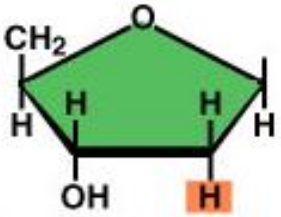
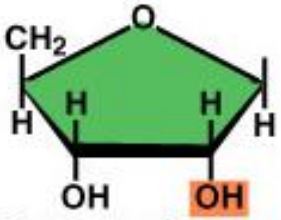
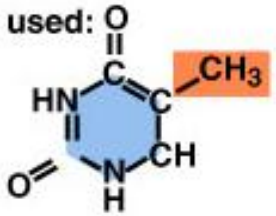
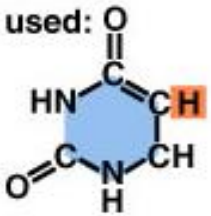
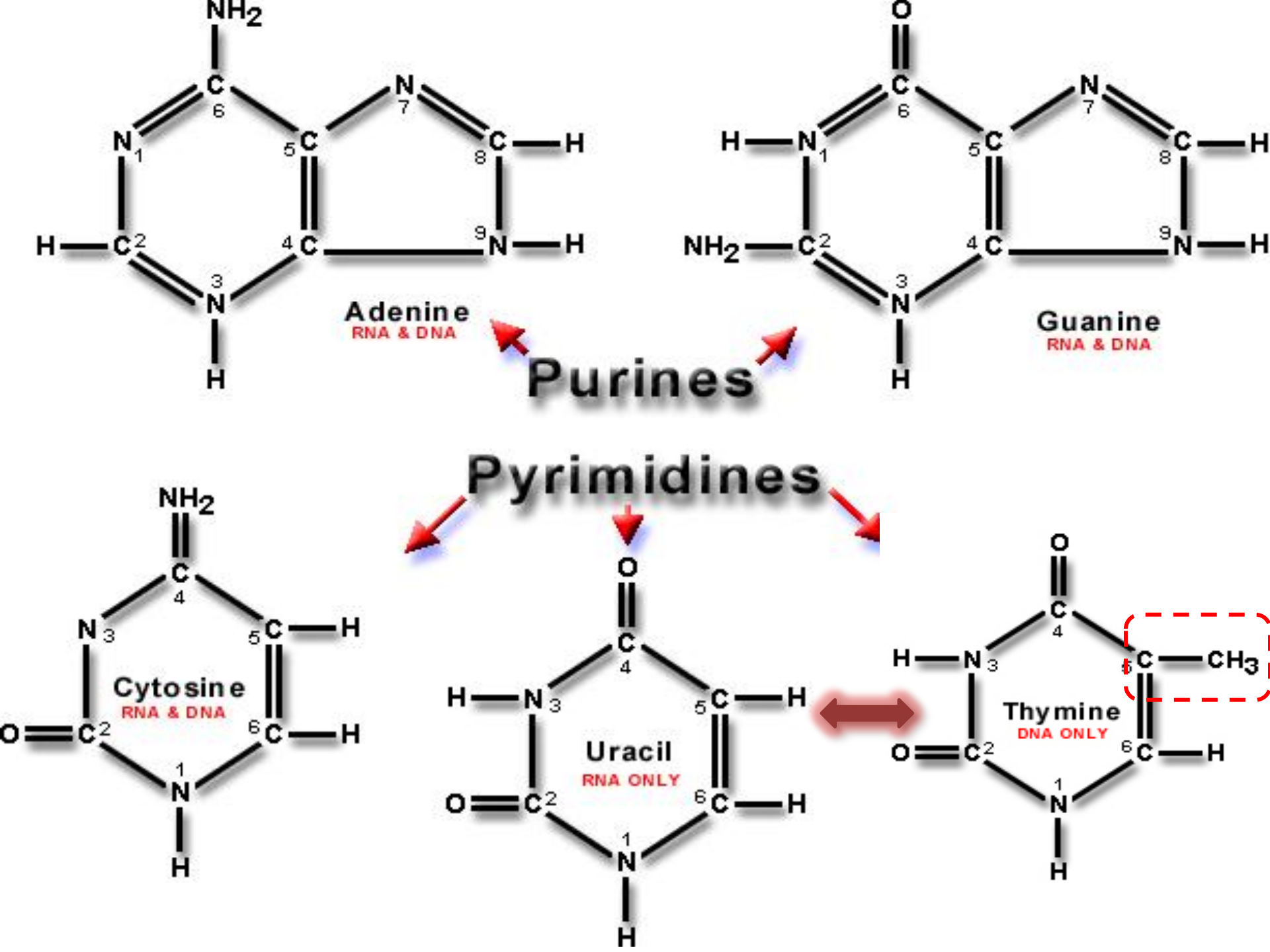
Lecture 1: From Genes to Genomes

Molecular Components of Biological System

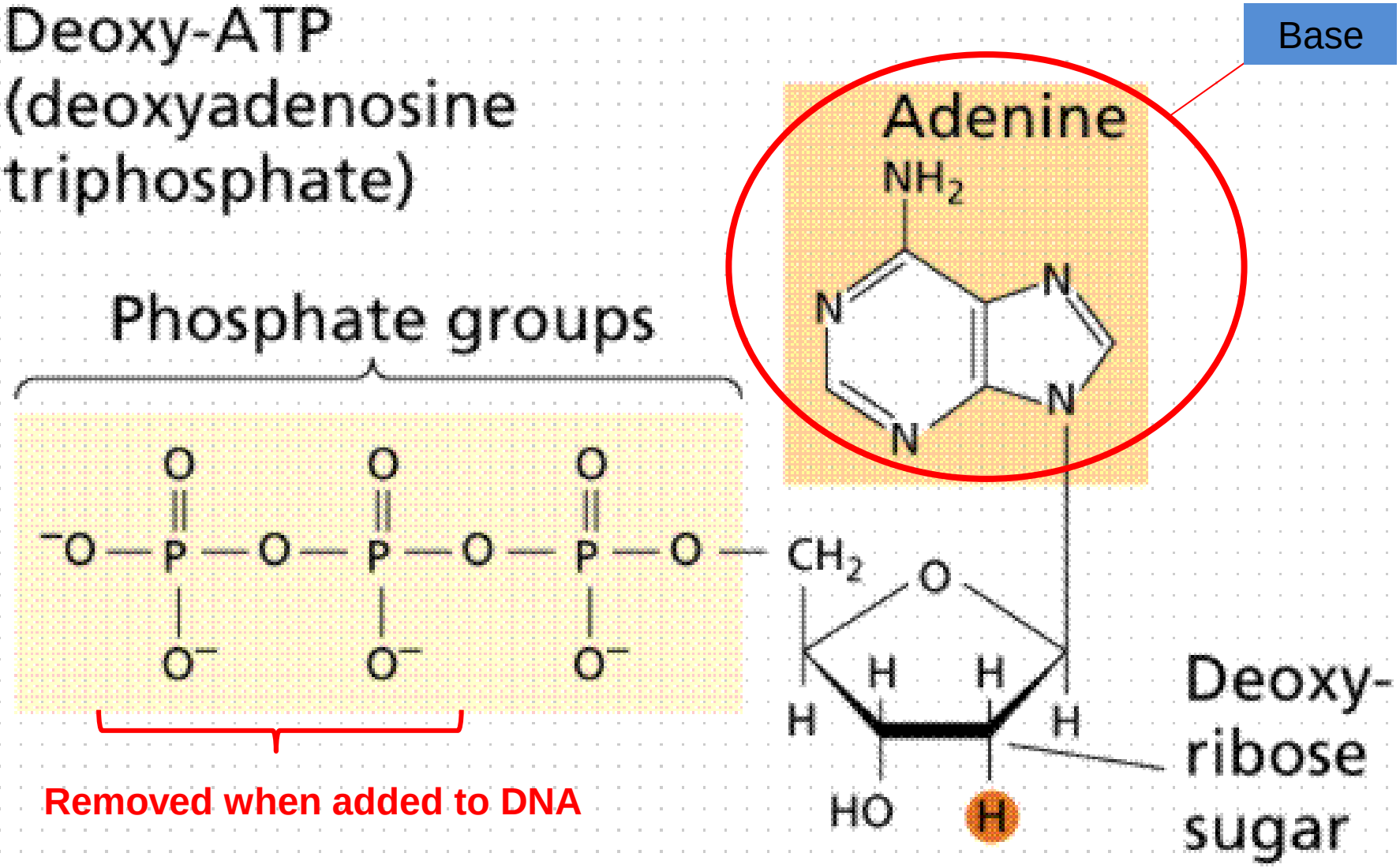
- **Nucleic acids**
  - **Deoxyribonucleic acid (DNA)**
  - **Ribonucleic acid (RNA)**
- **Proteins**
  - **Chains of amino acid residues**
  - **Single polypeptide or multiple polypeptides**
- **Lipids**
  - **Fatty acids, phospholipids and steroids**
- **Carbohydrates**
  - **Sugar, starches, and cellulose**

Two Types of Nucleic Acids

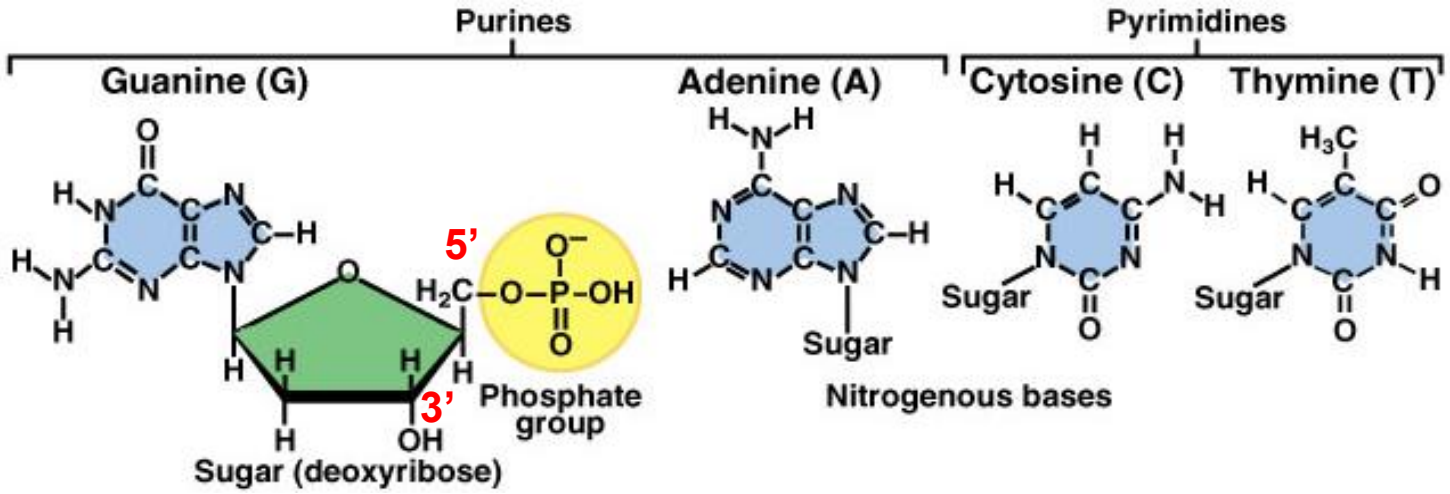
	DNA	RNA
# of strands	 Double-stranded	 Generally single-stranded
kind of sugar	 Deoxyribose as the sugar	 Ribose as the sugar
bases used	Bases used:  Thymine (T) Cytosine (C) Adenine (A) Guanine (G)	Bases used:  Uracil (U) Cytosine (C) Adenine (A) Guanine (G)
	• Carries RNA-encoding information • Not catalytic	• Carries Protein-encoding information • Can be catalytic

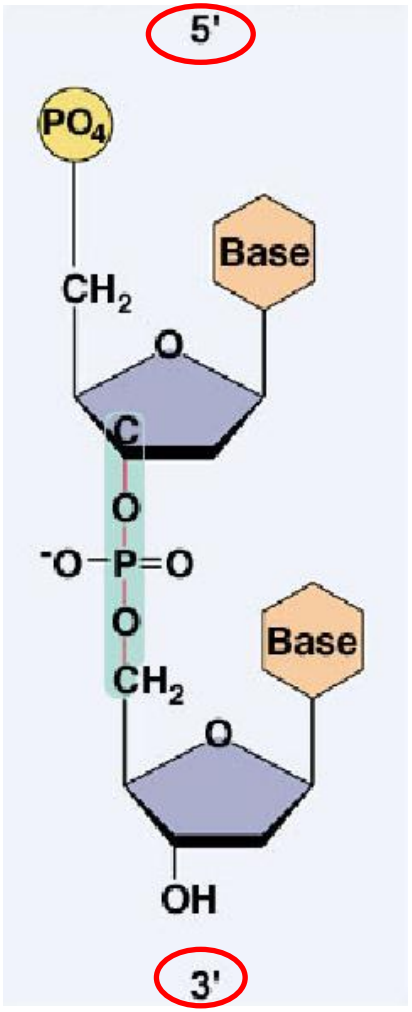

Adenine
RNA & DNA
Guanine
RNA & DNA
Purines
Pyrimidines
Cytosine
RNA & DNA
Uracil
RNA ONLY
Thymine
DNA ONLY

DNA Nucleotides



DNA (deoxyribonucleic acid) is a chain of nucleotides.





Formation of phosphodiester bond results in extension of the DNA chain.

DNA bases pair via hydrogen bonds.

Erwin Chargaff observed:

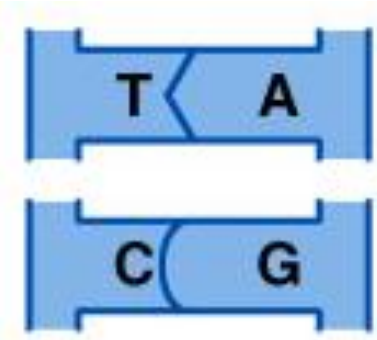
# of adenine = # of thymine

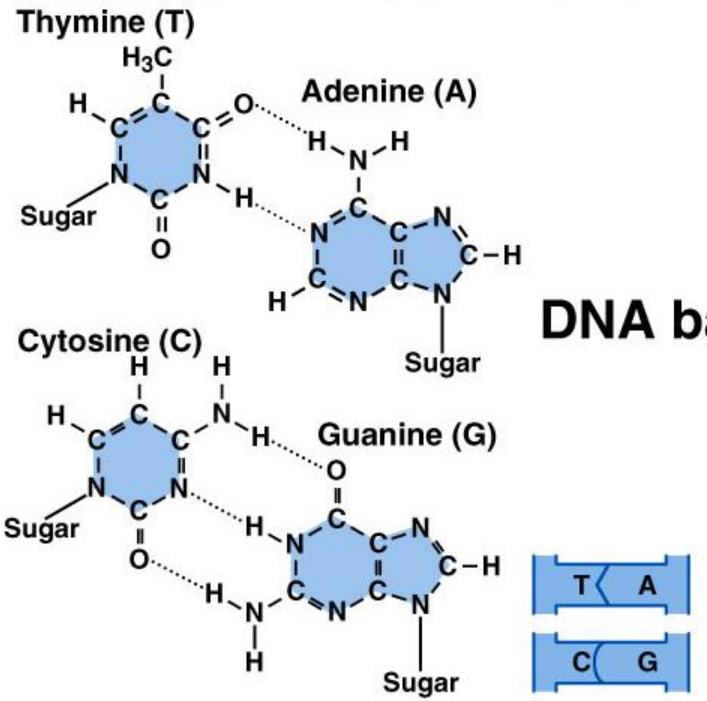
# of guanine = # of cytosine

Complementary bases pair:

A and T pair

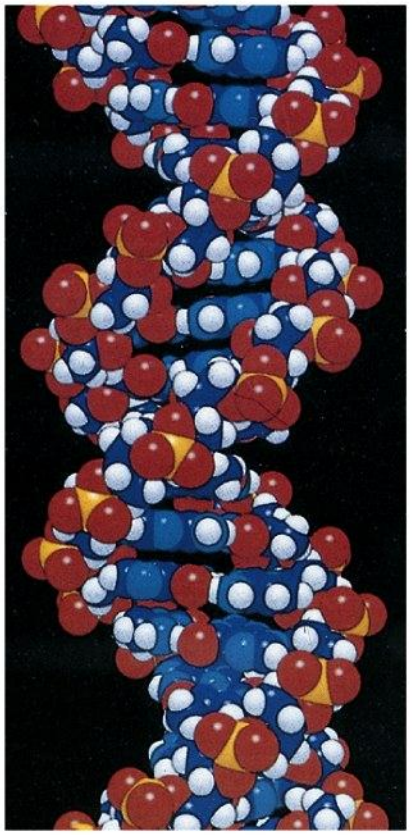
C and G pair



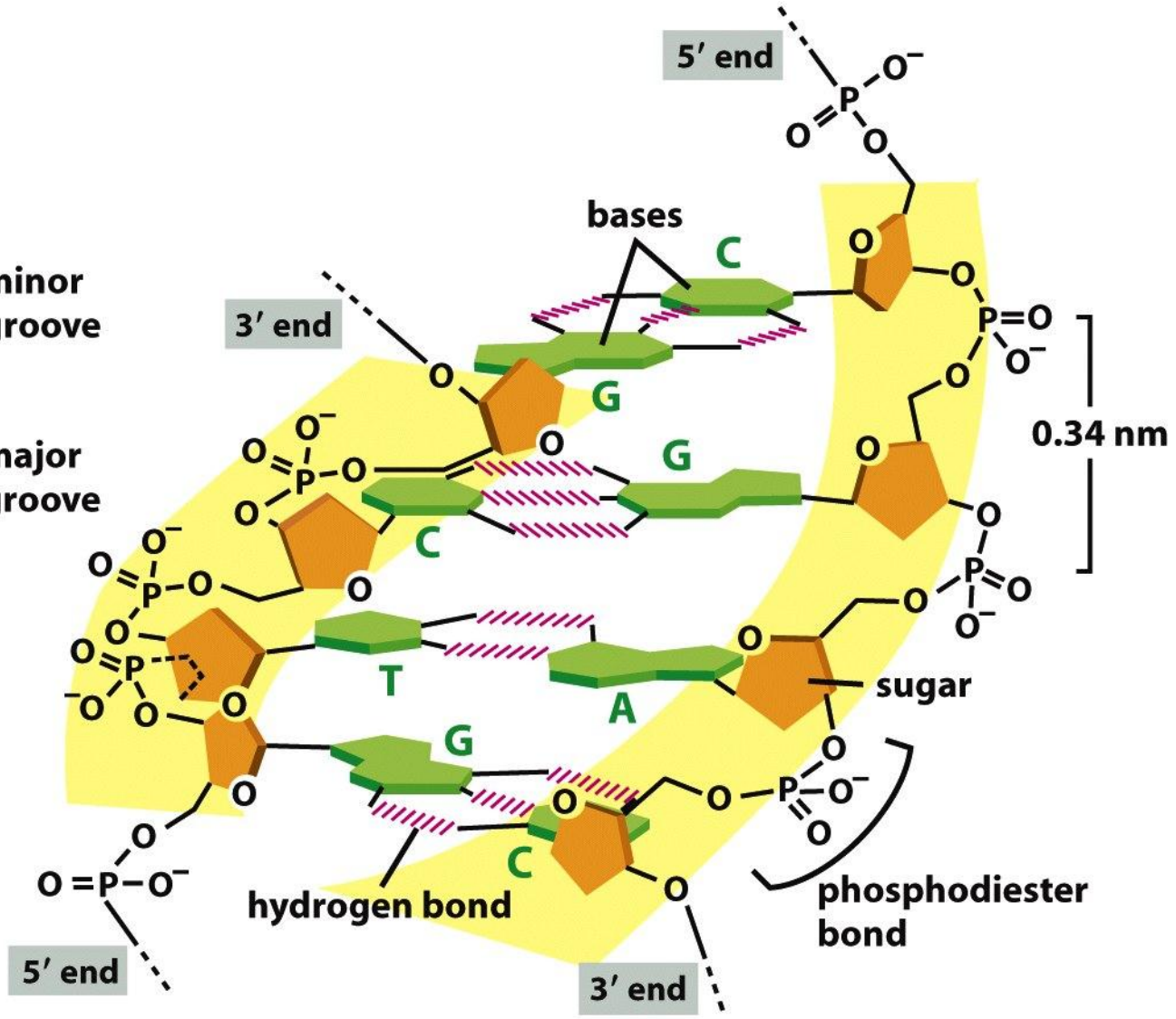


DNA base pairs

The DNA Double Helix



A Space-filling Model of 1.5 Turns



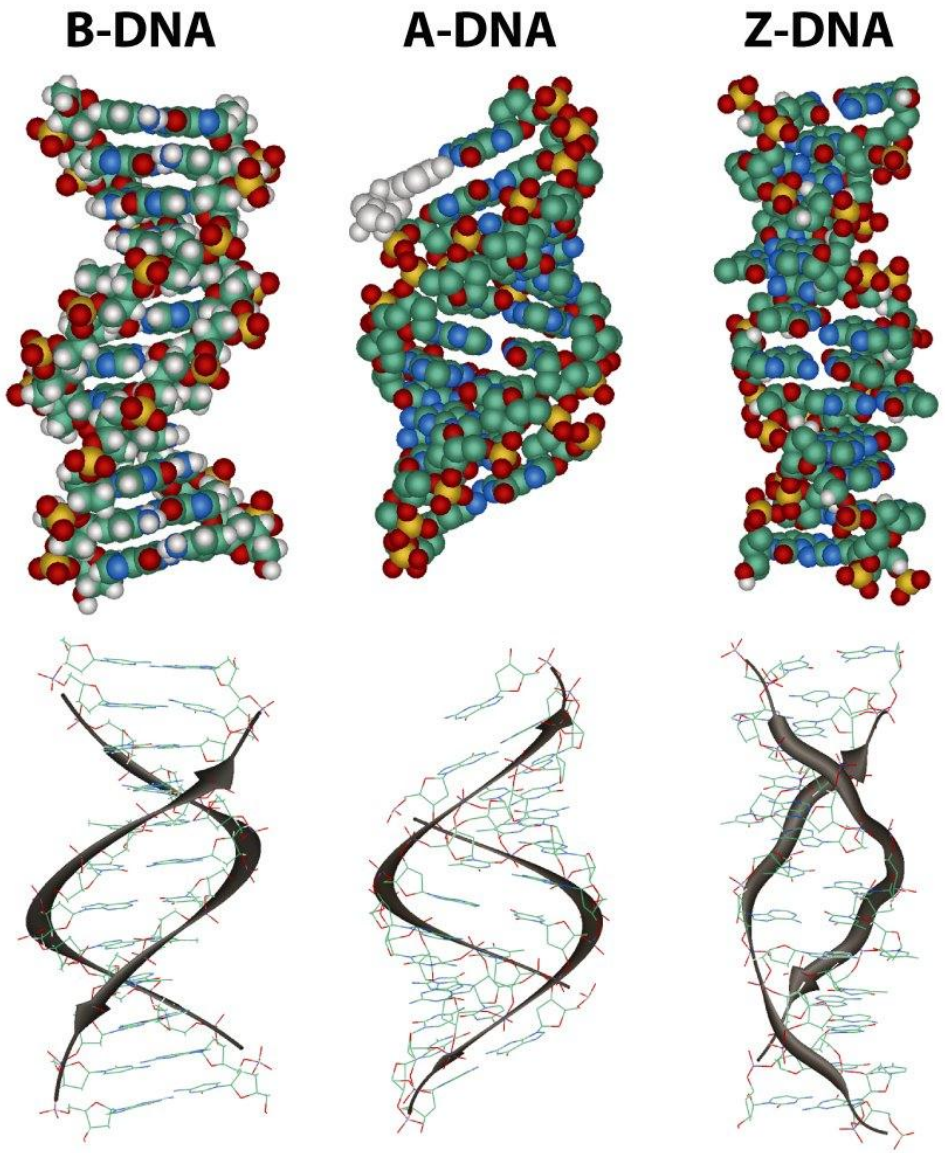
Short Section Viewed From Its Side

Nucleobases, Nucleosides and Nucleotides

- **Nucleobases (NB)**
  - ✓**purines (adenine, guanine); pyrimidines (uracil, thymine, cytosine)**
- **Nucleosides (NS): NB + pentose**
  - ✓**Ribonucleosides: adenosine, guanosine, uridine, cytidine**
  - ✓**Deoxyribonucleosides: deoxyadenosine, deoxyguanosine, thymidine, deoxyuridine, deoxycytidine**
- **Nucleotides: NS + phosphate**
  - ✓**Ribonucleotides: monophosphates (AMP, GMP, UMP, CMP); diphosphates (ADP, GDP, UDP, CDP); triphosphates (ATP, GTP, UTP, CTP)**
  - ✓**Deoxyribonucleotides: monophosphates (dAMP, dGMP, dUMP, TMP, dCMP); diphosphates (dADP, dGDP, TDP, dCDP); triphosphates (dATP, dGTP, TTP, dCTP)**
- **Cyclic**
  - ✓**cAMP, cGMP, c-di-GMP (cyclic diguanylate – involved in signal transduction), cADPR (cyclic adenosine diphosphoribose – a regulator of calcium signaling)**

Three Different Structures of DNA

- **B form – present in most DNA at neutral pH and physiological salt concentrations. The helix makes a right-handed turn every 3.4 nm, and the distance between two neighboring base pairs is 0.34 nm. The intertwined strands make two grooves of different width, i.e., major groove and minor groove, which may facilitate binding of specific proteins.**
- **A form – In a solution with higher salt concentration or with alcohol added, the DNA structure may change to an A form, which is still right-handed, but every 2.3 nm makes a turn and there are 11 base pairs per turn.**
- **Z form – formed by stretches of alternating purines and pyrimidines. Z DNA is left-handed. DNA with alternating G-C sequences in alcohol or high salt solution tends to have such structure.**



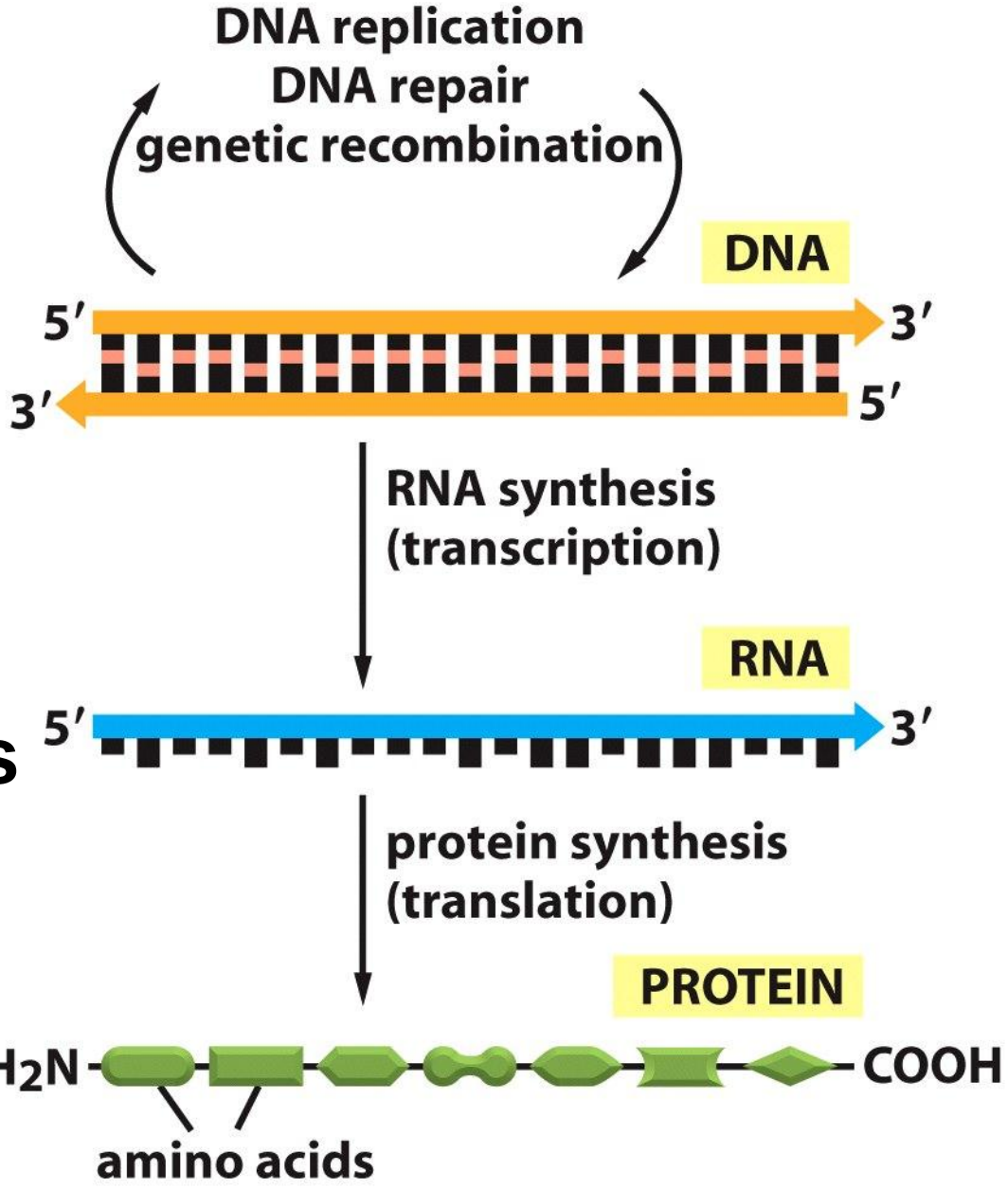
Differences include helical diameter, number of base pairs per complete turn and topology of the major and minor grooves.

Gene: molecular definition

- **A gene is a segment of DNA that directs the formation of RNA which, in turn, directs formation of a protein.**
- **The protein (or functional RNA) creates the phenotype.**
- **Information is conveyed by the sequence of the nucleotides.**

The Pathway from DNA to Protein

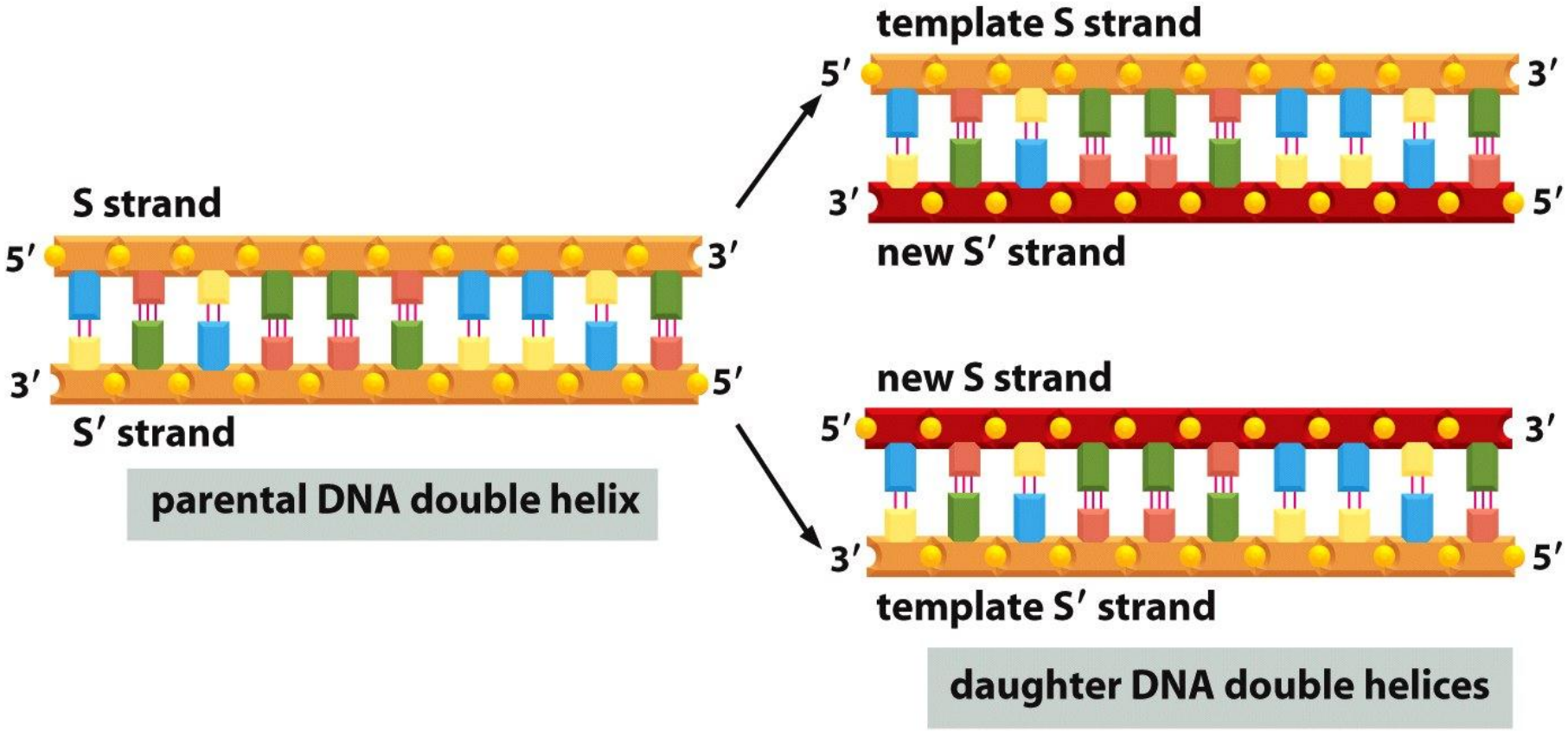
The flow of genetic information from DNA to RNA (transcription), and from RNA to protein (translation) occurs in all living cells.



Sense and Antisense Strands of DNA

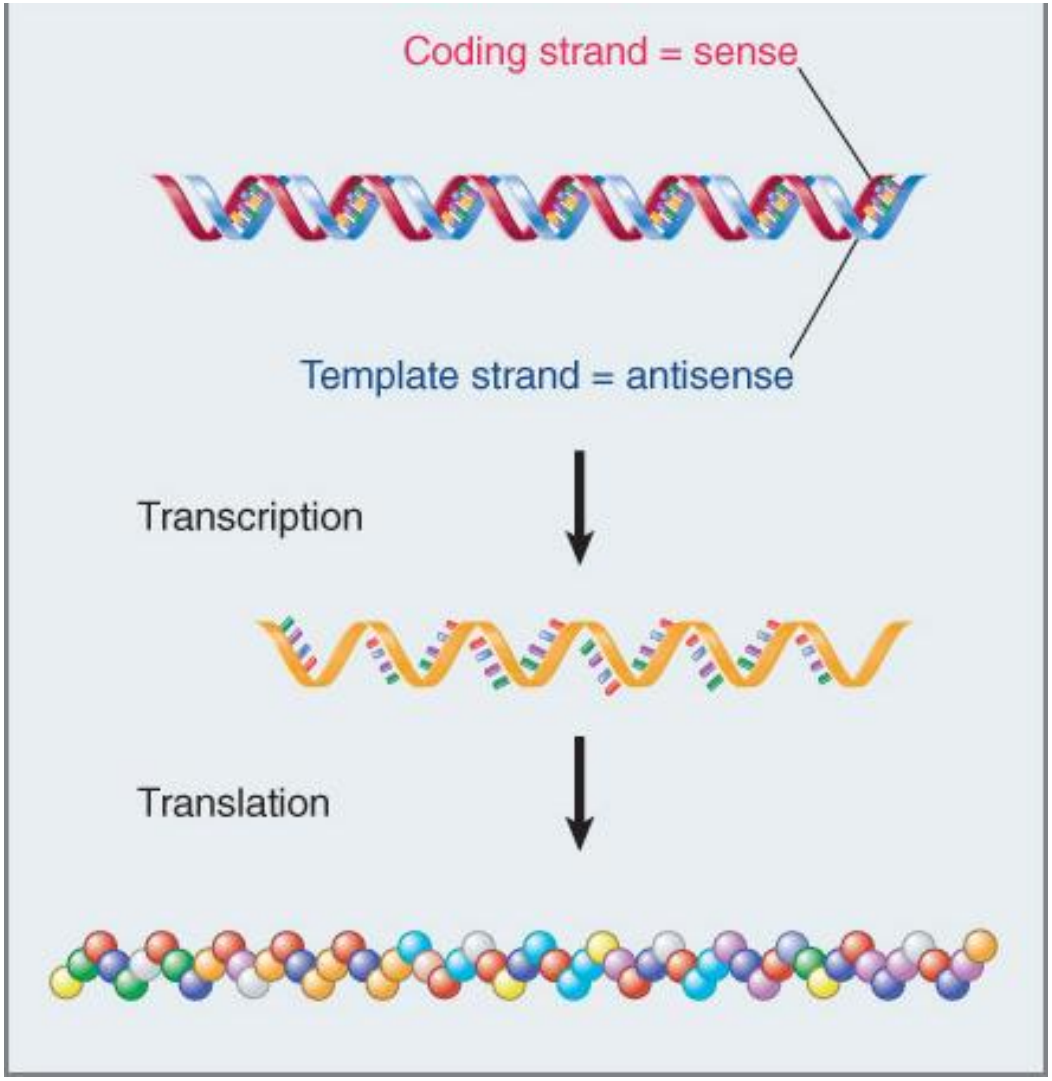
- The coding DNA sequence is called "sense“ [or positive (+) sense] if its sequence is the same as that of a messenger RNA copy that is translated into protein.
- The sequence on the opposite strand (template sequence) is its complement and is called the "antisense" [or negative (-) sense] sequence.
- Antisense molecules interact with complementary strands of nucleic acids, modifying expression of genes.
- The template DNA strand is called the transcribed strand with antisense sequence and the mRNA transcript is said to be sense sequence (the complement of antisense).
- Because DNA is double-stranded, the strand complementary to the antisense sequence is called the non-transcribed strand and has the same sense sequence as the mRNA transcript.
- Both sense and antisense sequences can exist on different parts of the same strand of DNA, i.e., both strands contain both sense and antisense sequences).
- In both prokaryotes and eukaryotes, antisense RNA sequences are produced, but the functions of these RNAs are not entirely clear.

DNA as a Template for Its Own Duplication (Replication)

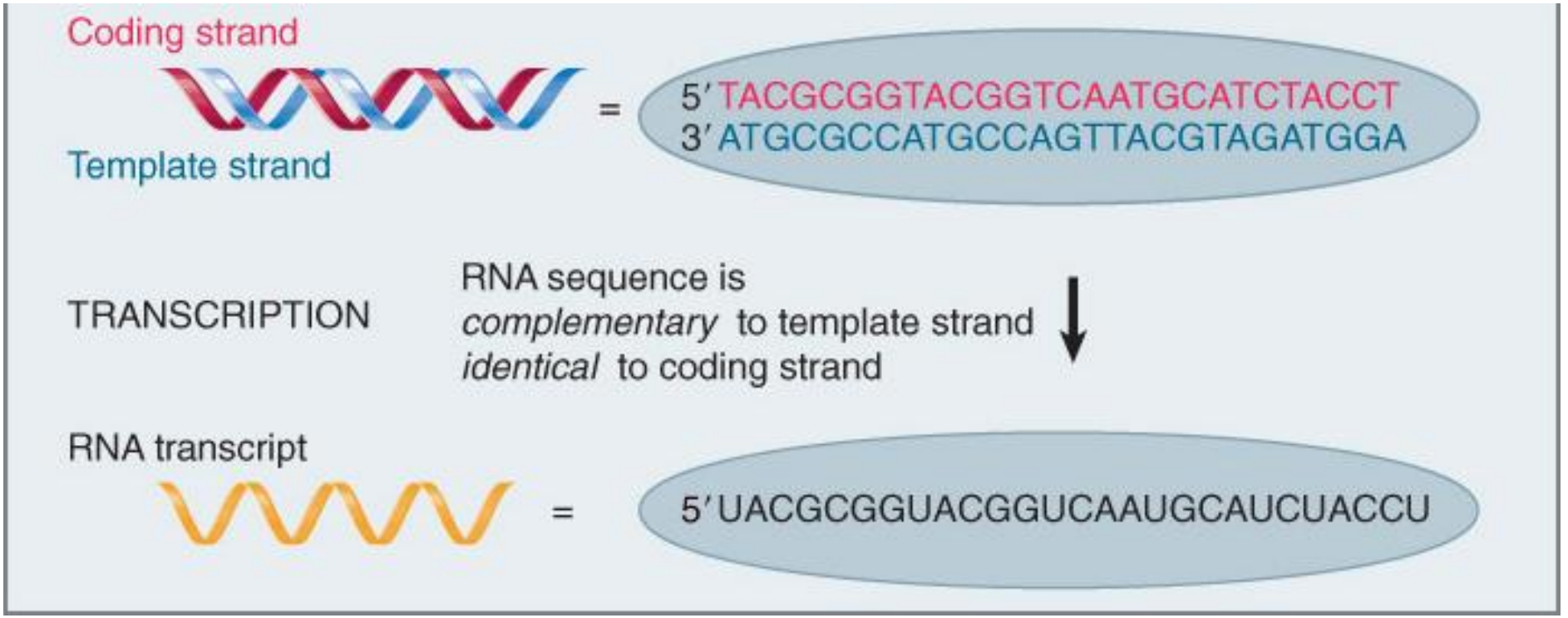


Gene expression = transcription + translation.

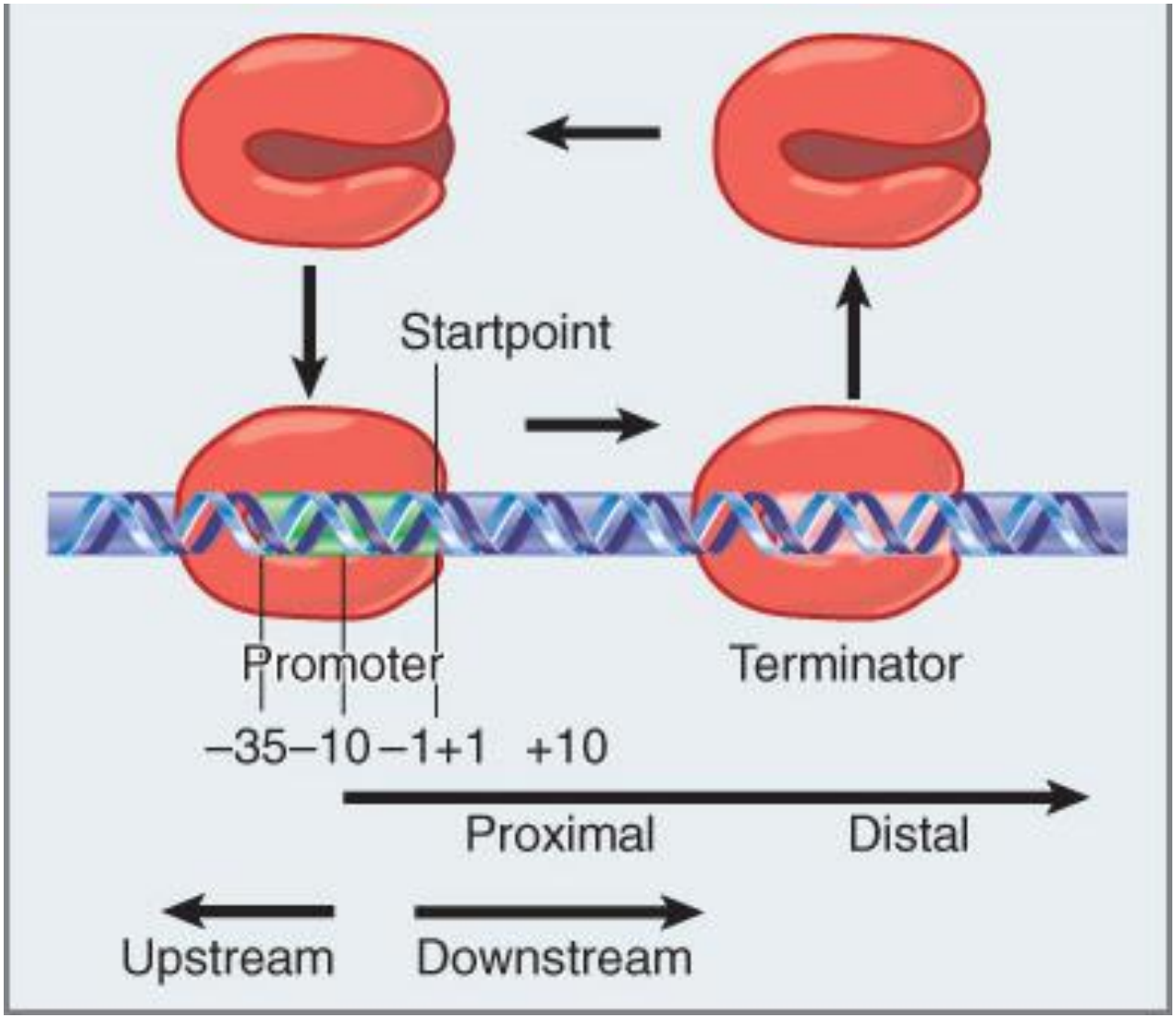
- Trancription generates an RNA that is complementary to the DNA template strand and has the same sequence as the DNA coding strand.
- Translation reads a triplet of bases into one amino acid.
- Three turns of the DNA double helix contain 30 bp, which code for ten amino acids.



One strand of DNA is transcribed into RNA.



Promoters and terminators define the transcription unit.

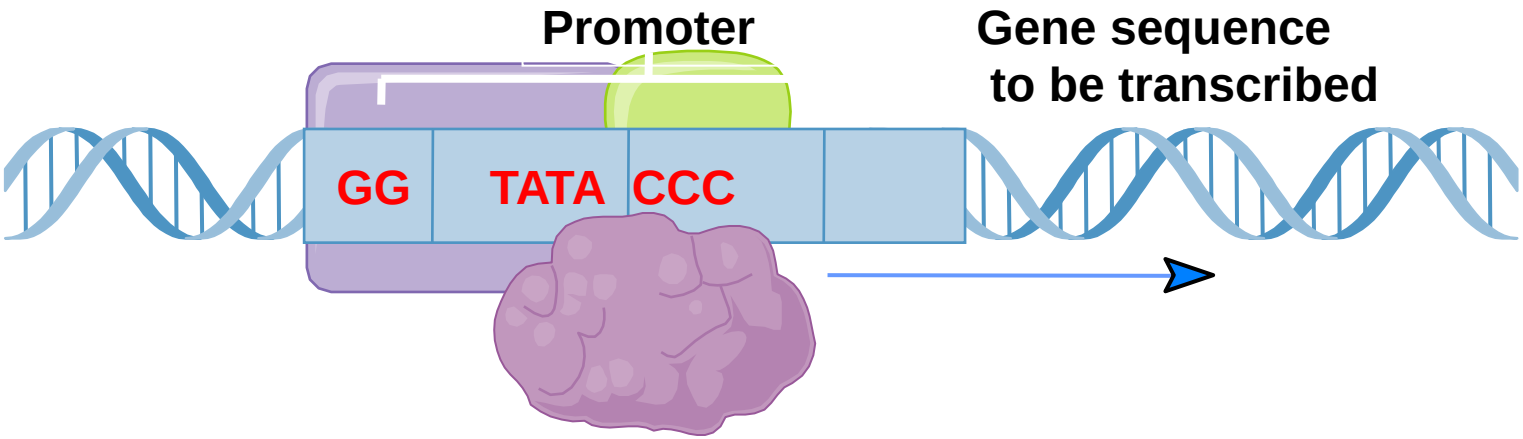


A transcription unit is a sequence of DNA transcribed into a single RNA, starting at the promoter and ending at the terminator.

Initiation of Transcription

Transcription begins at the 5′ end of the gene in a region called the promoter.

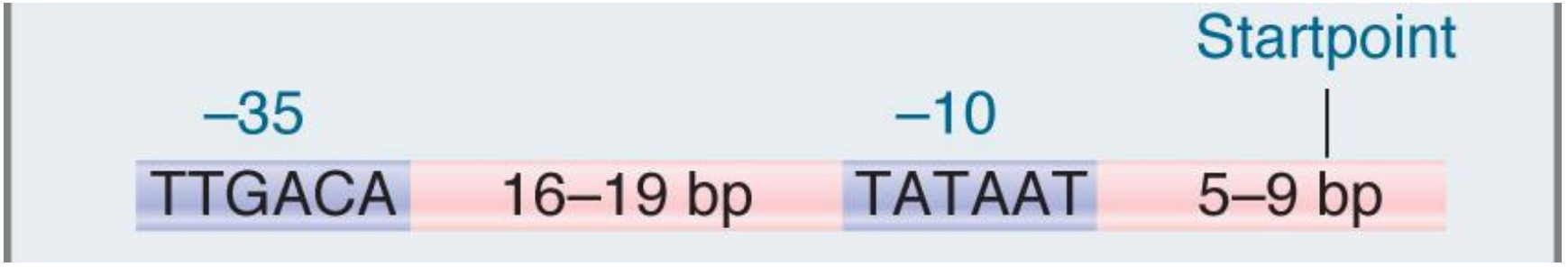
The promoter recruits TATA binding protein, a DNA binding protein, which in turn recruits other proteins including RNA polymerase.

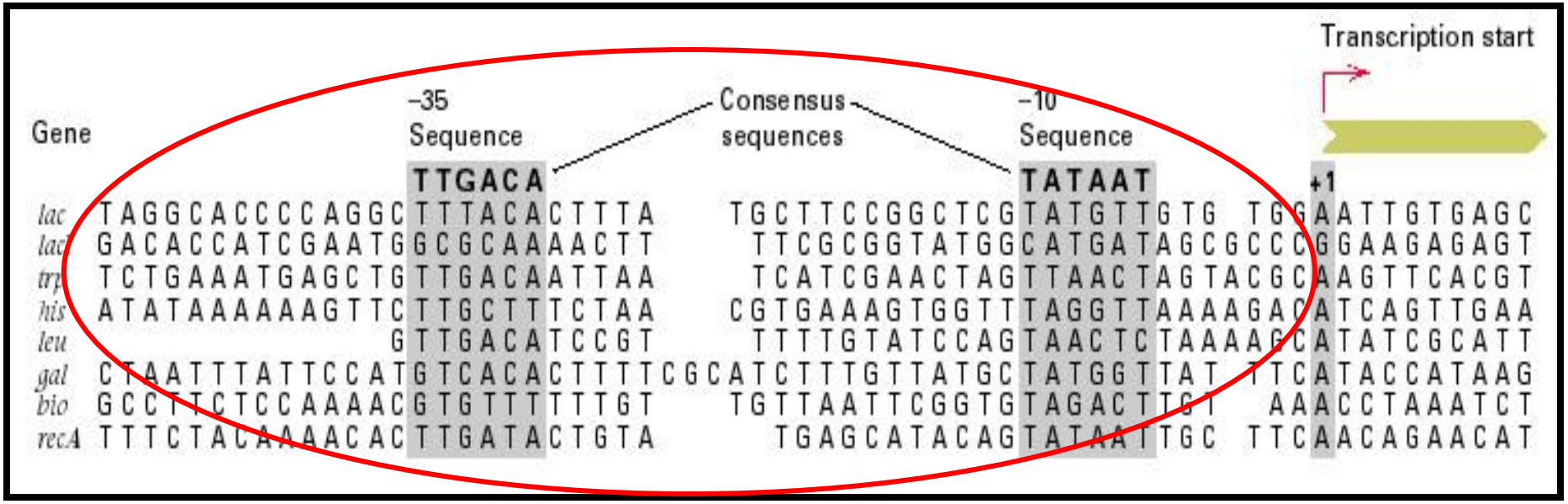


When a complete transcription complex is formed, RNA polymerase binds and transcription begins.

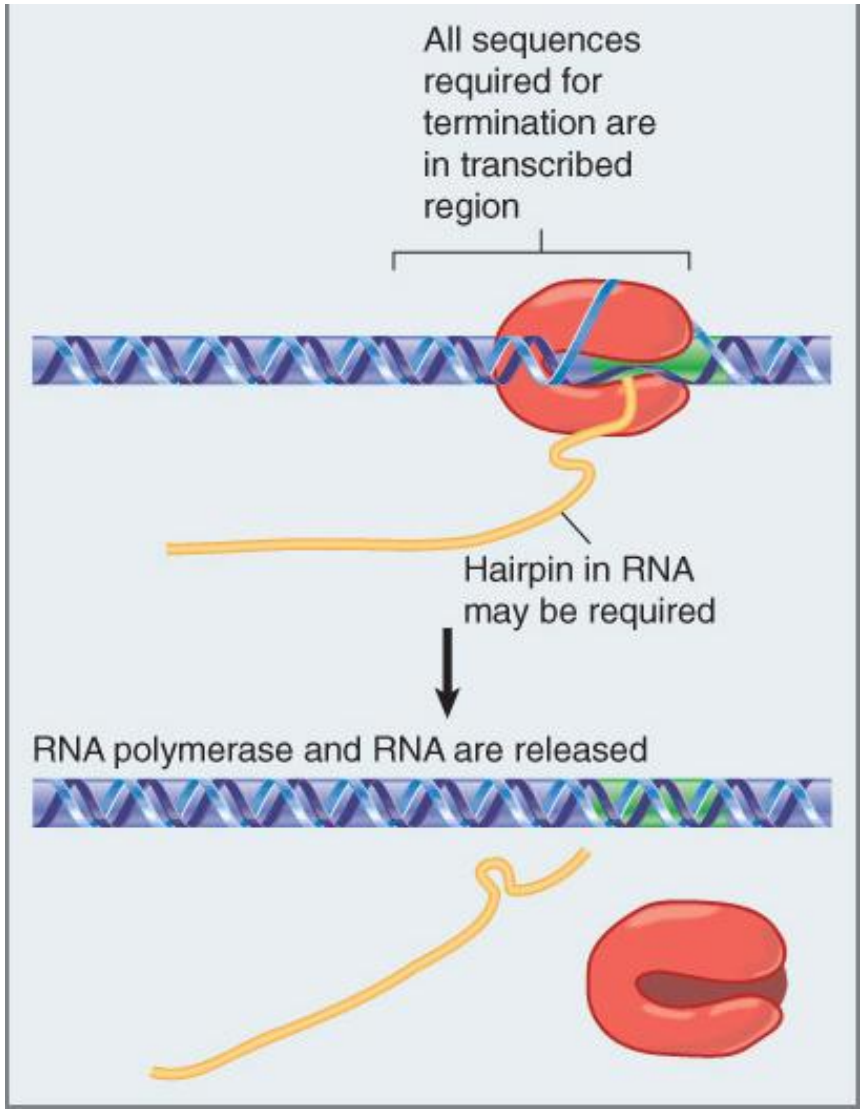
RNA Synthesis

- **Promoter = nucleotide sequence 5′ to the transcription start site which is the initial binding site of RNA polymerase and transcription initiation factors.**
- **Promoter recognition by RNA polymerase is a prerequisite for transcription initiation.**
- **A typical prokaryotic promoter has three components, consisting of consensus sequences at -35, -10 and the startpoint.**

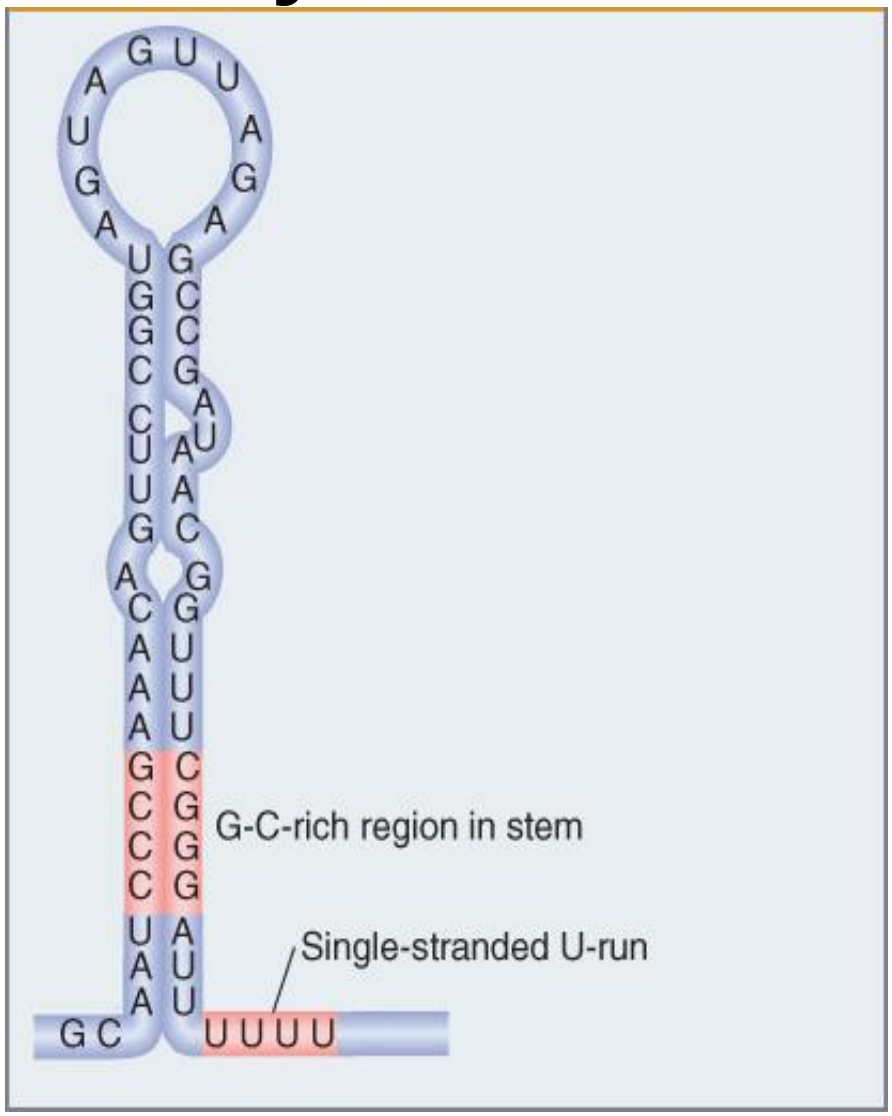




Termination of RNA Synthesis



The DNA sequences required for termination are located upstream of the terminator sequence.

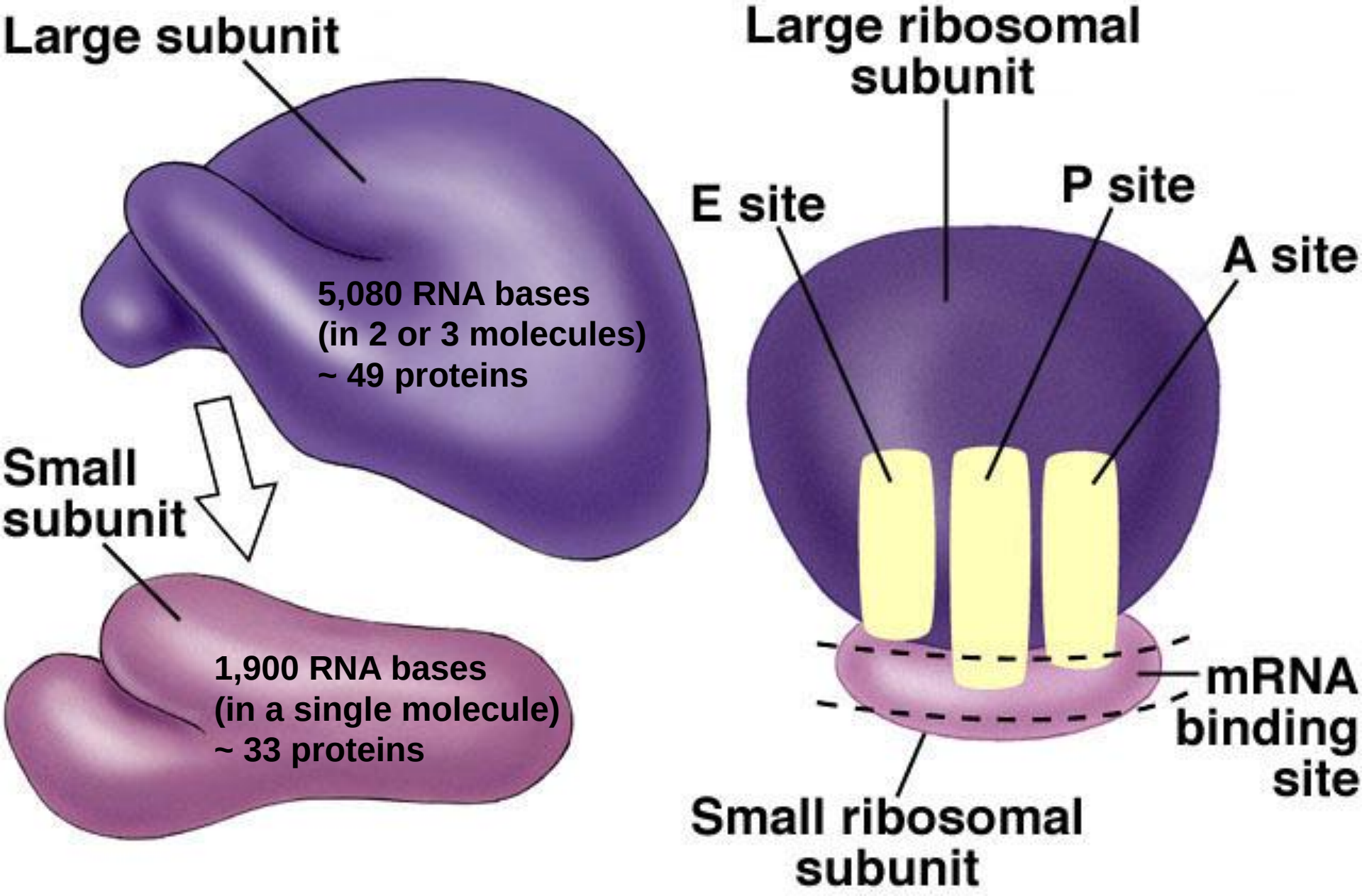


Formation of a hairpin may be necessary for termination.

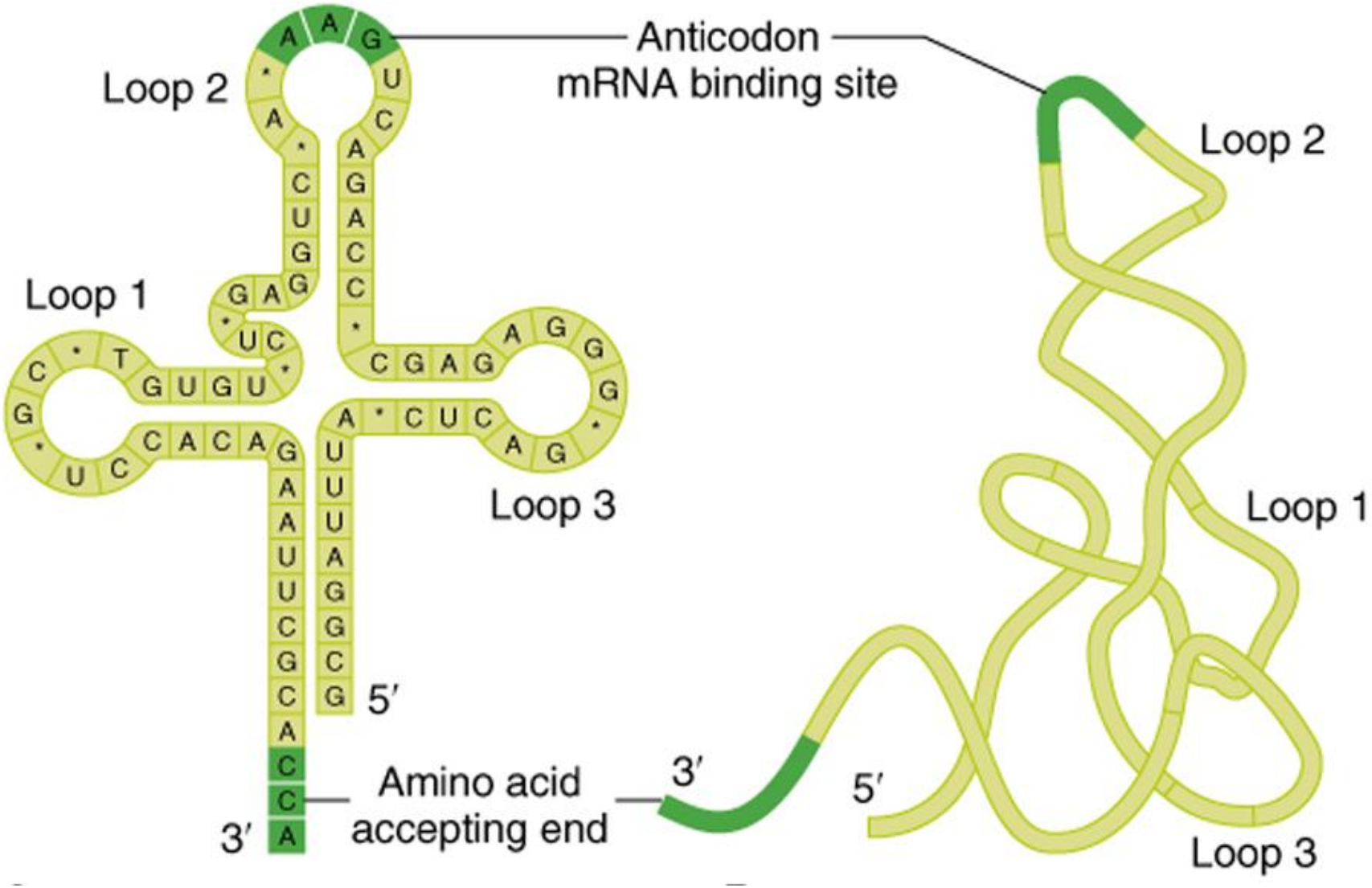
Types of RNA

RNA Type	Function
mRNA - Messenger RNA	encodes protein
rRNA - Ribosomal RNA	part of ribosome, used to translate mRNA into protein
tRNA - Transfer RNA	carries an amino acid residue and binds the mRNA codon

Ribosome Structure



tRNA is a connection between anticodon and amino acid.



Genetic Code

“The genetic code describes the way in which a sequence of twenty or more things is determined by a sequence of four things of a different type.”

Francis Crick - Nobel Lecture, December 11, 1962

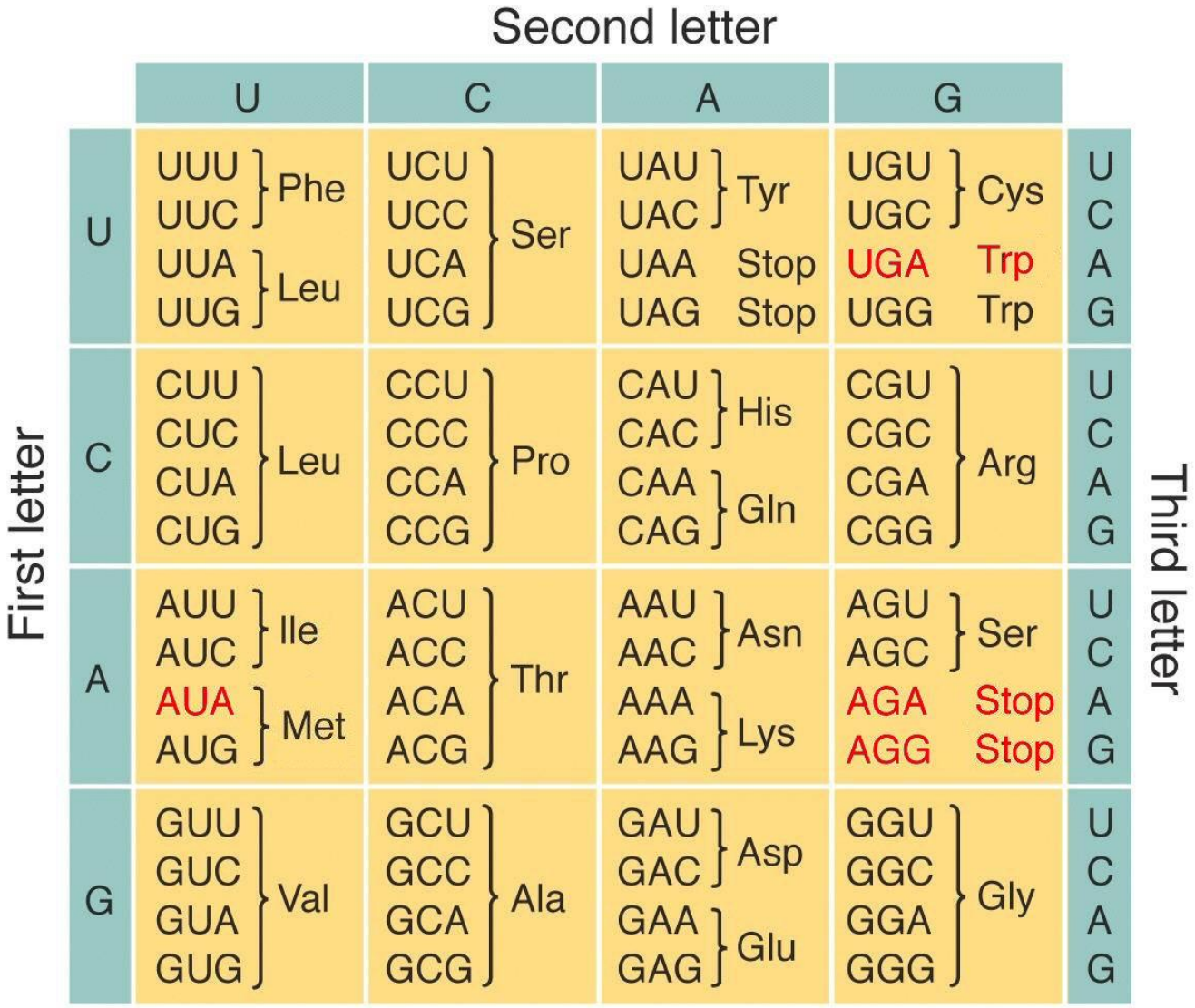
Genetic Code

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

Properties of the Genetic Code

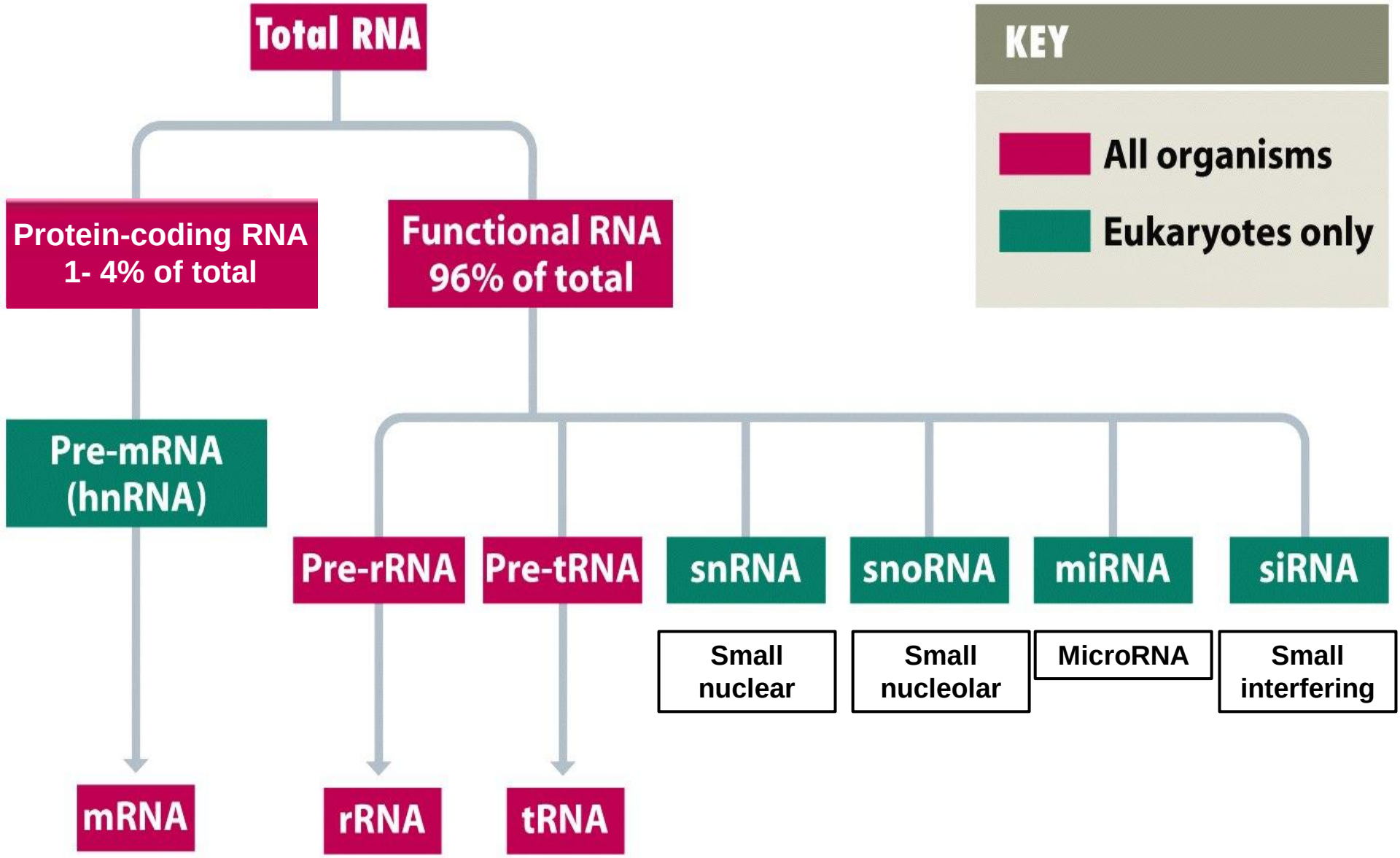
- **The genetic code is unambiguous: each codon specifies one amino acid only.**
- **The code is degenerate: one amino acid may be specified by more than one codon.**
- **Note that in most cases sufficient coding is performed by the first two bases, the third (or wobble) base playing a minor role.**
- **For instance the four codons that specify glycine (GGU, GGC, GGA and GGG) all start with GG.**
- **Codons with a similar sequence specify amino acids with similar chemical properties.**
- **The codons that specify threonine differ from those specifying serine by their 5′ nucleotide.**
- **The codons for aspartate and glutamate differ only by their 3′ position.**
- **Codons within the middle a pyrimidine generally specify for a hydrophobic amino acid.**
- **Thus, mutation of the 5′- or 3′-positions of these codons lead to a substitution of chemically similar amino acids.**
- **Note also the STOP codons, which cause termination of translation by the ribosome.**
- **The codon AUG for methionine is also used as start codon.**

The Vertebrate Mitochondrial DNA (mtDNA) Genetic Code

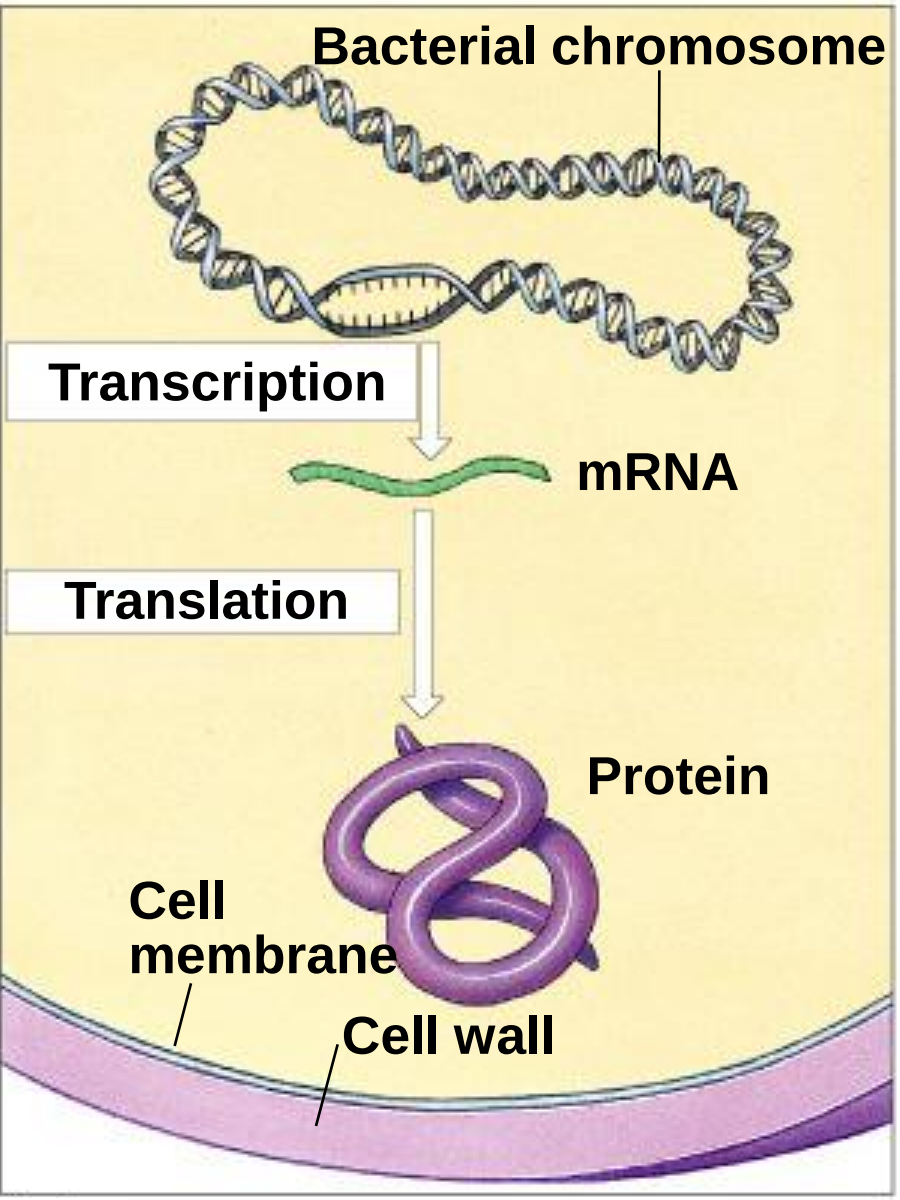


- **Differences between the vertebrate mtDNA code and the "universal" code are indicated in red.**
- **Note that UGA codes for Trp rather than being a stop codon.**
- **There are two Met codons and two AG*R* codons are read as Stops.**
- **Slightly different mtDNA codes are found in *Drosophila* and other invertebrate groups.**

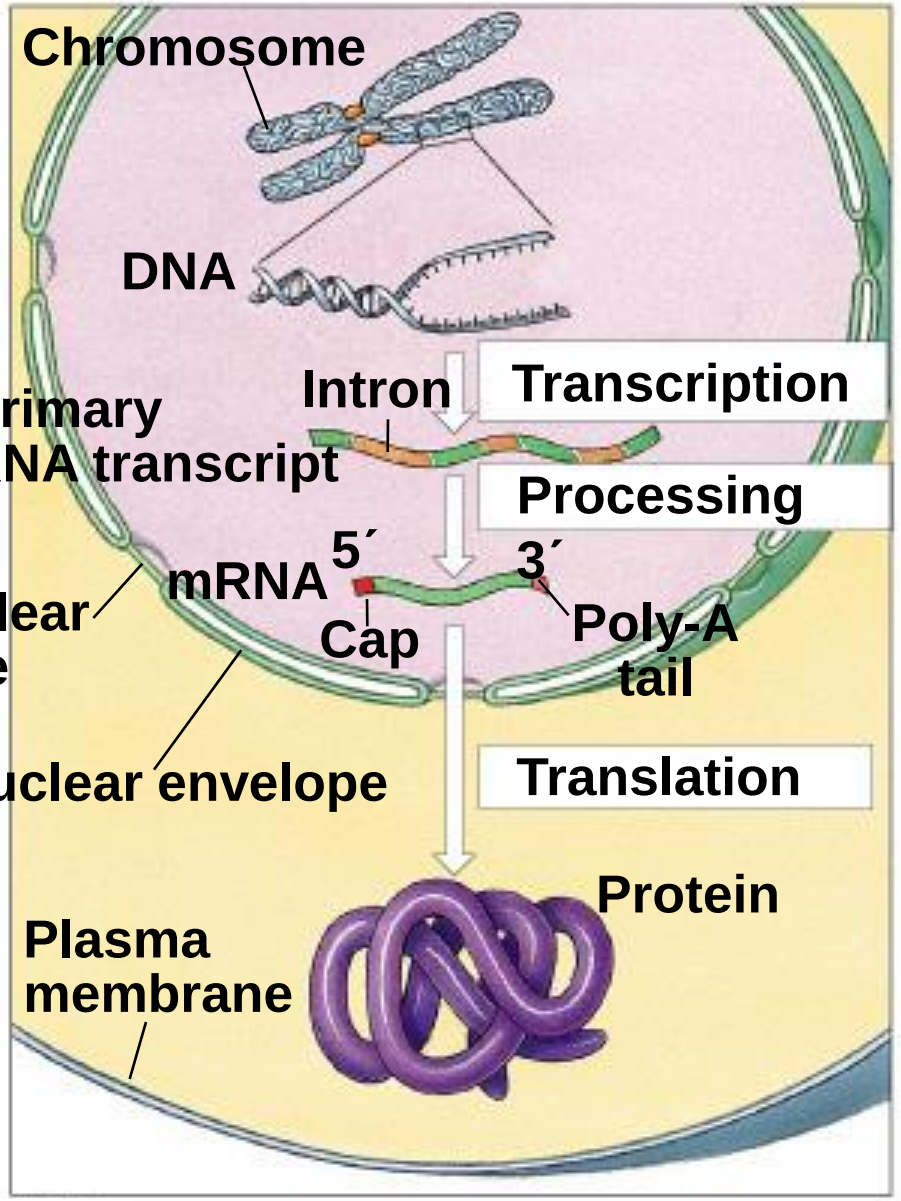
RNA Content of a Cell



Gene Expression

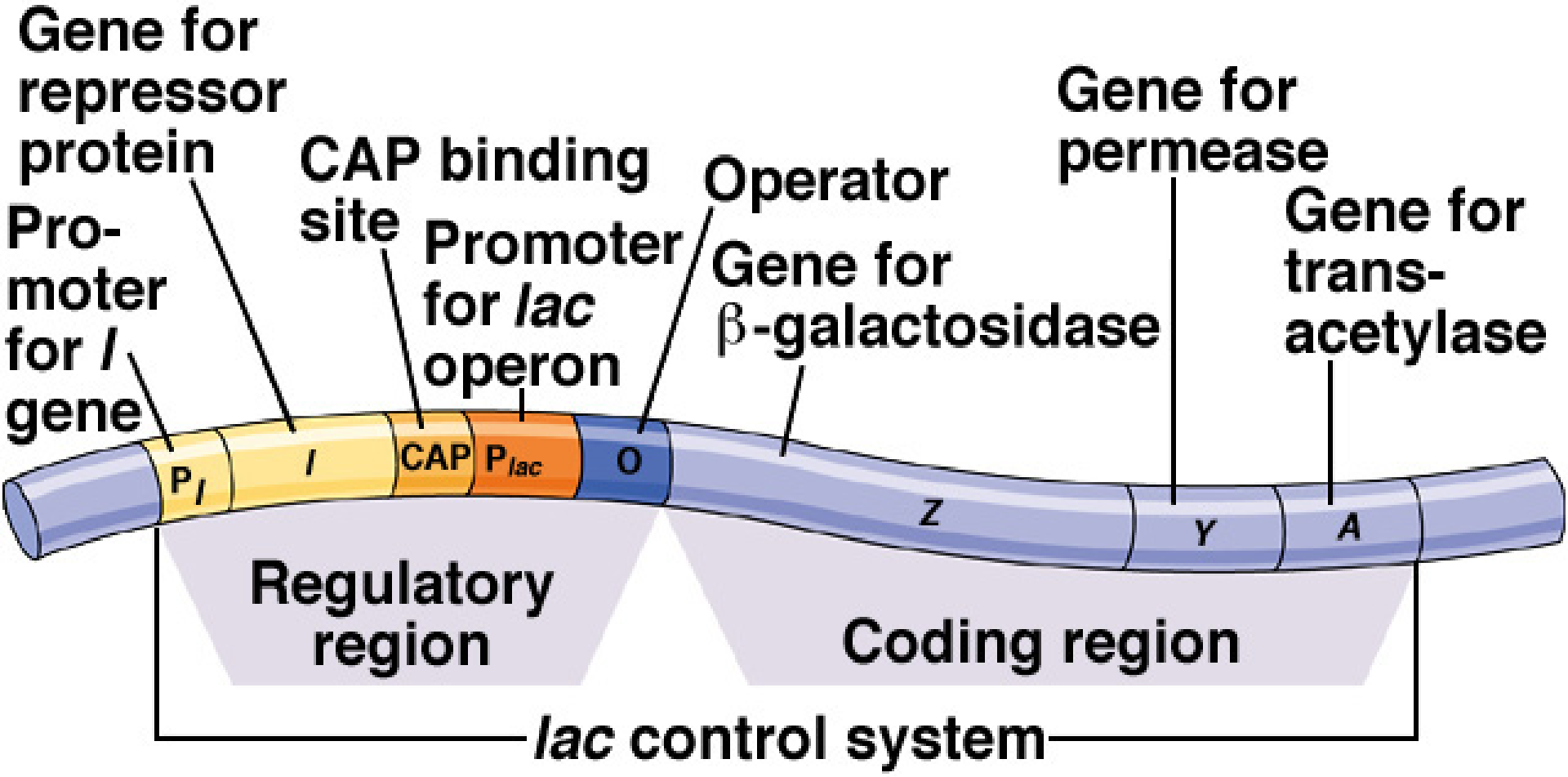


Prokaryotic Cell

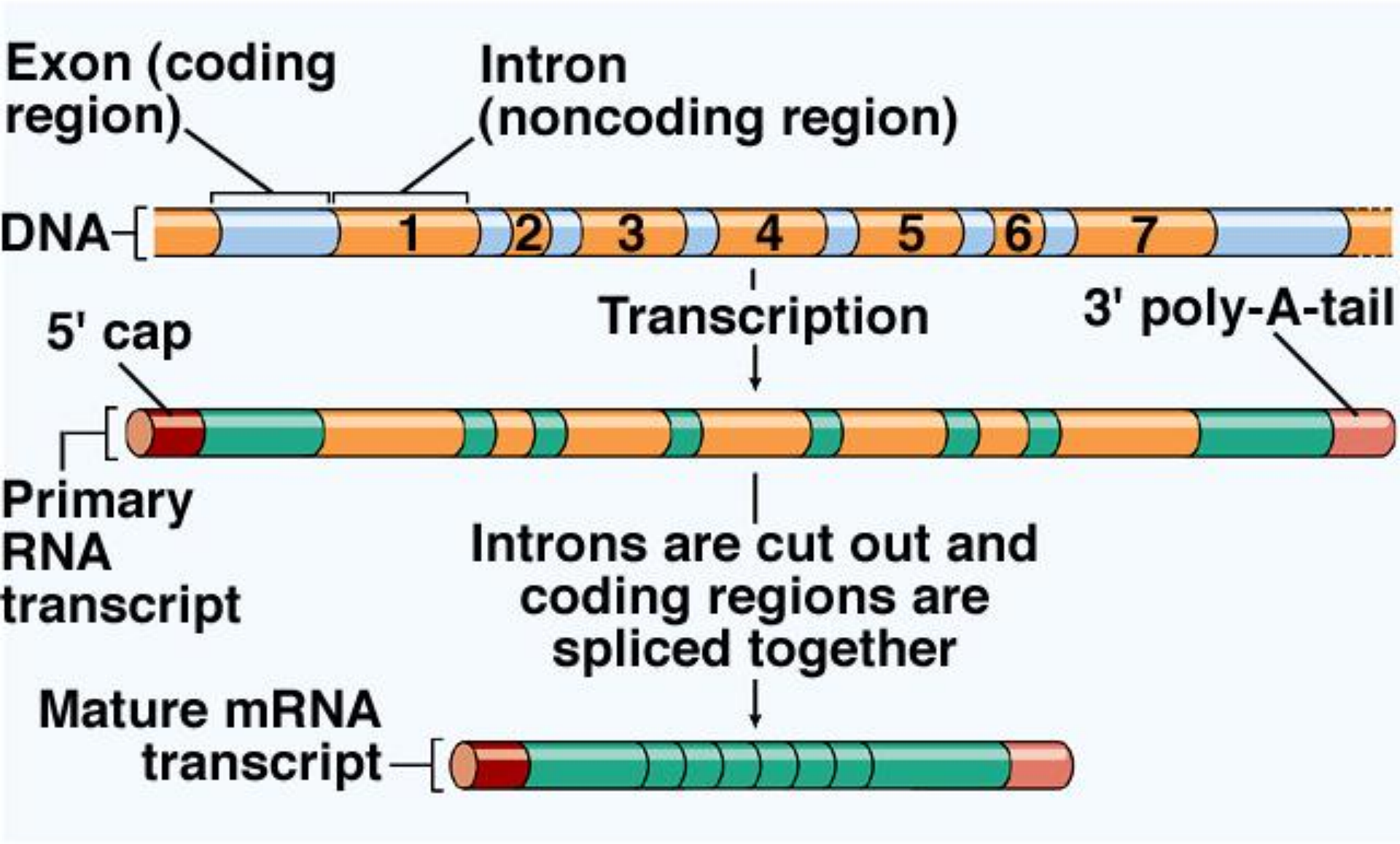


Eukaryotic Cell

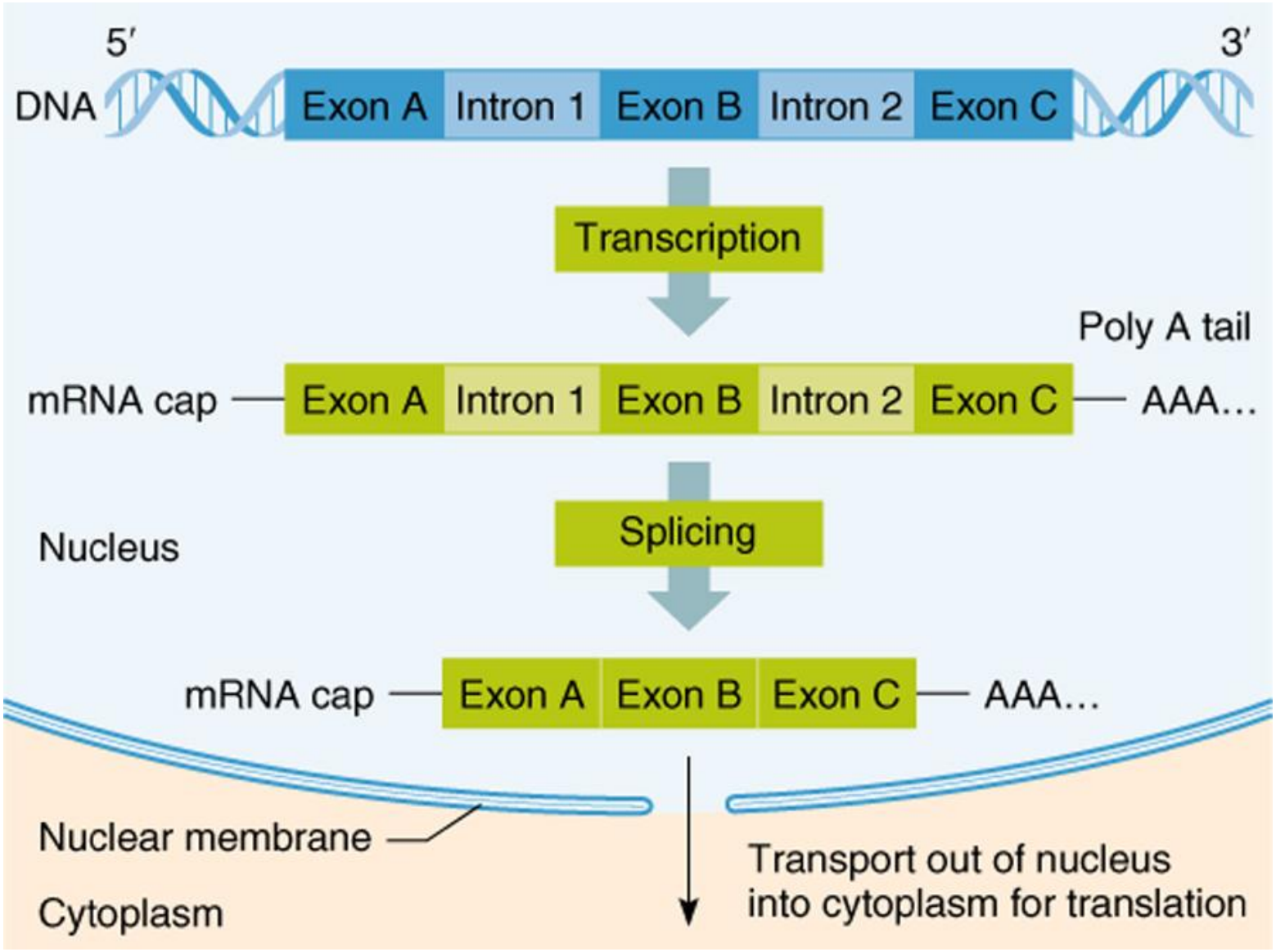
lac Operon of *E. coli*



Gene Structure in Eukaryotes (Intron and Exon)

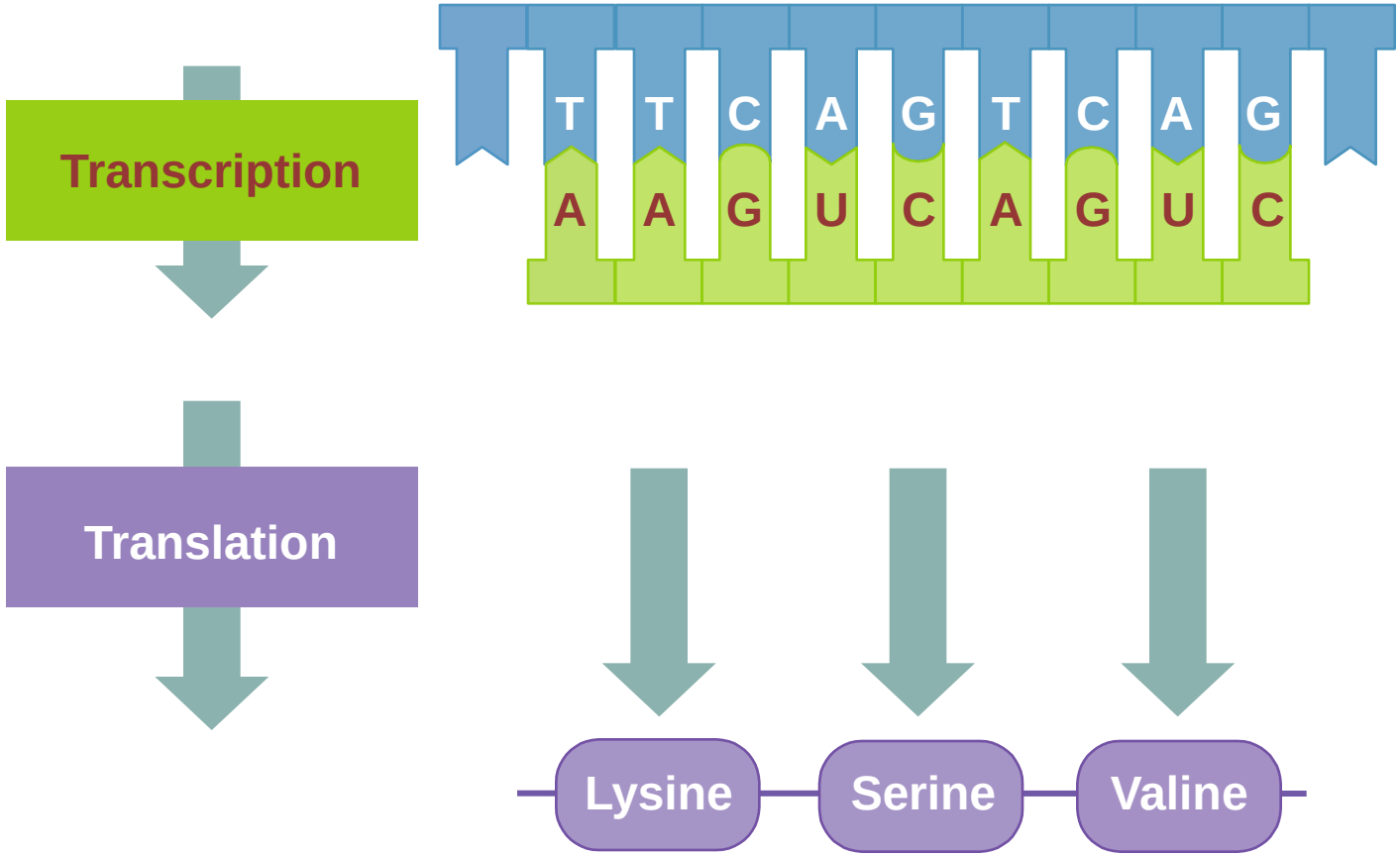


Eukaryotic RNA Processing



Translation

The process of reading the RNA sequence of an mRNA and creating the amino acid sequence of a protein is called translation.



-omes

- **genome – the complete set of sequences in the genetic material of an organism**
  - **It includes the sequence of each chromosome plus any DNA in organelles.**
- **transcriptome – the complete set of RNAs present in a cell, tissue, or organism**
  - **Its complexity is due mostly to mRNAs, but it also includes noncoding RNAs.**
- **proteome – the complete set of proteins that is expressed by the entire genome**
  - **The term is sometimes used to describe the complement of proteins expressed by a cell at any one time.**
- **interactome – the complete set of protein complexes/protein–protein interactions present in a cell, tissue or organism**

Genome

- A genome is a sequence of nucleotides arrayed across a linear scale of a start position to an end position.
- A genome sequence is a reference.
- As a reference, it cannot account for all the variability that exists in a species.
- The human genome sequence is haploid, which means that even if it were compiled from a single donor, the single reference sequence does not report the variation at millions of nucleotide positions between two donor's two copies (except for X and Y).
- Once the human genome is re-sequenced, to be reported as a diploid sequence, it will be done in a way that produces phased sequence, in which each chromosome is reported separately, rather than just identifying the two alleles at each variable site along the genome without specifying on which chromosome it lies.
- This format will represent sequences as they actually exist in a sequenced person, identifying which alleles go together on a chromosome, and are thus linked evolutionarily.