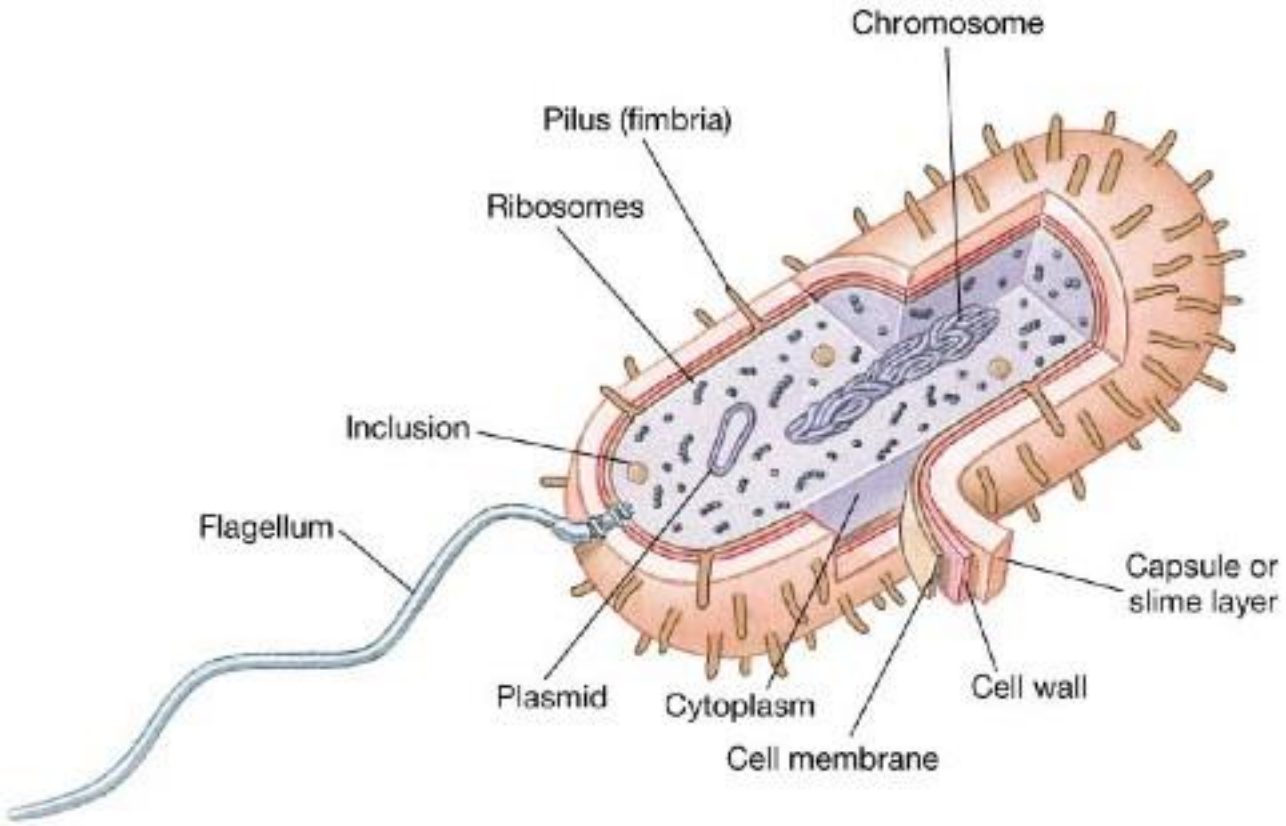
  **eukaryotic** cells.
- **Prokaryotes** and **Eukaryotes** are descended from the same primitive cell.
  - All extant prokaryotic and eukaryotic cells are the result of a total of 3.5 billion years of evolution.

# Evolutionary Tree of Life



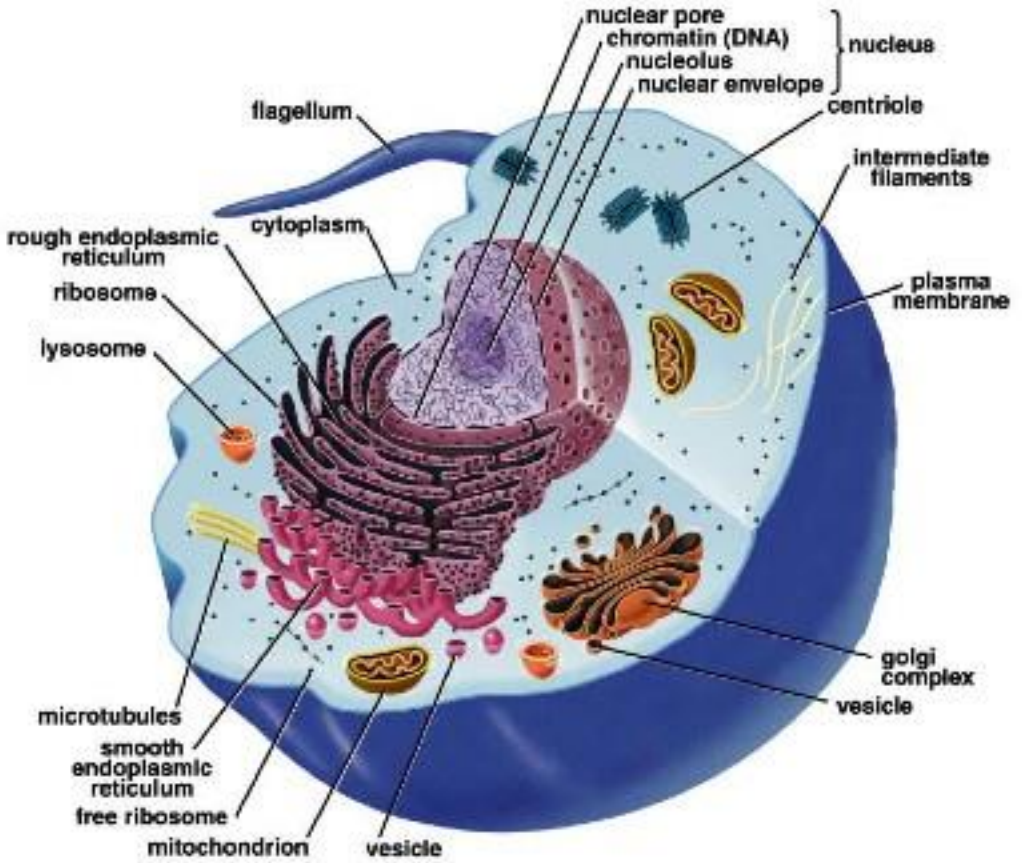

Njsas.org

# Bacterial Cell (Prokaryote)




Uccs.edu

# Animal Cell (Eukaryotic)




Faculty.southwest.tn.edu

# Prokaryotes and Eukaryotes

| **Prokaryotes** | **Eukaryotes** |
| --- | --- |
| Single cell | Single or multi cell |
| No nucleus | Nucleus |
| No organelles | Organelles |
| One piece of circular DNA | Chromosomes |
| No mRNA post transcriptional modification | Exons/Introns splicing |

# Cells: Information and Machinery

- Cells store all information to replicate itself
  - Human genome is around 3 billions base pair long
  - Almost every cell in human body contains same set of genes
  - But not all genes are used or expressed by those cells
- Machinery:
  - Collect and manufacture components
  - Carry out replication
  - Kick-start its new offspring

  (A cell is like a car factory)

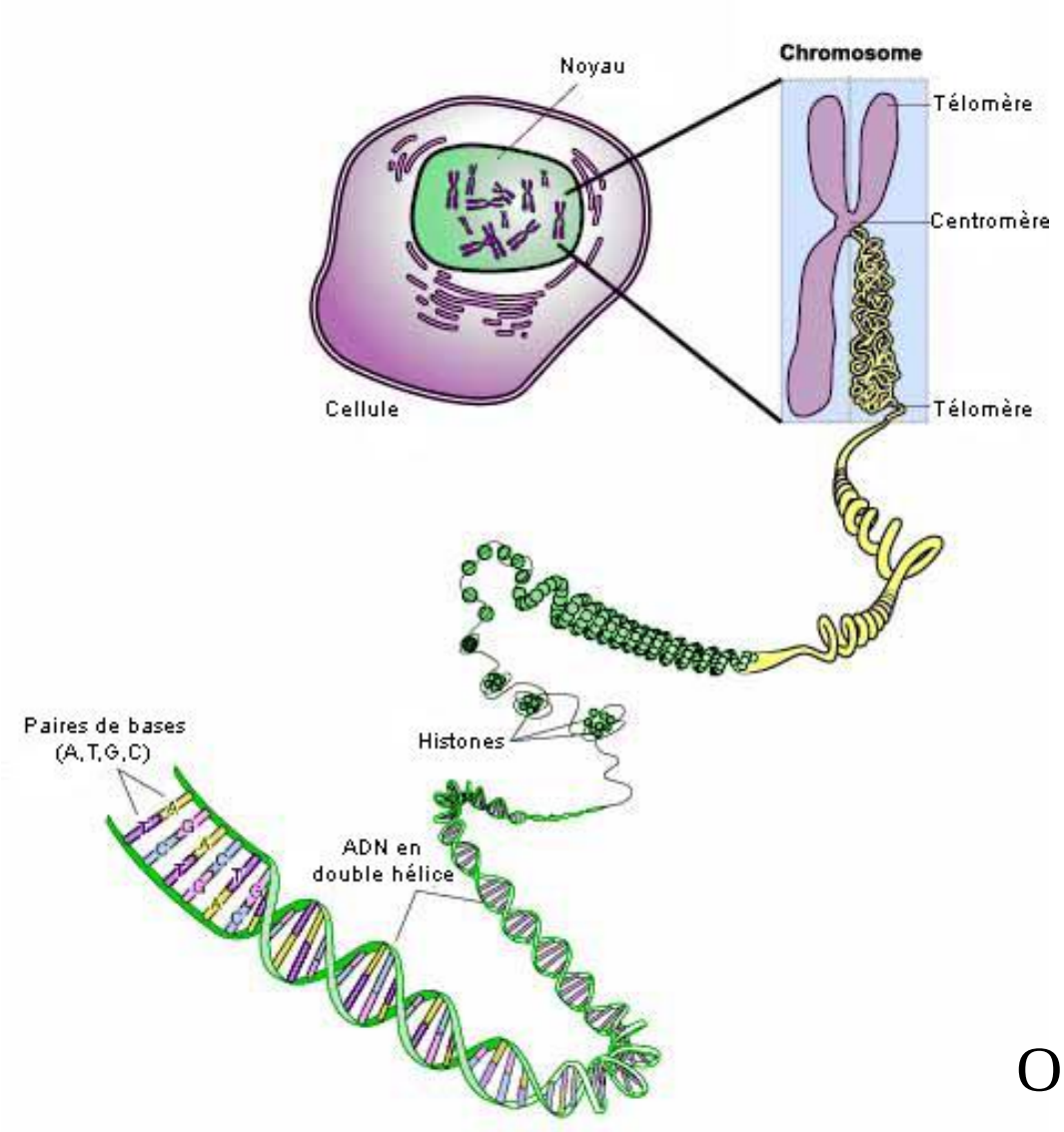
# Some Terminology

- **Genome**: **an organism's genetic material**
- **Gene**: **a discrete units of hereditary information located on the chromosomes and consisting of DNA.**
- **Genotype**: **The genetic makeup of an organism**
- **Phenotype**: **the physical expressed traits of an organism**
- **Nucleic acid**: **Biological molecules(RNA and DNA) that allow organisms to reproduce;**

# All Life depends on 3 critical molecules

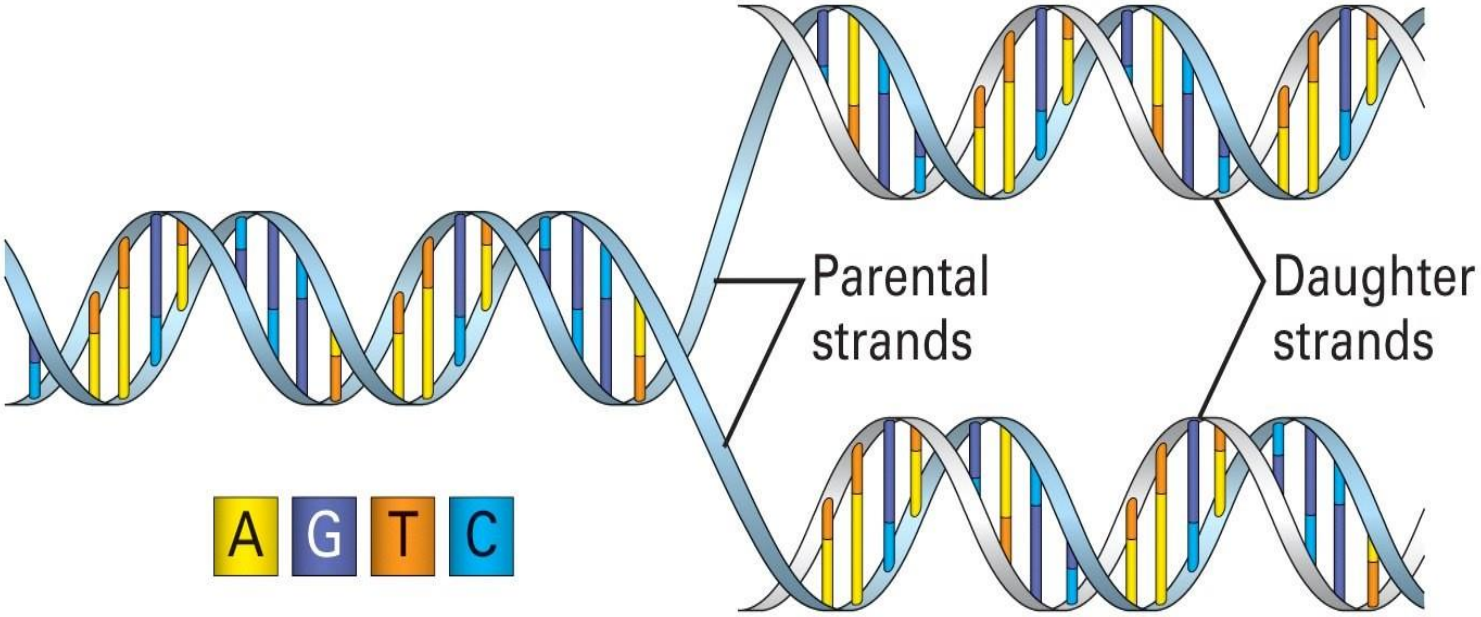
- DNAs
  - Hold information on how cell works
- RNAs
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body's major components (e.g., hair, skin, etc.)

# Cells & DNA



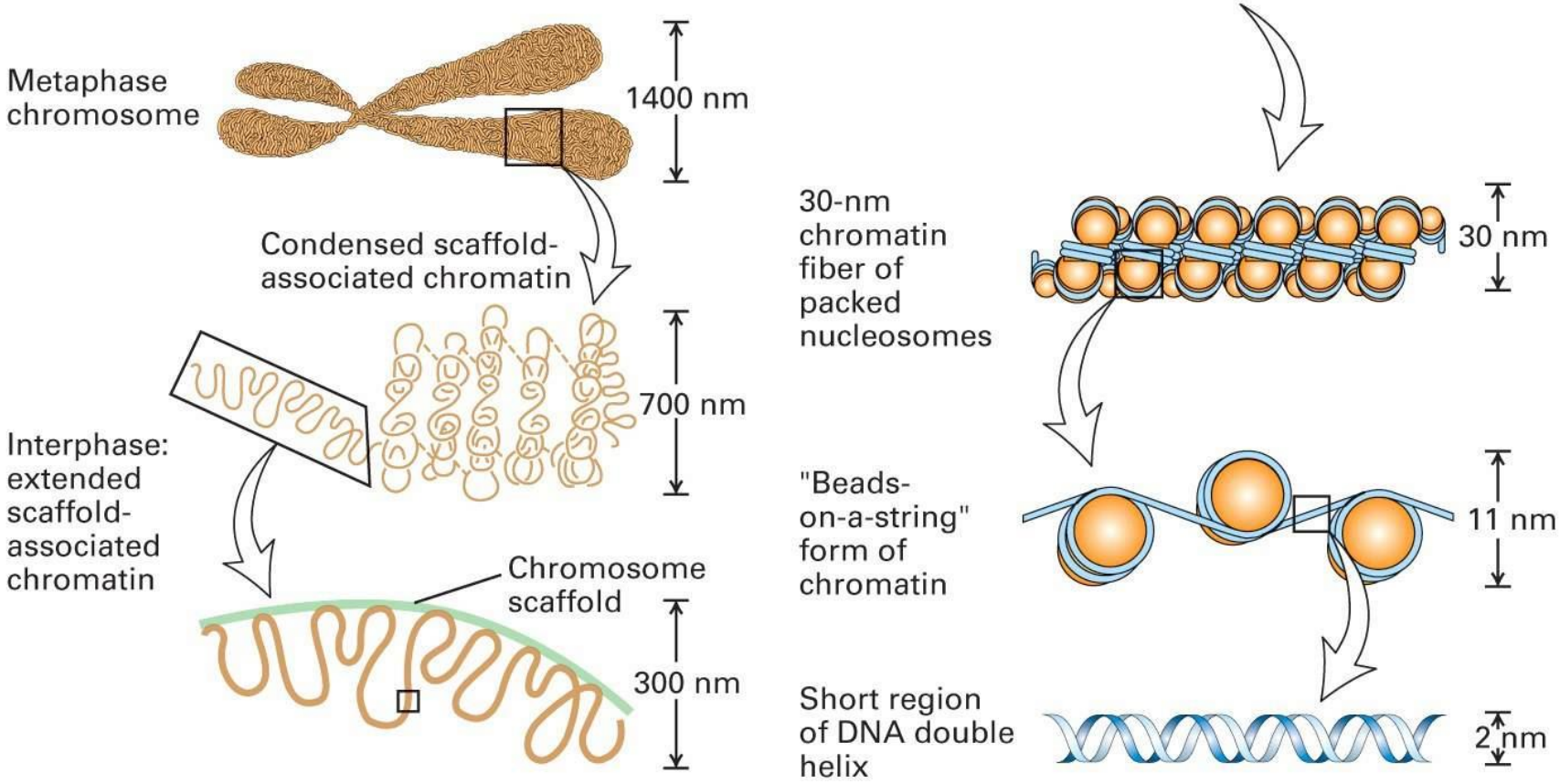

Ogm-info.com

# DNA: The Code of Life




- The structure and the four genomic letters code for all living organisms
- Adenine, Guanine, Thymine, and Cytosine which pair A-T and C-G on complimentary strands.

# Superstructure: DNA and Chromosome




Lodish et al. *Molecular Biology of the Cell* (5th ed.). W.H. Freeman & Co., 2003.

(length of 1 bp)(number of bp per cell)(number of cells in the body)

$(0.34 \times 10^{-9}\ \mathbf{m})(3 \times 10^{9})(10^{13}) = \mathbf{1.0 \times 10^{13}\ meters\ (35\ trips\ to\ Sun!)}$

# Chromosomes

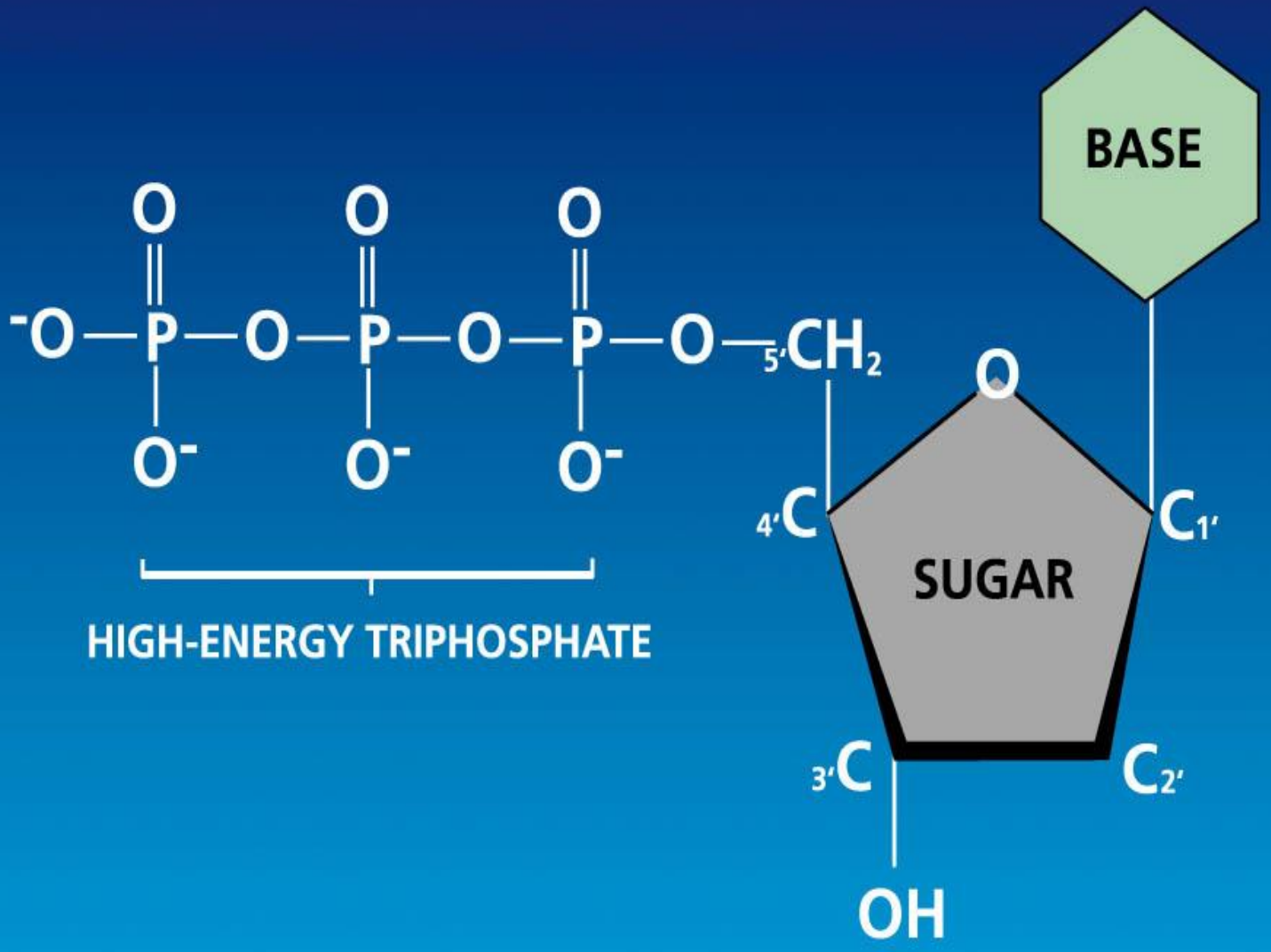
| Organism | Number of base pair | number of Chromosomes |
|---|---|---|
| Prokayotic | | |
| Escherichia coli (bacterium) | $4x10^6$ | 1 |
| | | |
| Eukaryotic | | |
| Saccharomyces cerevisiae (yeast) | $1.35x10^7$ | 17 |
| Drosophila melanogaster(insect) | $1.65x10^8$ | 4 |
| Homo sapiens(human) | $2.9x10^9$ | 23 |
| Zea mays(corn) | $5.0x10^9$ | 10 |

# Building Blocks of Biological Systems:

***nucleotides and amino acids***

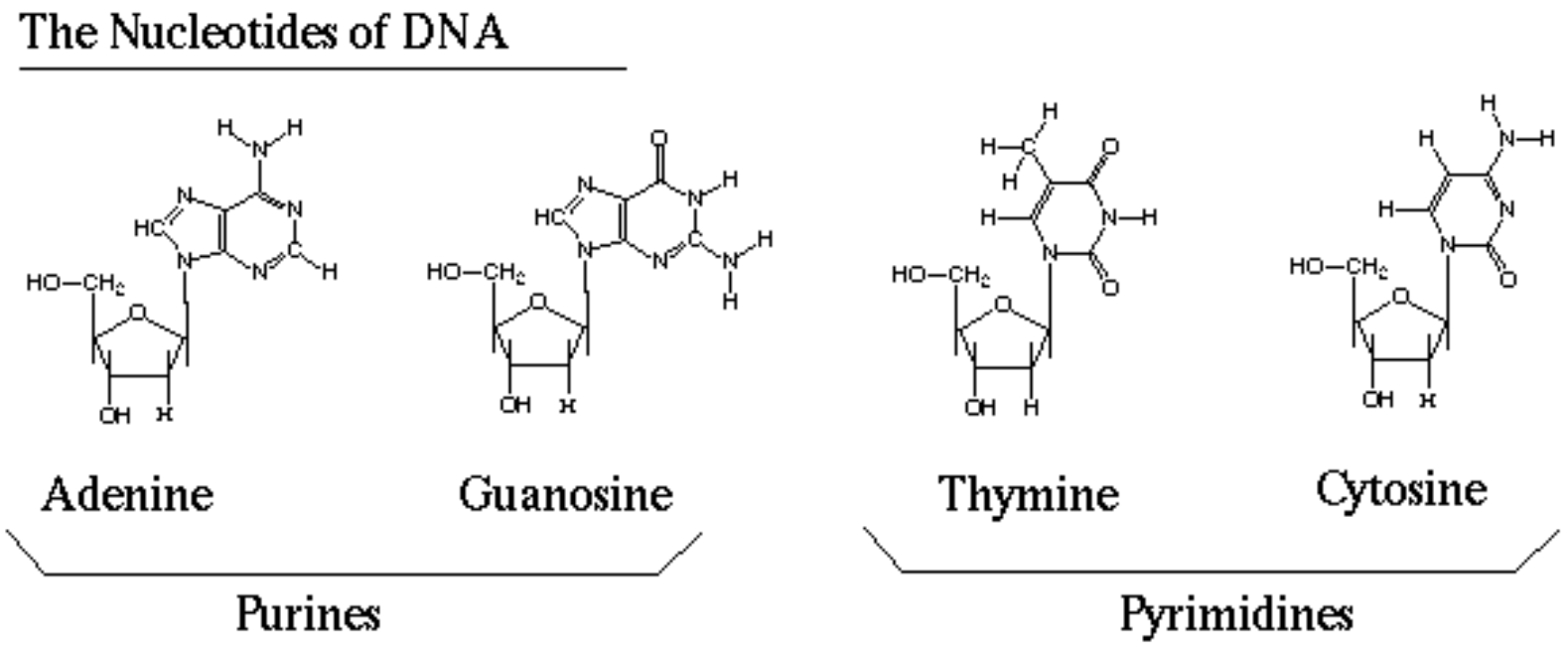
**DNA** (nucleotides, 4 types): information carrier/encoder.
**RNA**: bridge from DNA to protein.
**Protein** (amino acids, 20 types): action molecules.

# Structure of nucleotides



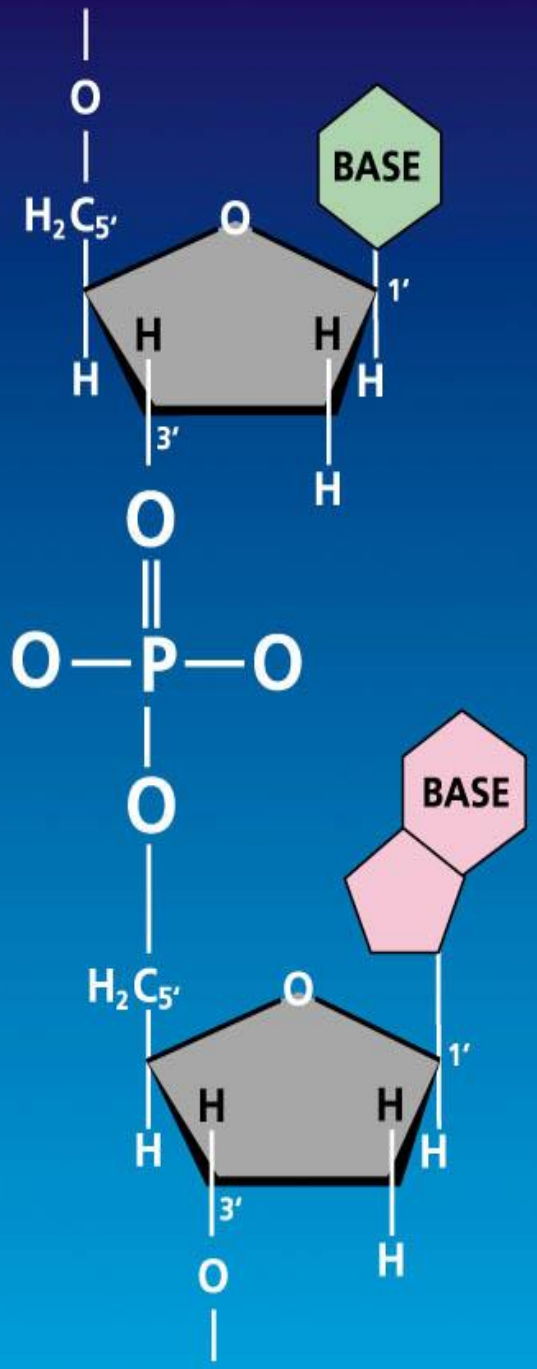
# Four types of nucleic acids of DNA

Deoxyribonucleic Acid (DNA) contains four nucleotide bases.




Note that A pairs with T; and G pairs with C.

# Nucleotides linked by phosphate bonds

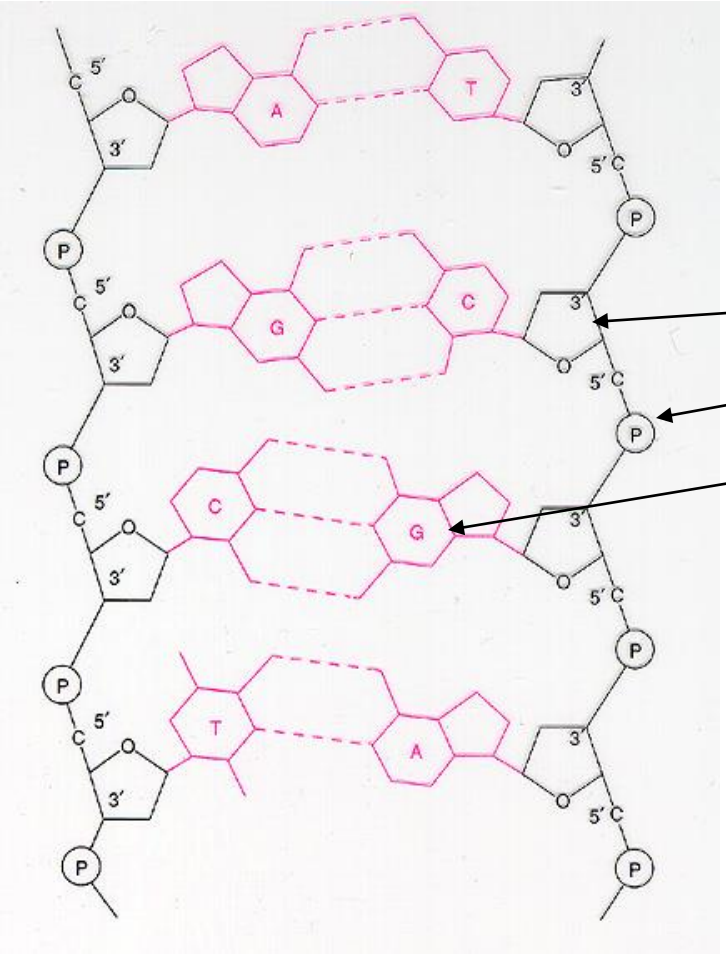


# Primary Structure of DNA

- Unbranched polymer
- Sequence of nucleotide bases
- Double stranded

```
atgaatcgta ggggtttgaa cgctggcaat
acgatgactt ctcaagcgaa cattgacgac
ggcagctgga aggcggtctc cgagggcgga ……
```

# DNA, continued




- DNA has a double helix structure which is composed of
  - sugar molecule
  - phosphate group
  - and a base (A,C,G,T)

- DNA always reads from 5’ end to 3’ end for transcription replication

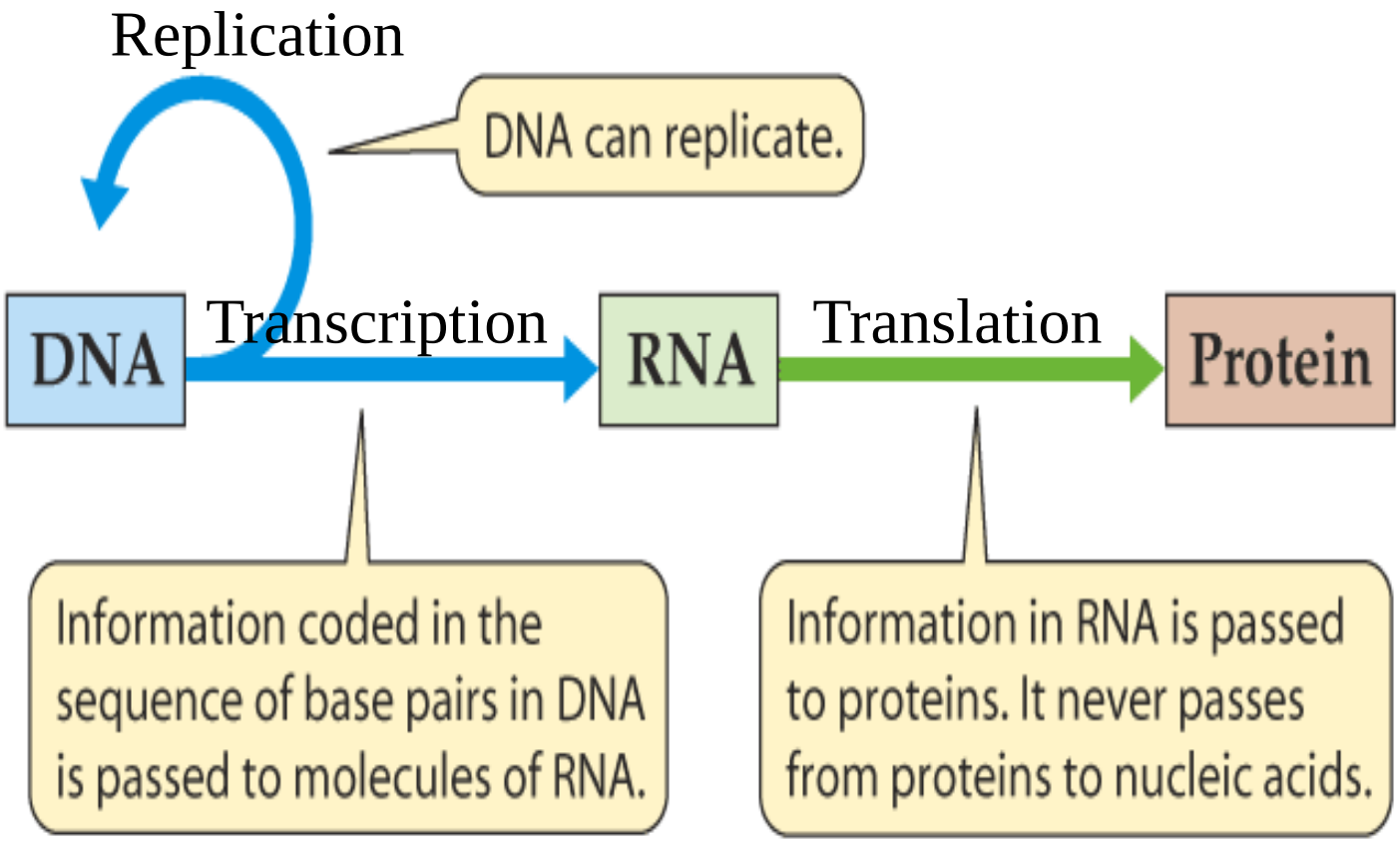
  5’ ATTTAGGCC 3’

  3’ TAAATCCGG 5’

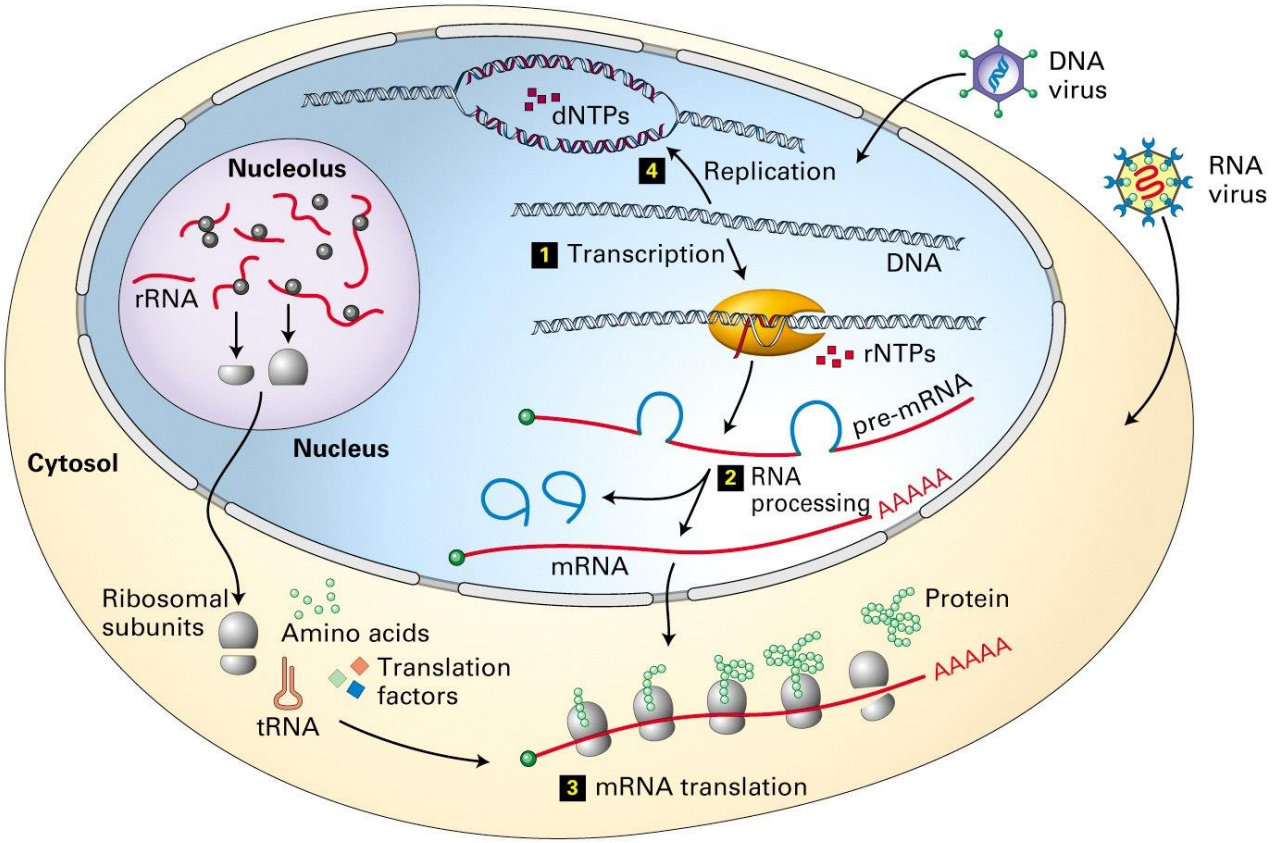
# Processes

- Replication of DNA
- Transcription of gene (DNA) to messenger RNA (mRNA)
- Translation of mRNA into proteins
- Folding of proteins into 3D from
- Biochemical or structural functions of proteins

# DNA, RNA, and the Flow of Information

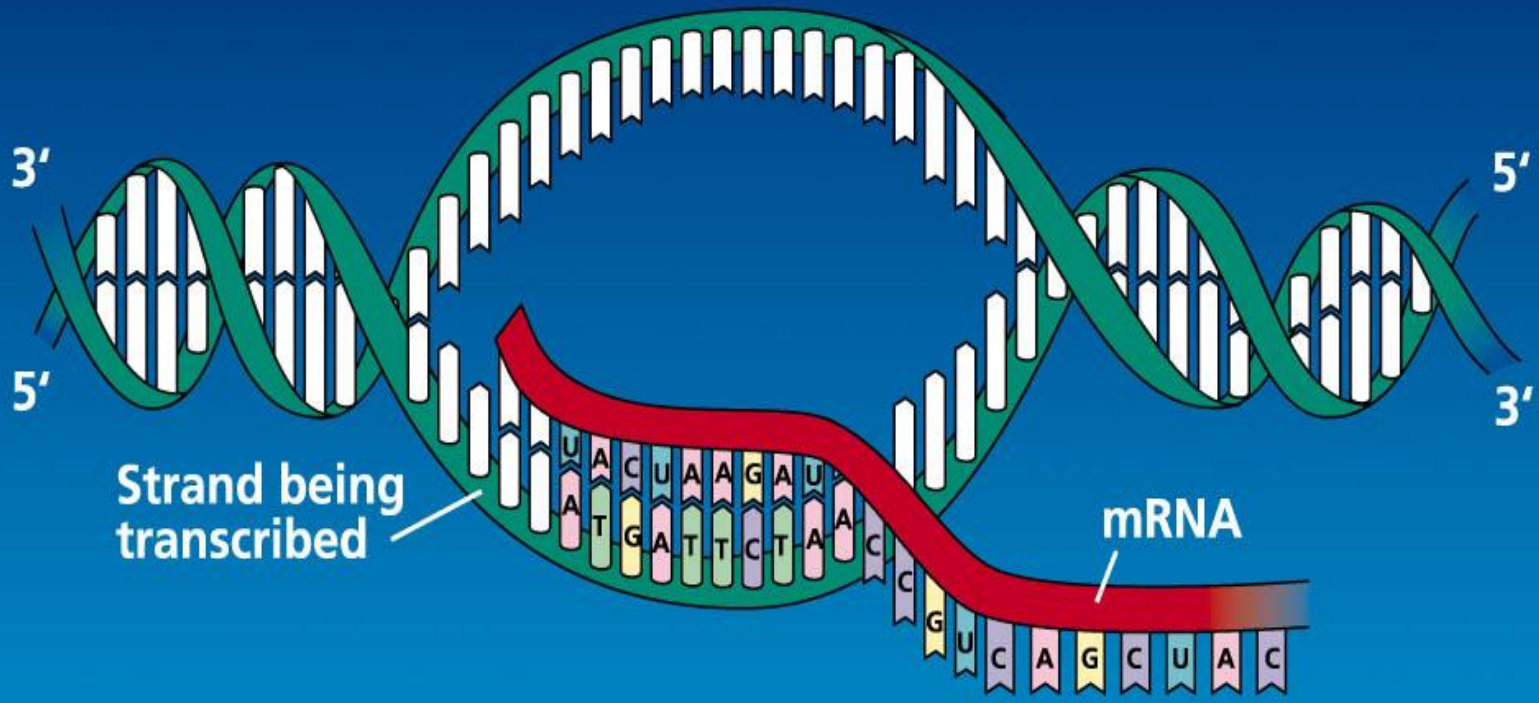


# Overview of DNA to RNA to Protein




- A gene is expressed in two steps
  1) Transcription: RNA synthesis
  2) Translation: Protein synthesis

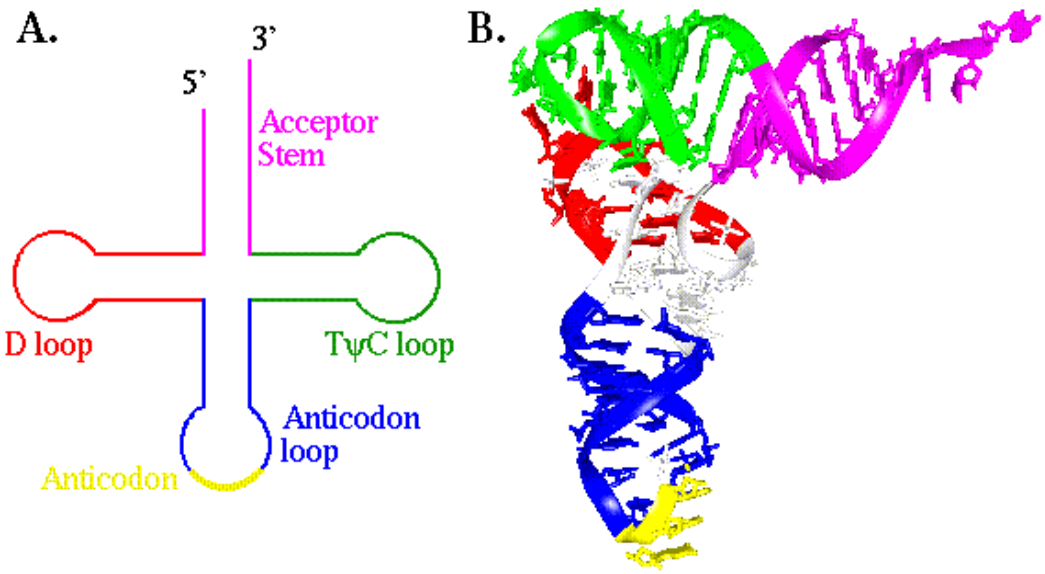
# Transcription



# RNA

- RNA is similar to DNA chemically. It is usually only a single strand. T(hyamine) is replaced by U(racil)
- Some forms of RNA can form 3D structures by "pairing up" with itself.

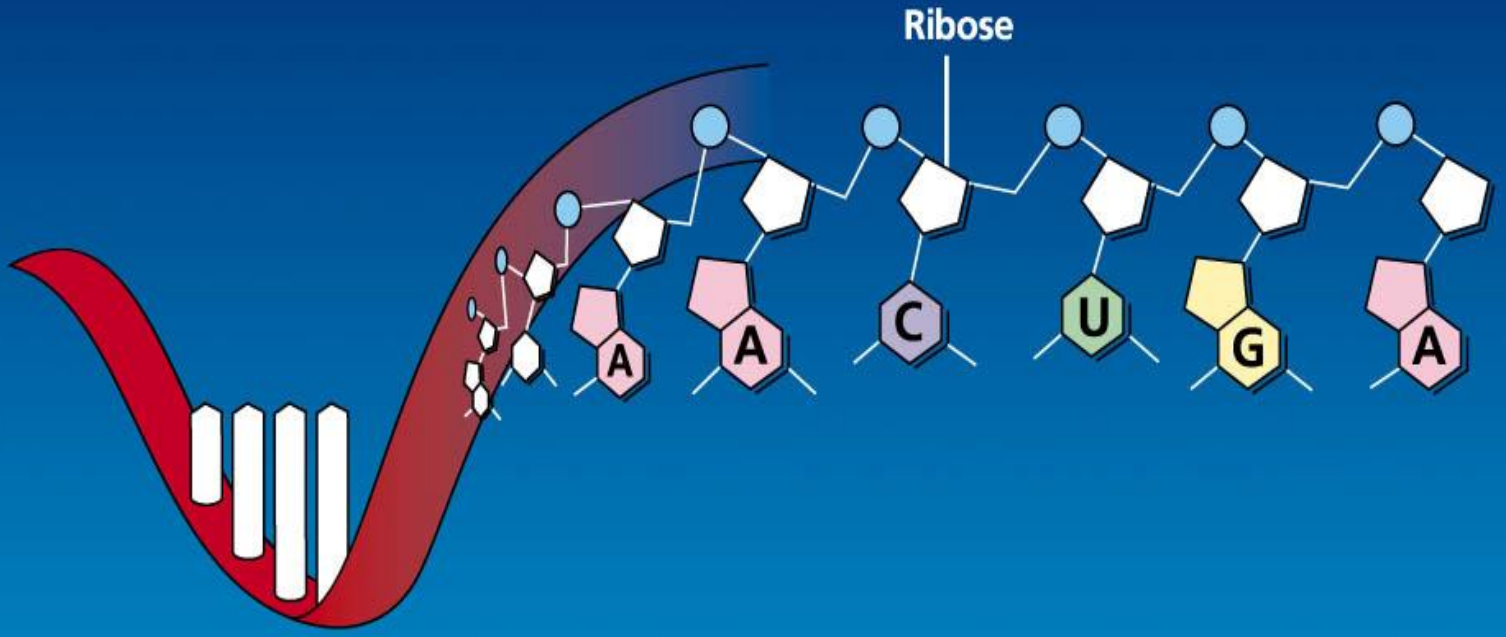
DNA and RNA can pair with each other.




tRNA linear and 3D view:

http://www.cgl.ucsf.edu/home/glasfeld/tutorial/trna/trna.gif

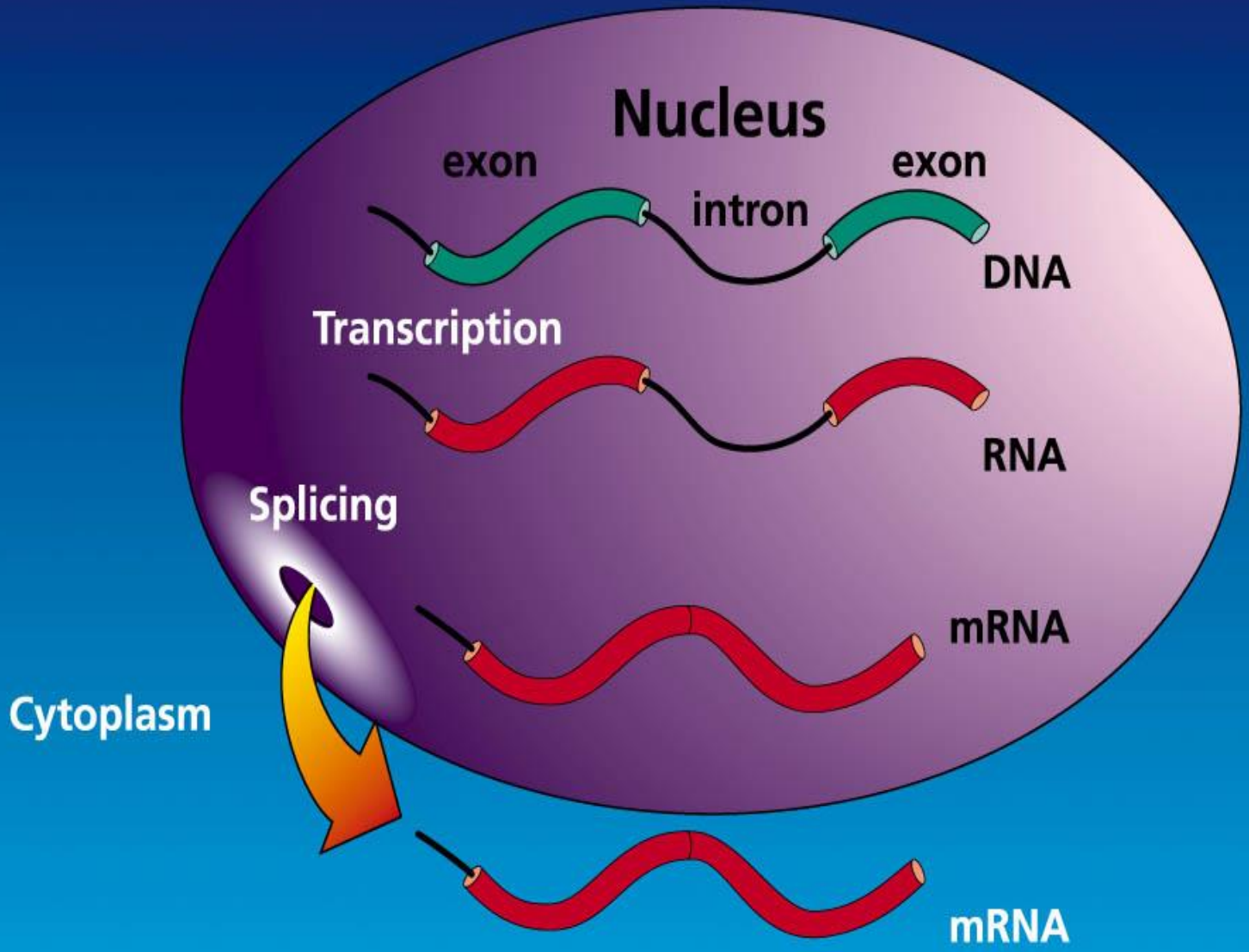
# RNA



# RNA, continued

- Several types exist, classified by function
- mRNA – this is what is usually being referred to when a Bioinformatician says “RNA”. This is used to carry a gene’s *m*essage out of the nucleus.
- tRNA – *t*ransfers genetic information from mRNA to an amino acid sequence
- rRNA – *r*ibosomal RNA. Part of the ribosome which is involved in translation.

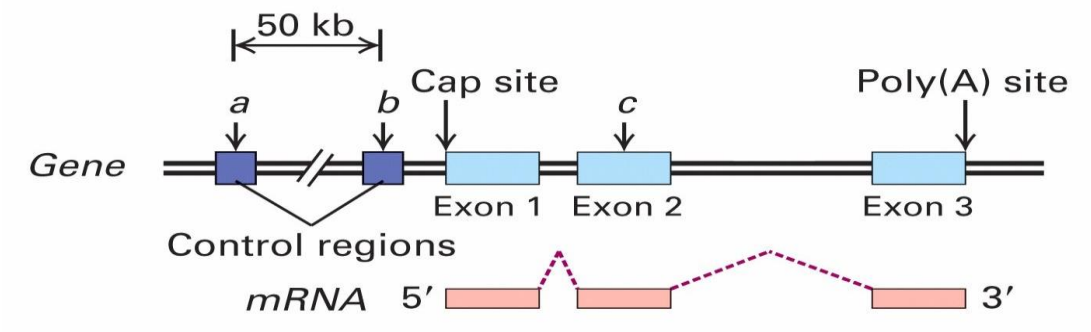
# Introns and exons



# Terminology for Splicing

- Exon: A portion of the gene that appears in both the primary and the mature mRNA transcripts.
- Intron: A portion of the gene that is transcribed but excised prior to translation.
- Lariat structure: The structure that an intron in mRNA takes during excision/splicing.
- Spliceosome: A organelle that carries out the splicing reactions whereby the pre-mRNA is converted to a mature mRNA.

# Definition of a Gene



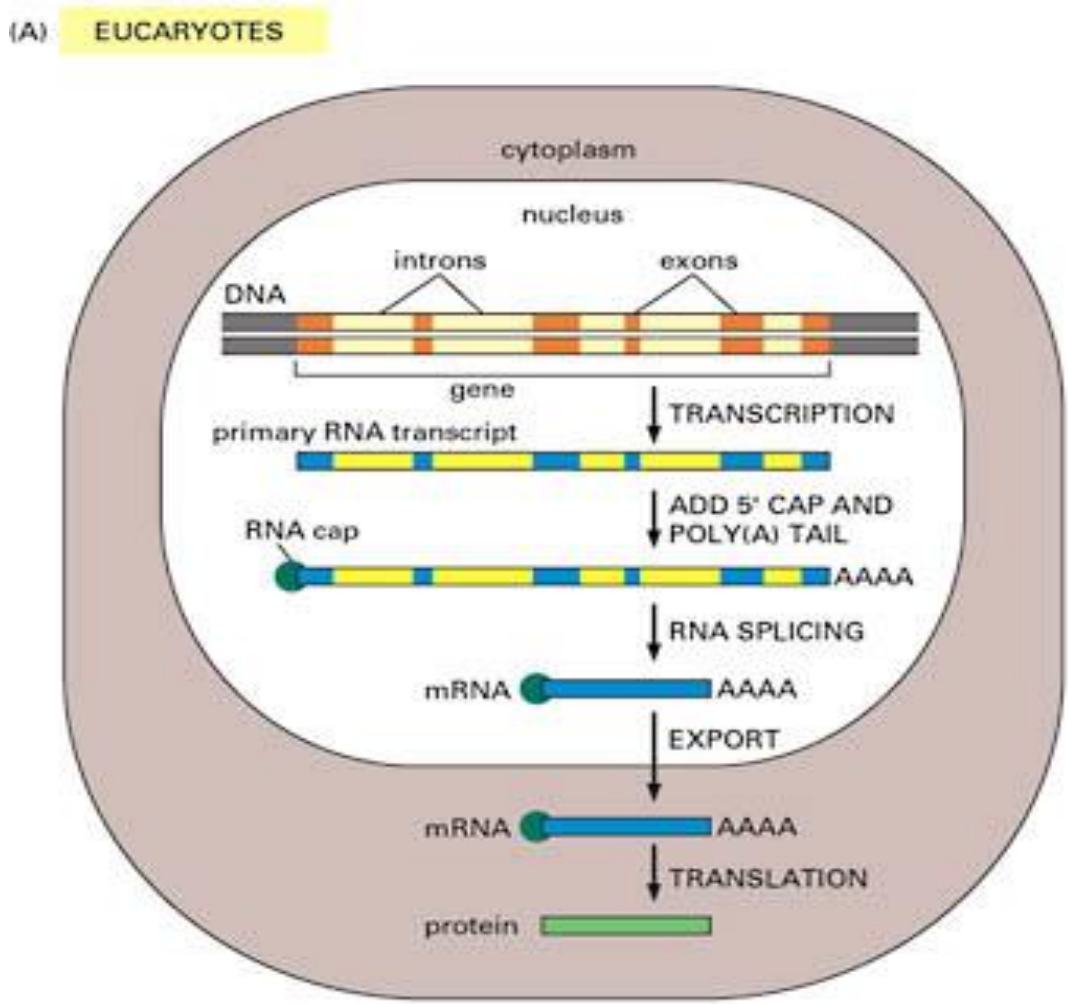

- Regulatory regions: up to 50 kb upstream of +1 site

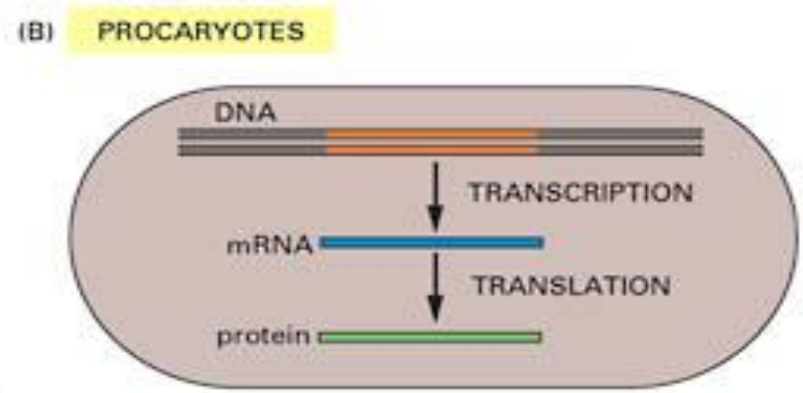
- Exons: protein coding and untranslated regions (UTR)
  1 to 178 exons per gene (mean 8.8)
  8 bp to 17 kb per exon (mean 145 bp)

- Introns: splice acceptor and donor sites, junk DNA
  average 1 kb – 50 kb per intron

- Gene size: Largest – 2.4 Mb (Dystrophin). Mean – 27 kb.

# Splicing


(A) EUCARYOTES
cytoplasm
nucleus
introns
exons
DNA
gene
TRANSCRIPTION
primary RNA transcript
ADD 5' CAP AND POLY(A) TAIL
RNA cap
AAAA
RNA SPLICING
mRNA
AAAA
EXPORT
mRNA
AAAA
TRANSLATION
protein

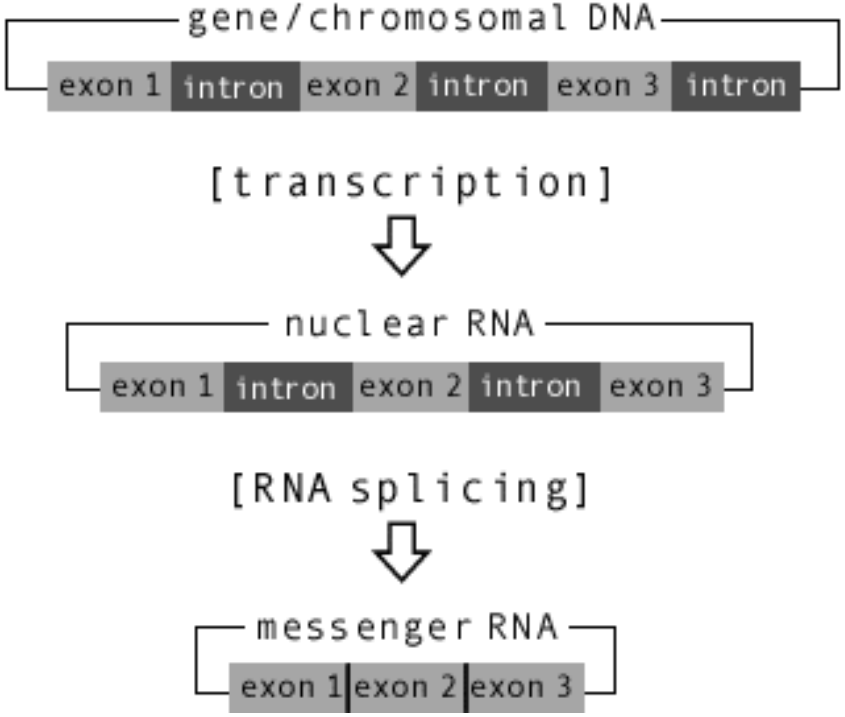

(B) PROCARYOTES
DNA
TRANSCRIPTION
mRNA
TRANSLATION
protein

# Splicing and other RNA processing

- In Eukaryotic cells, RNA is processed between transcription and translation.
- This complicates the relationship between a DNA gene and the protein it codes for.
- Sometimes alternate RNA processing can lead to an alternate protein as a result. This is true in the immune system.

# Splicing (Eukaryotes)

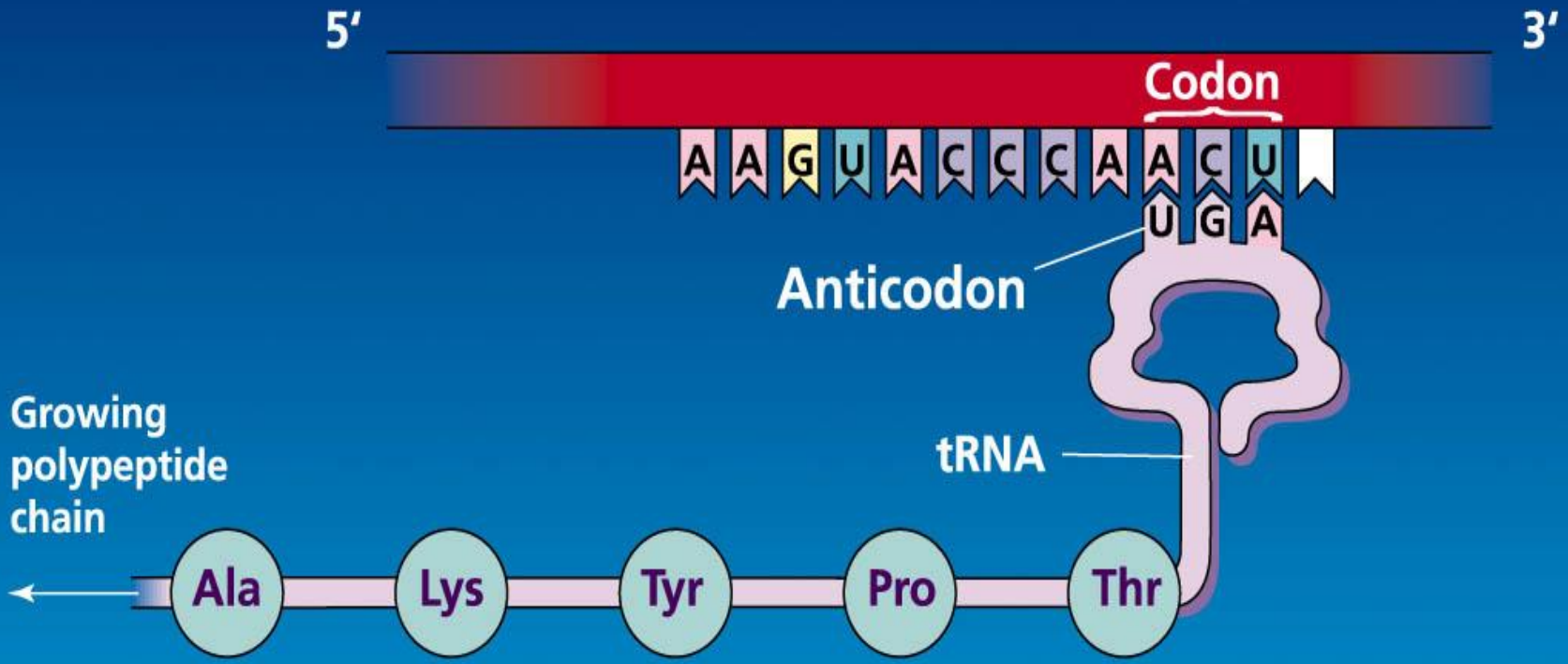
- Unprocessed RNA is composed of Introns and Extrons. Introns are removed before the rest is expressed and converted to protein.
- Sometimes alternate splicings can create different valid proteins.
- A typical Eukaryotic gene has 4-20 introns. Locating them by analytical means is not easy.



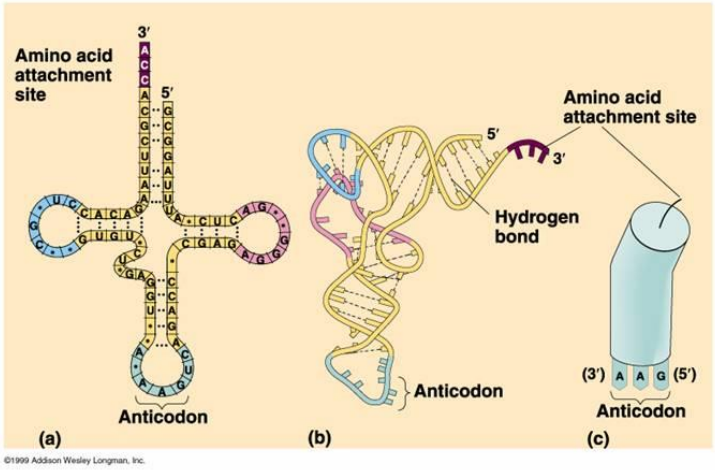
# Translation: Proteins

- Codon: The sequence of 3 nucleotides in DNA/RNA that encodes for a specific amino acid.
- mRNA (messenger RNA): A ribonucleic acid whose sequence is complementary to that of a protein-coding gene in DNA.
- Ribosome: The organelle that synthesizes polypeptides under the direction of mRNA
- rRNA (ribosomal RNA):The RNA molecules that constitute the bulk of the ribosome and provides structural scaffolding for the ribosome and catalyzes peptide bond formation.
- tRNA (transfer RNA): The small L-shaped RNAs that deliver specific amino acids to ribosomes according to the sequence of a bound mRNA.

# Translation



# Purpose of tRNA




- The proper tRNA is chosen by having the corresponding anticodon for the mRNA's codon.
- The tRNA then transfers its aminoacyl group to the growing peptide chain.
- For example, the tRNA with the anticodon UAC corresponds with the codon AUG and attaches methionine amino acid onto the peptide chain.

# Translation: Universal Genetic Code

- **Translation form nucleotide code to amino acid code**.

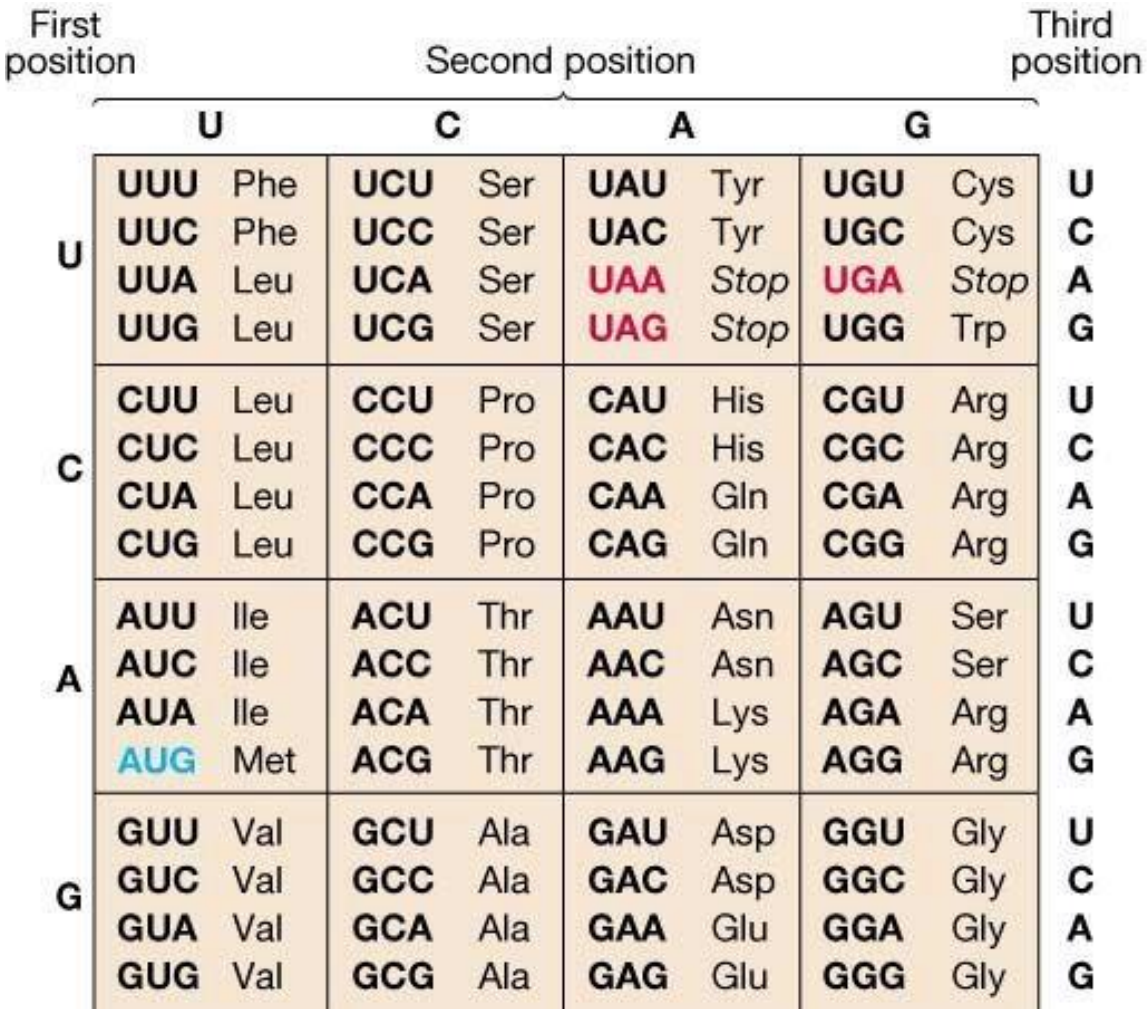
```
atgaatcgta ggggtttgaa cgctggcaat
acgatgactt ctcaagcgaa cattgacgac
ggcagctgga aggcggtctc cgagggcgga ....
```

```
MNRRGLNAGNTMTSQANIDDGSWKAVSEGG …
```

# Uncovering the code

- Scientists conjectured that proteins came from DNA; but how did DNA code for proteins?
- If one nucleotide codes for one amino acid, then there would be $4^1$ amino acids
- However, there are 20 amino acids, so at least 3 bases codes for one amino acid, since $4^2 = 16$ and $4^3 = 64$
  - This triplet of bases is called a “codon”
  - 64 different codons and only 20 amino acids means that the coding is degenerate: more than one codon sequence code for the same amino acid

# Genetic Code



| First position | Second position U | | Second position C | | Second position A | | Second position G | | Third position |
|---|---|---|---|---|---|---|---|---|---|
| U | UUU | Phe | UCU | Ser | UAU | Tyr | UGU | Cys | U |
| | UUC | Phe | UCC | Ser | UAC | Tyr | UGC | Cys | C |
| | UUA | Leu | UCA | Ser | UAA | *Stop* | UGA | *Stop* | A |
| | UUG | Leu | UCG | Ser | UAG | *Stop* | UGG | Trp | G |
| C | CUU | Leu | CCU | Pro | CAU | His | CGU | Arg | U |
| | CUC | Leu | CCC | Pro | CAC | His | CGC | Arg | C |
| | CUA | Leu | CCA | Pro | CAA | Gln | CGA | Arg | A |
| | CUG | Leu | CCG | Pro | CAG | Gln | CGG | Arg | G |
| A | AUU | Ile | ACU | Thr | AAU | Asn | AGU | Ser | U |
| | AUC | Ile | ACC | Thr | AAC | Asn | AGC | Ser | C |
| | AUA | Ile | ACA | Thr | AAA | Lys | AGA | Arg | A |
| | AUG | Met | ACG | Thr | AAG | Lys | AGG | Arg | G |
| G | GUU | Val | GCU | Ala | GAU | Asp | GGU | Gly | U |
| | GUC | Val | GCC | Ala | GAC | Asp | GGC | Gly | C |
| | GUA | Val | GCA | Ala | GAA | Glu | GGA | Gly | A |
| | GUG | Val | GCG | Ala | GAG | Glu | GGG | Gly | G |

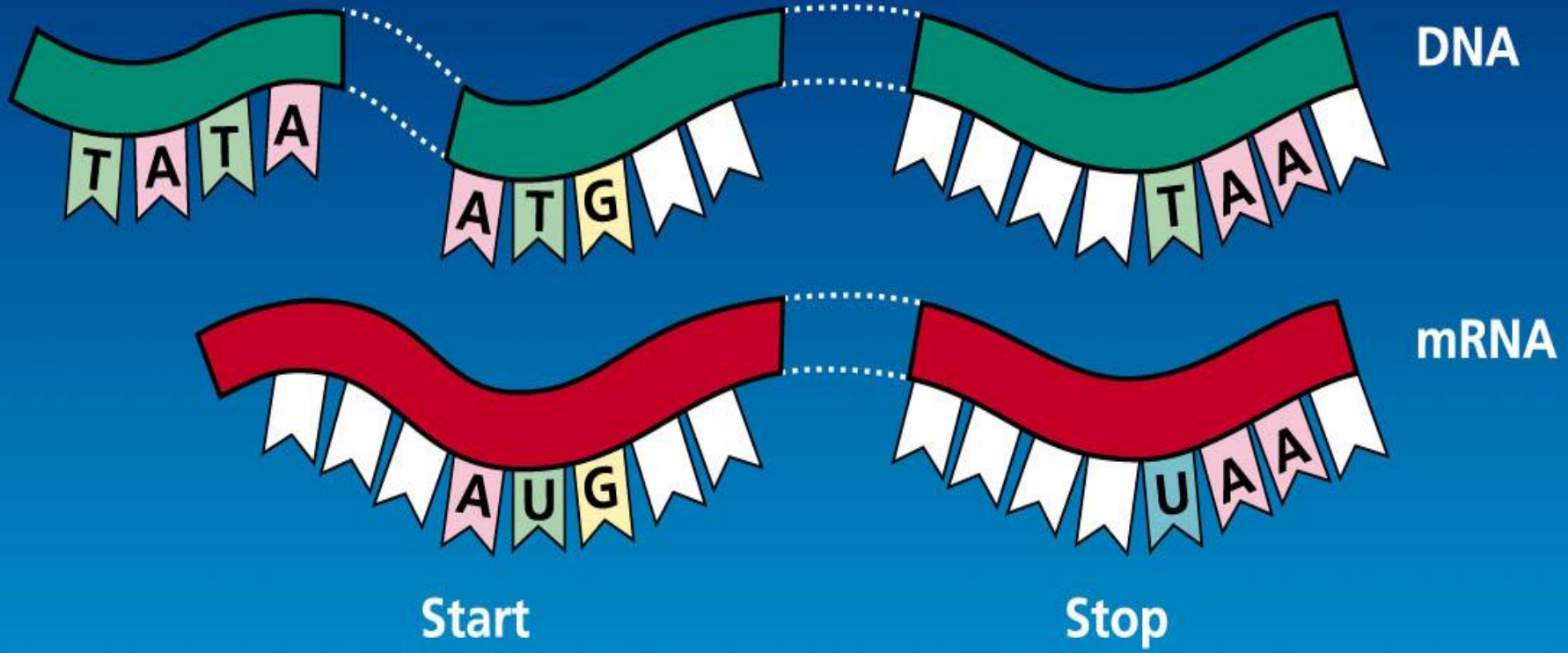
# Translation

- The process of going from RNA to polypeptide.
- Three base pairs of RNA (called a codon) correspond to one amino acid based on a fixed table.
- Always starts with Methionine and ends with a stop codon

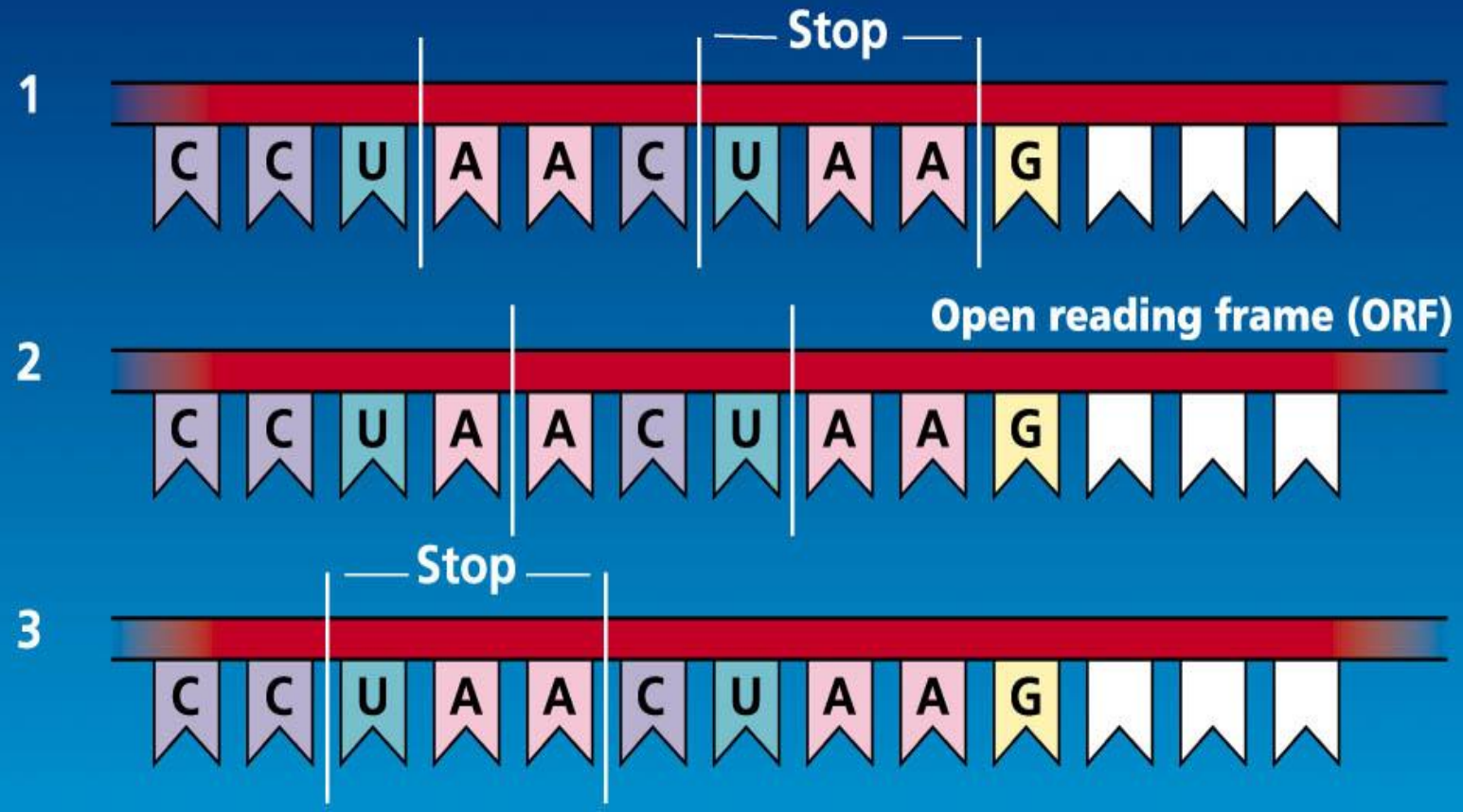
| FIRST POSITION | SECOND POSITION: U | SECOND POSITION: C | SECOND POSITION: A | SECOND POSITION: G | THIRD POSITION |
|---|---|---|---|---|---|
| U | phenyl-alanine | serine | tyrosine | cysteine | U |
| | | | | | C |
| | leucine | | stop | stop | A |
| | | | stop | tryptophan | G |
| C | leucine | proline | histidine | arginine | U |
| | | | | | C |
| | | | glutamine | | A |
| | | | | | G |
| A | isoleucine | threonine | asparagine | serine | U |
| | | | | | C |
| | * methionine | | lysine | arginine | A |
| | | | | | G |
| G | valine | alanine | aspartic acid | glycine | U |
| | | | | | C |
| | | | glutamic acid | | A |
| | | | | | G |

* and start

# Control of translation



# Open reading frames



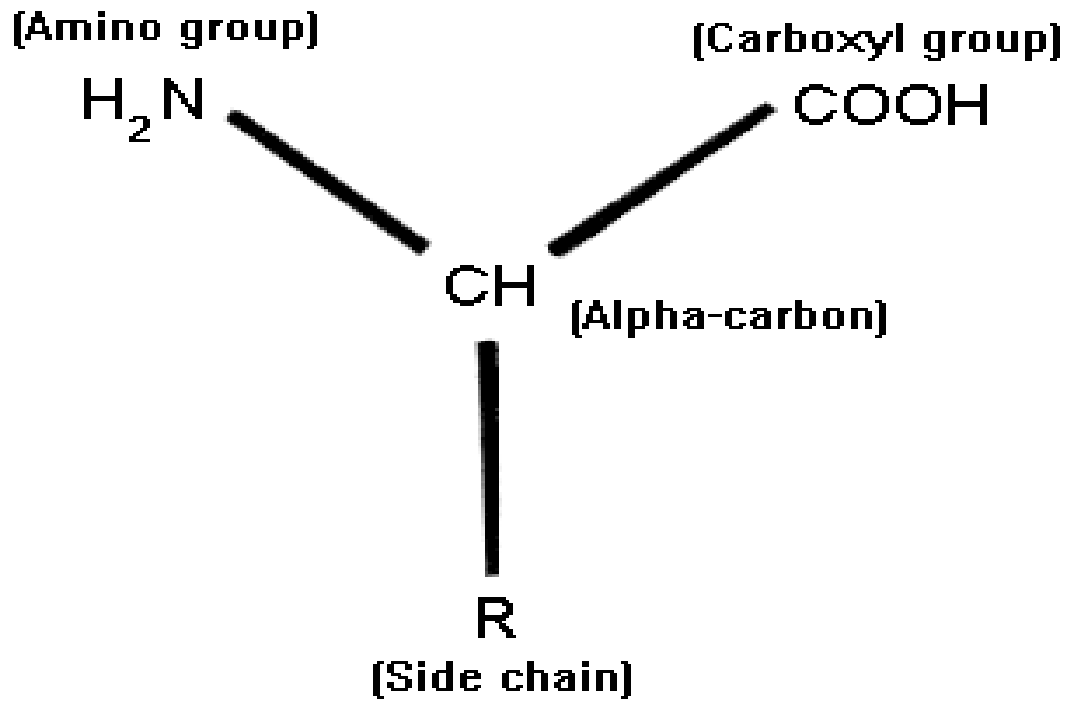
# Proteins: Workhorses of the Cell

- 20 different **amino acids**
  - different chemical properties cause the protein chains to fold up into specific three-dimensional structures that define their particular functions in the cell.
- Proteins do all essential work for the cell
  - build cellular structures
  - digest nutrients
  - execute metabolic functions
  - Mediate information flow within a cell and among cellular communities.
- Proteins work together with other proteins or nucleic acids as "molecular machines"
  - structures that fit together and function in highly specific, lock-and-key ways.

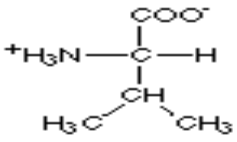
# Sequence of Amino Acids: Protein

- Unbranched polymer
- Peptide backbone
- Twenty side chain types
- 3D structure the key

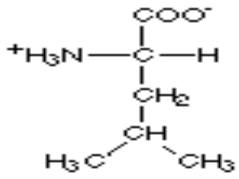
# Amino Acid


(Amino group)
$H_2N$
(Carboxyl group)
COOH
CH
(Alpha-carbon)
R
(Side chain)

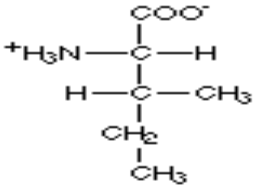
Amino acids with hydrophobic side groups



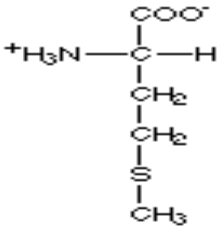
Valine
(val)



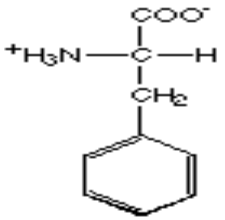
Leucine
(leu)



Isoleucine
(ile)

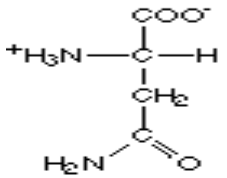


Methionine
(met)

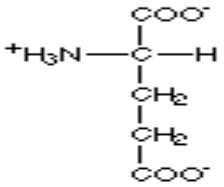


Phenylalanine
(phe)

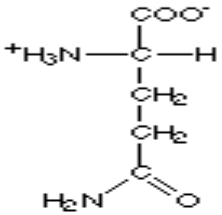
Amino acids with hydrophilic side groups



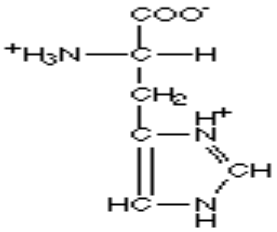
Asparagine
(asn)



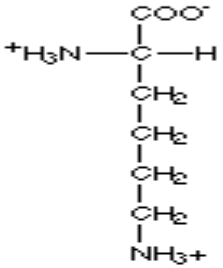
Glutamic acid
(glu)



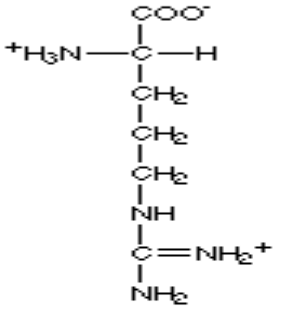
Glutamine
(gln)



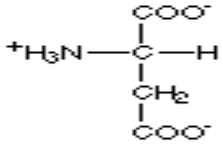
Histidine
(his)



Lysine
(lys)

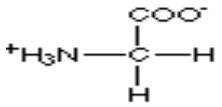


Arginine
(arg)

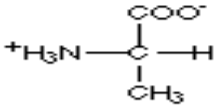


Aspartic acid
(asp)

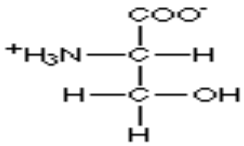
Amino acids that are in between



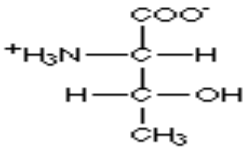
Glycine
(gly)



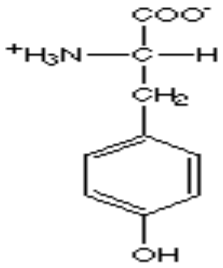
Alanine
(ala)



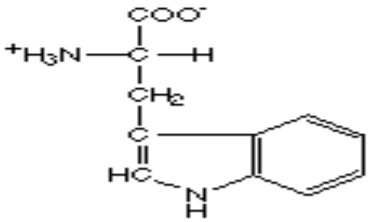
Serine
(ser)



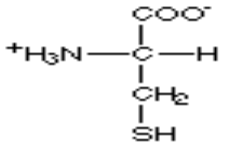
Threonine
(thr)



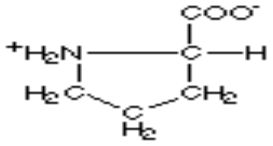
Tyrosine
(tyr)



Tryptophan
(trp)

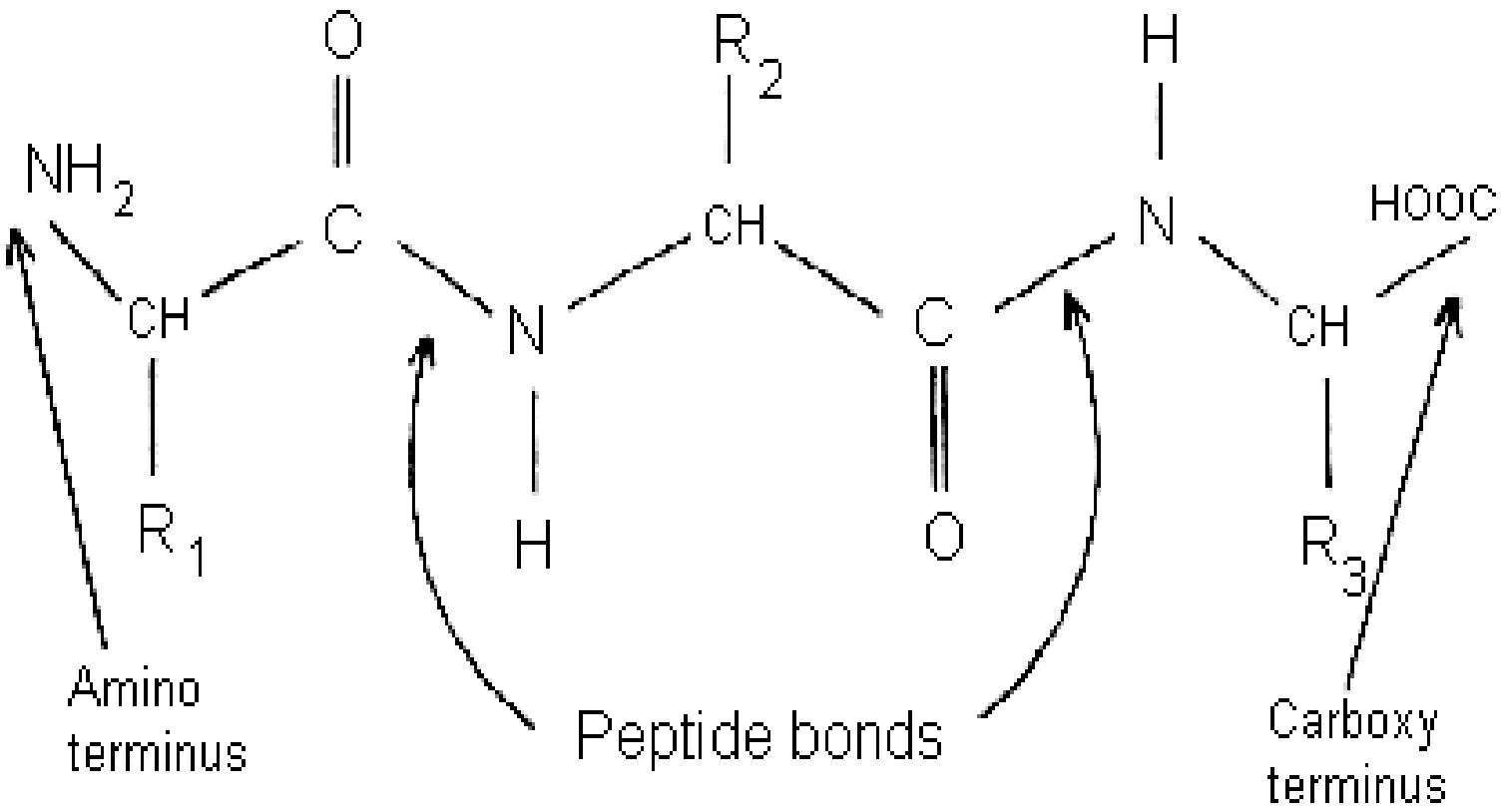


Cysteine
(cys)

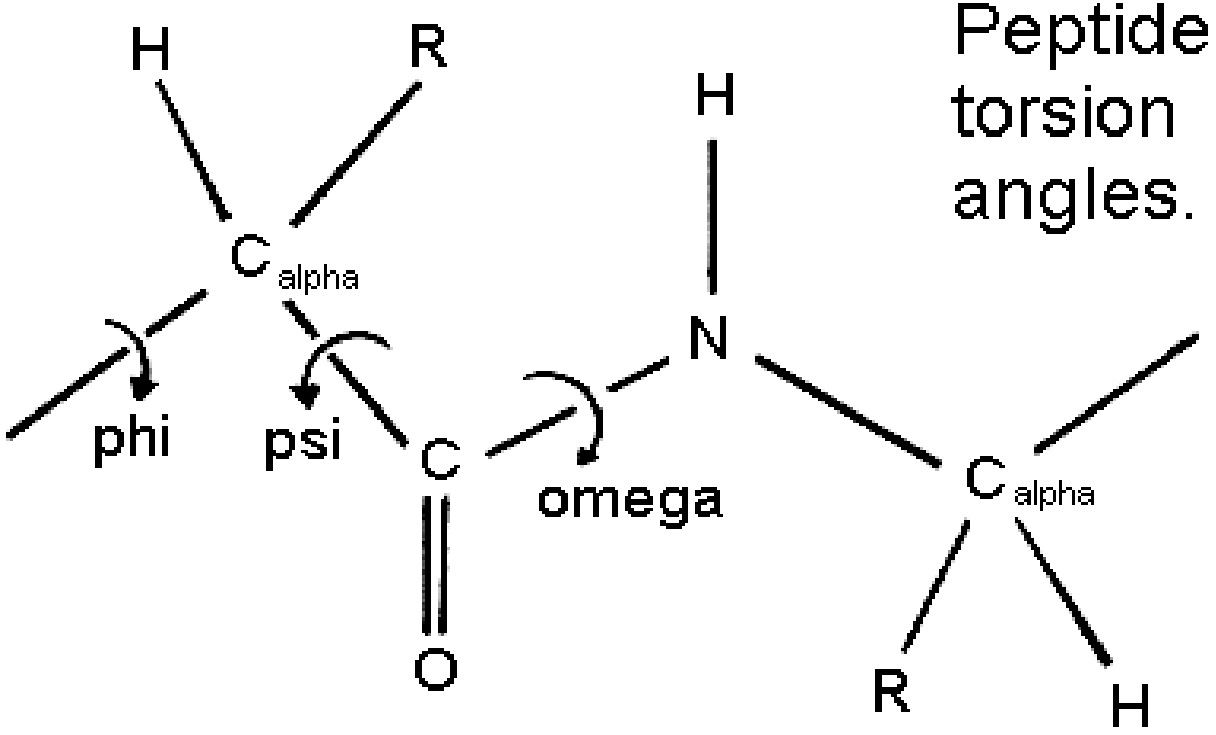


Proline
(pro)

# Polypeptide Chain



# Torsion Angles



# Protein Primary Structure

- Unbranched polymer
- 20 side chains

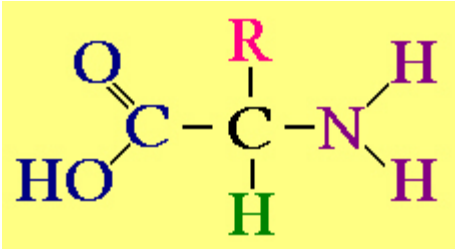
MNRRGLNAGNTMTSQANIDDGSWKAVSEGG …

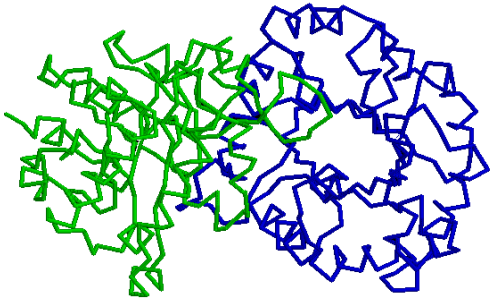
# 3D: Higher Order Protein Structures

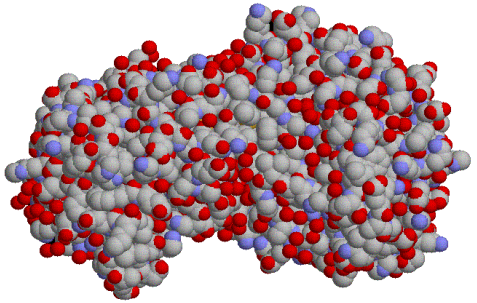
- Higher order structures
  - Secondary: local in sequence
  - Tertiary:  3D fold of one polypeptide chain
  - Quaternary: Chains packing together

# Protein Structure

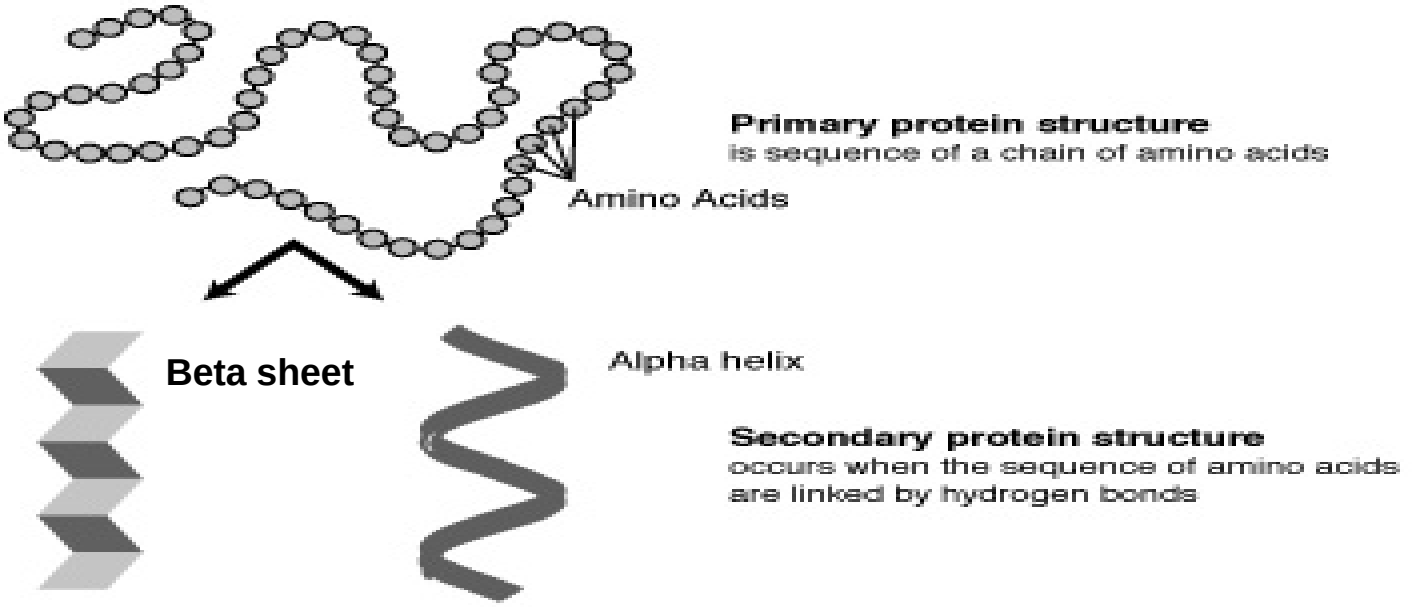
- **Proteins are composed of twenty amino acids**
- **Amino acids *are composed of a central carbon atom attached to four groups of atoms.***
- ***Amino acids (AA)* combine in sequence to form a polymer (*polypeptide chain*).**
- **A chain folds into a specific 3D shape by the laws of nature.**
- **A protein is made up of one or more chains.**





# Protein Structure Hierarchy




***Primary structure:*** **The linear sequence of amino acids comprising a protein:**

**AGVGTVPMTAYGNDIQYYGQVT…**

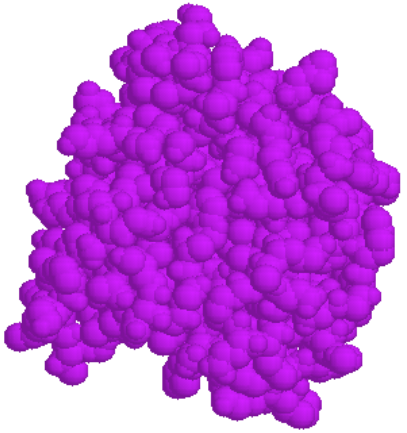
***Secondary structure:*** **local ordered sub-structures**

**two common types: *α-Helix* and *β-Sheet***

# Protein Structure Hierarchy

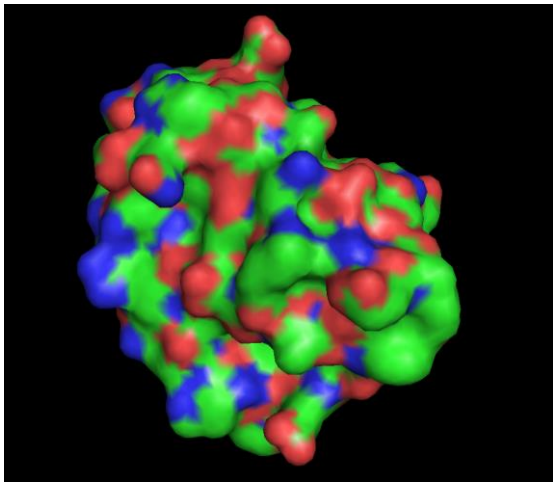
***Tertiary structure:***

**"global" folding of a single chain in 3D form.**

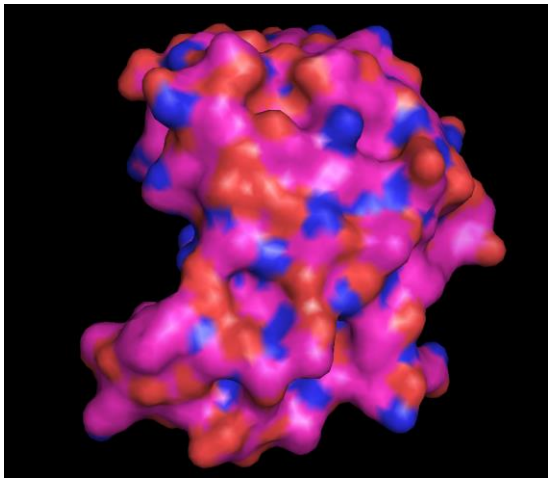


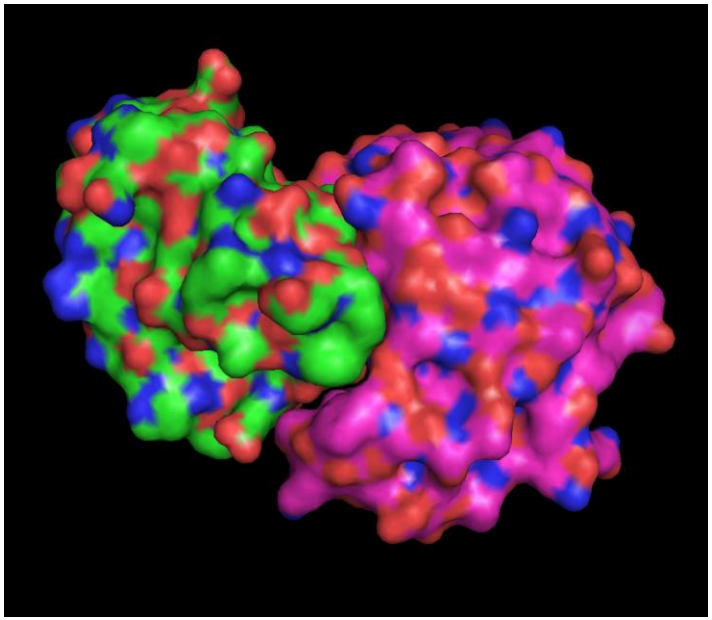
MDL

1BGS-E (BARNASE)



1BGS-A (BARSTAR)





1BGS-AE

# Protein Folding

- Proteins the action molecules of life
- 3D Structure critical to function
- Many protein sequences fold spontaneously to native 3D fold

# Protein Folding

- The structure that a protein adopts is vital to it's chemistry
- Its structure determines which of its amino acids are exposed carry out the protein's function
- Its structure also determines what substrates it can react with