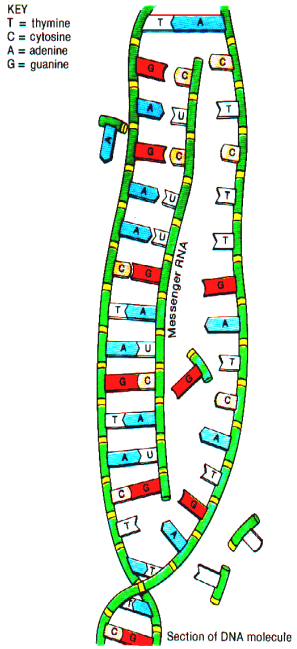
Bioinformatics

ONE

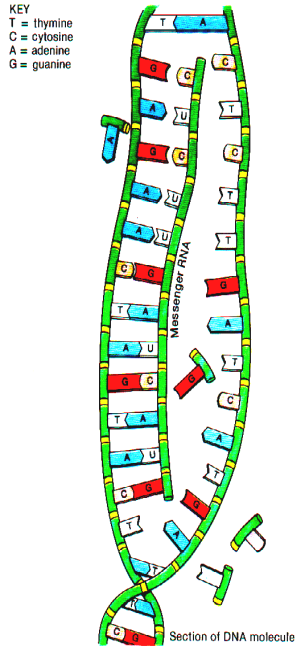
Introduction to Biology



Wendy Lee
Dept of Computer Science
San José State University

Biology/CS/SE 123A

Fall 2014



bi·o·in·for·mat·ics

/ˌbīōˌinfərˈmatiks/

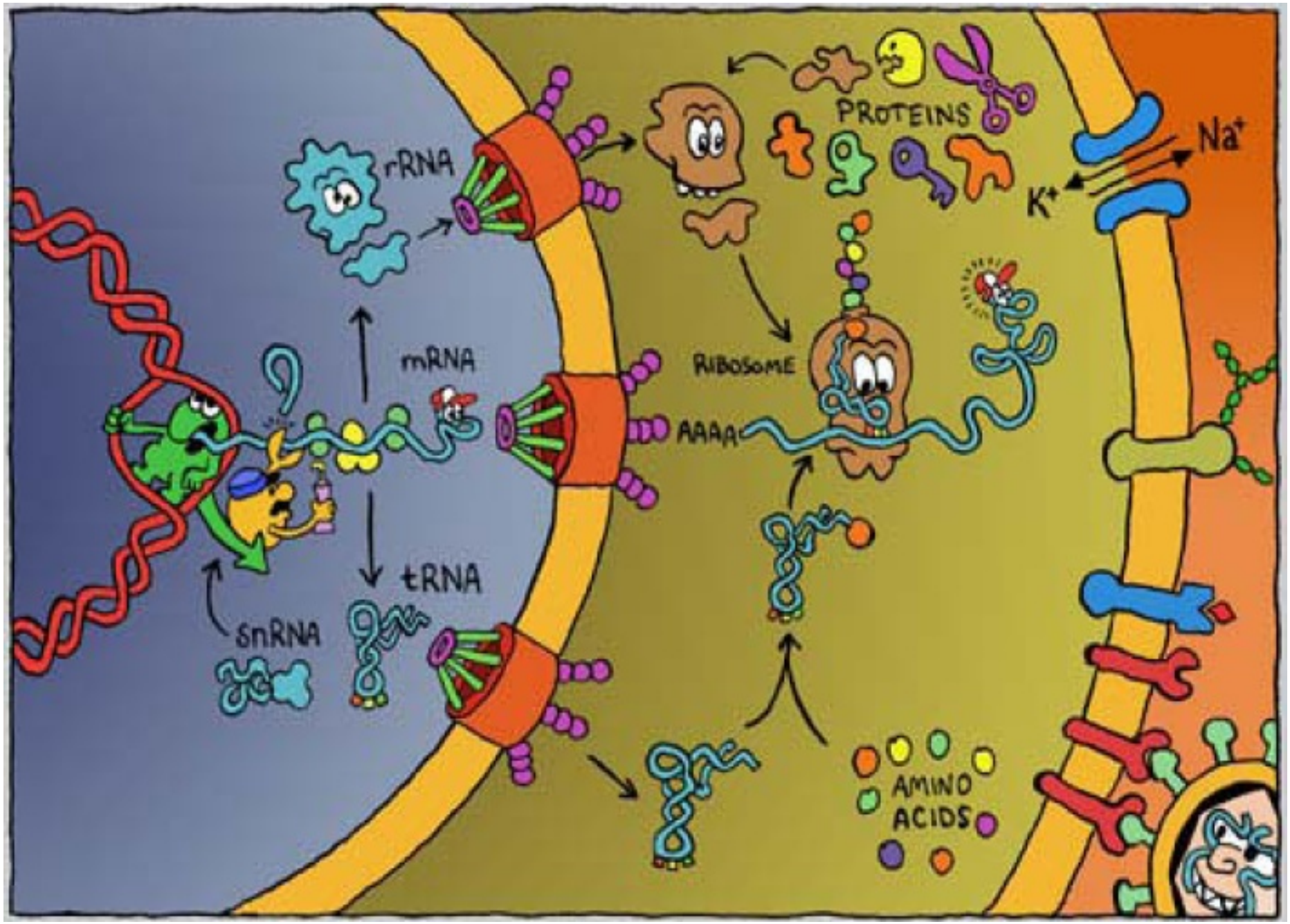
noun

noun: **bioinformatics**; noun: **bio-informatics**

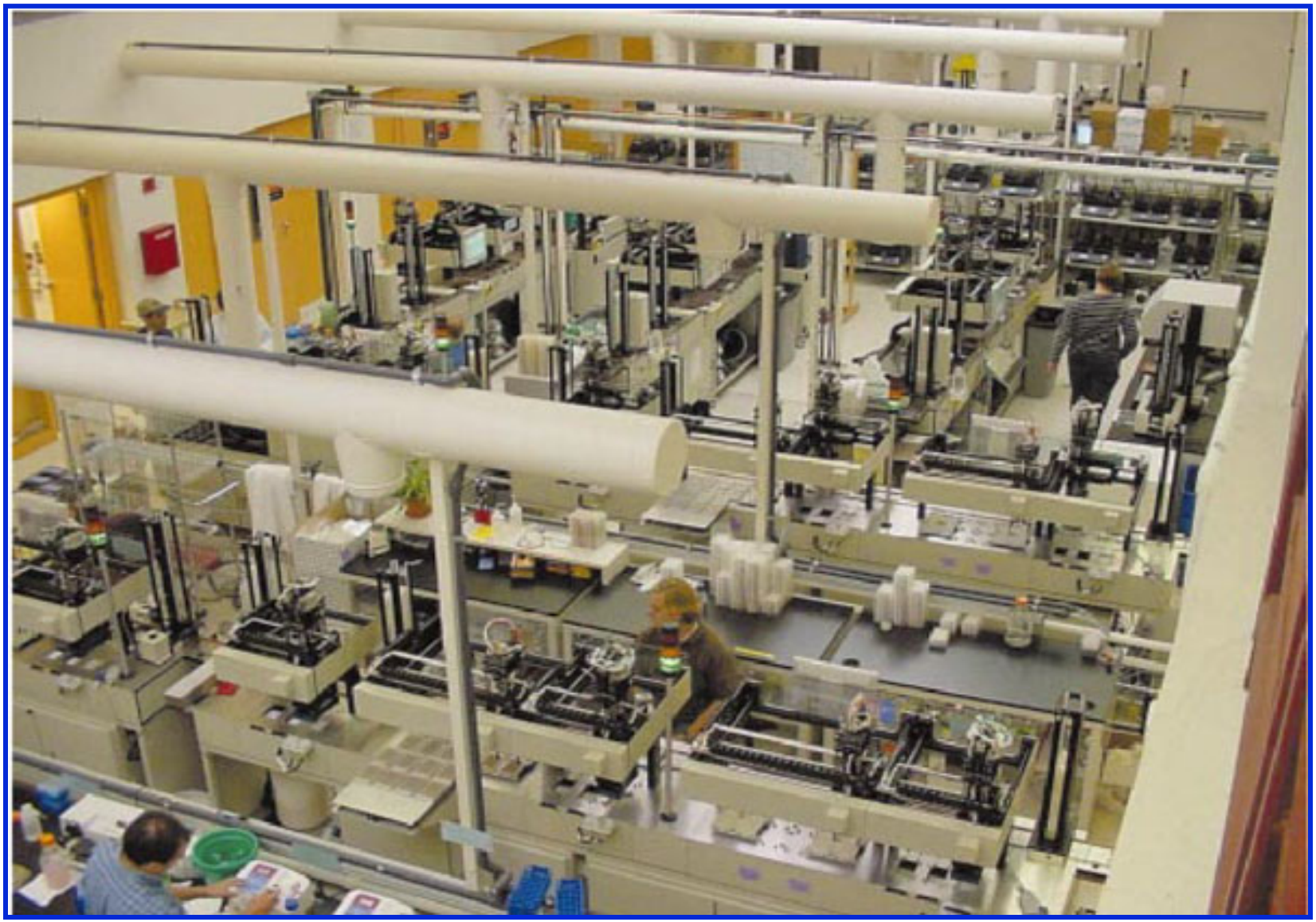
the science of collecting and analyzing complex biological data such as genetic codes.

Translate bioinformatics to Choose language

Use over time for: bioinformatics



The New Face of Biology



Sequencing SARS

http://www.bcgsc.ca/bioinfo/SARS

Some Biological Questions (I)

- Why don't all tumors of the same type respond to the same chemotherapies?
  - Could we individualize treatment to increase the chances of successful therapy?
- What are the causes of the hundreds of genetic diseases?
  - How do genes influence complex diseases such as diabetes, obesity, heart disease, cancer, and alcoholism?
  - What can we do about these diseases?

Some Biological Questions (II)

- How should we define a species?
  - How do we determine whether two different organisms belong to the same species?
- In the human genome, how does a surprisingly small set of genes (probably fewer than 25,000) result in the synthesis of hundreds of thousands of proteins?
  - What functions remain to be discovered for the 99% of the genome that apparently does not encode proteins?

Some Biological Questions (III)

- How do bacteria and viruses cause disease?
  - How can a better understanding of their physiology improve prevention and treatment?
- Can DNA evidence be relied upon in convicting (or exonerating) those accused of crimes?
- What is the risk that a child will inherit a genetic disease or the susceptibility for the disease?

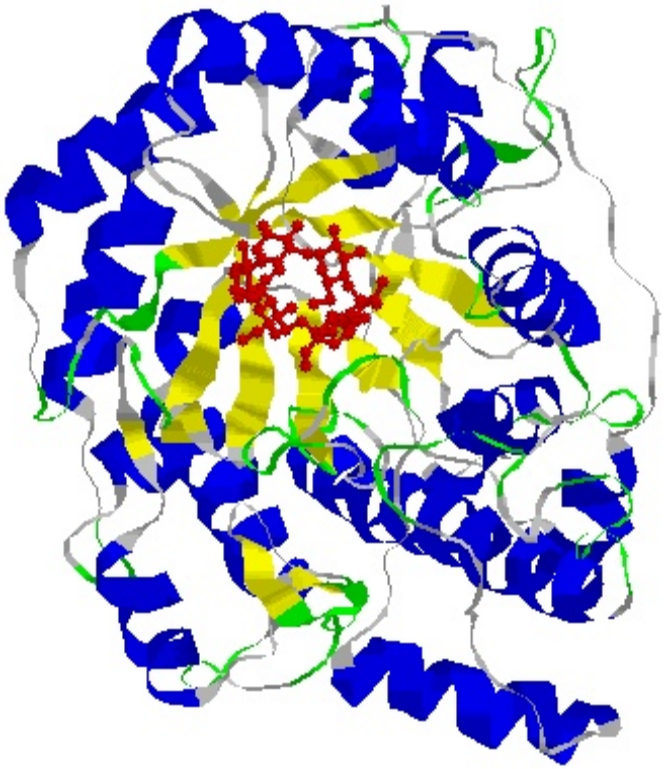
Some Biological Questions (IV)

- What are the evolutionary pathways that have led to the development of the diversity of organisms that we see today?
- How can drug design be improved so that specific drugs can be made to work on specific targets?
- What can the genes of bacteria, yeast, flies, worms, plants, and mice tell us about the functions of human beings?

Biggest Challenge for Contemporary Cell Biology

- Most often, cell biologists studying the cell's control systems sum up their knowledge in simple schematic diagrams rather than in numbers, graphs, and differential equations.
- To progress from qualitative descriptions and intuitive reasoning to quantitative descriptions and mathematical deduction is one of the biggest challenges for contemporary cell biology.

Biology Review

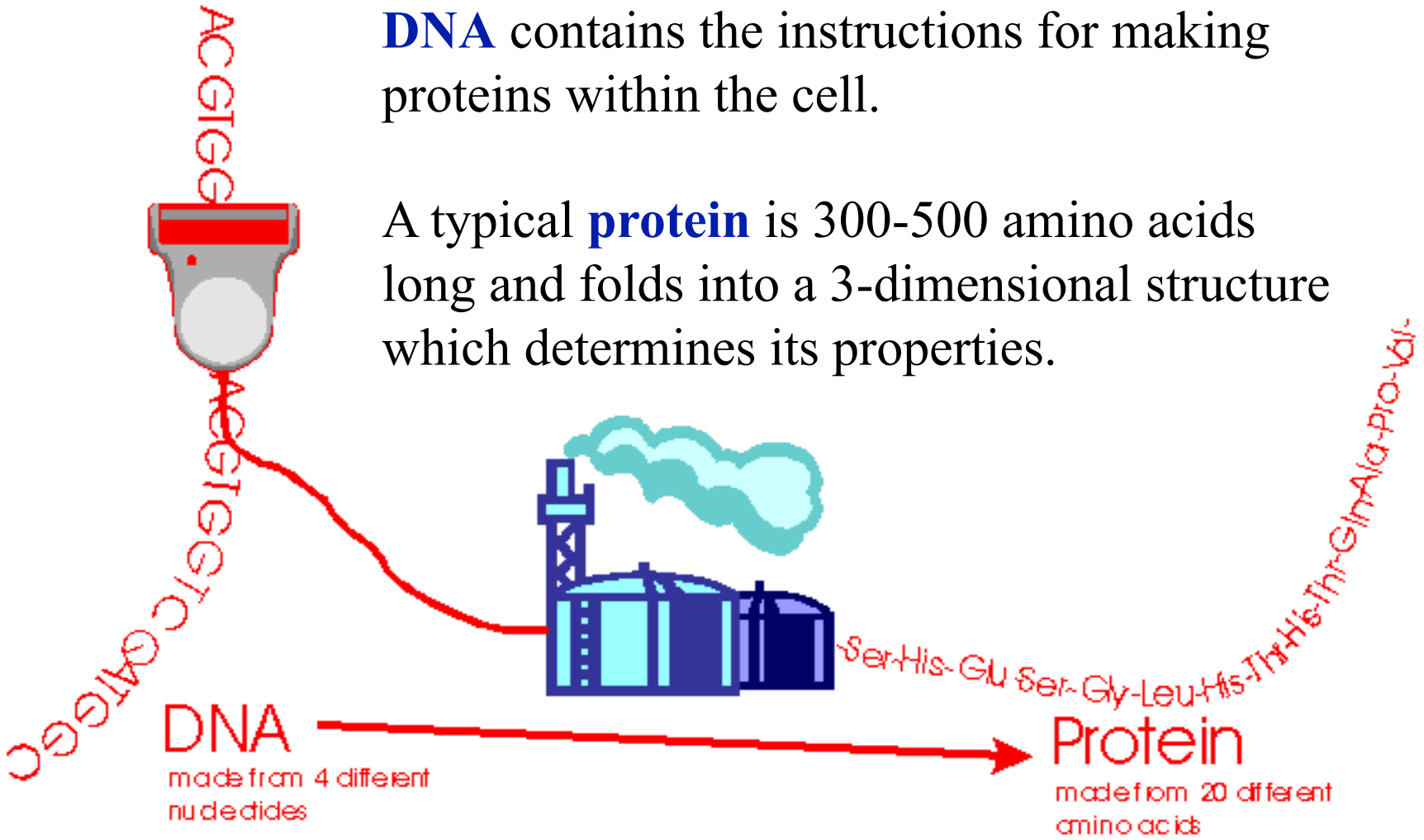


- DNA
- RNA
- Proteins
- Central Dogma
- Transcription
- Translation

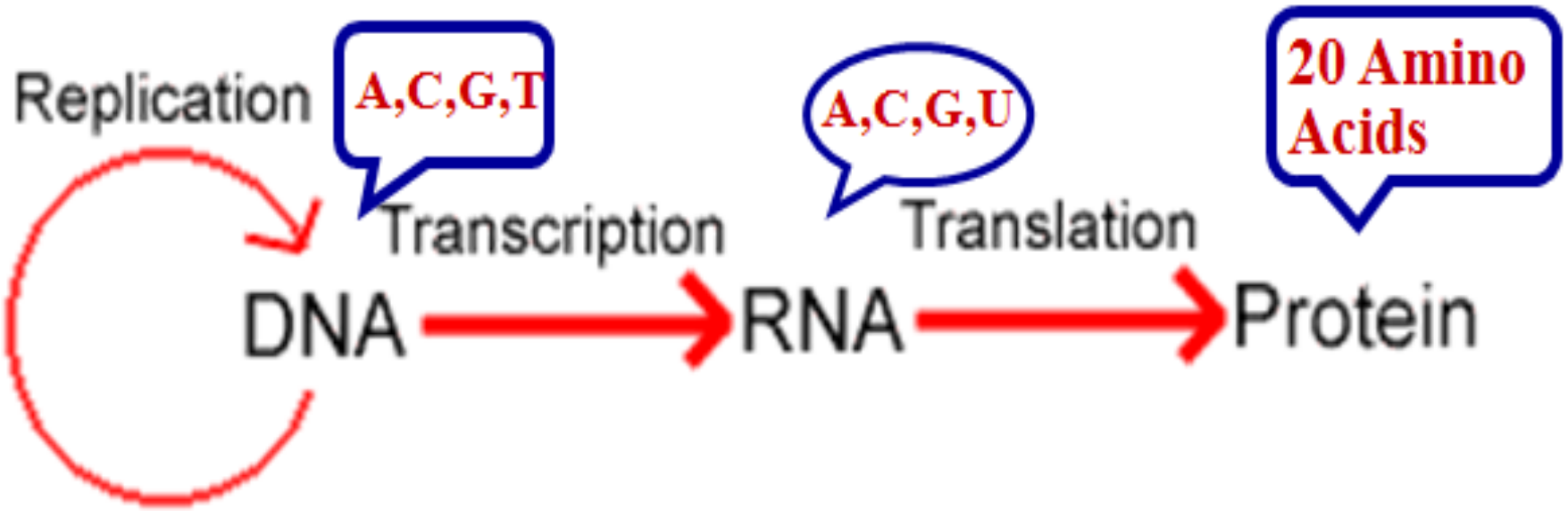
Protein Factory

DNA contains the instructions for making proteins within the cell.

A typical **protein** is 300-500 amino acids long and folds into a 3-dimensional structure which determines its properties.

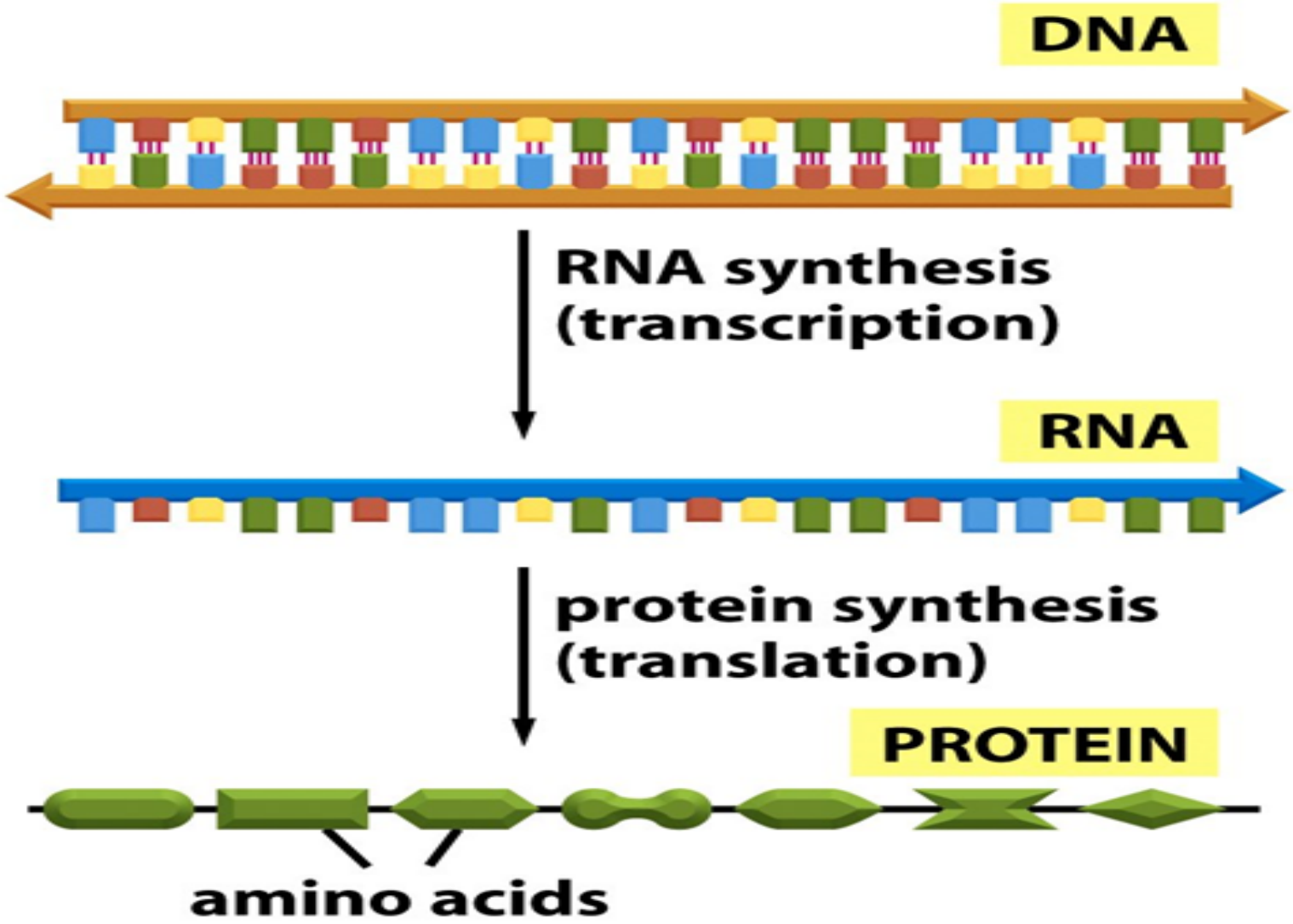


DNA → RNA → Protein



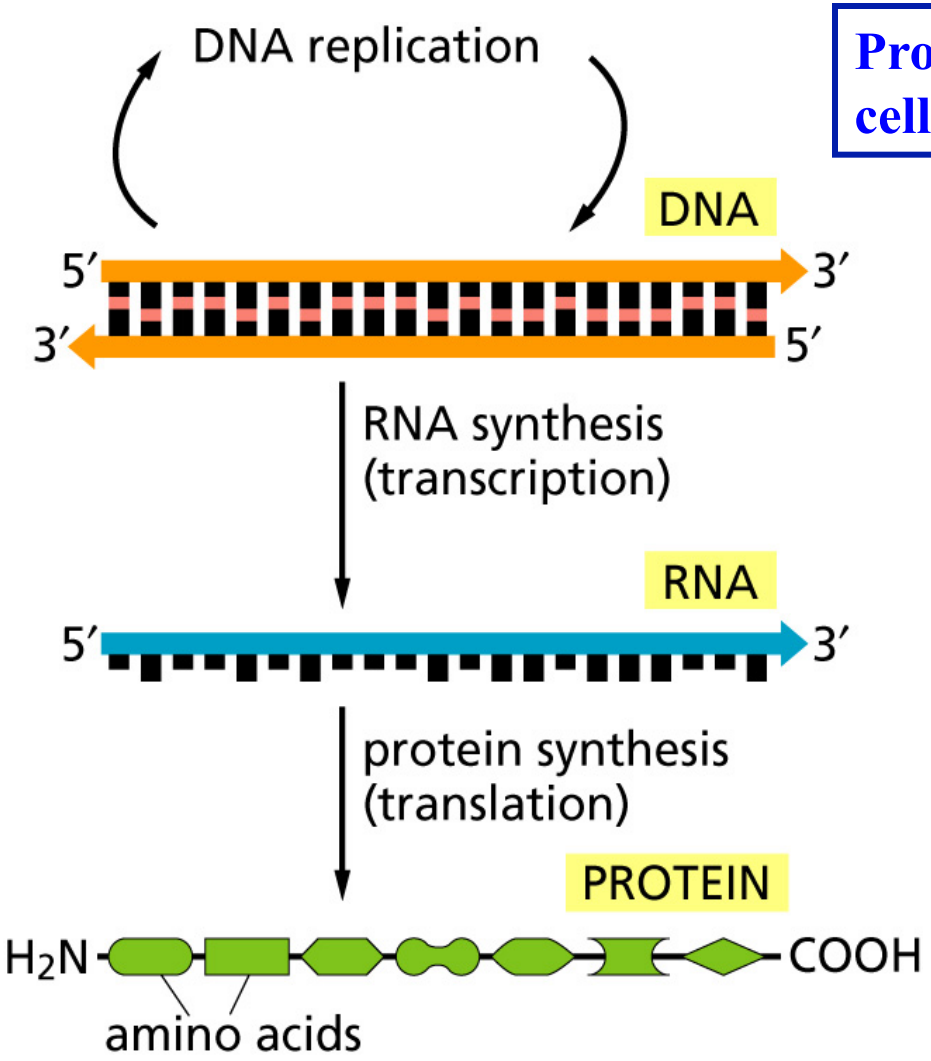
Adenine	(A)	Adenine	(A)	{A-Y}-{BJOUX}
Guanine	(G)	Guanine	(G)	
Cytosine	(C)	Cytosine	(C)	
Thymine	(T)	Uracil	(U)	

From DNA to Protein



Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Central Dogma of Molecular Biology



Proteins, are the molecules that put the cell's genetic information into action.

- **Traits**
- **Diseases**
- **Drug Resistance**
- **Physiology**
- **Metabolism**

"We *are* our Proteins" Doolittle

Limitless Diversity From Combinatorial Assemblies of Limited Building Blocks

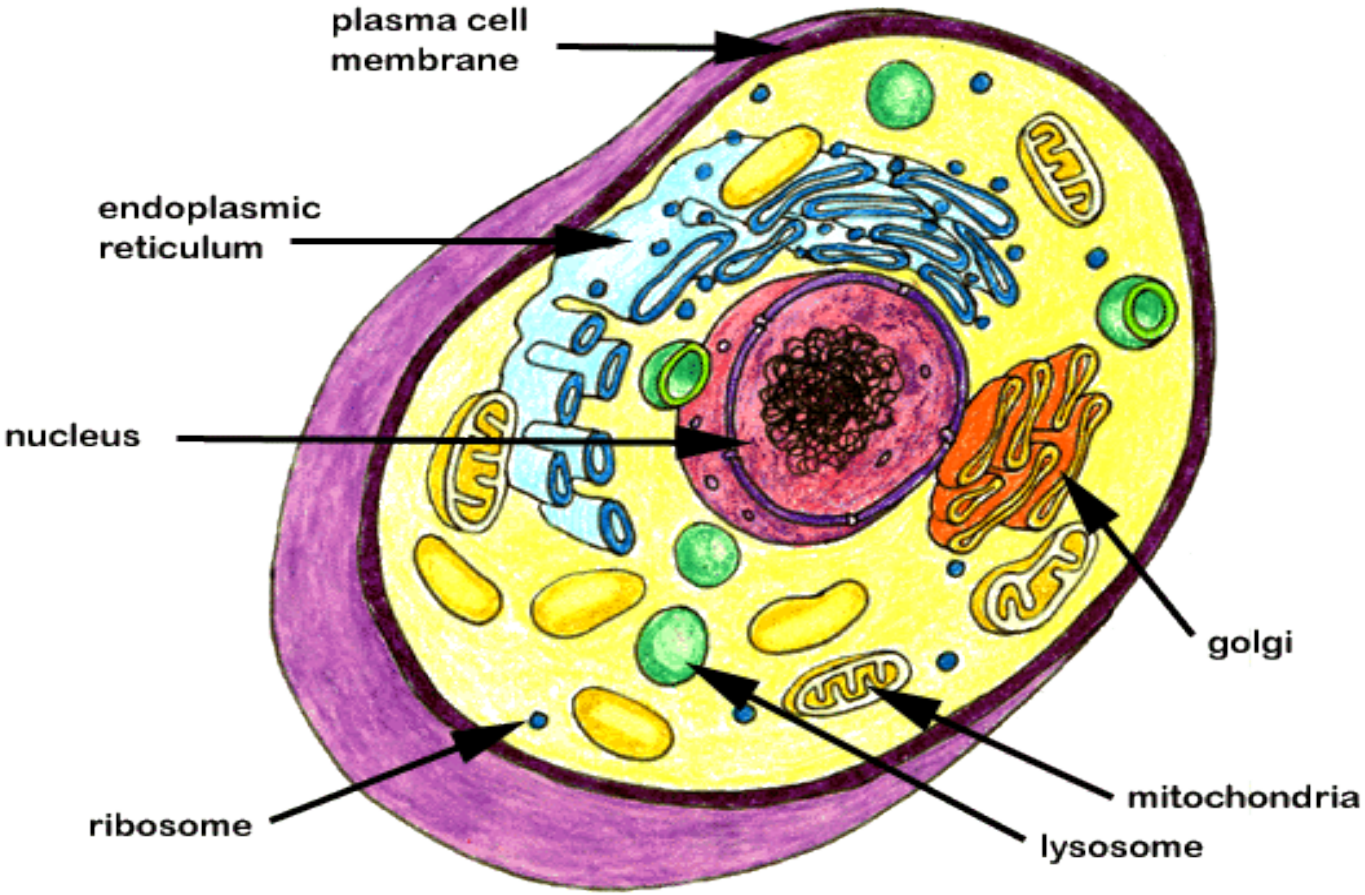
Prokaryotes and Eukaryotes

A **cell** is the fundamental working unit of every living organism.

There are two kinds of cells:

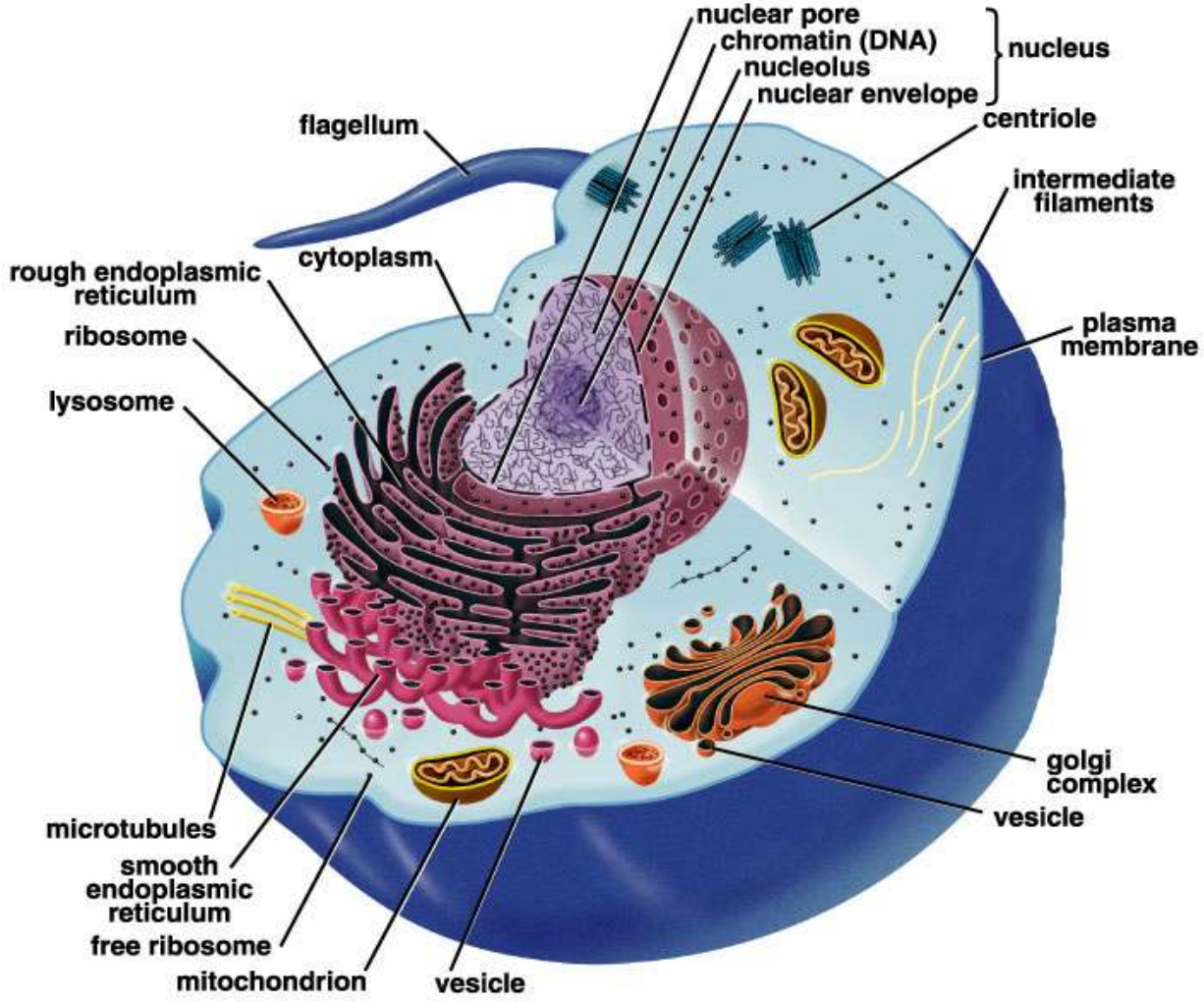
- **prokaryotes**, which are single-celled organisms with no cell nucleus: archea and bacteria.
- **eukaryotes**, which are higher level organisms, and their cells have nuclei: animals and plants.

Eukaryotic Cell (I)



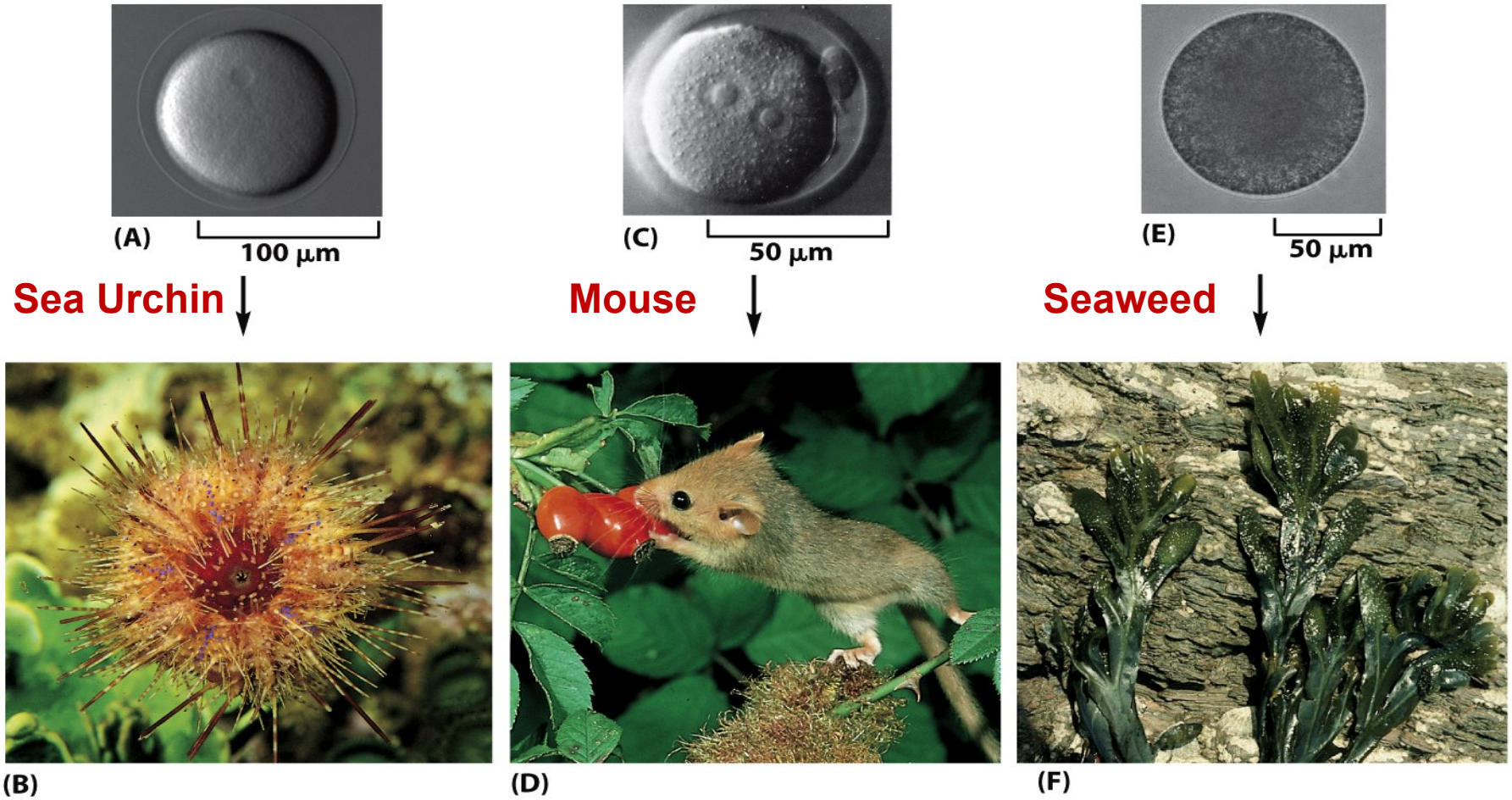
local.brookings.k12.sd.us/biology/CellStructure.htm

Eukaryotic Cell (II)



A **cell** carries the entire set of genetic instructions:
the genome, that makes an entire organism.
The **instructions** are encoded in DNA as genes and packaged as chromosomes in the nucleus.

Hereditary Information



The hereditary information in the fertilized egg cell determines the nature of the whole multicellular organism.

Proteins and Nucleic Acids

All living organisms have a similar molecular chemistry. The main actors in the chemistry of life are molecules:

- **proteins**: which are responsible for what a living being is and does in a physical sense.

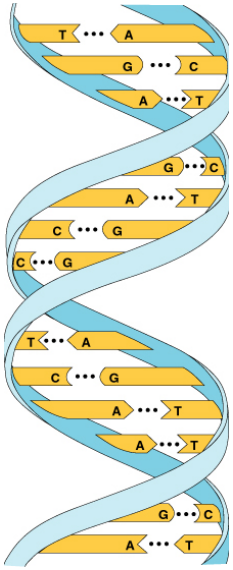
  "We are our proteins" R. Doolittle.

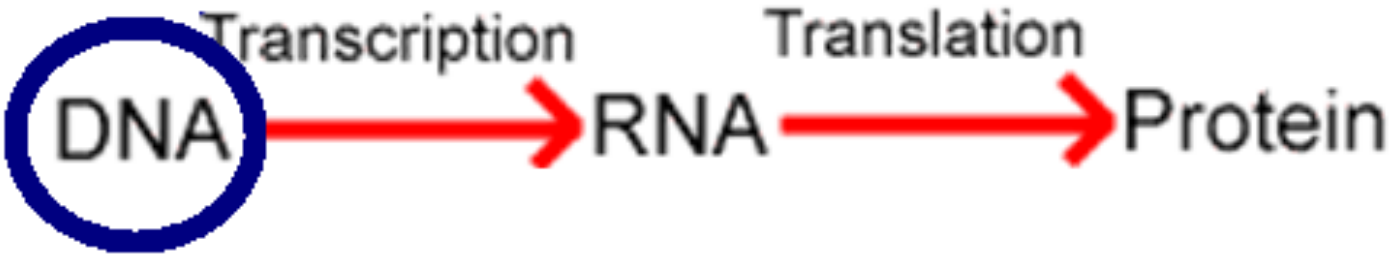
- **nucleic acids**: which encode the information necessary to produce proteins and are responsible for passing the "recipe" to subsequent generations.

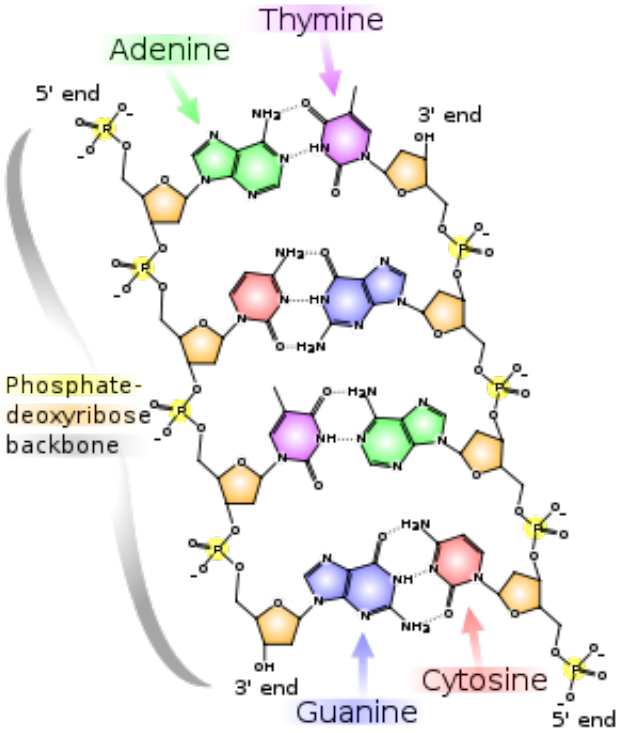
Living organisms contain 2 kinds of nucleic acids:

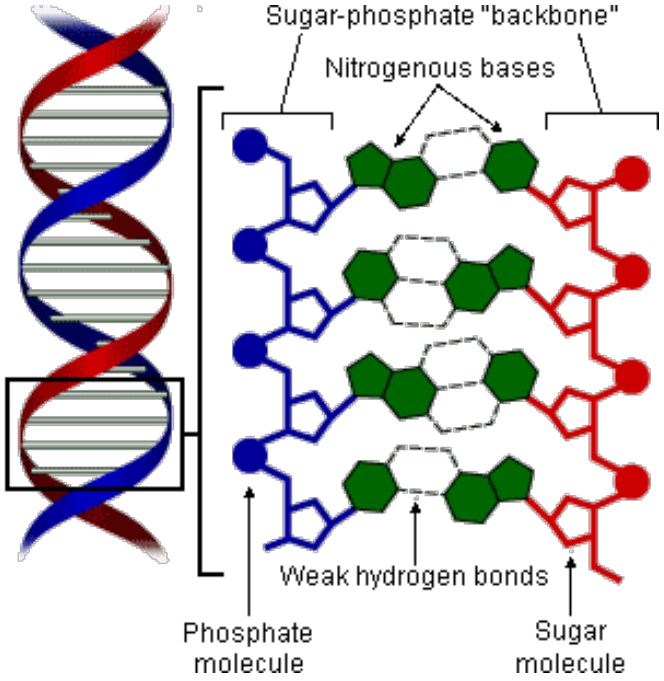
- **Ribonucleic acid** (**RNA**)
- **Deoxyribonucleic acid** (**DNA**)

DNA

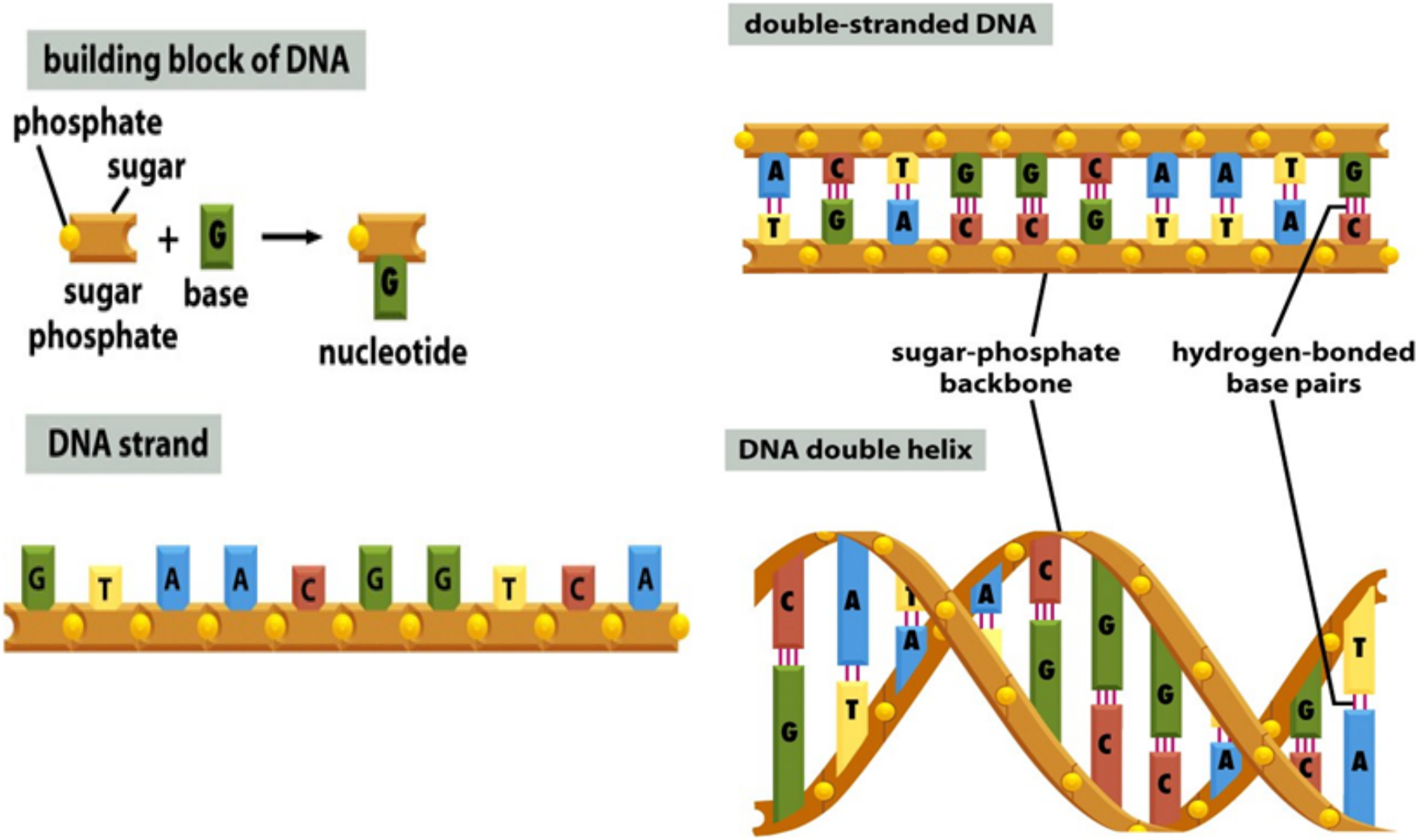






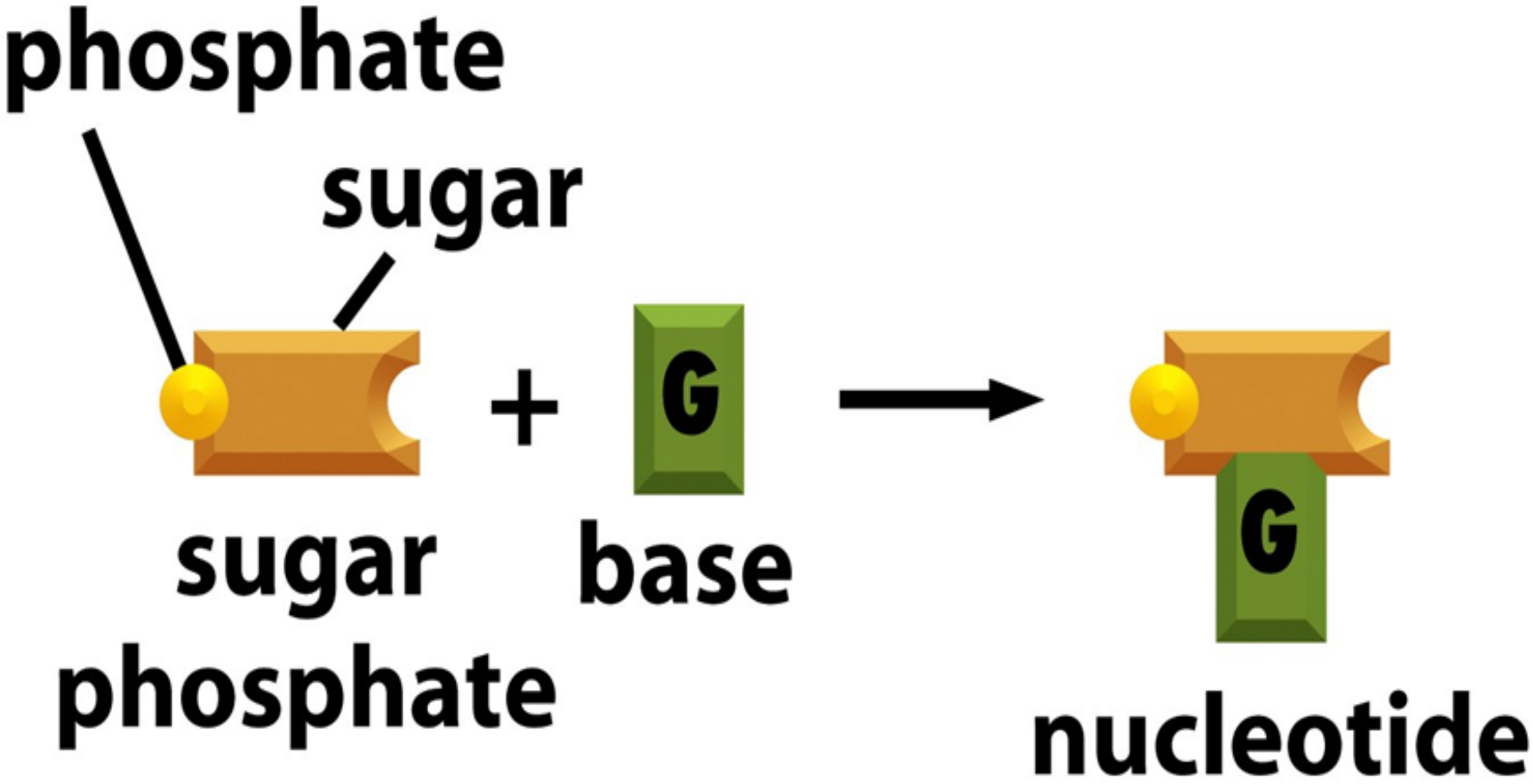


DNA Double Helix



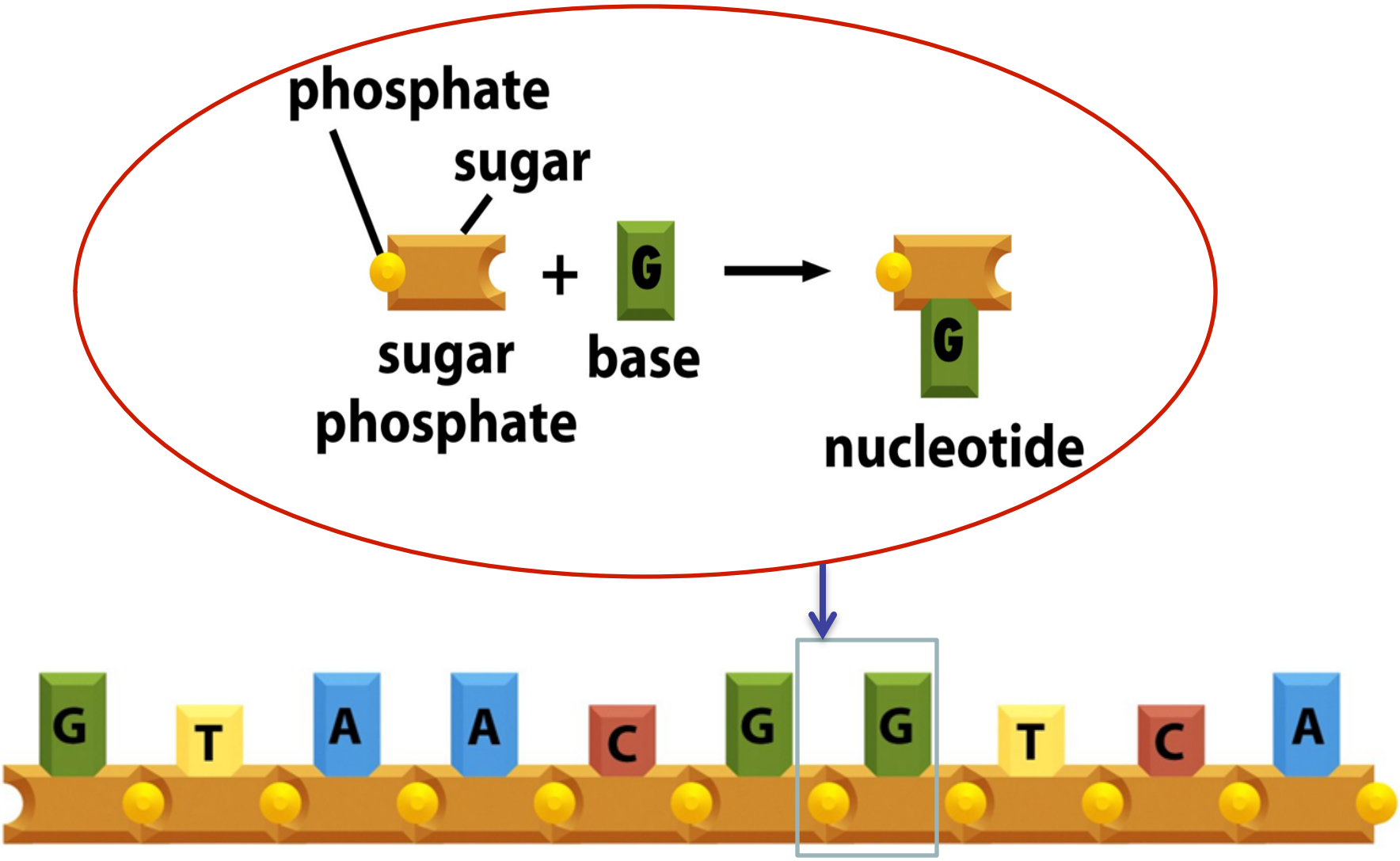
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Building Block of DNA



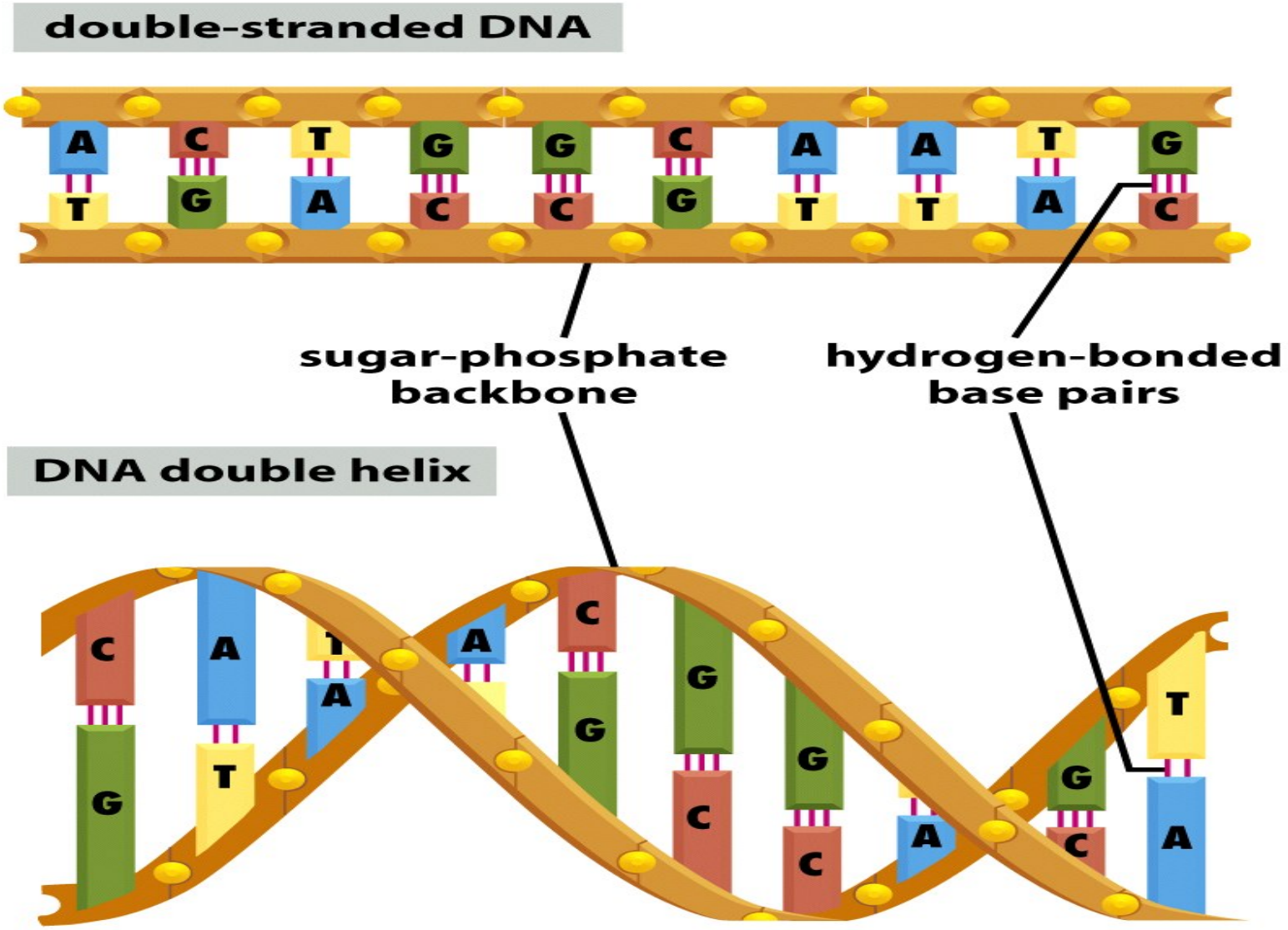
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

DNA Strand



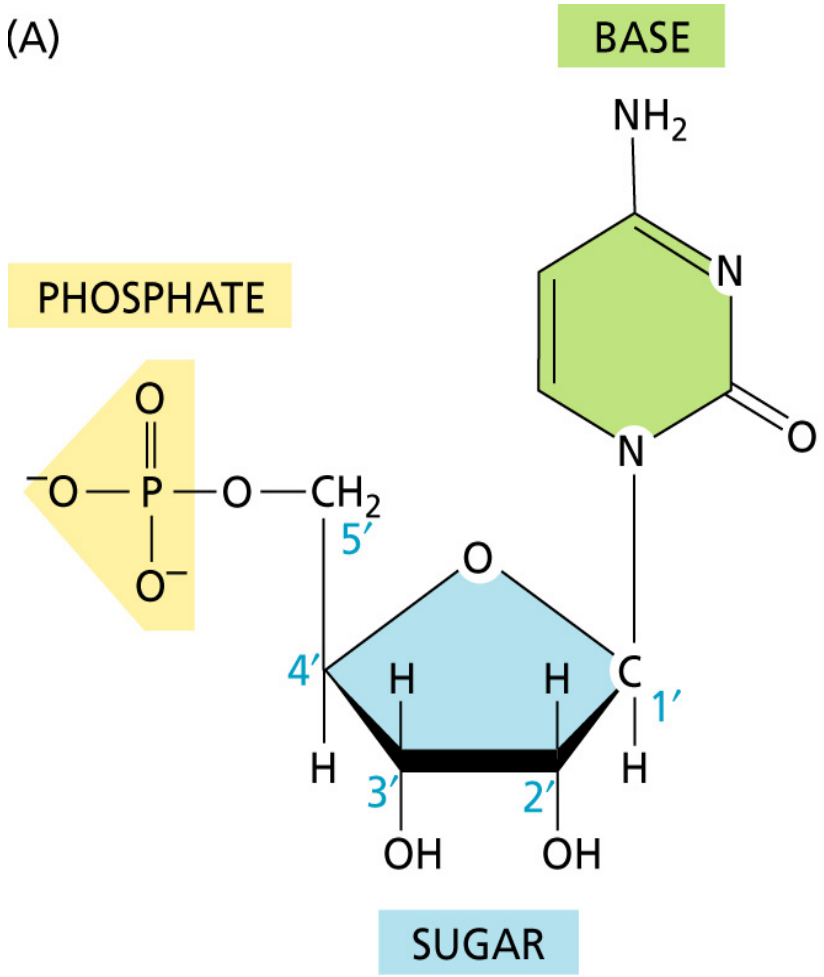
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Double-Stranded DNA

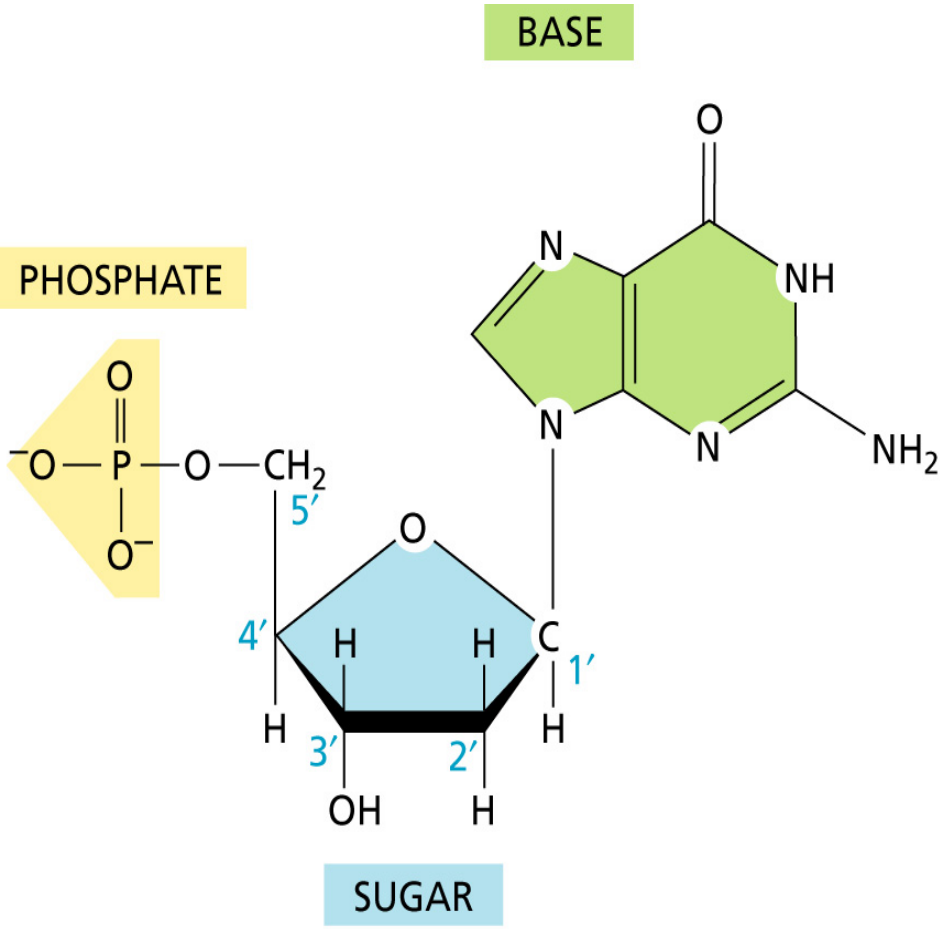


Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Sugars Found in Nucleic Acids

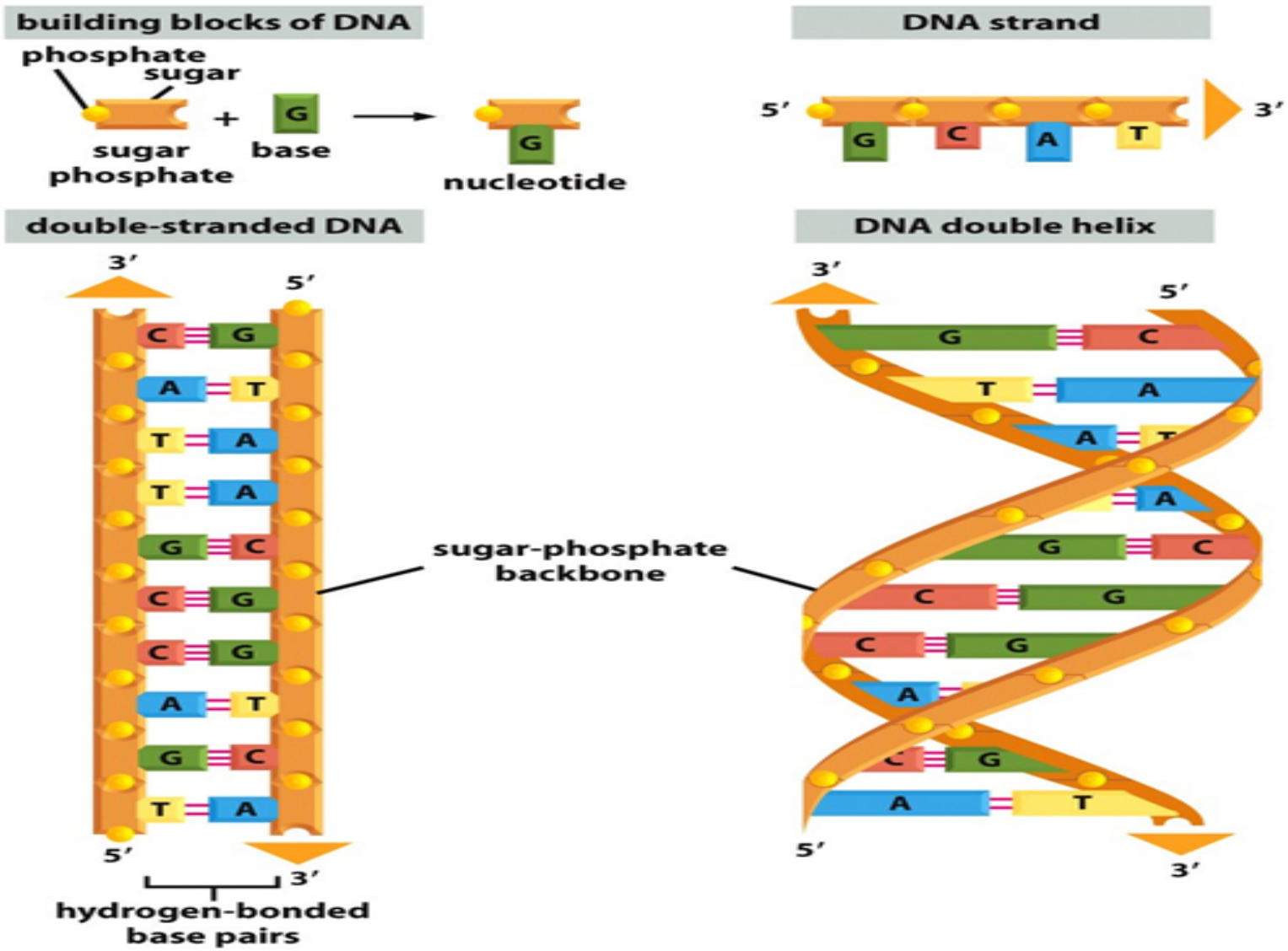


Pentose sugar present in **RNA**



Pentose sugar present in **DNA**

Double-Stranded DNA



Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

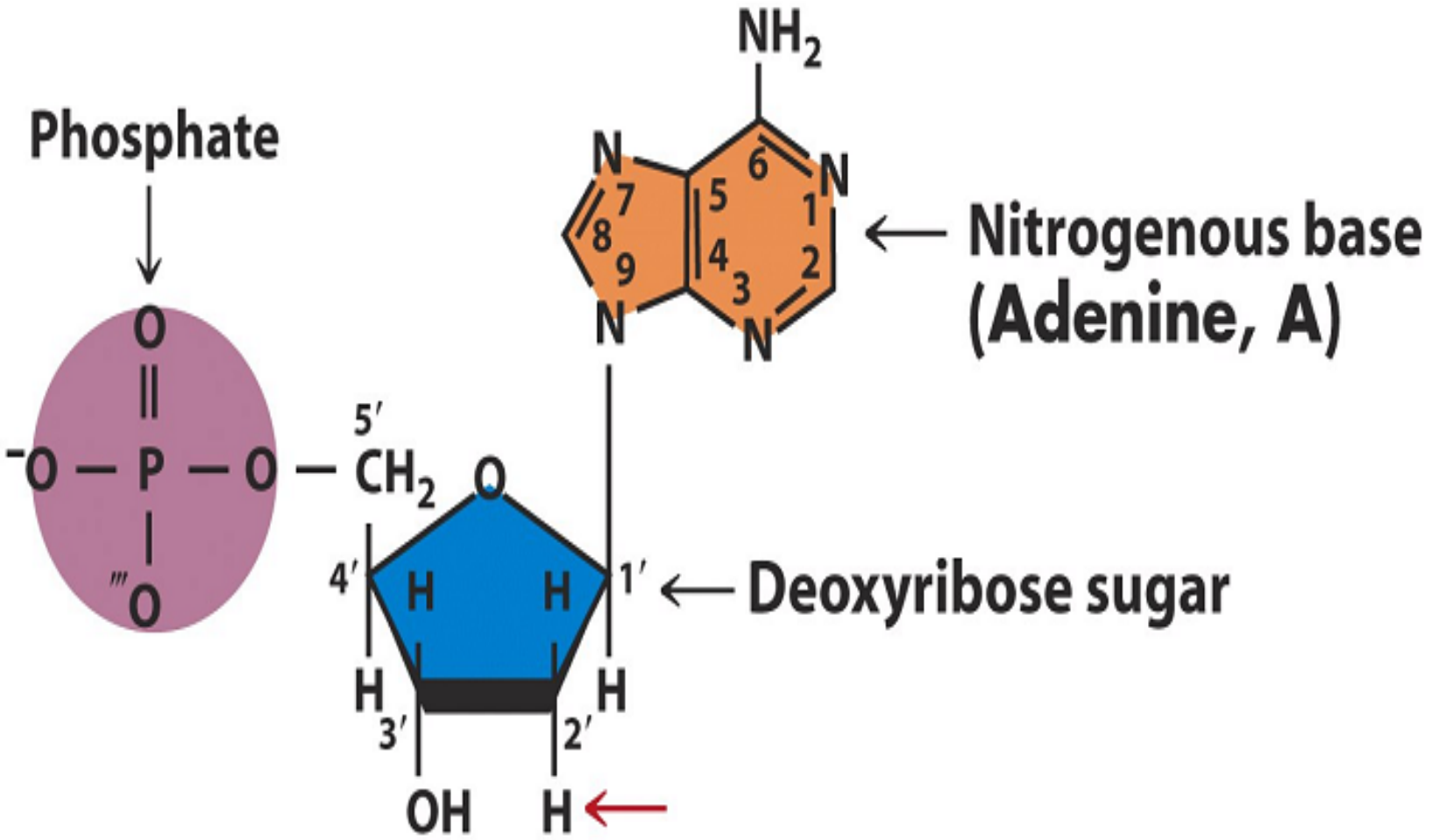
DNA Structure

- A **deoxyribonucleic acid** or **DNA** molecule is a double-stranded polymer composed of four basic molecular units called nucleotides.
- Each DNA nucleotide comprises
  - a phosphate group
  - a deoxyribose sugar
  - one of four nitrogen bases:

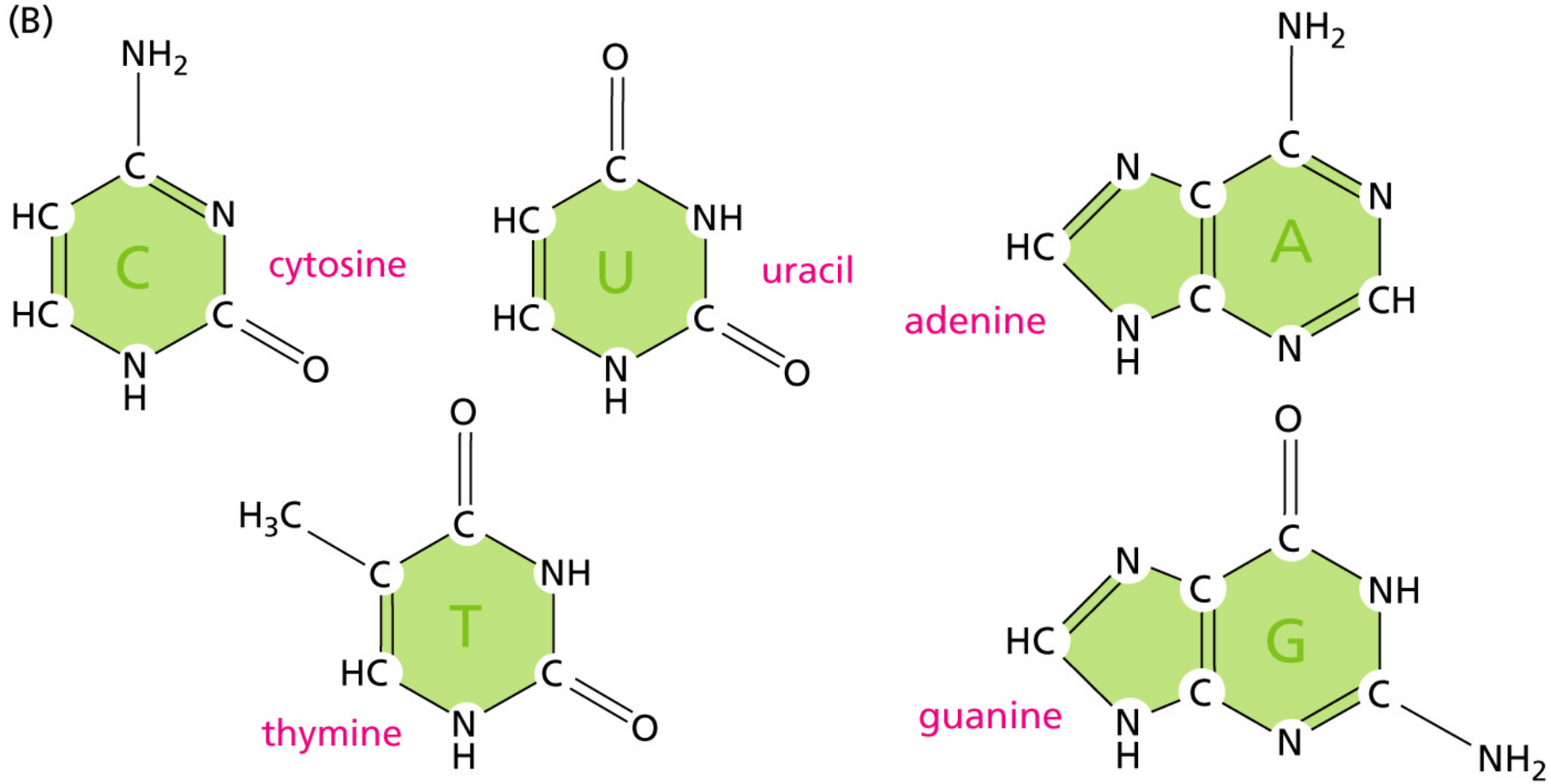
    purines: **adenine** (**A**) and **guanine** (**G**)

    pyrimidines: **cytosine** (**C**) and **thymine** (**T**).

A Nucleotide



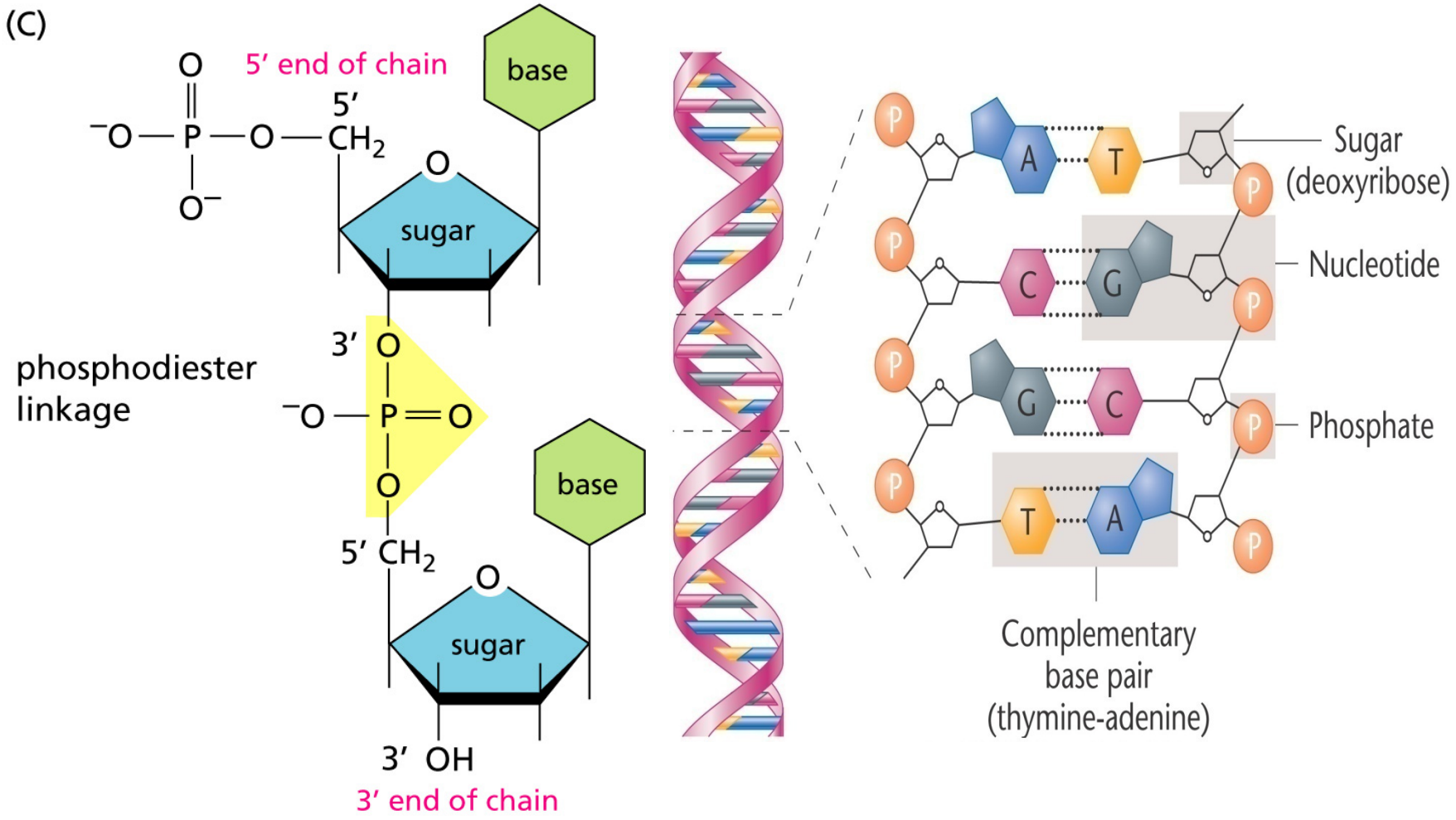
Purines and Pyrimidines



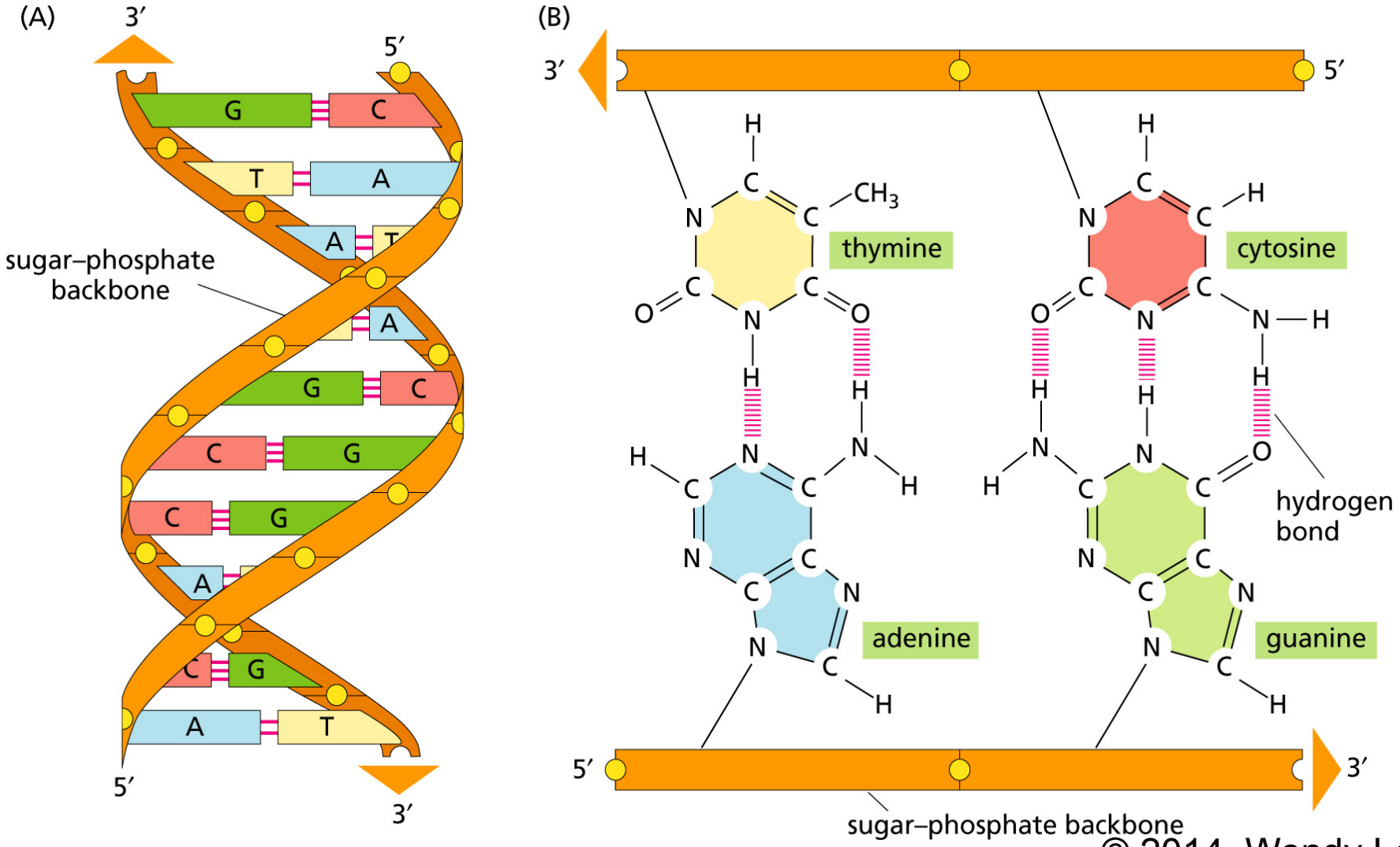
Double Helix

- The binding of two nucleotides forms a base pair.
- The double helix is formed by connecting complementary nucleotides A-T and C-G on two strands with hydrogen bonds.
- Knowledge of the sequence on one strand allows us to infer the sequence of the other strand.
- The bases are arranged along the sugar phosphate backbone in a particular order, known as the DNA sequence, encoding all genetic instructions for an organism.

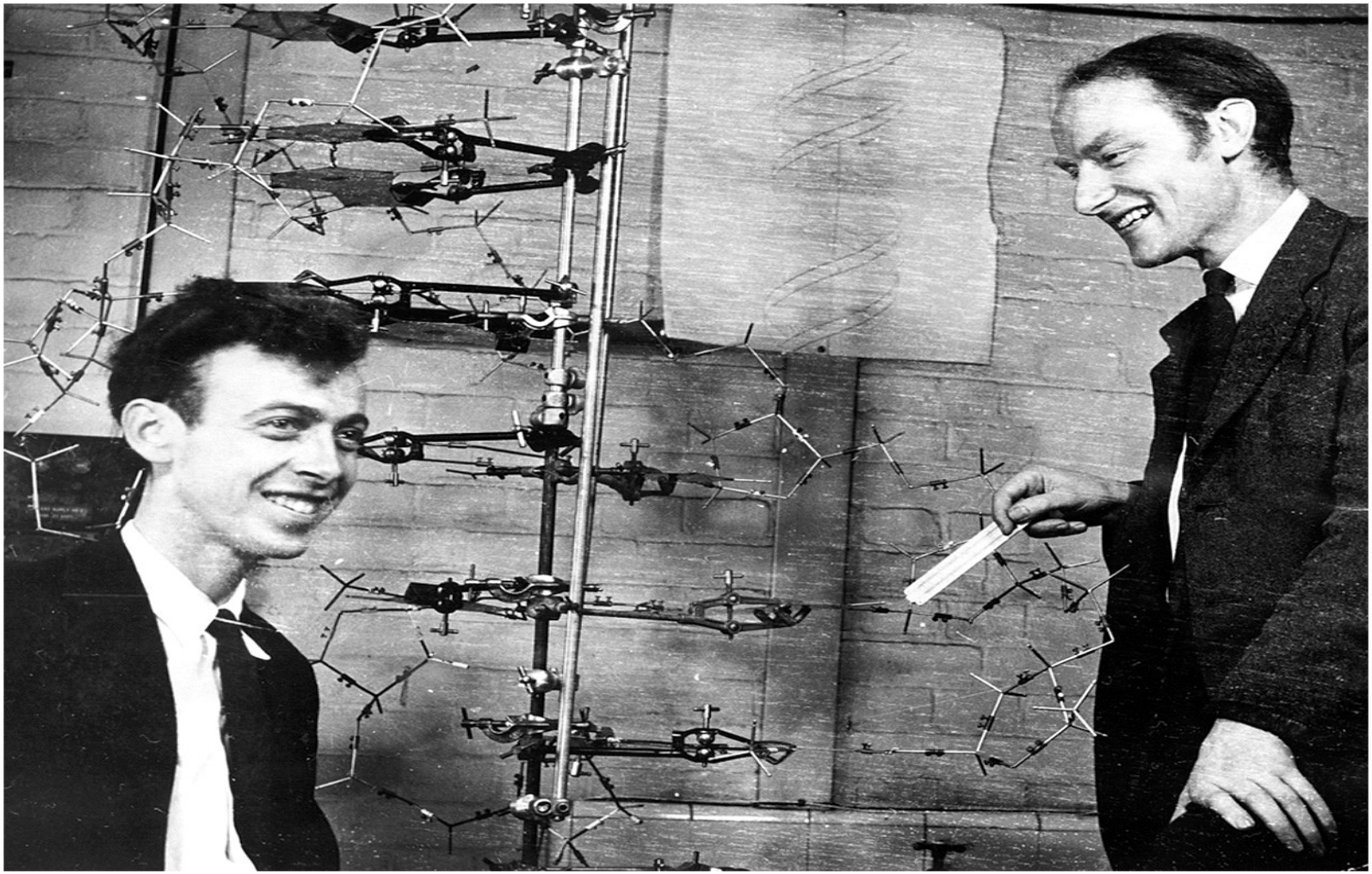
DNA Phosphodiester Backbone



Double Helical Structure of DNA

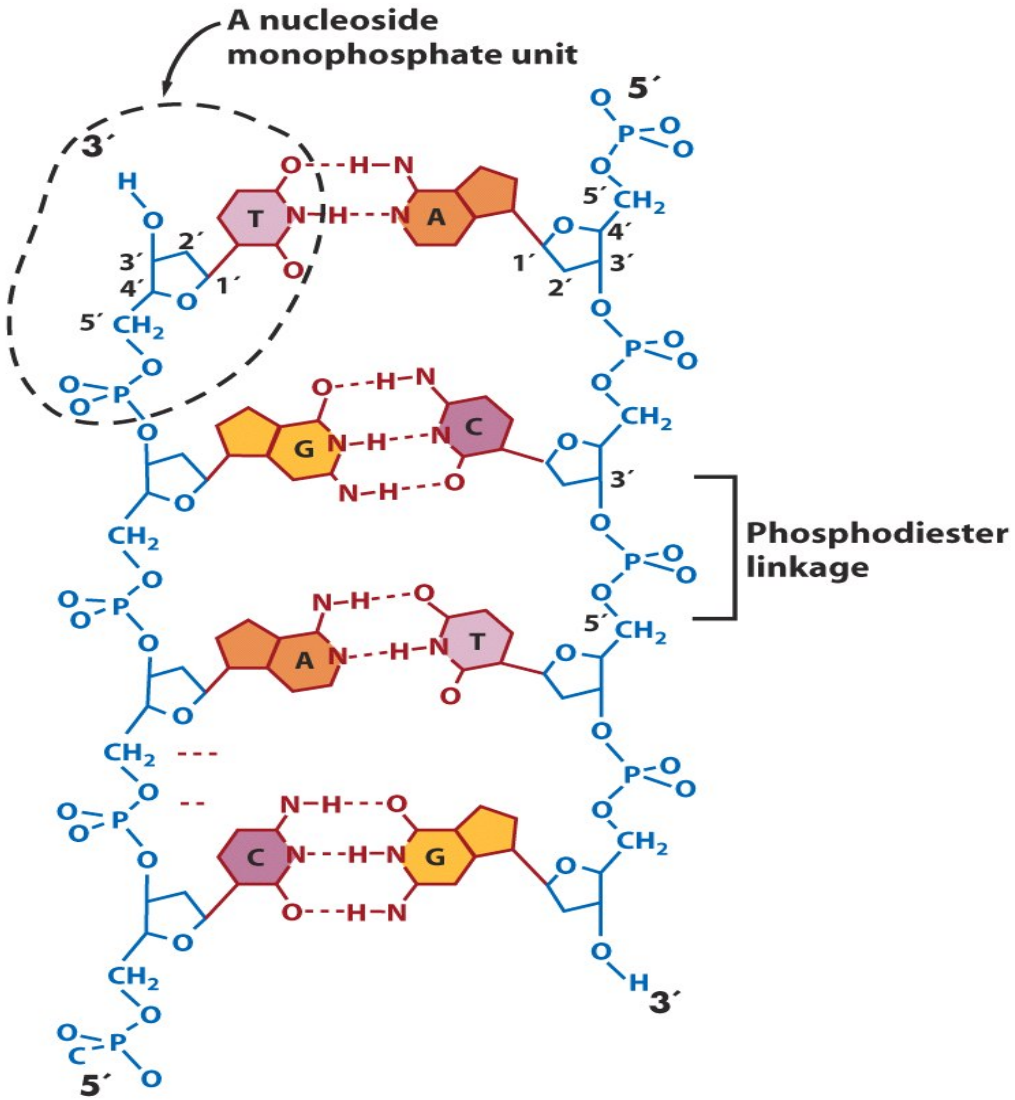


J. Watson and F. Crick



The Two Backbones of DNA

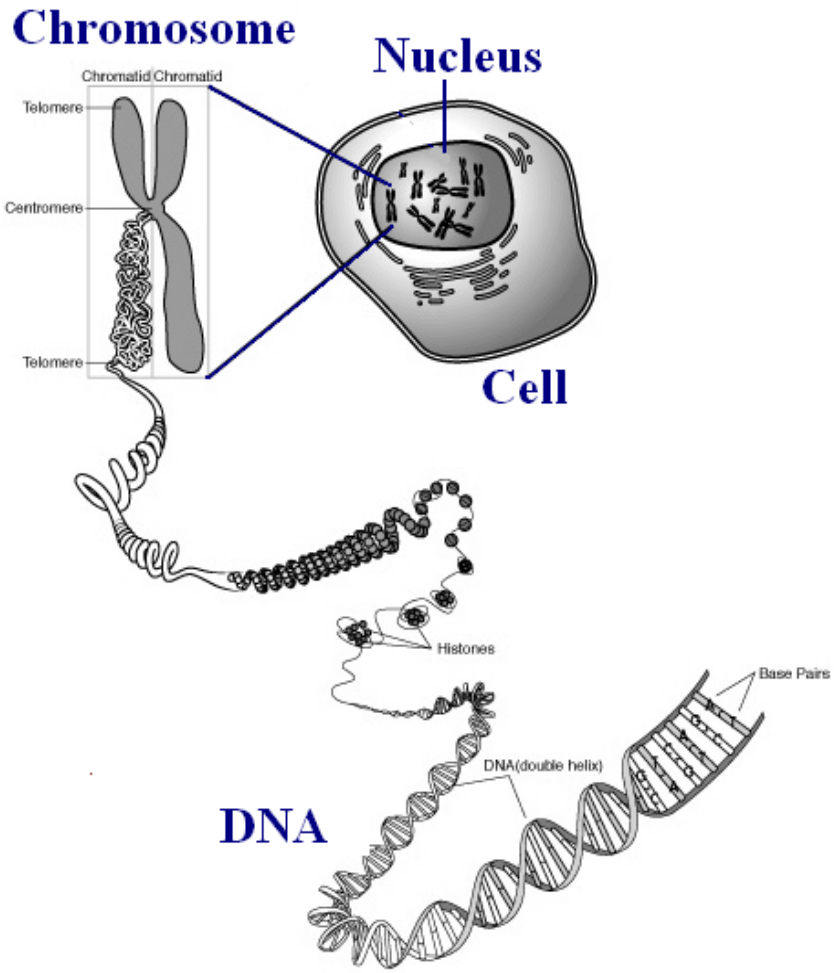
The DNA backbones have alternating sugar-phosphate components. The backbones run in opposite directions.

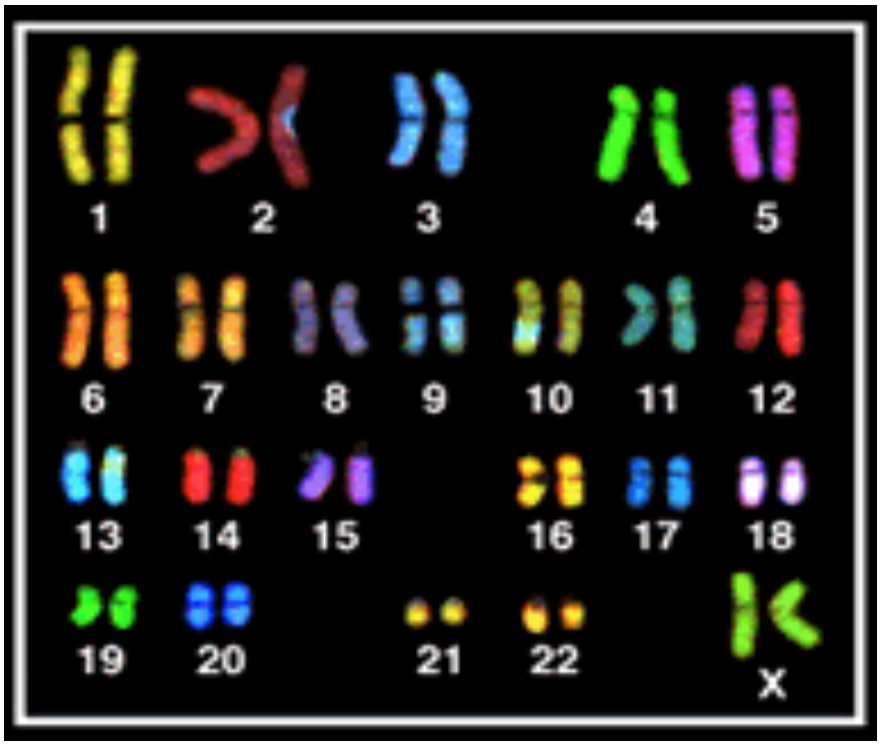


DNA and Chromosomes

- The **genome** is a complete set of instructions for making an organism, consists of tightly coiled threads of **DNA** organized into structures called **chromosomes**.
- Besides the reproductive cell and red blood cell, every single cell in the human body contains the human genome.

Chromosomes and Genome





Karyotype of a Male

Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Human Chromosomes

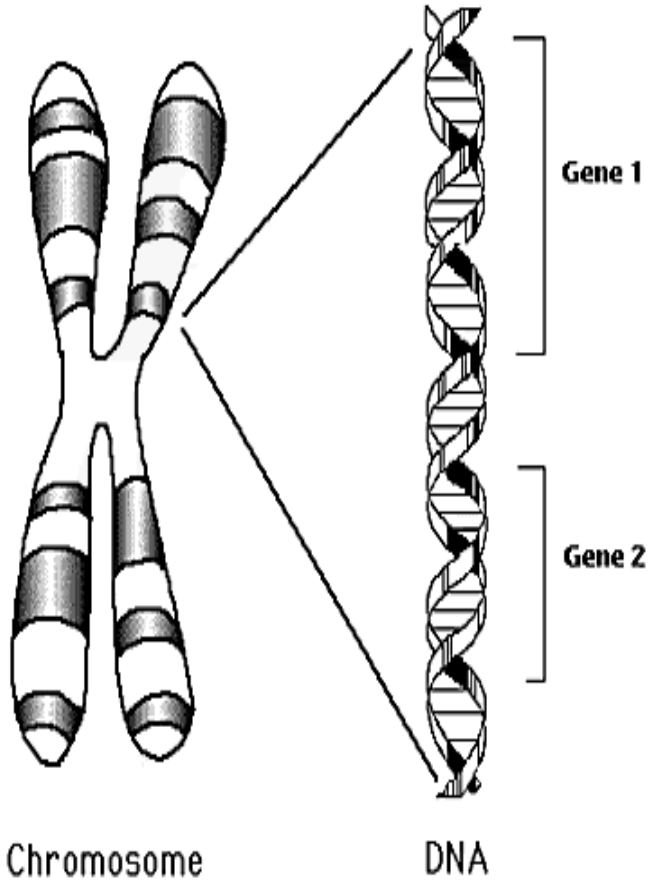
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Pairs of Chromosomes in Species

Table 3-2 Numbers of Pairs of Chromosomes in Different Species of Plants and Animals

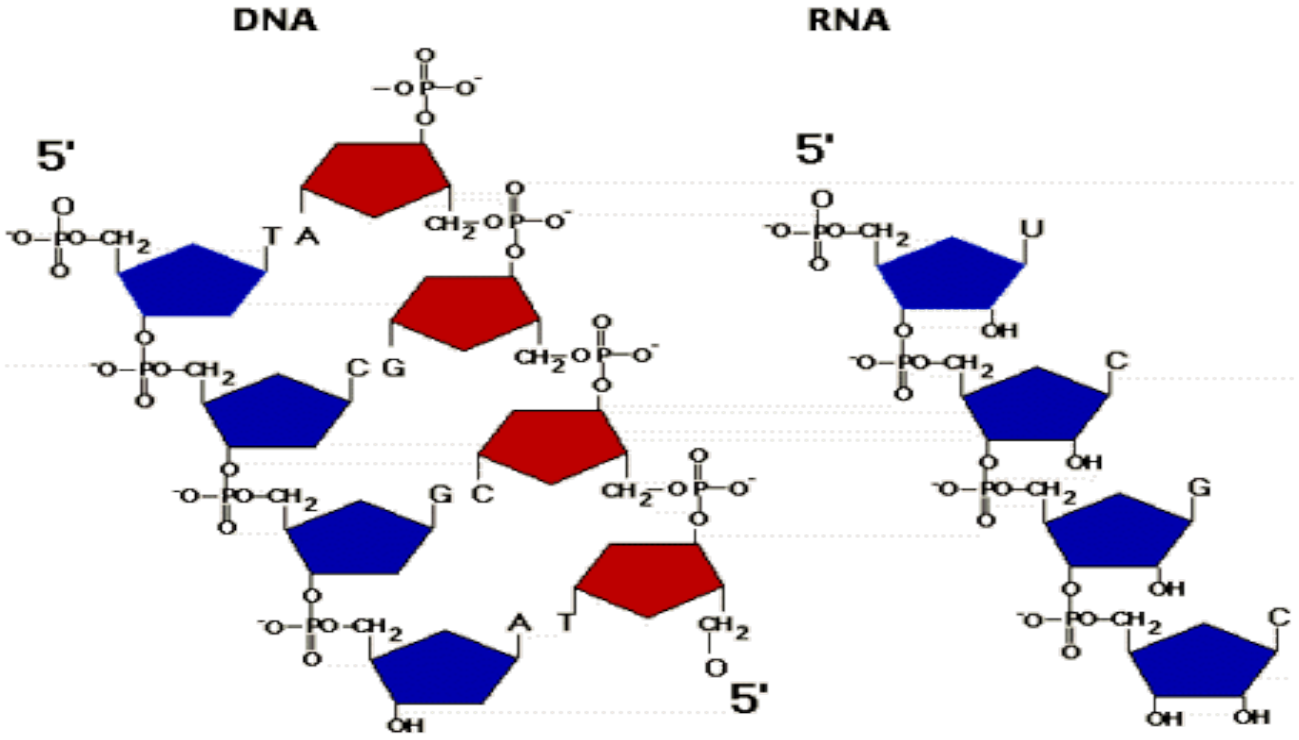
Common name	Scientific name	Number of chromosome pairs	Common name	Scientific name	Number of chromosome pairs
Mosquito	*Culex pipiens*	3	Wheat	*Triticum aestivum*	21
Housefly	*Musca domestica*	6	Human	*Homo sapiens*	23
Garden onion	*Allium cepa*	8	Potato	*Solanum tuberosum*	24
Toad	*Bufo americanus*	11	Cattle	*Bos taurus*	30
Rice	*Oryza sativa*	12	Donkey	*Equus asinus*	31
Frog	*Rana pipiens*	13	Horse	*Equus caballus*	32
Alligator	*Alligator mississipiensis*	16	Dog	*Canis familiaris*	39
Cat	*Felis domesticus*	19	Chicken	*Gallus domesticus*	39
House mouse	*Mus musculus*	20	Carp	*Cyprinus carpio*	52
Rhesus monkey	*Macaca mulatta*	21			

Genes



- A **gene** is a specific sequence of nucleotide bases along a chromosome carrying information for constructing a protein. A gene encodes a protein (or an RNA).
- The distance between **genes** is often much larger than the genes themselves.
- The human genome has around 23,500 genes.

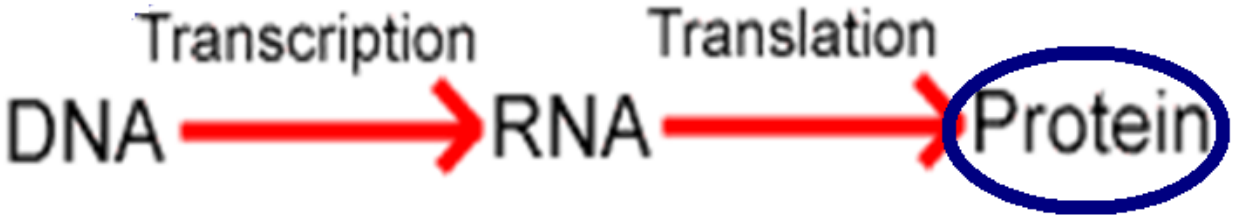
RNA

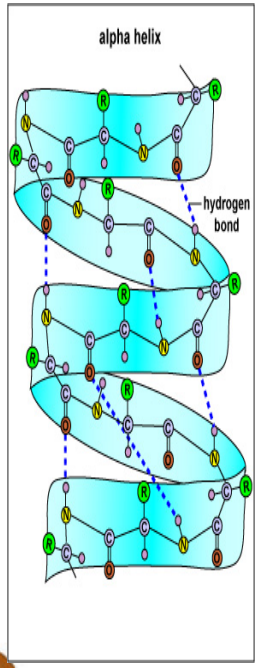


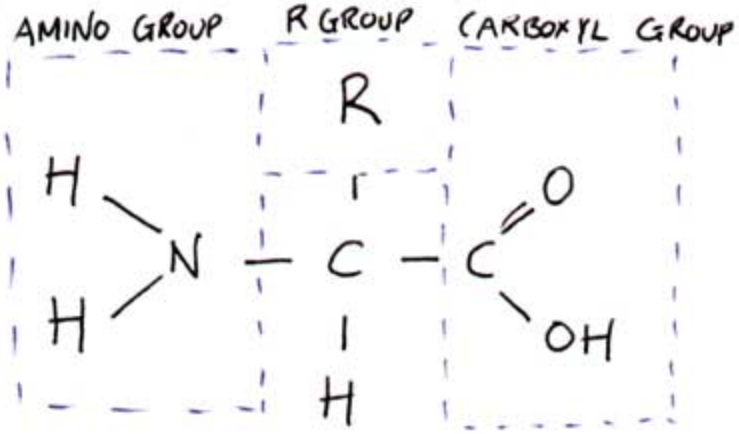
Ribonucleic Acid - RNA

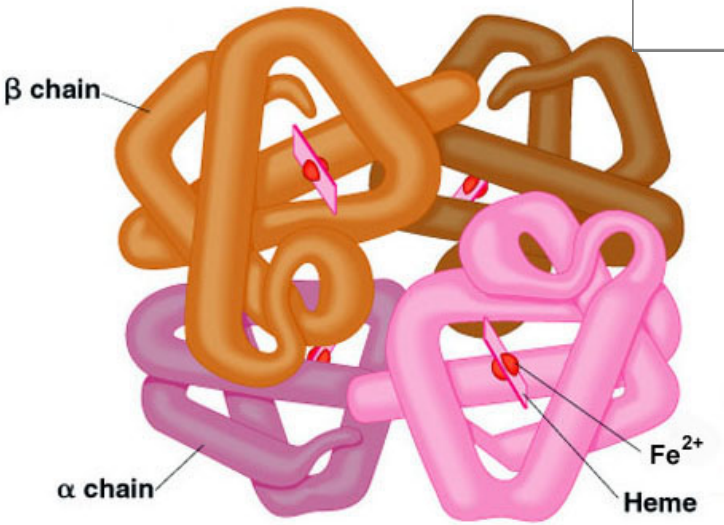
- **RNA** is found in the cell and can also carry genetic information.
- While DNA is located primarily in the nucleus, **RNA** can also be found in the cytoplasm.
- **RNA** is built from the nucleotides cytosine, guanine, adenine and uracil (U) (instead of thymine).
- **RNA** has its sugar phosphate backbone containing ribose.
- **RNA** forms a single strand.
- **RNA** molecules tend to have a less-regular three-dimensional structure than DNA.

Proteins





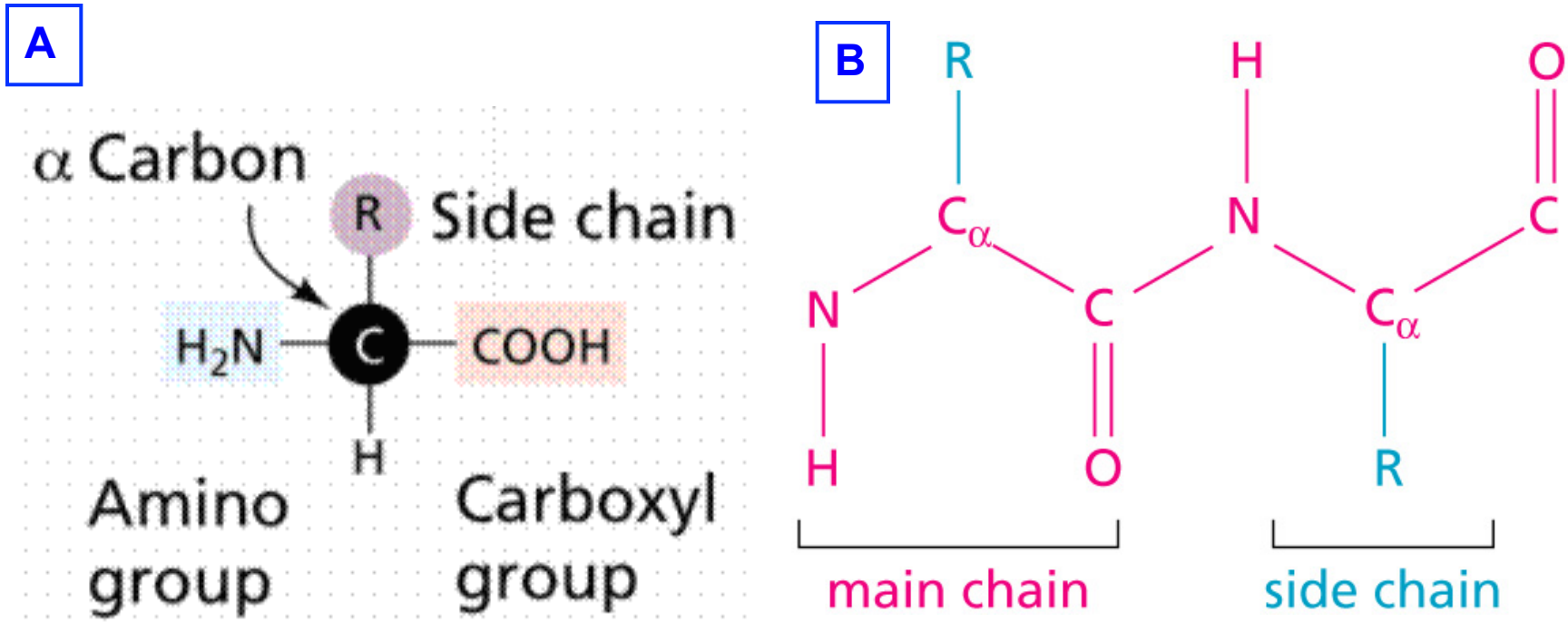




Proteins

- 20 different **amino acids** are used to synthesize **proteins**.
- The shape and other properties of each **protein** is dictated by the precise sequence of **amino acids** in it.
- The function of a **protein** is determined by its unique three-dimensional structure.

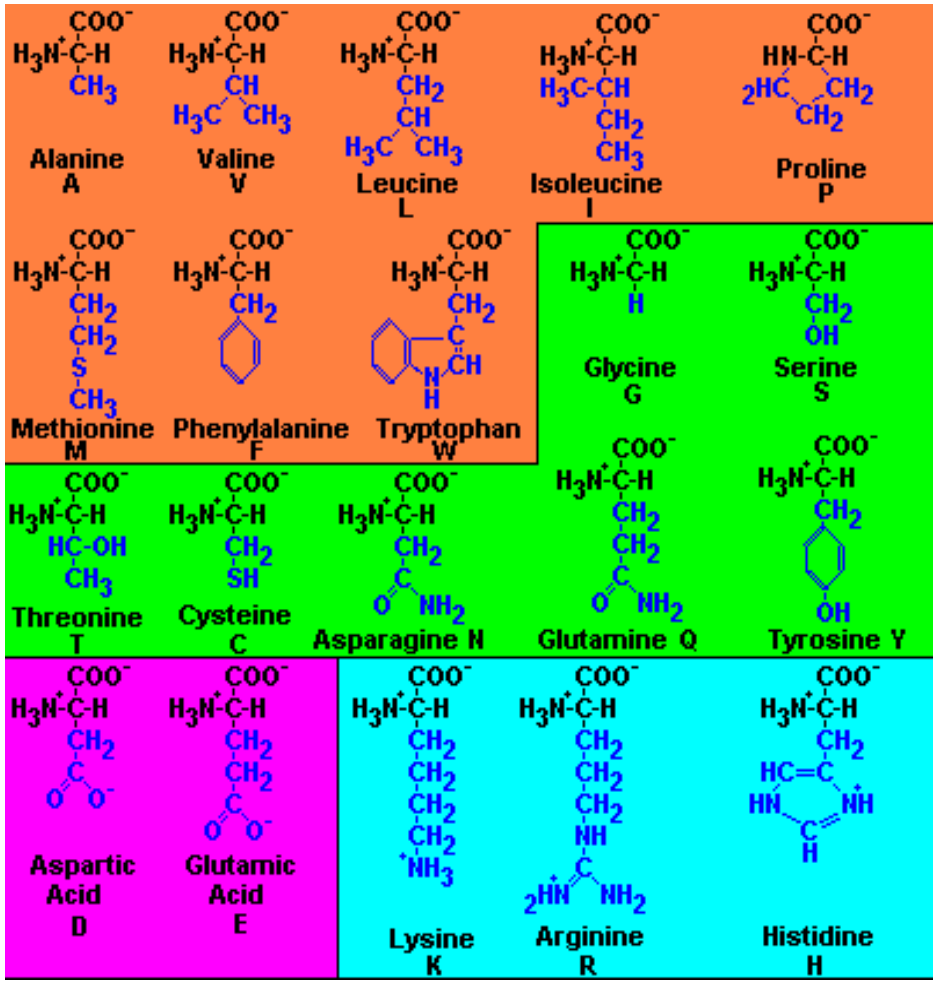
Structure of the Amino Acid



A) The functional properties of proteins are almost entirely due to the side chains in the amino acids.

B) Two amino acids: the main chain is in red and the side chain in blue.

The Twenty Amino Acids



Orange:

nonpolar and hydrophobic.

The other amino acids are: polar and hydrophilic - "water loving".

Magenta:

acidic - "carboxy" group in the side chain.

Light blue:

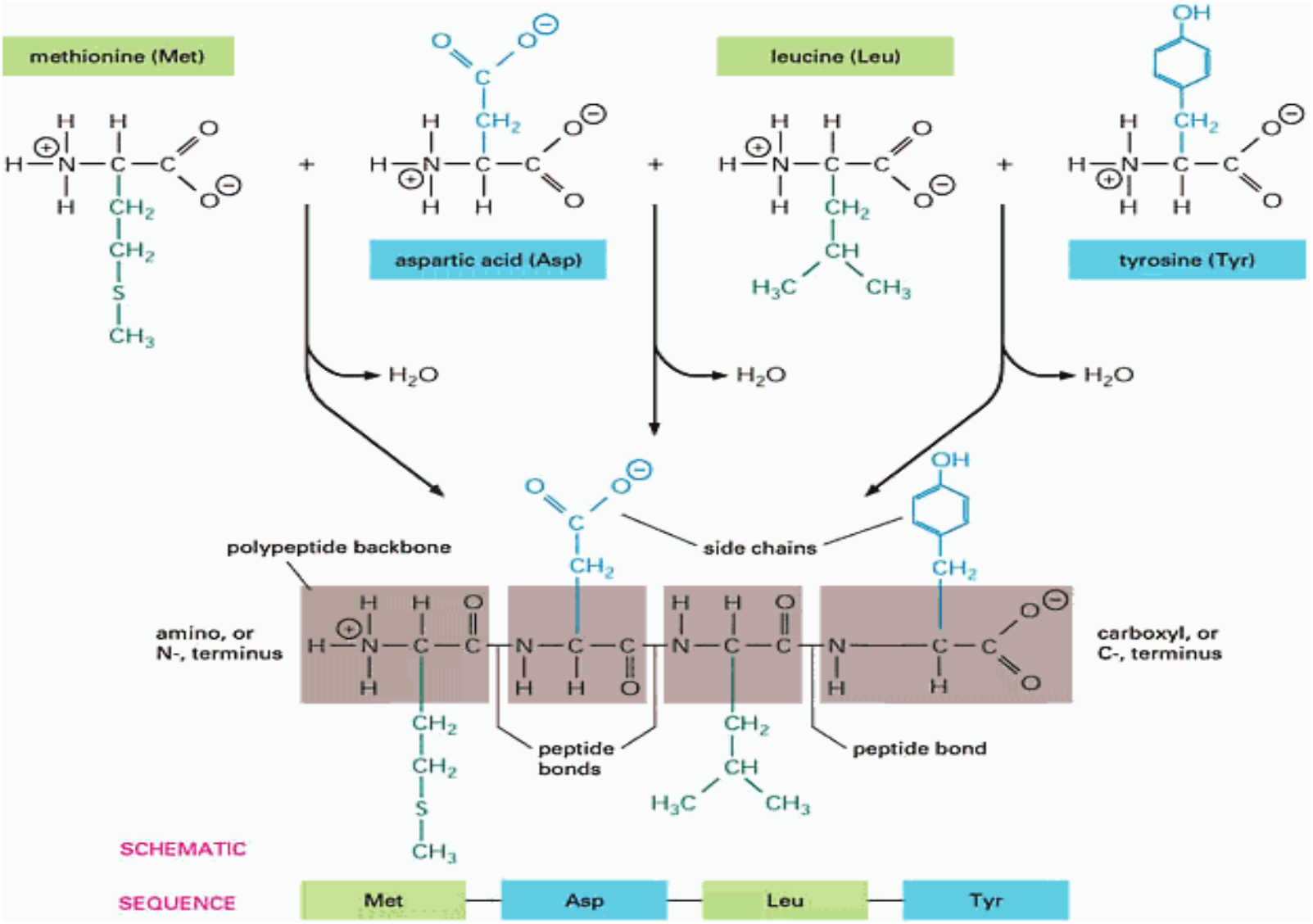
basic - "amine" group in the side chain.

The 20 Amino Acids

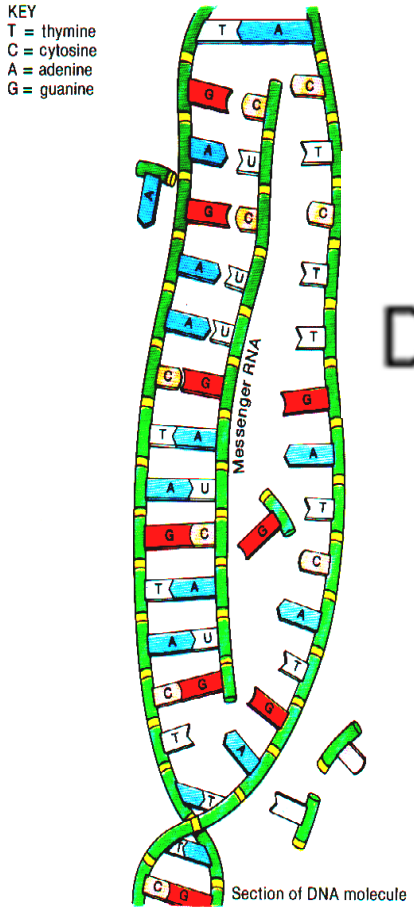
1-letter	3-letter	Amino acid
A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic Acid
E	Glu	Glutamic Acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine

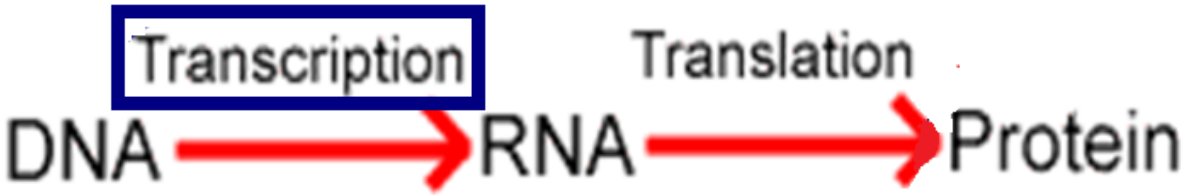
1-letter	3-letter	Amino Acid
M	Met	Methionin
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonin
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

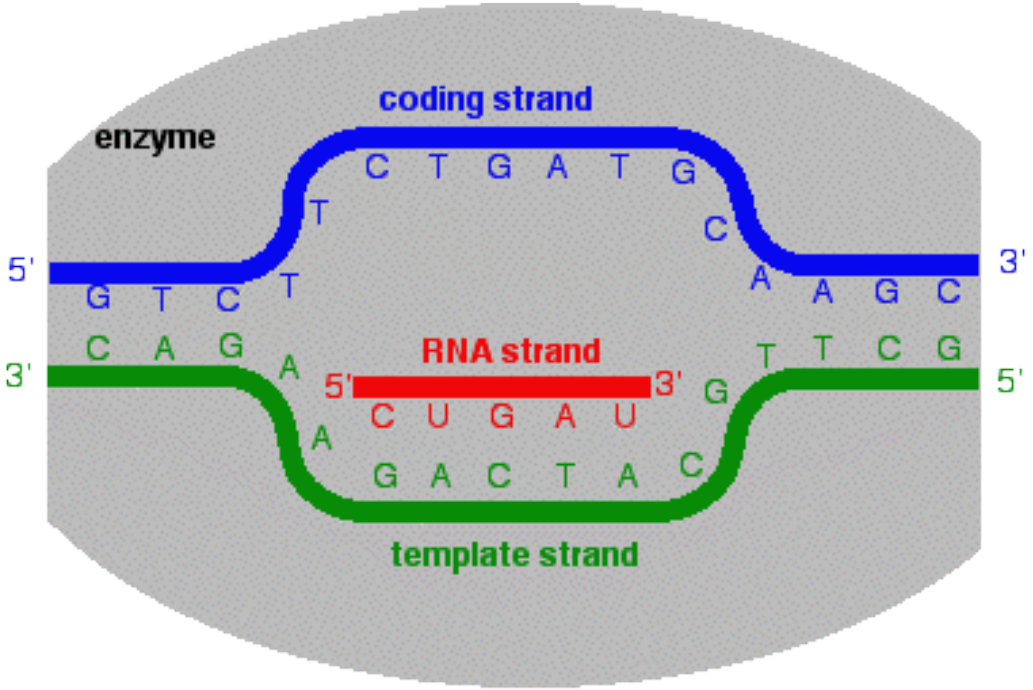
Protein Structure

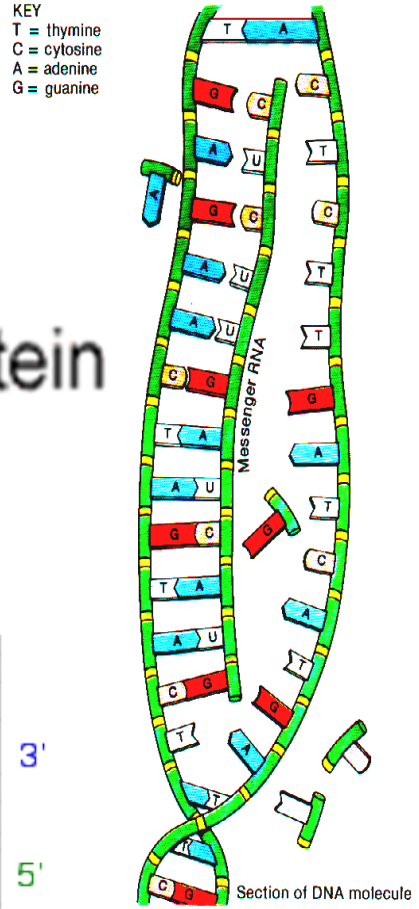


Transcription

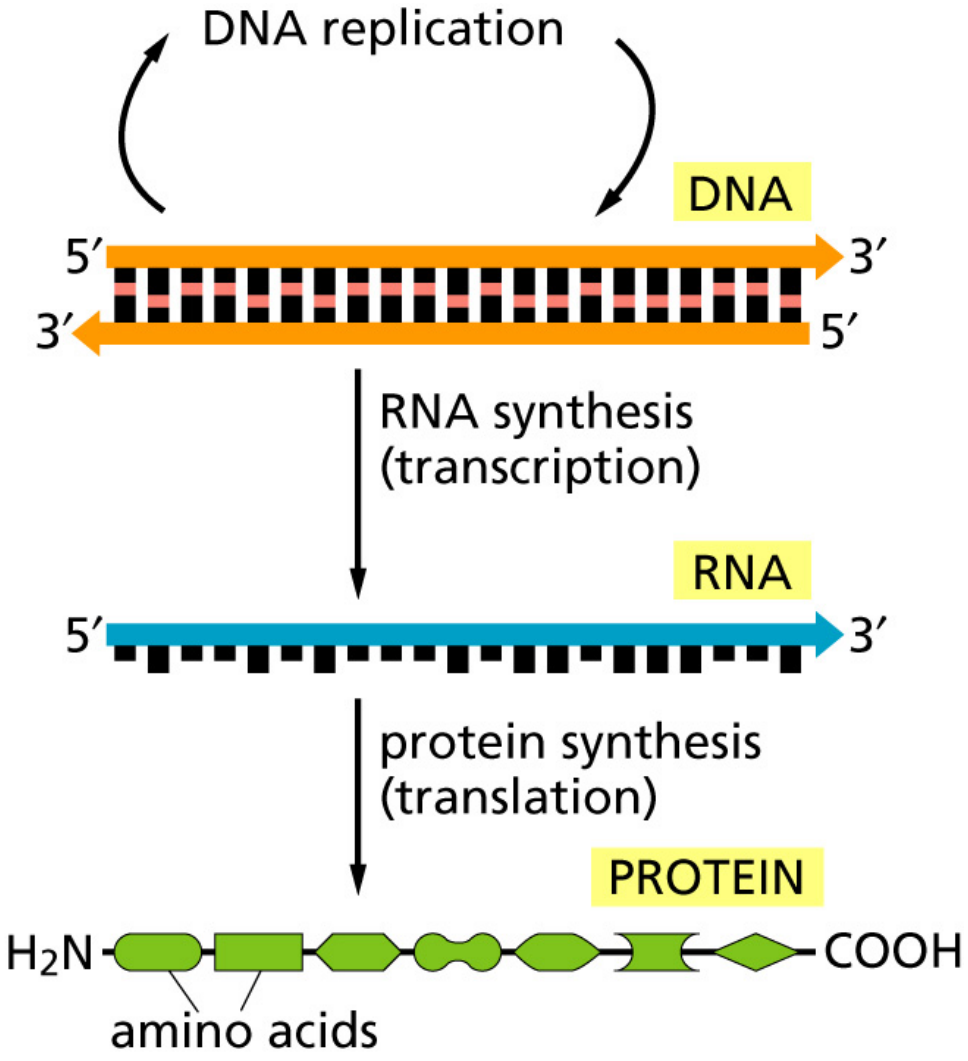






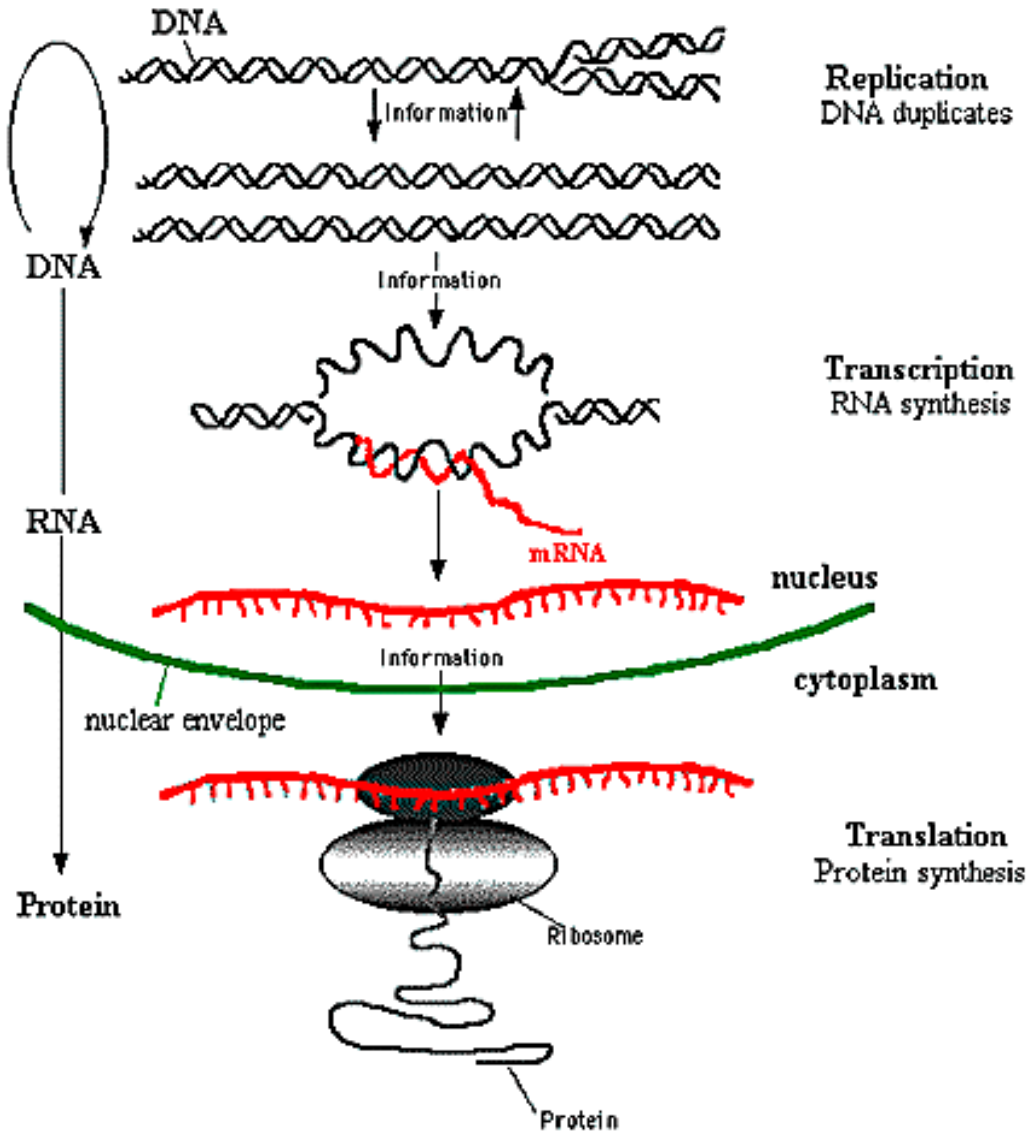


Central Dogma of Molecular Biology



According to the **central dogma of molecular biology**, there is a single direction of flow of genetic information from the **DNA**, which acts as the information store, through **RNA** molecules from which the information is translated into **proteins**.

Steps of the Central Dogma



Replication
DNA duplicates

Genetic information embodied in DNA is **replicated** into more DNA

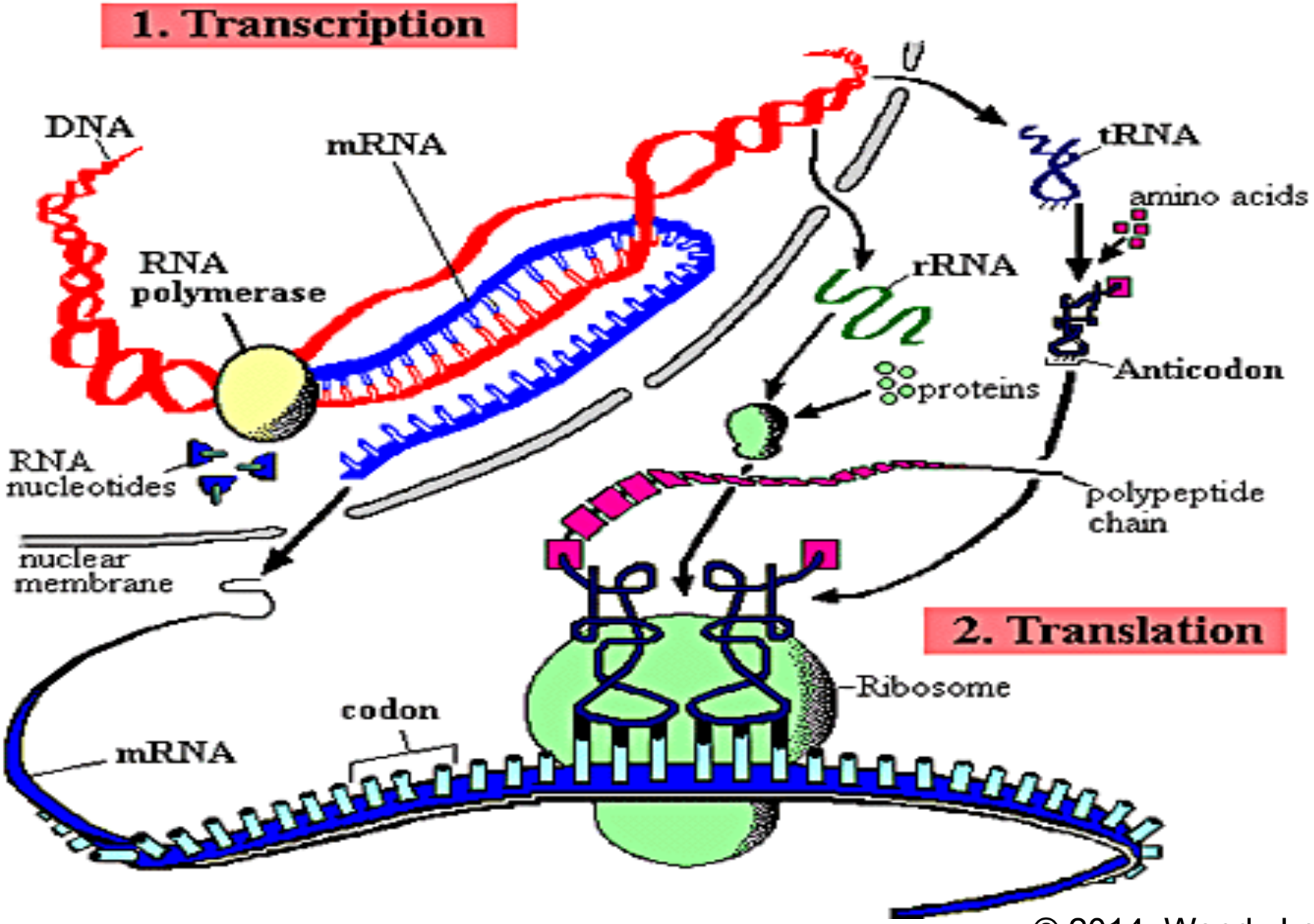
Transcription
RNA synthesis

The synthesis of an RNA from a sequence of DNA.
The resulting RNA is **mRNA**.

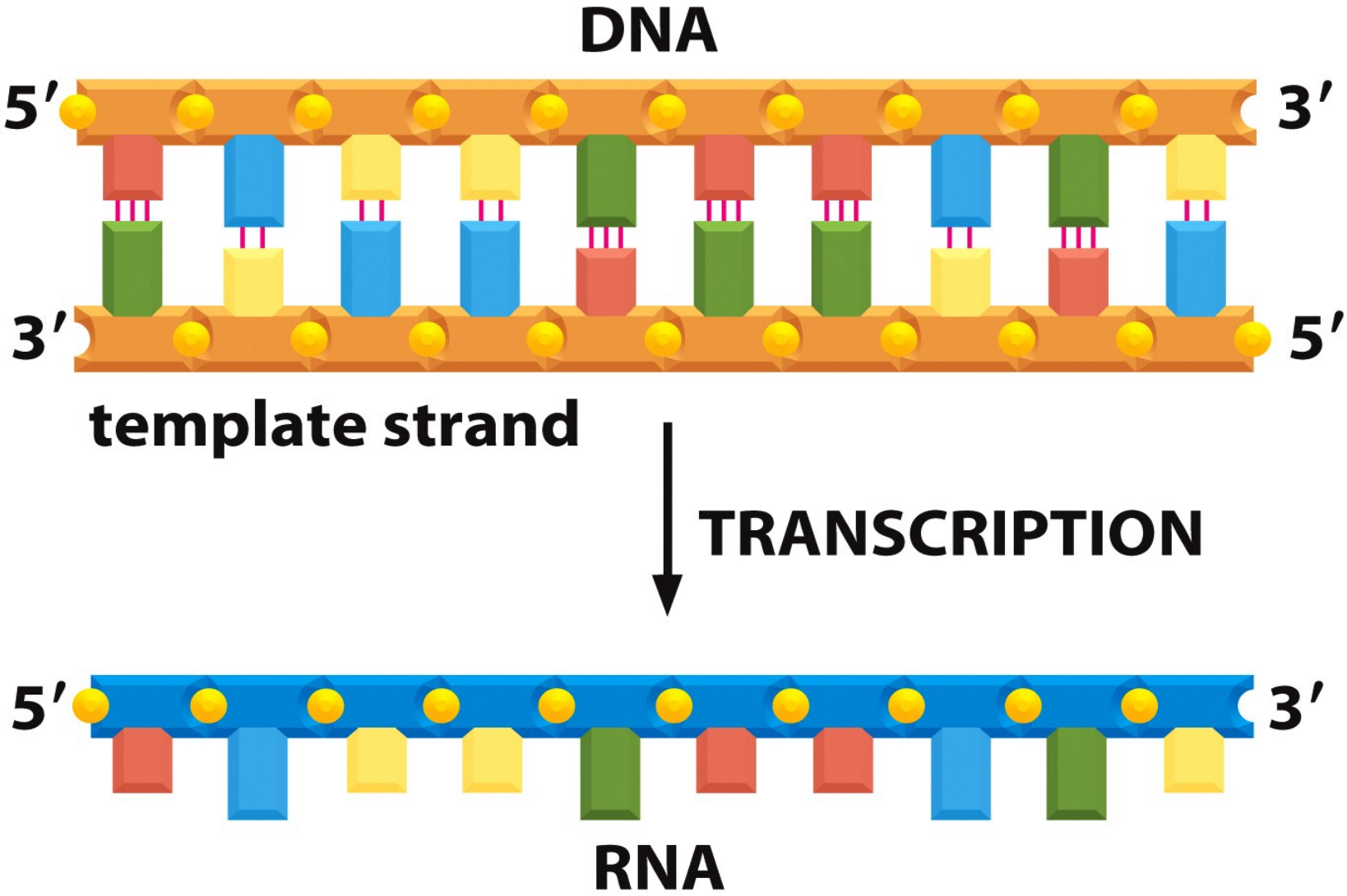
In eukaryotic cells, the mRNA is **spliced** and it migrates from the nucleus to the cytoplasm.

Translation
Protein synthesis

Messenger RNA carries coded information to ribosomes that "read"and use it for **protein synthesis**.

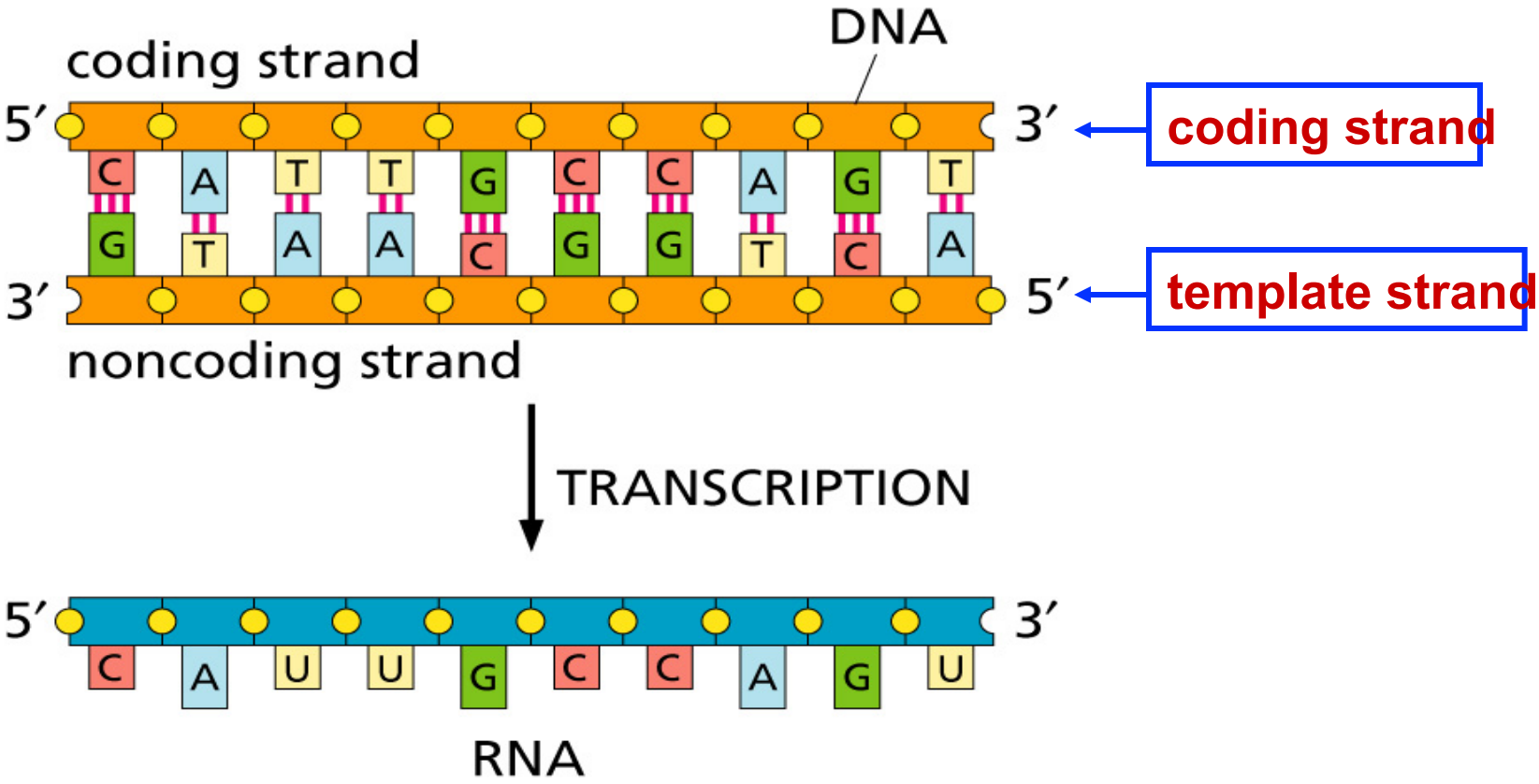


Transcription



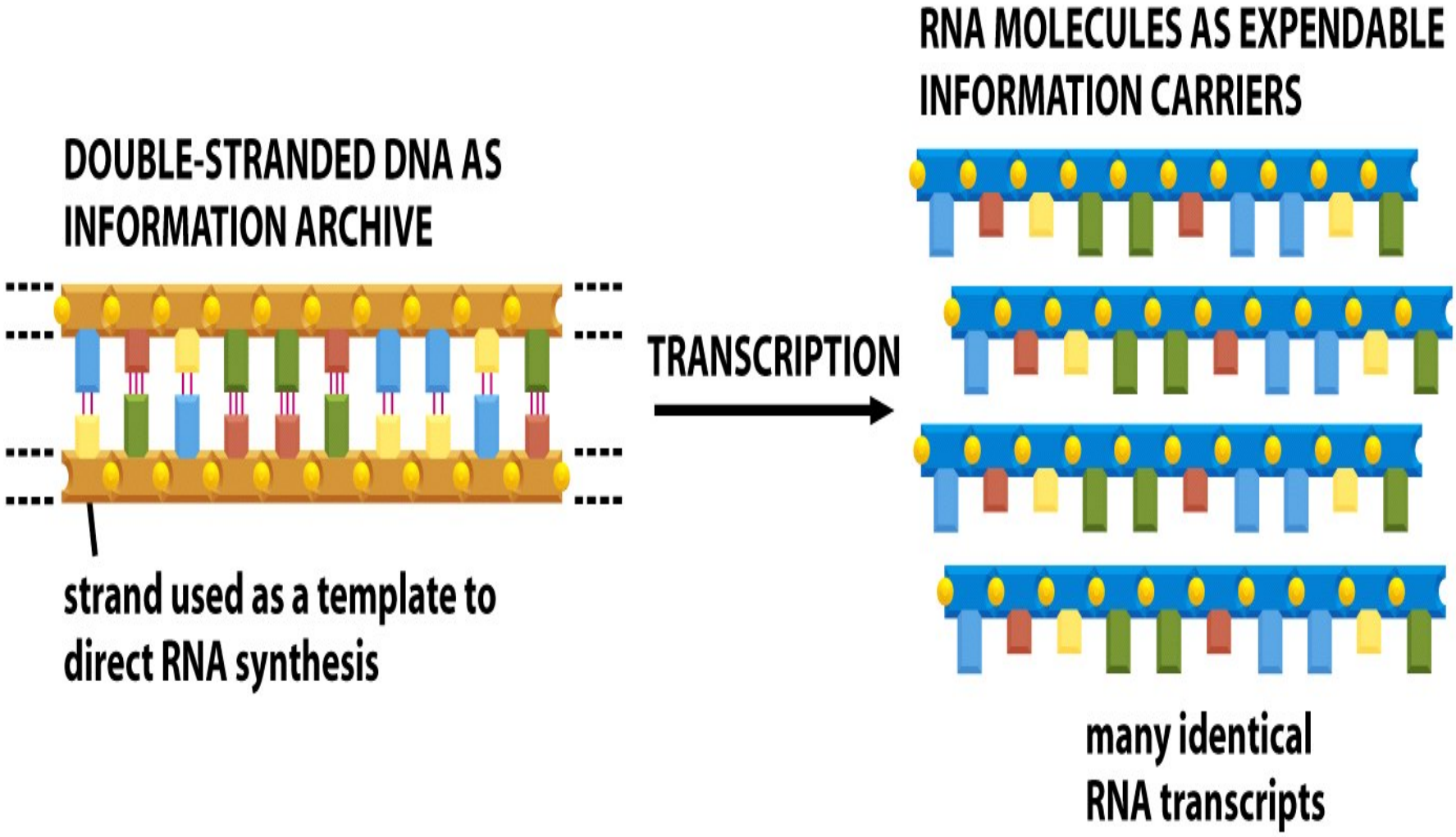
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Transcription



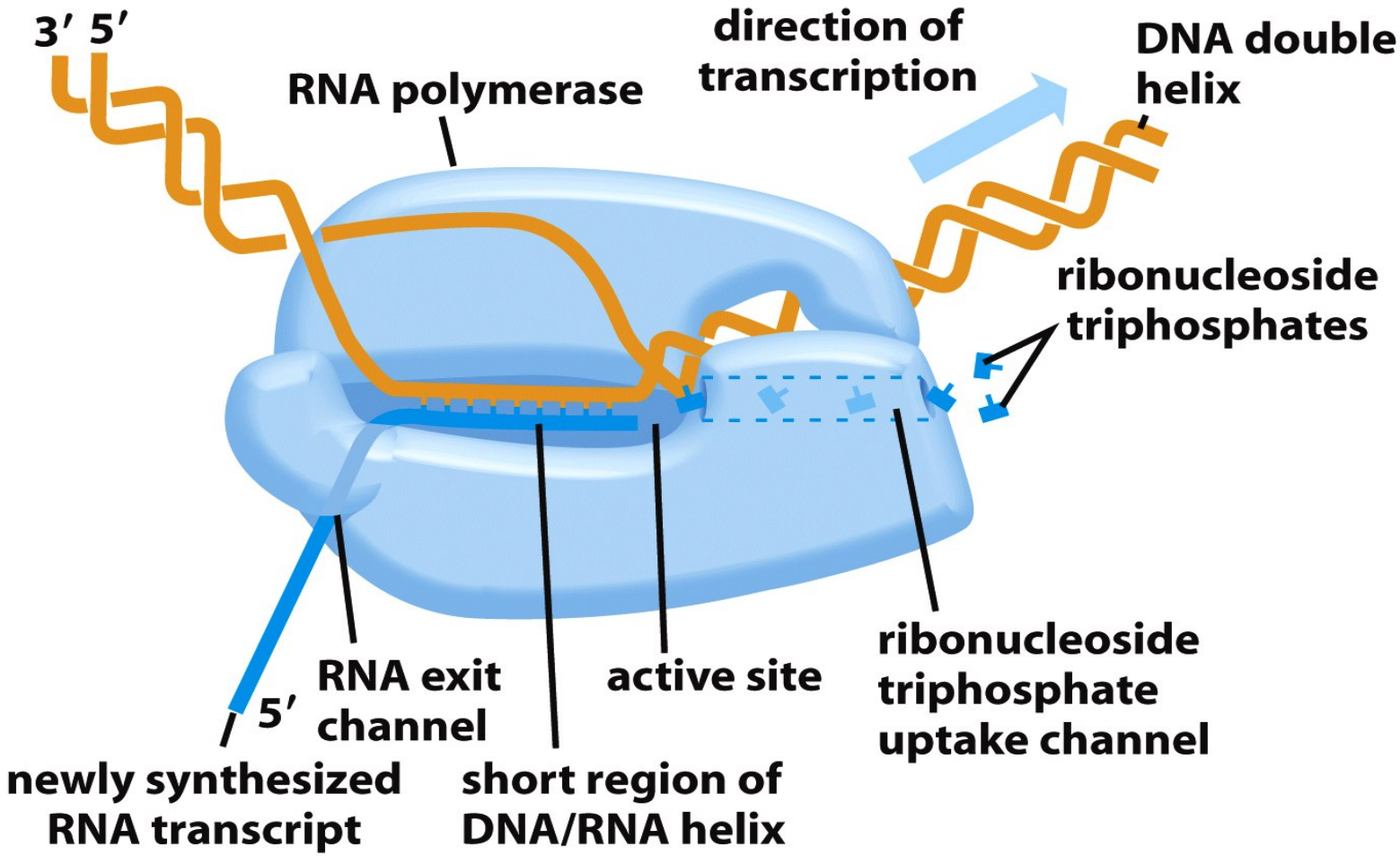
Transcription is the process in which one DNA strand: the **template strand**, is used to synthesize a complementary RNA.

RNA: Carrier of Information



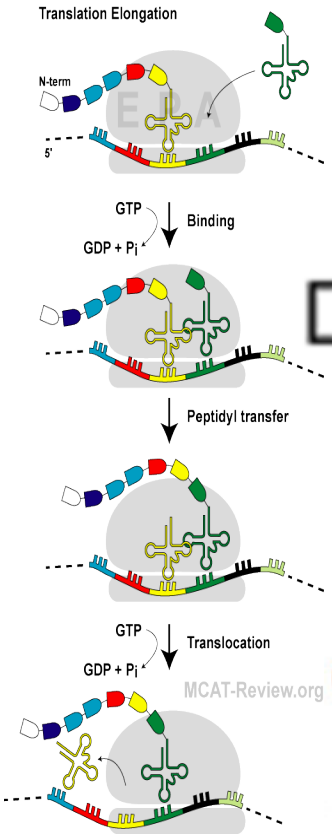
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

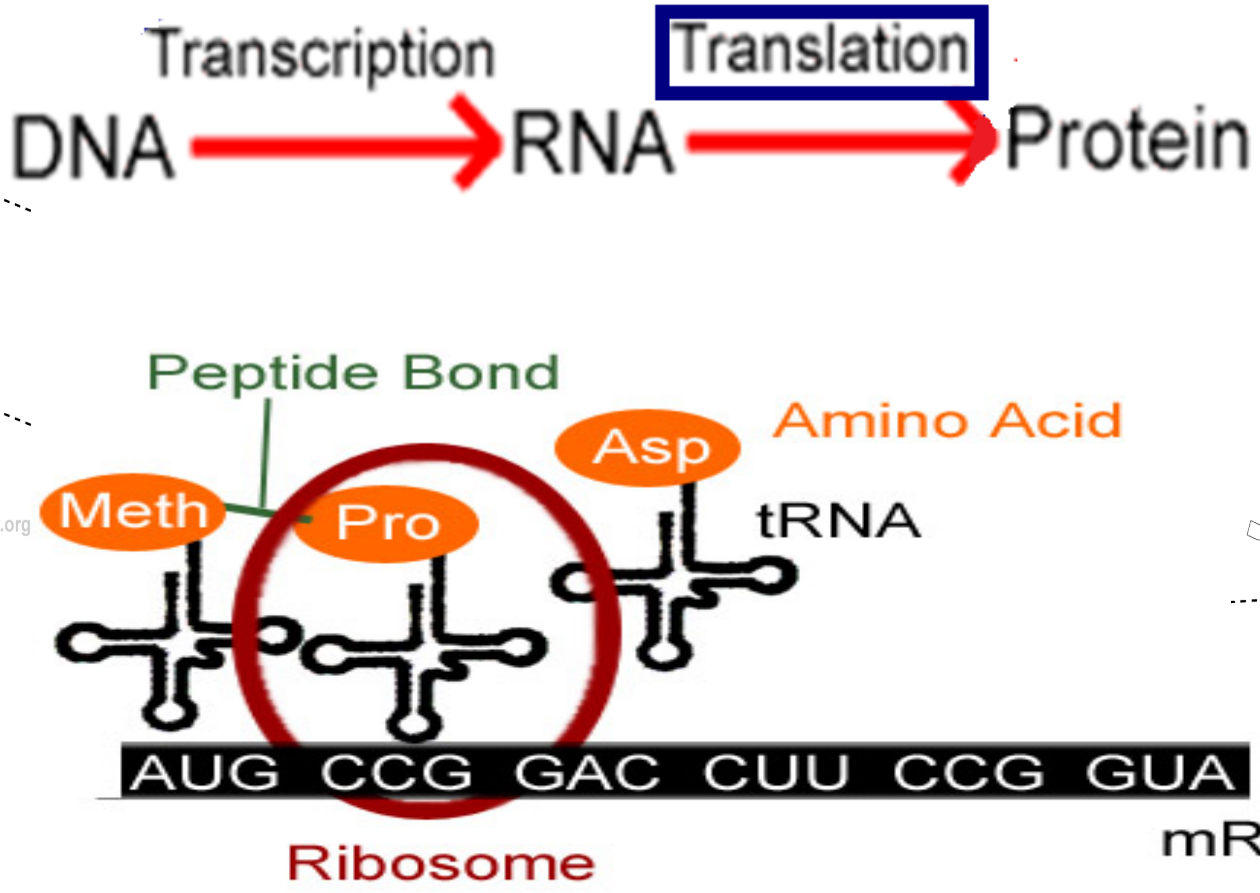
Synthesizing RNA from 5' to 3'

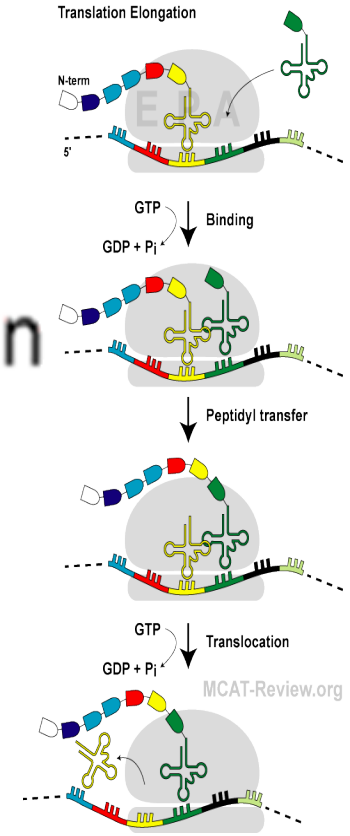


Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

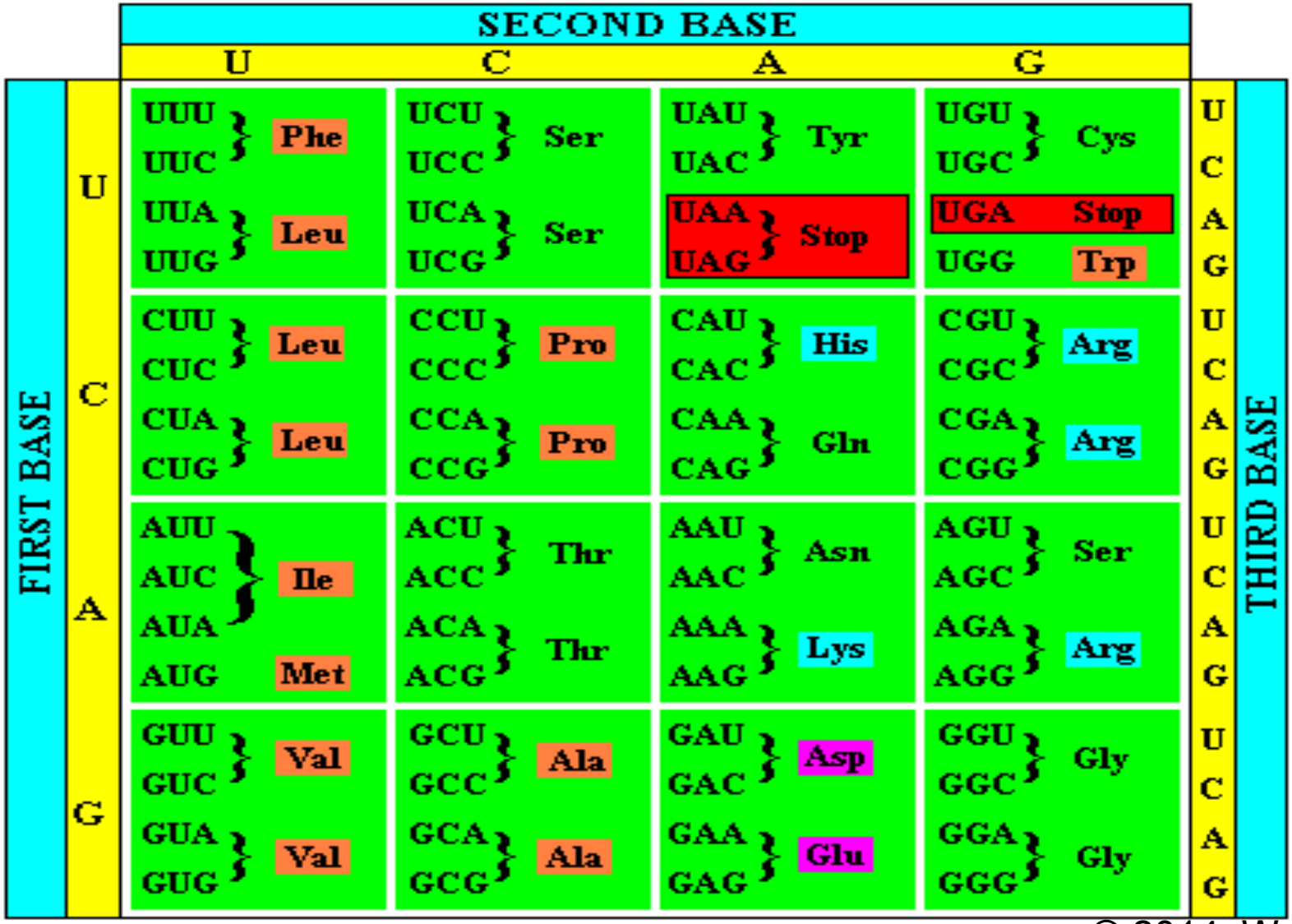
Translation







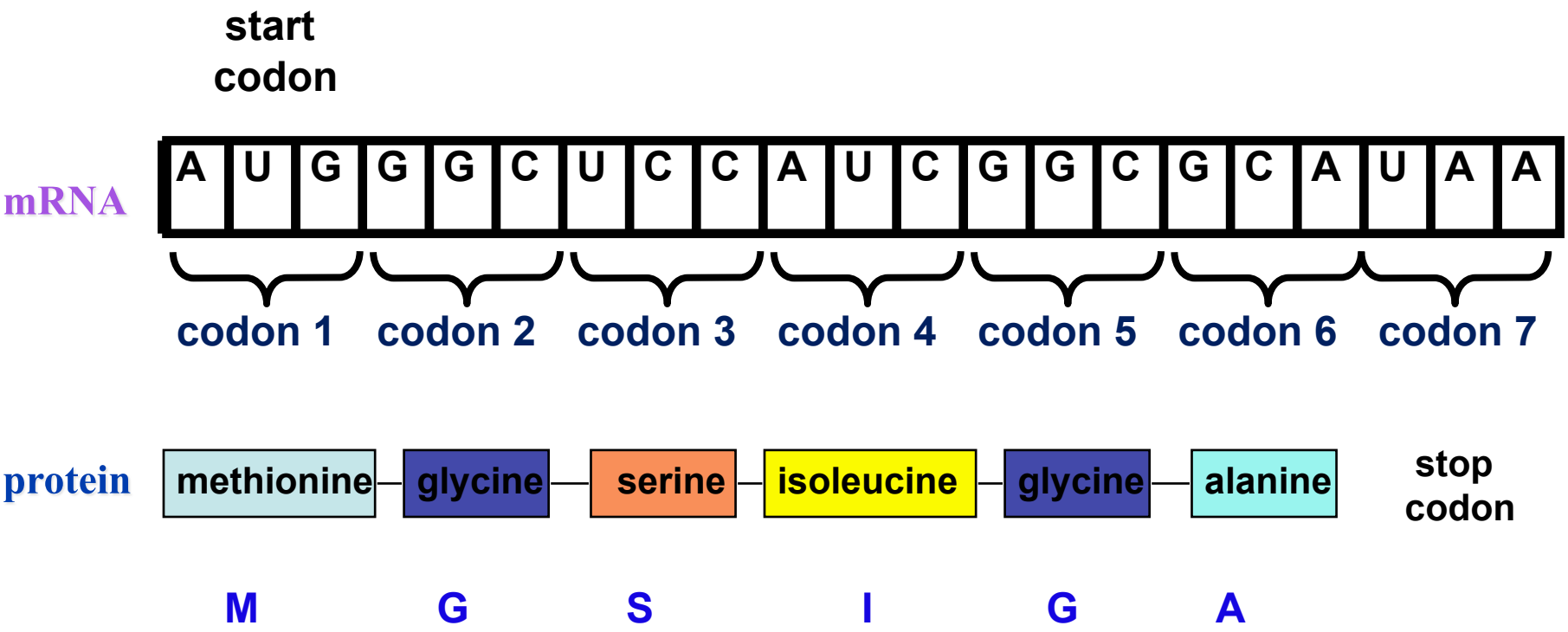
The Genetic Code



The Genetic Code

The Genetic Code							
		Second Codon Position					
		U	C	A	G		
First Codon Position (5' End)	U	UUU Phe (F) UUC Phe (F) UUA Leu (L) UUG Leu (L)	UCU Ser (S) UCC Ser (S) UCA Ser (S) UCG Ser (S)	UAU Tyr (Y) UAC Tyr (Y) UAA *Stop* UAG *Stop*	UGU Cys (C) UGC Cys (C) UGA *Stop* UGG Trp (W)	U C A G	Third Codon Position (3' End)
	C	CUU Leu (L) CUC Leu (L) CUA Leu (L) CUG Leu (L)	CCU Pro (P) CCC Pro (P) CCA Pro (P) CCG Pro (P)	CAU His (H) CAC His (H) CAA Gln (Q) CAG Gln (Q)	CGU Arg (R) CGC Arg (R) CGA Arg (R) CGG Arg (R)	U C A G	
	A	AUU Ile (I) AUC Ile (I) AUA Ile (I) AUG Met (M)	ACU Thr (T) ACC Thr (T) ACA Thr (T) ACG Thr (T)	AAU Asn (N) AAC Asn (N) AAA Lys (K) AAG Lys (K)	AGU Ser (S) AGC Ser (S) AGA Arg (R) AGG Arg (R)	U C A G	
	G	GUU Val (V) GUC Val (V) GUA Val (V) GUG Val (V)	GCU Ala (A) GCC Ala (A) GCA Ala (A) GCG Ala (A)	GAU Asp (D) GAC Asp (D) GAA Glu (E) GAG Glu (E)	GGU Gly (G) GGC Gly (G) GGA Gly (G) GGG Gly (G)	U C A G	

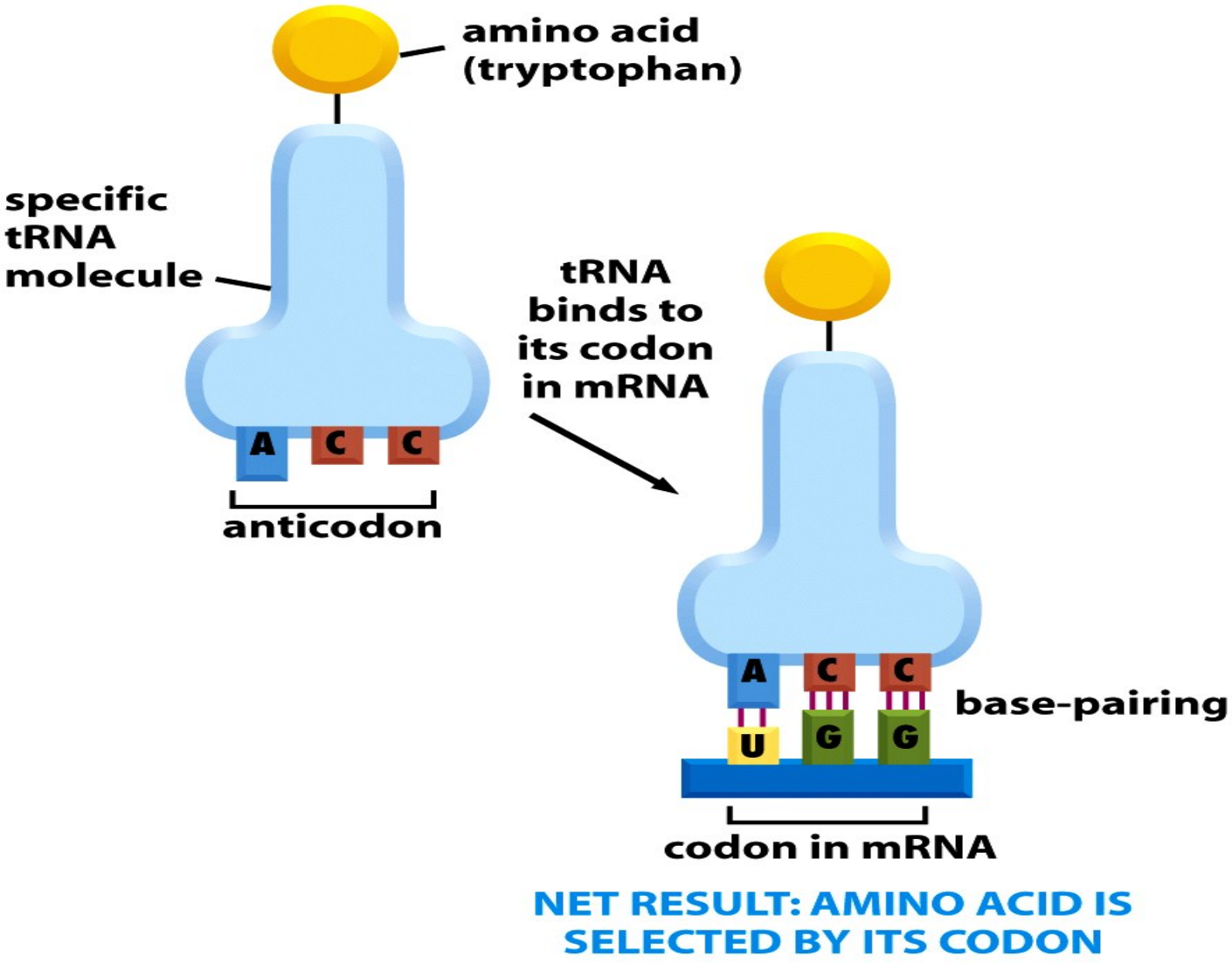
Translation: An Example



Transfer RNA and Translation

- The translation from nucleotides to amino acid is done by means of **transfer RNA** (**tRNA**) molecules, each specific for one amino acid and for a particular **triplet** of nucleotides in mRNA called a **codon**.
- The family of tRNA molecules enables the codons in a mRNA molecule to be **translated** into the sequence of amino acids in the protein.

tRNA, Anticodon and Codon



Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

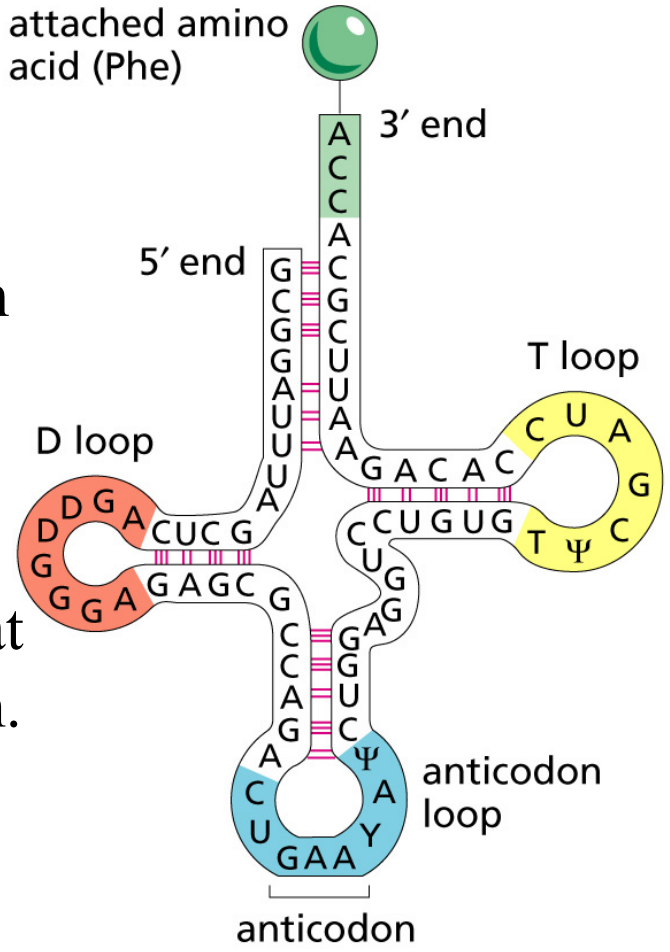
Codons and Anticodons

At least one kind of tRNA is present for each of the 20 amino acids used in protein synthesis.

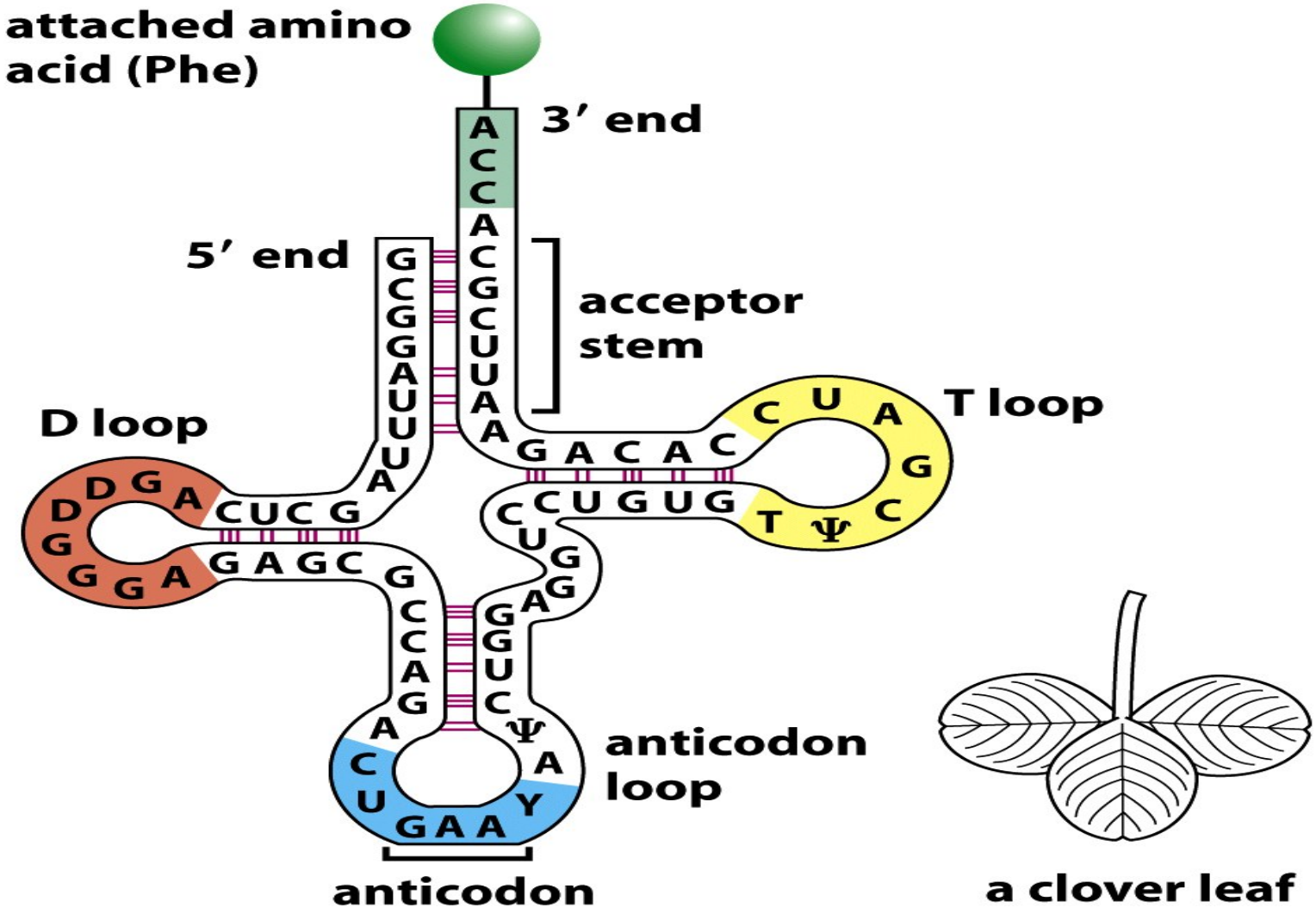
Each kind of tRNA has a sequence of 3 unpaired nucleotides - the anticodon - which can bind to the complementary triplet of nucleotides - the codon - in an mRNA molecule.

The reading of codons in mRNA requires that the anticodons bind in the opposite direction.

```
      Anticodon:      3' AAG  5'
                Codon:     5'
UUC  3'
```

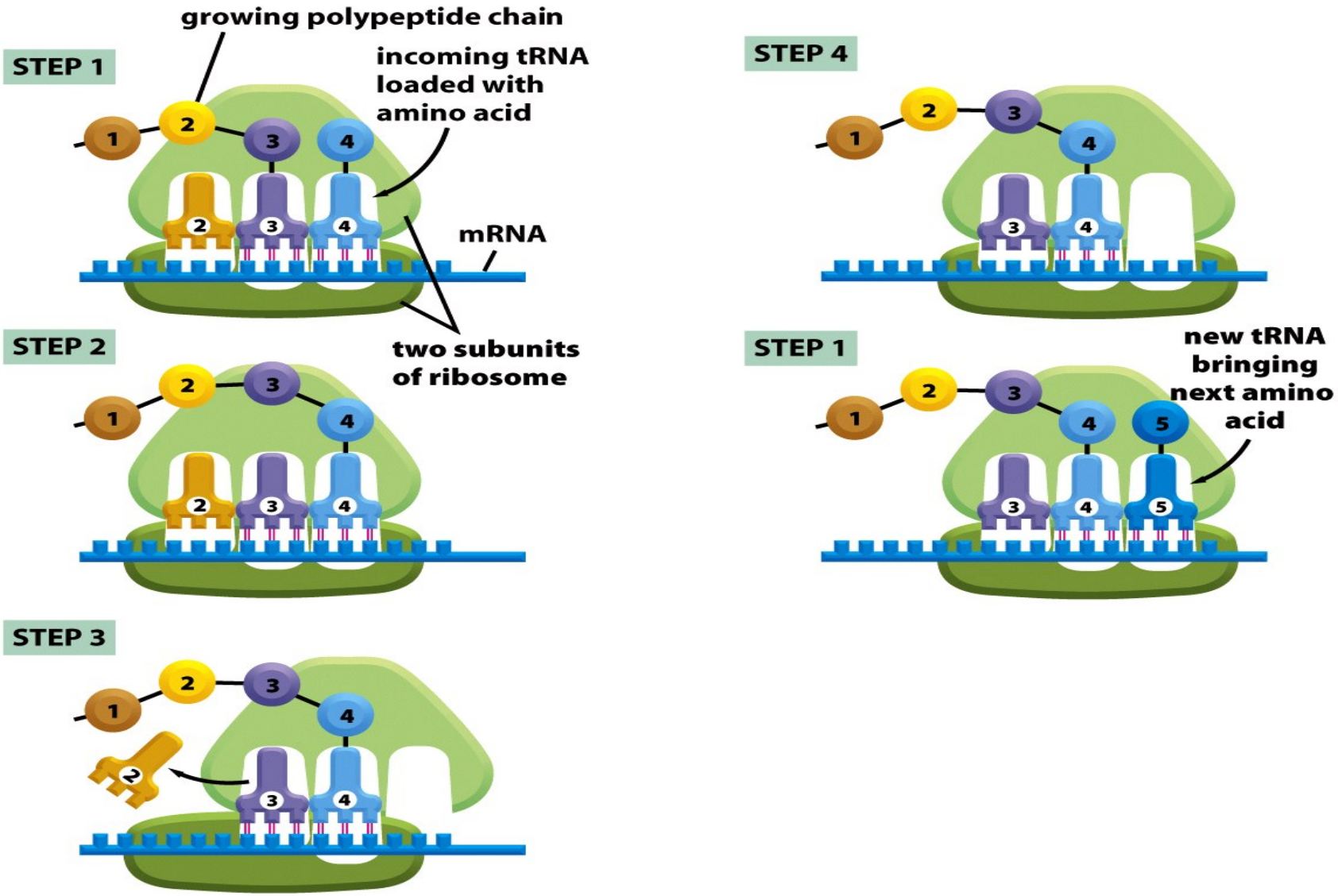


tRNA and Anticodon



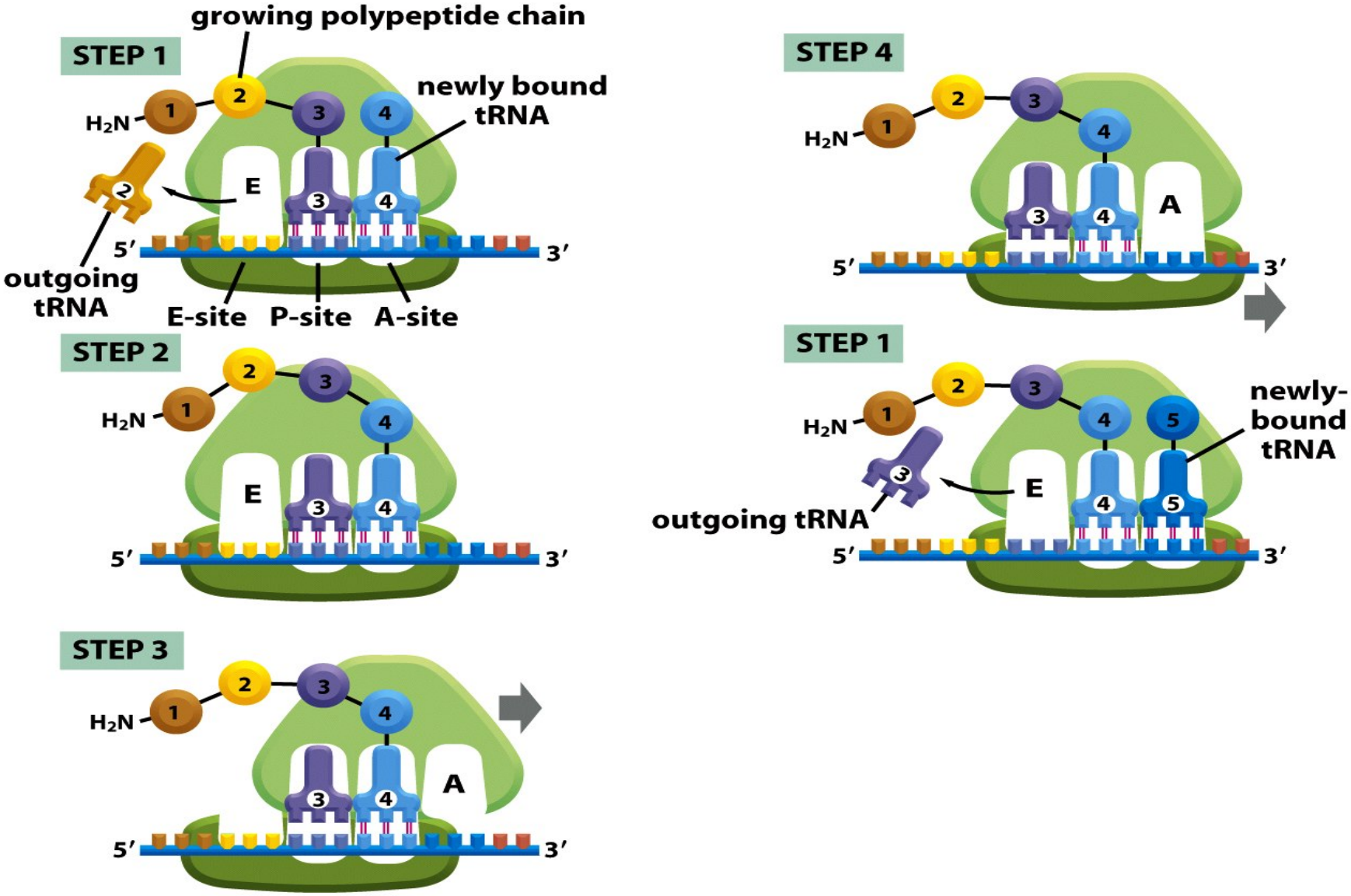
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Steps of Translation



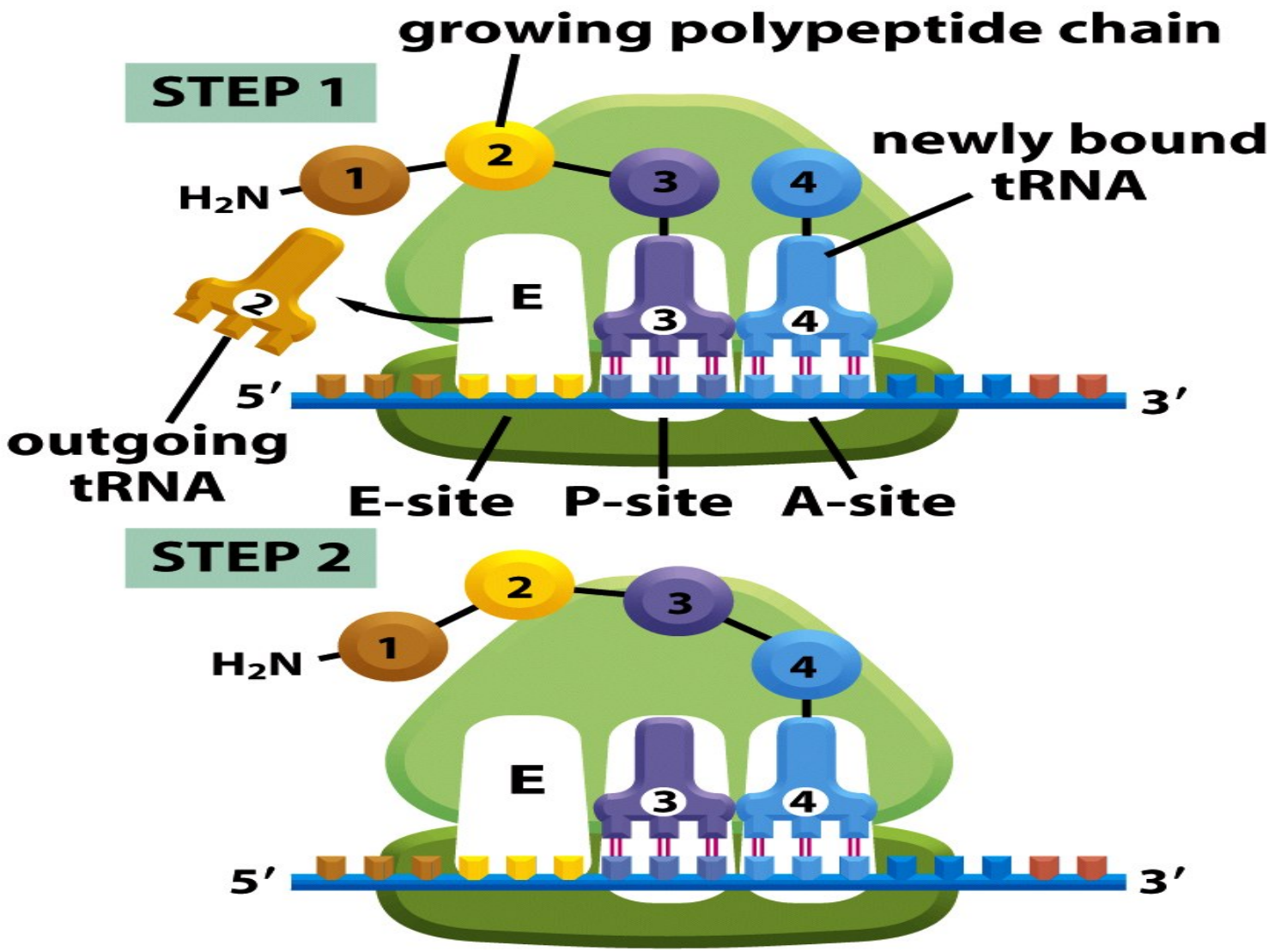
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Translation Showing Ends



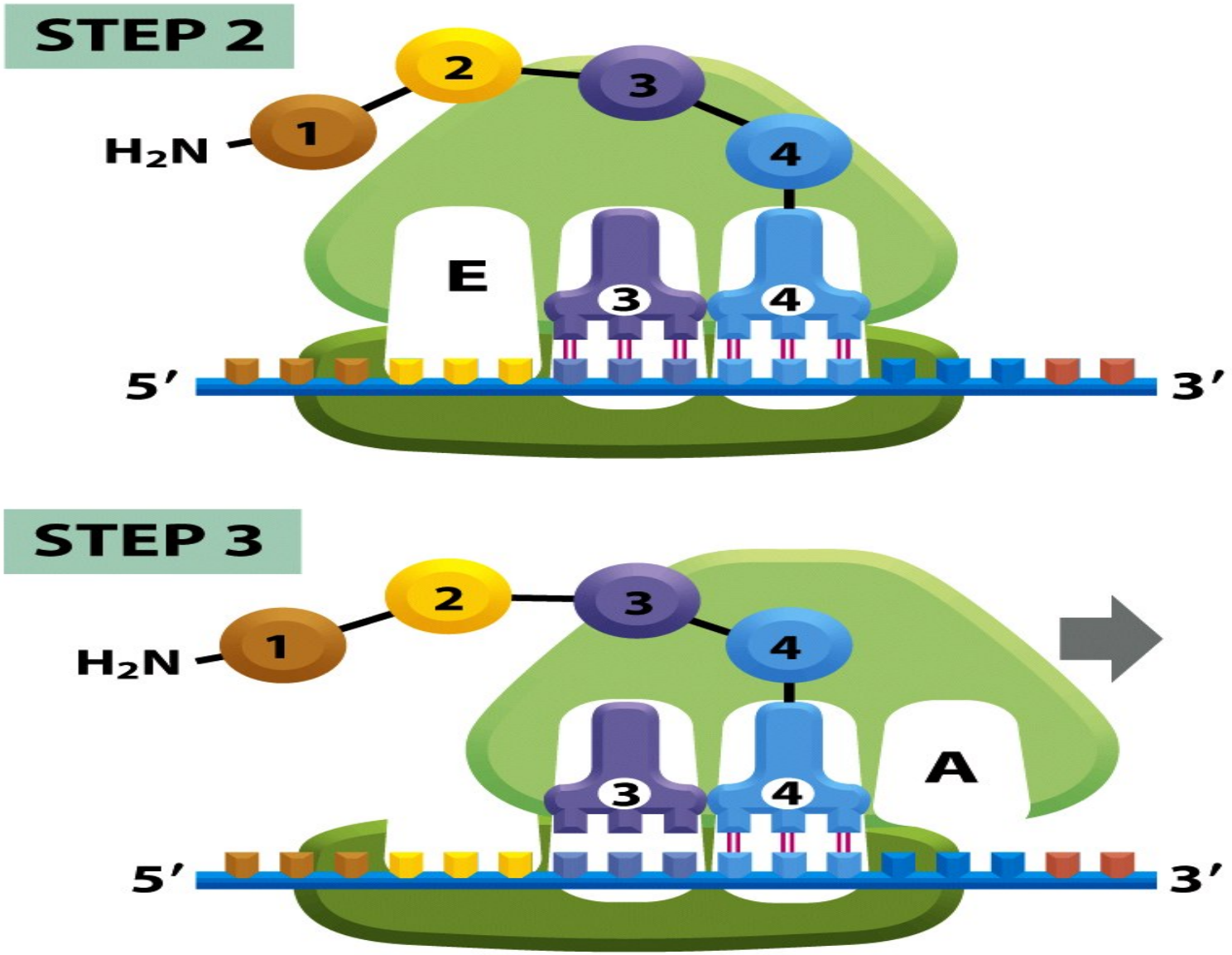
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Translation: Steps 1 and 2



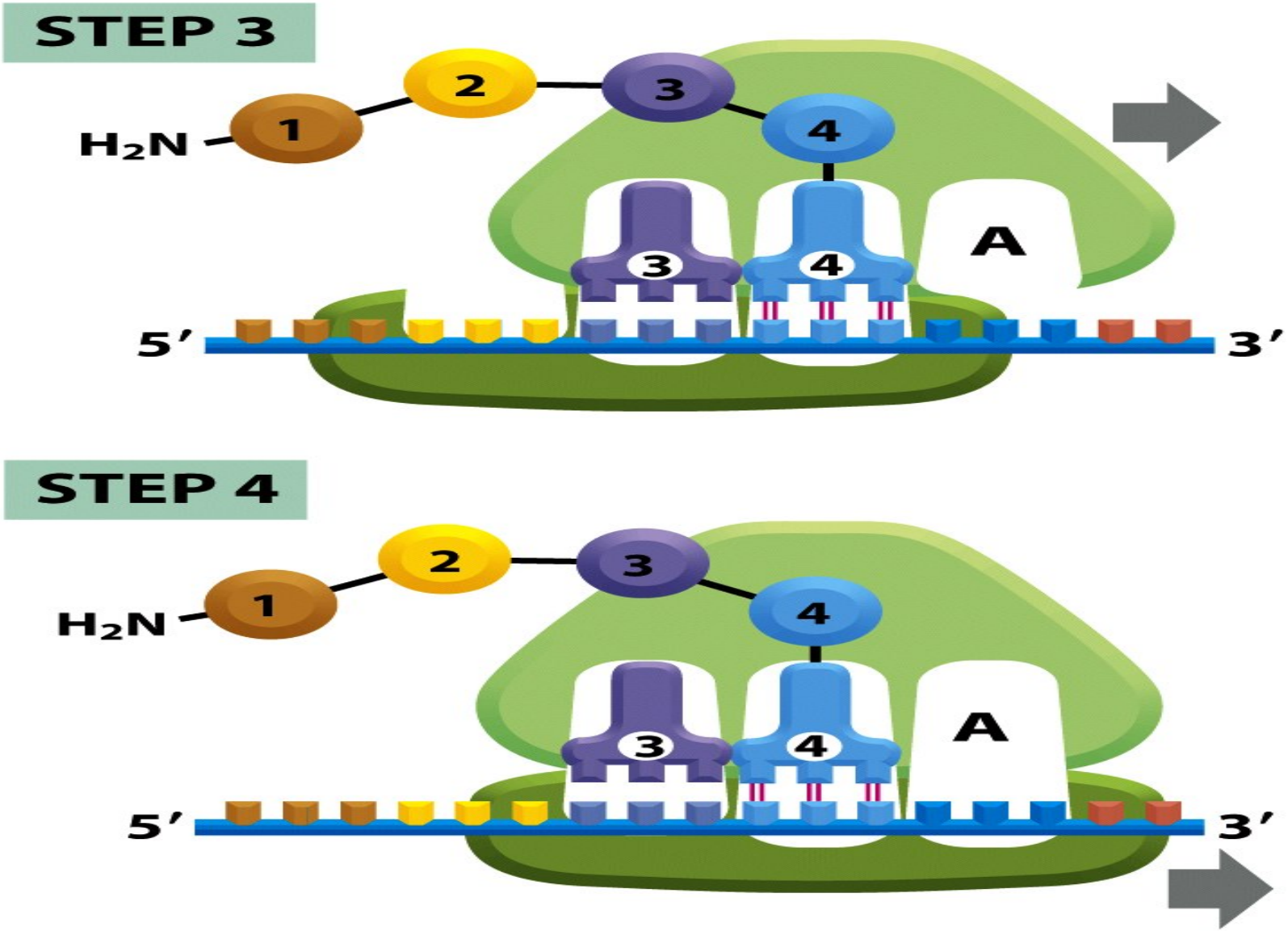
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Translation: Steps 2 and 3



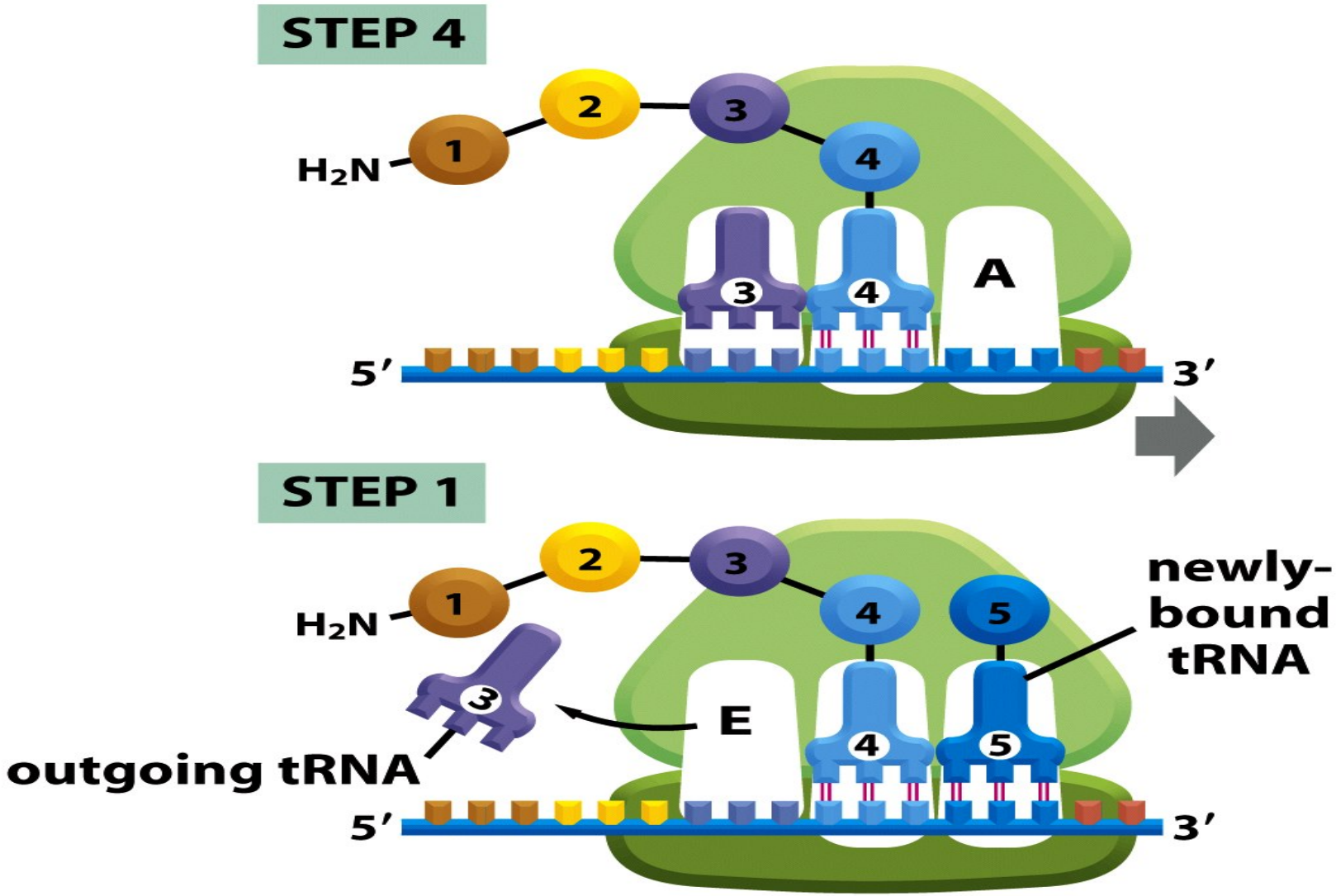
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Translation: Steps 3 and 4



Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Translation: Steps 4 and 1

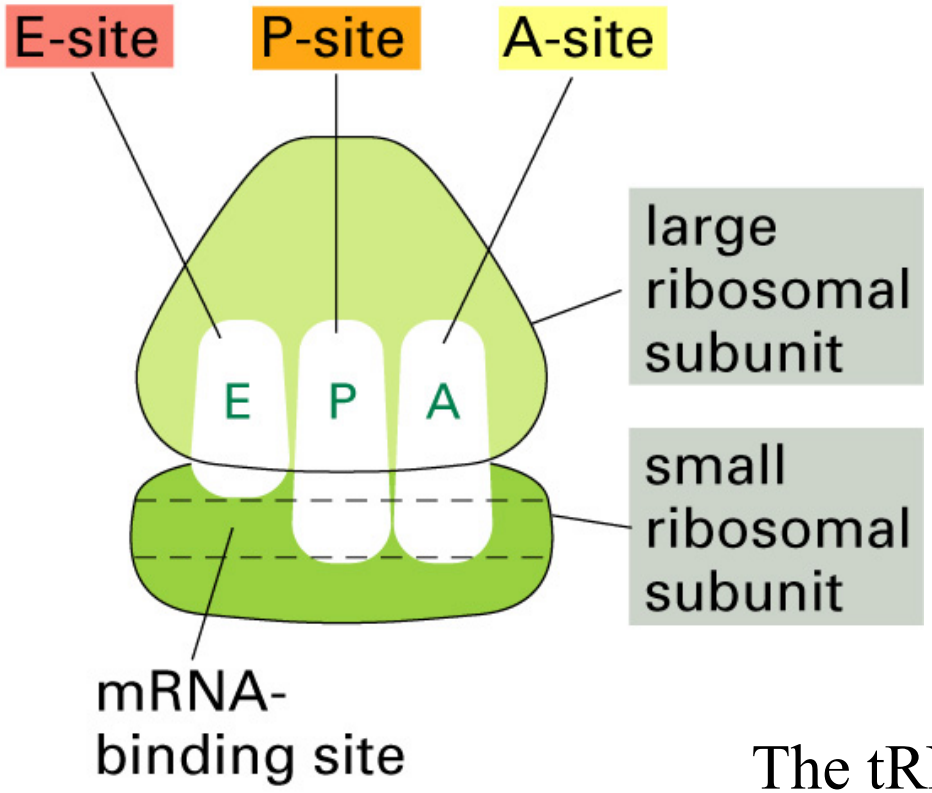


Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

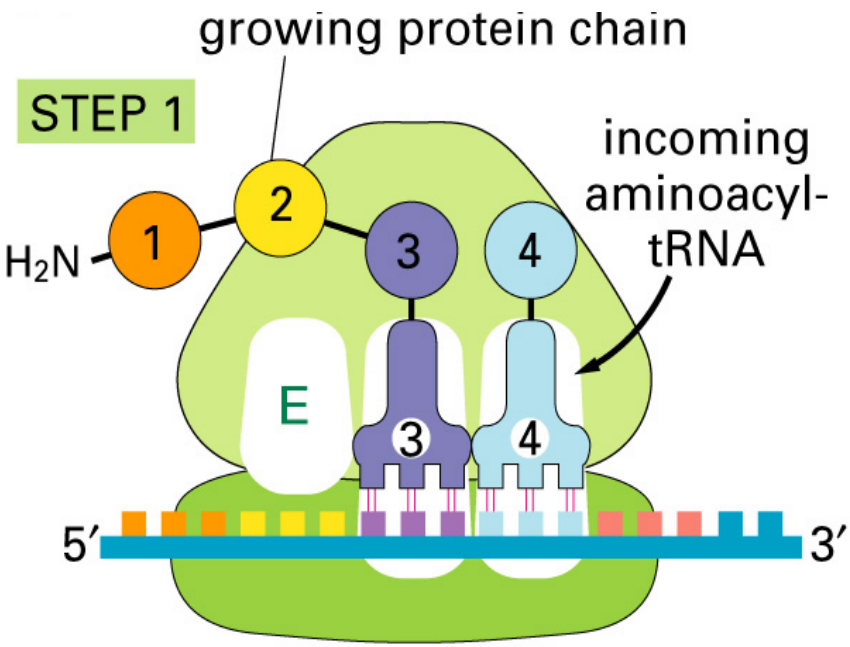
Start and Stop Codons

- The codon **AUG** serves two related functions
  - It begins most messages; that is, it signals the start of translation placing the amino acid methionine at the amino terminal of the polypeptide to be synthesized.
  - When it occurs within the message, it guides the incorporation of methionine.
- Three **codons**, UAA, UAG, and UGA, act as signals to terminate translation. They are called **STOP codons**.

Translation with Details



Binding site of ribosome for the mRNA and the three tRNA binding sites.



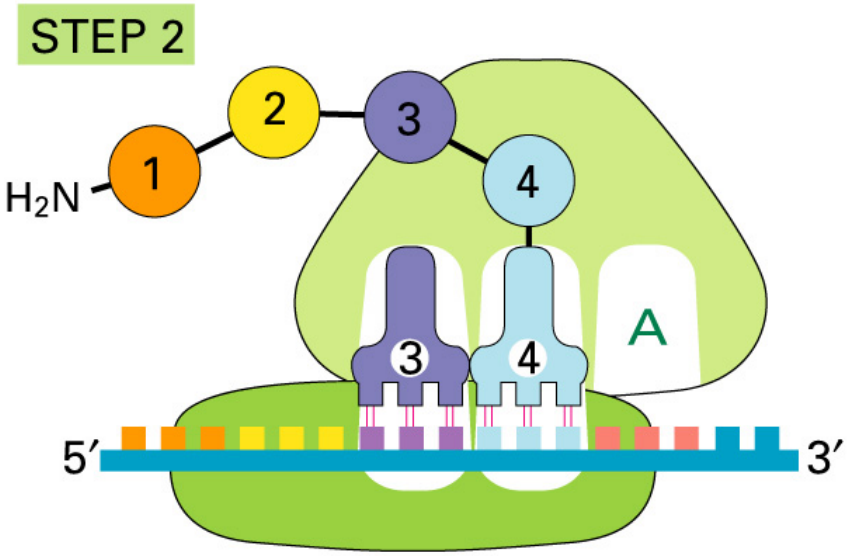
The tRNA molecules bind to the ribosome and are the physical link between the mRNA and the growing protein chain.

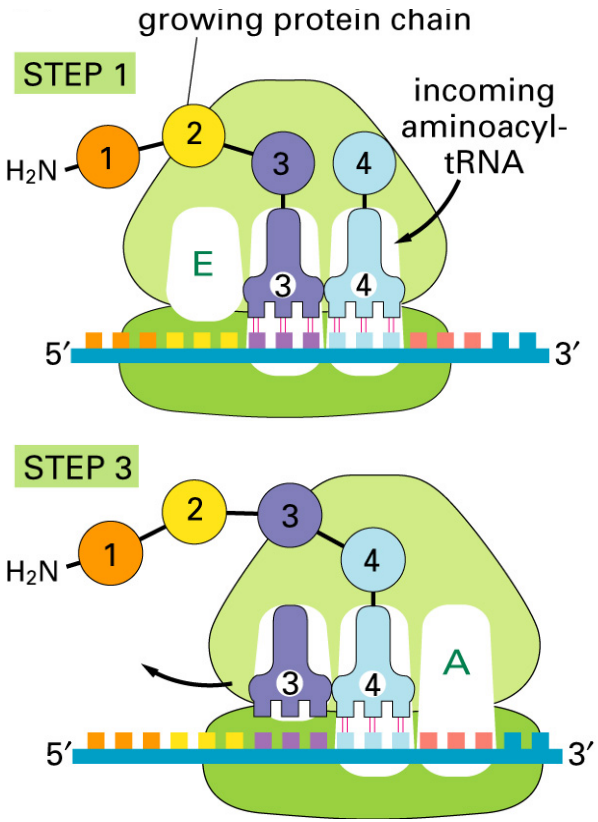
Steps of Translation: Initiation

- The small subunit of the ribosome binds to a site "upstream" of the start of the message.
- It proceeds downstream until it encounters the start codon **AUG**.
- It is then joined by the large subunit and a special initiator tRNA. The initiator tRNA binds to the P site on the ribosome.
- In eukaryotes, initiator tRNA generally carries methionine (Met).

Steps of Translation: Elongation

An aminoacyl-tRNA able to base pair with the next codon on the mRNA arrives at the A site.



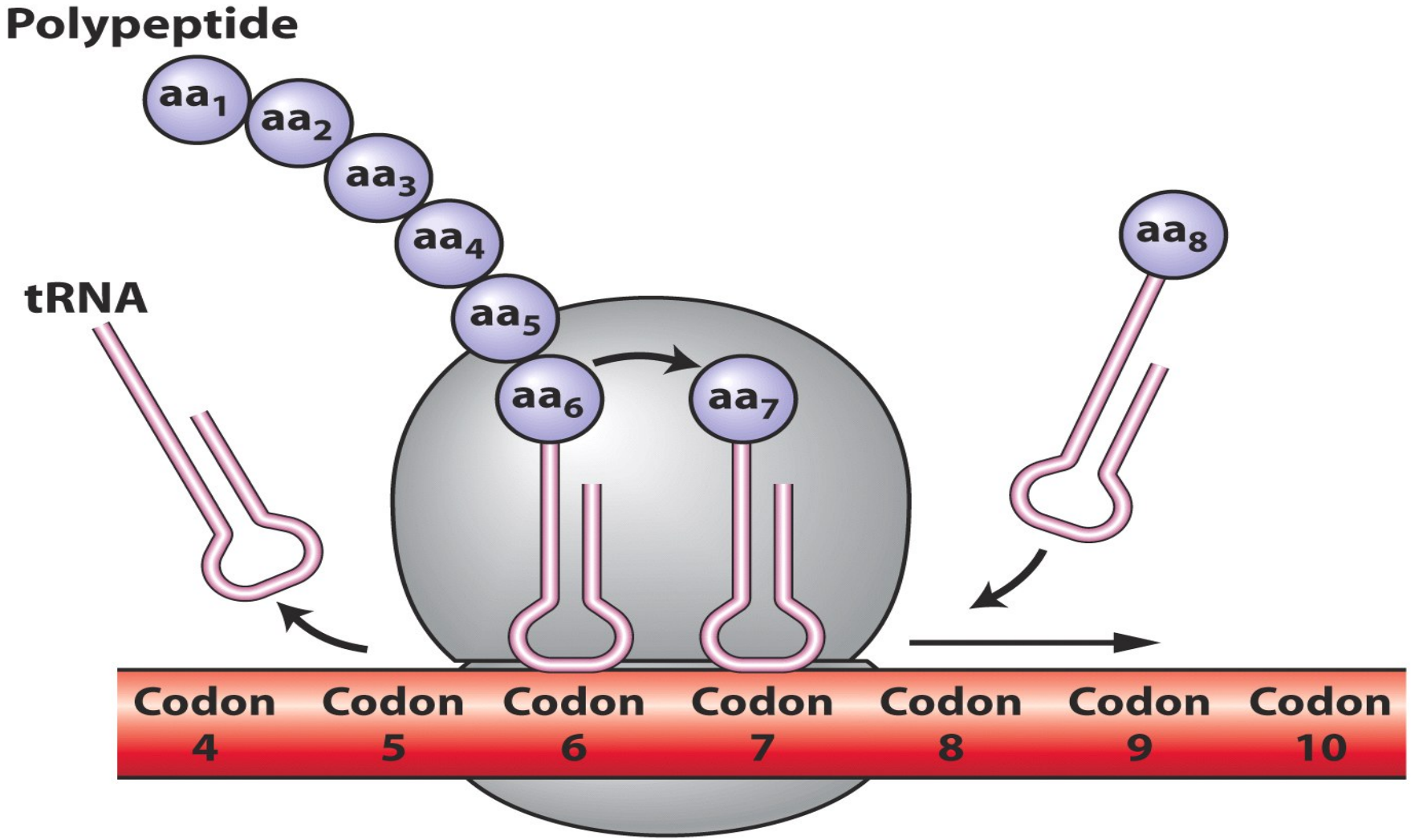


The preceding amino acid is linked to the incoming amino acid with a peptide bond.

Steps of Translation: Termination

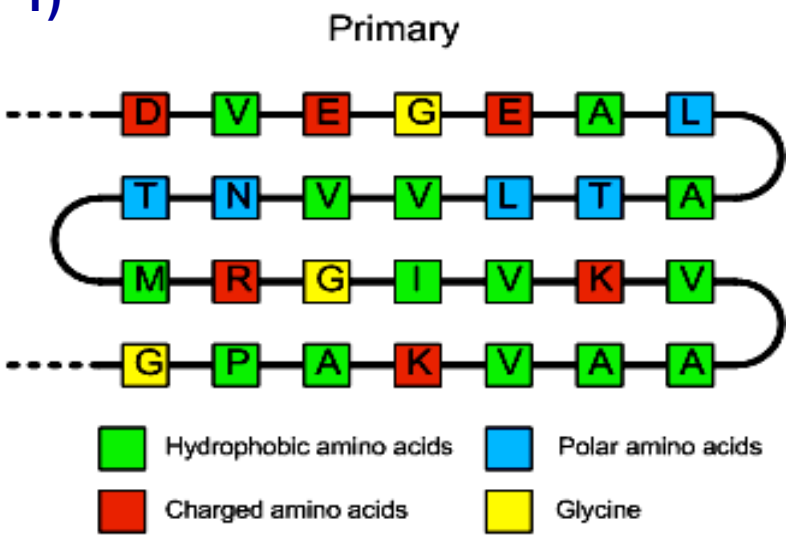
- The end of the message is marked by a STOP codon: UAA, UAG, UGA.
- No tRNA molecules have anticodons for STOP codons. Instead, protein release factor recognizes these codons when they arrive at the A site.
- Binding of this protein releases the polypeptide from the ribosome.
- The ribosome splits into its subunits, which can later be reassembled for another round of protein synthesis.

Chain of Amino Acids

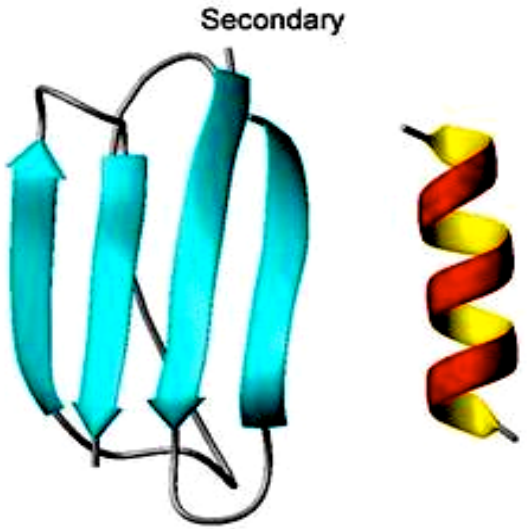


Four Structures of Proteins

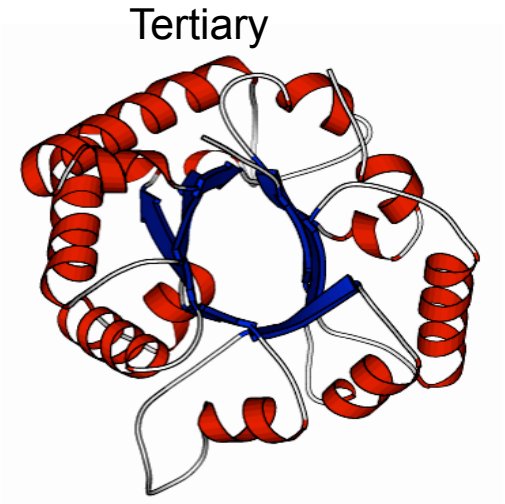
1)



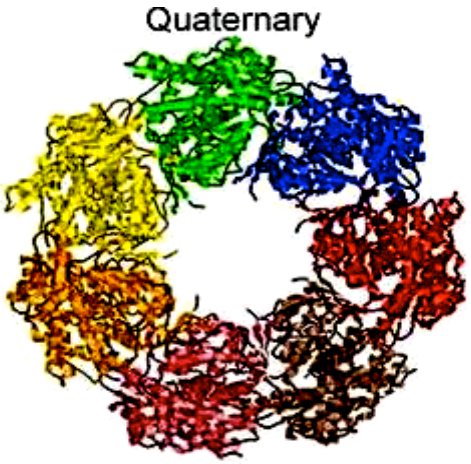
2)



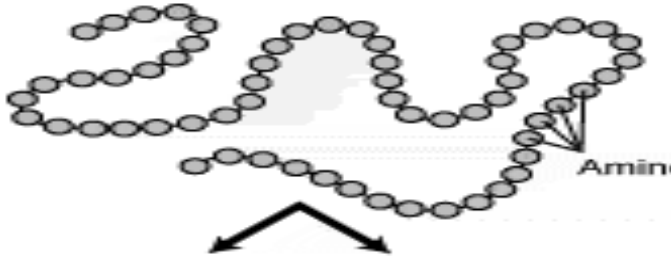
3)



4)



Four Protein Structures



Primary protein structure
is sequence of a chain of amino acids

Amino Acids

Pleated sheet

Alpha helix

Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds

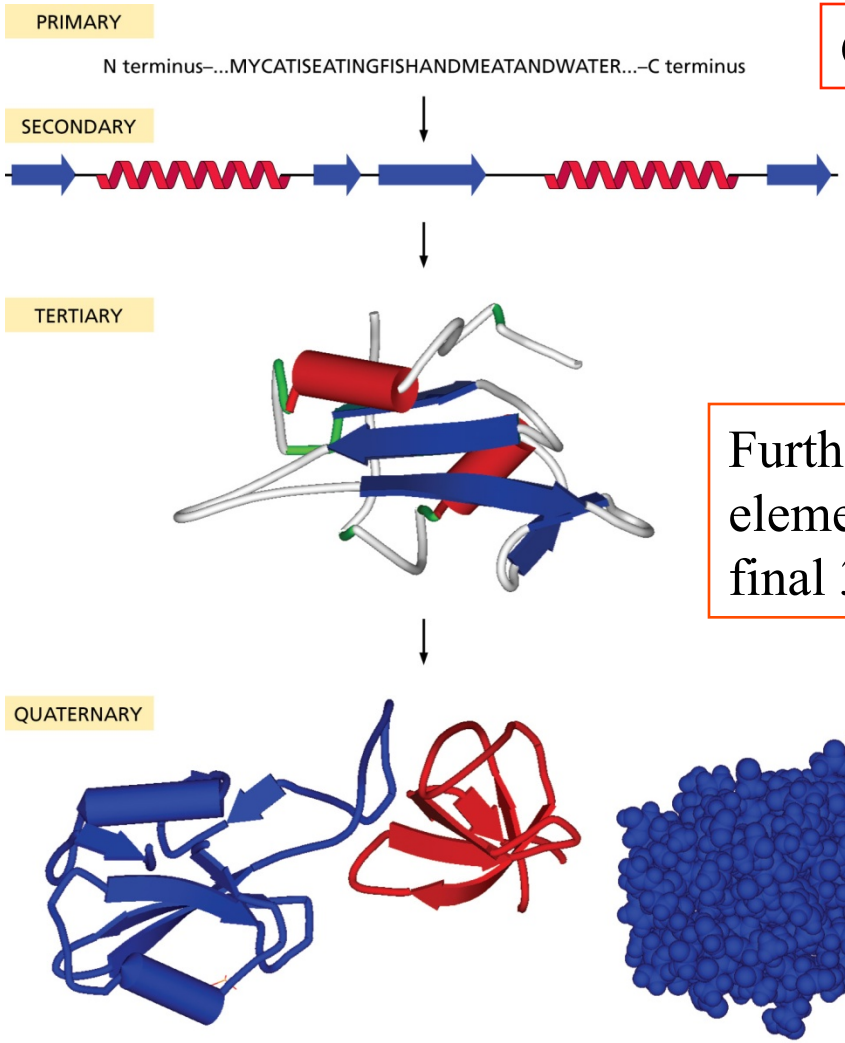
Pleated sheet

Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.

Alpha helix

Quaternary protein structure
is a protein consisting of more than one amino acid chain.

Four Levels of Protein Structure



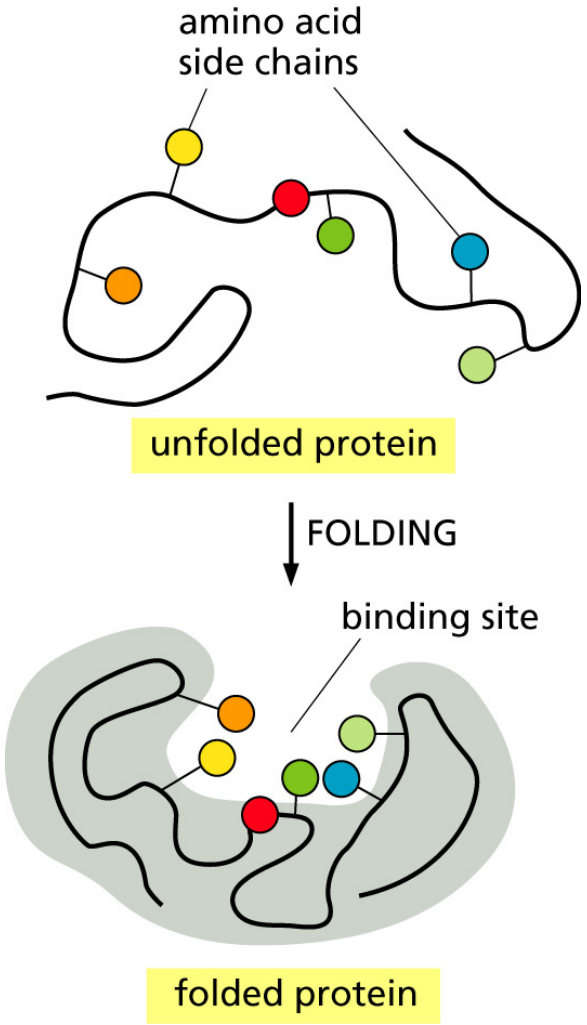
Chain of amino acids

First level of protein folding
Parts of the chain fold to form genetic structures found in all proteins

Further folding and packing together of the elements of secondary structure to produce the final 3-D conformation unique to the protein

Multisubunit protein formed of more than one protein chain.
Each protein chain is called a subunit

Folded Protein



Distant residues (in the primary structure of a protein) can come close in the folded structure.

When a polypeptide chain (primary structure) folds into a tertiary structure, residues that are far apart from each other in the sequence can come close together to form a binding site.

Understanding Bioinformatics by Zvelebil and Baum, 2008

Protein Function

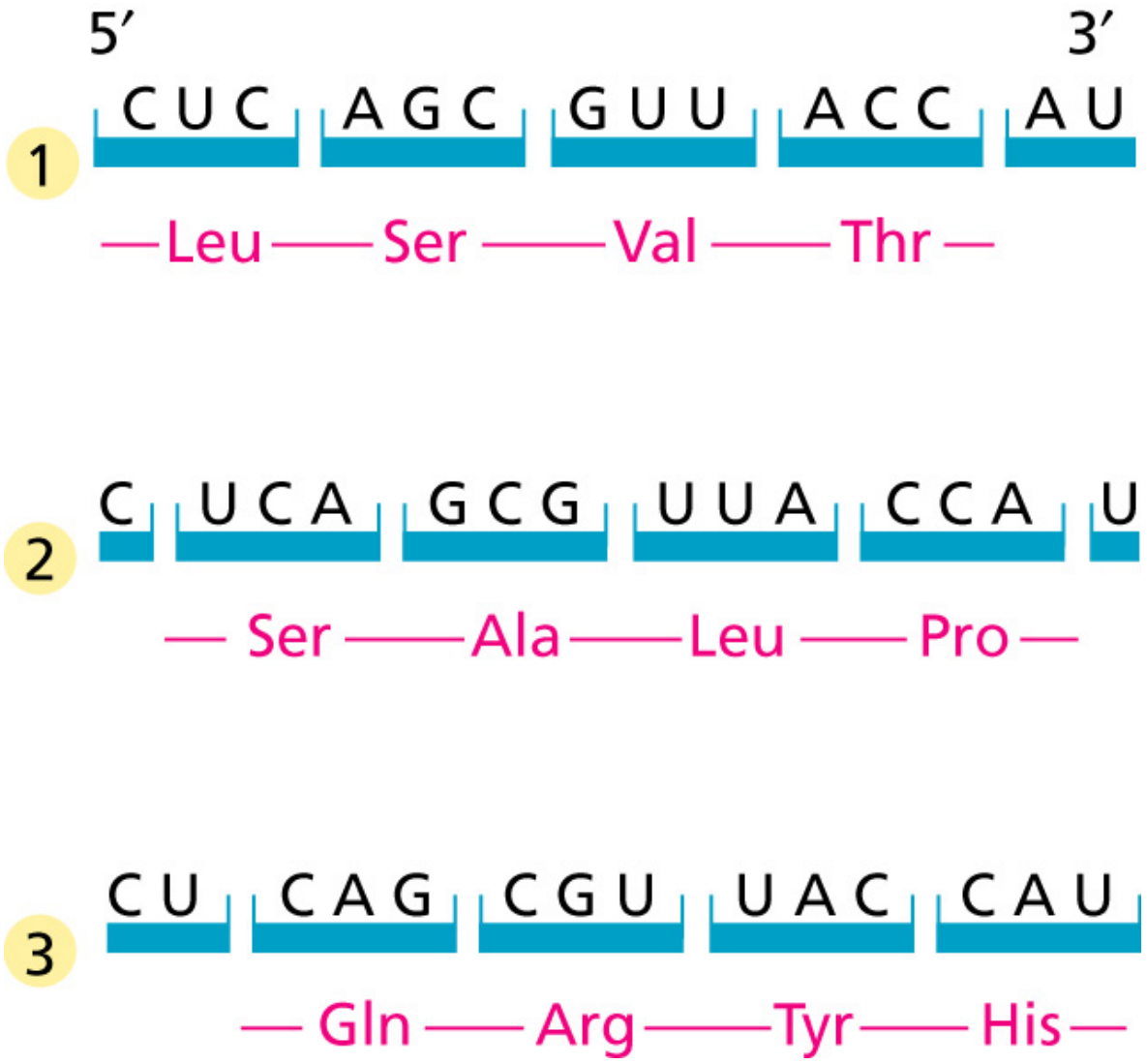
- The amino acid sequence of the protein is critical to its function in 2 ways:
  - The protein can only **interact** with other molecules if it has the correct amino acids with the correct side chains to **bind** those molecules.
  - The sequence of amino acids also determines how the protein can **fold**.

    The ability of the protein to **interact** with specific molecules is dependent on the **folding** of its amino acid chain into a very specific three-dimensional shape.

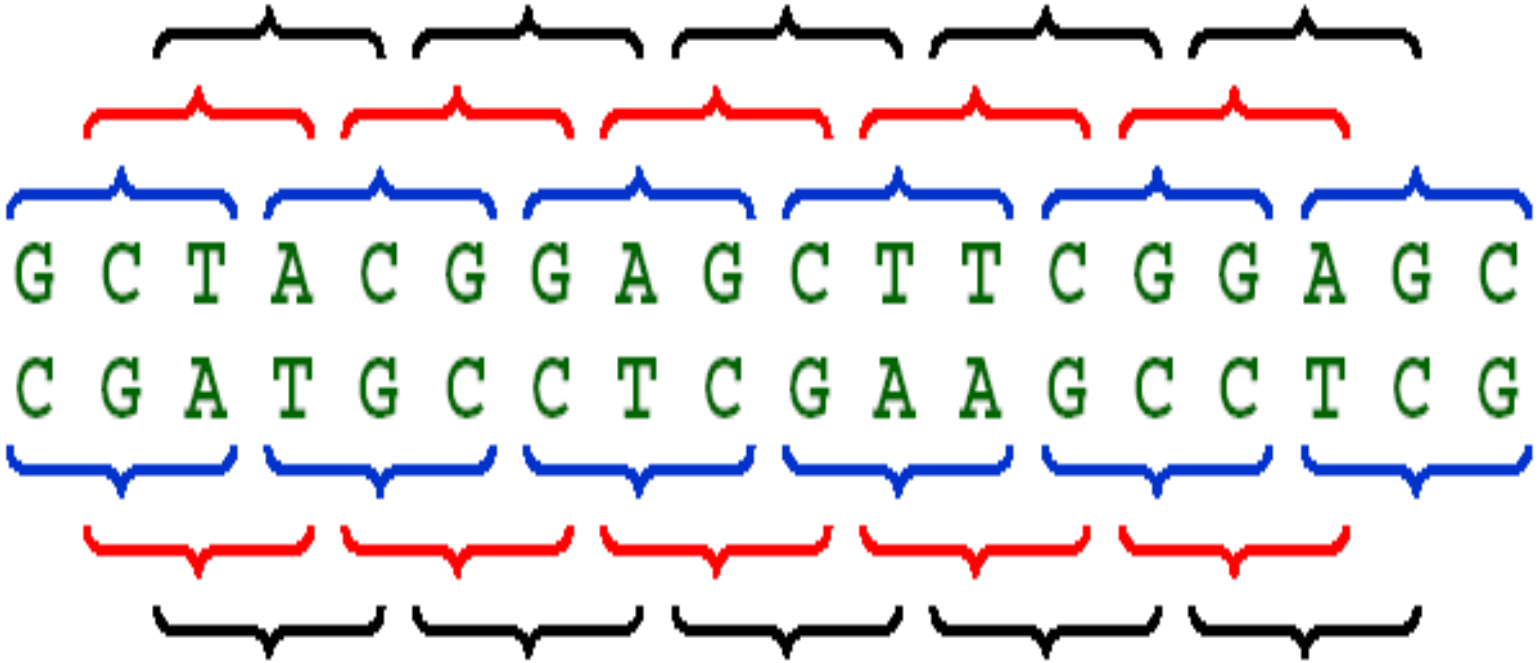
Structure and Function Relationship

- Knowing the relationship between a protein's structure and its function provides a greater understanding of how modifying the structure will affect the function.
- As the vast majority of currently marketed pharmaceuticals act by interacting with proteins, structure-function studies are vital to the design of new drugs, and bioinformatics has an important role in speeding up this process and enabling computer modeling of these interactions.

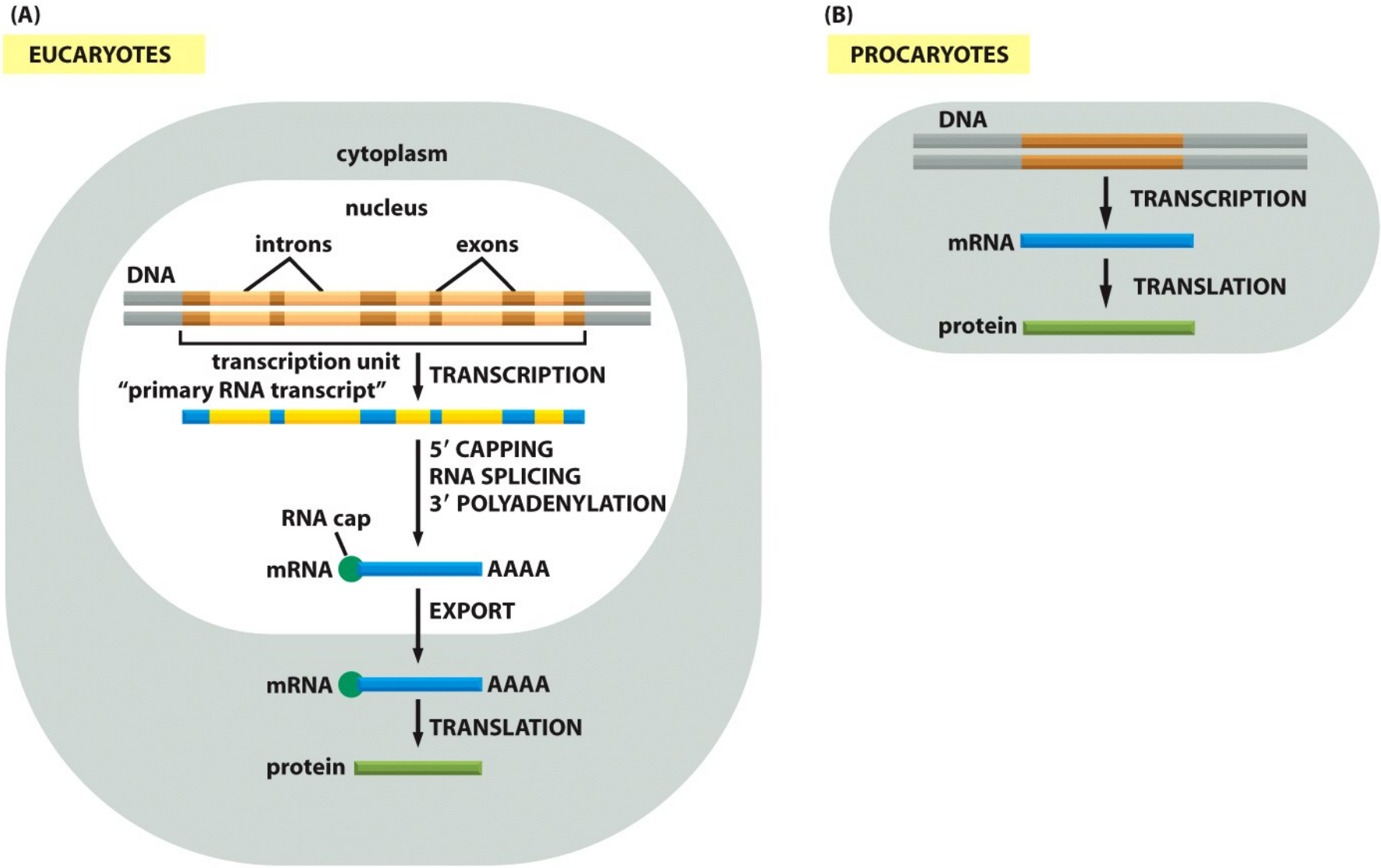
3 Reading Frames of mRNA



Six Reading Frames

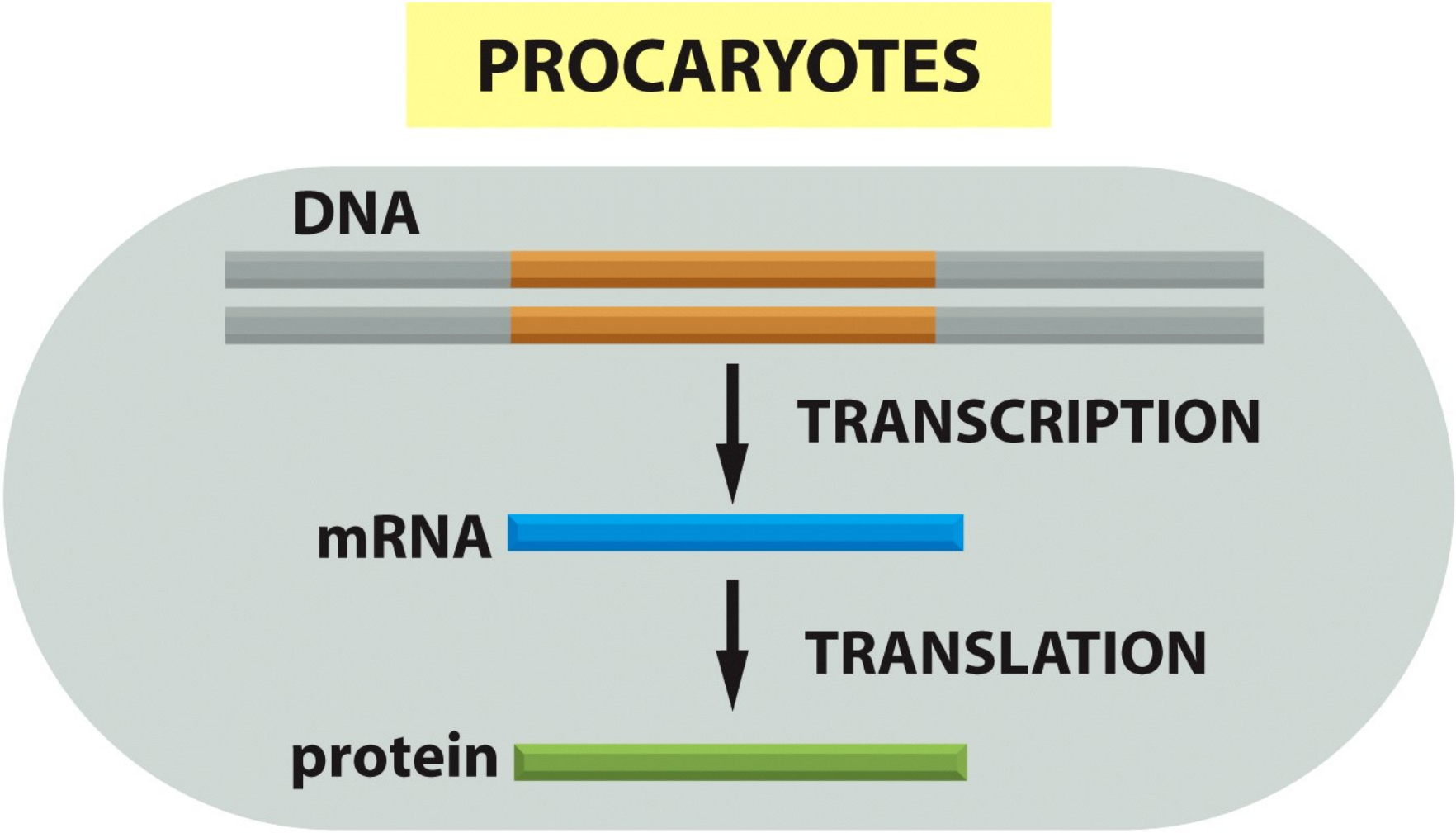


Dogma Revisited



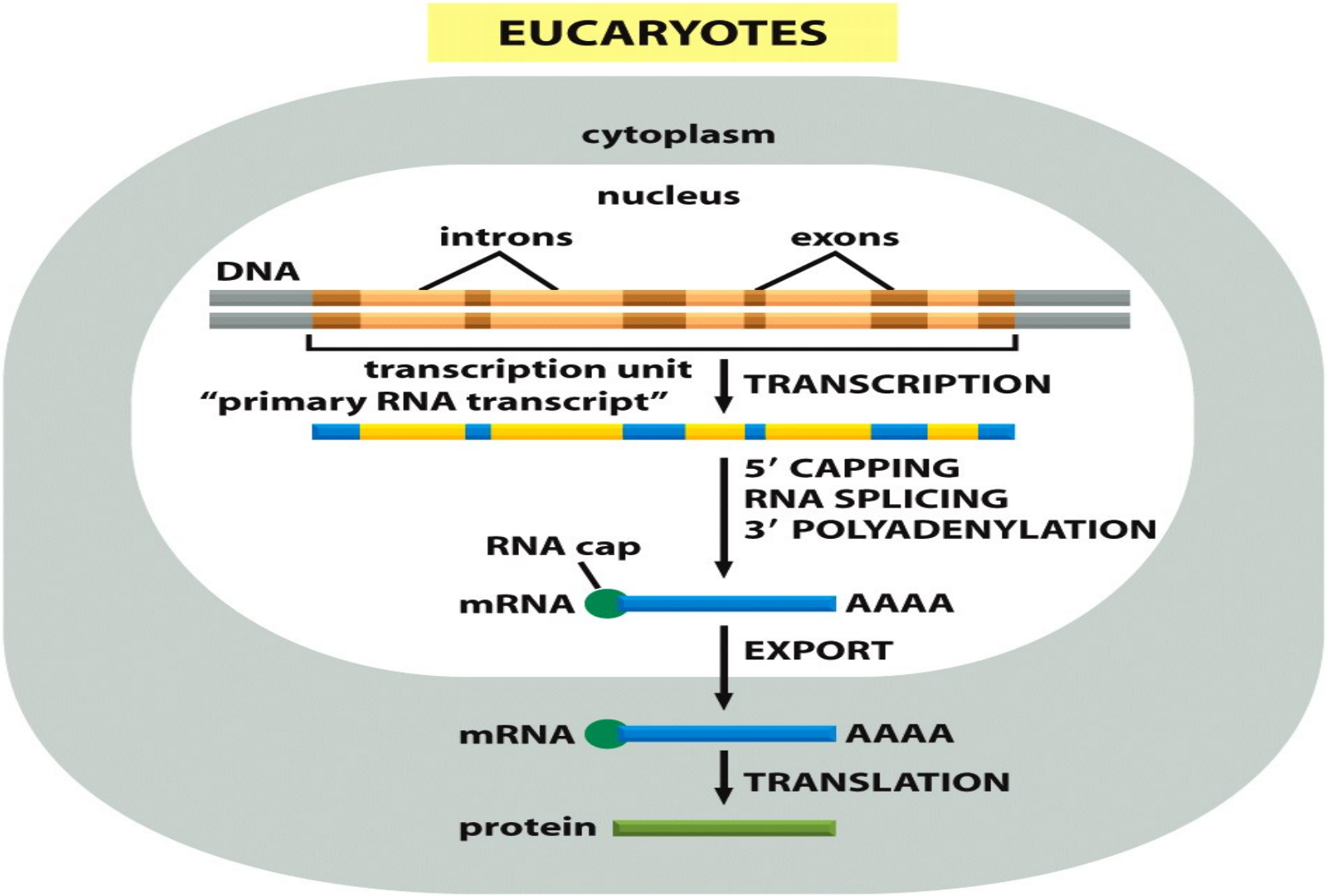
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Dogma Revisited: Prokaryotes



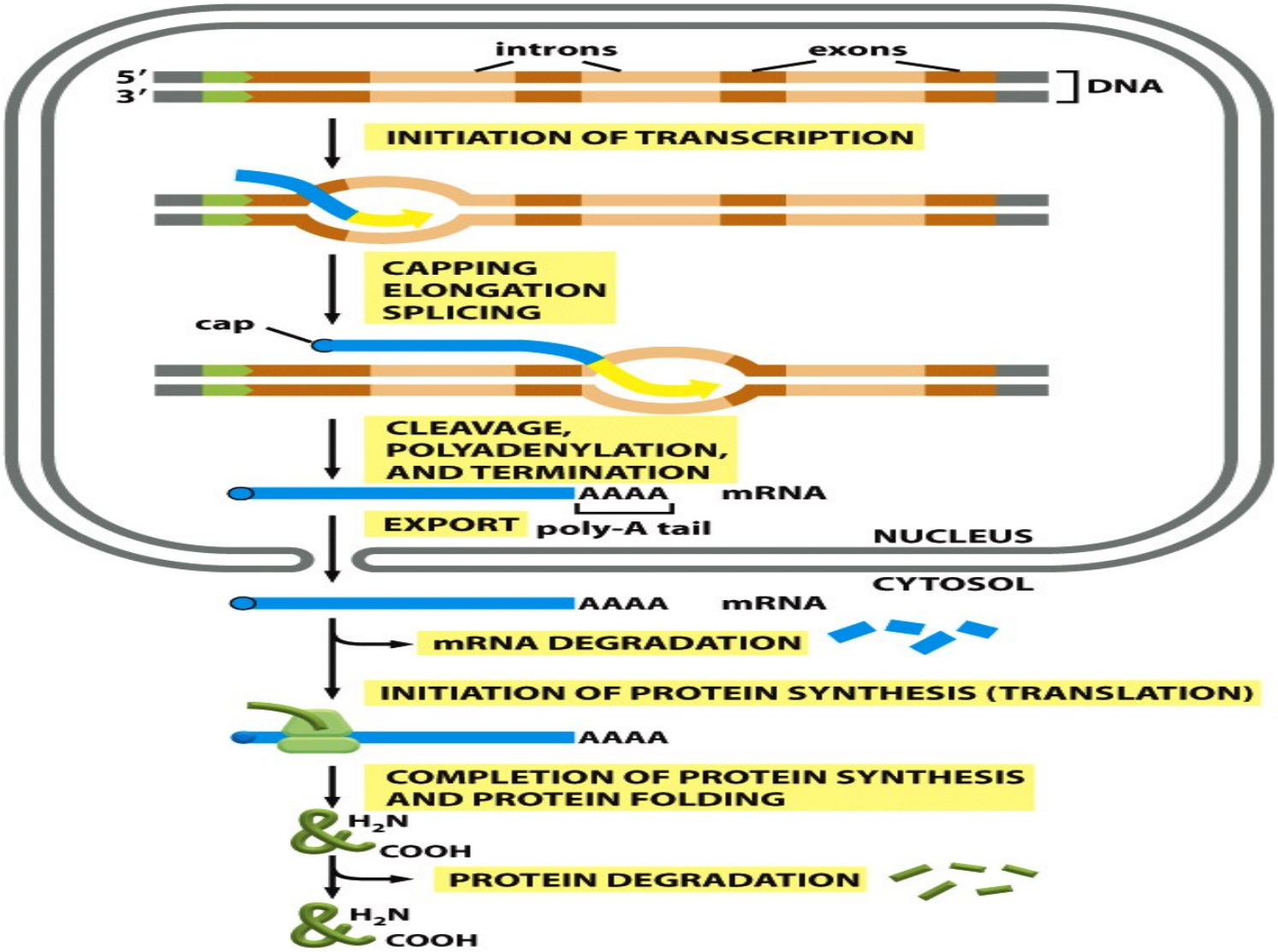
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Dogma Revisited: Eukaryotes



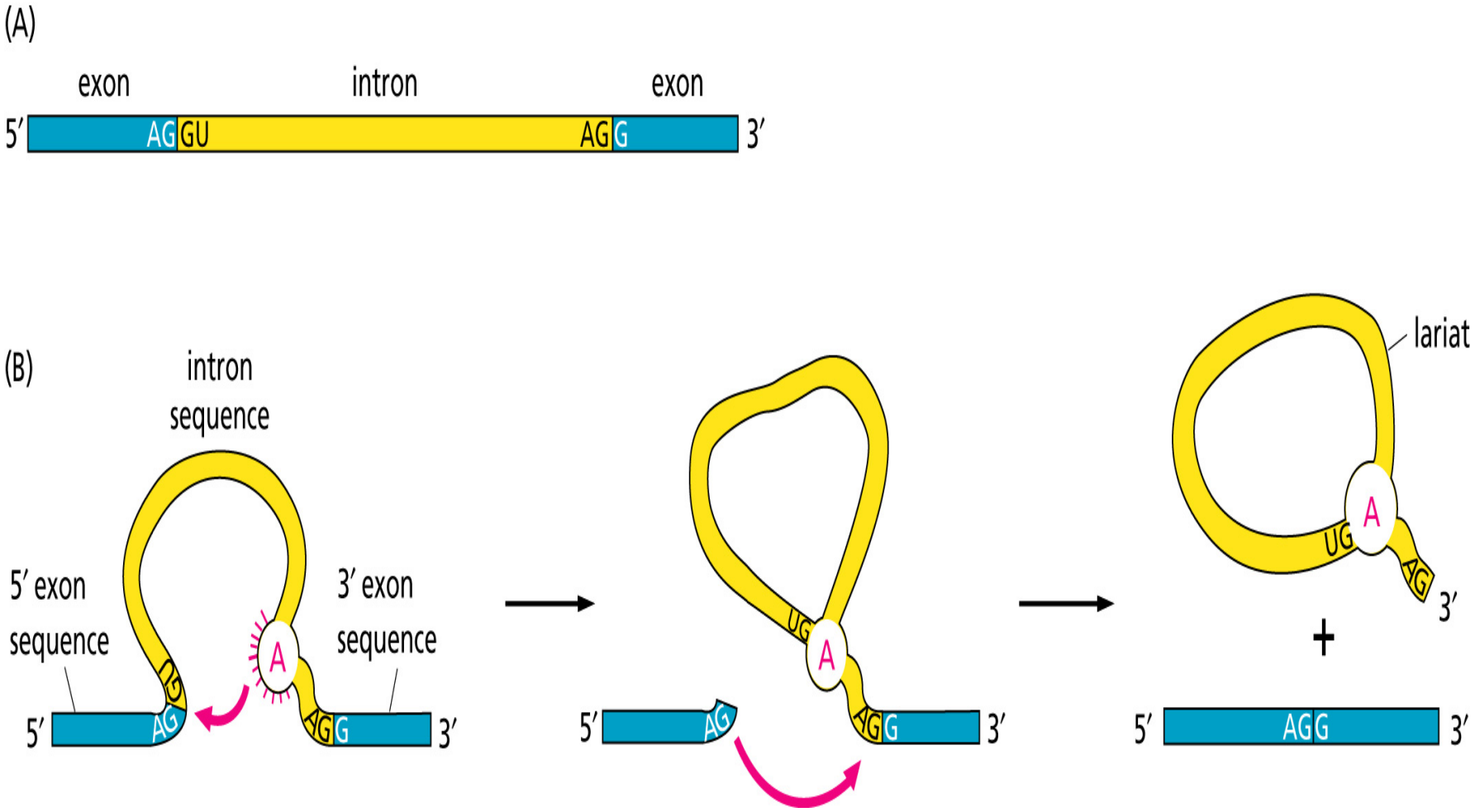
Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

Dogma Re-revisited

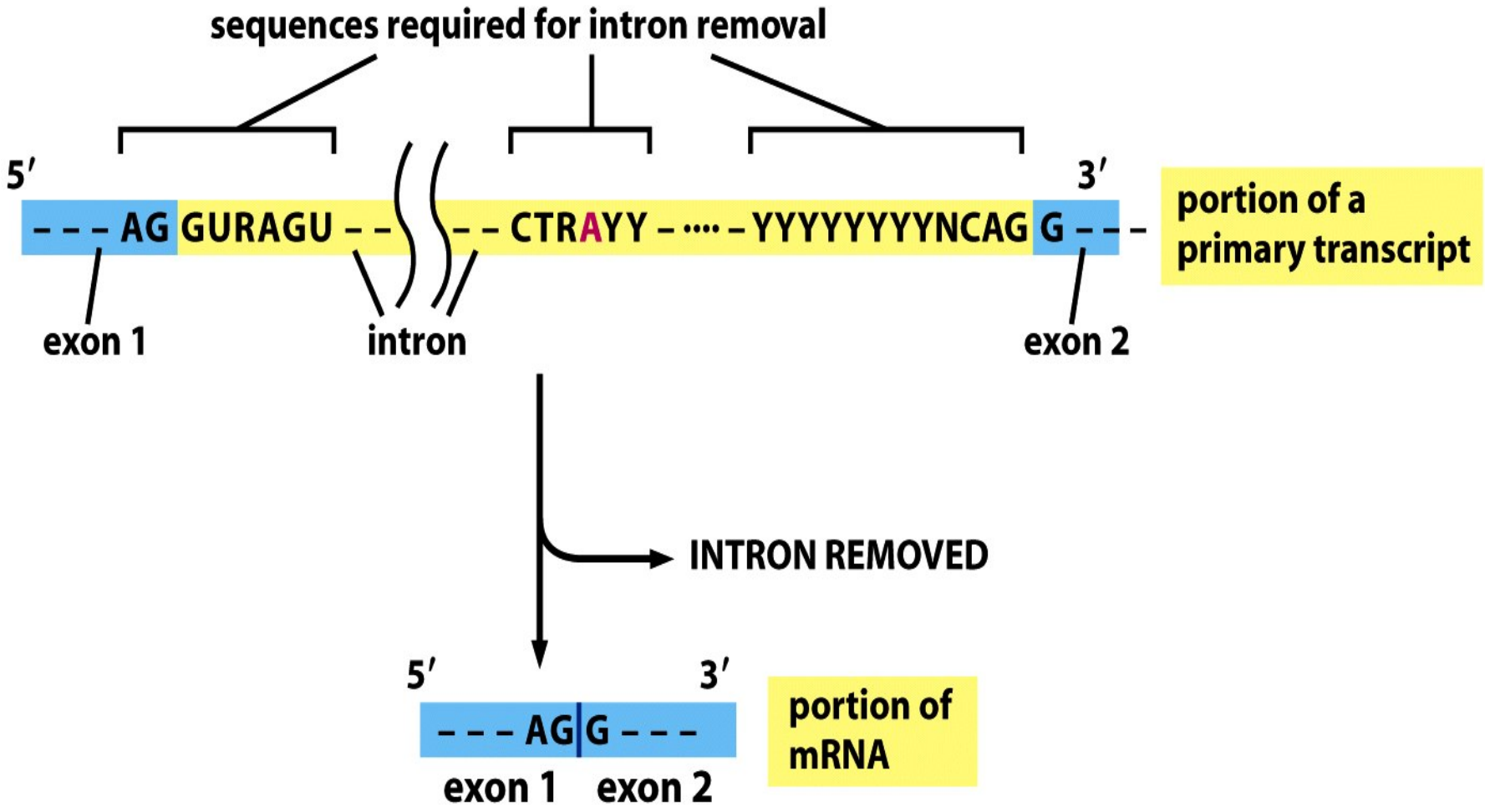


Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)

The Splicing of an Intron



Splicing in Eukaryotes



Molecular Biology of the Cell, Fifth Edition (© Garland Science 2008)