CISC 636 Computational Biology & Bioinformatics
(Fall 2016)

Molecular Biology
A Primer

What is Life
  - Three kingdoms
  - The Cell thoery

Central Dogma
  - Genetic code
  - Transcription
  - Translation

Organisms: three kindoms of life -- eukaryotes, eubacteria, and archea

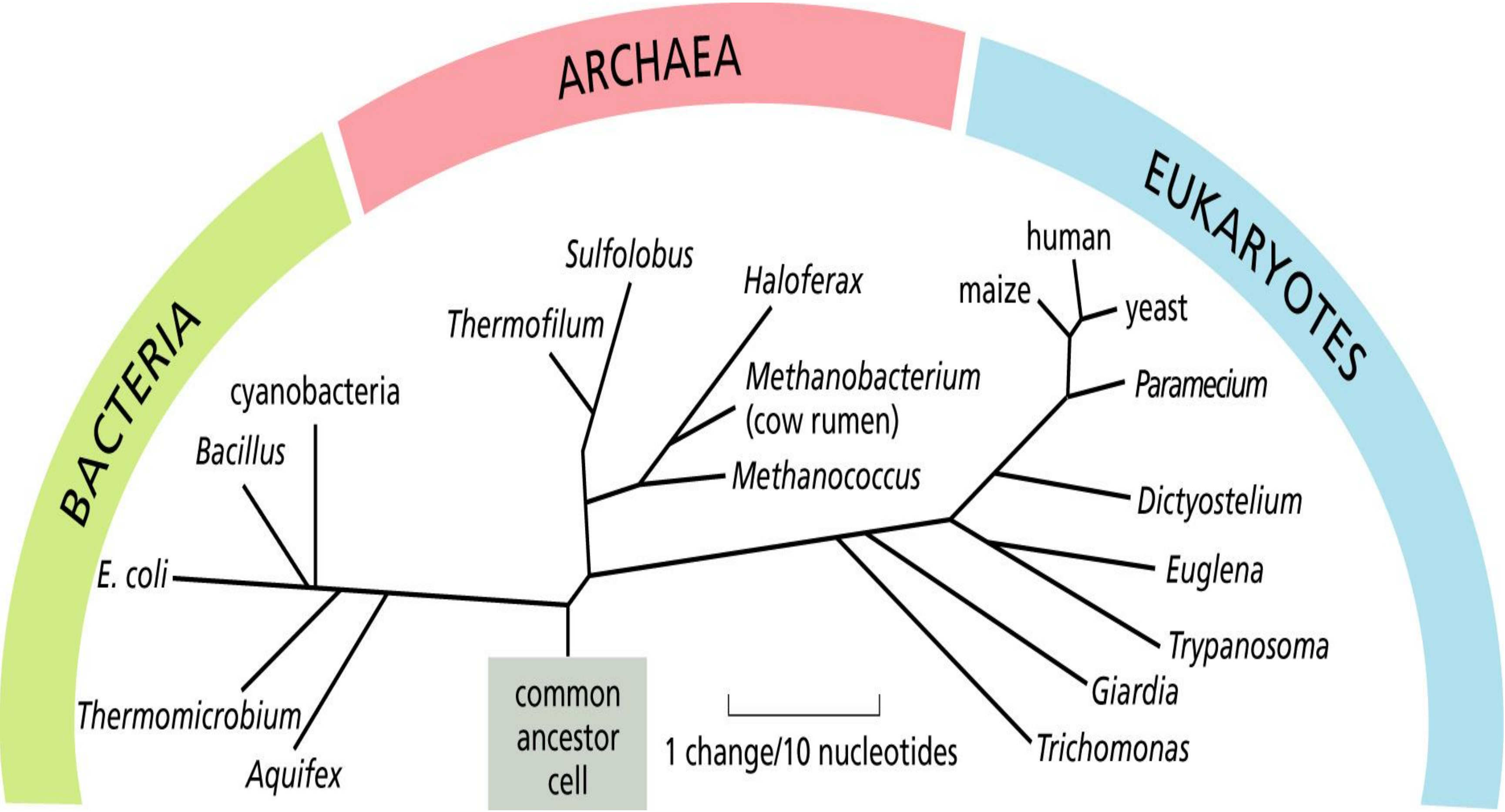
- Observation: a lot of living things
- Why does Mother nature have this biodiversity?
- Answers
  - Simple classification based on morphological features
  - Theory: evolution – mutations, natural selection, …
- Tree of life
  - NCBI Taxonomy http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy

Model organisms: E. coli, Drosophila, C. elegans, Yeast, Arabidopsis, Mouse, …

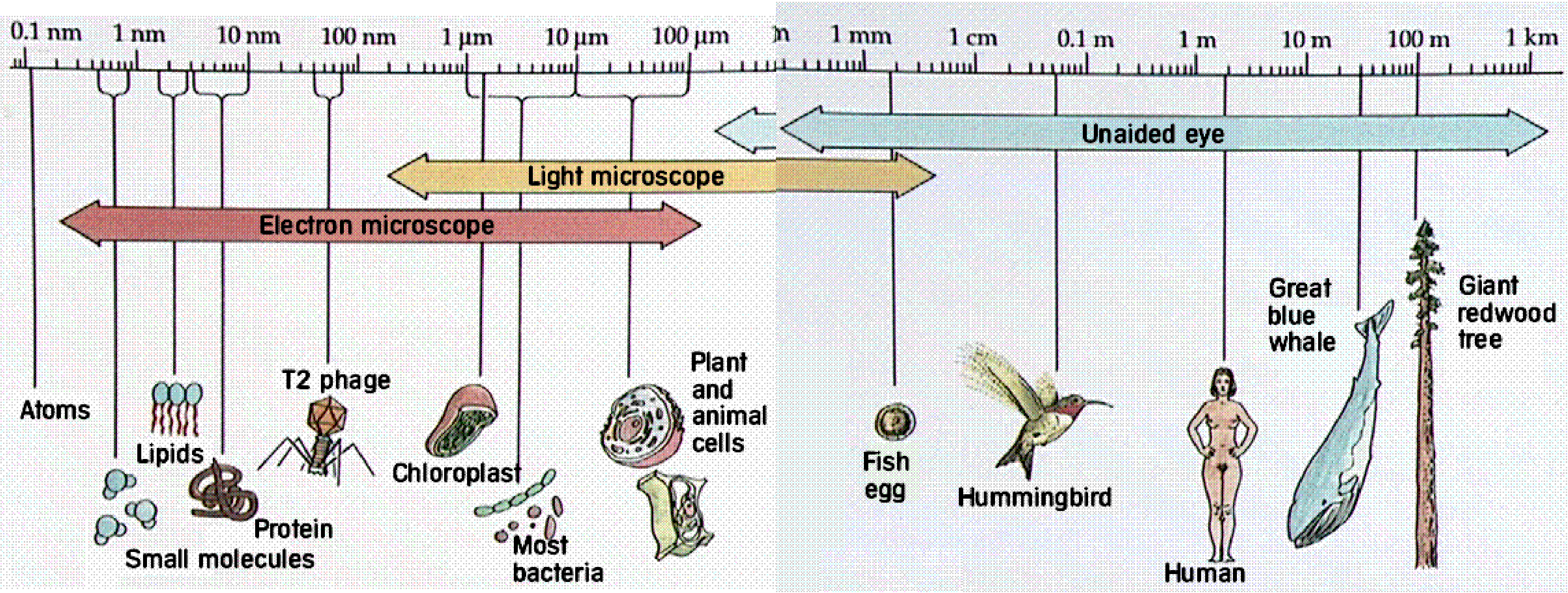
Cell: the basic unit of life

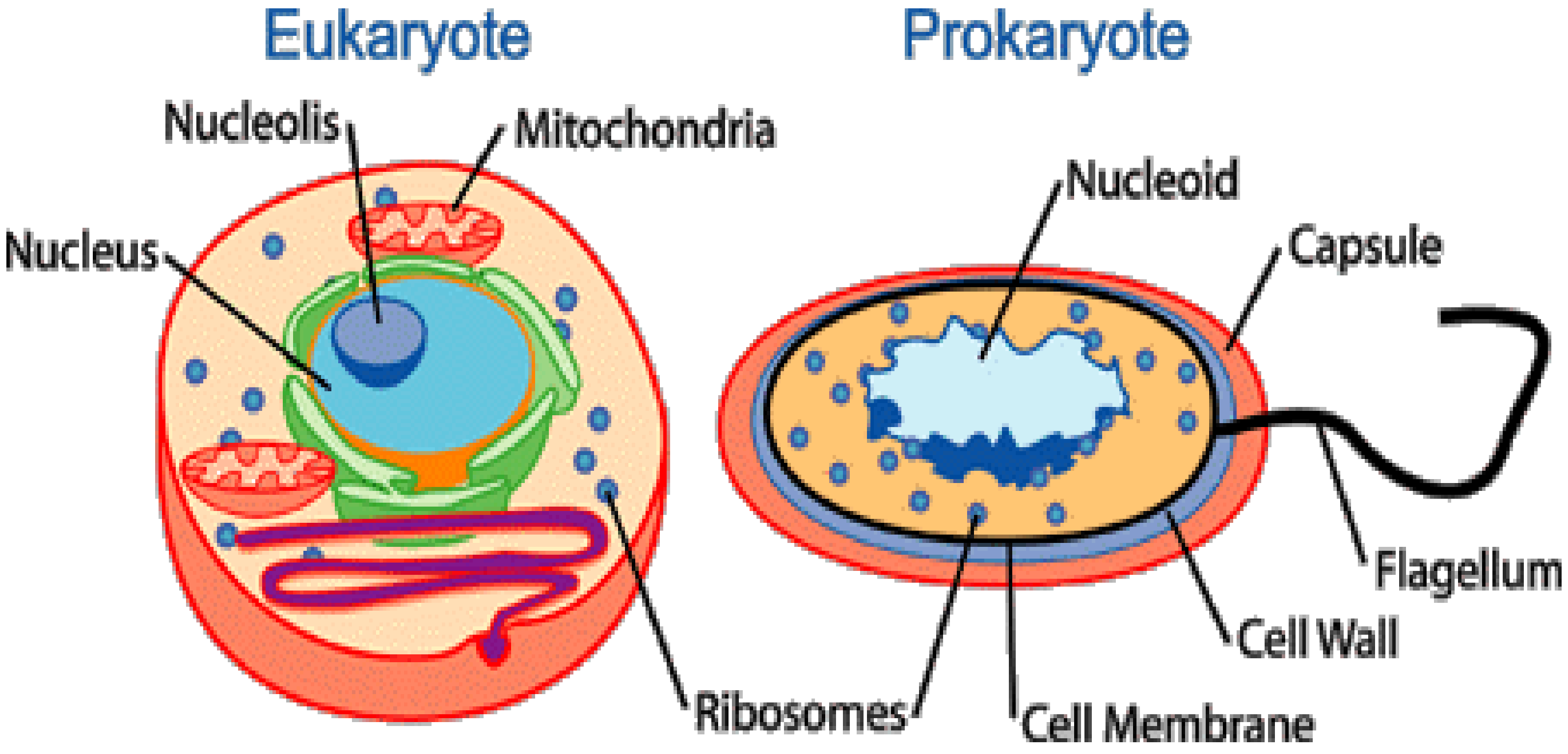
- Every living thing is made of cells.
- Every cell comes from a pre-existing cell.

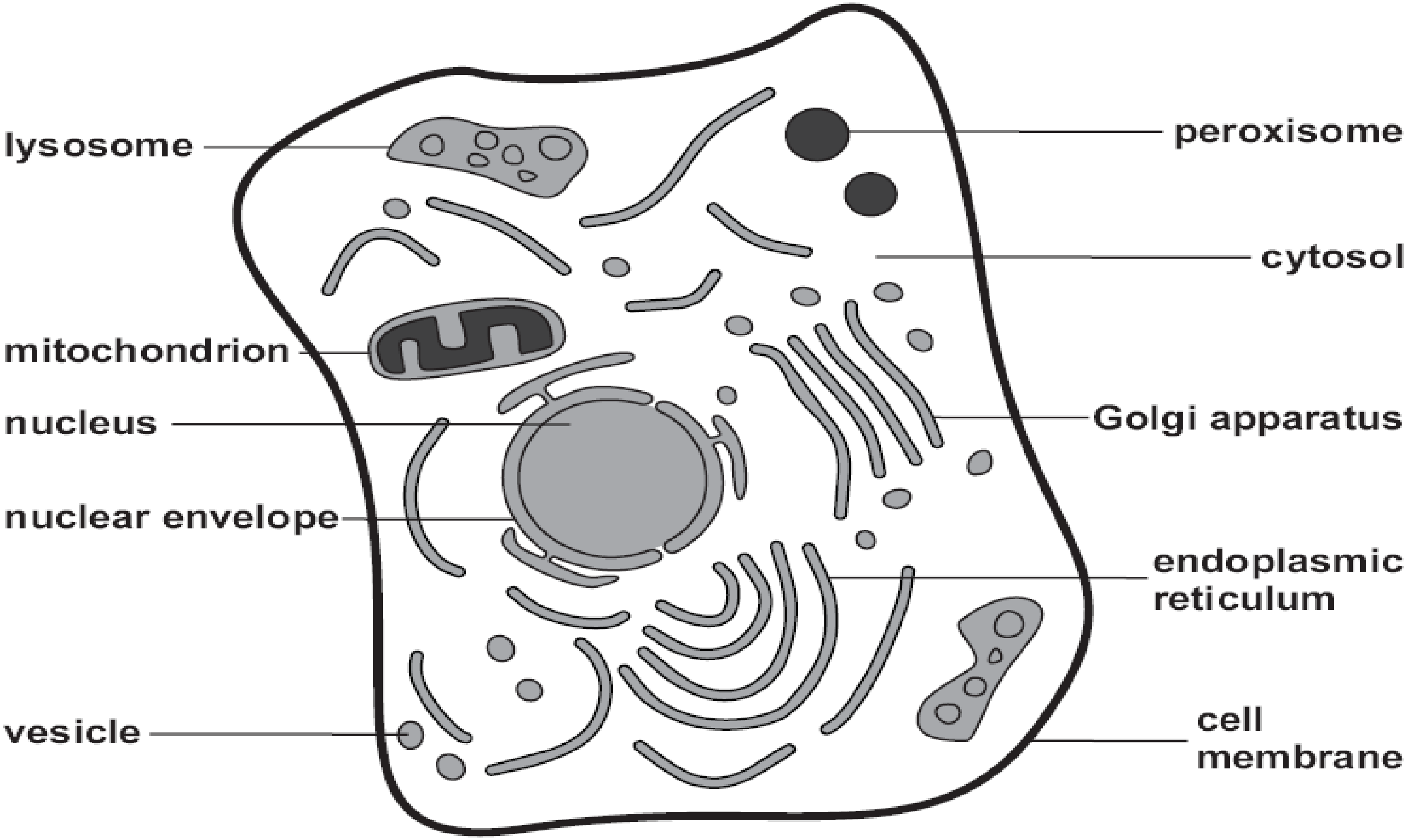
Tree of Life

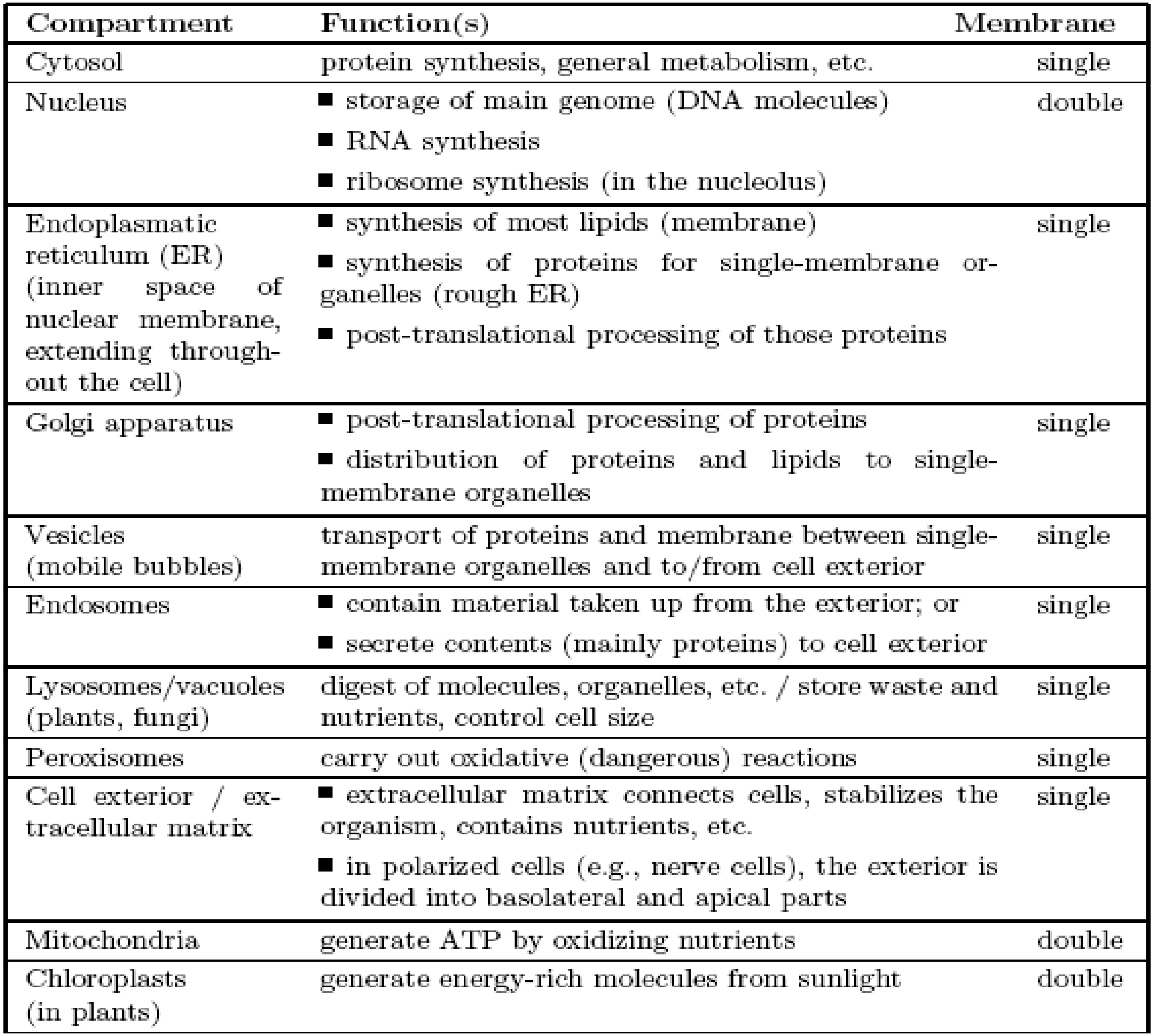

ARCHAEA
BACTERIA
EUKARYOTES
Sulfolobus
Haloferax
Thermofilum
Methanobacterium
(cow rumen)
Methanococcus
human
maize
yeast
Paramecium
Dictyostelium
Euglena
Trypanosoma
Giardia
Trichomonas
cyanobacteria
Bacillus
E. coli
Thermomicrobium
Aquifex
common
ancestor
cell
1 change/10 nucleotides

10^{-9} 10^{-6} 10^{-3}

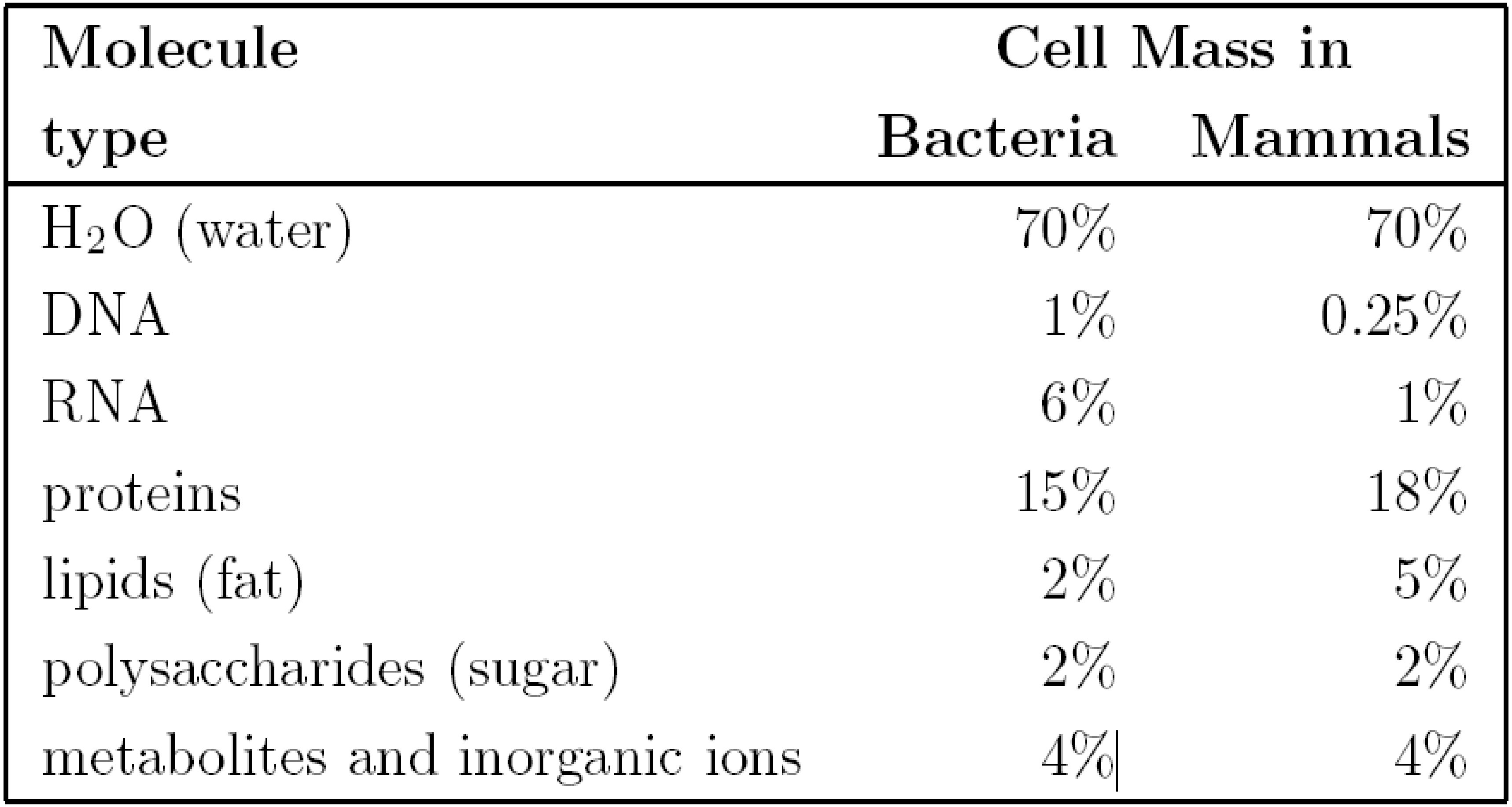

0.1 nm
1 nm
10 nm
100 nm
1 µm
10 µm
100 µm
1 mm
1 cm
0.1 m
1 m
10 m
100 m
1 km
Unaided eye
Light microscope
Electron microscope
Atoms
Lipids
Small molecules
Protein
T2 phage
Chloroplast
Most bacteria
Plant and animal cells
Fish egg
Hummingbird
Human
Great blue whale
Giant redwood tree


Eukaryote
Prokaryote
Nucleolis
Mitochondria
Nucleus
Nucleoid
Capsule
Flagellum
Cell Wall
Ribosomes
Cell Membrane

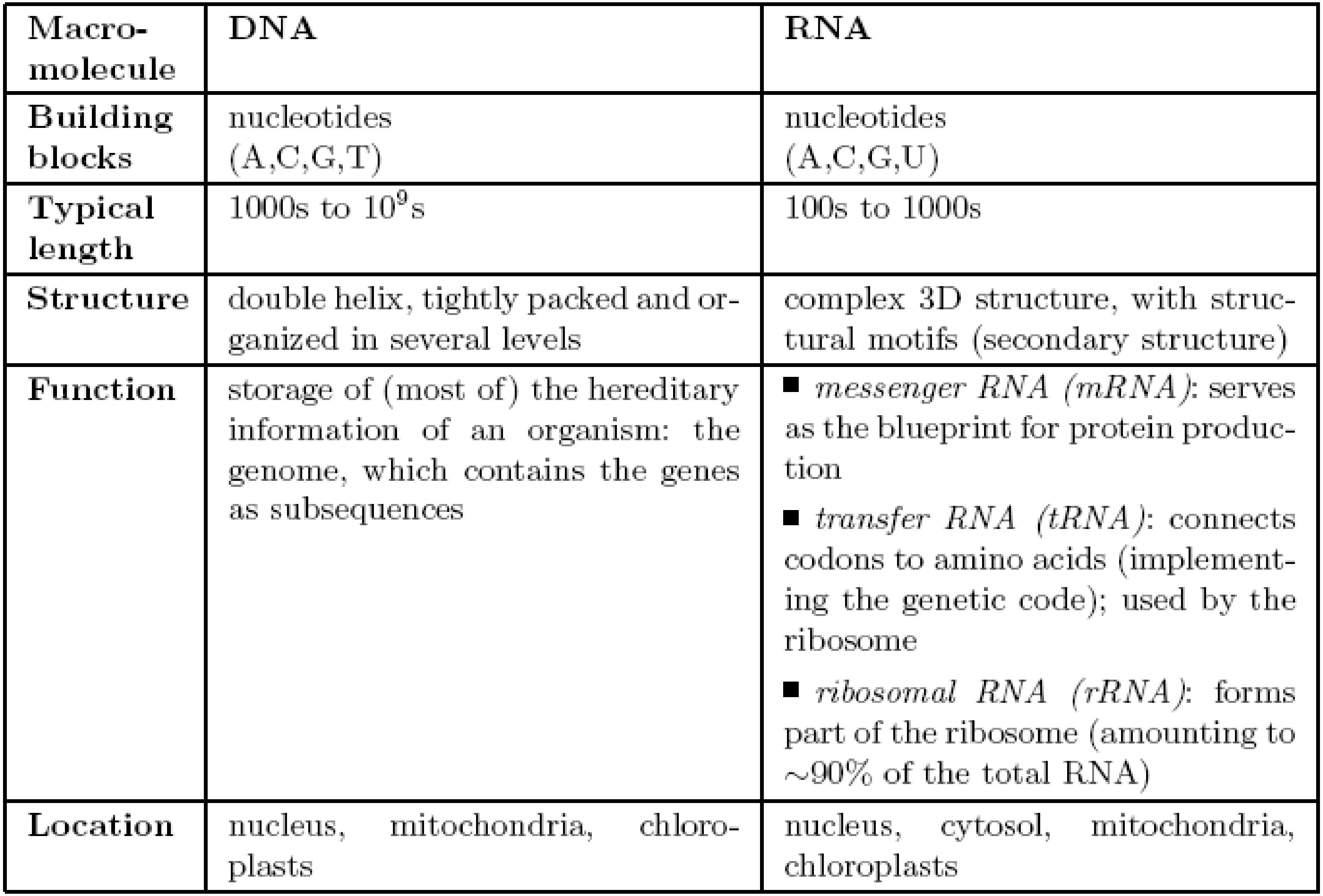

lysosome
peroxisome
cytosol
mitochondrion
nucleus
Golgi apparatus
nuclear envelope
endoplasmic reticulum
vesicle
cell membrane



Compartment	Function(s)	Membrane
Cytosol	protein synthesis, general metabolism, etc.	single
Nucleus	■ storage of main genome (DNA molecules) ■ RNA synthesis ■ ribosome synthesis (in the nucleolus)	double
Endoplasmatic reticulum (ER) (inner space of nuclear membrane, extending throughout the cell)	■ synthesis of most lipids (membrane) ■ synthesis of proteins for single-membrane organelles (rough ER) ■ post-translational processing of those proteins	single
Golgi apparatus	■ post-translational processing of proteins ■ distribution of proteins and lipids to single-membrane organelles	single
Vesicles (mobile bubbles)	transport of proteins and membrane between single-membrane organelles and to/from cell exterior	single
Endosomes	■ contain material taken up from the exterior; or ■ secrete contents (mainly proteins) to cell exterior	single
Lysosomes/vacuoles (plants, fungi)	digest of molecules, organelles, etc. / store waste and nutrients, control cell size	single
Peroxisomes	carry out oxidative (dangerous) reactions	single
Cell exterior / extracellular matrix	■ extracellular matrix connects cells, stabilizes the organism, contains nutrients, etc. ■ in polarized cells (e.g., nerve cells), the exterior is divided into basolateral and apical parts	single
Mitochondria	generate ATP by oxidizing nutrients	double
Chloroplasts (in plants)	generate energy-rich molecules from sunlight	double



Molecule type	Cell Mass in	
	Bacteria	Mammals
H_2O (water)	70%	70%
DNA	1%	0.25%
RNA	6%	1%
proteins	15%	18%
lipids (fat)	2%	5%
polysaccharides (sugar)	2%	2%
metabolites and inorganic ions	4%	4%



Macro-molecule	**DNA**	**RNA**
Building blocks	nucleotides (A,C,G,T)	nucleotides (A,C,G,U)
Typical length	1000s to 10^9s	100s to 1000s
Structure	double helix, tightly packed and organized in several levels	complex 3D structure, with structural motifs (secondary structure)
Function	storage of (most of) the hereditary information of an organism: the genome, which contains the genes as subsequences	■ *messenger RNA (mRNA)*: serves as the blueprint for protein production ■ *transfer RNA (tRNA)*: connects codons to amino acids (implementing the genetic code); used by the ribosome ■ *ribosomal RNA (rRNA)*: forms part of the ribosome (amounting to ~90% of the total RNA)
Location	nucleus, mitochondria, chloroplasts	nucleus, cytosol, mitochondria, chloroplasts

Chromosome (DNA)

- > circular, also called plasmid when small (for bacteria)
- > linear (for eukaryotes)

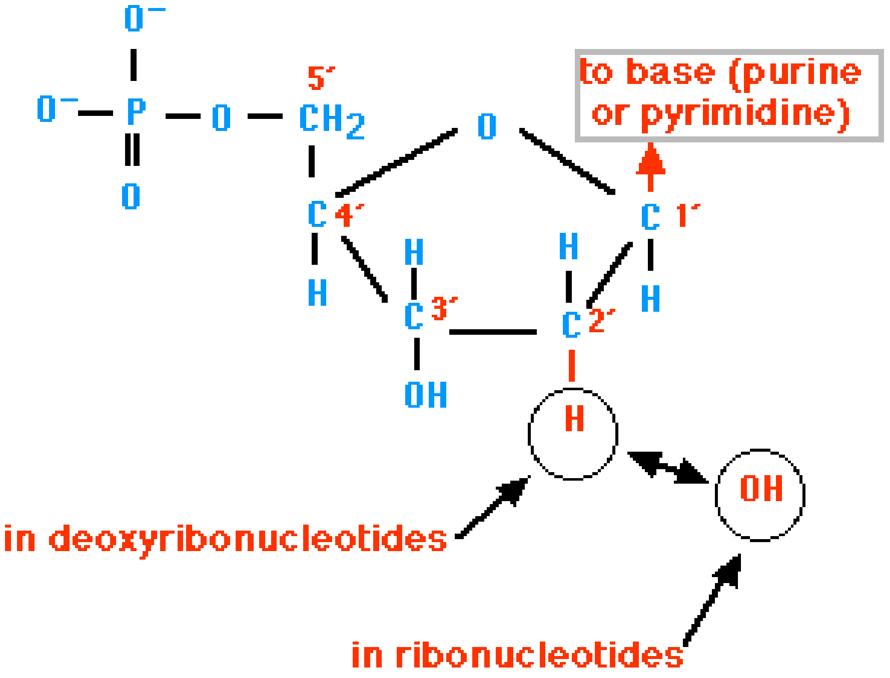
Genes: segments on DNA that contain the instructions for organism's structure and function

Proteins: the workhorse for the cell.

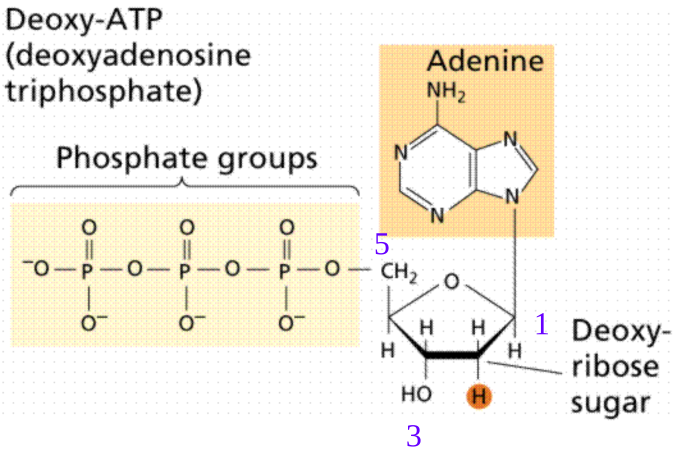
- > establishment and maintenance of structure
- > transport. e.g., hemoglobin, and integral transmembrane proteins
- > protection and defense. e.g., immunoglobin G
- > Control and regulation. e.g., receptors, and DNA binding proteins
- > Catalysis. e.g., enzymes

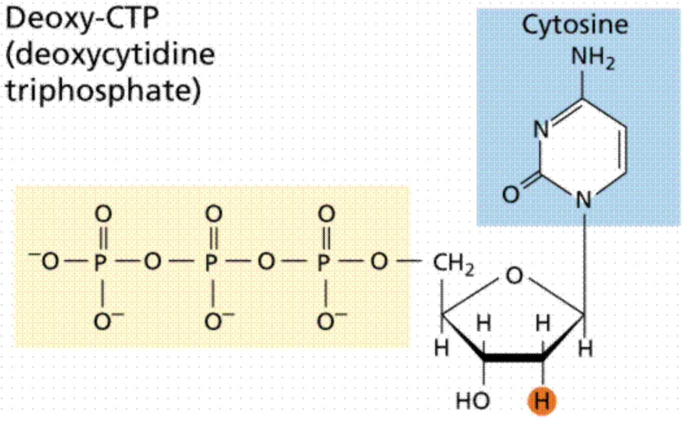
Small molecules:

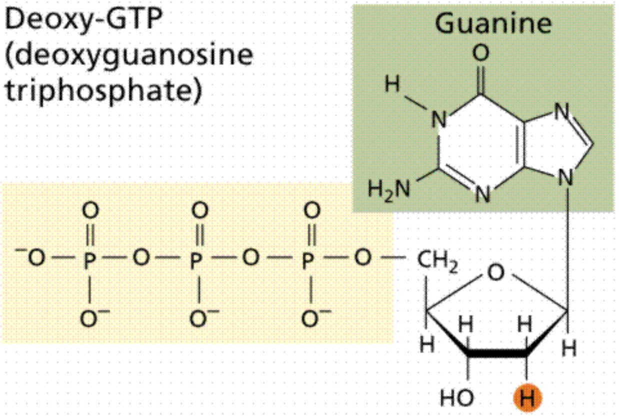
> sugar: carbohydrate
> fatty acids
> nucleotides: A, C, G, T (Purines: A and G; Pyrimidines: C and T)



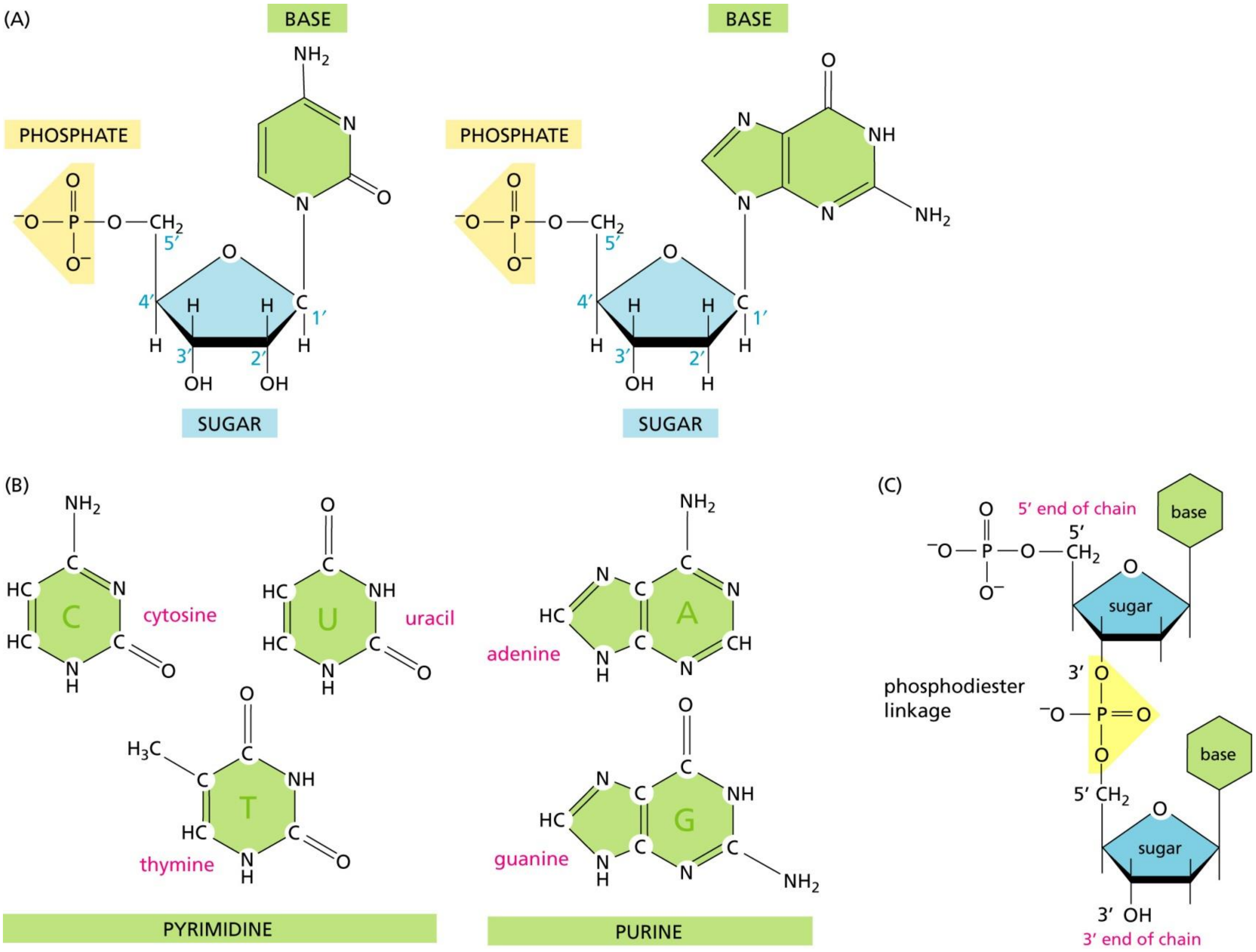
Structure of the bases (Thymine is not shown here)

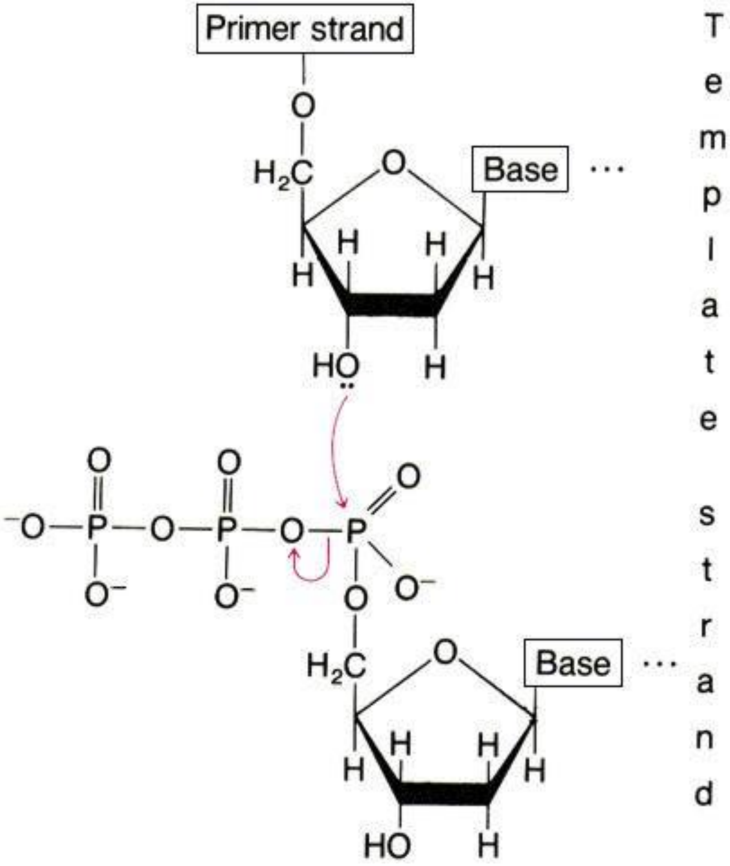






- Purines:A and G
- Pyrimidines: C and T
- Oligonucleotide: a DNA of a few tens of nucleotides
- ATP, ADP, AMP


(A)
BASE
PHOSPHATE
SUGAR
BASE
PHOSPHATE
SUGAR
(B)
cytosine
uracil
adenine
thymine
guanine
PYRIMIDINE
PURINE
(C)
5′ end of chain
base
sugar
phosphodiester linkage
base
sugar
3′ end of chain


Primer strand
O
H2C
O
Base
H
H
H
H
HO
H
O
O
O
⁻O
P
O
P
O
P
O⁻
O⁻
O
O⁻
H2C
O
Base
H
H
H
H
HO
H
Template strand

DNA (double helix, hydrogen bond, complementary bases A-T, G-C)

5' end phosphate group

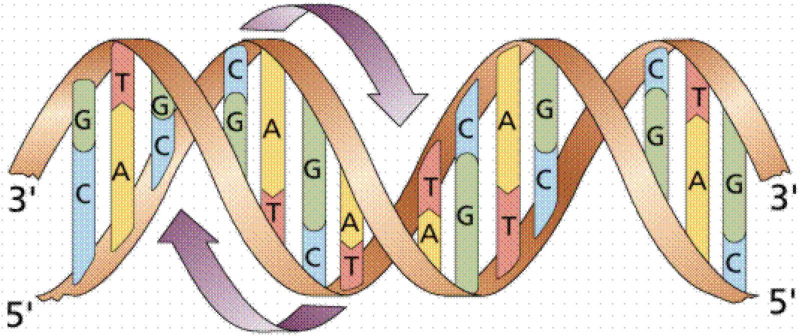
3' end is free

1' position is attached with the base

double strand DNA sequences form a helix via hydrogen bonds between complementary bases

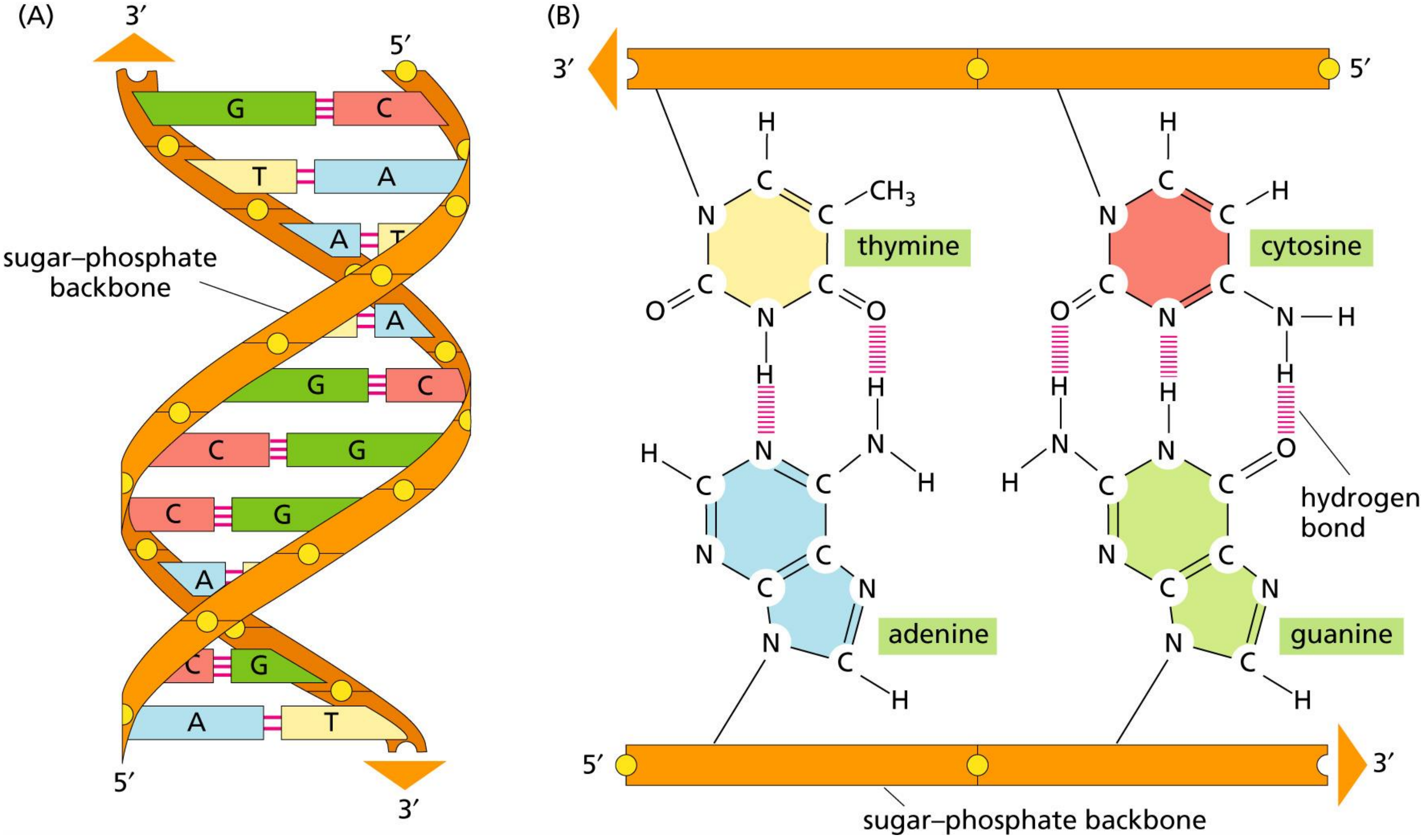
hydrogen bond:

- weak: about 3~5 kJ/mol (A covalent C-C bond has 380 kJ/mol), will break when heated
- saturation:
- spefic:

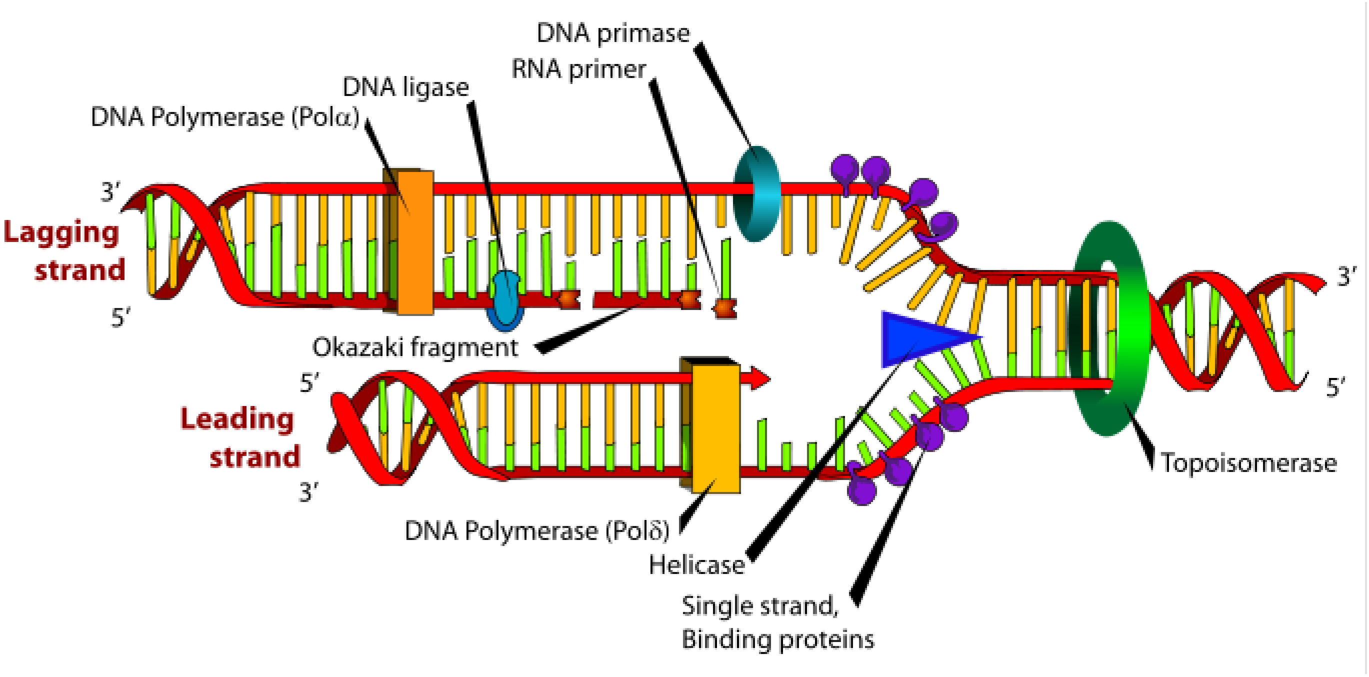


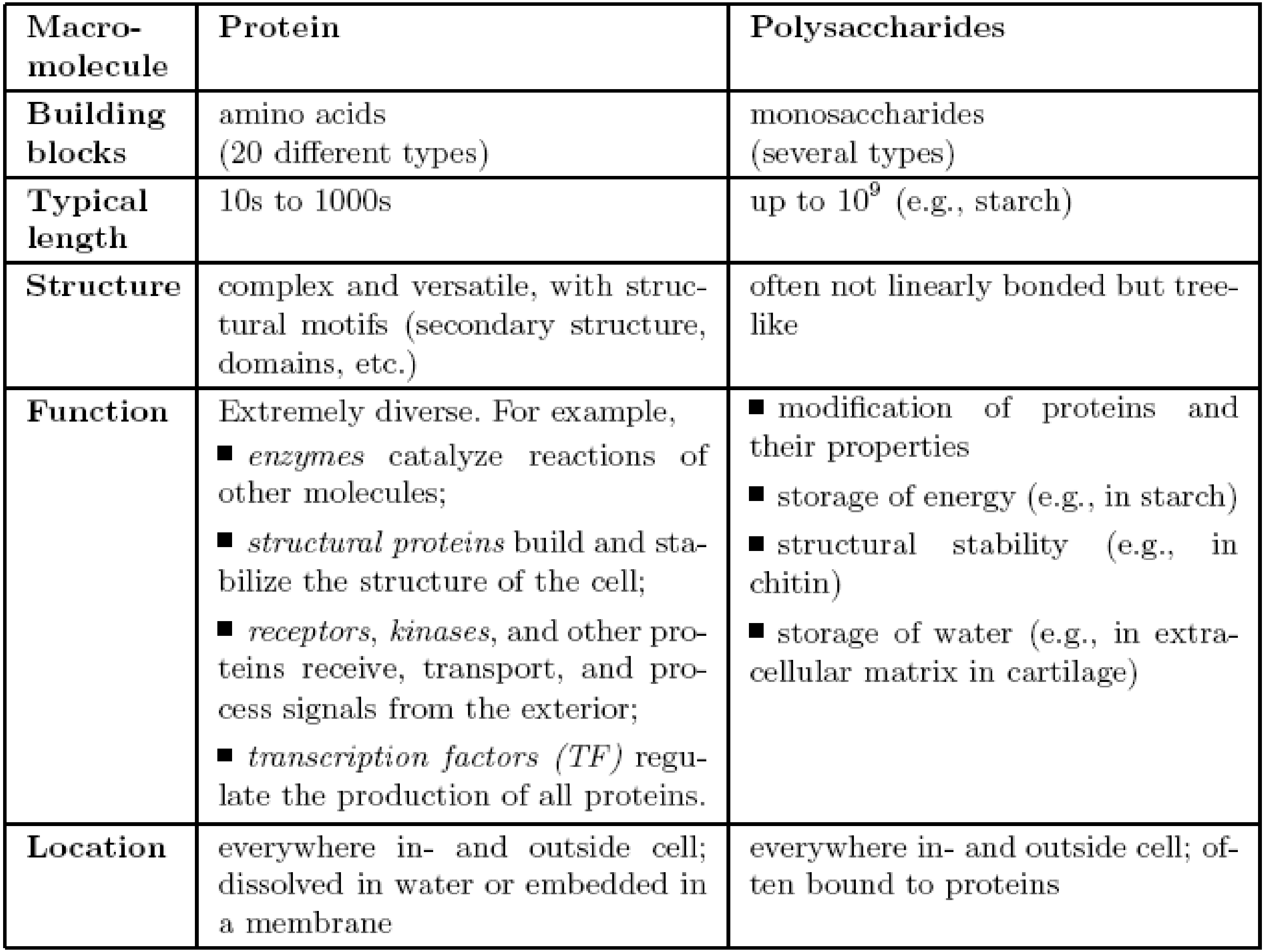
The rules for base pairing (**Watson-Crick** base pairing) :

A with **T**: the purine **adenine** (A) always pairs with the pyrimidine **thymine** (T)
C with **G**: the pyrimidine **cytosine** (C) always pairs with the purine **guanine** (G)

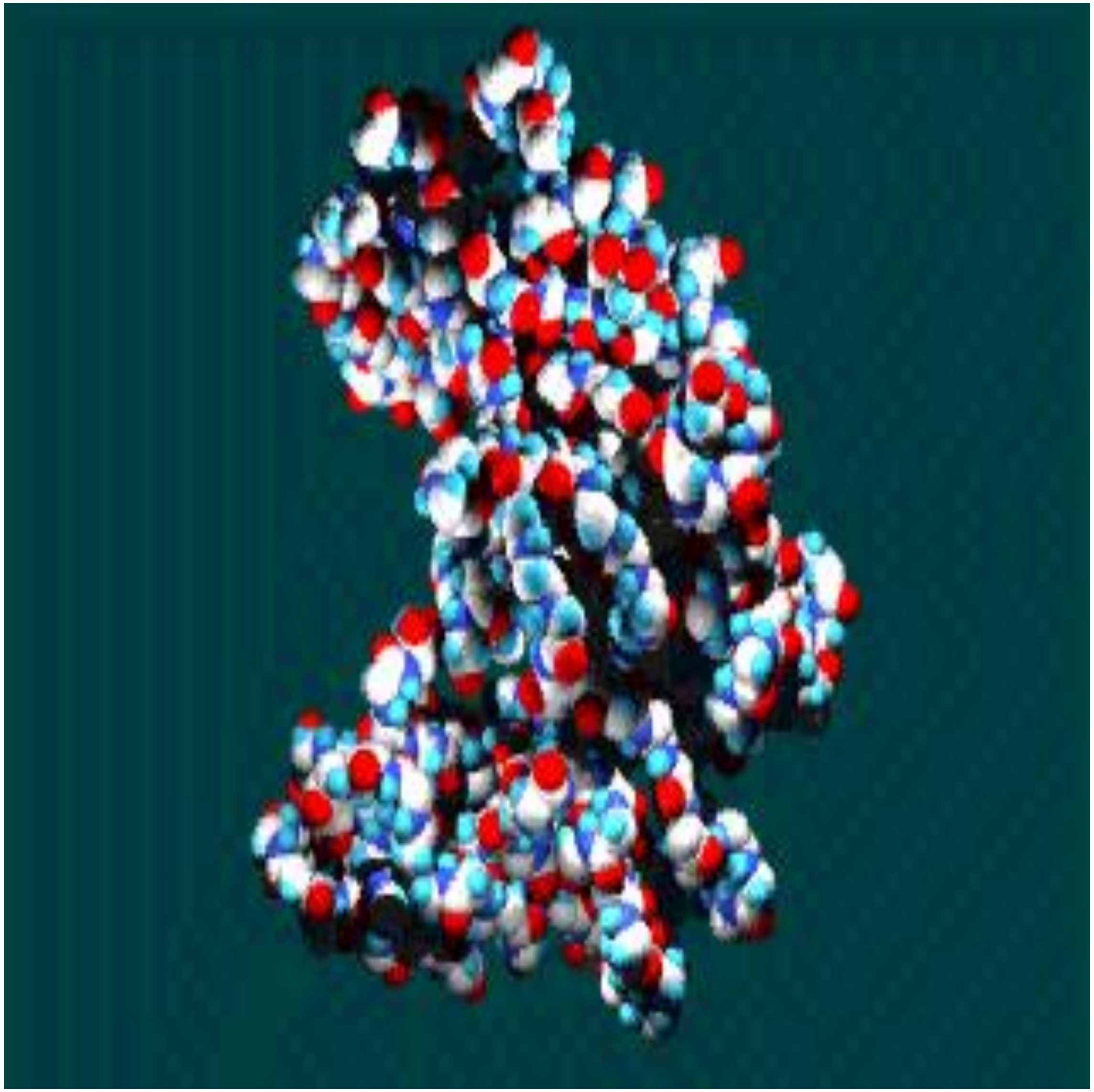


DNA replication



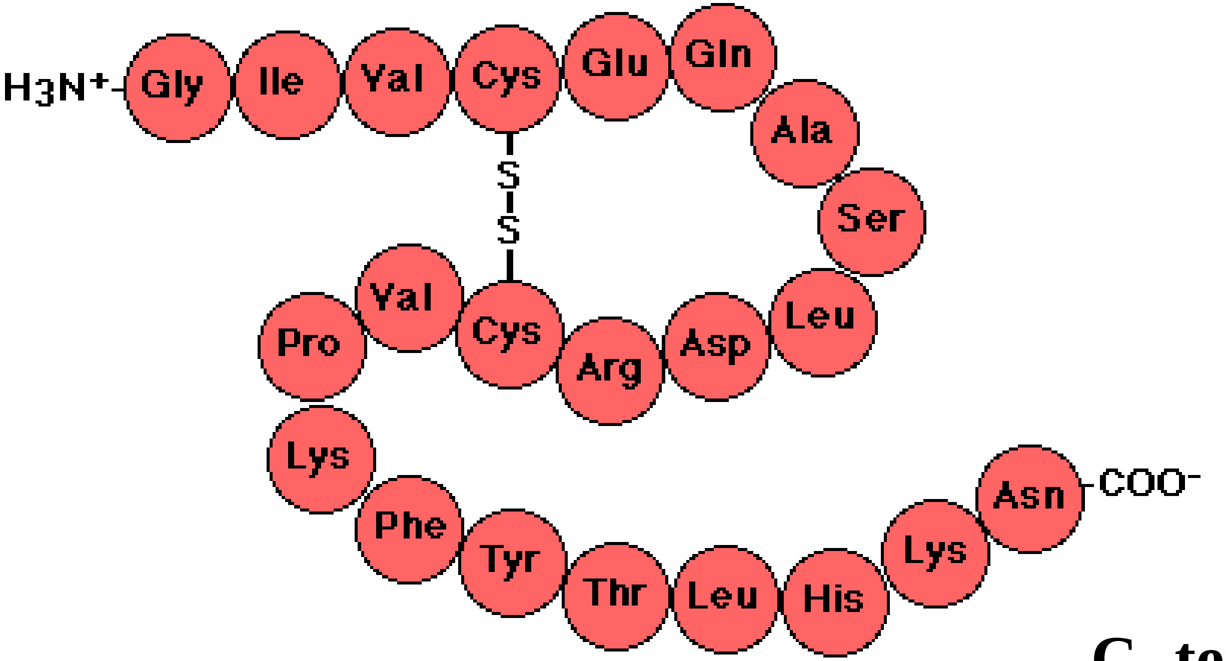


Macro-molecule	**Protein**	**Polysaccharides**
Building blocks	amino acids (20 different types)	monosaccharides (several types)
Typical length	10s to 1000s	up to 10^9 (e.g., starch)
Structure	complex and versatile, with structural motifs (secondary structure, domains, etc.)	often not linearly bonded but tree-like
Function	Extremely diverse. For example, ■ *enzymes* catalyze reactions of other molecules; ■ *structural proteins* build and stabilize the structure of the cell; ■ *receptors*, *kinases*, and other proteins receive, transport, and process signals from the exterior; ■ *transcription factors (TF)* regulate the production of all proteins.	■ modification of proteins and their properties ■ storage of energy (e.g., in starch) ■ structural stability (e.g., in chitin) ■ storage of water (e.g., in extracellular matrix in cartilage)
Location	everywhere in- and outside cell; dissolved in water or embedded in a membrane	everywhere in- and outside cell; often bound to proteins

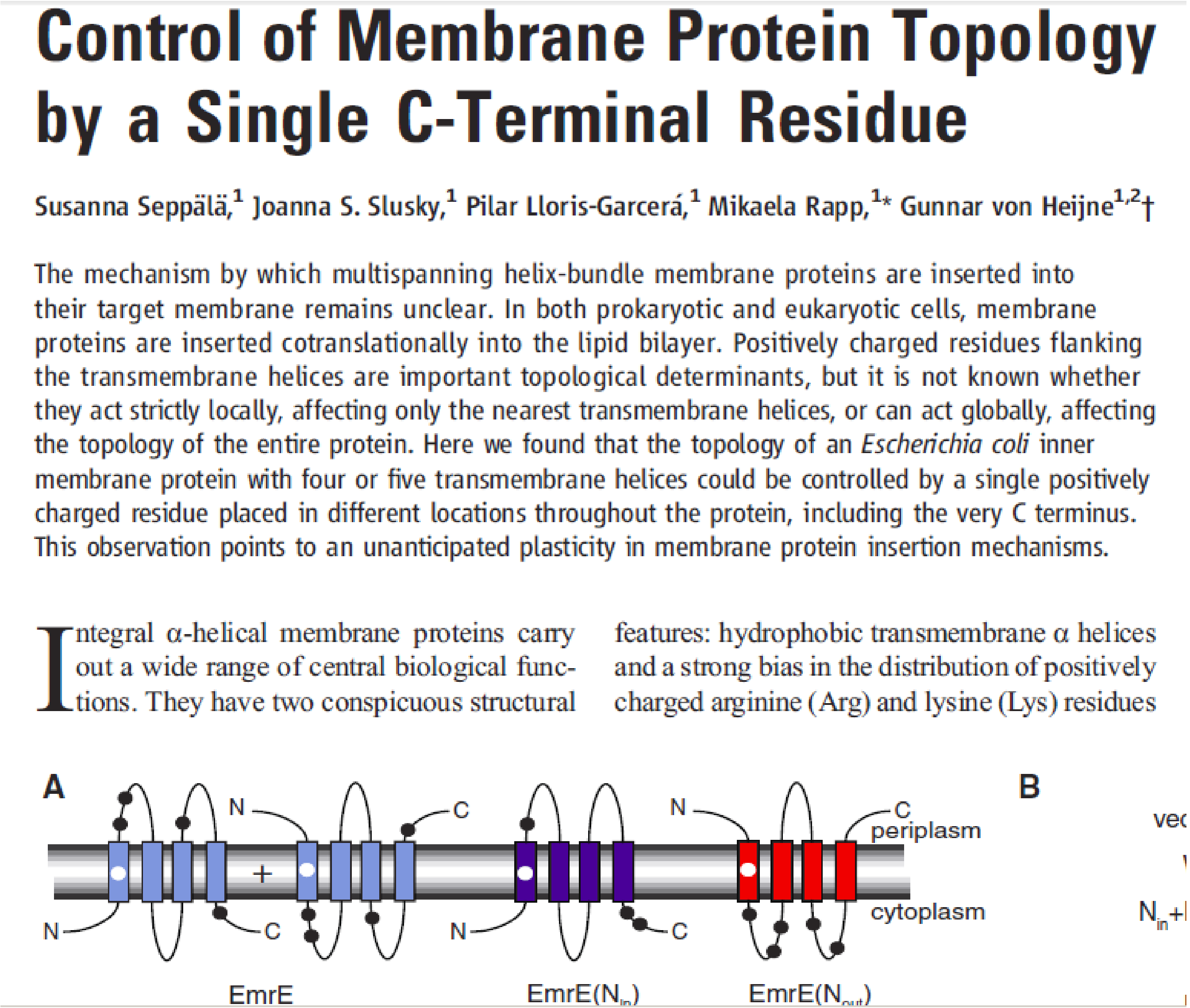


Polypeptide

N-terminal



C- terminal

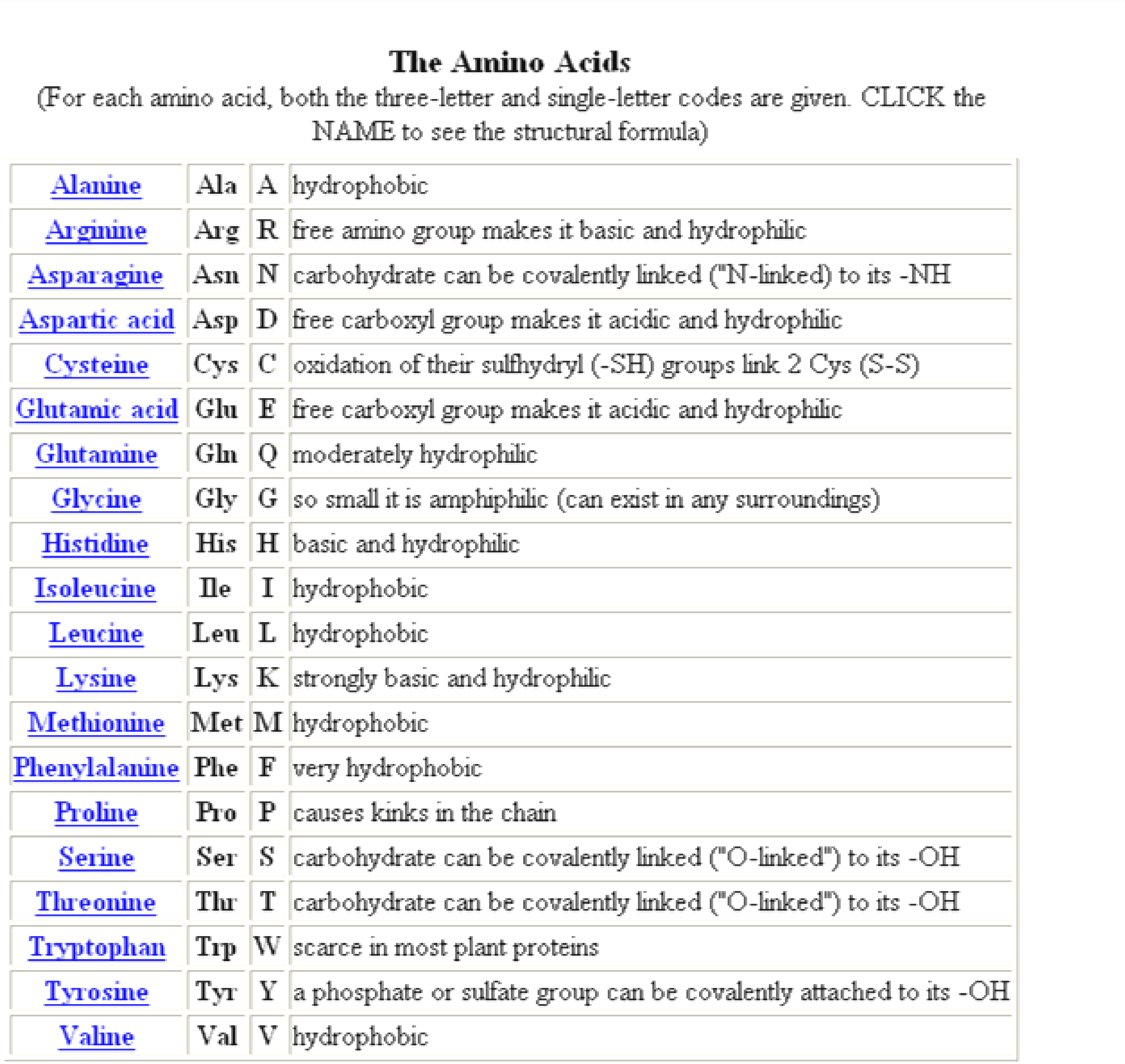


Control of Membrane Protein Topology by a Single C-Terminal Residue

Susanna Seppälä,[1] Joanna S. Slusky,[1] Pilar Lloris-Garcerá,[1] Mikaela Rapp,[1]* Gunnar von Heijne[1,2]†

The mechanism by which multispanning helix-bundle membrane proteins are inserted into their target membrane remains unclear. In both prokaryotic and eukaryotic cells, membrane proteins are inserted cotranslationally into the lipid bilayer. Positively charged residues flanking the transmembrane helices are important topological determinants, but it is not known whether they act strictly locally, affecting only the nearest transmembrane helices, or can act globally, affecting the topology of the entire protein. Here we found that the topology of an *Escherichia coli* inner membrane protein with four or five transmembrane helices could be controlled by a single positively charged residue placed in different locations throughout the protein, including the very C terminus. This observation points to an unanticipated plasticity in membrane protein insertion mechanisms.

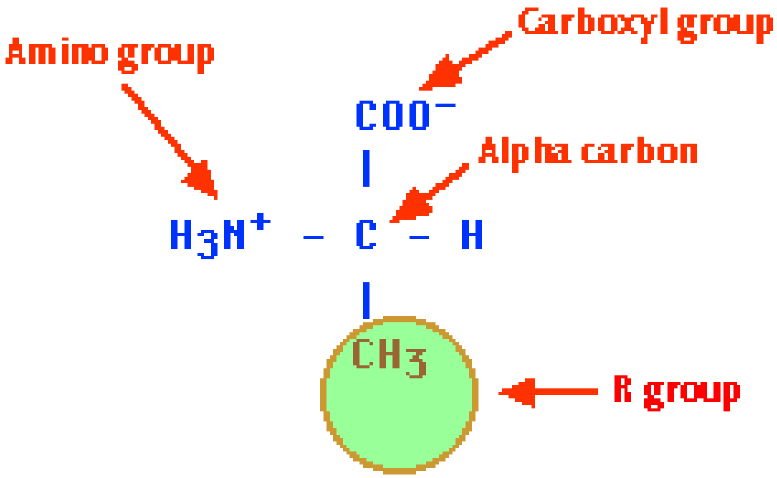
Integral α-helical membrane proteins carry out a wide range of central biological functions. They have two conspicuous structural features: hydrophobic transmembrane α helices and a strong bias in the distribution of positively charged arginine (Arg) and lysine (Lys) residues

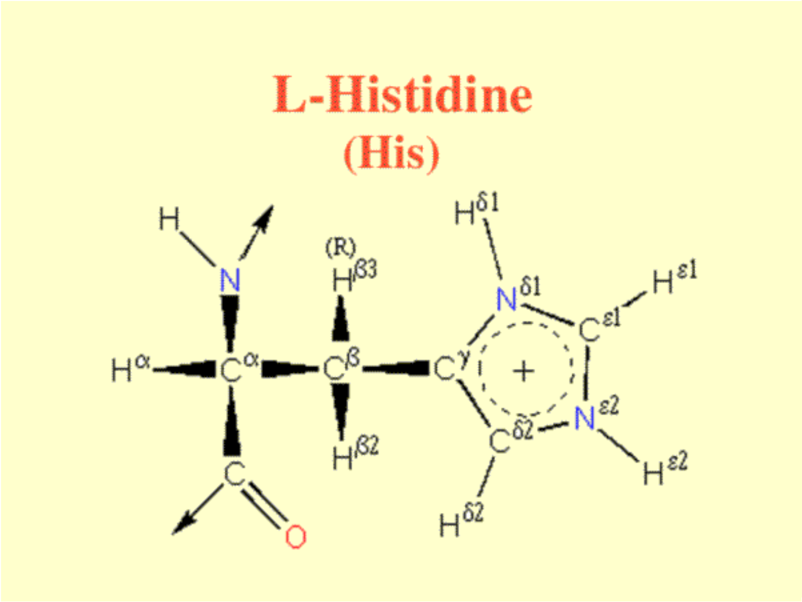


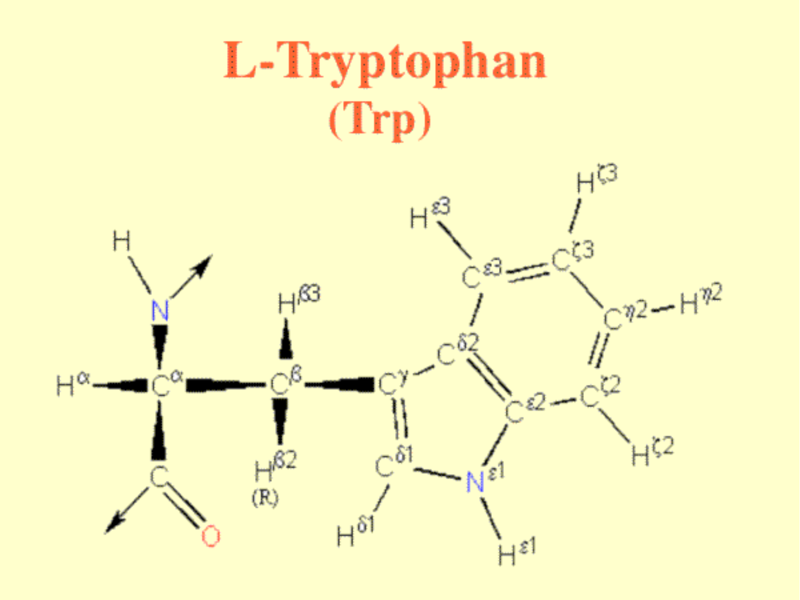
The Amino Acids

(For each amino acid, both the three-letter and single-letter codes are given. CLICK the NAME to see the structural formula)

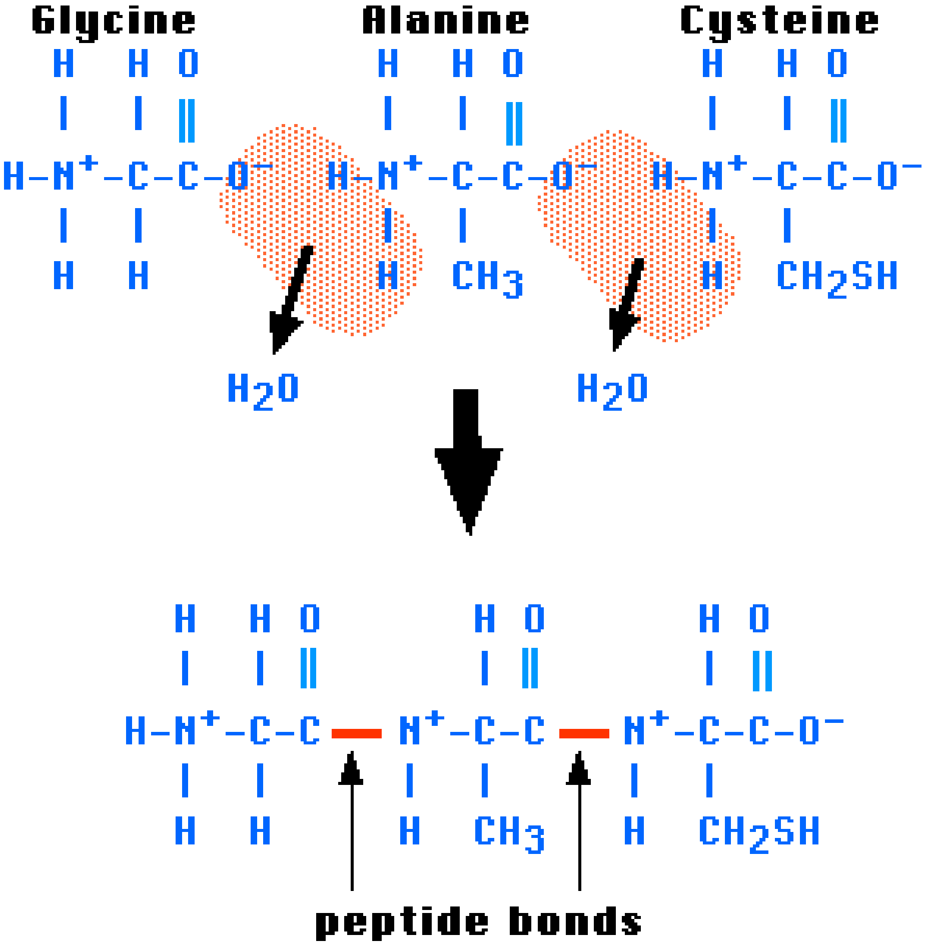
Alanine	Ala	A	hydrophobic
Arginine	Arg	R	free amino group makes it basic and hydrophilic
Asparagine	Asn	N	carbohydrate can be covalently linked ("N-linked) to its -NH
Aspartic acid	Asp	D	free carboxyl group makes it acidic and hydrophilic
Cysteine	Cys	C	oxidation of their sulfhydryl (-SH) groups link 2 Cys (S-S)
Glutamic acid	Glu	E	free carboxyl group makes it acidic and hydrophilic
Glutamine	Gln	Q	moderately hydrophilic
Glycine	Gly	G	so small it is amphiphilic (can exist in any surroundings)
Histidine	His	H	basic and hydrophilic
Isoleucine	Ile	I	hydrophobic
Leucine	Leu	L	hydrophobic
Lysine	Lys	K	strongly basic and hydrophilic
Methionine	Met	M	hydrophobic
Phenylalanine	Phe	F	very hydrophobic
Proline	Pro	P	causes kinks in the chain
Serine	Ser	S	carbohydrate can be covalently linked ("O-linked") to its -OH
Threonine	Thr	T	carbohydrate can be covalently linked ("O-linked") to its -OH
Tryptophan	Trp	W	scarce in most plant proteins
Tyrosine	Tyr	Y	a phosphate or sulfate group can be covalently attached to its -OH
Valine	Val	V	hydrophobic


Amino group
Carboxyl group
COO−
Alpha carbon
H3N+ – C – H
CH3
R group


L-Histidine
(His)

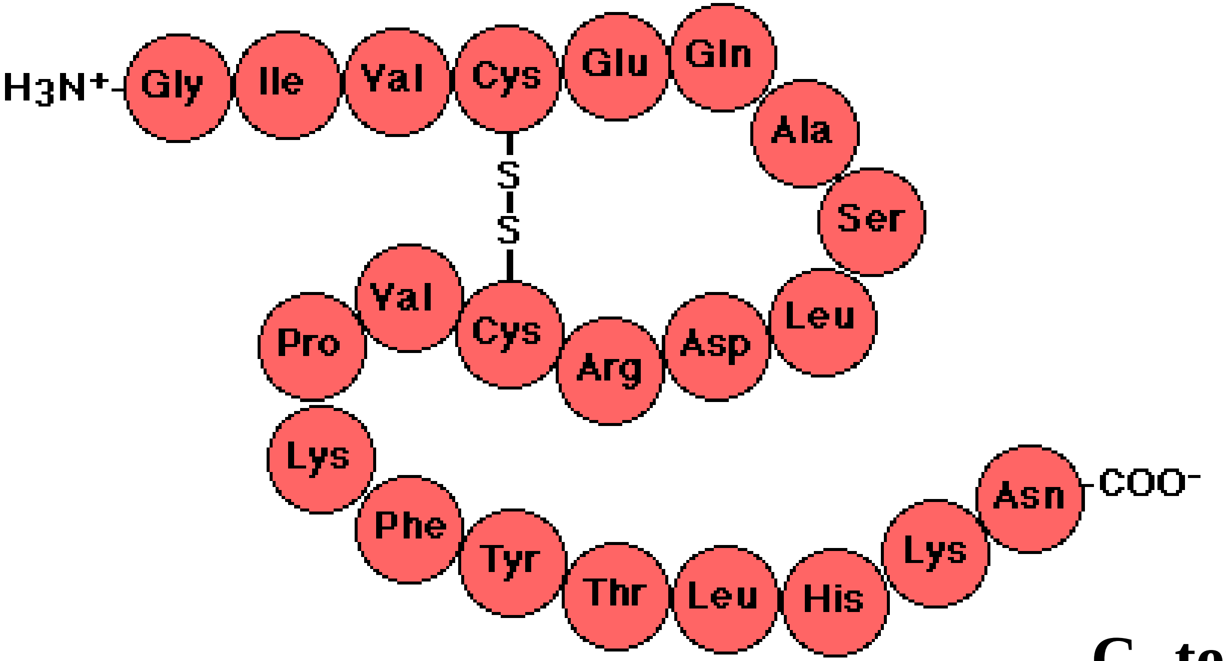

L-Tryptophan
(Trp)

Peptide bond

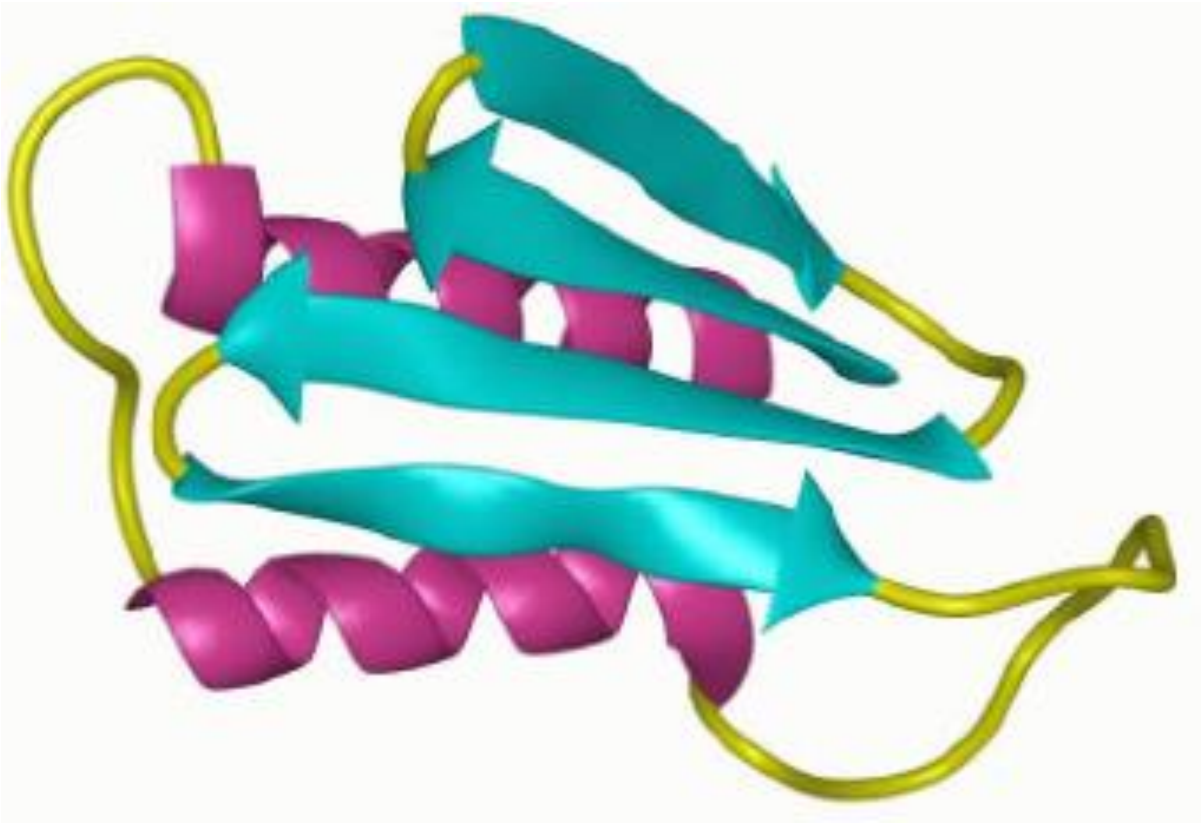


Polypeptide

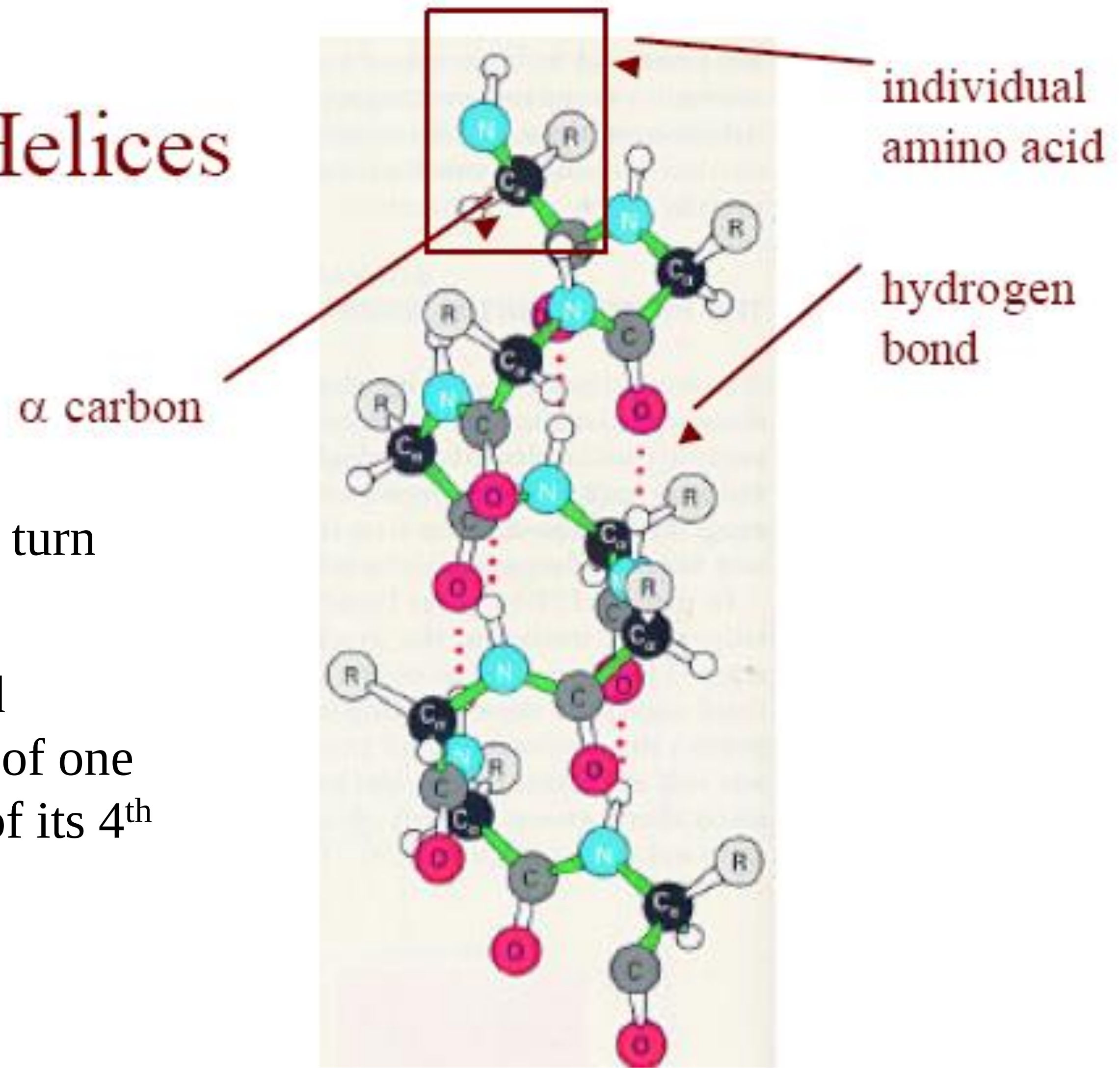
N-terminal



C- terminal

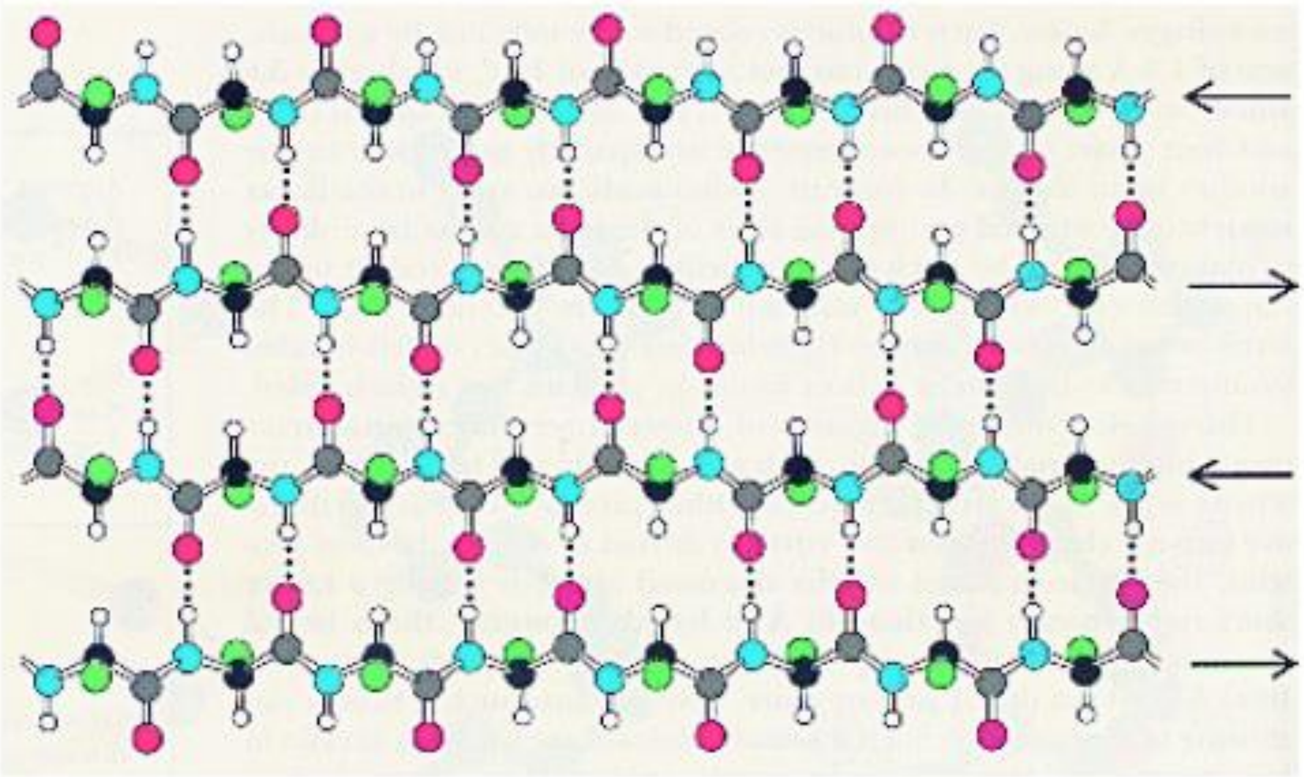


α Helices



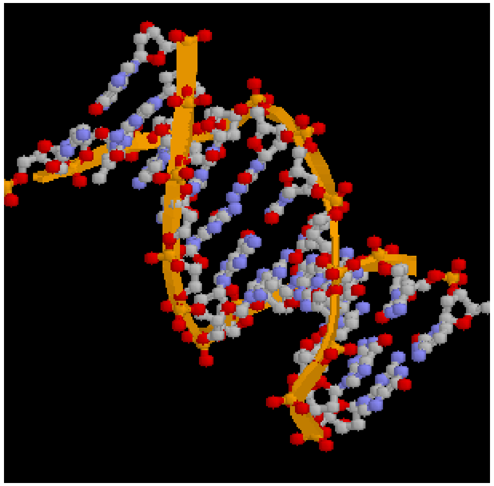
- Helix complete turn every 3.6 AAs
- Hydrogen bond between (-C=O) of one AA and (-N-H) of its 4th neighboring AA

β Strands

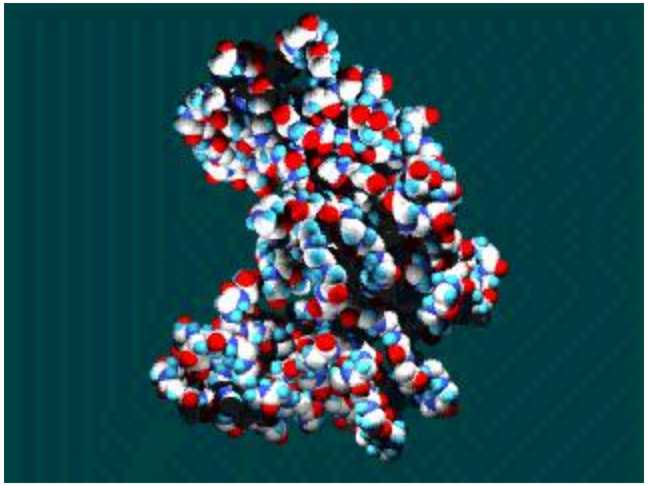


Hydrogen bond b/w carbonyl oxygen atom on one chain and NH group on the adjacent chain

Information Expression



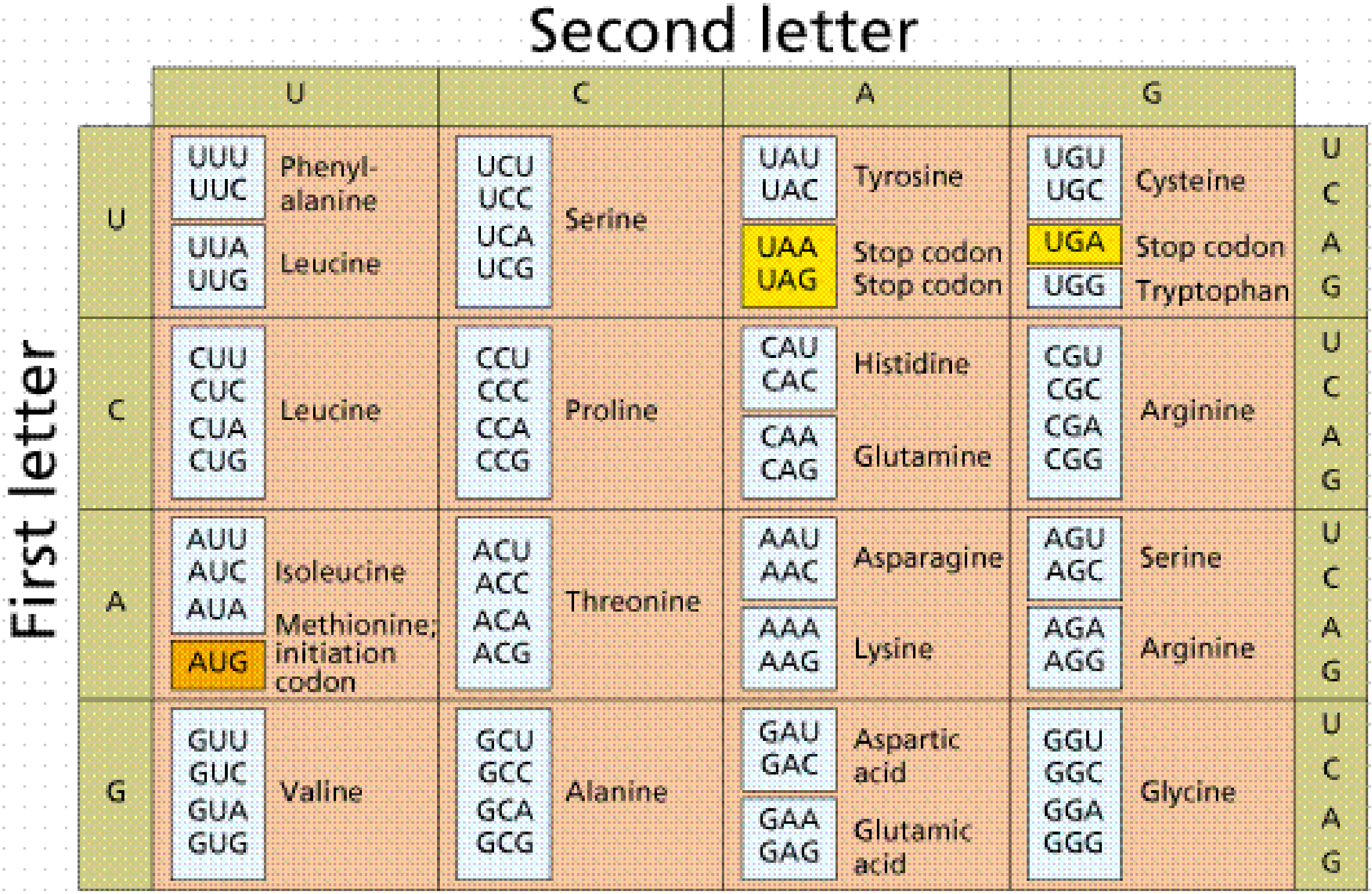
1-D information array



3-D biochemical structure

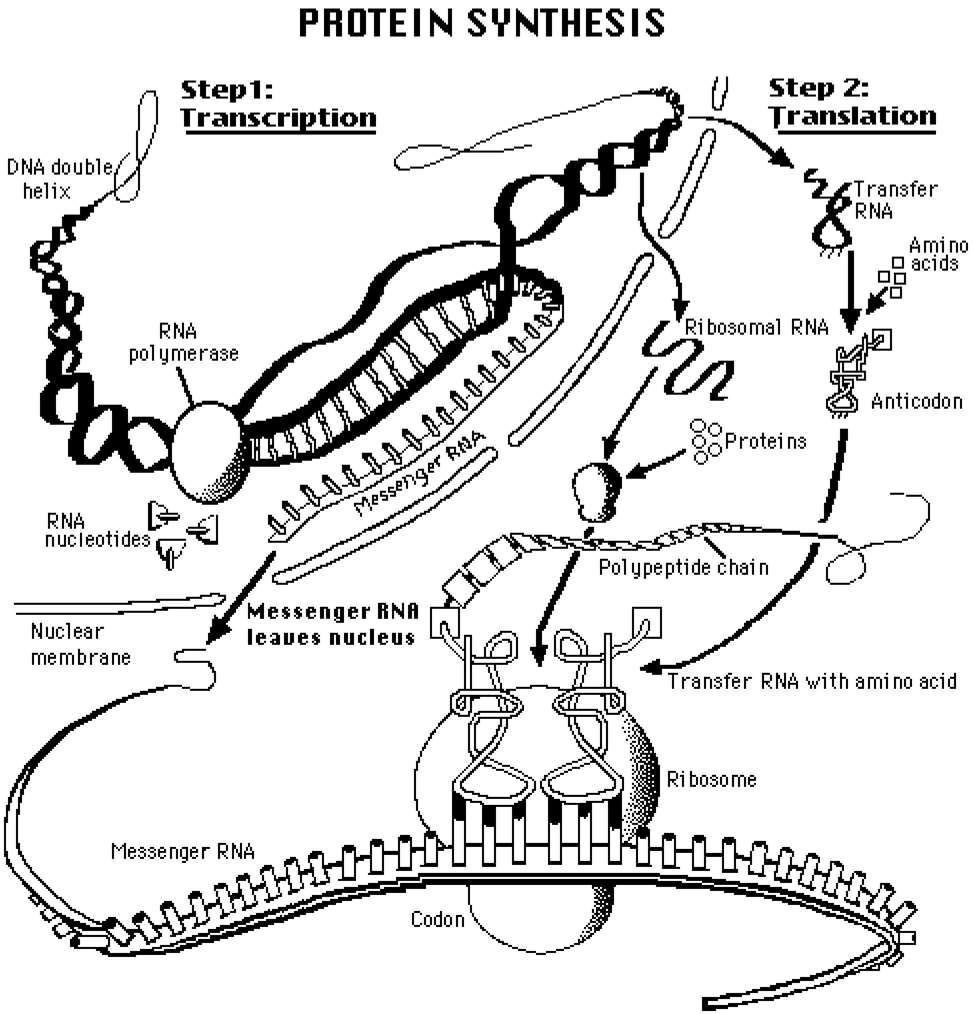
Central Dogma: DNA → RNA → Protein

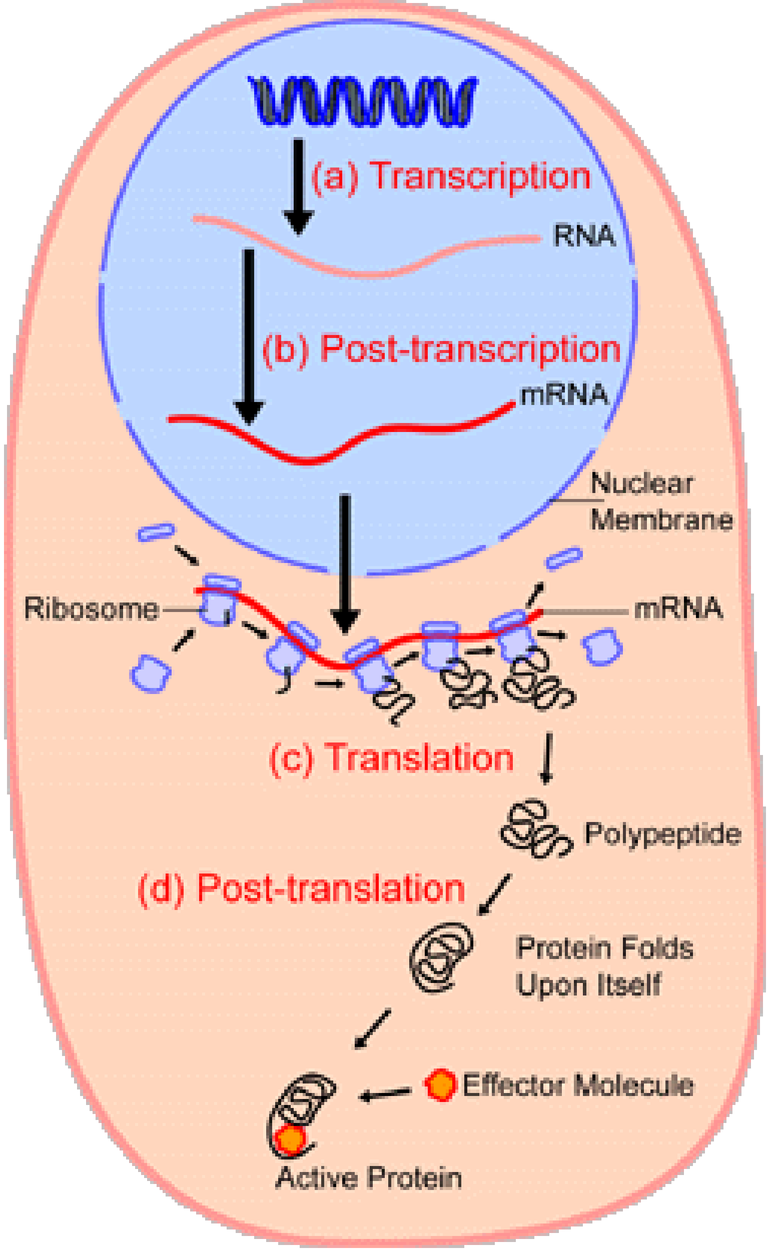
Genetic Code: codons



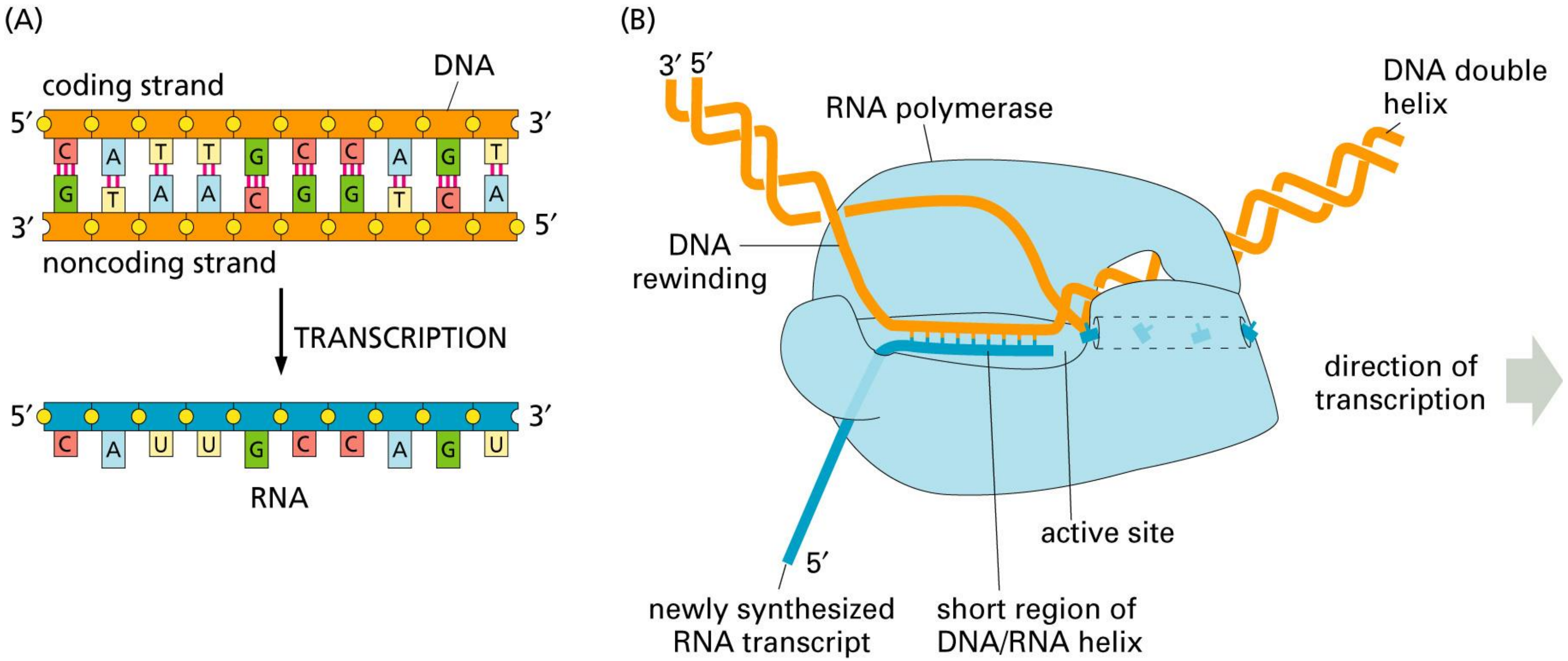
Second letter

First letter	U	C	A	G	
U	UUU UUC Phenyl-alanine UUA UUG Leucine	UCU UCC UCA UCG Serine	UAU UAC Tyrosine UAA Stop codon UAG Stop codon	UGU UGC Cysteine UGA Stop codon UGG Tryptophan	U C A G
C	CUU CUC CUA CUG Leucine	CCU CCC CCA CCG Proline	CAU CAC Histidine CAA CAG Glutamine	CGU CGC CGA CGG Arginine	U C A G
A	AUU AUC AUA Isoleucine AUG Methionine; initiation codon	ACU ACC ACA ACG Threonine	AAU AAC Asparagine AAA AAG Lysine	AGU AGC Serine AGA AGG Arginine	U C A G
G	GUU GUC GUA GUG Valine	GCU GCC GCA GCG Alanine	GAU GAC Aspartic acid GAA GAG Glutamic acid	GGU GGC GGA GGG Glycine	U C A G


PROTEIN SYNTHESIS
Step1: Transcription
Step 2: Translation
DNA double helix
RNA polymerase
RNA nucleotides
Messenger RNA
Transfer RNA
Amino acids
Ribosomal RNA
Anticodon
Proteins
Polypeptide chain
Nuclear membrane
Messenger RNA leaves nucleus
Transfer RNA with amino acid
Ribosome
Messenger RNA
Codon

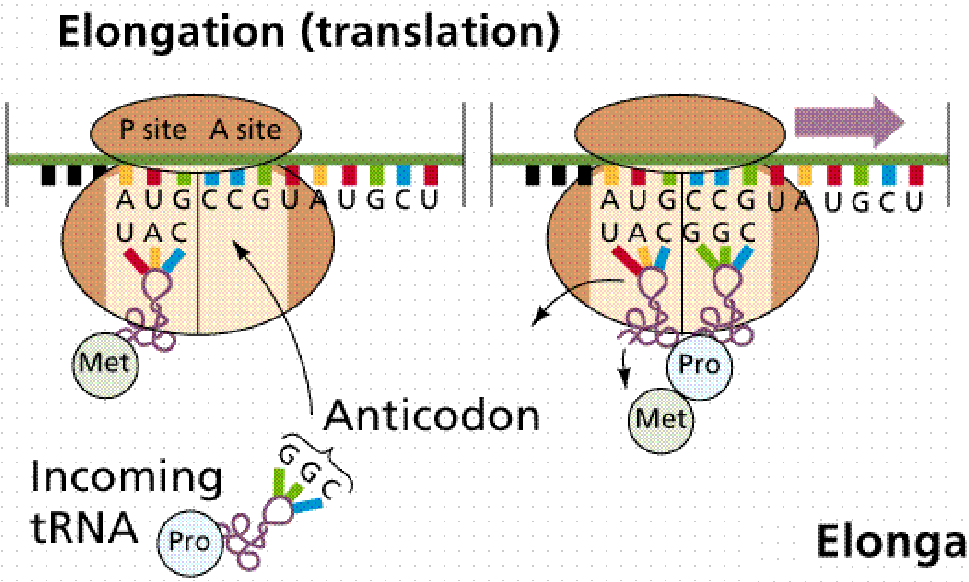

(a) Transcription
RNA
(b) Post-transcription
mRNA
Nuclear Membrane
Ribosome
mRNA
(c) Translation
Polypeptide
(d) Post-translation
Protein Folds Upon Itself
Effector Molecule
Active Protein

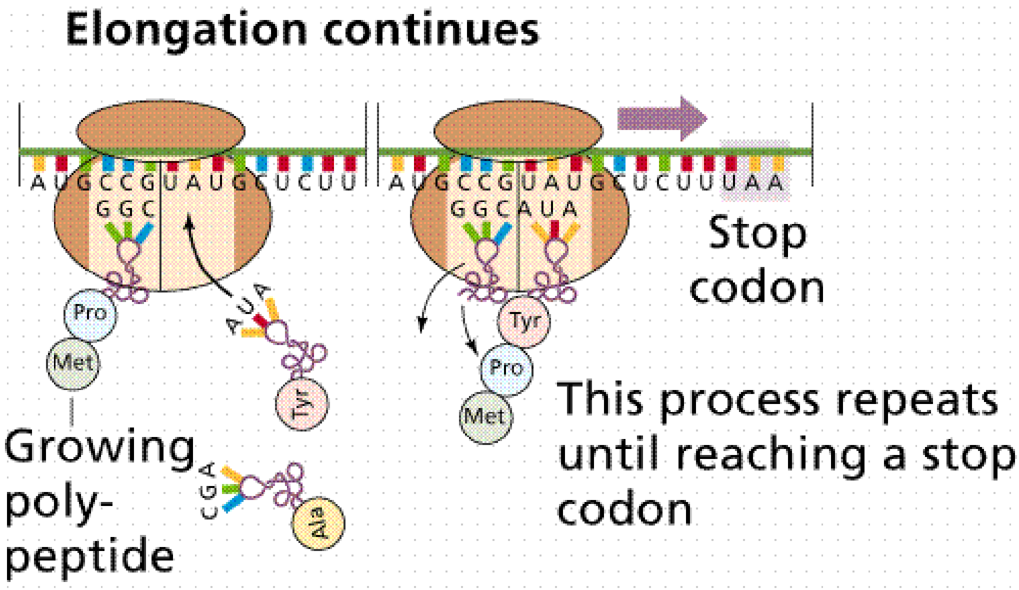
Transcription

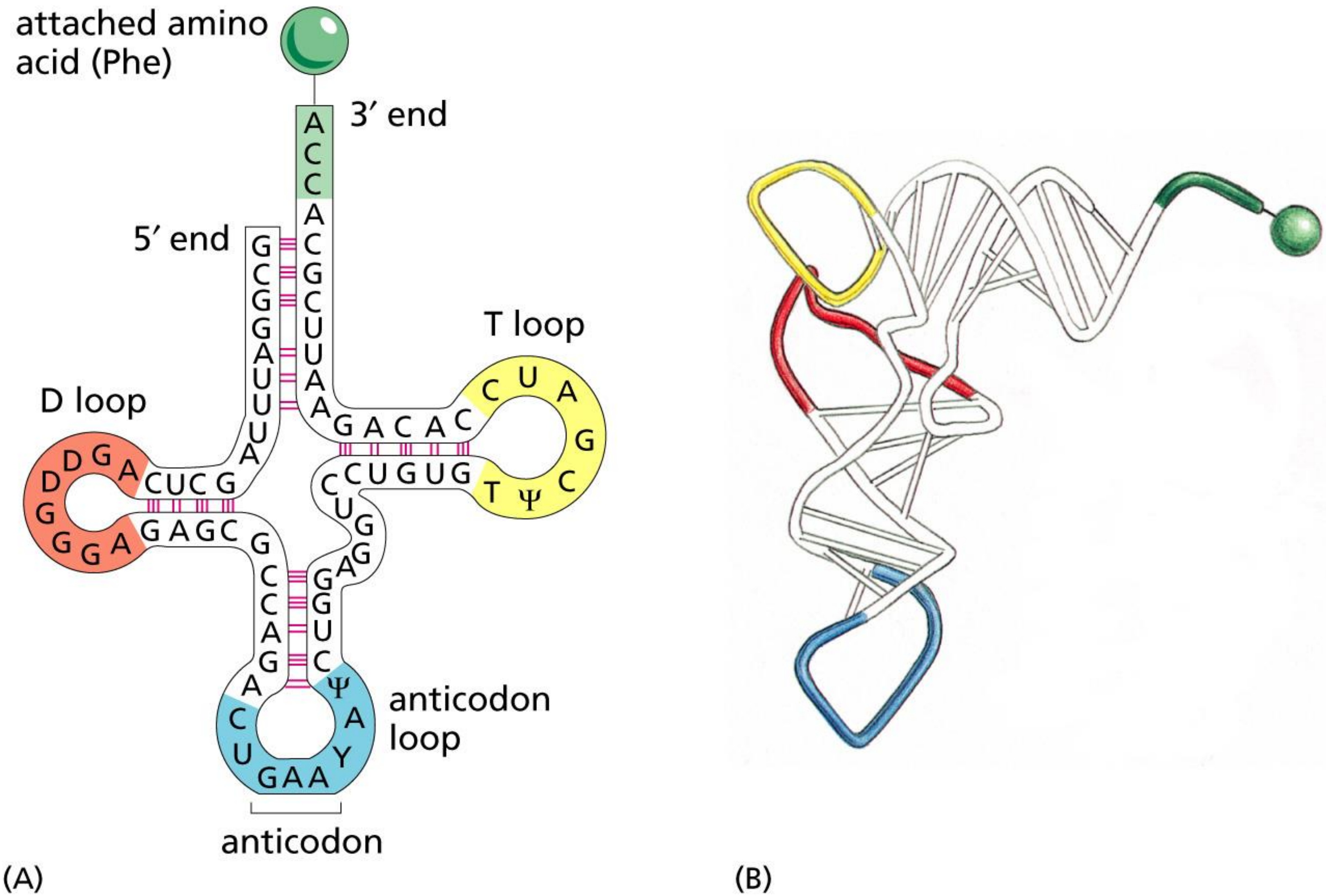


Translation

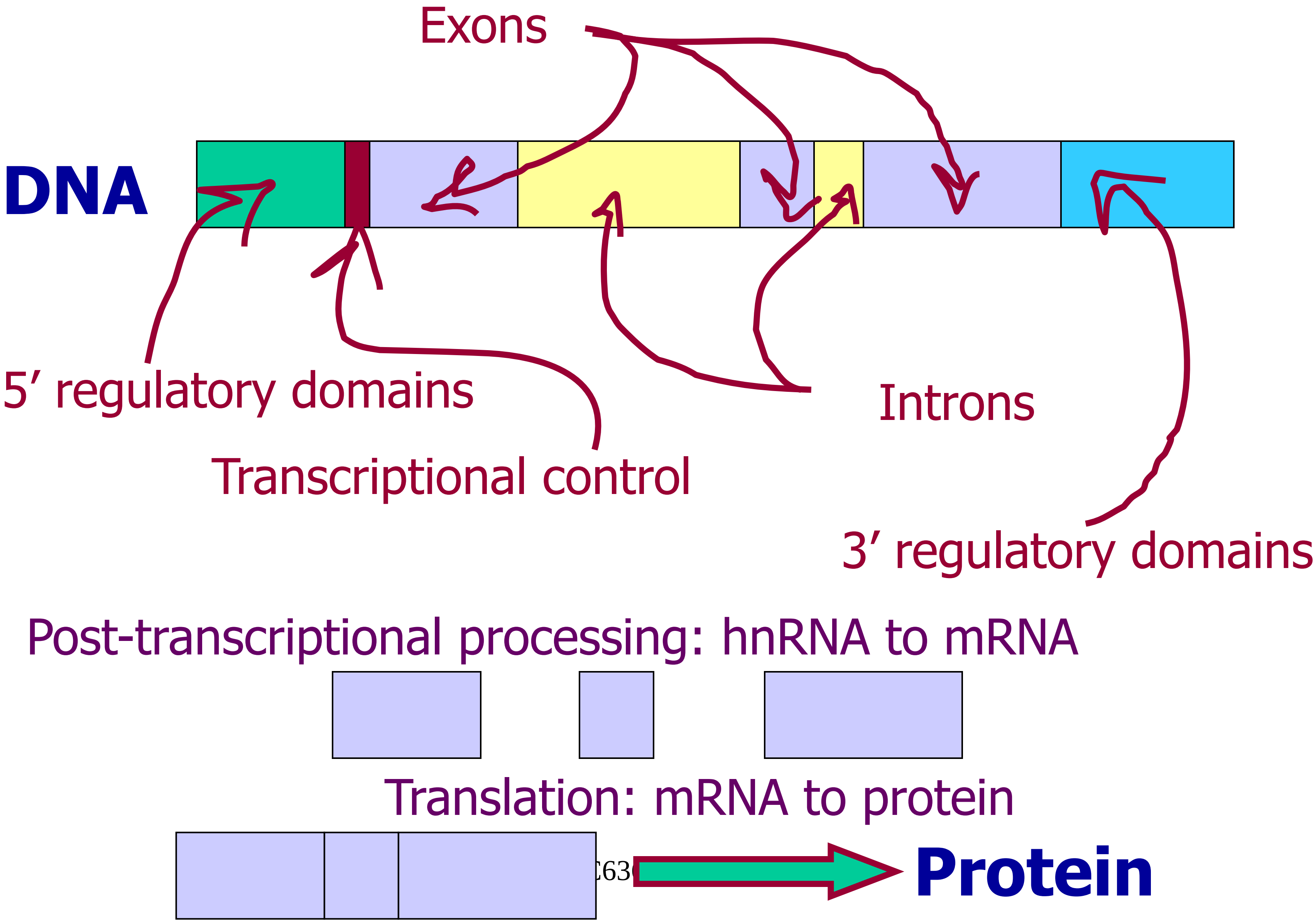
http://www.youtube.com/watch?v=B6O6uRb1D38






attached amino acid (Phe)
3′ end
5′ end
A C C A
G C G G A U U U A
A C G C U U A A
D loop
D D G A
D G G G A
C U C G
G A G C
T loop
C U A G C Ψ T
G A C A C
C U G U G
C U G G A G
G C C A G A
G G U C
C U G A A Y A Ψ
anticodon loop
anticodon
(A)
(B)

Gene Structure



Exons

DNA

5′ regulatory domains

Transcriptional control

Introns

3′ regulatory domains

Post-transcriptional processing: hnRNA to mRNA

Translation: mRNA to protein

Protein

How complex can a 4 letter code really be?

atcgggctatcgatagctatagcgcgatatatcgcgcgtatatgcgcgcatattag
tagctagtgctgattcatctggactgtcgtaatatatacgcgcccggctatcgcgct
atgcgcgatatcgcgcggcgctatataaatattaaaaaataaaatatatatatatgc
tgcgcgatagcgctataggcgcgctatccatatataggcgctcgcccgggcgcga
tgcatcggctacggctagctgtagctagtcggcgattagcggcttatatgcggcga
gcgatgagagtcgcggctataggcttaggctatagcgctagtatatagcggctagc
cgcgtagacgcgatagcgtagctagcggcgcgcgtatatagcgcttaagagcca
aaatgcgtctagcgctataatatgcgctatagctatatgcggctattatatagcgca
gcgctagctagcgtatcaggcgaggagatcgatgctactgatcgatgctagagca
gcgtcatgctagtagtgccatatatatgctgagcgcgcgtagctcgatattacgcta
cctagatgctagcgagctatgatcgtagca...

- Alternative splicing
  - Exception to the “One gene one protein” rule.
- Codon usage
  - http://www.kazusa.or.jp/codon/
- EST (expressed sequence tag)
- Reverse translation and transcription
- cDNA

Given a DNA sequence, we like to computationally

- Identify genes,
  - introns, exons, alternative splicing sites, promoters, …
- Determine the functions of the protein that a gene encodes
- Identify functional signatures, e.g., motifs
- Determine the structure of proteins