Molecular Biology 101

Fall 2017

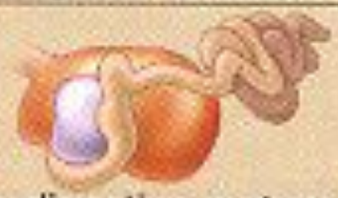
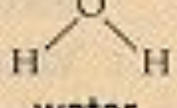
BMI/CS 576

www.biostat.wisc.edu/bmi576/

Irene Ong

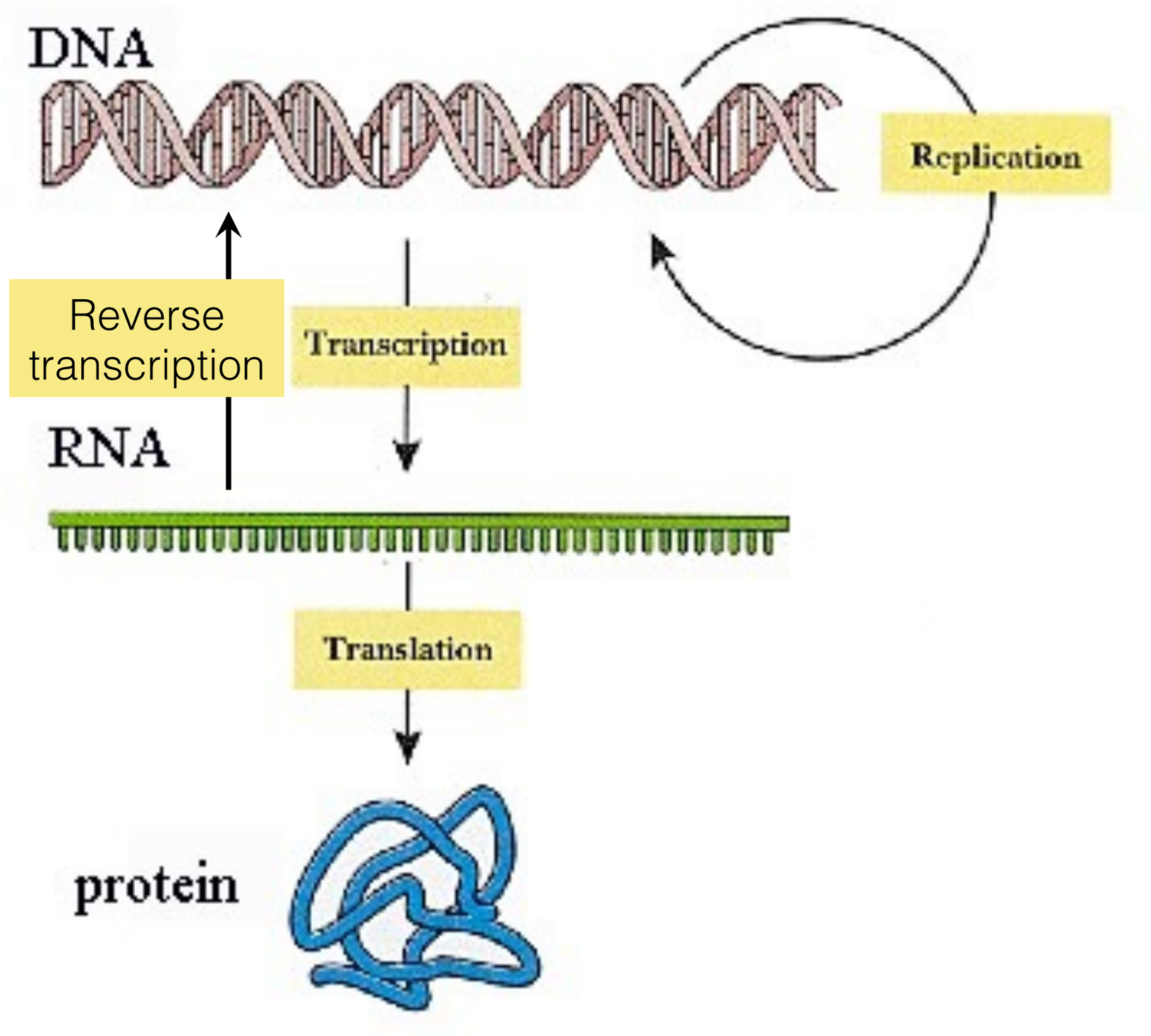
irene.ong@wisc.edu

Levels of the biological hierarchy

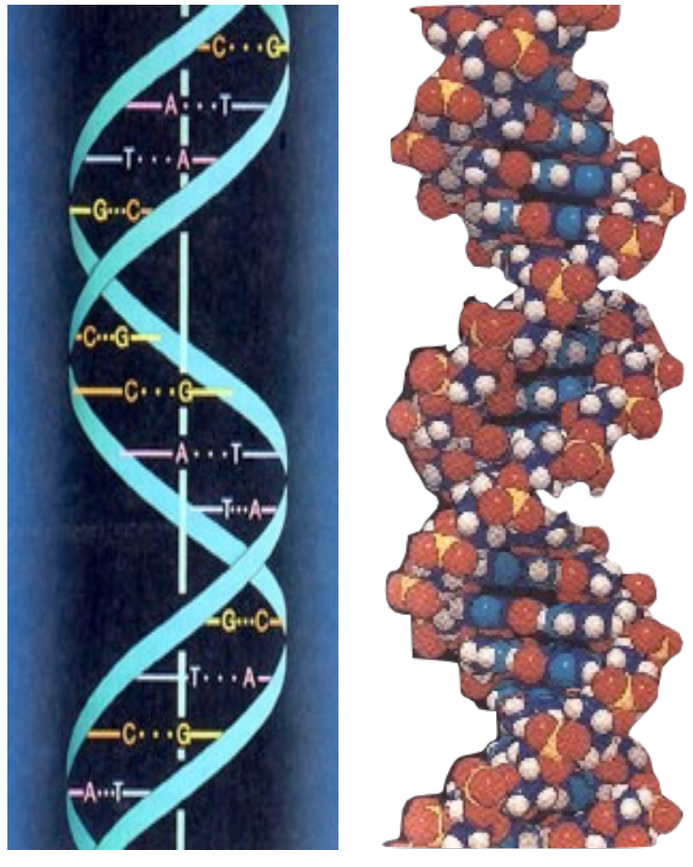
Population	Members of one species inhabiting the same area	 herd of pronghorn antelope
Multicellular organism	An individual living thing composed of many cells	 pronghorn antelope
Organ system	Two or more organs working together in the execution of a specific bodily function	 the digestive system
Organ	A structure usually composed of several tissue types that form a functional unit	 the stomach
Tissue	A group of similar cells that perform a specific function	 epithelial tissue
Cell	The smallest unit of life	 red blood cell  epithelial cell  nerve cell
Molecule	A combination of atoms	 water  glucose  DNA

Edition 9 of Biology, by Audesirk et. al.

The Central Dogma



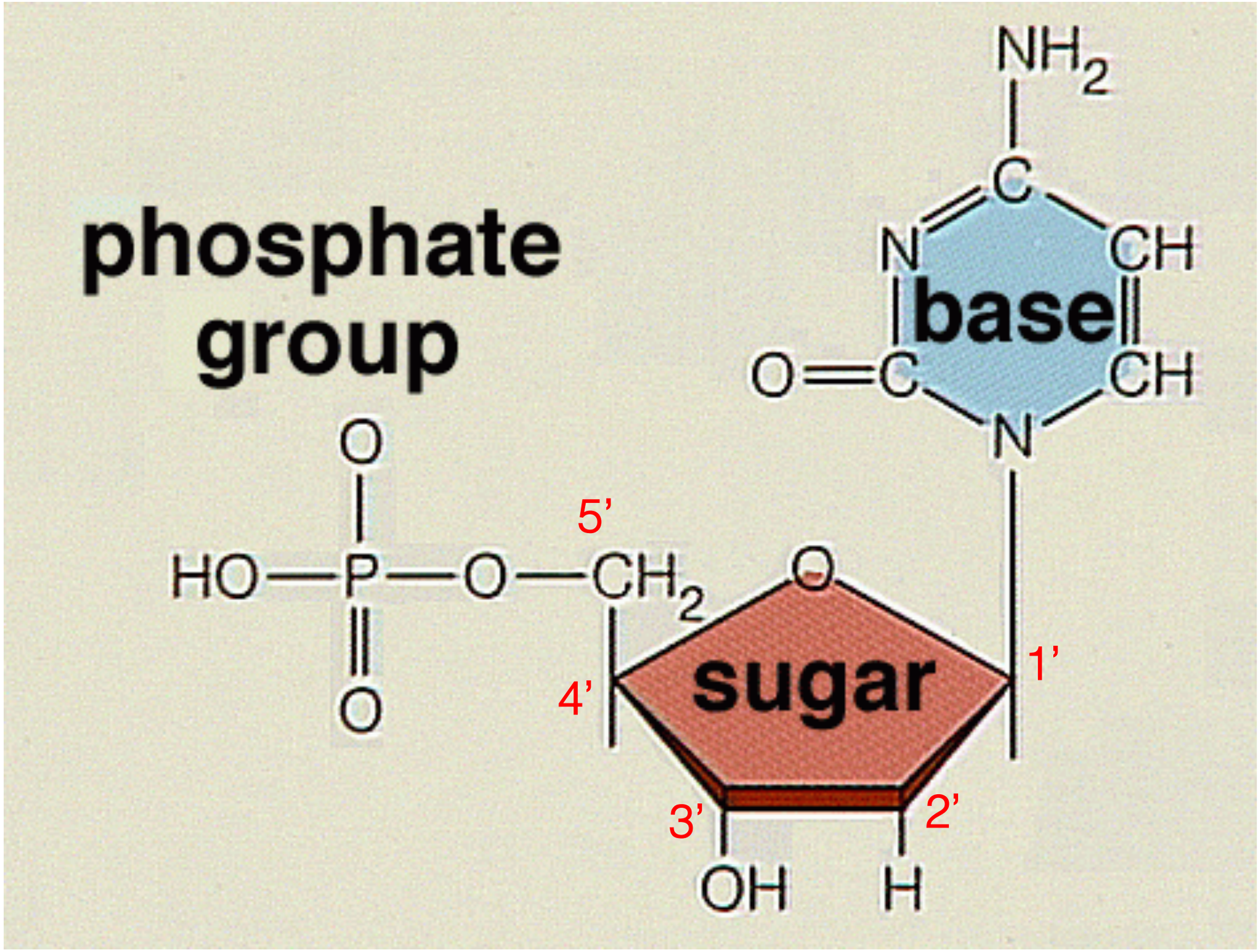
DNA



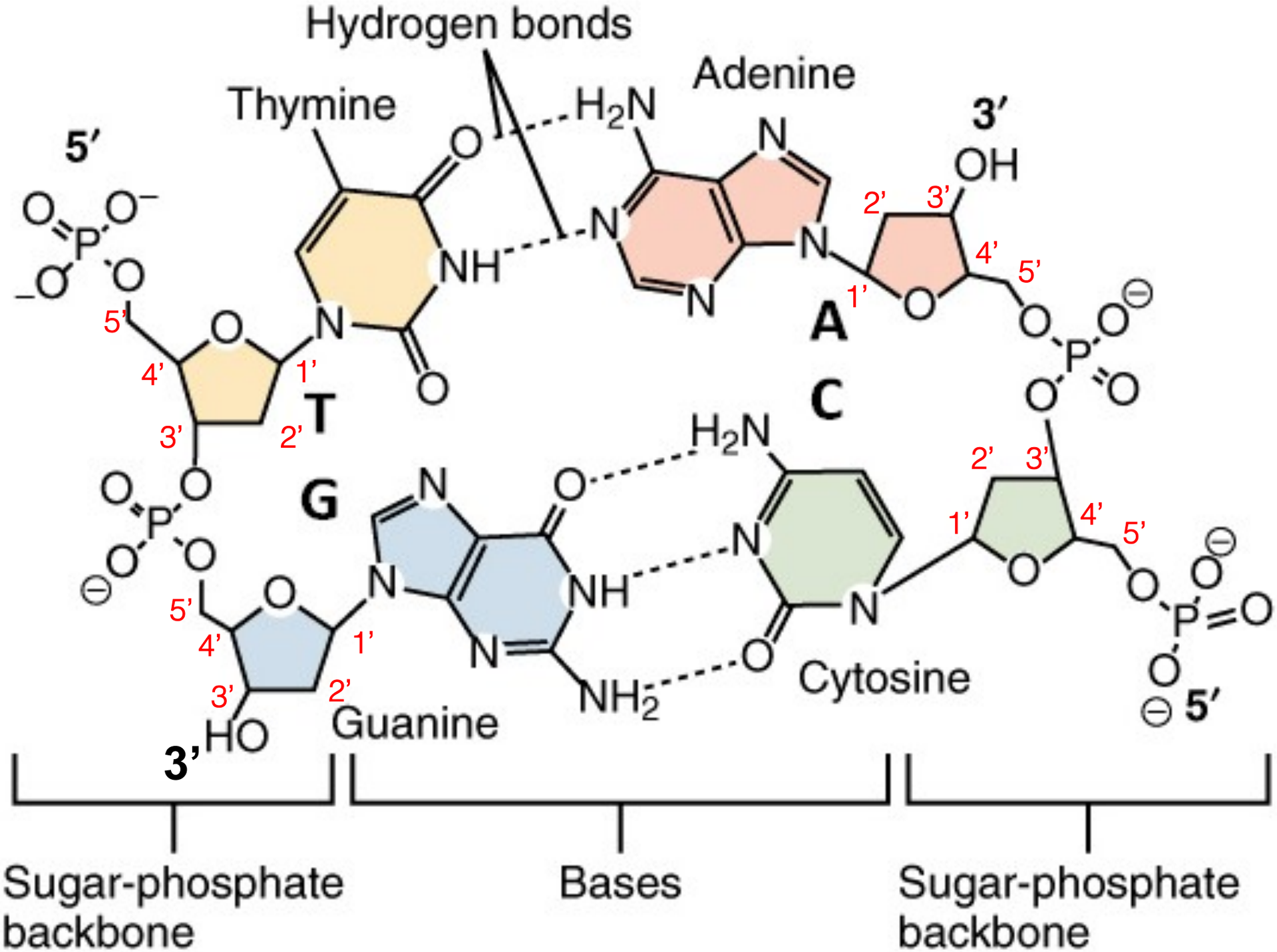
- DeoxyriboNucleic Acid – two strands - double helix
- can be thought of as the “blueprint” for an organism
- a linear chain of small molecules called nucleotides
  - four different nucleotides:
    - adenine (A), cytosine (C), guanine (G) and thymine (T)
- is a polymer: large molecule consisting of similar units (nucleotides in this case)
- a single strand of DNA can be thought of as a string composed of the four letters: A, C, G, T

  CTGCTGGACCGGGTGCTAGGACCCTGACTGCCCGGGGCCGGGGGT
  GCGGGGCCCGCTGAG…

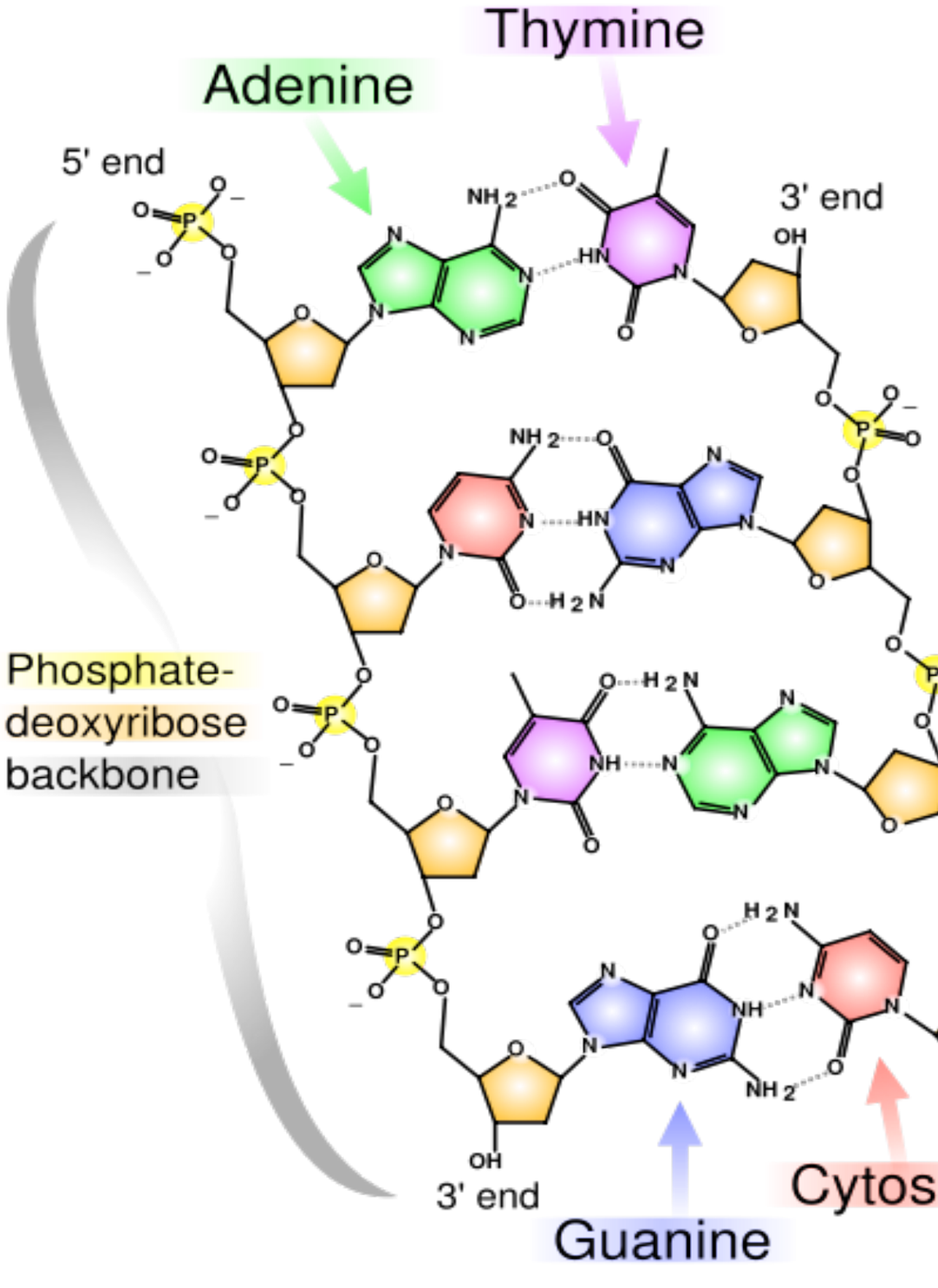
Nucleotides: the subunits of DNA



The four DNA bases



DNA strand: polymer of nucleotides

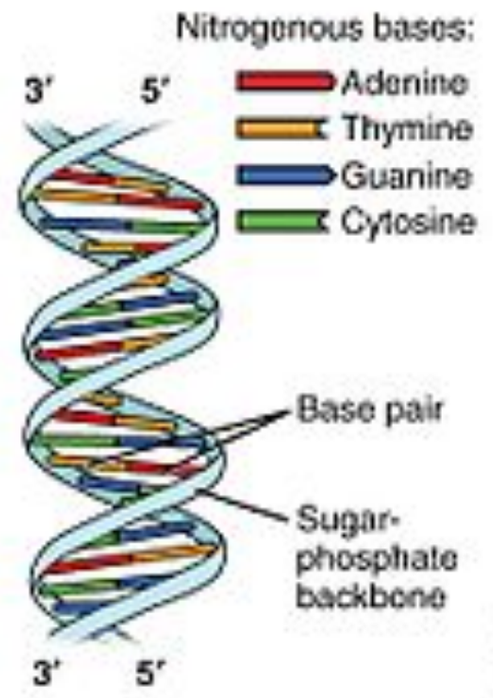


- in double-stranded DNA

  A always bonds to T

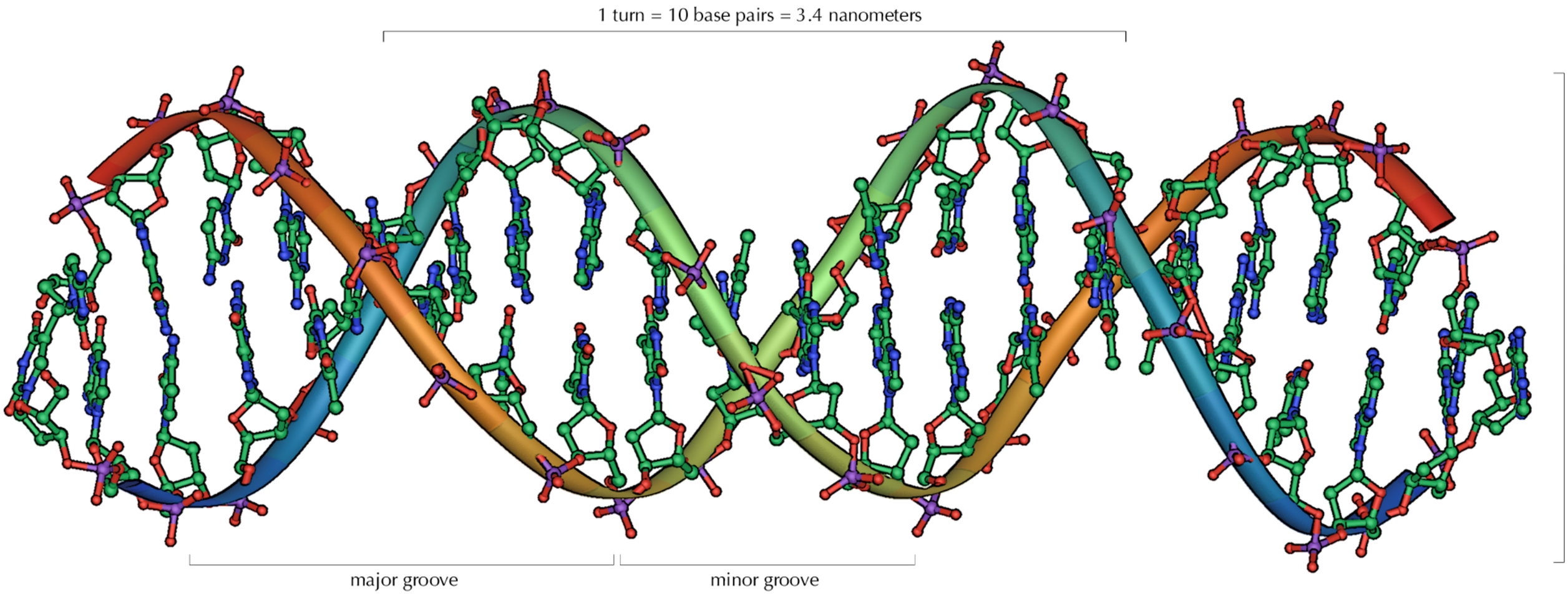
  C always bonds to G

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)

DNA dimensions



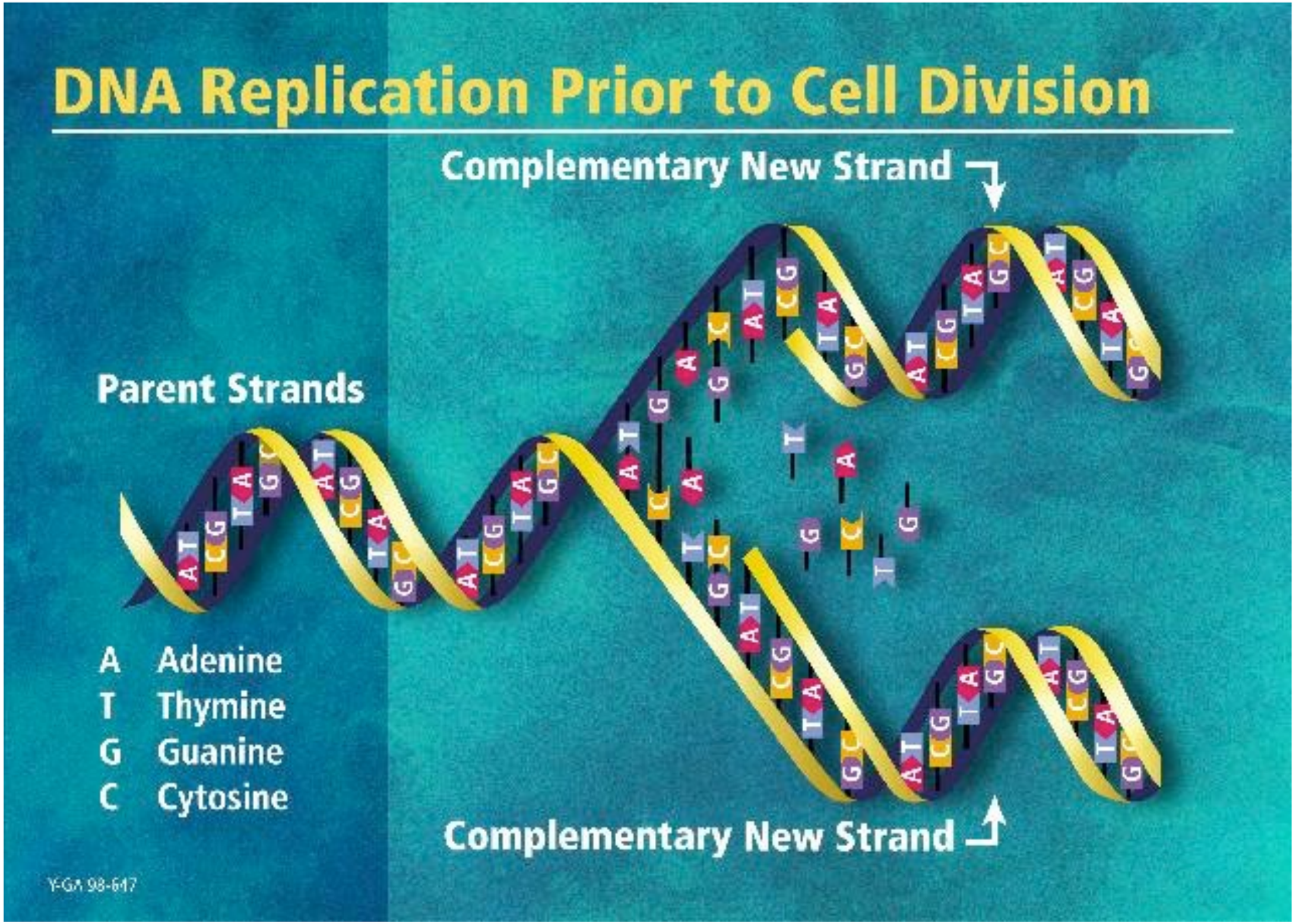


image from the DOE Human Genome Program
http://www.ornl.gov/hgmis

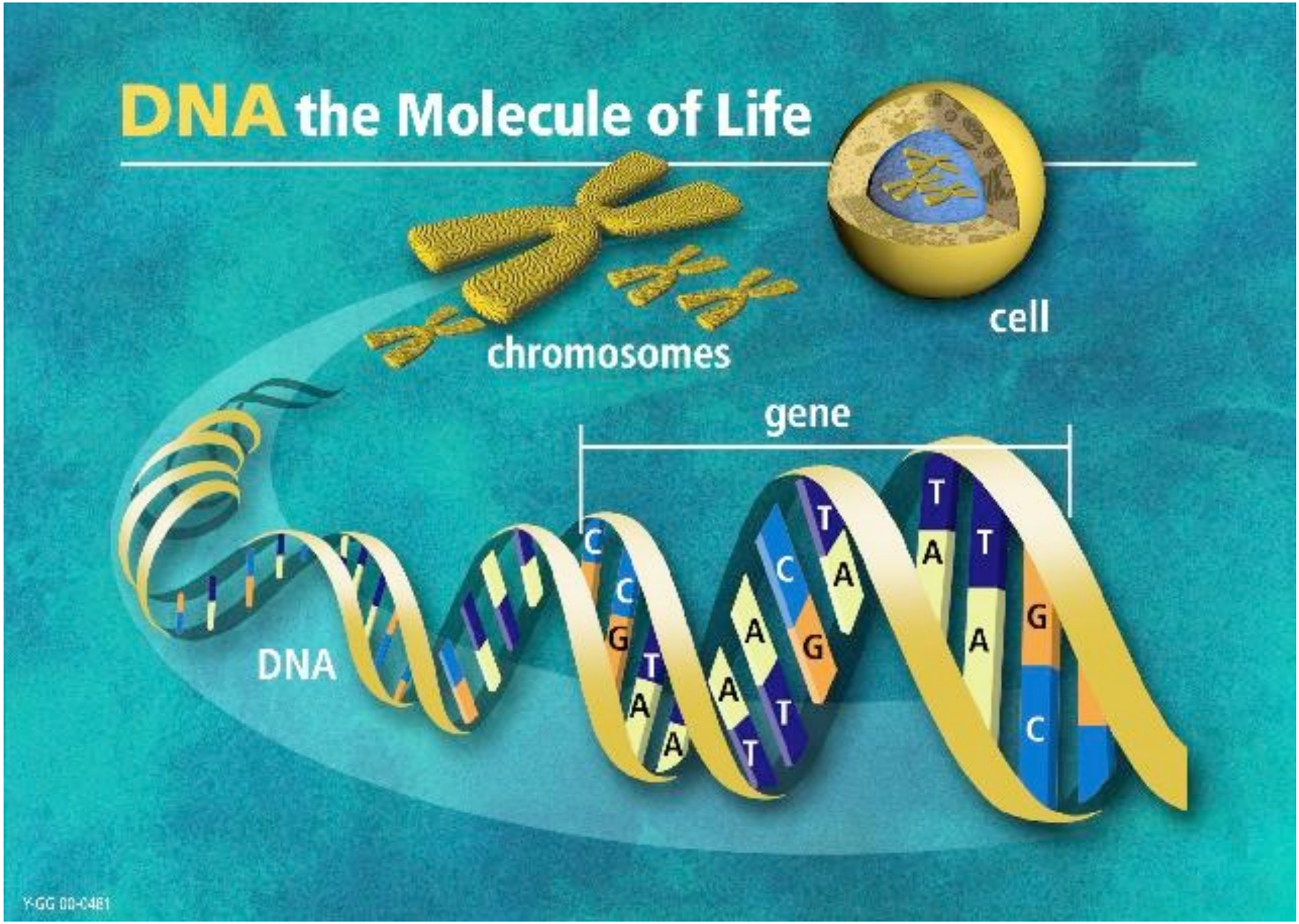


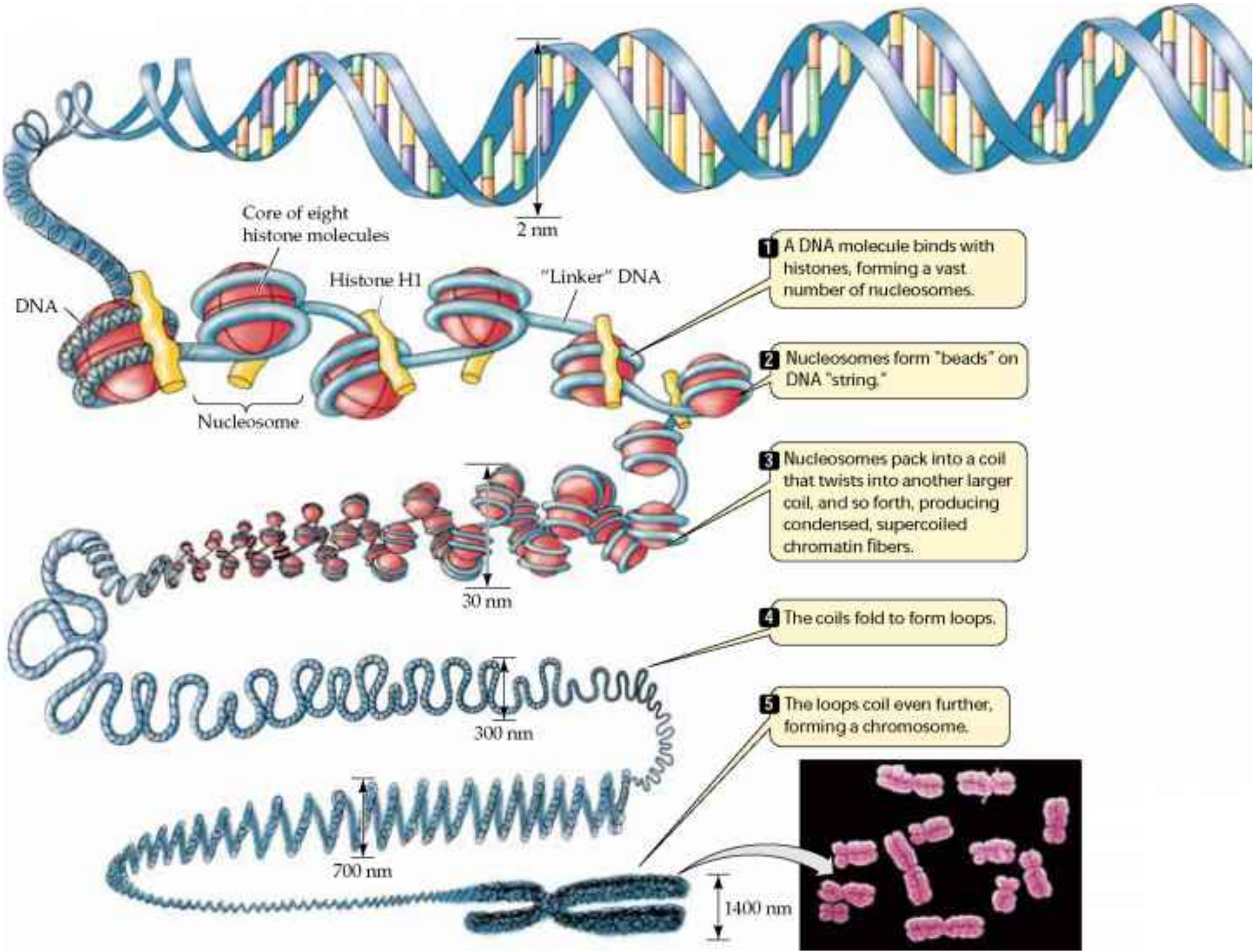
image from the DOE Human Genome Program
http://www.ornl.gov/hgmis

Chromosomes

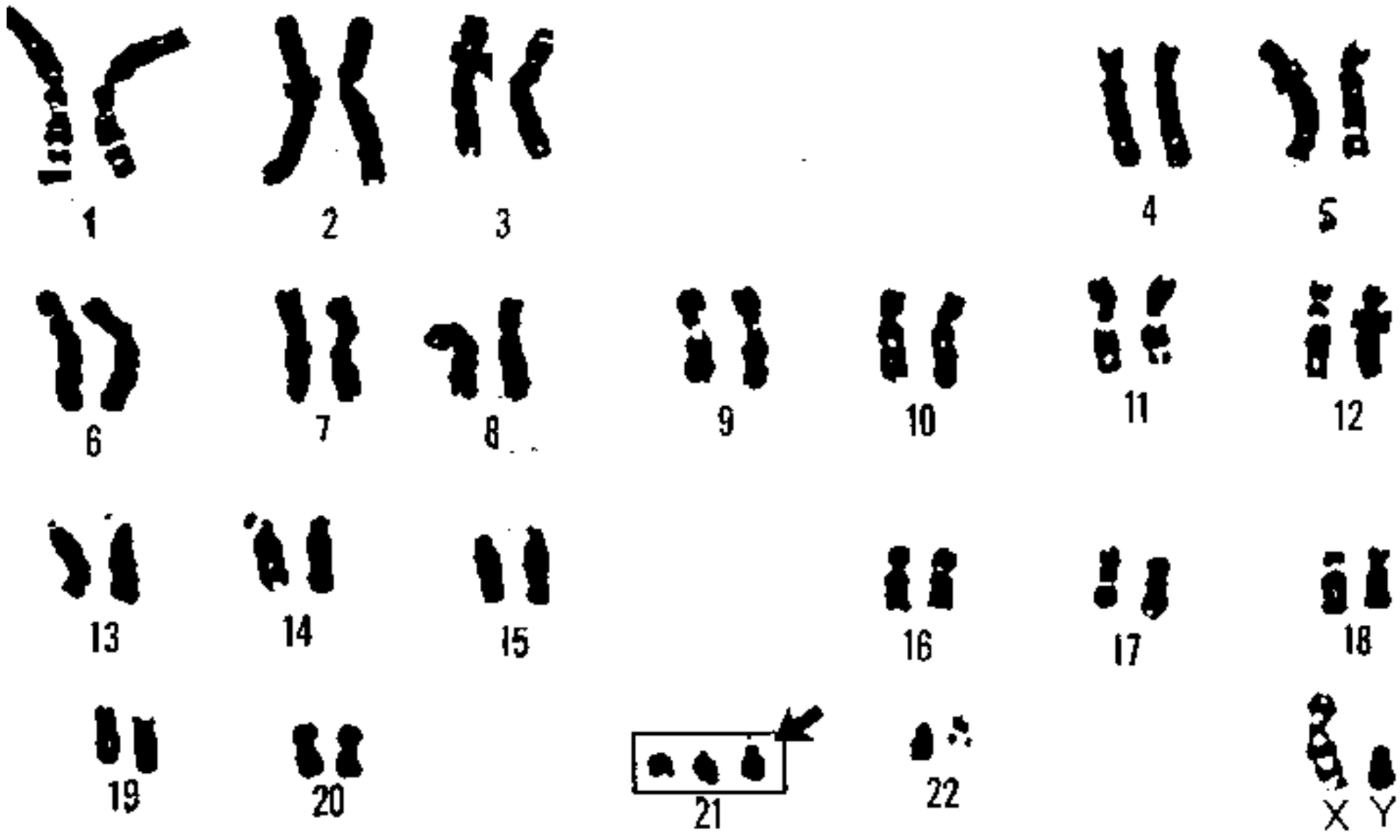
- DNA is packaged into individual chromosomes (along with proteins)

- 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
Core of eight histone molecules
Histone H1
"Linker" DNA
DNA
Nucleosome
1 A DNA molecule binds with histones, forming a vast number of nucleosomes.
2 Nucleosomes form "beads" on DNA "string."
3 Nucleosomes pack into a coil that twists into another larger coil, and so forth, producing condensed, supercoiled chromatin fibers.
30 nm
4 The coils fold to form loops.
300 nm
5 The loops coil even further, forming a chromosome.
700 nm
1400 nm

Human Chromosomes



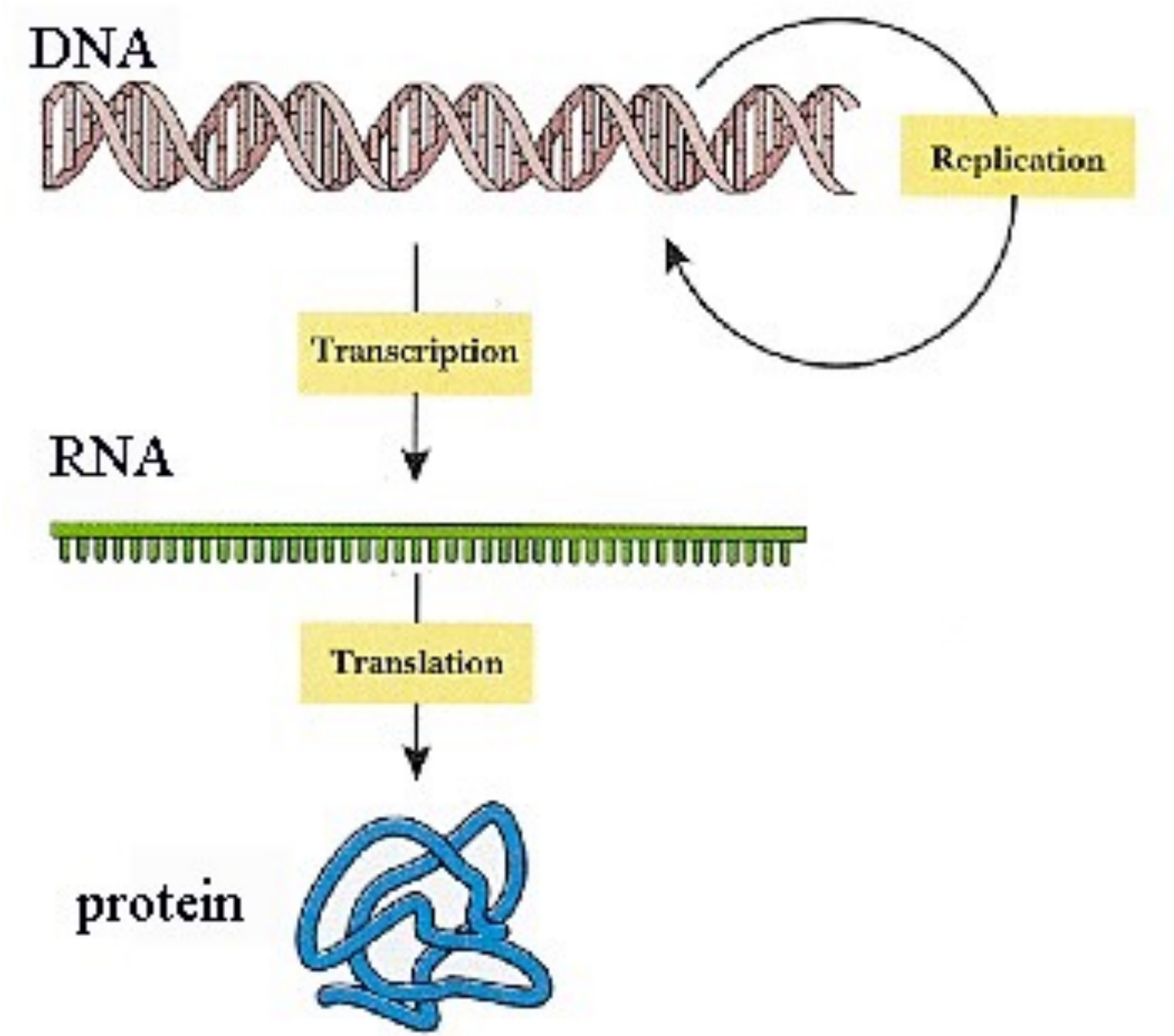
Genomes

- 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

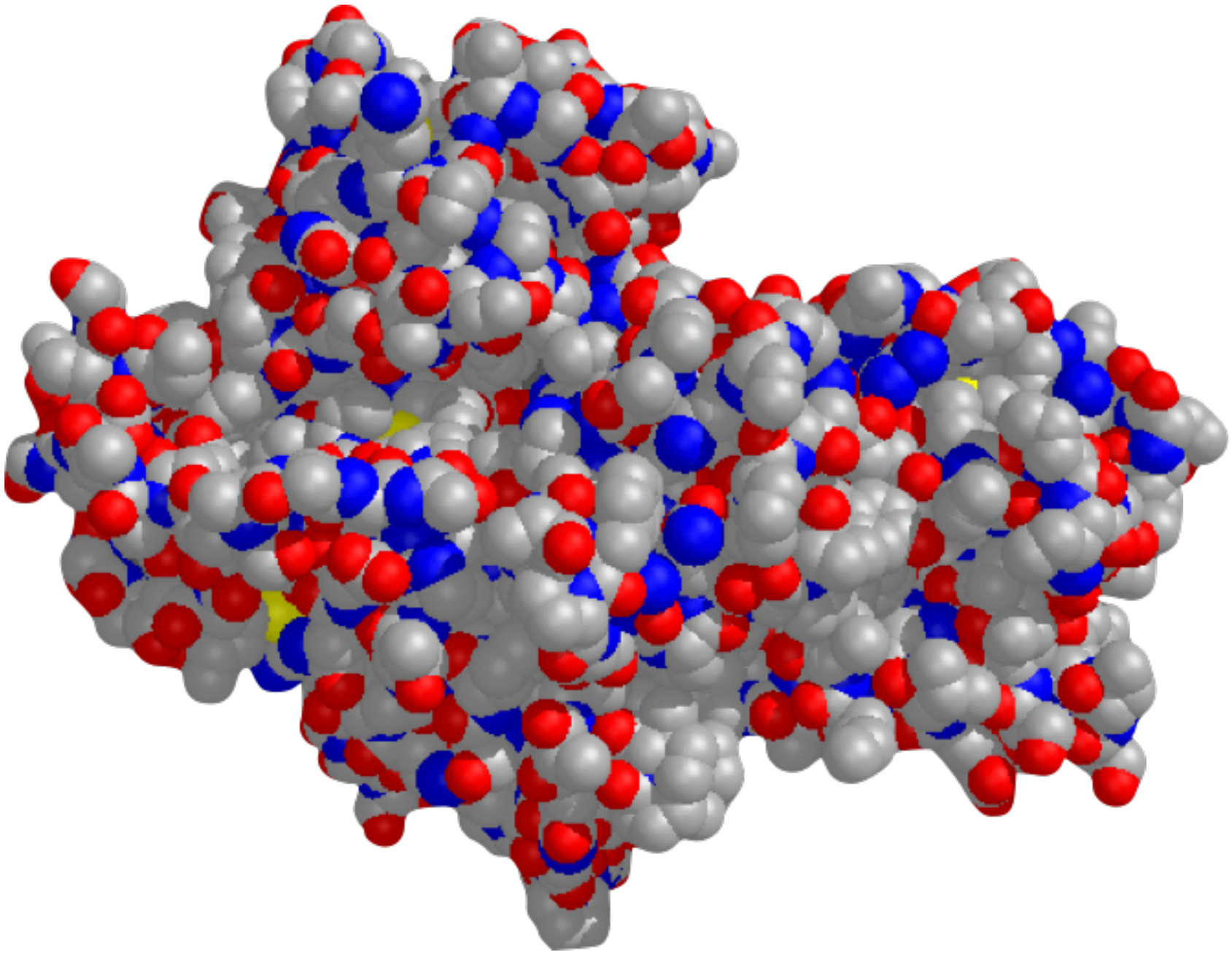
The Central Dogma



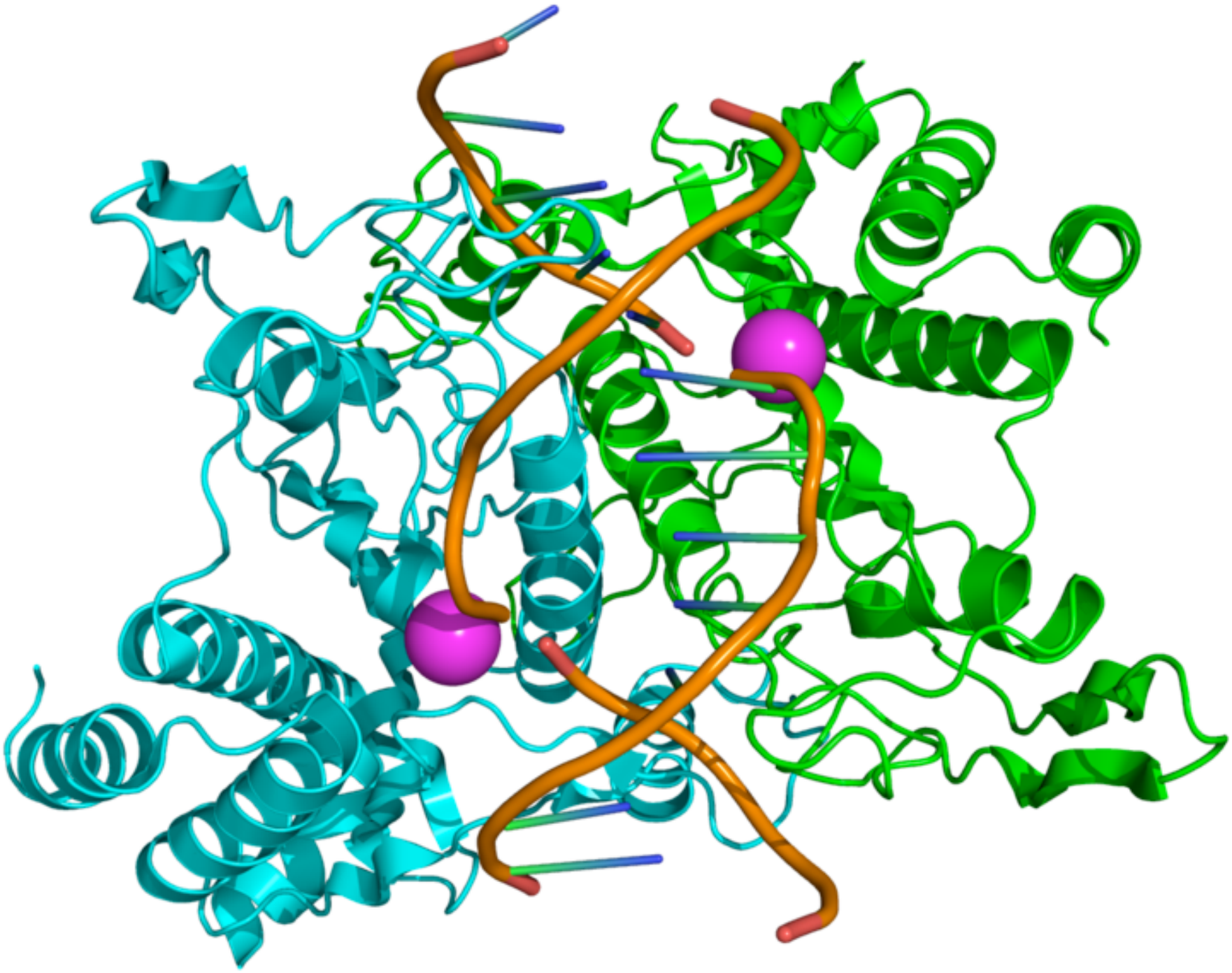
Examples of proteins

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)

Space-Filling Model of Hexokinase

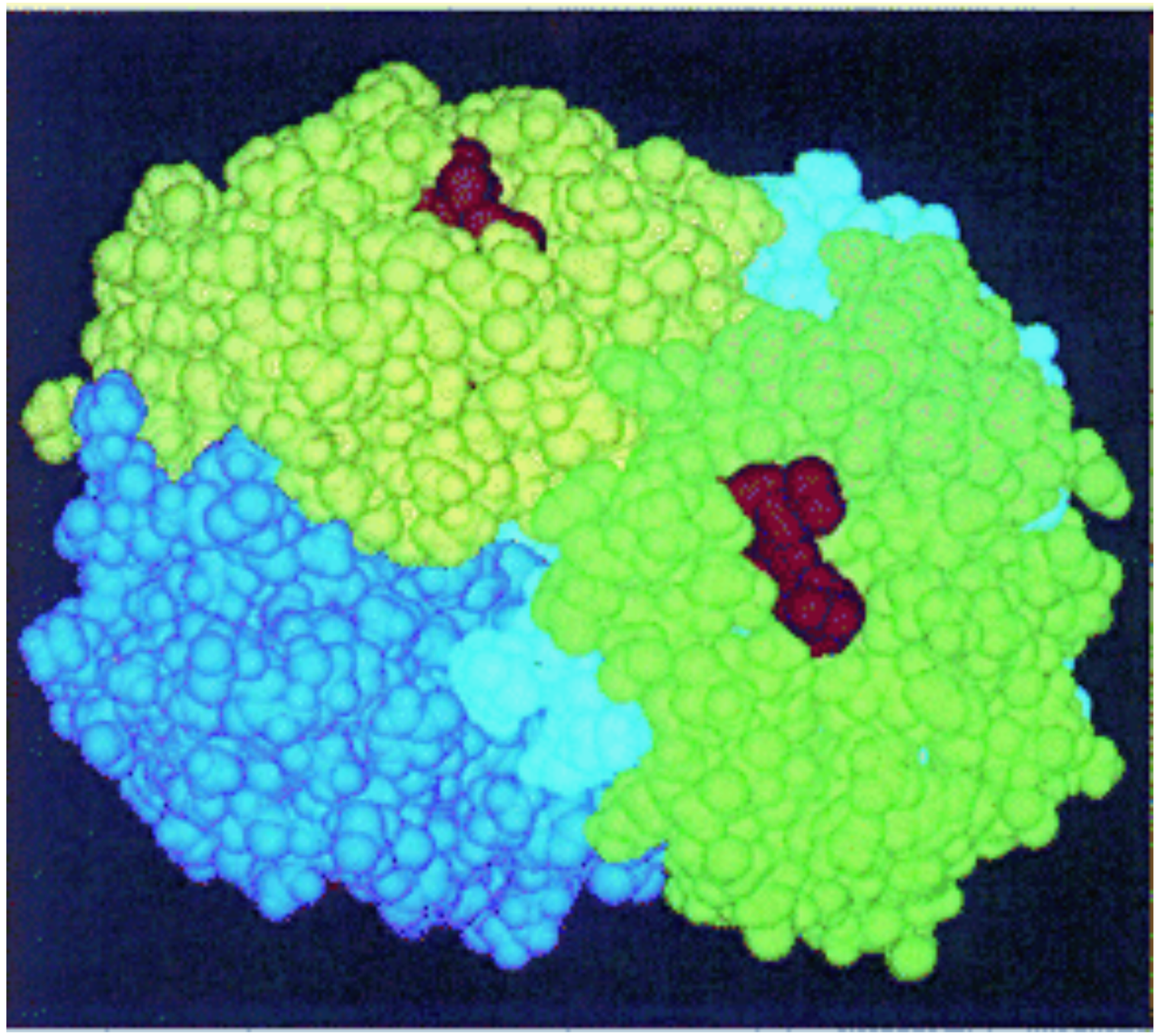


EcoRI – restriction enzyme

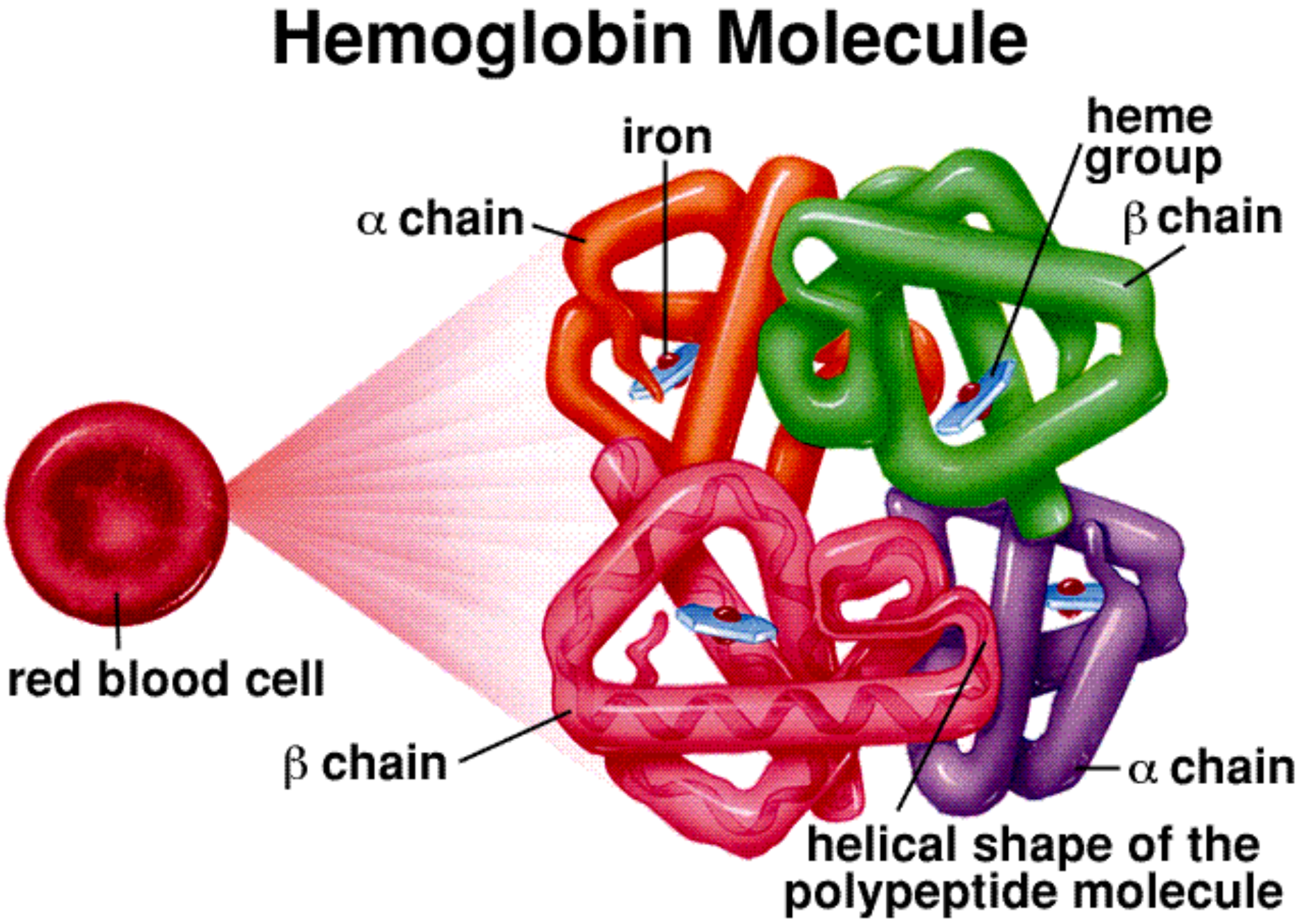


Hemoglobin

- protein built from 4 polypeptides
- responsible for carrying oxygen in red blood cells



Hemoglobin: carrier of oxygen

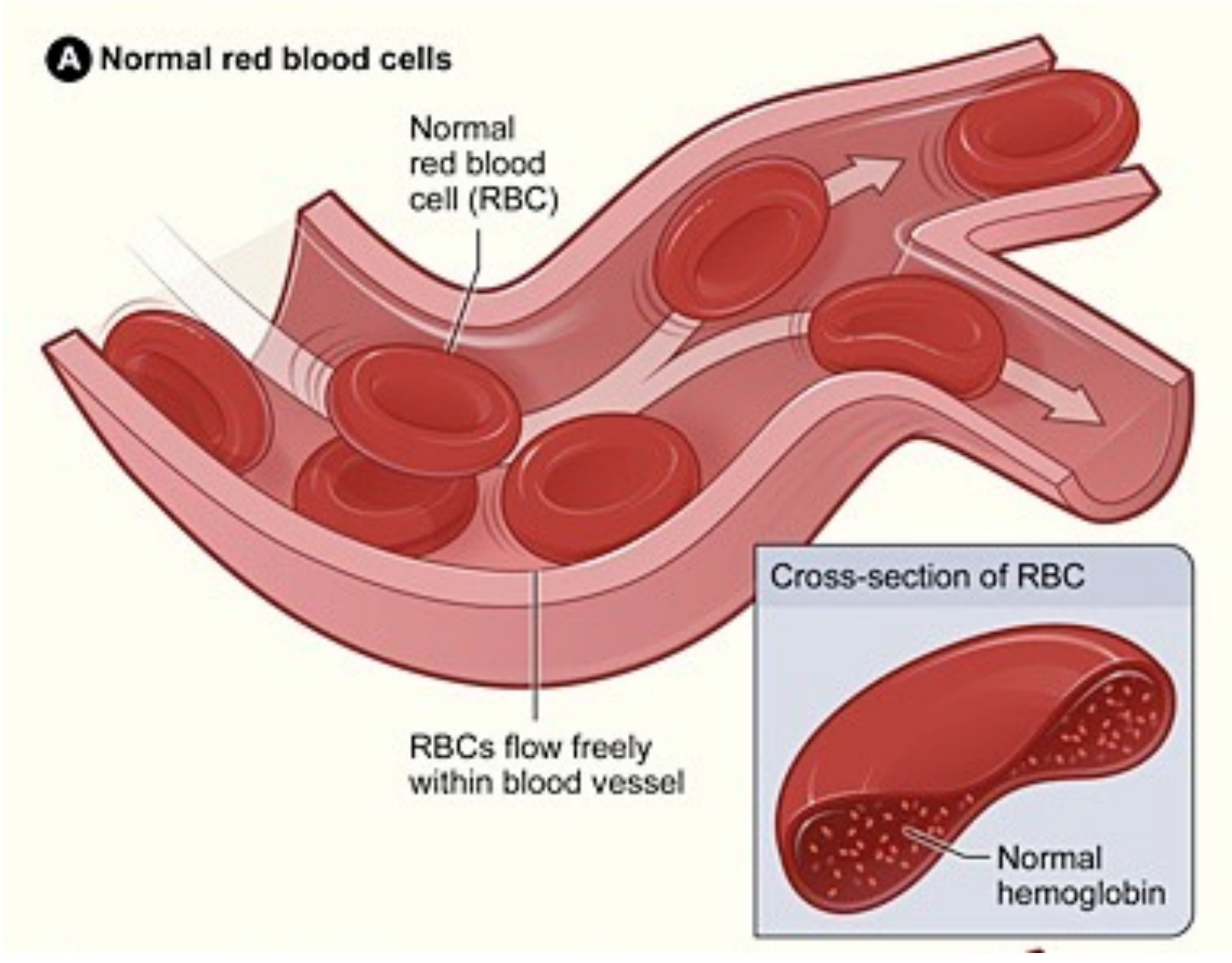


Mutant β-globin → Sickle blood cells

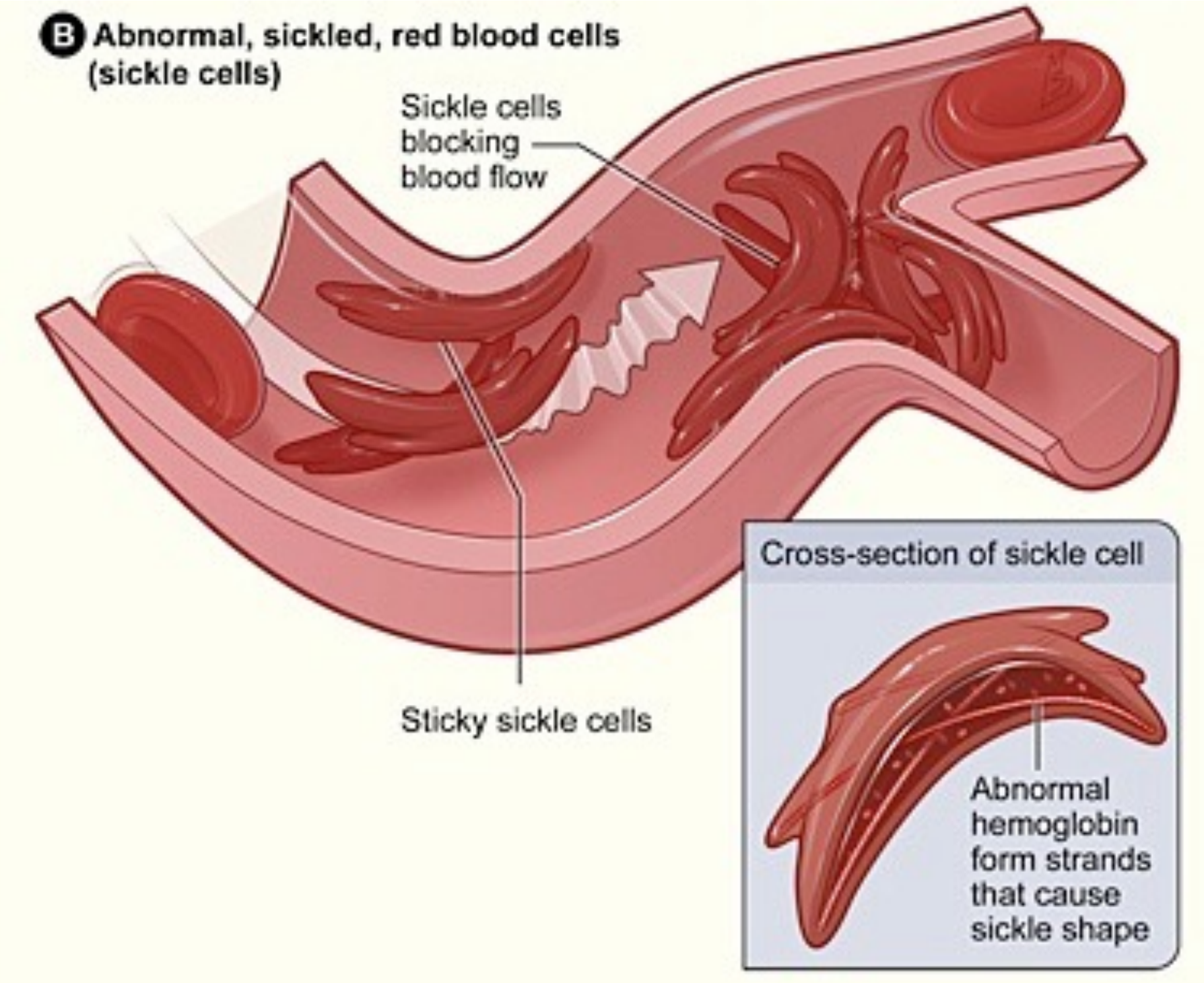
Fiber of sickle hemoglobin

Sickle and normal blood cells

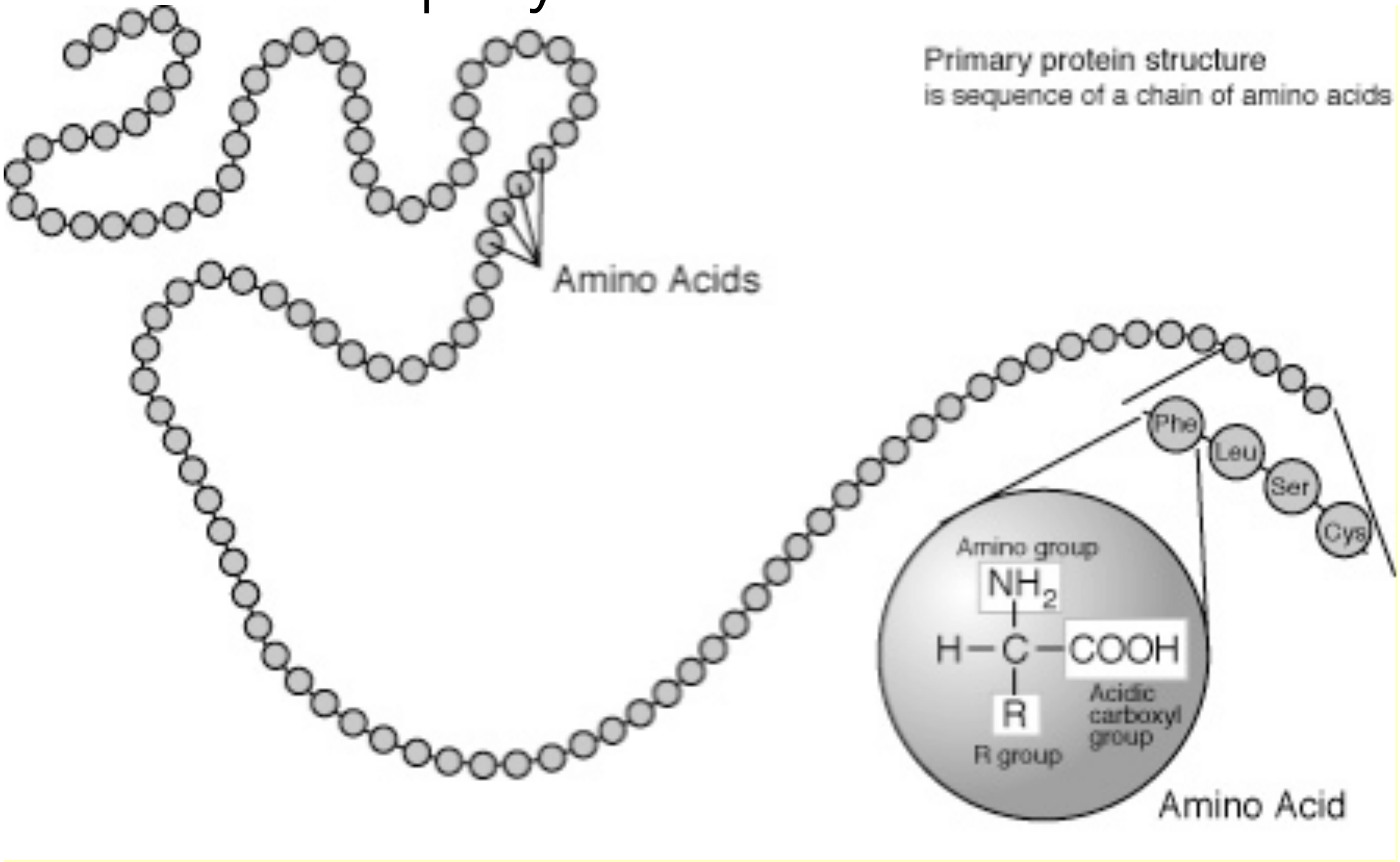
Normal blood flow



Sickle cell complications



Protein: polymers of amino acids



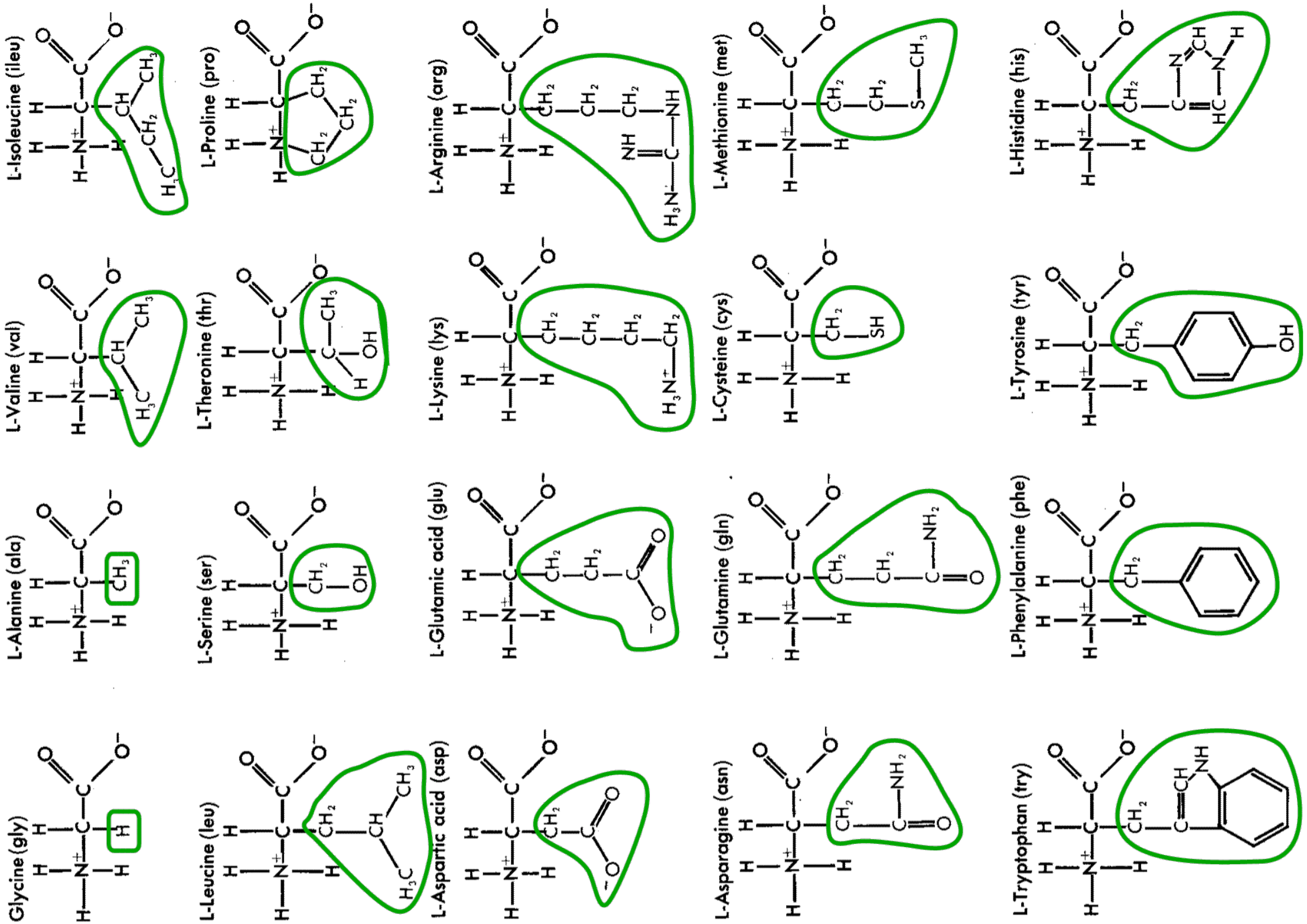
Proteins

- 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

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

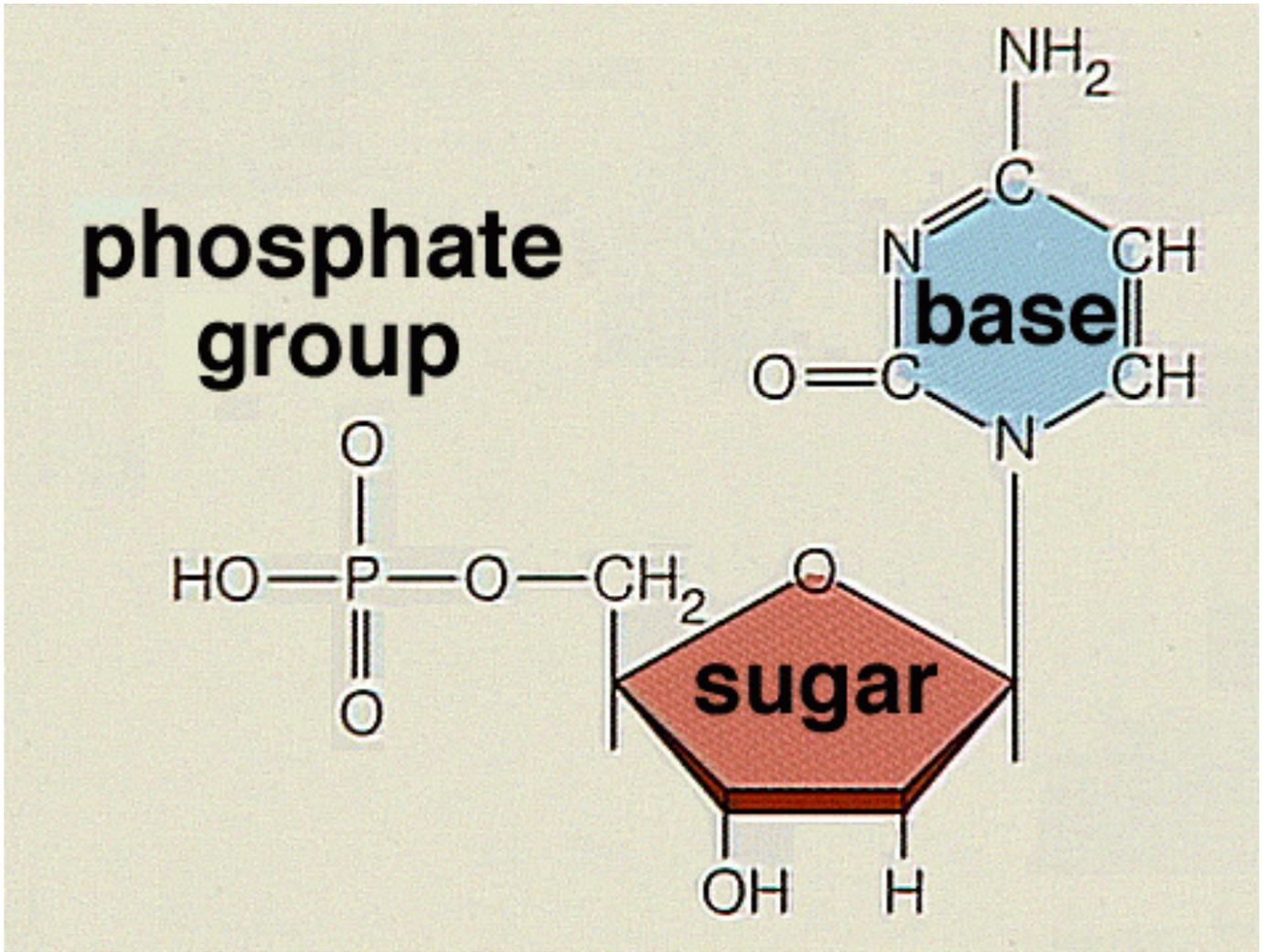
The alphabet of protein: 20 amino acids



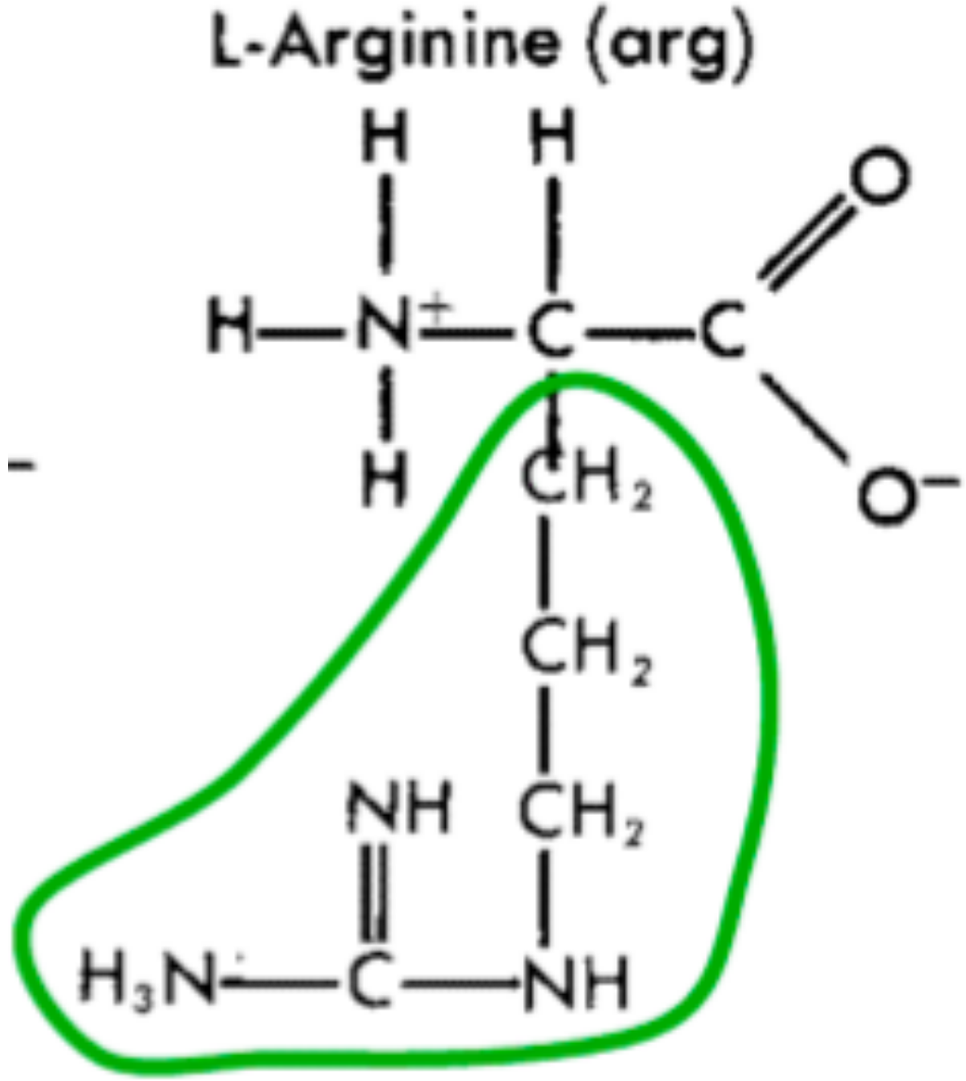
Amino Acid Sequence of Hexokinase

```
  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
```

Nucleotides vs. Amino Acids



Nucleotide

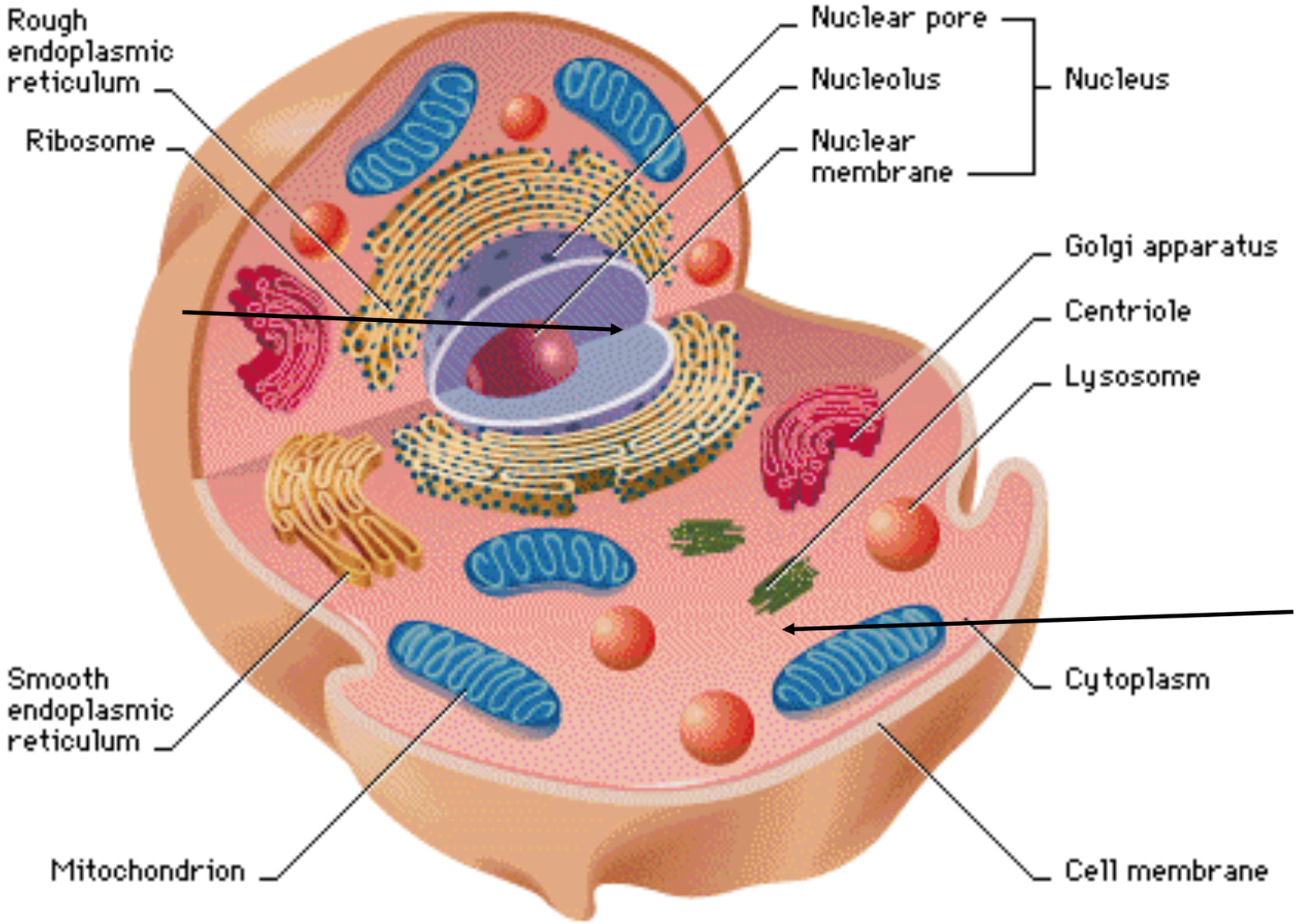


Amino Acid

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

The inaccessible code

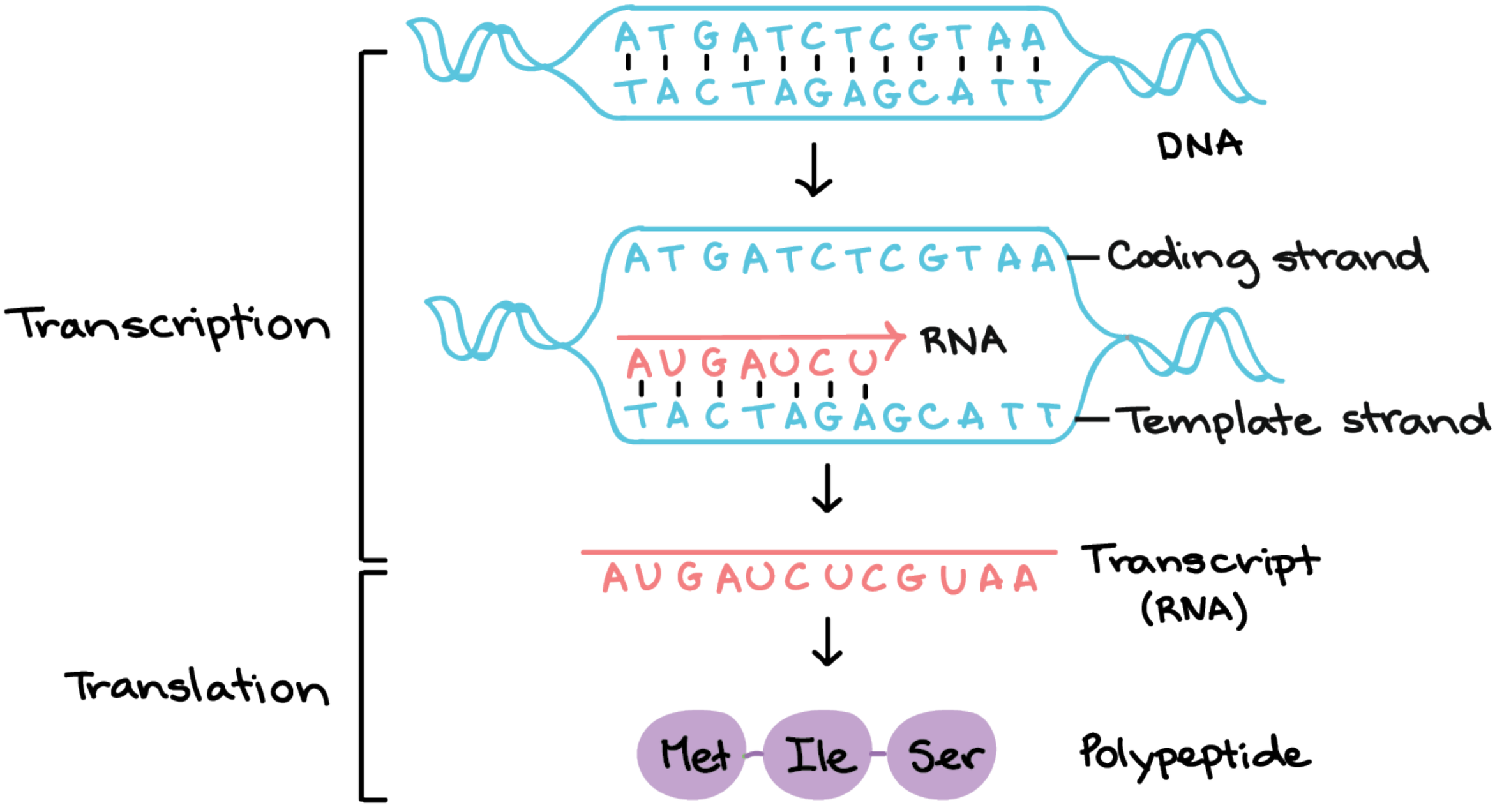
DNA is in the nucleus



Proteins are (mostly) made in the cytoplasm

(eukaryotic cell)

The Central Dogma

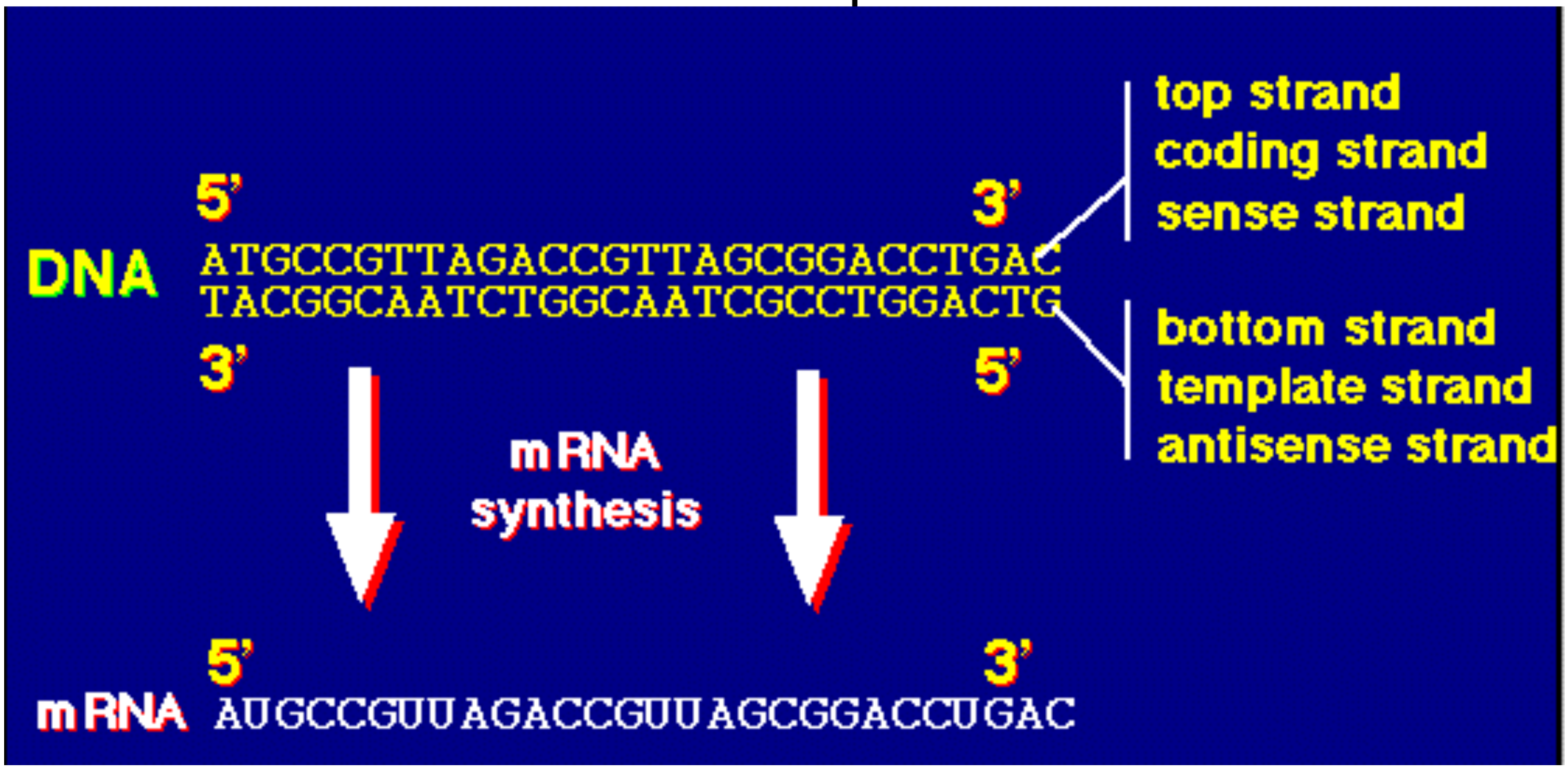


Transcription

- 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)

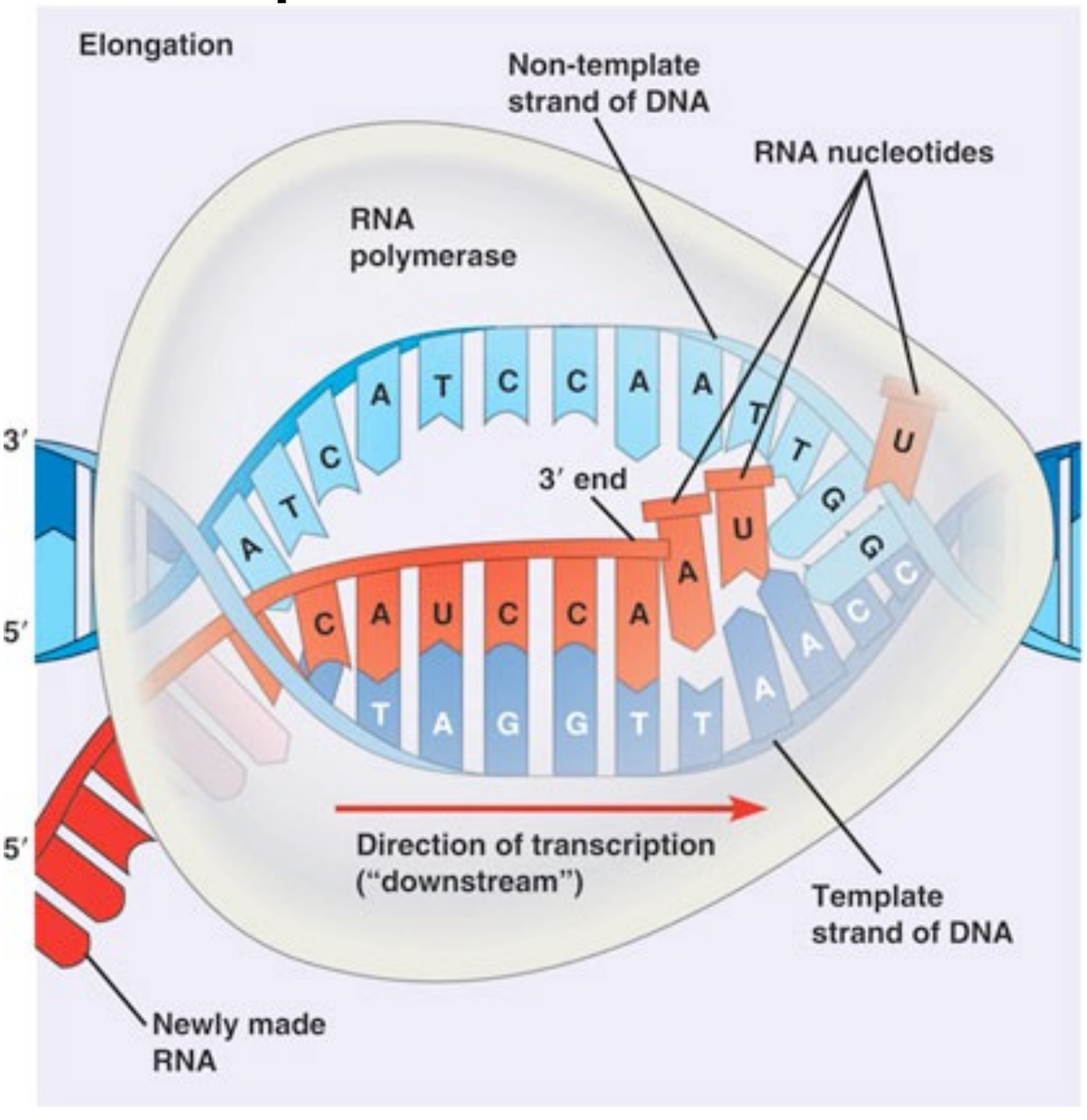
Transcription



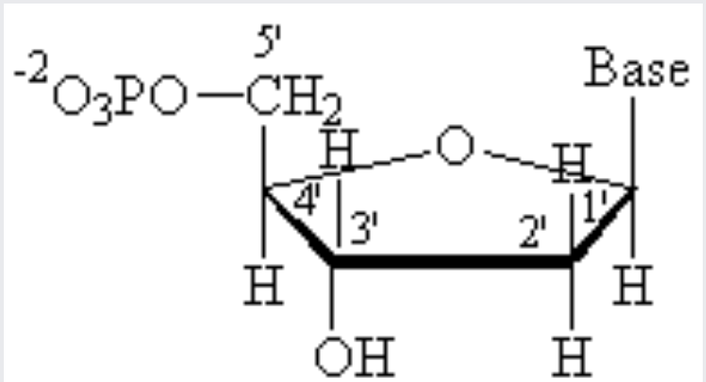
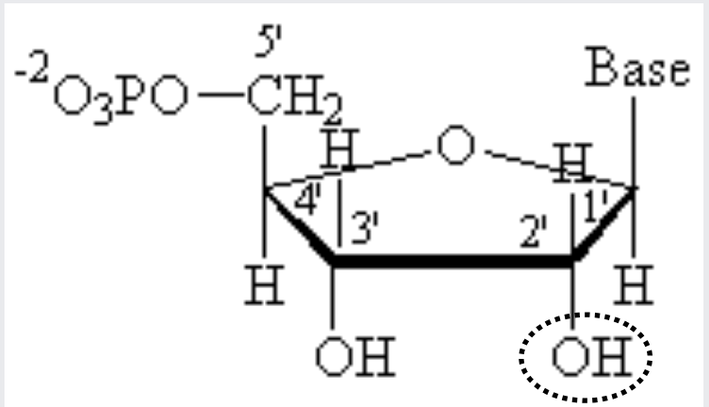
- RNA that is transcribed from a protein-coding gene is called messenger RNA (mRNA)

- RNA polymerase is the enzyme that builds an RNA molecule from a gene

Transcription: DNA→RNA



RNA vs. DNA structure

DNA	RNA
linear polymer	linear polymer
double-stranded	single-stranded
deoxyribonucleotide monomer	ribonucleotide monomer
	
A,C,G,T bases	A,C,G,U bases

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

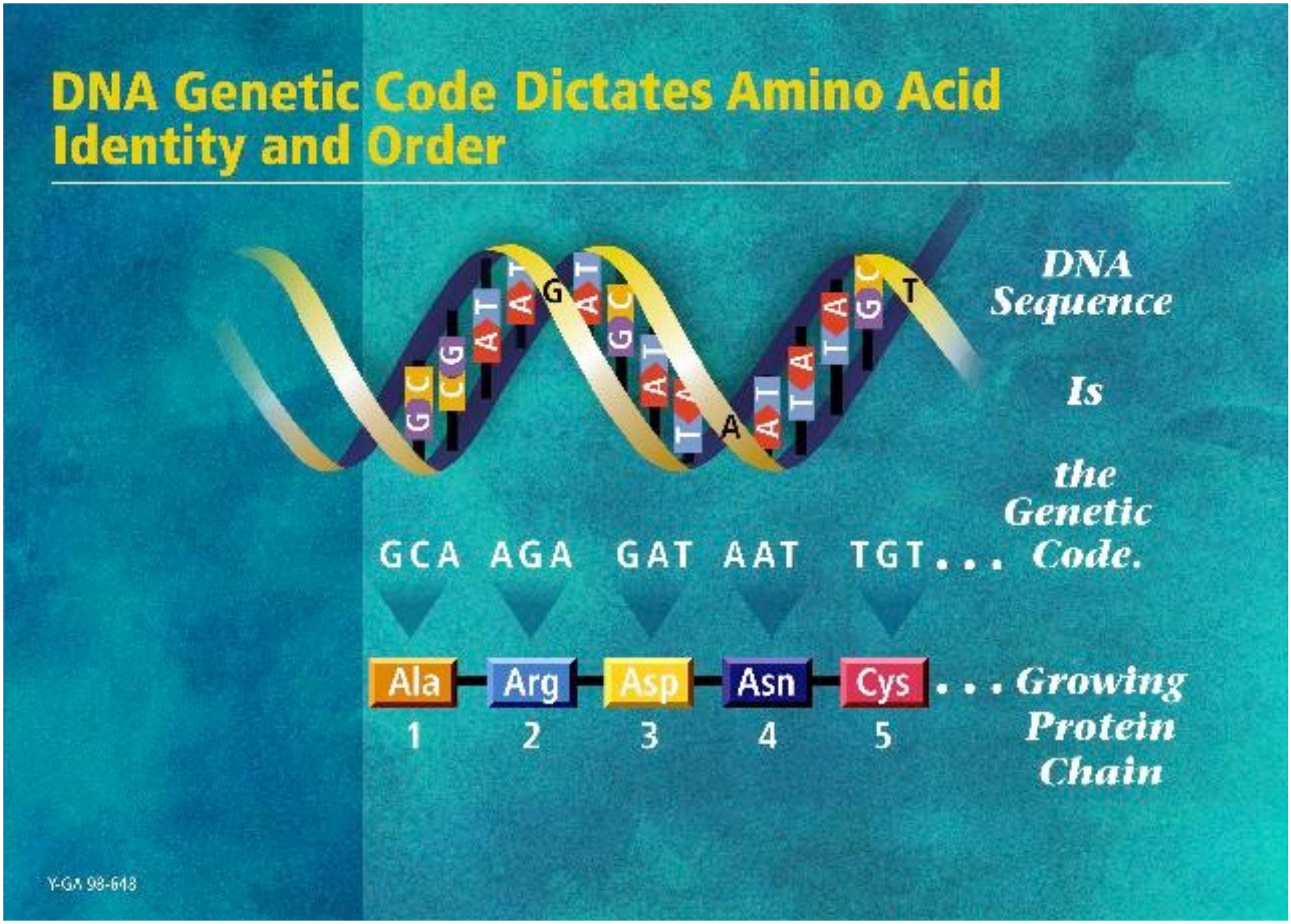
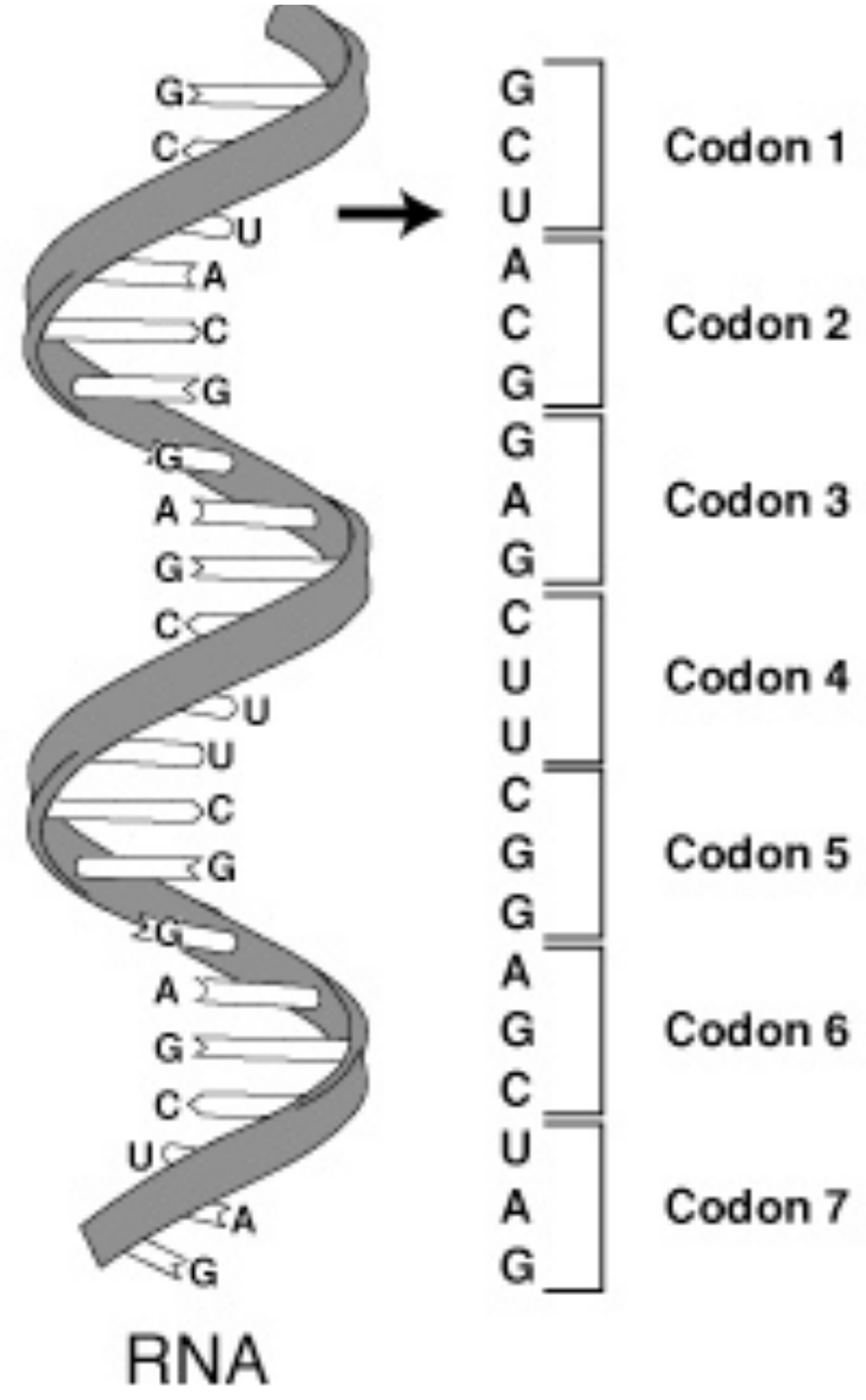


image from the DOE Human Genome Program
http://www.ornl.gov/hgmis

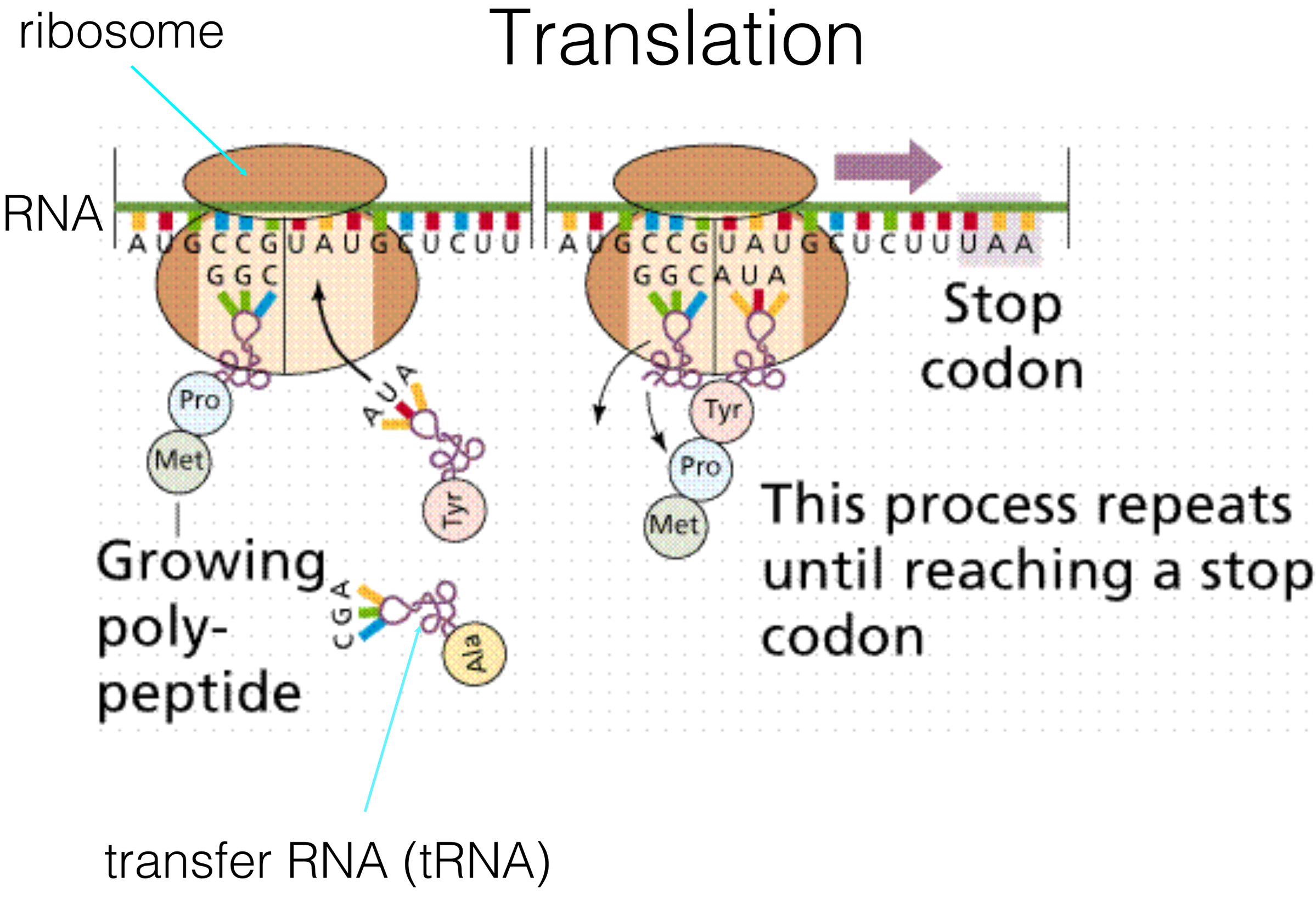
Translation

- ribosomes are the machines that synthesize proteins from mRNA
- the grouping of codons is called the reading frame
- translation begins with the start codon: AUG
- translation ends with the stop codons: UAA, UAG, UGA

Codons and Reading Frames



Translation



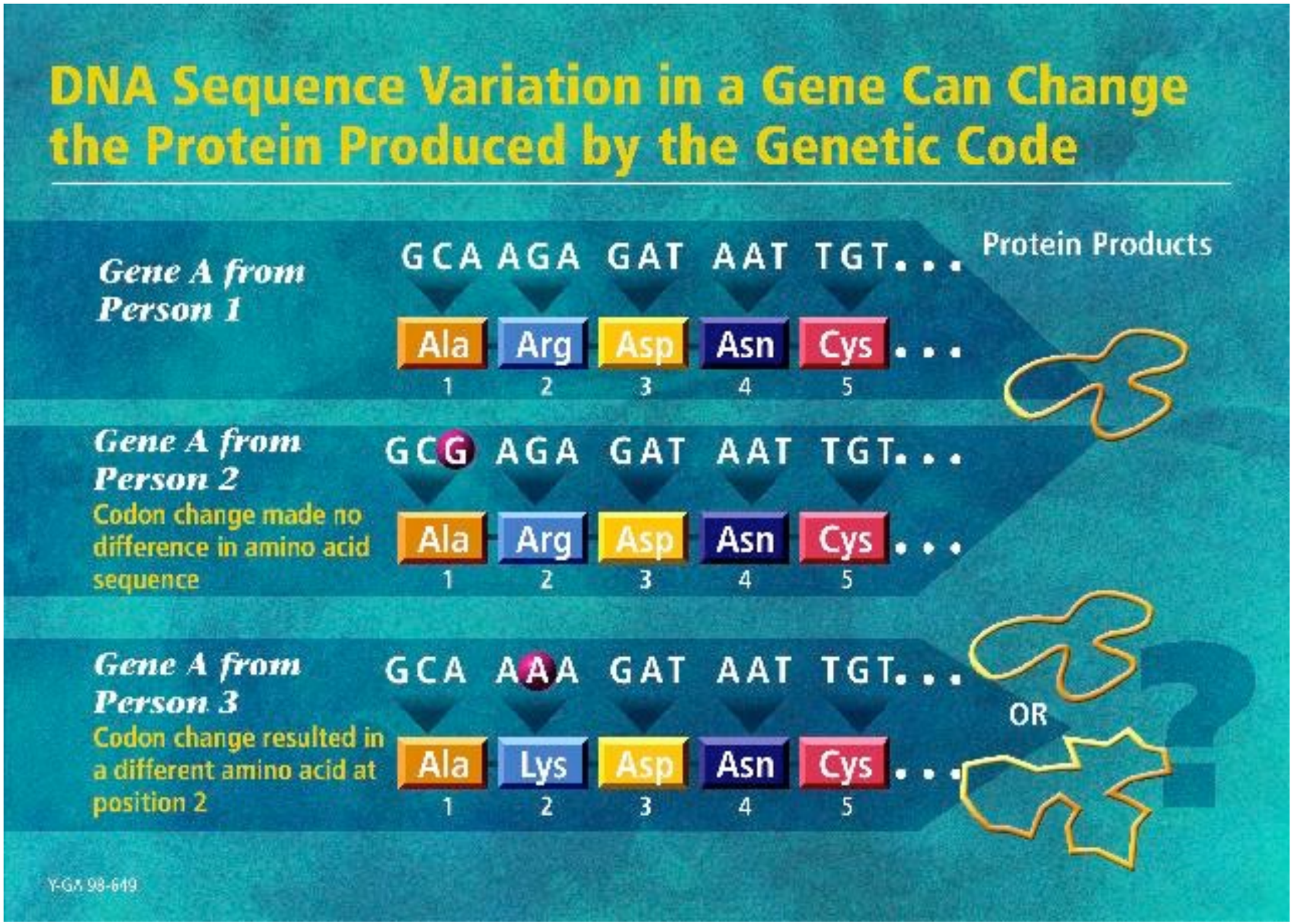


image from the DOE Human Genome Program
http://www.ornl.gov/hgmis

Genes

- genes are the basic units of heredity
- they are generally the intervals of the genome that are transcribed into RNA
- a protein-coding gene is a gene whose RNA carries the information required for constructing a particular protein (polypeptide really)
- the human genome comprises ~20,000 protein-coding genes

Gene Density

- not all of the DNA in a genome encodes protein:

bacteria	~90% coding	gene/kb
human	~1.5% coding	gene/35kb

RNA Processing in Eukaryotes

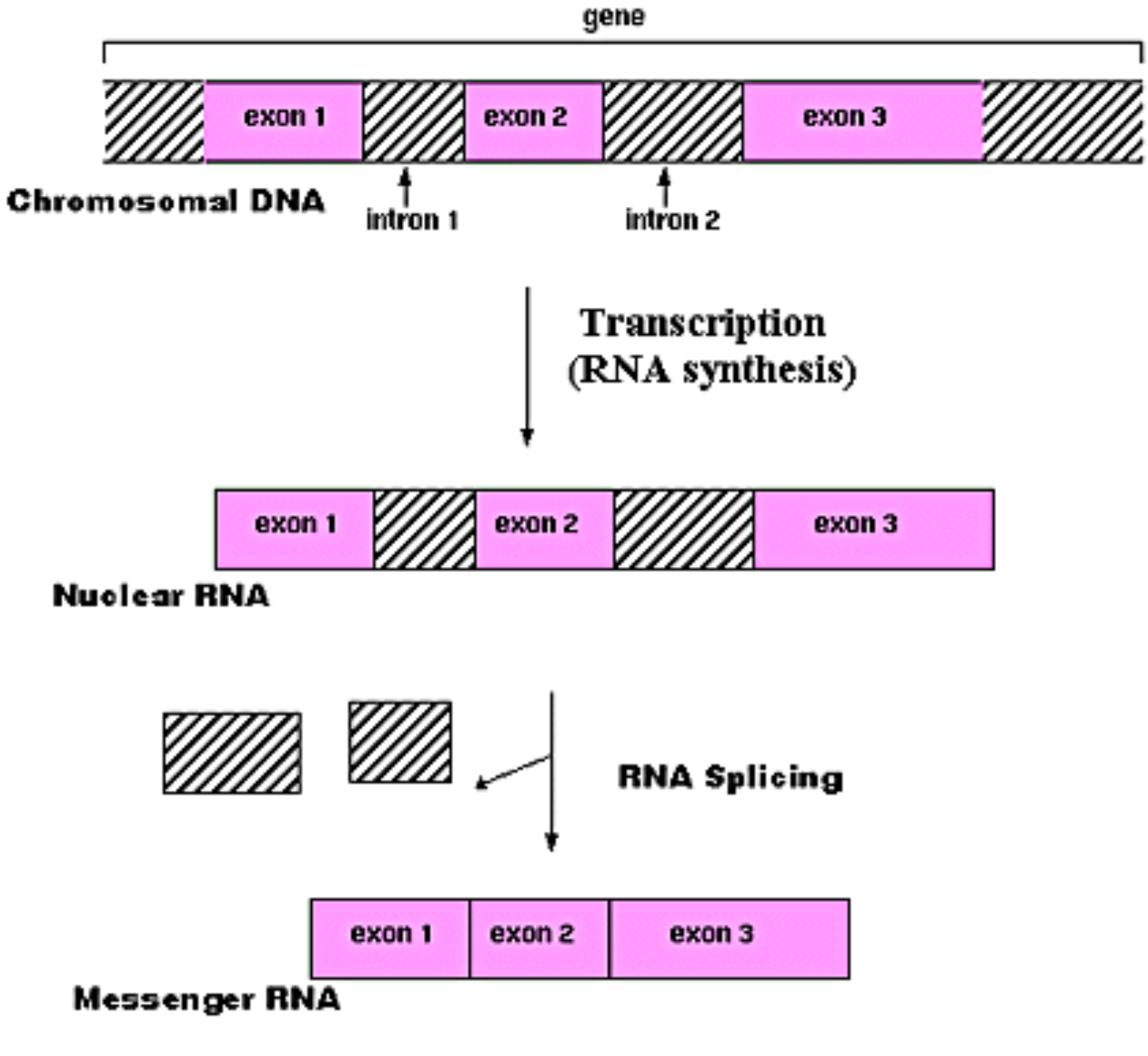
- eukaryotes are organisms that have enclosed nuclei in their cells

- in many eukaryotes, genes/mRNAs consist of alternating exon/intron segments

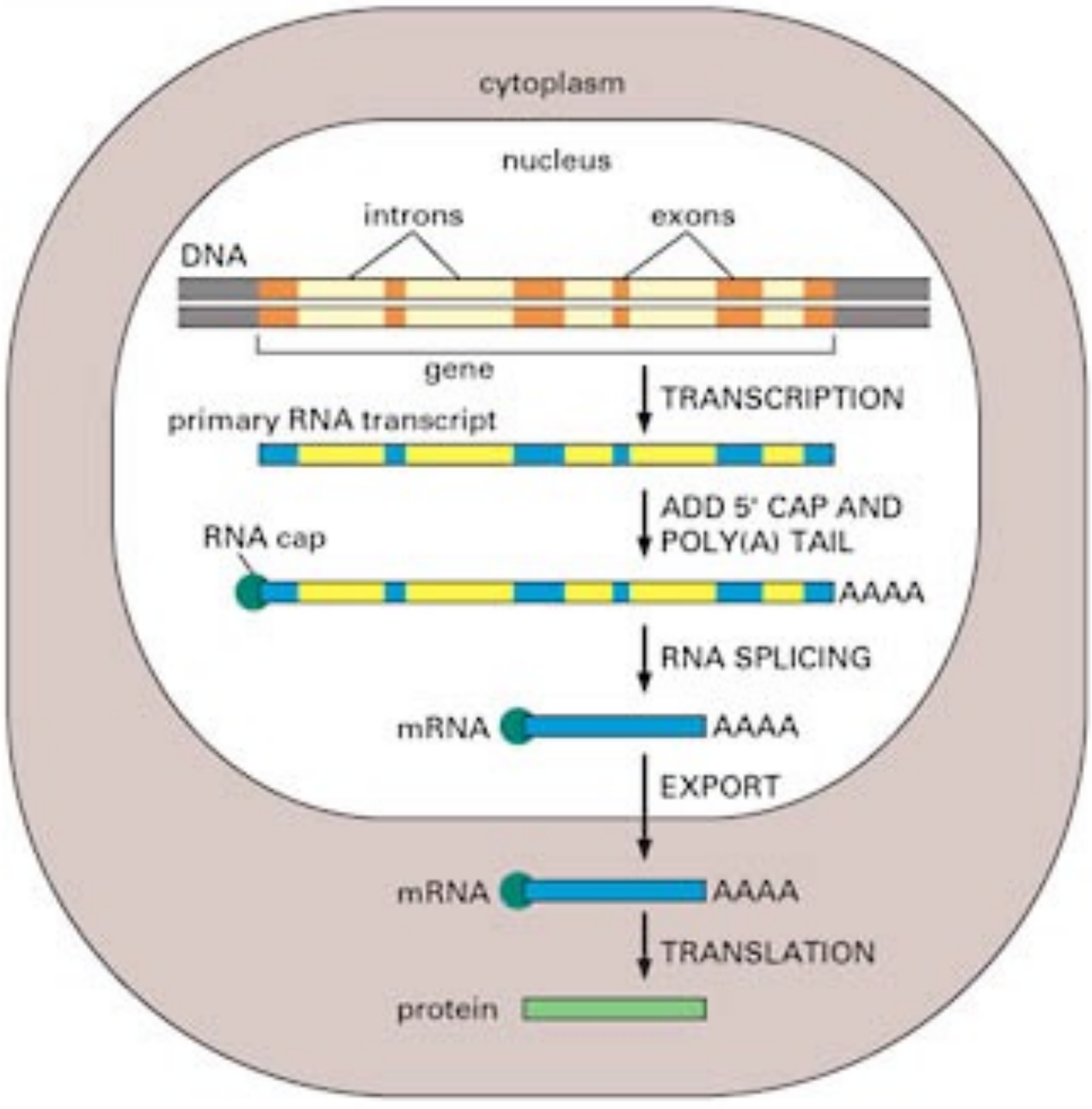
- exons are the coding parts

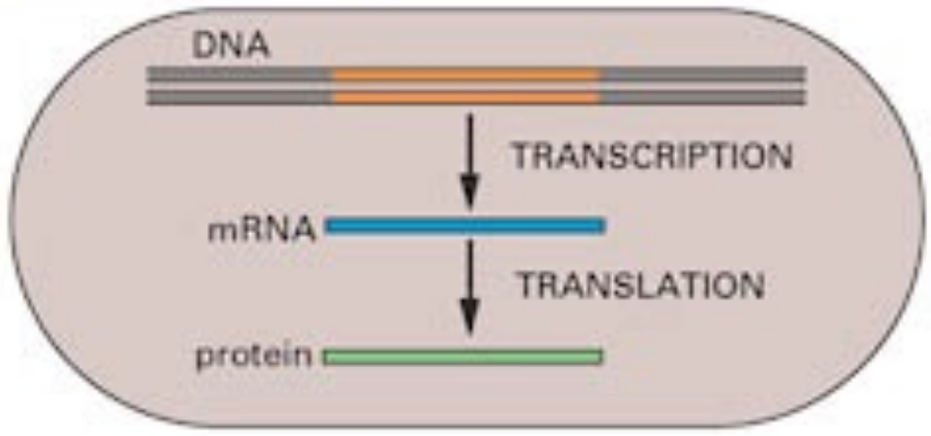
- introns are spliced out before translation

RNA Splicing



Protein Synthesis in Eukaryotes vs. Prokaryotes





RNA Genes

- not all genes encode proteins
- for some genes the end product is RNA
  - ribosomal RNA (rRNA), which includes major constituents of ribosomes
  - transfer RNAs (tRNAs), which carry amino acids to ribosomes
  - micro RNAs (miRNAs), which play an important regulatory role in various plants and animals
- etc.

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

Overview of the E. coli Metabolic Pathway Map

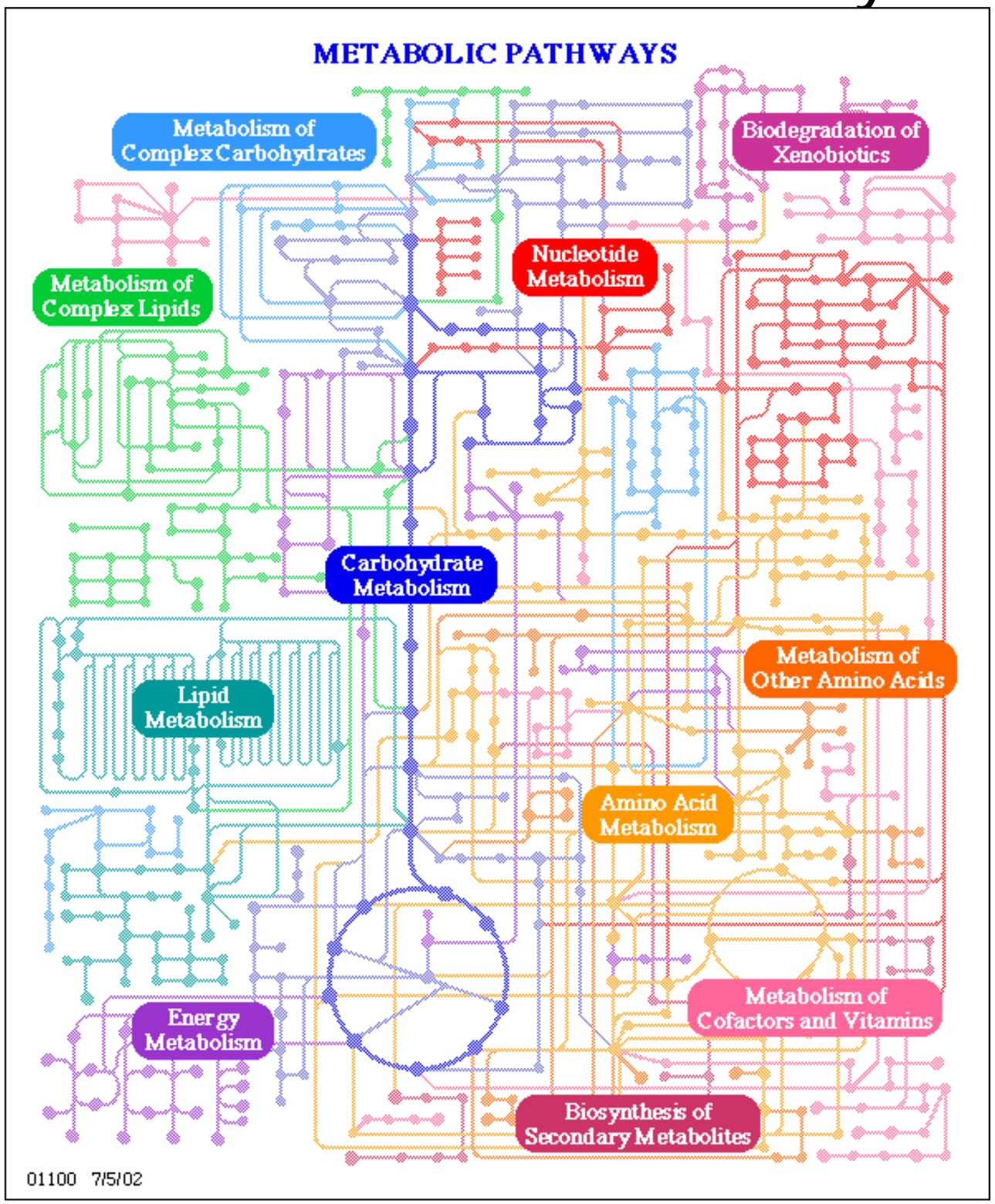
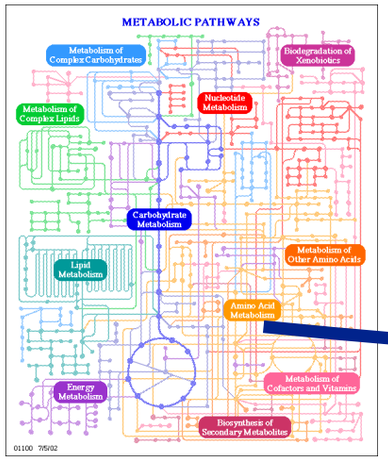


image from the KEGG database

The Metabolic Pathway for Synthesizing the Amino Acid Alanine



reactions

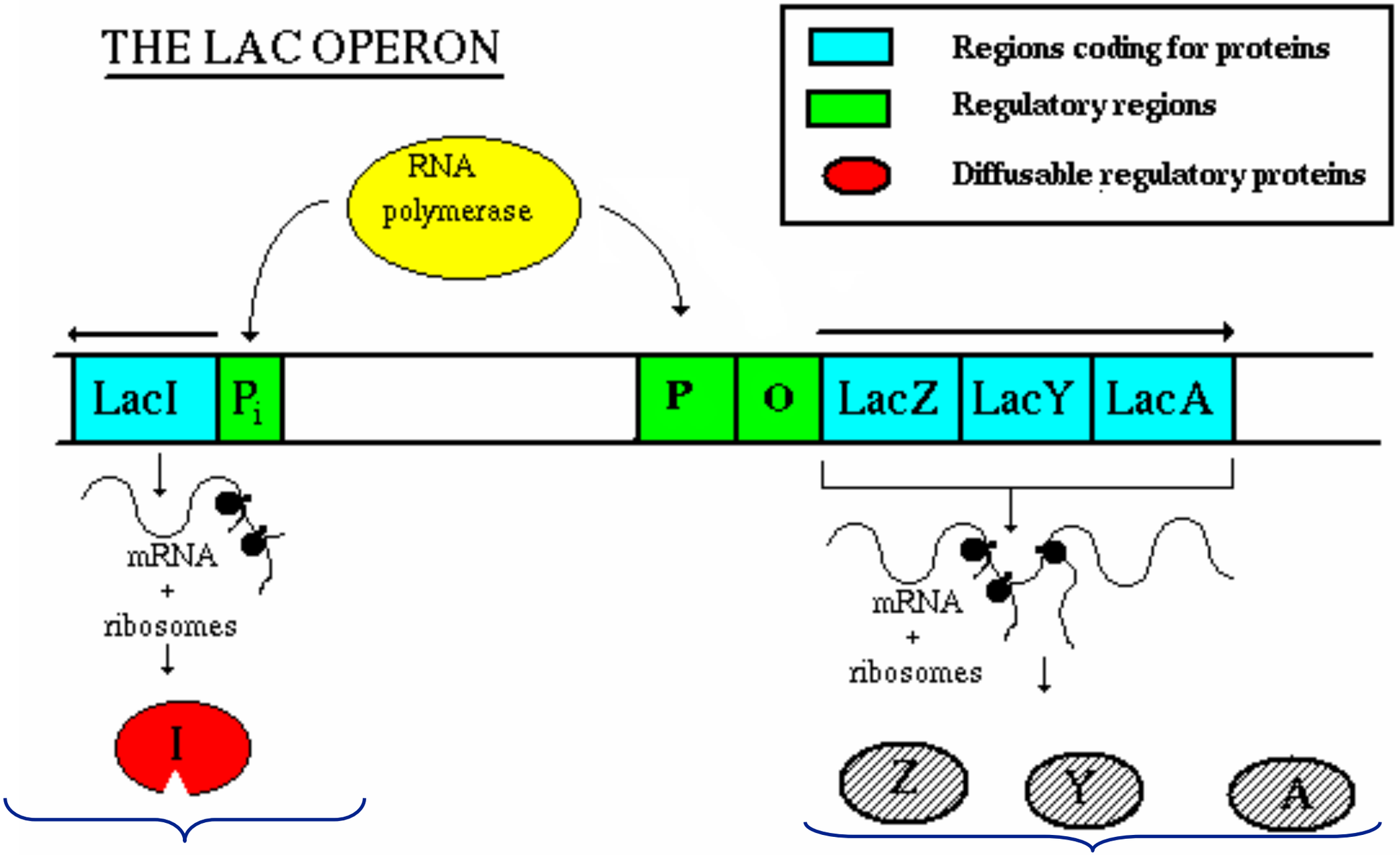
metabolites

enzymes (proteins that catalyze reactions)

genes encoding the enzymes

image from the Ecocyc database
www.biocyc.org

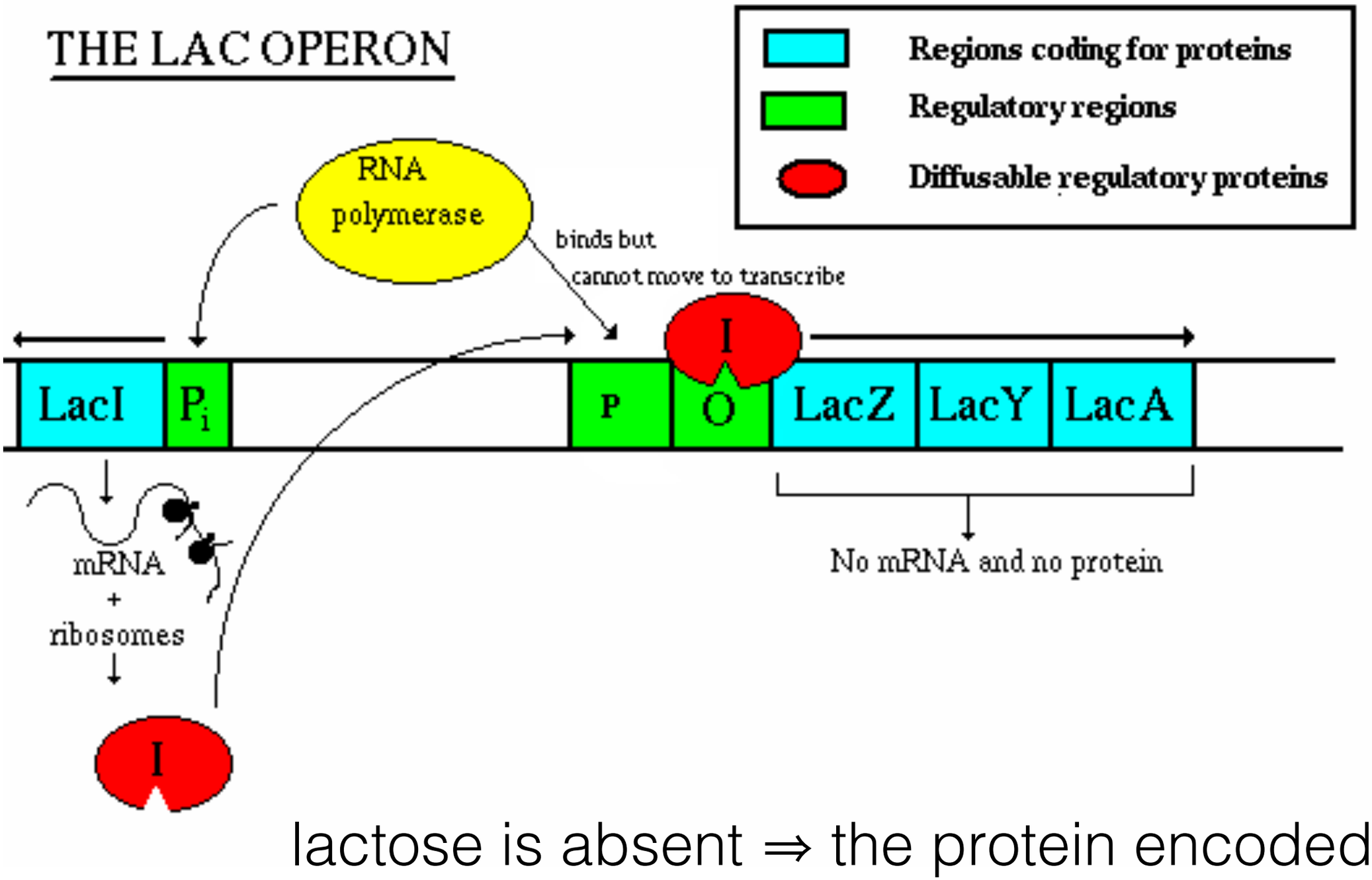
Gene Regulation Example: the lac Operon



this protein regulates the transcription of LacZ, LacY, LacA

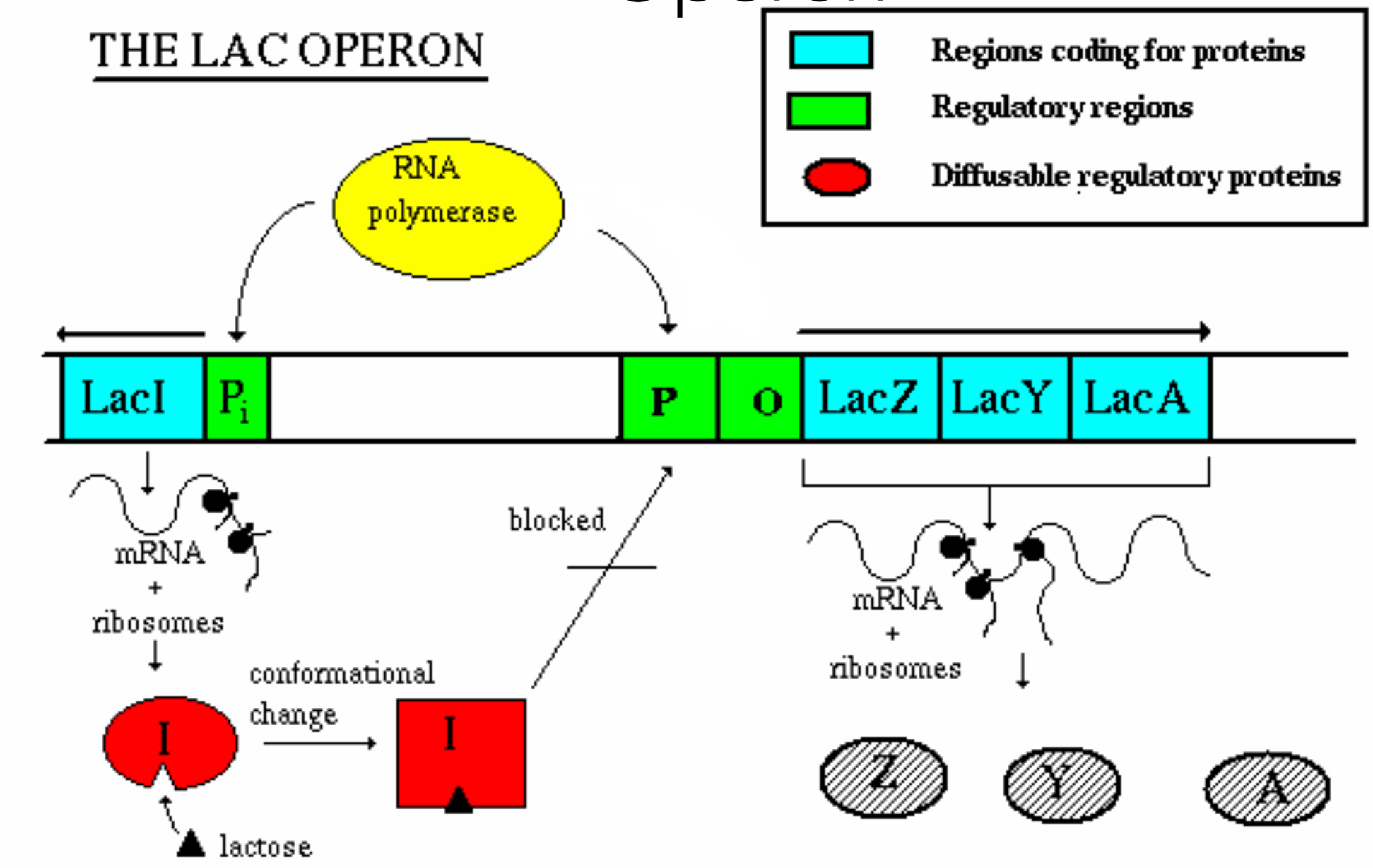
these proteins metabolize lactose

Gene Regulation Example: the lac Operon



lactose is absent ⇒ the protein encoded by lacI represses transcription of the lac operon

Gene Regulation Example: the lac Operon



lactose is present ⇒ it binds to the protein encoded by lacI changing its shape; in this state, the protein doesn't bind upstream from the lac operon; therefore the lac operon can be transcribed

Gene Regulation Example: the lac Operon

- this example provides a simple illustration of how a cell can regulate (turn on/off) certain genes in response to the state of its environment

    - an operon is a sequence of genes transcribed as a unit

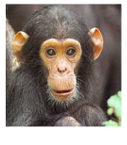
        - the lac operon is involved in metabolizing lactose

            - it is “turned on” when lactose is present in the cell

            - the lac operon is regulated at the transcription level

- the depiction here is incomplete; for example, the level of glucose in the cell also influences transcription of the lac operon

Selected milestones in genome sequencing

	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

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 (see pointer to Nucleic Acids Research directory on course home page)
- GenBank

  > 201 million sequence entries

  > 235 billion bases
- UniProtKB / Swis-Prot

  > 89 million protein sequence entries

  > 30 billion amino acids
- Protein Data Bank

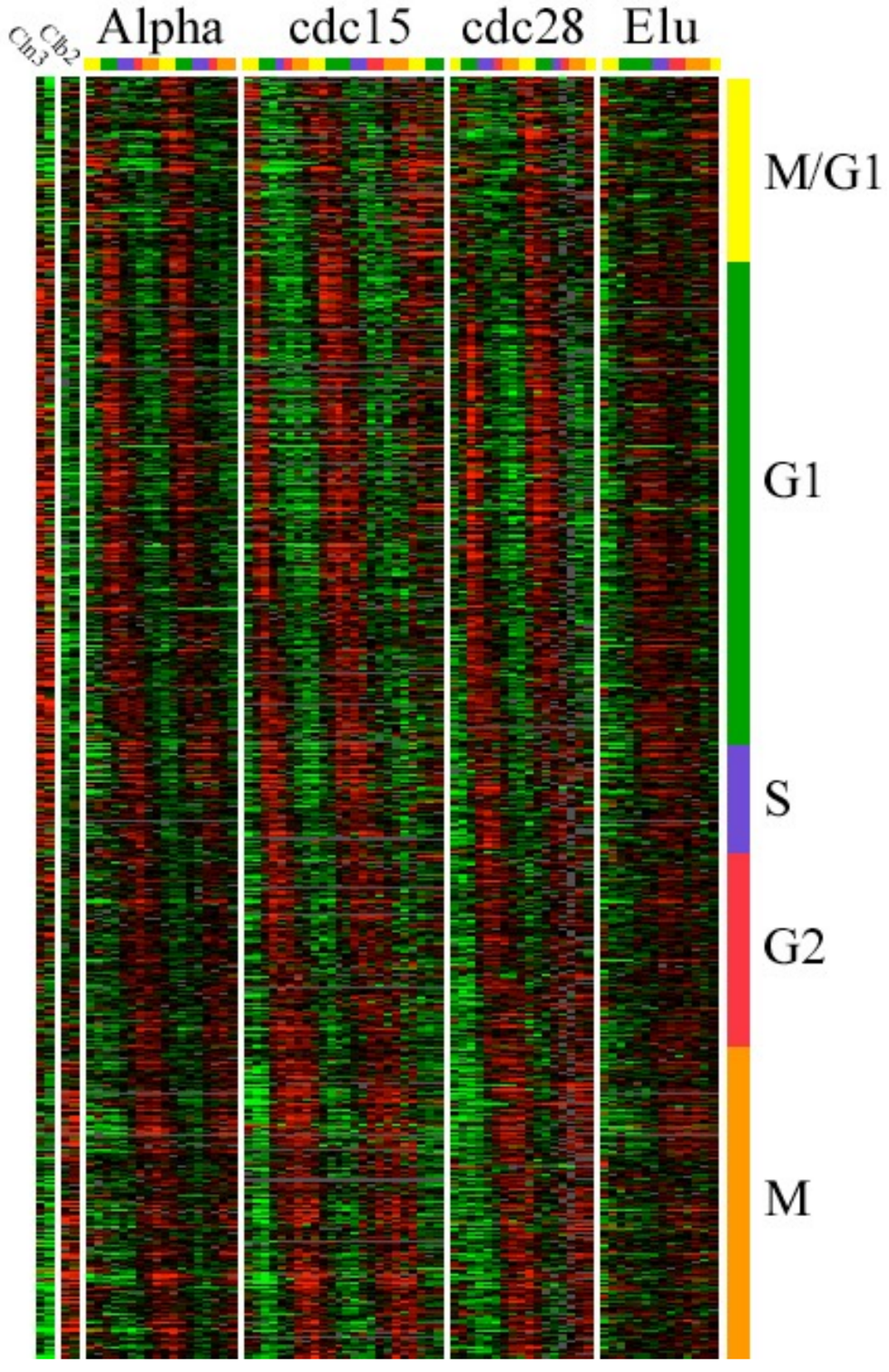
  123,837 protein (and related) structures

* all numbers current as of 9/17

More Data:
High-Throughput Experiments

- RNA abundances
- protein abundances
- small molecule abundances
- protein-protein interactions
- protein-DNA interactions
- protein-small molecule interactions
- genetic variants of an individual (e.g. which DNA base does the individual have at a few million selected positions)
- something (e.g. viral replication) measured across thousands of genetic variants
- etc.

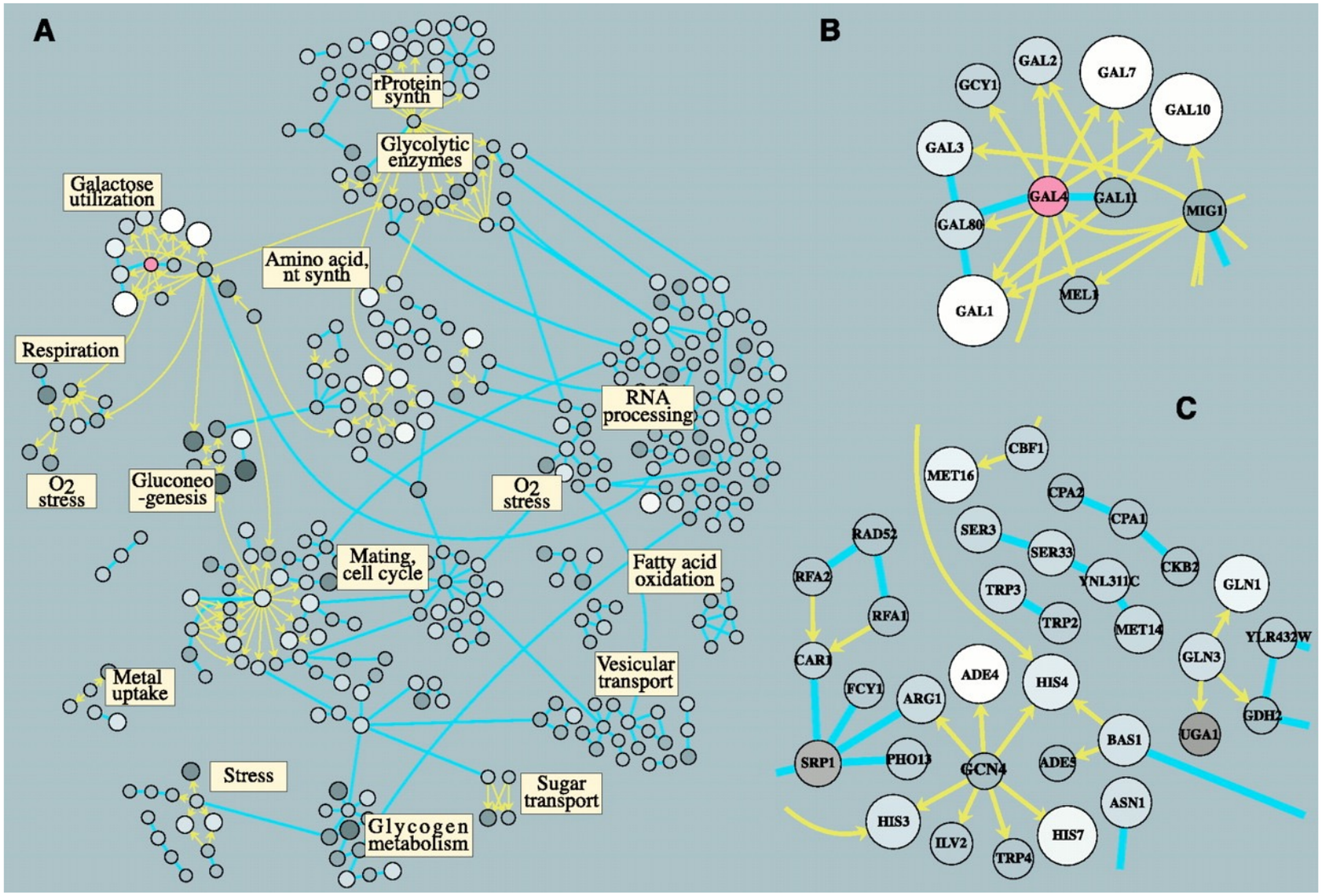
Example HT Experiment



- this figure depicts one yeast gene-expression data set
  - each row represents a gene
  - each column represents a measurement of gene expression (mRNA abundance) at some time point
    - red indicates that a gene is being expressed more than some baseline; green means less

Figure from Spellman et al., Molecular Biology of the Cell, 9:3273-3297, 1998

More Data: Interactions



- each node represents a gene product (protein)
- blue edges show direct protein-protein interactions
- yellow edges show interactions in which one protein binds to DNA and affects the expression of another

Figure from Ideker et al., Science 292(5518):929-934, 2001

Significance of the Genomics Revolution

- Data driven biology
  - functional genomics
  - comparative genomics
  - systems biology
- Molecular medicine
  - identification of genetic components of various maladies
  - diagnosis/prognosis from sequence or expression
  - gene therapy
- Pharmacogenomics
  - developing highly targeted drugs
- Toxicogenomics
  - elucidating which genes are affected by various chemicals

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