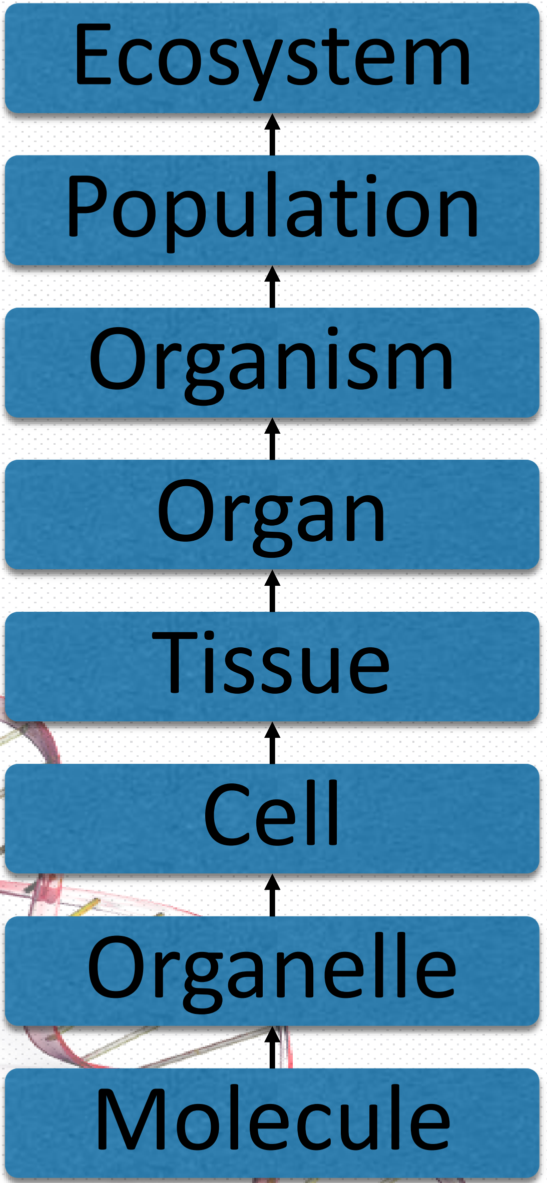
Background: Molecular Biology

Hammad Naveed
08/25/2023

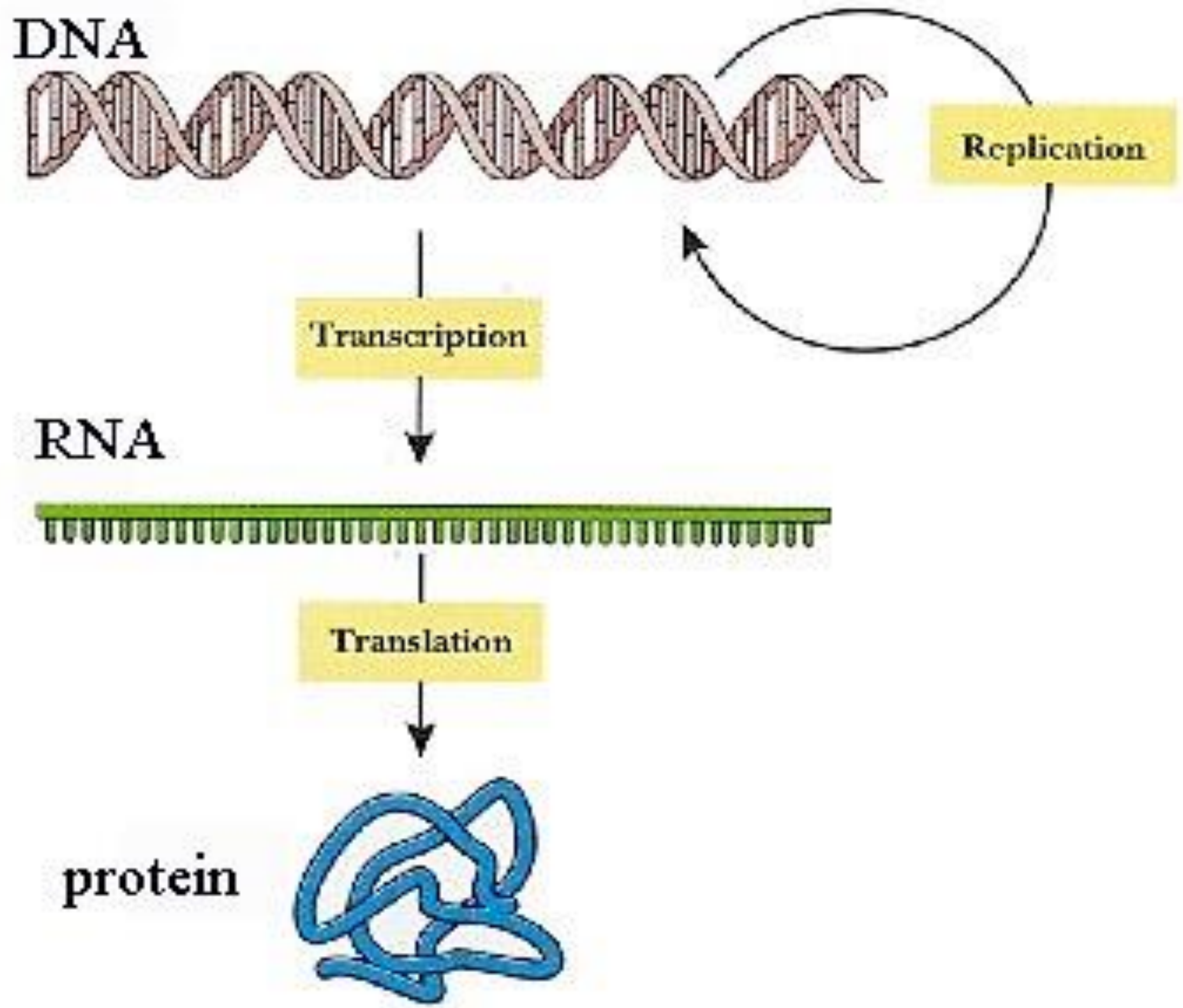
Levels of the biological hierarchy

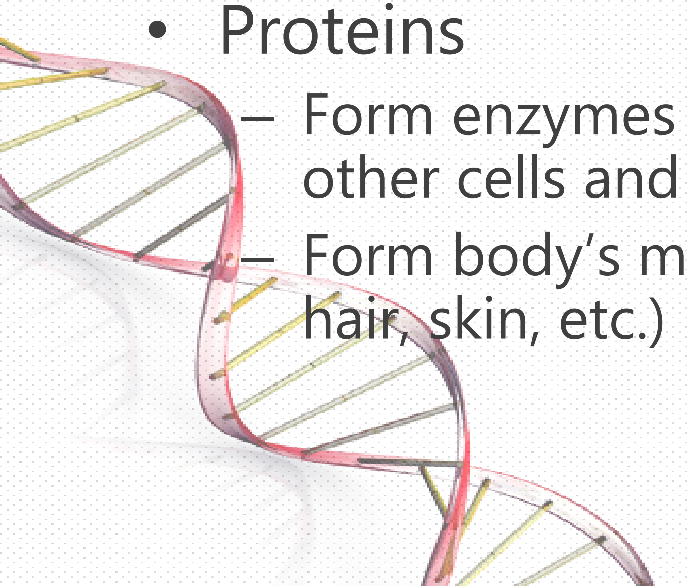


Ecosystem	Lake Monona, Mojave desert, …
Population	Madison Humans, Monona Muskie
Organism	Pine tree, Lizard, Human…
Organ	Heart, Lungs, Brain…
Tissue	Adipose tissue, Blood, Nerve tissue…
Cell	Fat cell, Blood cell, Muscle cell…
Organelle	Nucleus, Mitochondrion…
Molecule	DNA, RNA, Protein, Lipids…

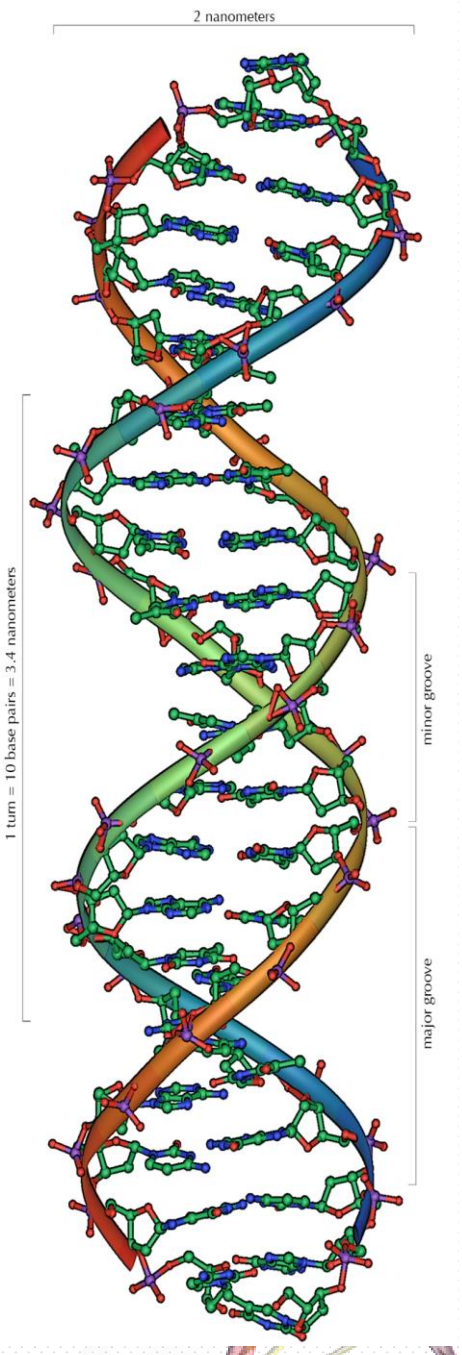
All Life depends on 3 critical molecules

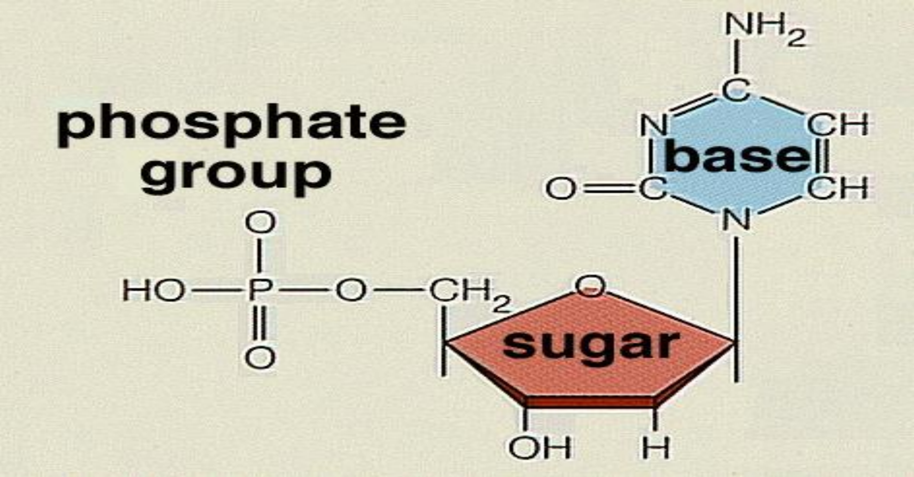
- DNAs
  - Hold information on how cell works
- RNAs
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body’s major components (e.g.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

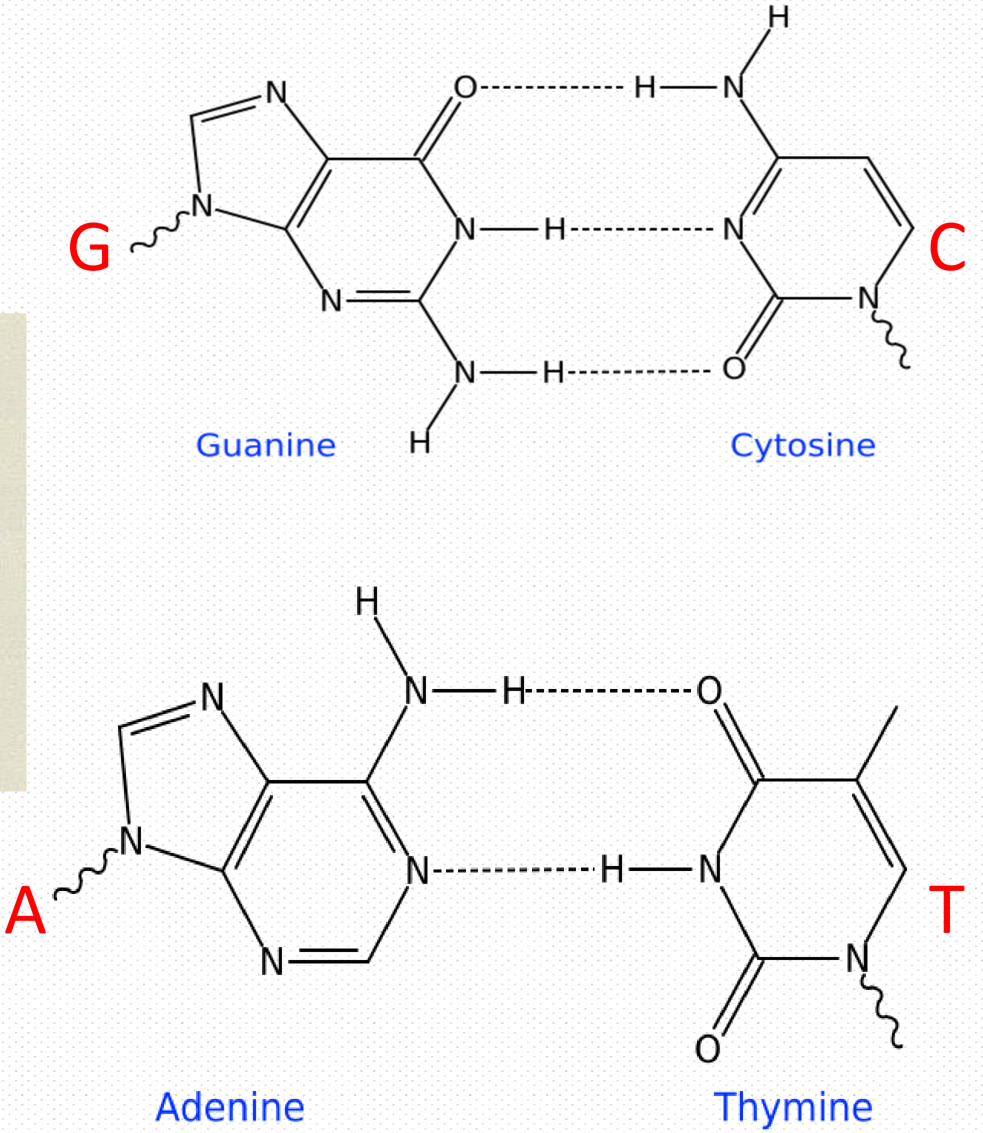




DNA

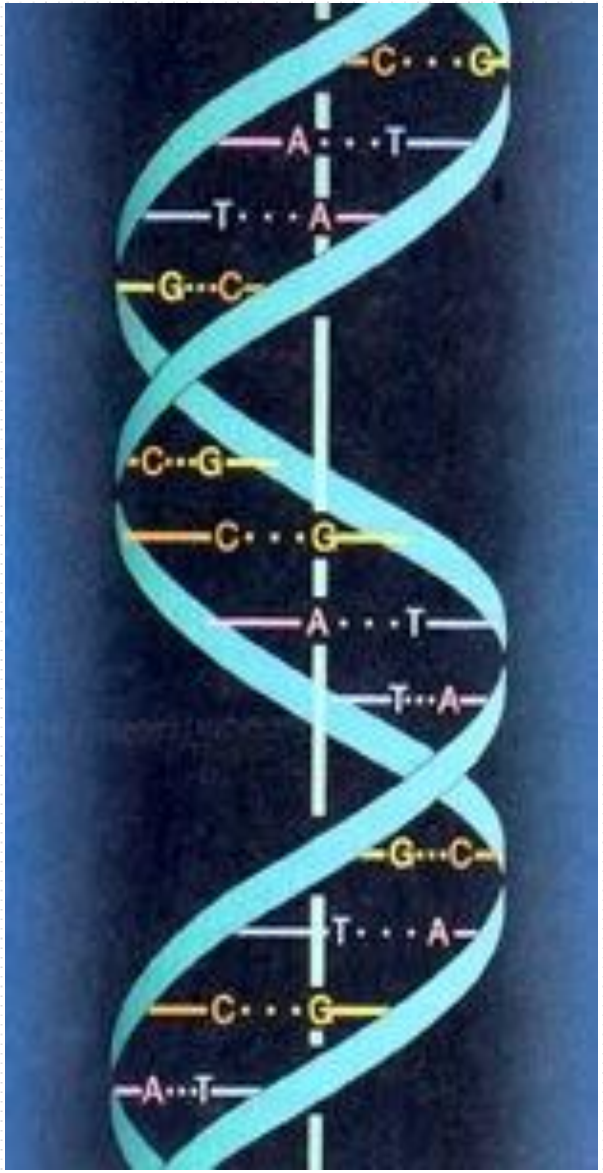

2 nanometers
1 turn = 10 base pairs = 3.4 nanometers
minor groove
major groove

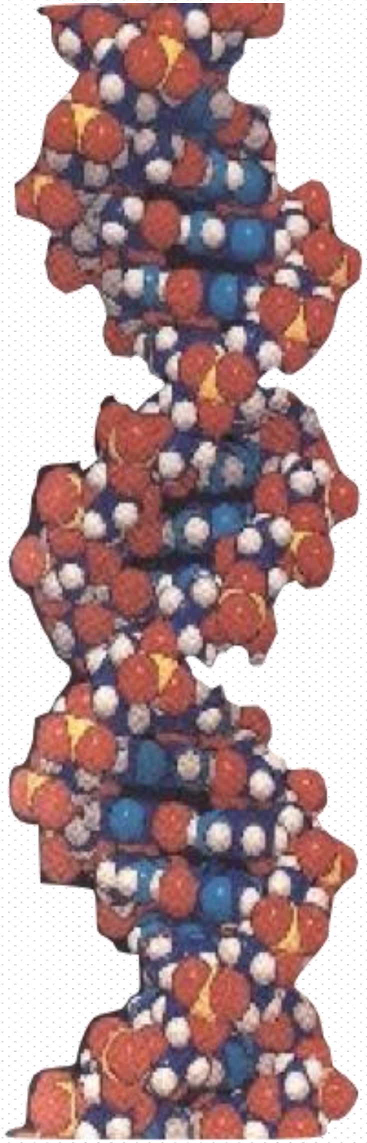

phosphate group
base
sugar
NH2
HO—P—O—CH2
OH
H

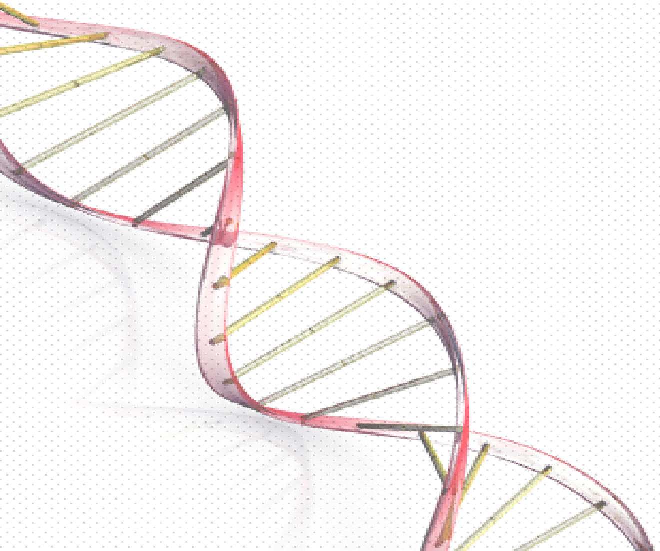

G
C
Guanine
Cytosine
A
T
Adenine
Thymine

The Double Helix

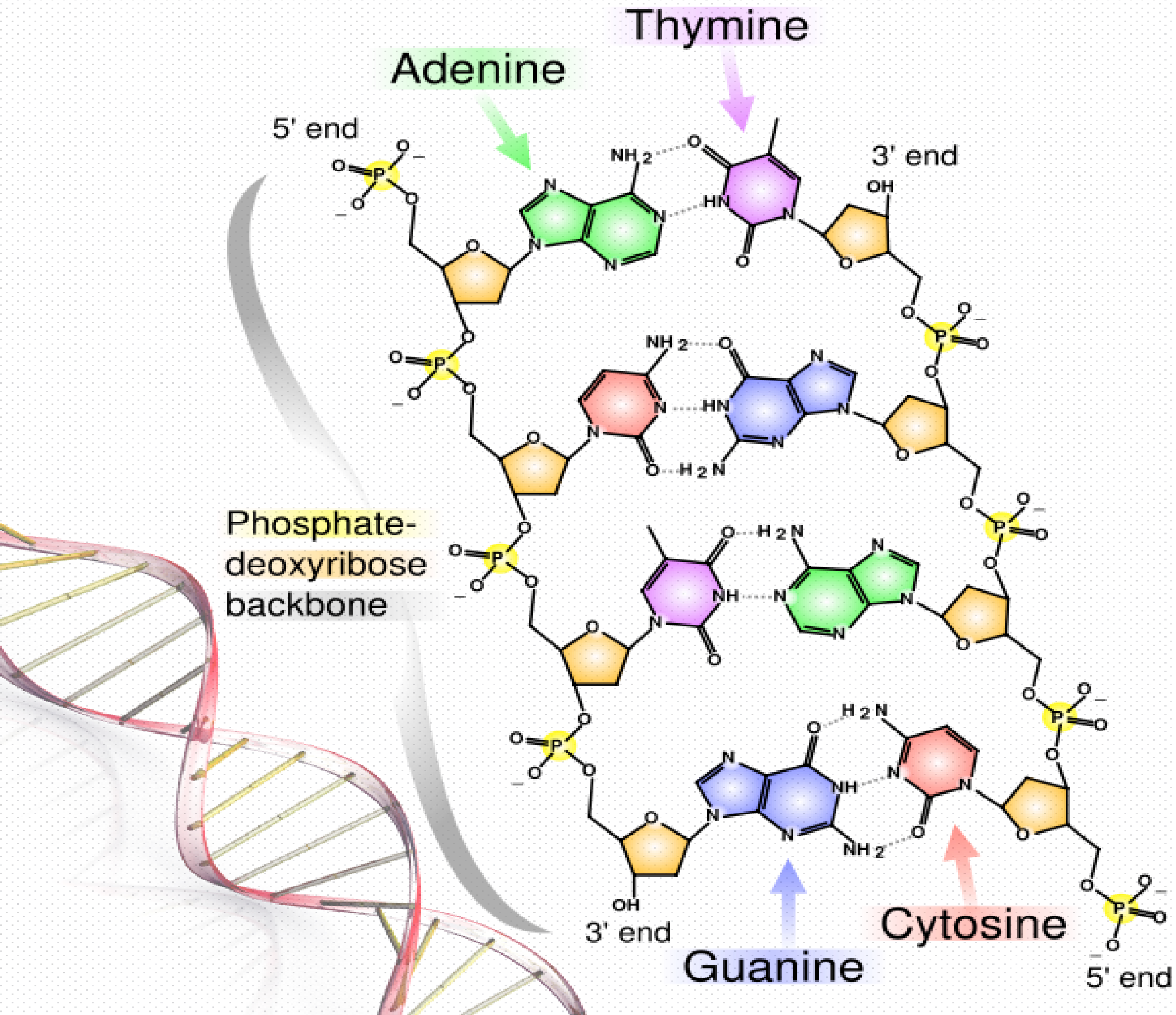
- DNA molecules usually consist of two strands arranged in the famous double helix







DNA strand: polymer of nucleotides



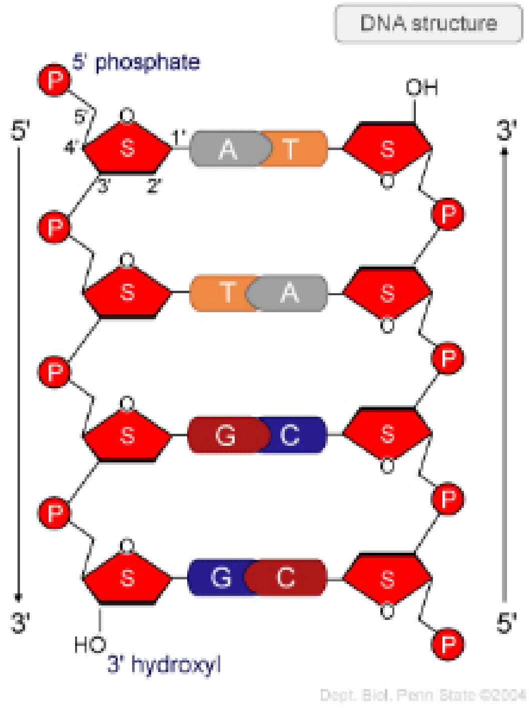
In double-stranded DNA:

A always bonds to T

C always bonds to G

The Double Helix

- each strand of DNA has a "direction"
  - at one end, the terminal carbon atom in the backbone is the 5′ carbon atom of the terminal sugar
  - at the other end, the terminal carbon atom is the 3′ carbon atom of the terminal sugar
- therefore we can talk about the 5′ and the 3′ ends of a DNA strand
- in a double helix, the strands are antiparallel (arrows drawn from the 5′ end to the 3′ end go in opposite directions)



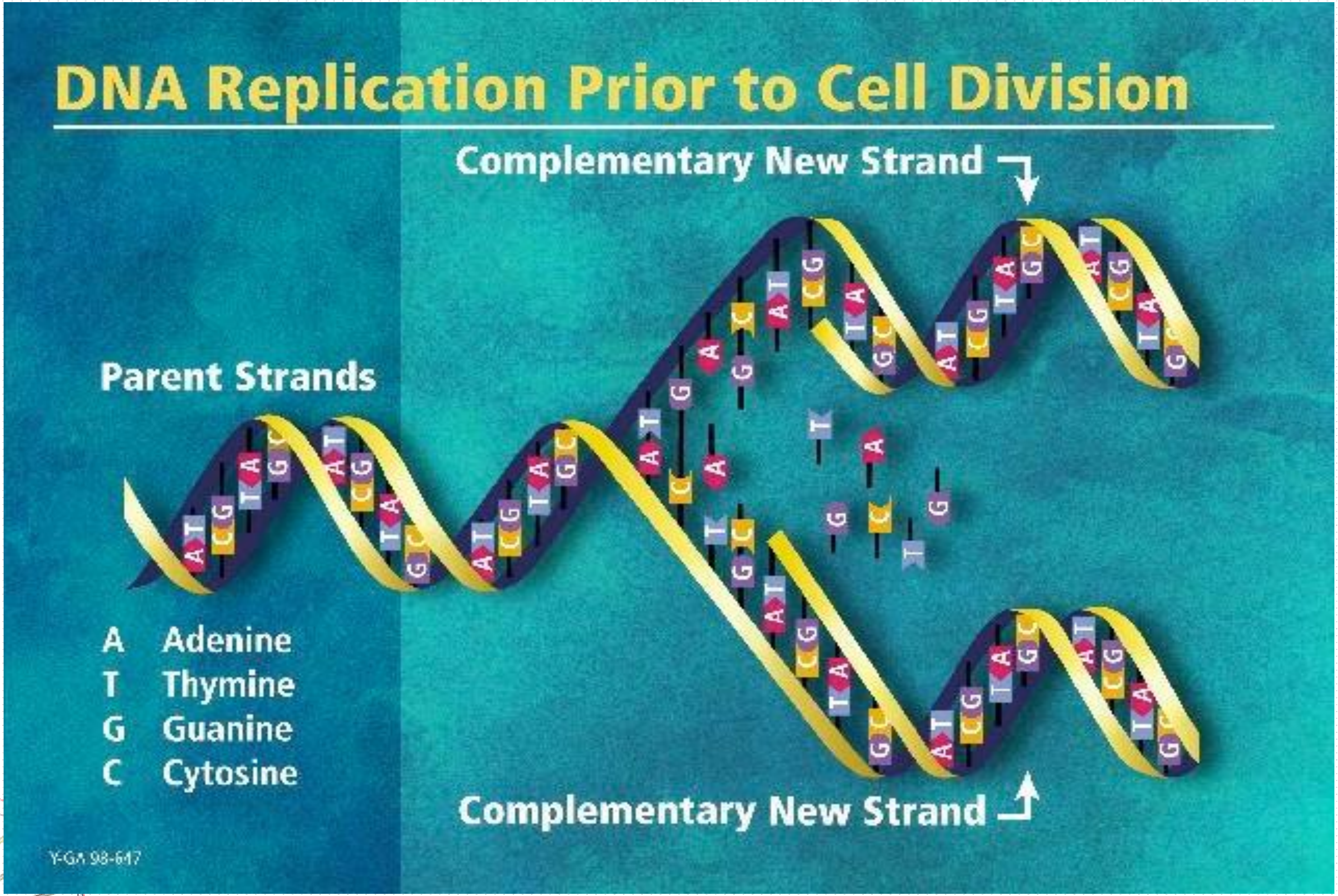


image from the DOE Human Genome Program
http://www.ornl.gov/hgmis

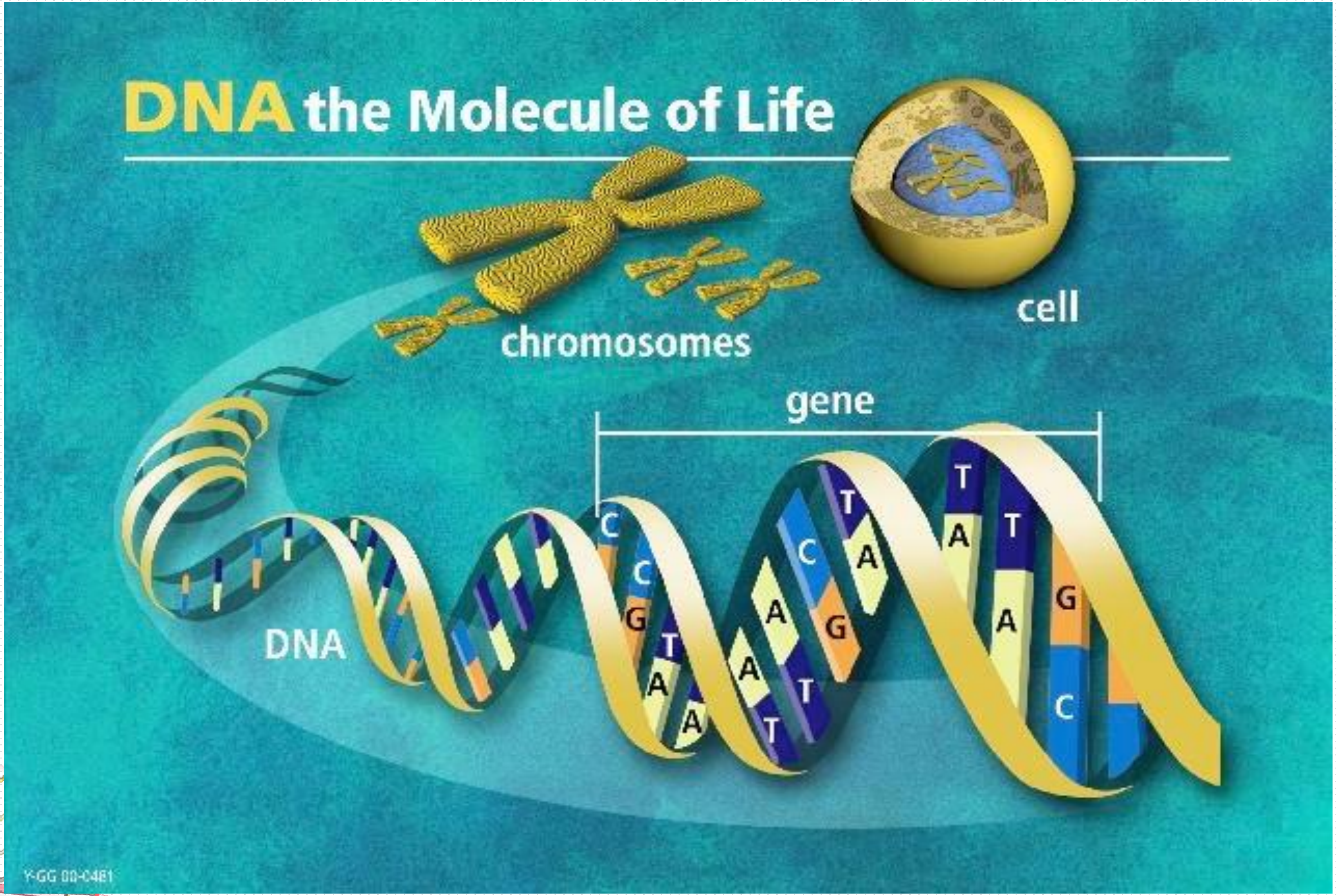


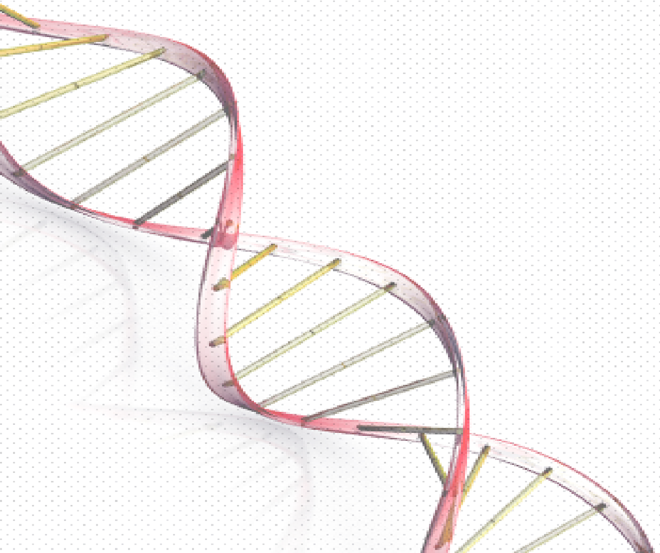
image from the DOE Human Genome Program
http://www.ornl.gov/hgmis

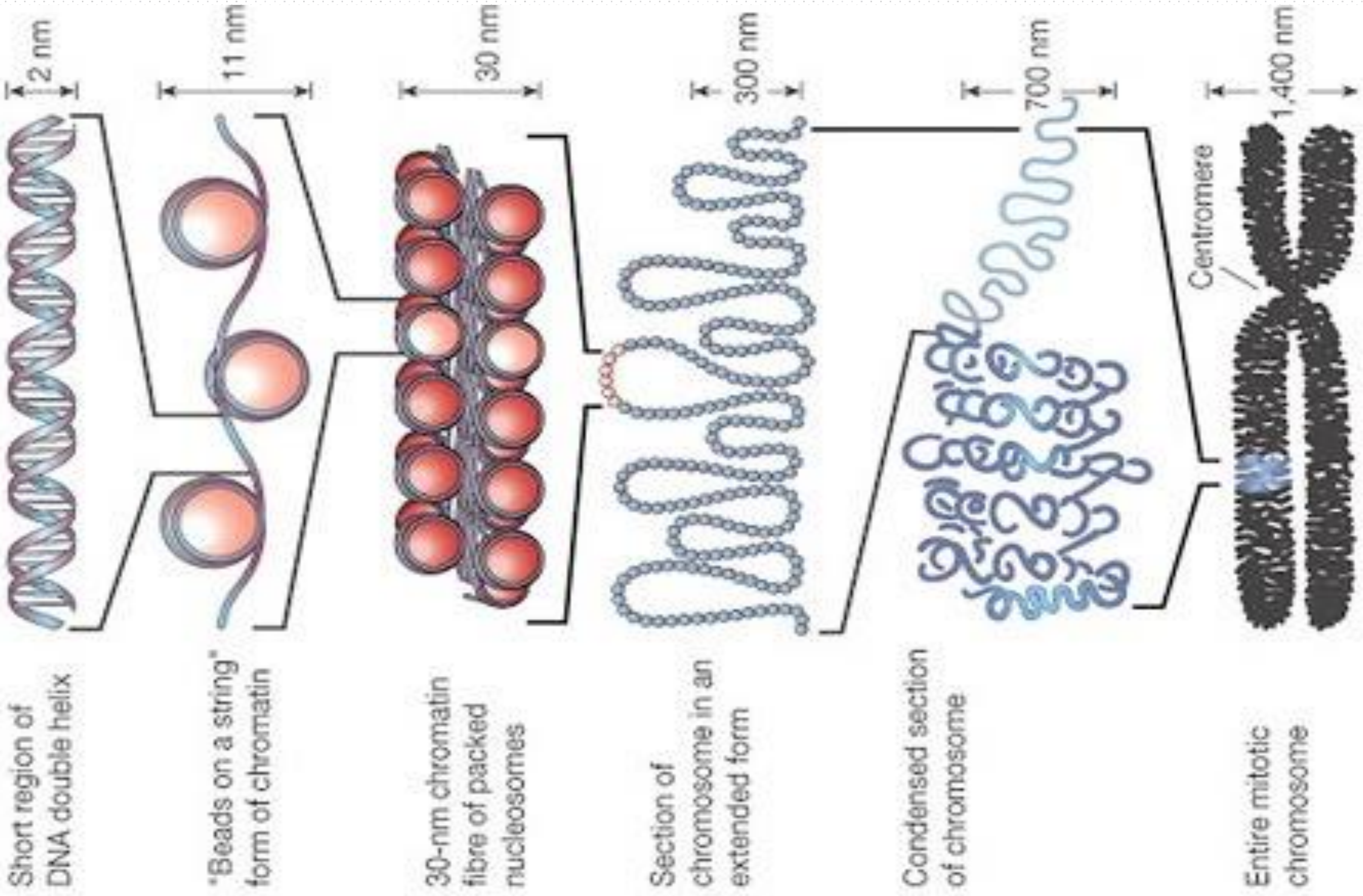
Chromosomes

- DNA is packaged into individual chromosomes (along with proteins)

- prokaryotes (single-celled organisms lacking nuclei) typically have a single circular chromosome

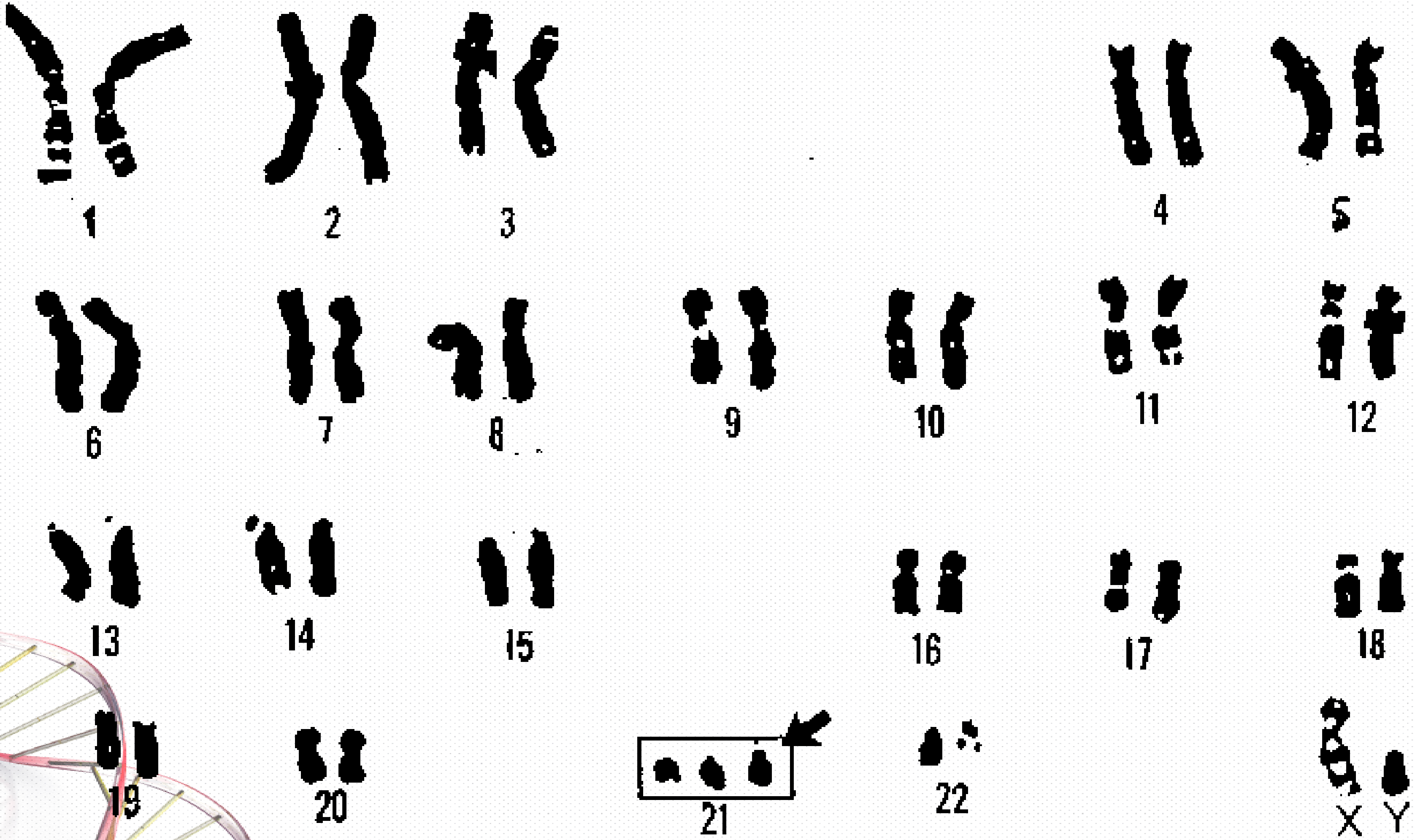
- eukaryotes (organisms with nuclei) have a species-specific number of linear chromosomes




2 nm
Short region of DNA double helix
11 nm
"Beads on a string" form of chromatin
30 nm
30-nm chromatin fibre of packed nucleosomes
300 nm
Section of chromosome in an extended form
700 nm
Condensed section of chromosome
1,400 nm
Centromere
Entire mitotic chromosome

DNA is tightly packed!

Human Chromosomes



Genomes

- the term genome refers to the complete complement of DNA for a given species

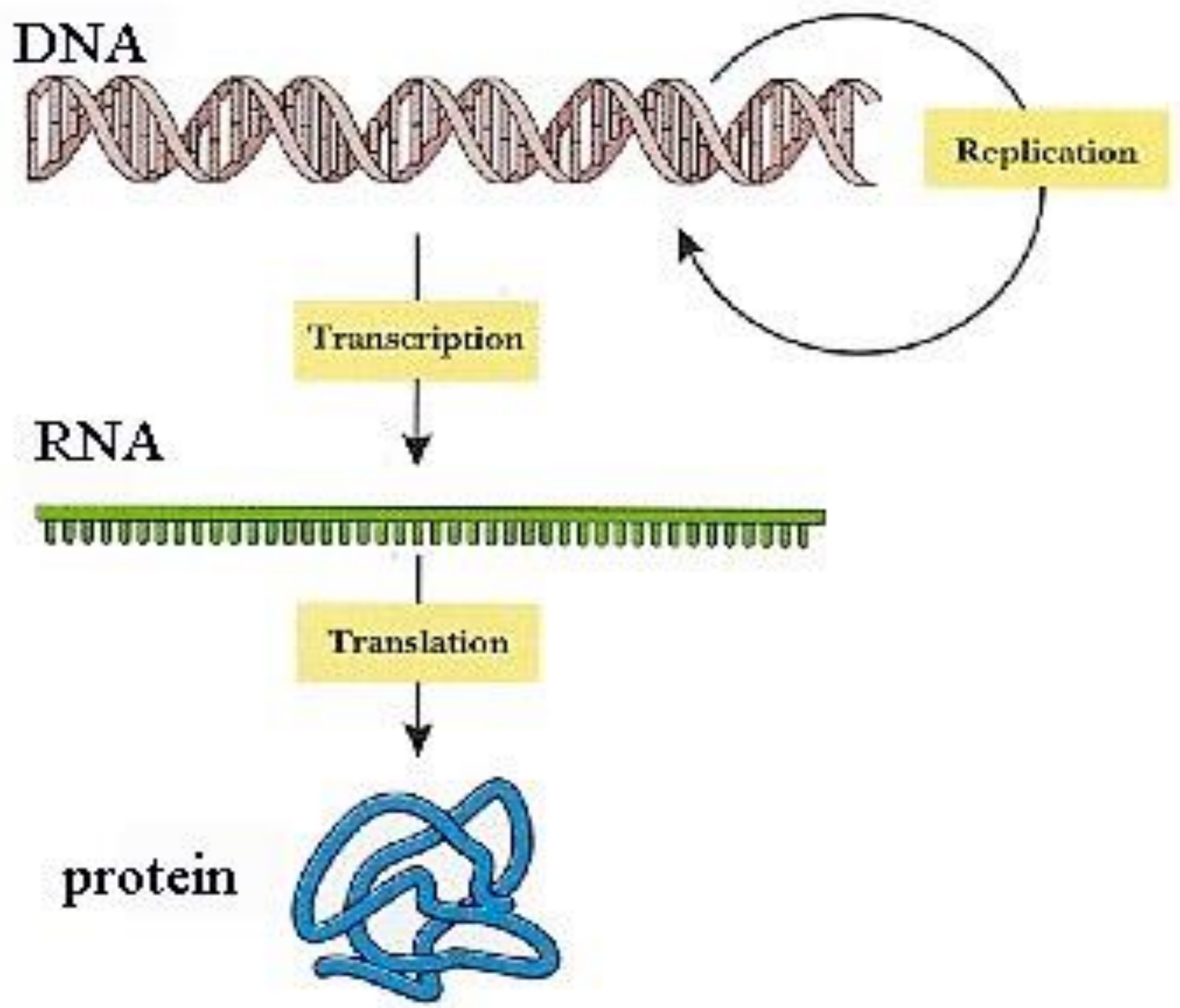
- the human genome consists of 46 chromosomes (23 pairs)

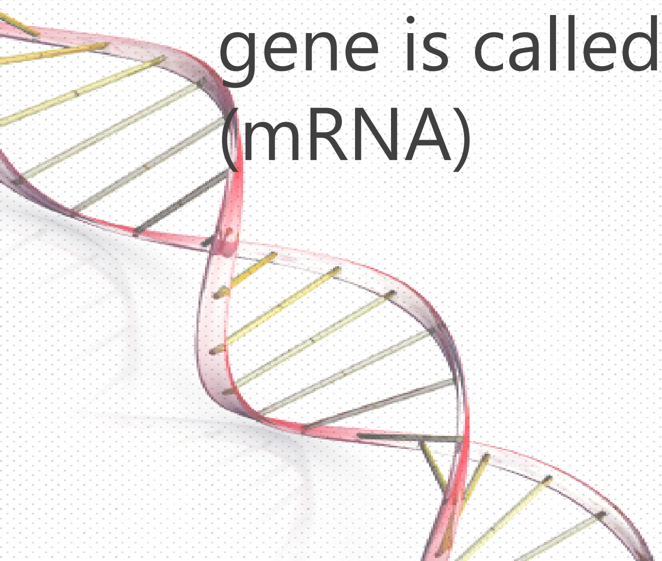
- every cell (except sex cells and mature red blood cells) contains the complete genome of an organism

Transcription

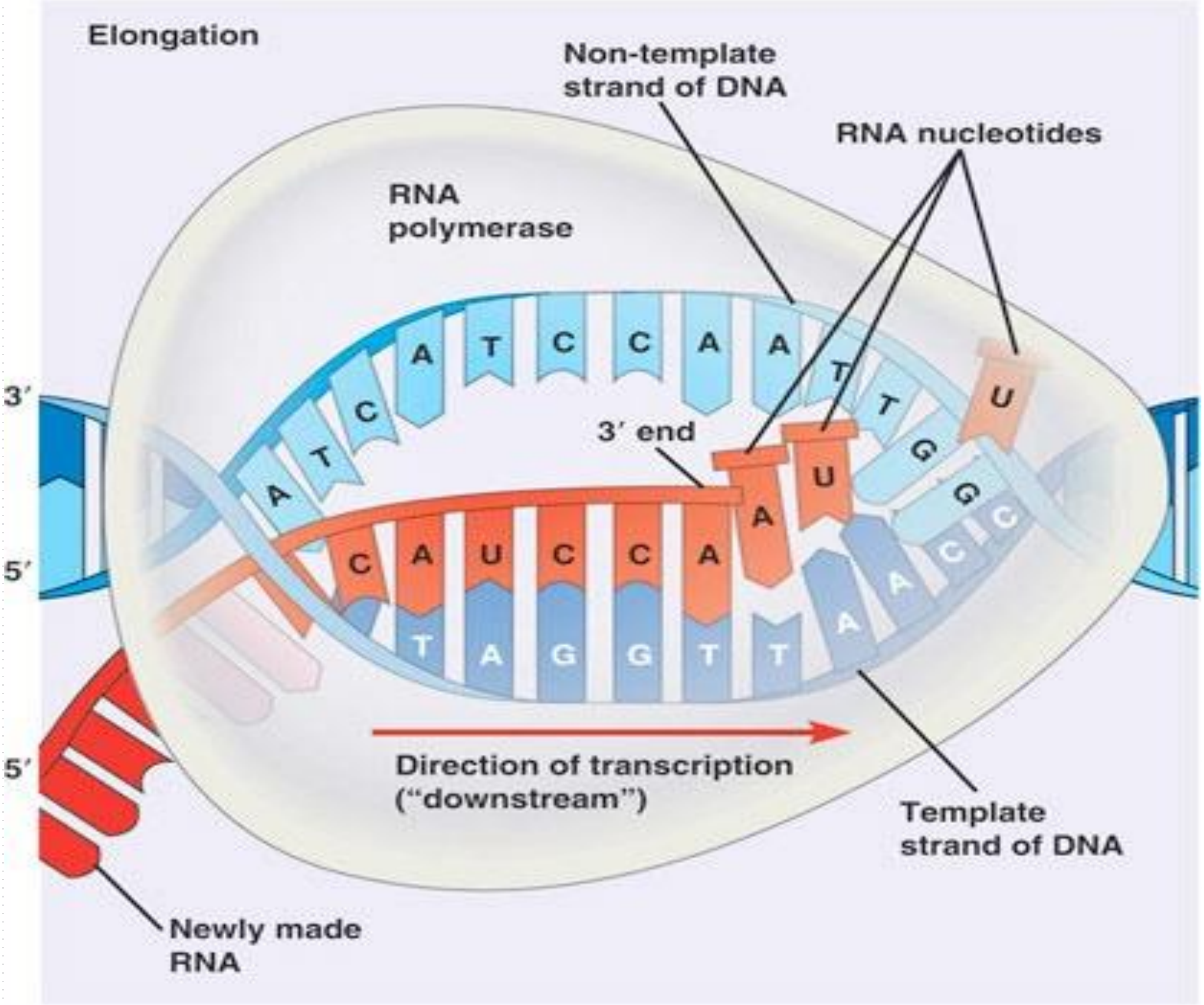
- RNA polymerase is the enzyme that builds an RNA strand from a gene within DNA

- RNA that is transcribed from a gene is called messenger RNA (mRNA)



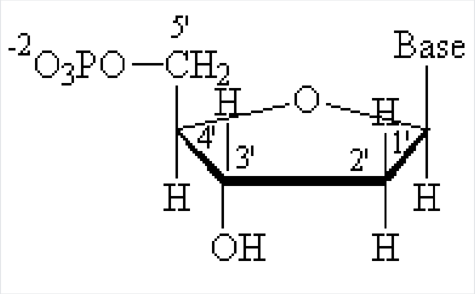
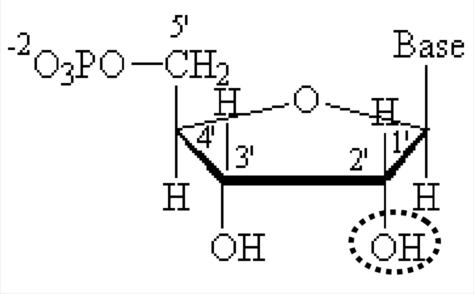


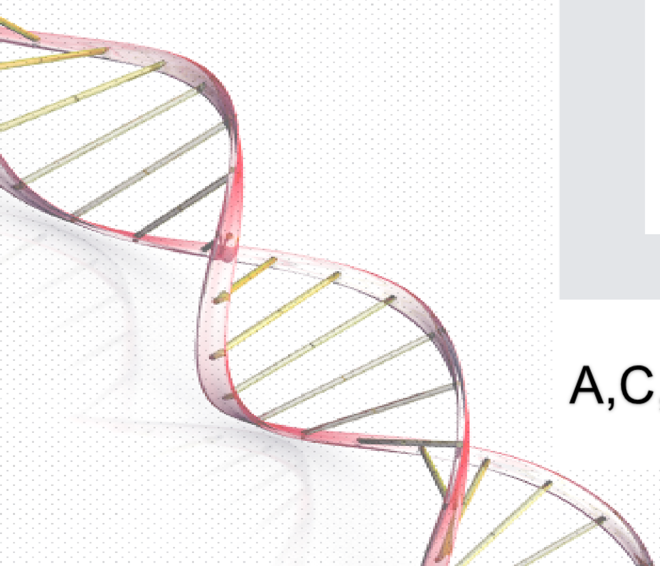
Transcription: DNA→RNA



T is replaced by U
U: Uridine

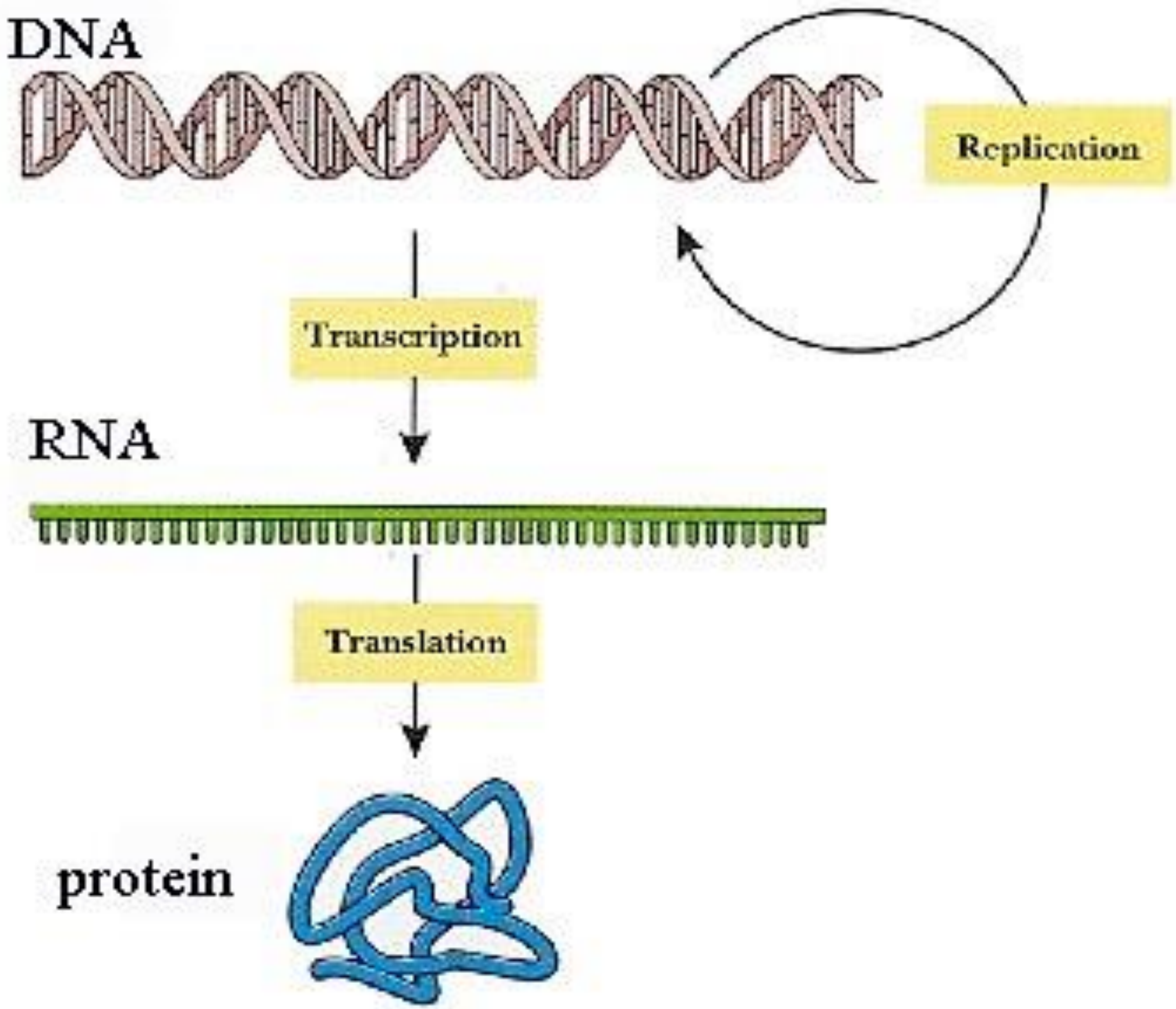
RNA vs DNA structure

DNA	RNA
linear polymer	linear polymer
double-stranded	single-stranded
deoxyribonucleotide monomer  -2 O3PO—CH2 5' Base H O H 4' 3' 2' 1' H H OH H	ribonucleotide monomer  -2 O3PO—CH2 5' Base H O H 4' 3' 2' 1' H H OH OH
A,C,G,T bases	A,C,G,U bases



Translation

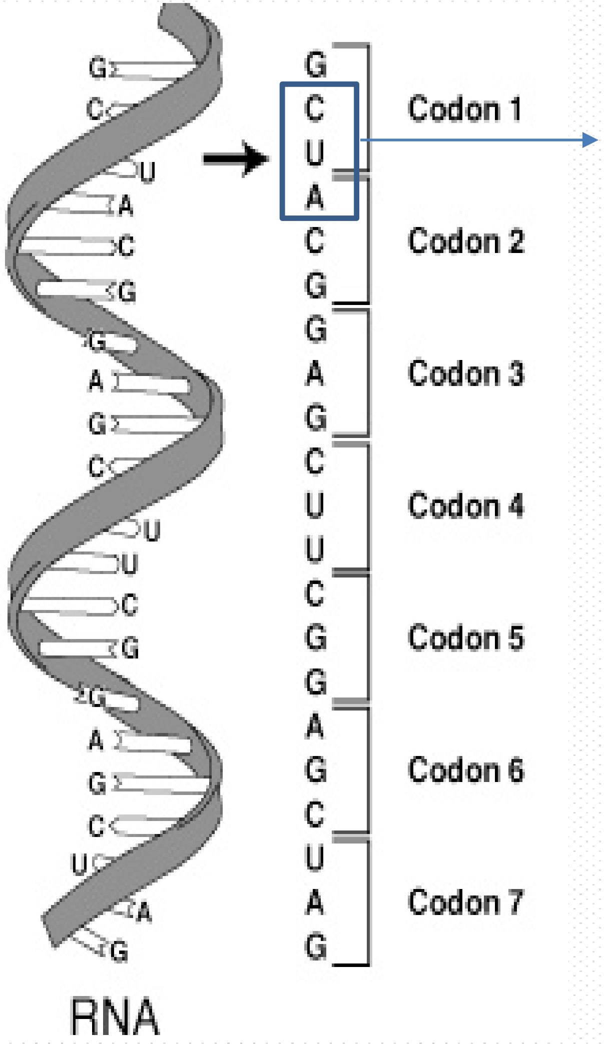
- proteins are molecules composed of one or more polypeptides
- a polypeptide is a polymer composed of amino acids
- cells build their proteins from 20 different amino acids
- a polypeptide can be thought of as a string composed from a 20-character alphabet



The Genetic Code

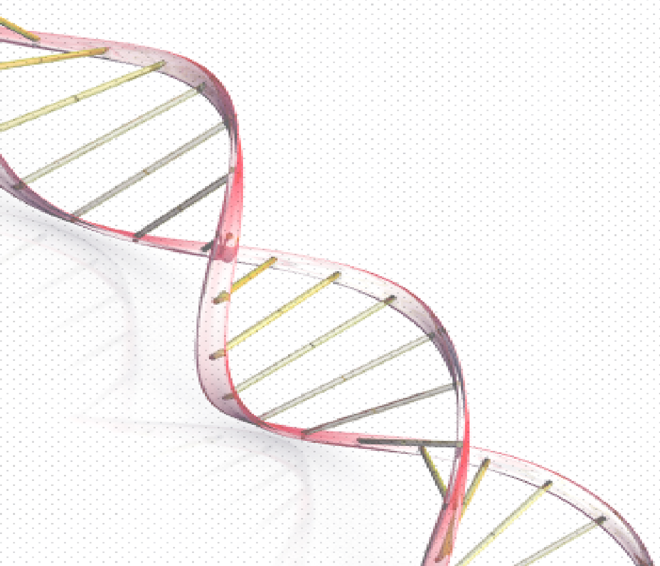
	Second letter				
First letter	U	C	A	G	
U	UUU UUC Phenylalanine UUA UUG Leucine	UCU UCC UCA UCG Serine	UAU UAC Tyrosine UAA Stop codon UAG Stop codon	UGU UGC Cysteine UGA Stop codon UGG Tryptophan	U C A G
C	CUU CUC CUA CUG Leucine	CCU CCC CCA CCG Proline	CAU CAC Histidine CAA CAG Glutamine	CGU CGC CGA 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

Codons and Reading Frames

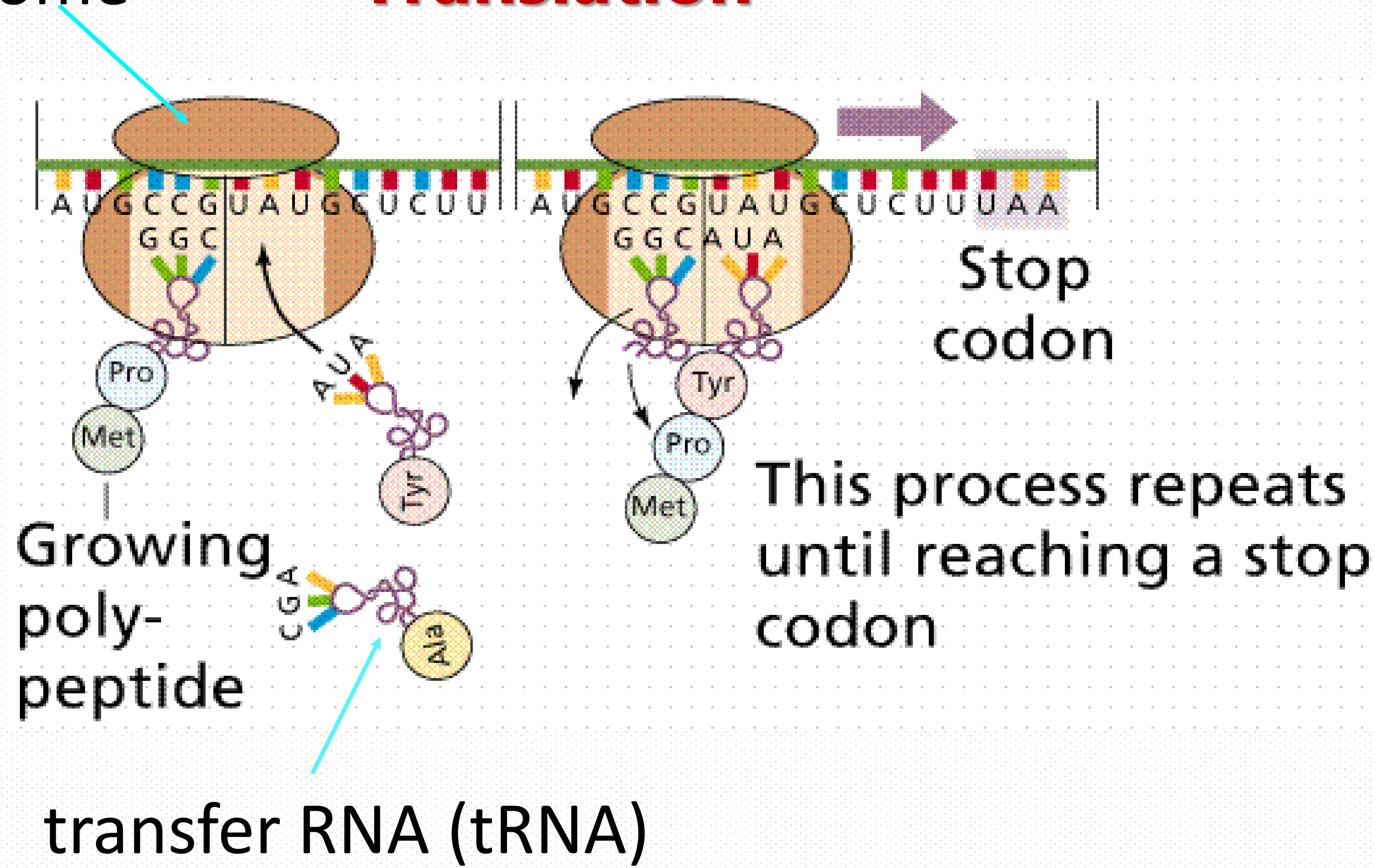


Alt Codon 1

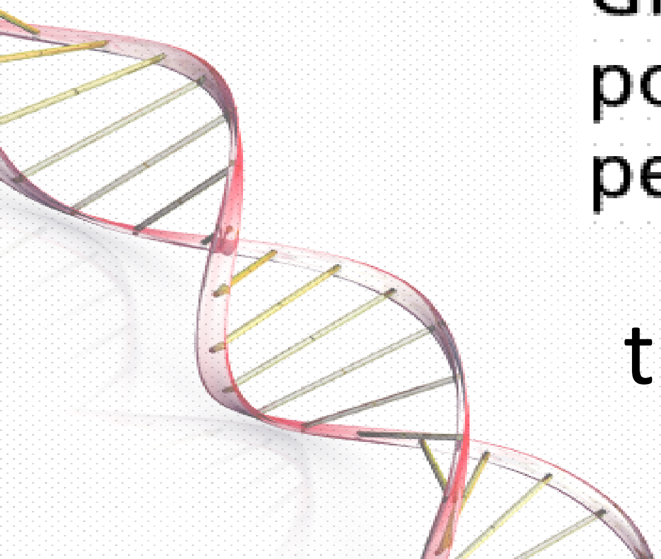
Open reading frames



Translation

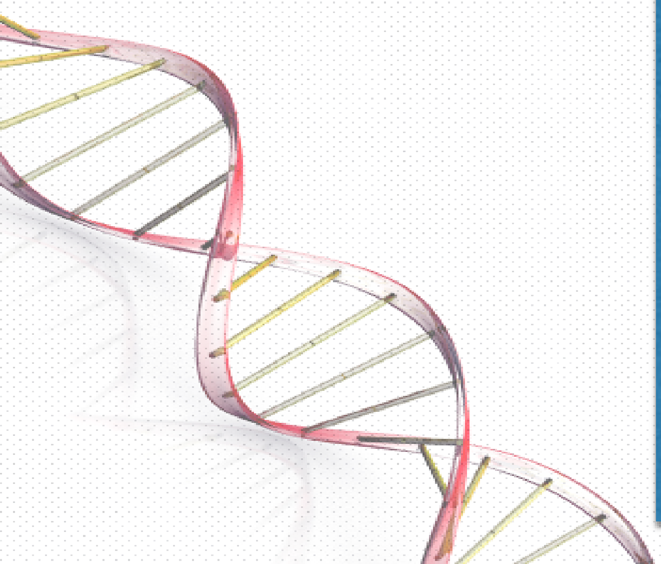


This process repeats until reaching a stop codon

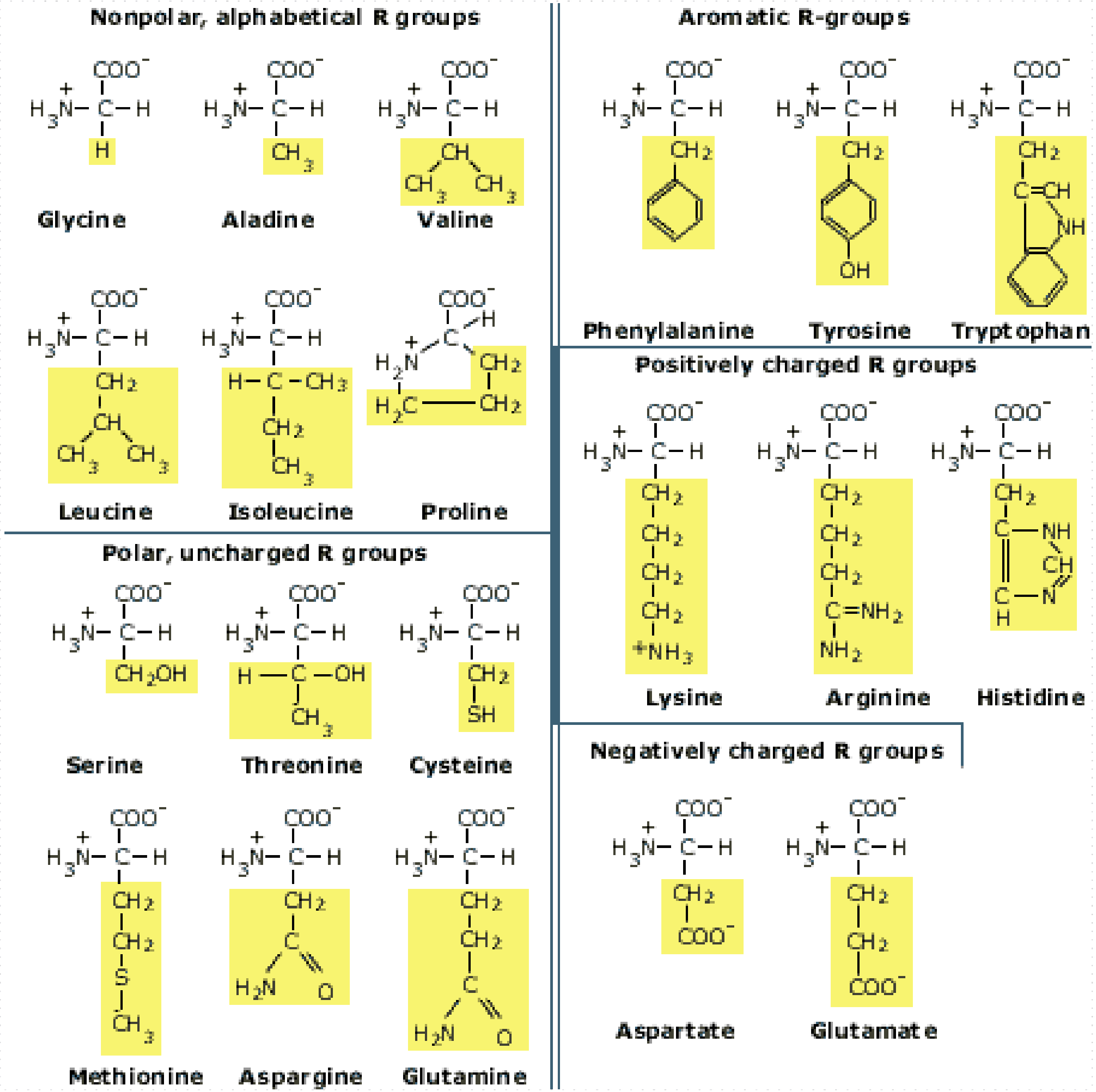


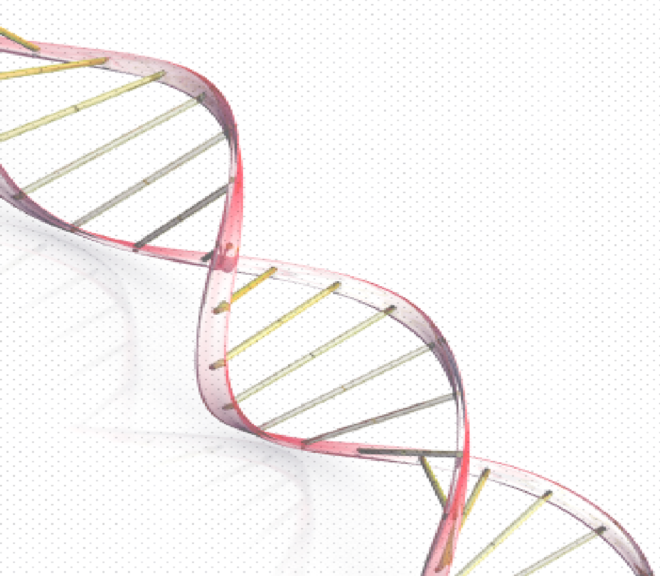
Amino Acids

Alanine	Ala	A
Arginine	Arg	R
Aspartic Acid	Asp	D
Asparagine	Asn	N
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

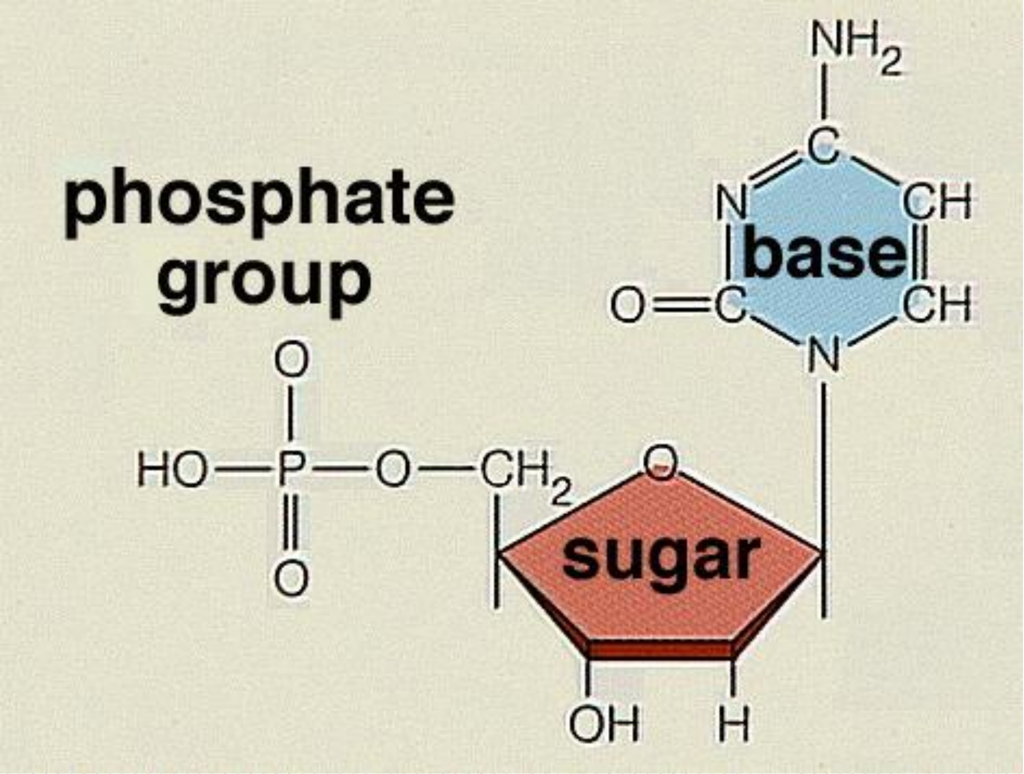


Amino Acids

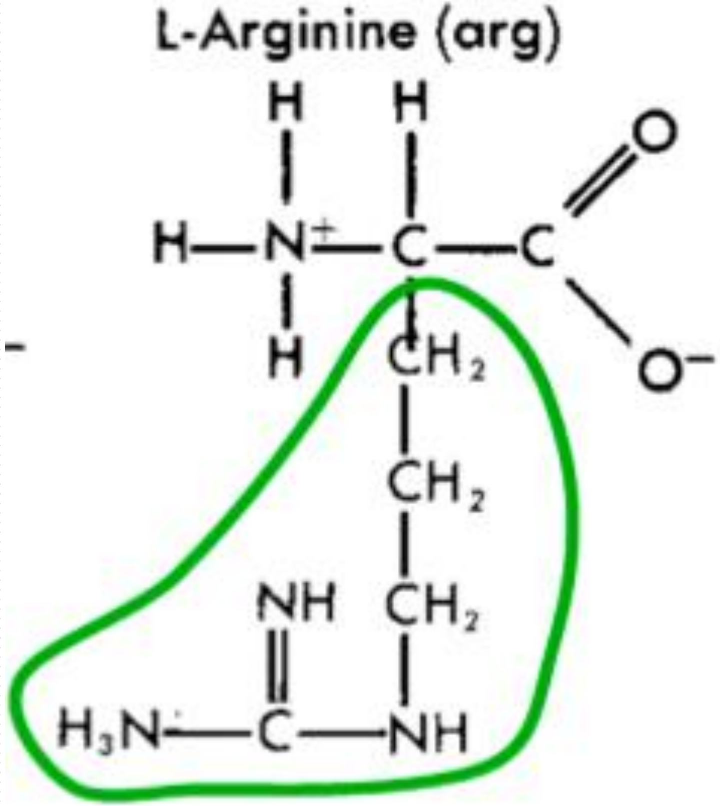

Nonpolar, alphabetical R groups
Glycine
Aladine
Valine
Leucine
Isoleucine
Proline
Polar, uncharged R groups
Serine
Threonine
Cysteine
Methionine
Aspargine
Glutamine
Aromatic R-groups
Phenylalanine
Tyrosine
Tryptophan
Positively charged R groups
Lysine
Arginine
Histidine
Negatively charged R groups
Aspartate
Glutamate



Nucleotides vs Amino Acids



Nucleotide

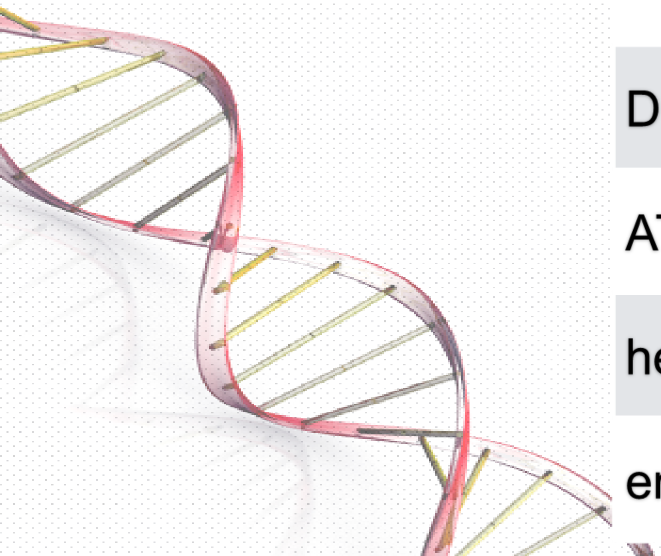


Amino Acid

Both made up of “backbone” and “residue” parts

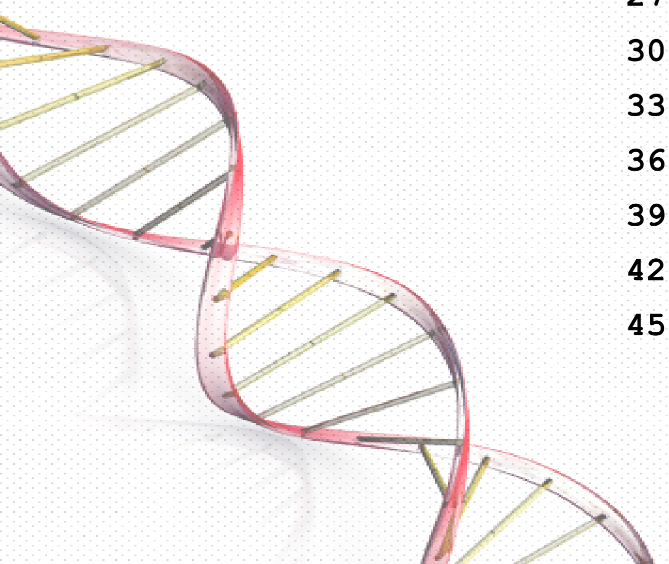
Examples of proteins

Protein	Role
alpha-keratin	component of hair
beta-keratin	component of scales
insulin	regulates blood glucose level
actin & myosin	muscle contraction
DNA polymerase	synthesis of DNA
ATP synthase	makes ATP
hemoglobin	transport of oxygen
endonuclease	cuts DNA (restriction enzyme)

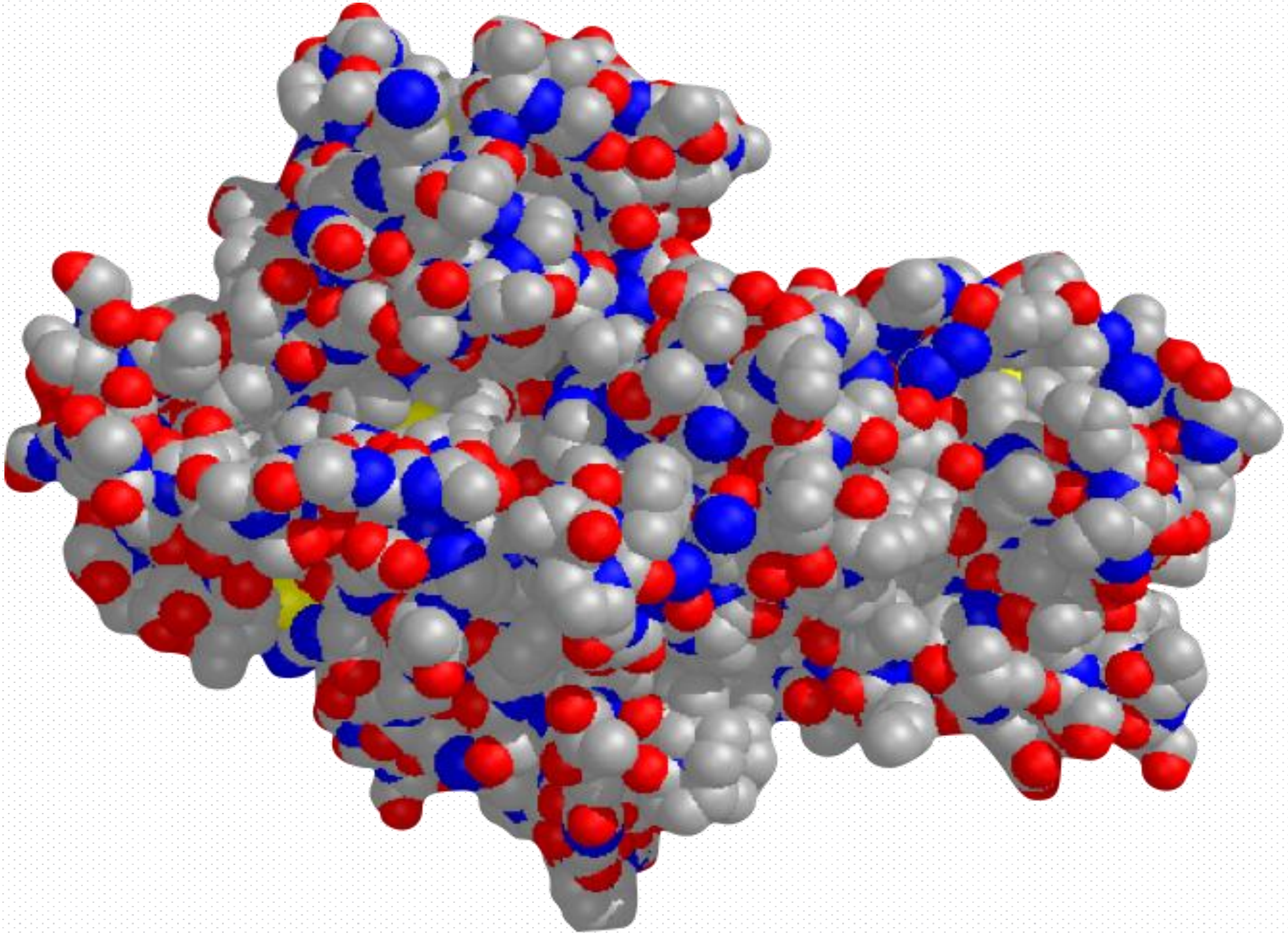


Amino Acid Sequence of Hexokinase

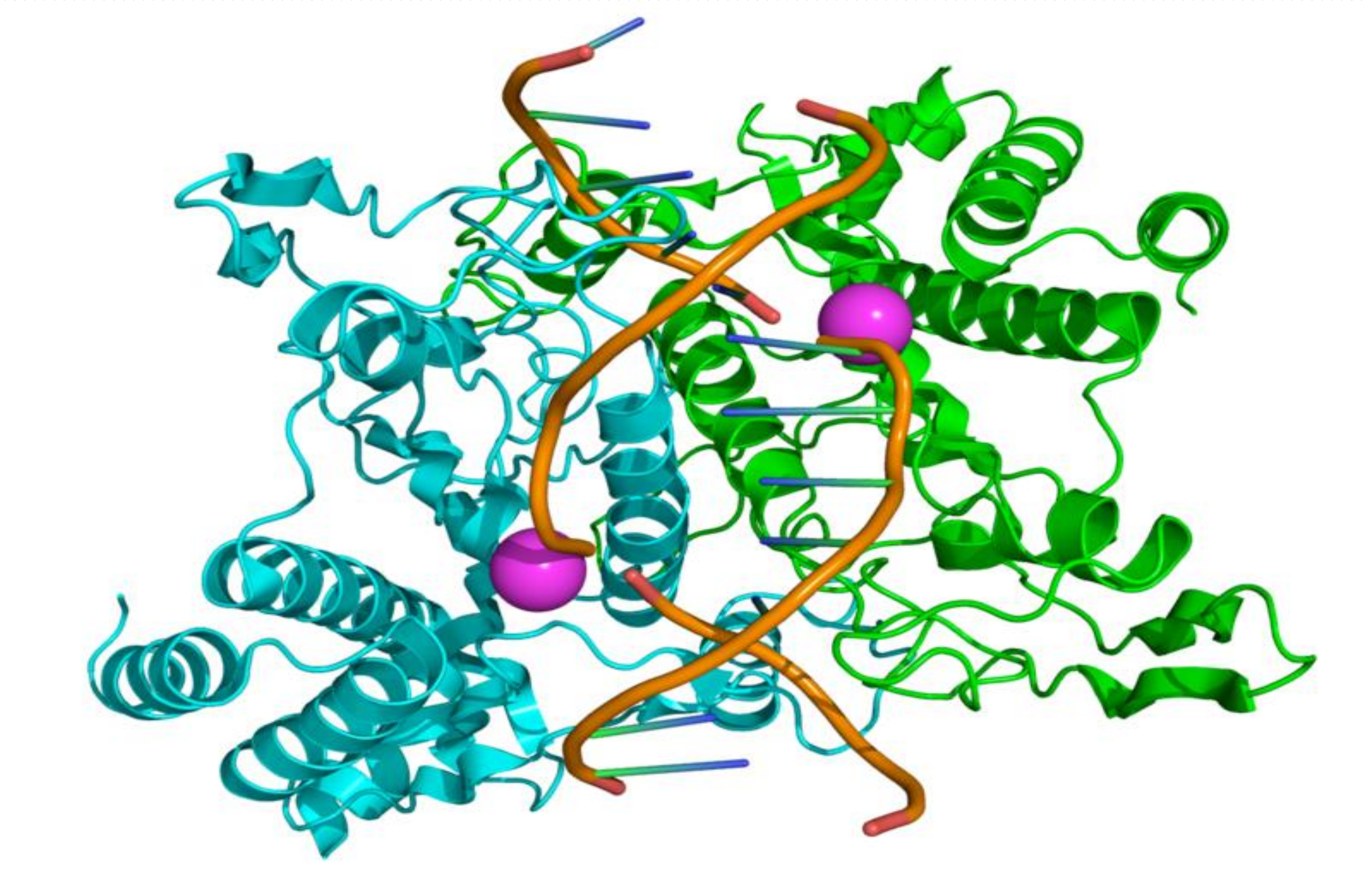
```
  1 A A S X D X S L V E V H X X V F I V P P X I L Q A V V S I A
 31 T T R X D D X D S A A A S I P M V P G W V L K Q V X G S Q A
 61 G S F L A I V M G G G D L E V I L I X L A G Y Q E S S I X A
 91 S R S L A A S M X T T A I P S D L W G N X A X S N A A F S S
121 X E F S S X A G S V P L G F T F X E A G A K E X V I K G Q I
151 T X Q A X A F S L A X L X K L I S A M X N A X F P A G D X X
181 X X V A D I X D S H G I L X X V N Y T D A X I K M G I I F G
211 S G V N A A Y W C D S T X I A D A A D A G X X G G A G X M X
241 V C C X Q D S F R K A F P S L P Q I X Y X X T L N X X S P X
271 A X K T F E K N S X A K N X G Q S L R D V L M X Y K X X G Q
301 X H X X X A X D F X A A N V E N S S Y P A K I Q K L P H F D
331 L R X X X D L F X G D Q G I A X K T X M K X V V R R X L F L
361 I A A Y A F R L V V C X I X A I C Q K K G Y S S G H I A A X
391 G S X R D Y S G F S X N S A T X N X N I Y G W P Q S A X X S
421 K P I X I T P A I D G E G A A X X V I X S I A S S Q X X X A
451 X X S A X X A
```



Space-Filling Model of Hexokinase



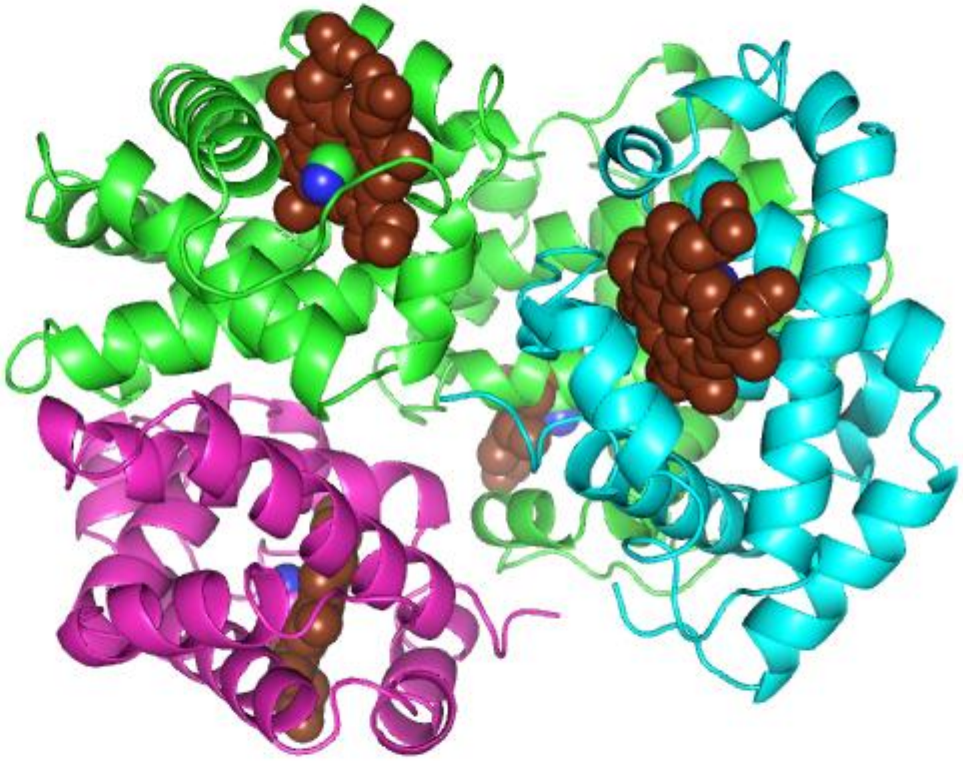
EcoRI – restriction enzyme



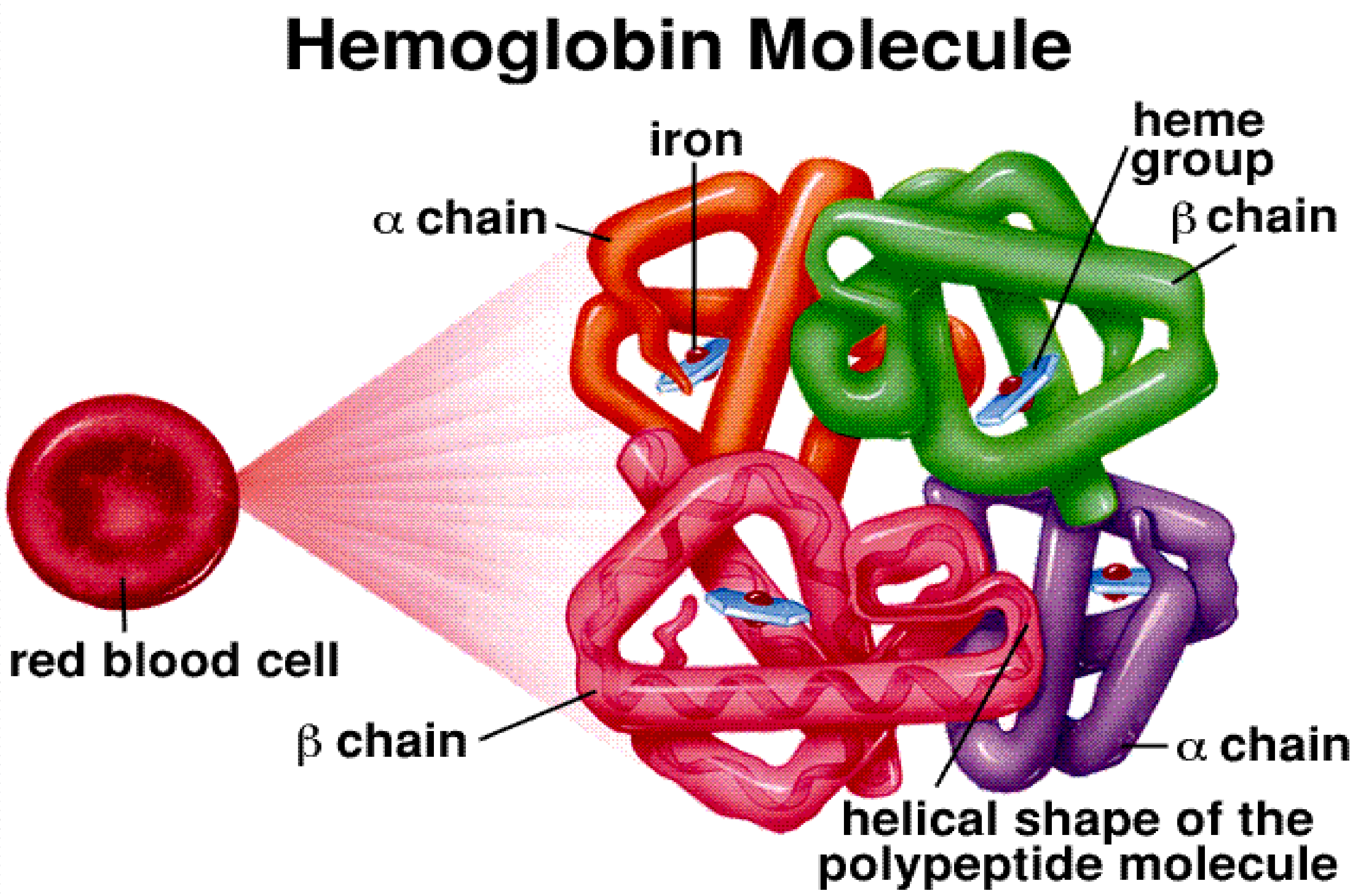
Hemoglobin

- protein built from 4 polypeptides

- responsible for carrying oxygen in red blood cells



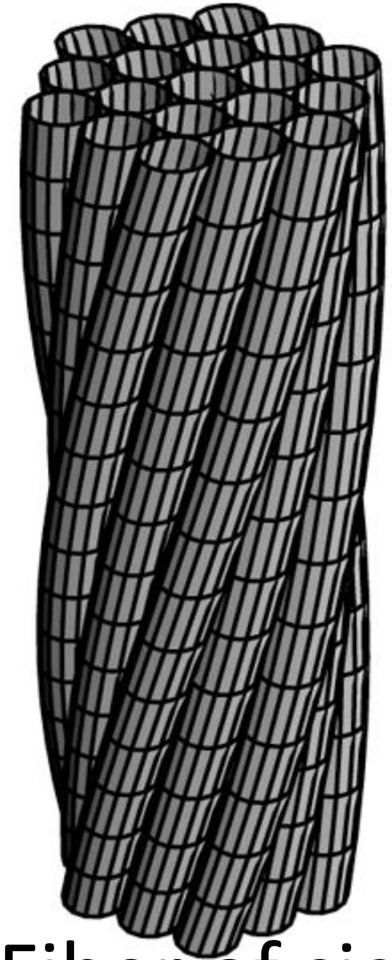
Hemoglobin: carrier of oxygen



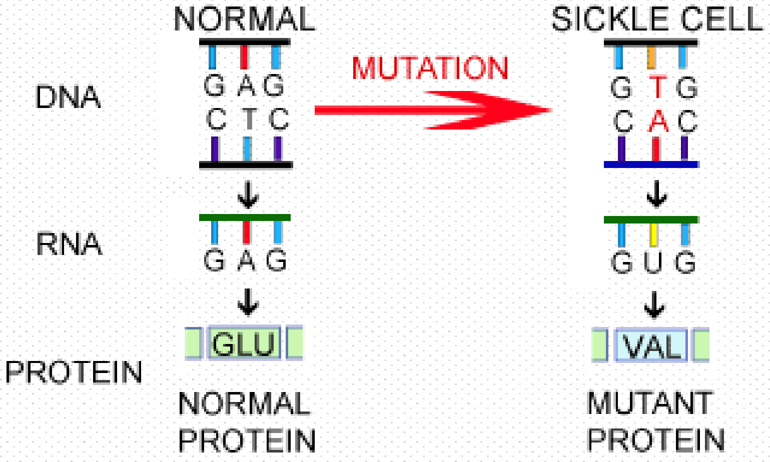
Sylvia S. Mader, Inquiry into Life, 8th edition. Copyright © 1997 The McGraw-Hill Companies, Inc. All rights reserved.

Hemoglobin Molecule

Mutant β-globin → Sickle blood cells



Fiber of sickle hemoglobin



NORMAL

SICKLE CELL

MUTATION

DNA

G A G
C T C

G T G
C A C

RNA

G A G

G U G

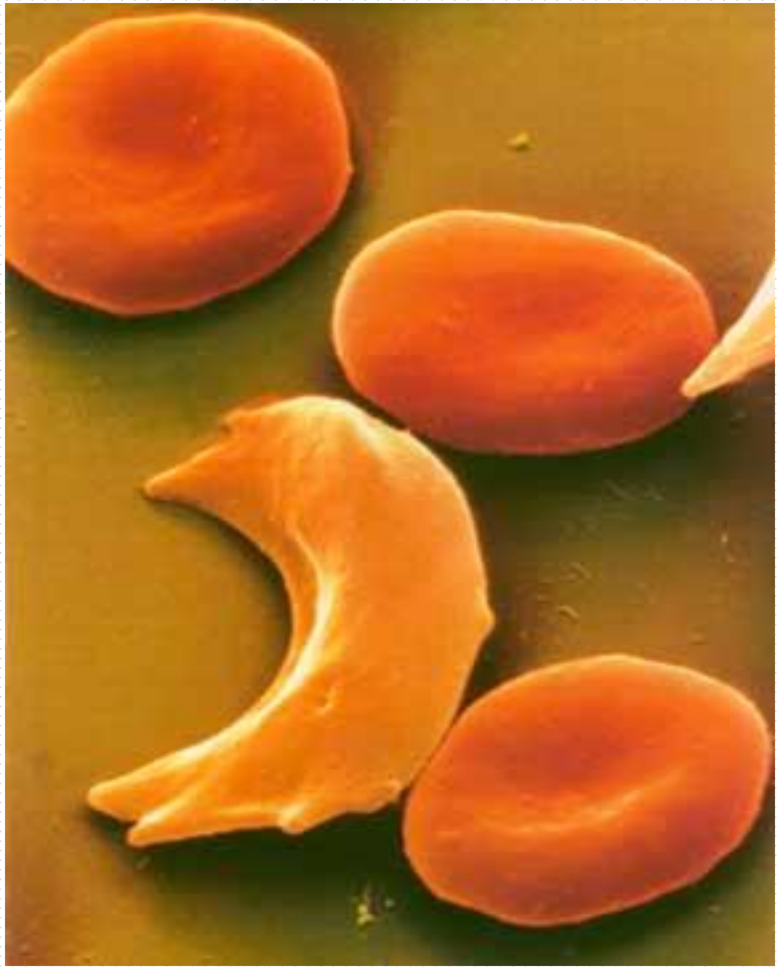
PROTEIN

GLU

VAL

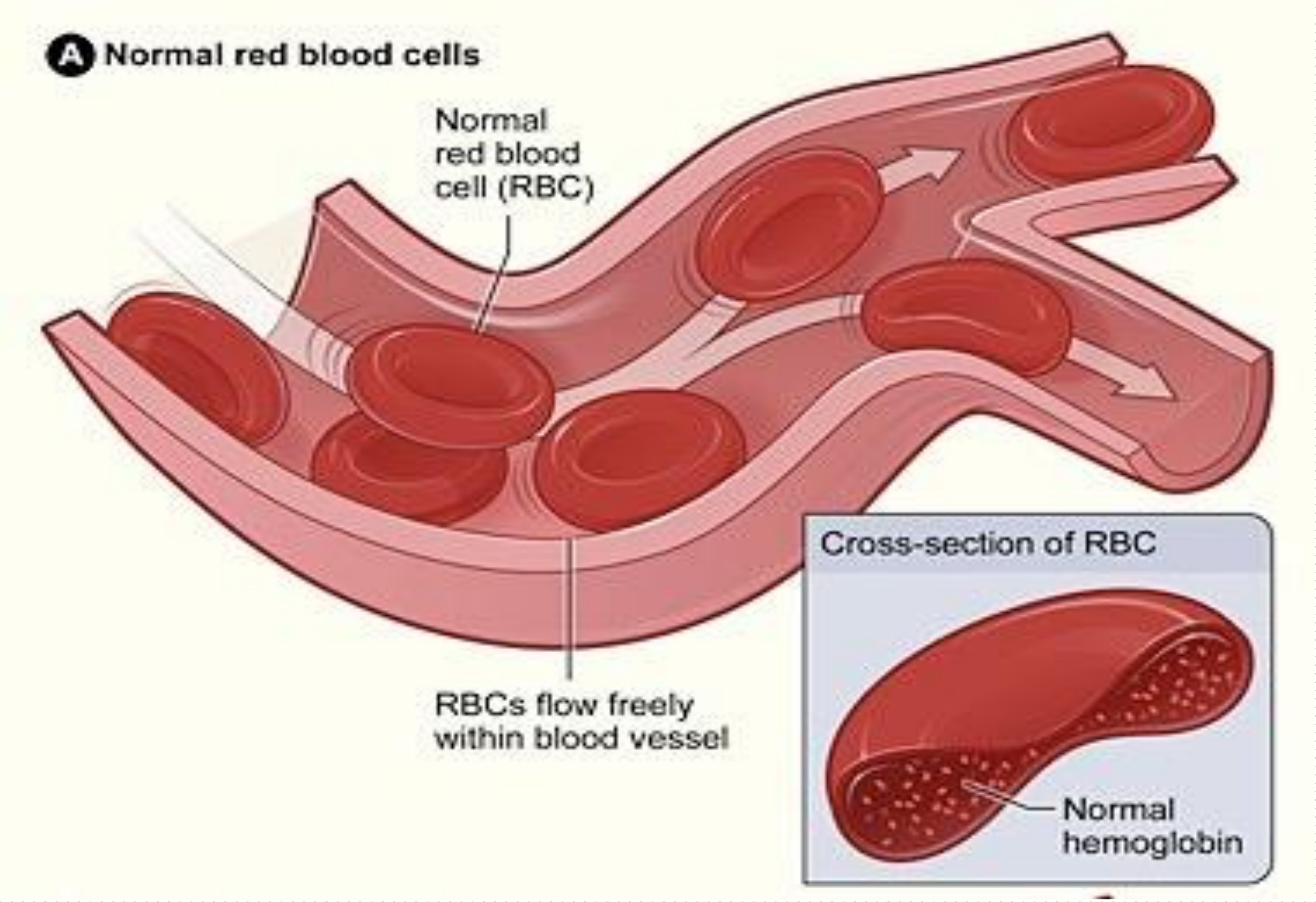
NORMAL PROTEIN

MUTANT PROTEIN

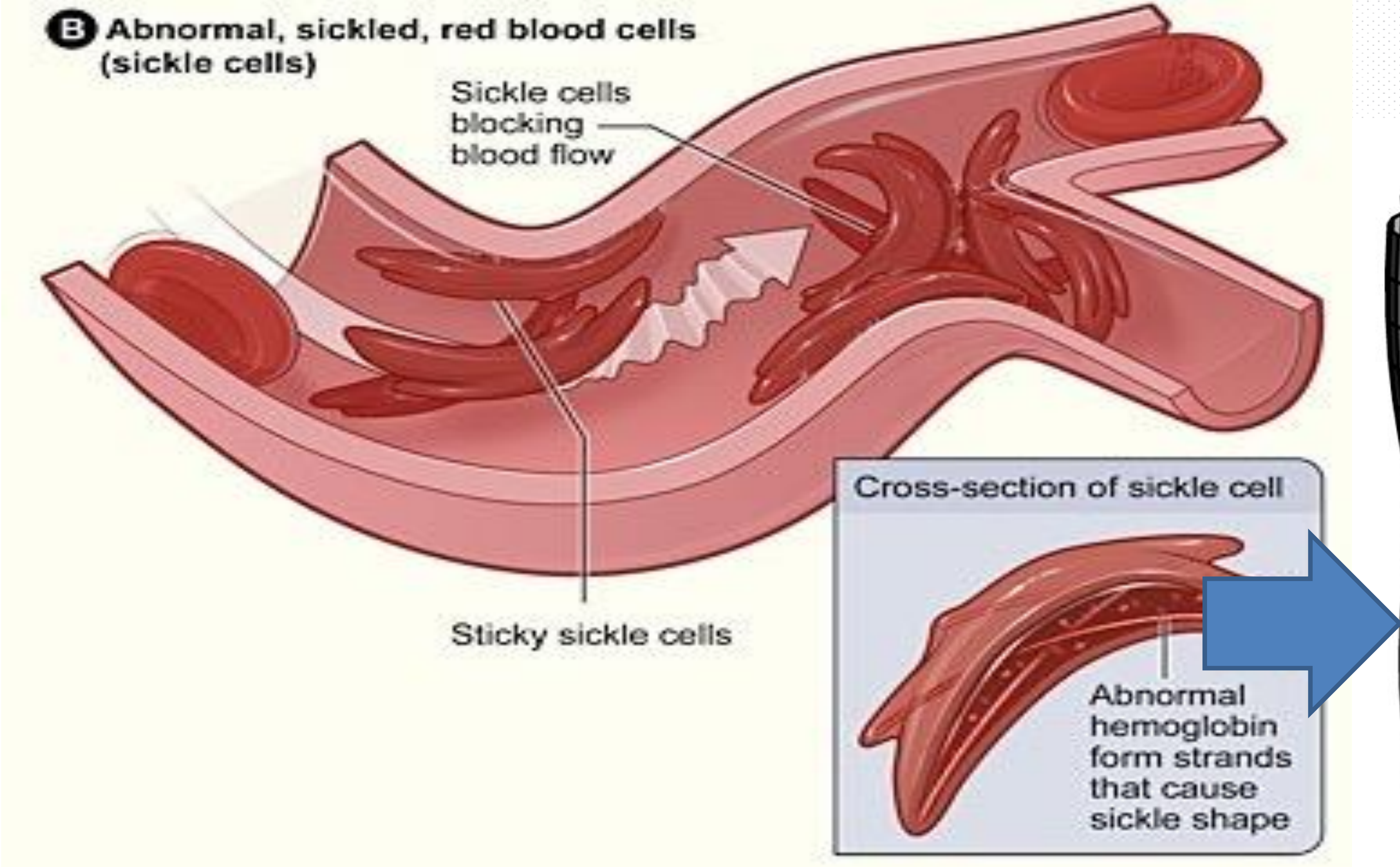


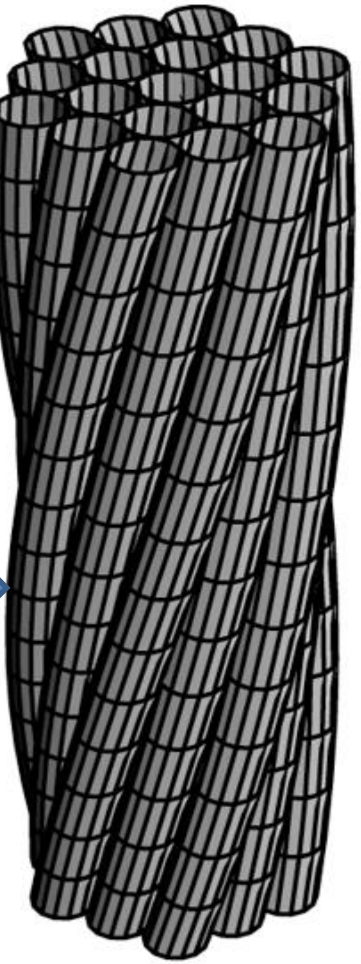
Sickle and normal blood cells

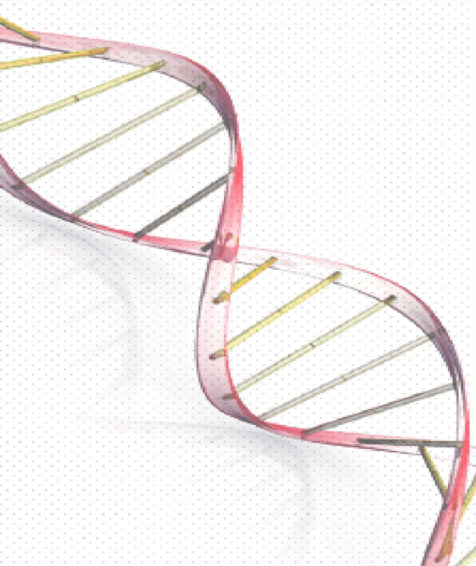
Normal blood flow


A Normal red blood cells
Normal red blood cell (RBC)
RBCs flow freely within blood vessel
Cross-section of RBC
Normal hemoglobin

Sickle cell complications


B Abnormal, sickled, red blood cells (sickle cells)
Sickle cells blocking blood flow
Sticky sickle cells
Cross-section of sickle cell
Abnormal hemoglobin form strands that cause sickle shape





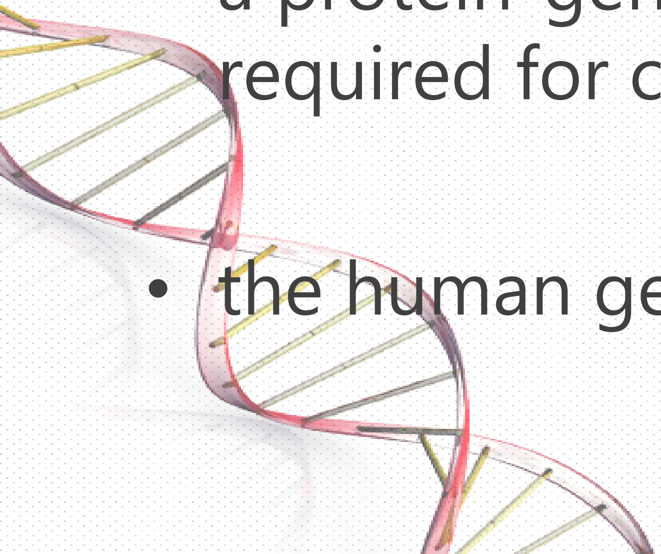
Genes

- genes are the basic units of heredity

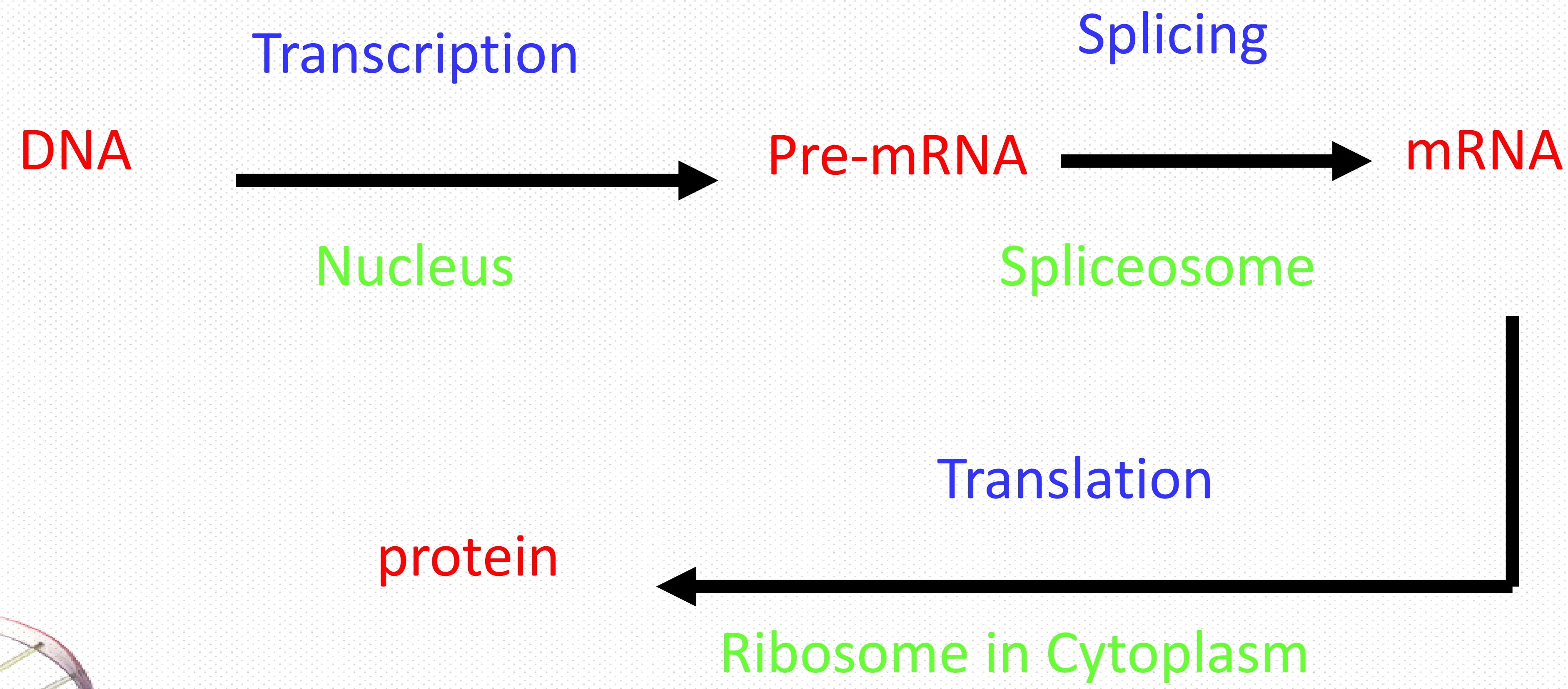
- they are generally the intervals of the genome that are transcribed into RNA

- a protein-gene is a gene whose RNA carries the information required for constructing a particular protein (polypeptide really)

- the human genome comprises ~30,000 protein-coding genes



Central Dogma Revisited



Transcription

DNA

Nucleus

Splicing

Pre-mRNA

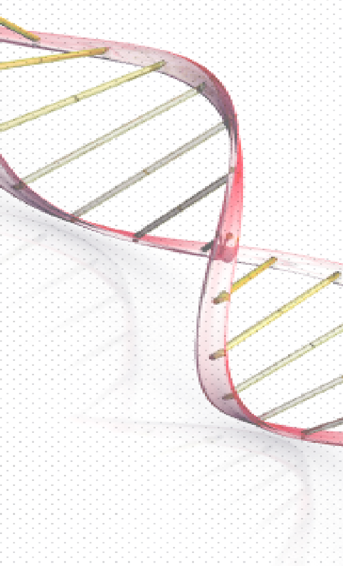
Spliceosome

mRNA

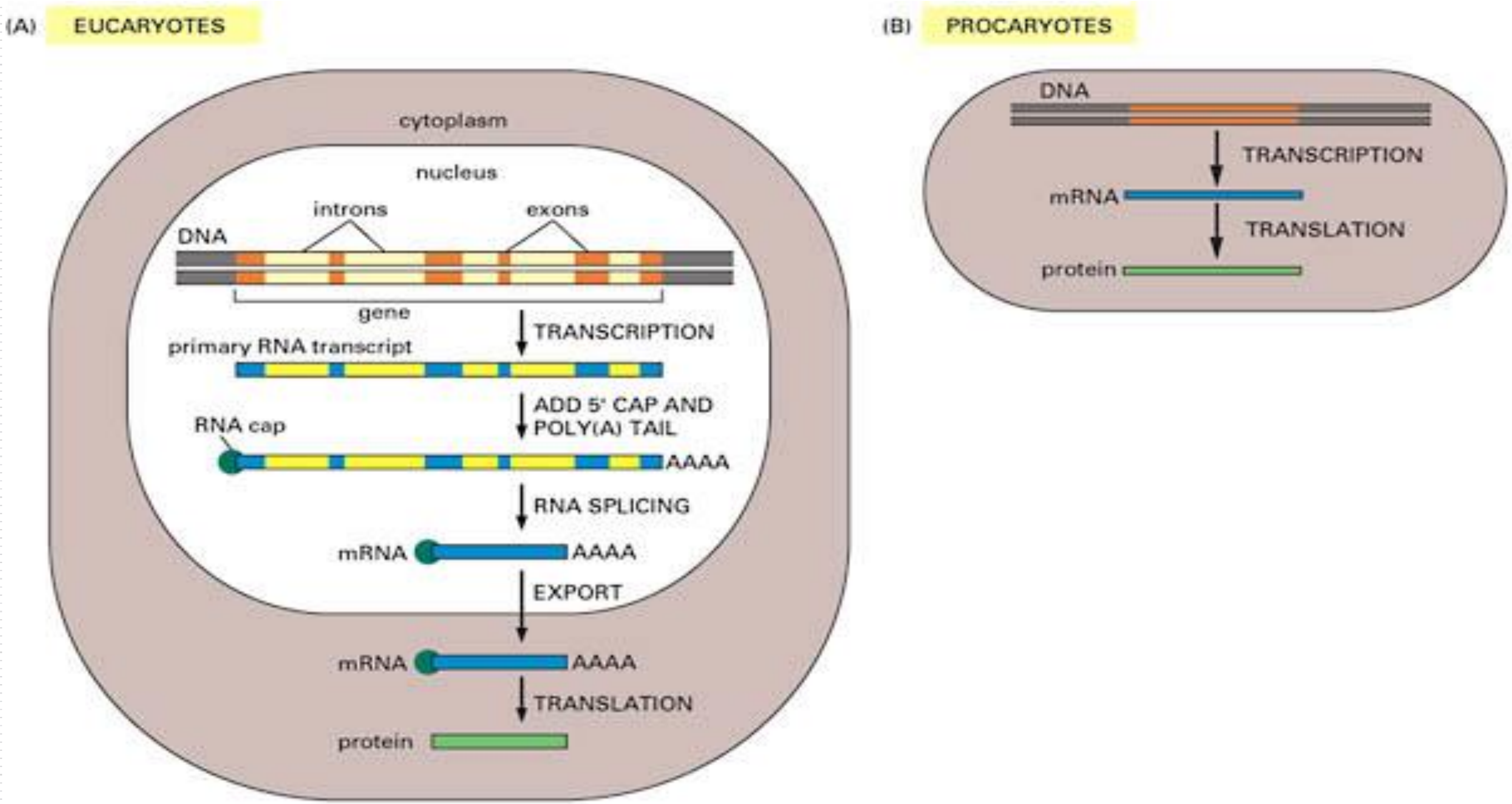
Translation

protein

Ribosome in Cytoplasm

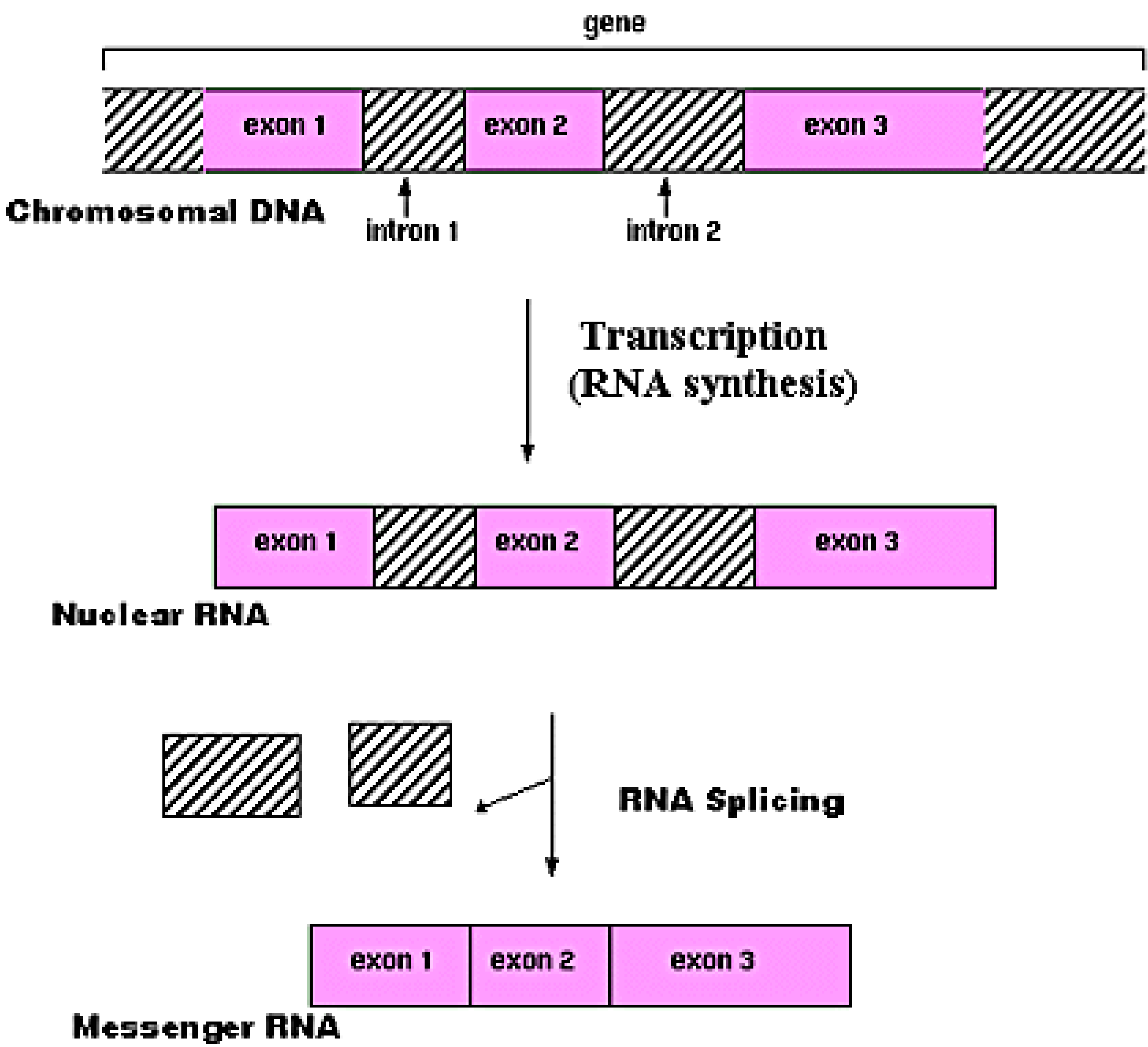


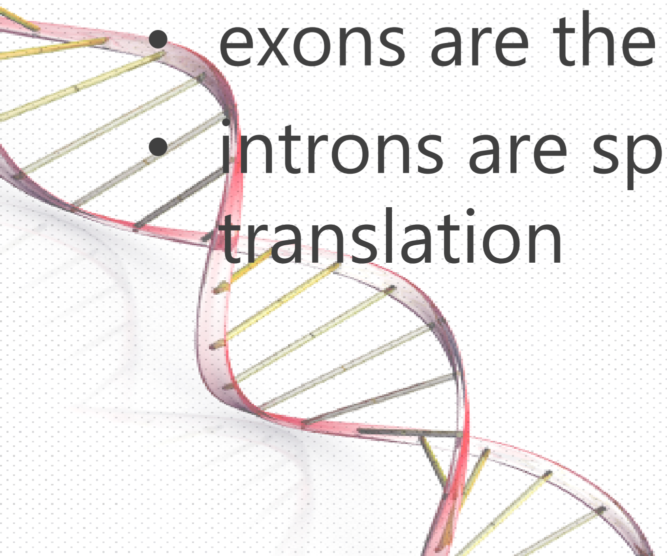
Splicing


(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

- eukaryotes are organisms that have enclosed nuclei in their cells
- in many eukaryotes, genes/mRNAs consist of alternating exon/intron segments
- exons are the coding parts
- introns are spliced out before translation





Alternate Splicing

The Dynamics of Cells

- all cells in an organism have the same genomic data, but the genes expressed in each vary according to cell type, time, and environmental factors

- there are networks of interactions among various biochemical entities in a cell (DNA, RNA, protein, small molecules) that carry out processes such as

  - metabolism

  - intra-cellular and inter-cellular signaling

  - regulation of gene expression

Selected milestones

	Year	Common Name	Species	# of Chromosomes	Size (base pairs)
	1995	Bacterium	Haemophilus influenzae	1	1.8×10^6
	1996	Yeast	Saccharomyces cerevisiae	16	1.2×10^7
	1998	Worm	Caenorhabditis elegans	6	1.0×10^8
	1999	Fruit Fly	Drosophila melanogaster	4	1.3×10^8
	2000	Human	Homo sapiens	23	3.1×10^9
	2002	Mouse	Mus musculus	20	2.6×10^9
	2004	Rat	Rattus norvegicus	21	2.8×10^9
	2005	Chimpanzee	Pan troglodytes	24	3.1×10^9

Bigger genome than Humans

SIZE DOES NOT MATTER

☺

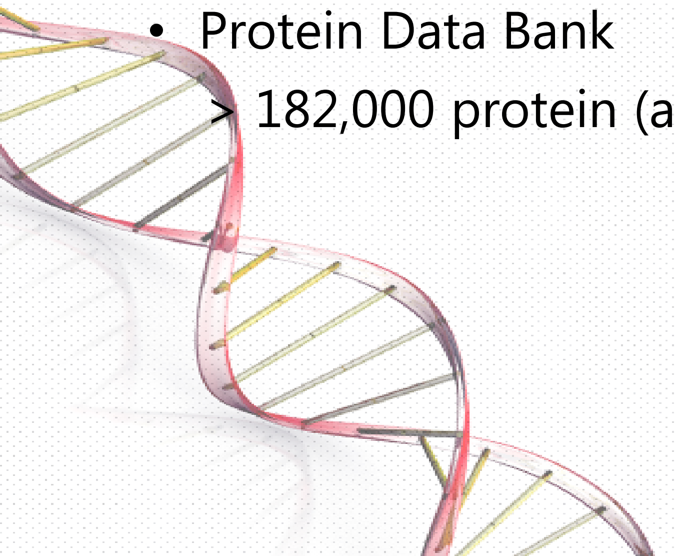
Sequence is freely available

NCBI - http://www.ncbi.nlm.nih.gov

UCSC - http://genome.ucsc.edu

But Wait, There’s More…

- > 1000 other publicly available databases pertaining to molecular biology
- GenBank
  > 231 million sequence entries
  > 940 billion bases
- UniProtKB / Swiss-Prot
  > 565k protein sequence entries
  > 200 million amino acids
- Protein Data Bank
  > 182,000 protein (and related) structures



Bioinformatics Revisited

Representation/storage/retrieval/ analysis of biological data concerning

– sequences (DNA, protein, RNA)

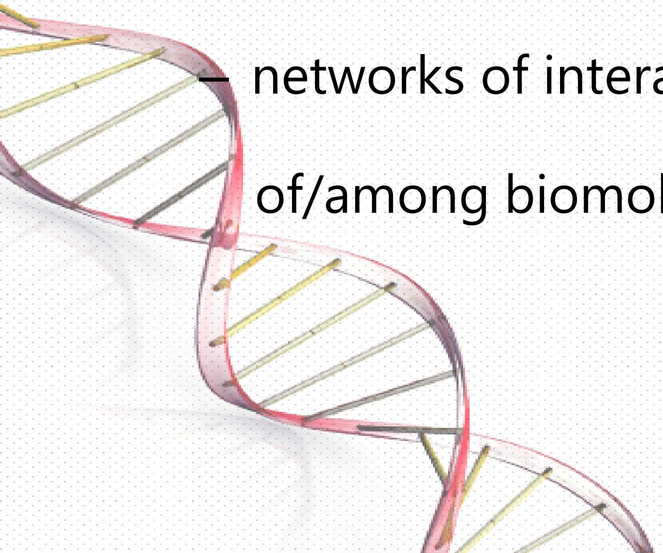
– structures (protein, RNA)

– functions (protein, sequence signals)

– activity levels (mRNA, protein, metabolites)

– networks of interactions (metabolic pathways, regulatory pathways, signaling pathways)

of/among biomolecules



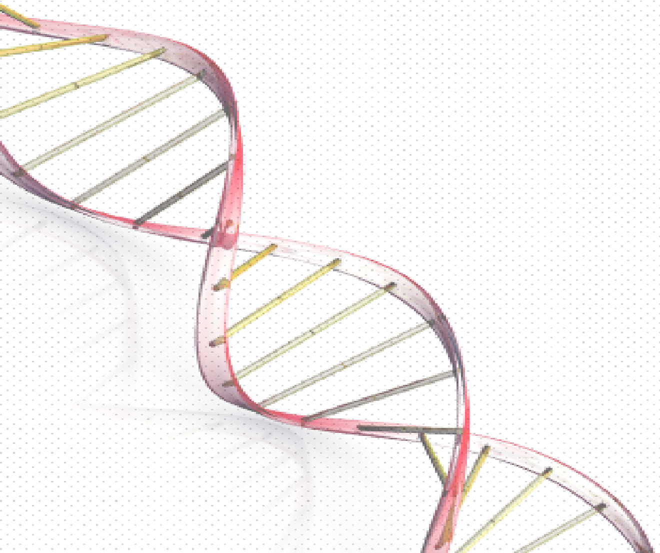
Data and Timeline

- 12,000 enzymes
  - Purified and characterized
  - Function and effect of mutations has been understood

~100 years

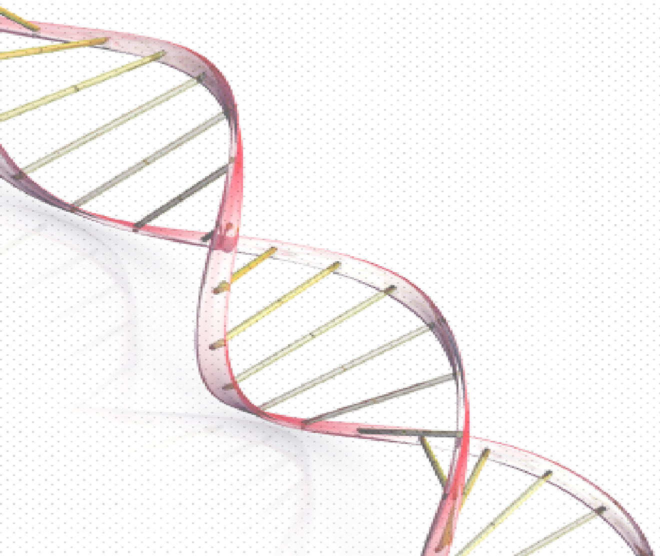
- 95,000 protein structures

~50 years



Data and Timeline

- 1,500,000 genes and their protein products
  - the complete genetic information of several viral, bacterial genomes, and more to come.



Lets try to cure CANCER

- Have access only to a limited amount of information

↓

- Focus on parameters (e.g. genes) with dominant effects

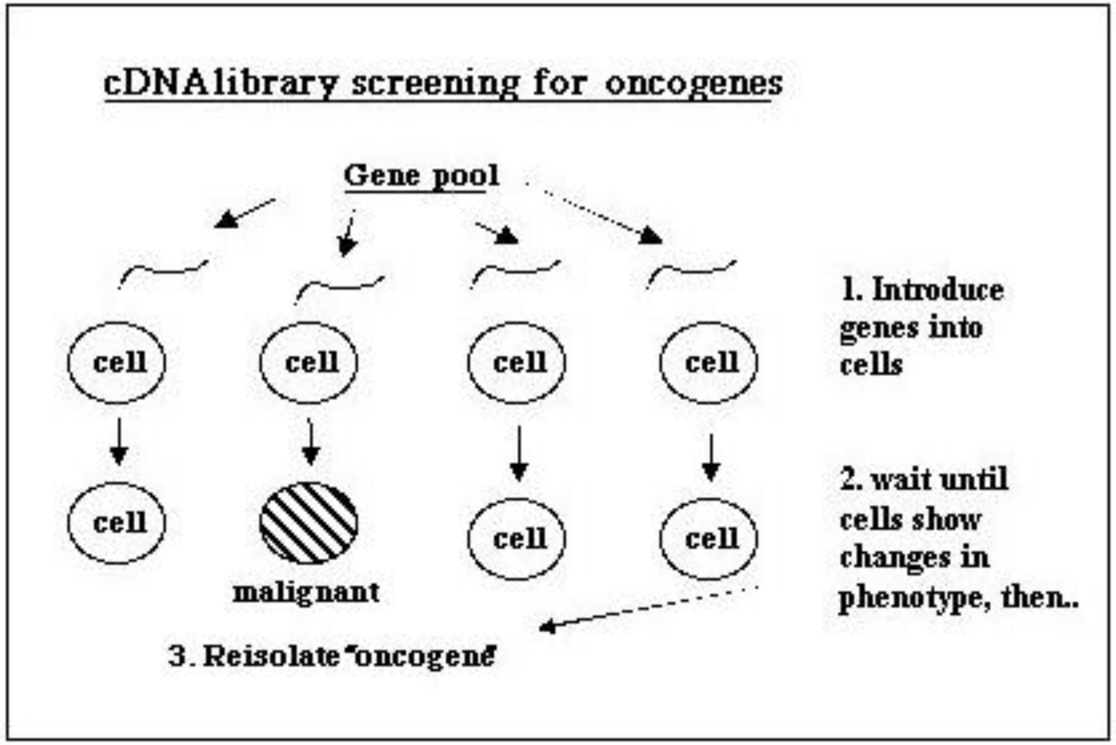
↓

- Design experimental strategies that help to sort out the dominant parameters before information is extracted - (information is gained only about "important" parameter)

↓

- **Find somehow an/the important parameter (eg. Gene)**

Lets try to cure CANCER



cDNAlibrary screening for oncogenes

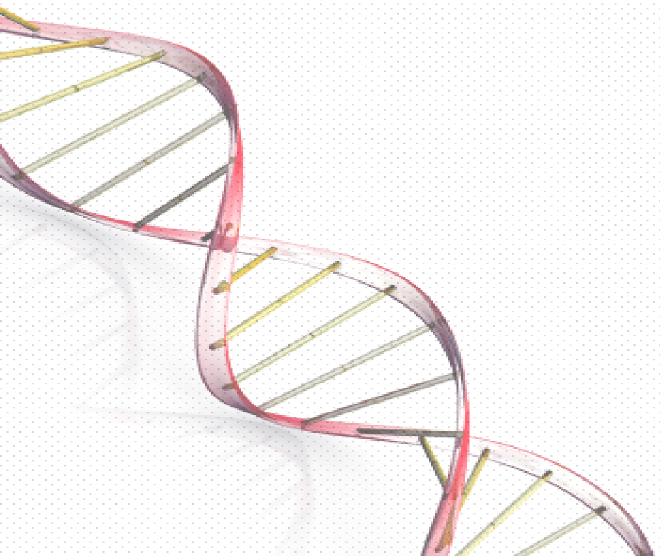
1. Introduce genes into cells

2. wait until cells show changes in phenotype, then..

3. Reisolate "oncogene"

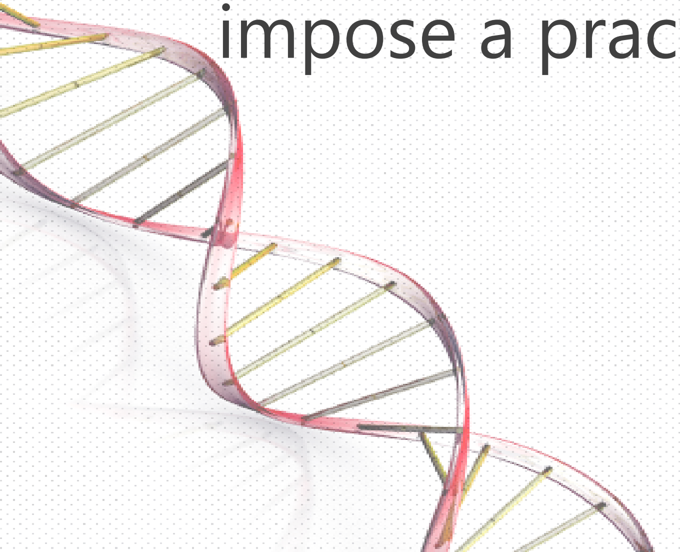
Lets try to cure CANCER

- Limitations
  - Many phenotype (eg. cancer) are due to multiple factors
  - Each factor alone may not have any effects e.g. non-dorminant oncogenes
  - Only specific combination of such cooperating factors lead to phenotype



Lets try to cure CANCER

- Suppose 3 cooperative genes are responsible for cancer.
- Assume that there are 15,000 human genes.
- Need:
  - 15,000*14,999*14998=3.3 x 10^{13} experiments (or 10^{4N} for N cooperating genes).
- The exponential increase in the number of samples to be tested impose a practical and conceptual limitation.



Molecular Biology:

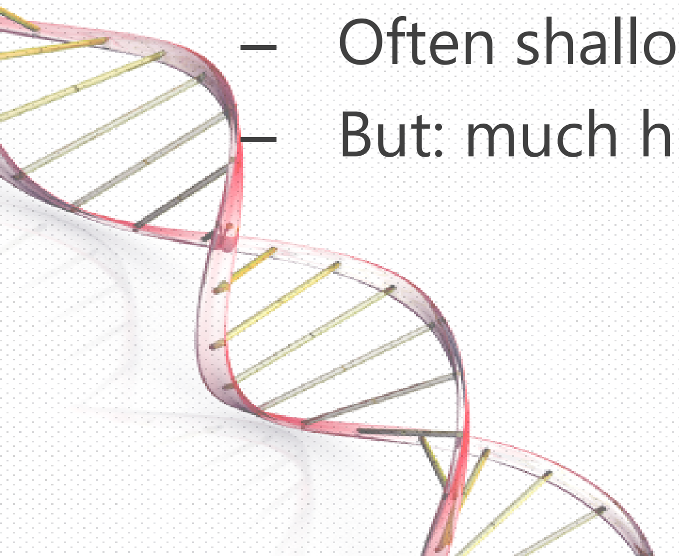
- Often deep understanding of the function of one or several gene/protein,
- But: low ratio of

$$\frac{\textbf{Information extracted}}{\textbf{Potentially relevant information}}$$

Genomics, Gene expression, Bioinformatics:

- Often shallower understanding of the functions
- But: much higher ratio of

$$\frac{\textbf{Information extracted}}{\textbf{Potentially relevant information}}$$



References

- Lecture notes of Colin Dewey @ University of Wisconsin-Madison
- Lecture notes of Arne Elofsson @ Stockholm University
- Lecture notes of Yuzhen Ye @ Indiana University
- http://www.ornl.gov/hgmis