Bioinformatics

Lecture 01

Course Code: CSE 469

Credit: 3.0

Total Course Hour: 36

Reference Books

- **Lehninger Principles of Biochemistry** by Albert L. Lehninger, David L. Nelson, and Michael M. Cox
- **An Introduction to Genetic Analysis** by Anthony J. F. Griffiths, Susan R. Wessler, John Doebley and Sean B. Carroll

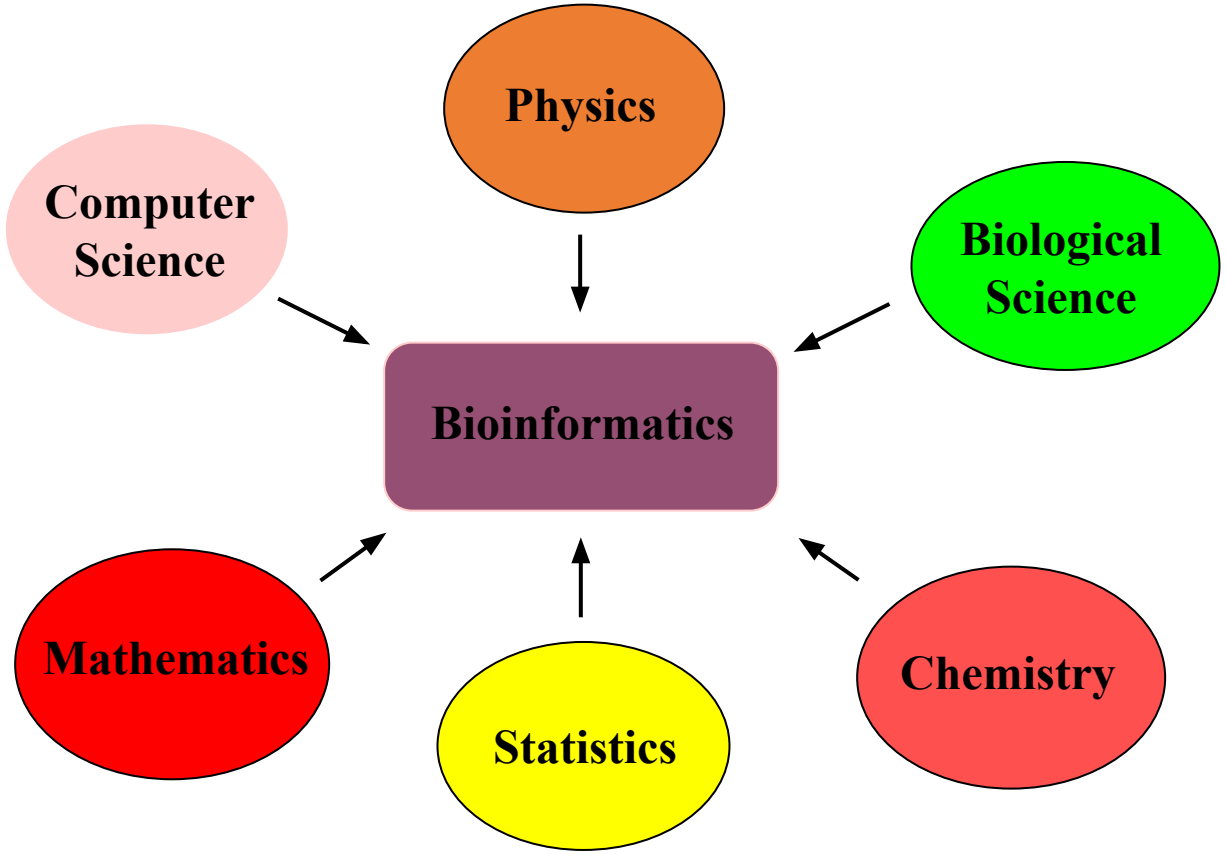
- **Bioinformatics Algorithms: An Active Learning Approach Vol. 1 & 2 (2nd Edition)** by Phillip Compeau, Pavel Pevzner
- **An Introduction to Bioinformatics Algorithms** by Neil C. Jones and Pavel A. Pevzner.

What is Bioinformatics?

- Bioinformatics is defined as an academic field that seeks to create and advance algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of **biological data**.
- Before the era of bioinformatics, only two ways of performing biological experiments were available: within a living organism (so-called ***in vivo***) or in an artificial environment (so-called ***in vitro***, from the Latin in glass). Taking the analogy further, we can say that bioinformatics is in fact ***in silico*** biology, from the silicon chips on which microprocessors are built.

Bioinformatics: Integration of several fields

As an **interdisciplinary field** of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data.



Biological information

Central Dogma

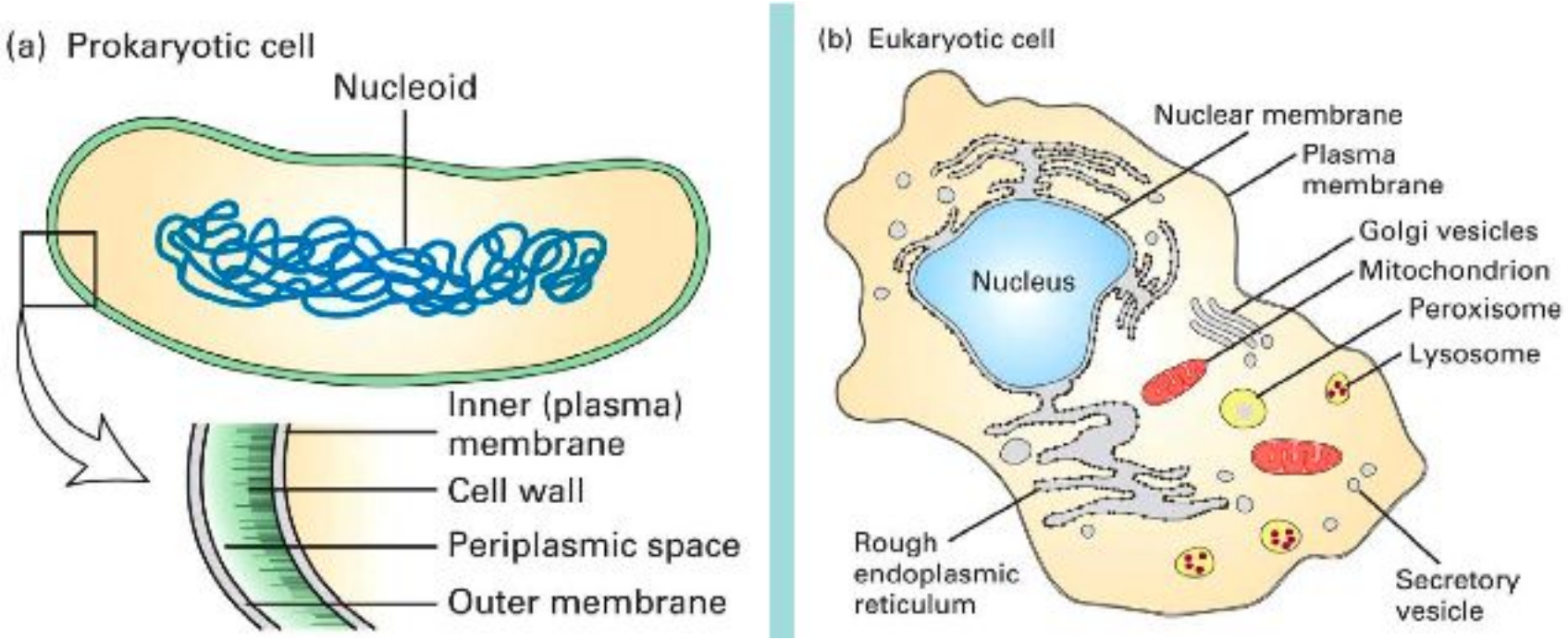
- DNA
- RNA
- Protein

Cell

- Origin of life on Earth about 3.5 billion years ago
- Fundamental working units of every living system.
- A cell is the smallest structural unit of an organism that is capable of functioning independently.
- Every organism is composed of one of two radically different types of cells: **Prokaryotic** cells or **Eukaryotic** cells.
- Prokaryotes and Eukaryotes have all evolved from the same primitive cell 3.5 billion years ago.
- Chemical composition—by weight:
  - 70% water
  - 7% small molecules: salts, lipids, amino acids, nucleotides
  - 23% macromolecules: proteins, polysaccharides, lipids

Cell Structure

- Most cells have unique properties and become a specific type of cell to make up a part of the organism. For example, human skin cells remain skin cells throughout their life cycle. Stem cells are the only type of cells that have the ability to turn into any other type of cell.

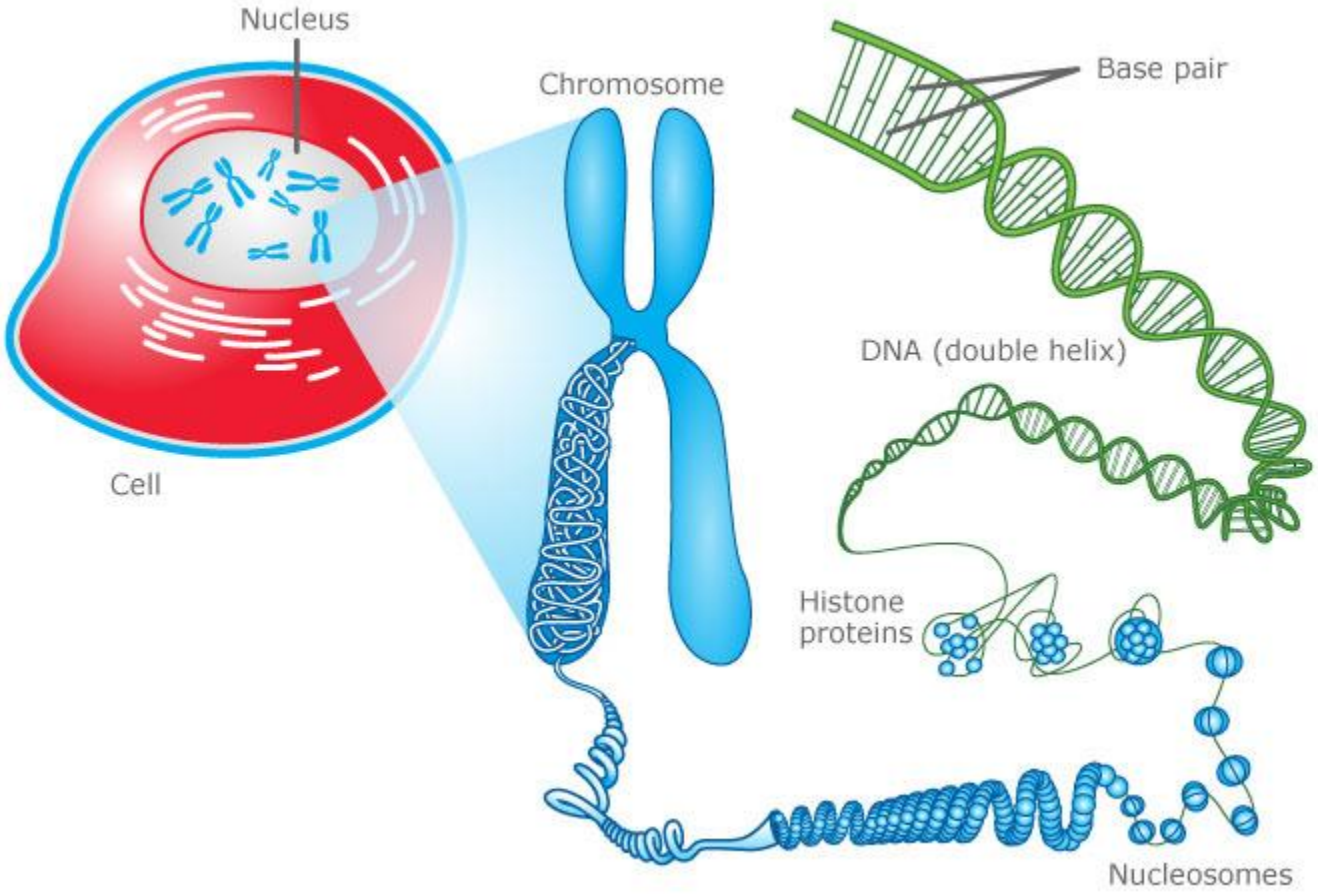


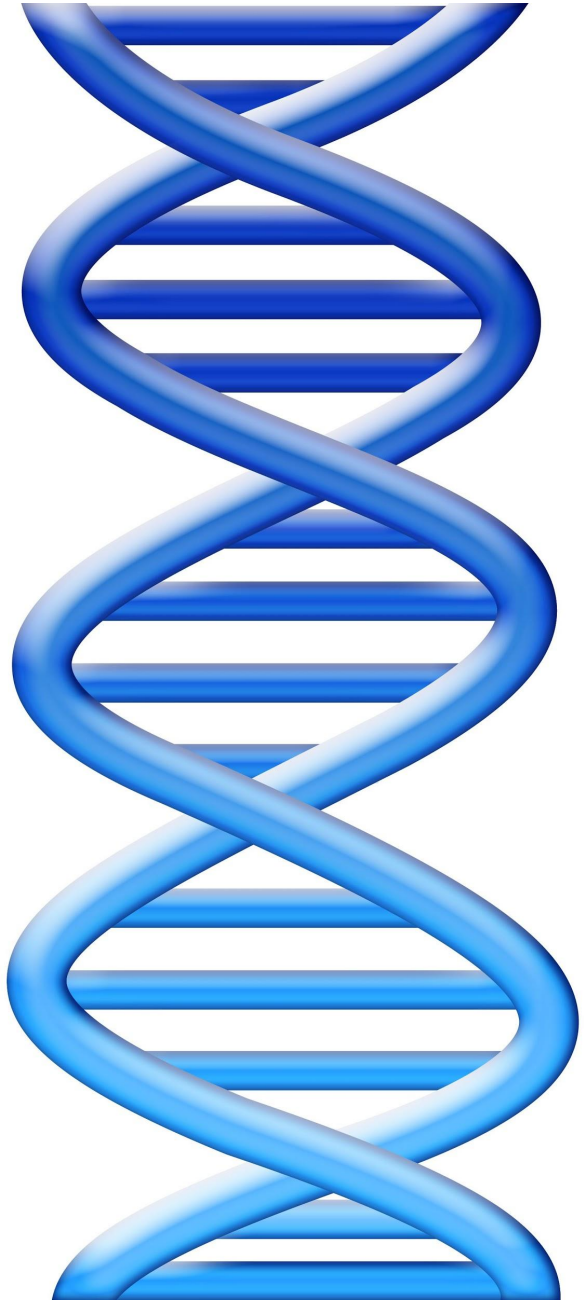
Prokaryote Vs Eukaryote

	Prokaryotes	Eukaryotes
Cell	Single	Single or Multi
Size	Generally small (1–10 µm)	Generally large (5–100 µm)
Membrane-bounded organelles	Absent	Present
DNA	Circular Chromosome	Linear Chromosomes with histone
Example	Bacteria	Plants, Animals

Chromosome

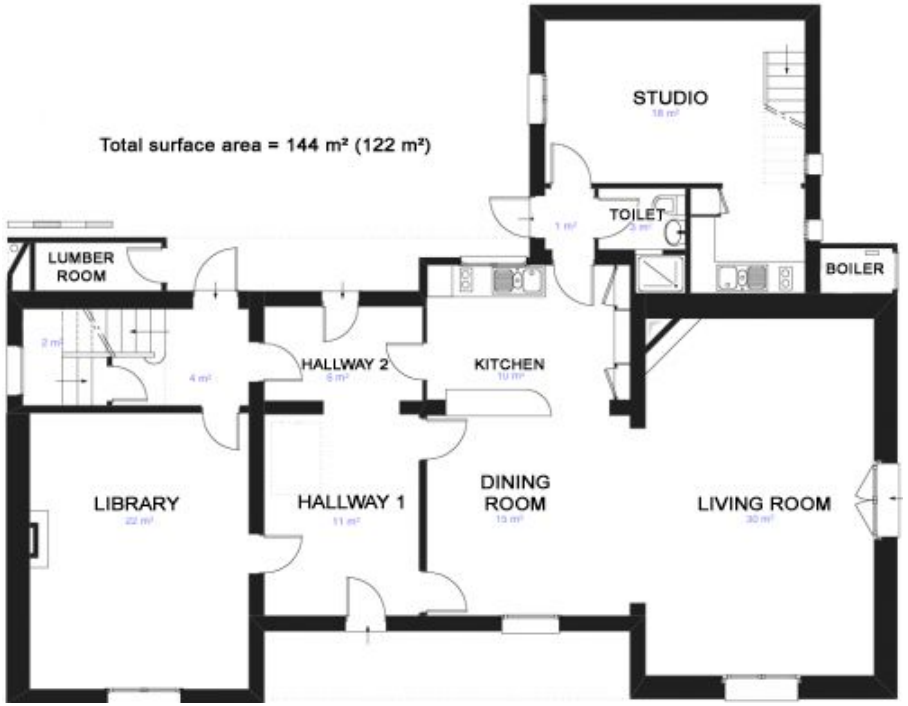
- All eukaryotic cells store genetic information in chromosomes
- Most eukaryotes have between 10 and 50 chromosomes in their body cells
- Human body cells have 46 chromosomes or 23 nearly-identical pairs
- Chromosomes are composed of a complex of DNA and protein called chromatin that condenses during cell division



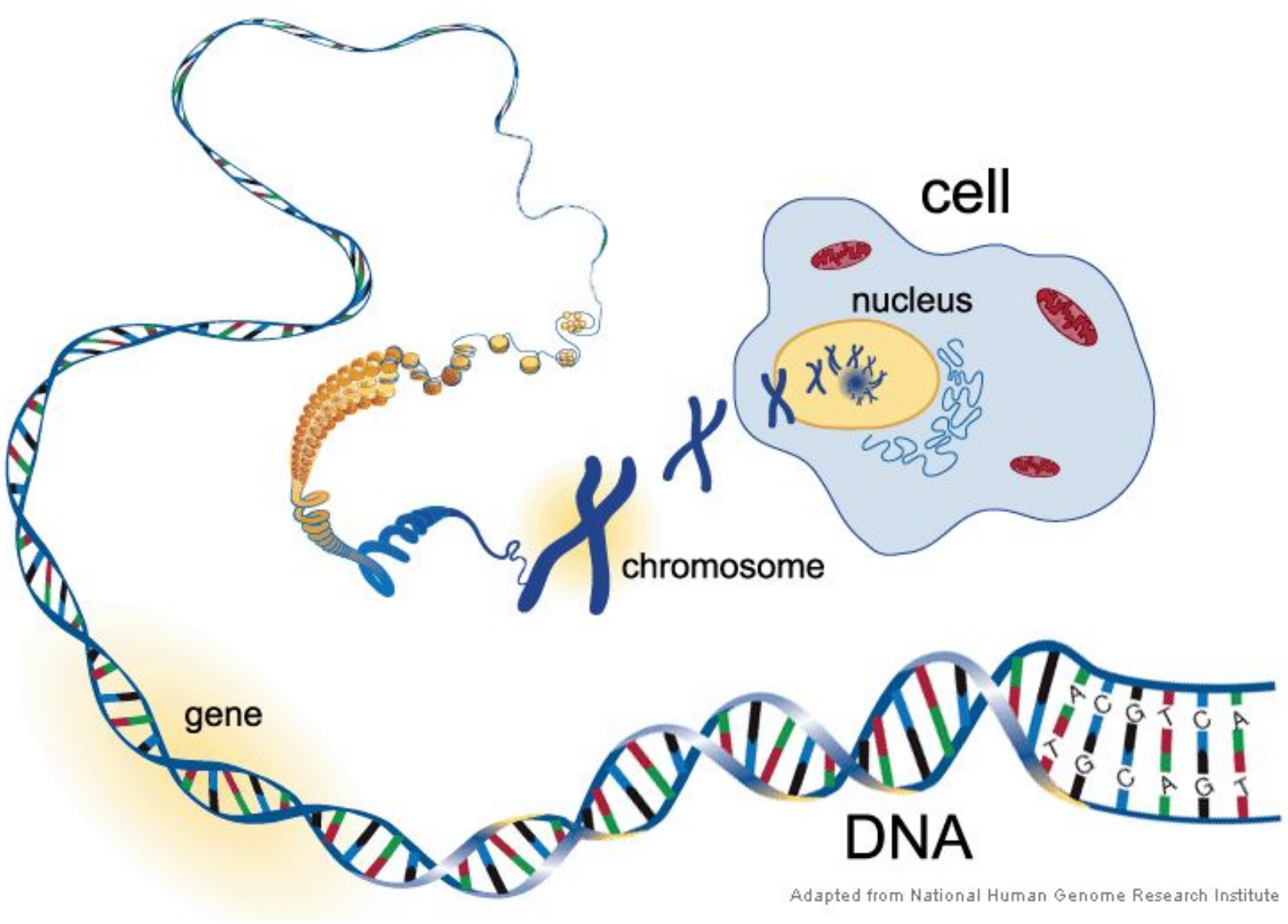


DNA

"The Blueprint of Life"



DNA stands for... **D**eoxyribo**N**ucleic**A**cid



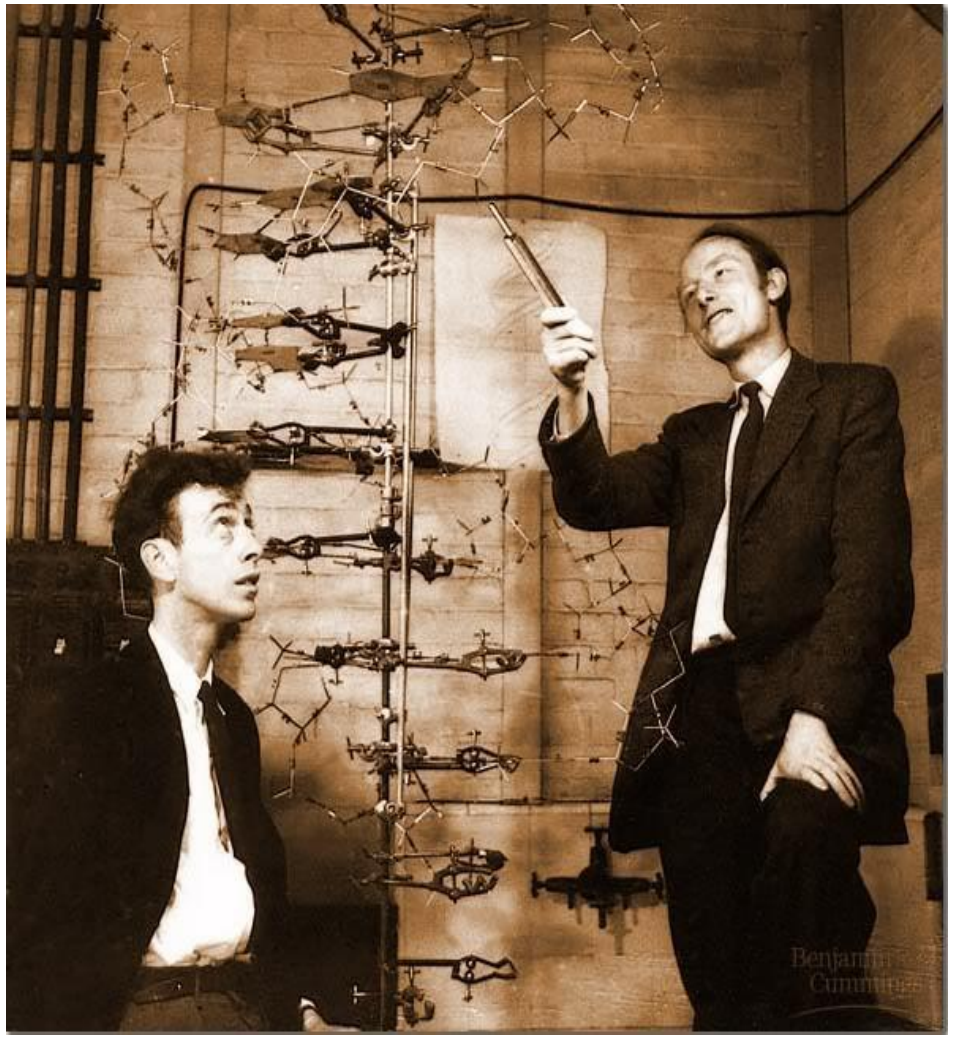
DNA Facts

- DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. Nearly every cell in a person's body has the same DNA. Most DNA is located in the cell nucleus (where it is called nuclear DNA), but a small amount of DNA can also be found in the mitochondria (where it is called mitochondrial DNA or mtDNA).
- DNA is arranged as a coil of coils of coils of coils! This allows the 3 billion base pairs in each cell to fit into a space just 6 microns across.
- If you stretched the DNA in one cell all the way out, it would be about 2m long and all the DNA in all your cells put together would be about twice the diameter of the Solar System.

DNA Facts

Established by James Watson and Francis Crick

Shape of a double helix



DNA made of repeating subunits called NUCLEOTIDES (4 Types)

1. Adenine ------ A

2.Thymine ------ T

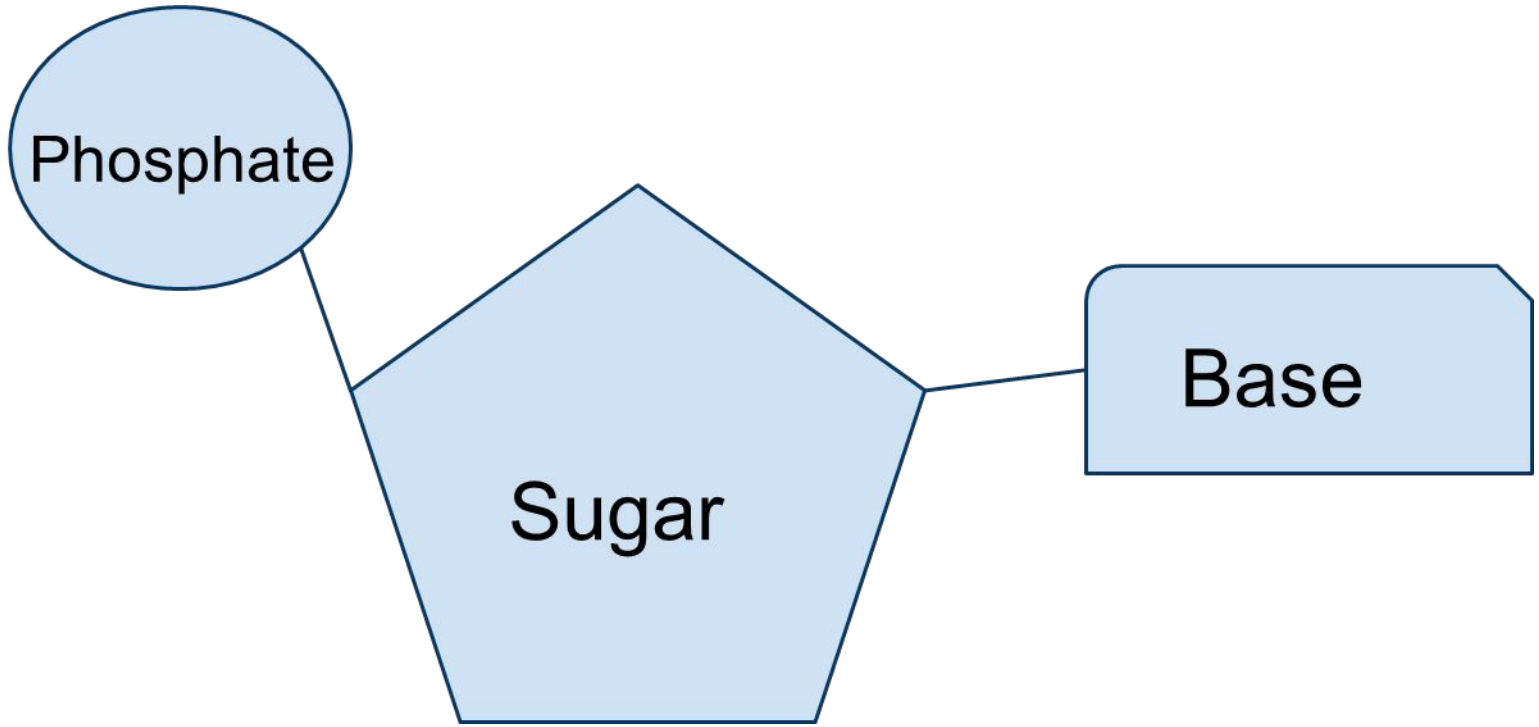
3.Guanine ------ G

4.Cytosine ------ C

What is a nucleotide?

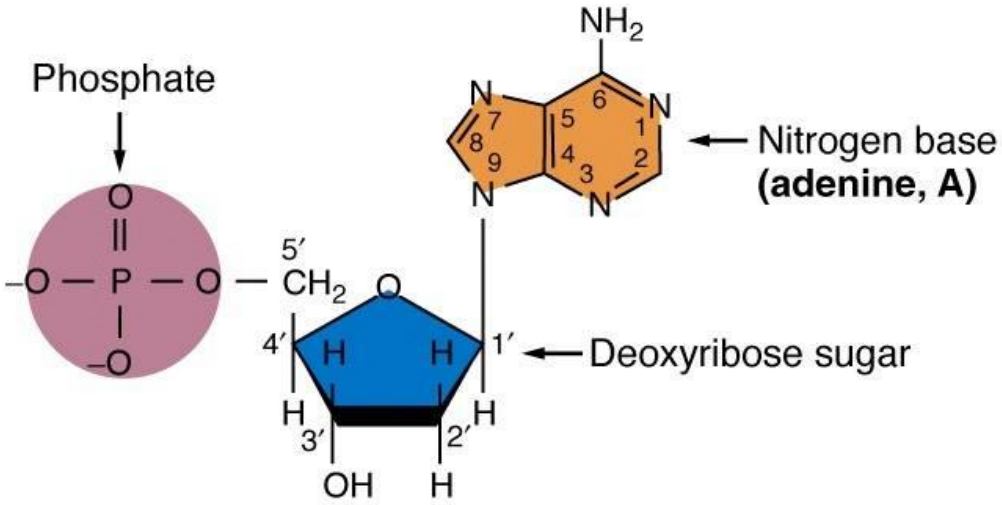
Nucleotides have three characteristic components:

1. a nitrogenous (nitrogen-containing) base,
2. a pentose (sugar),
3. a phosphate

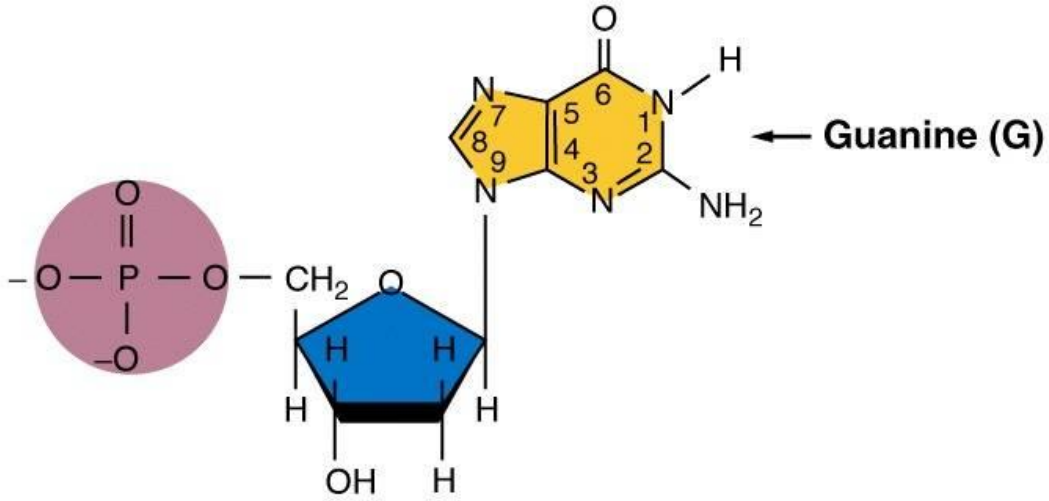


The nitrogenous bases are derivatives of two parent compounds, **pyrimidine** and **purine**.

Purine nucleotides

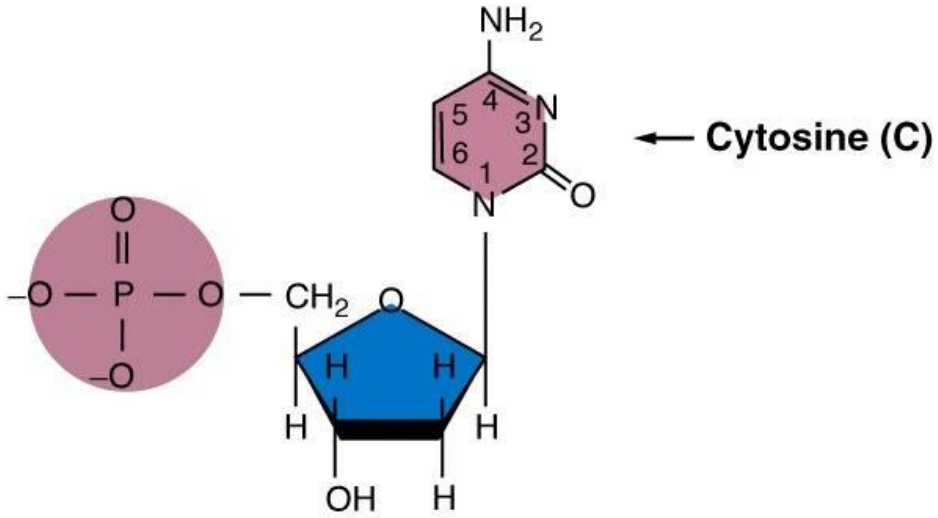


Deoxyadenosine 5′-phosphate (dAMP)

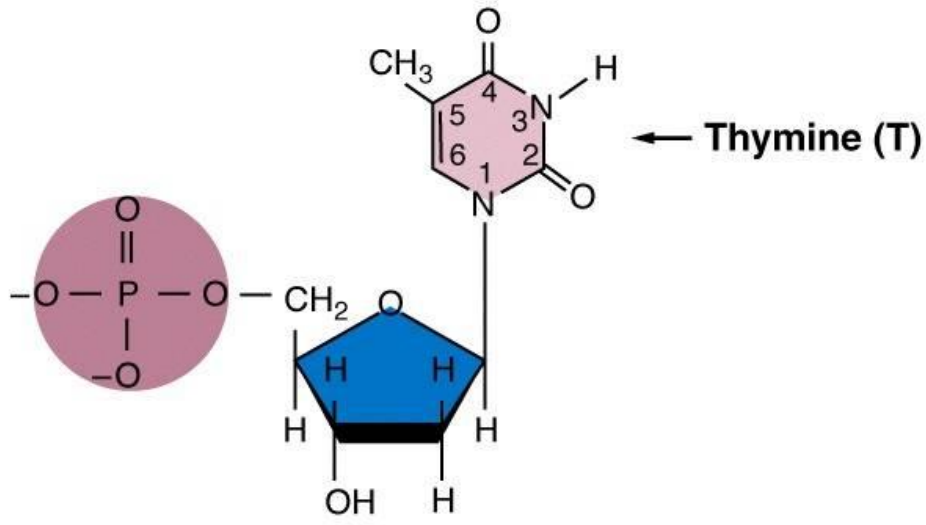


Deoxyguanosine 5′-phosphate (dGMP)

Pyrimidine nucleotides



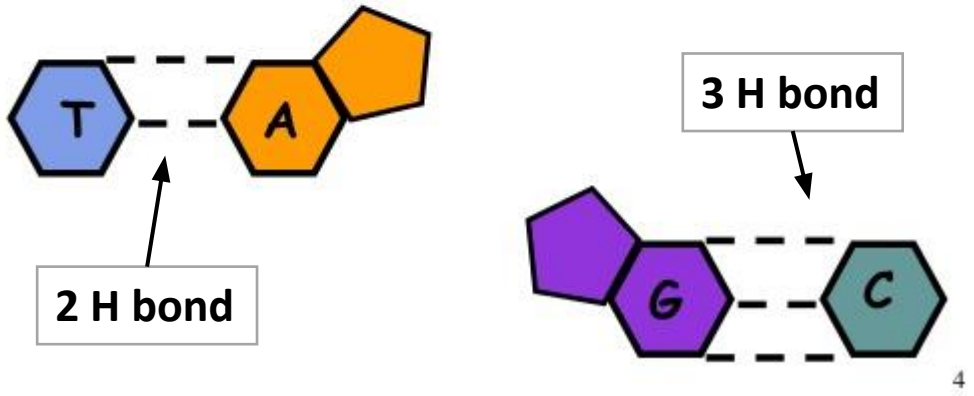
Deoxycytidine 5′-phosphate (dCMP)

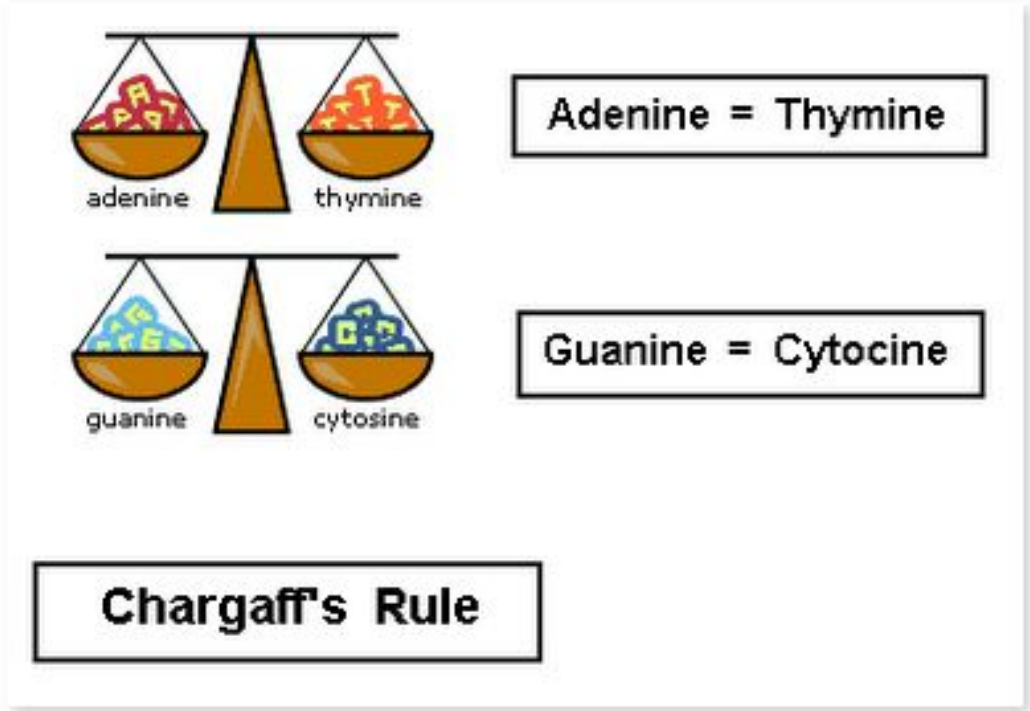


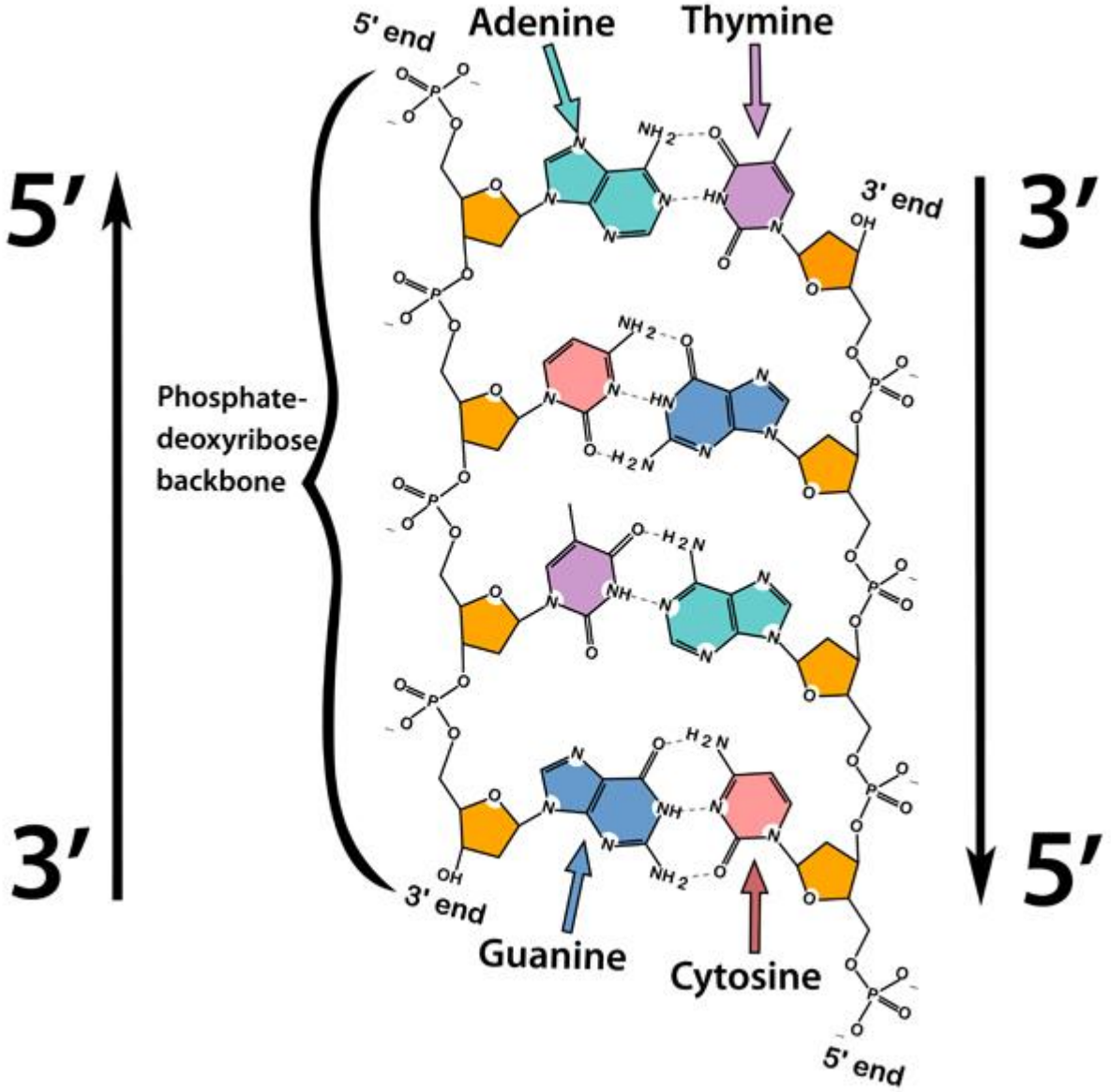
Deoxythymidine 5′-phosphate (dTMP)

Chargaff's Rule of Base Pairing

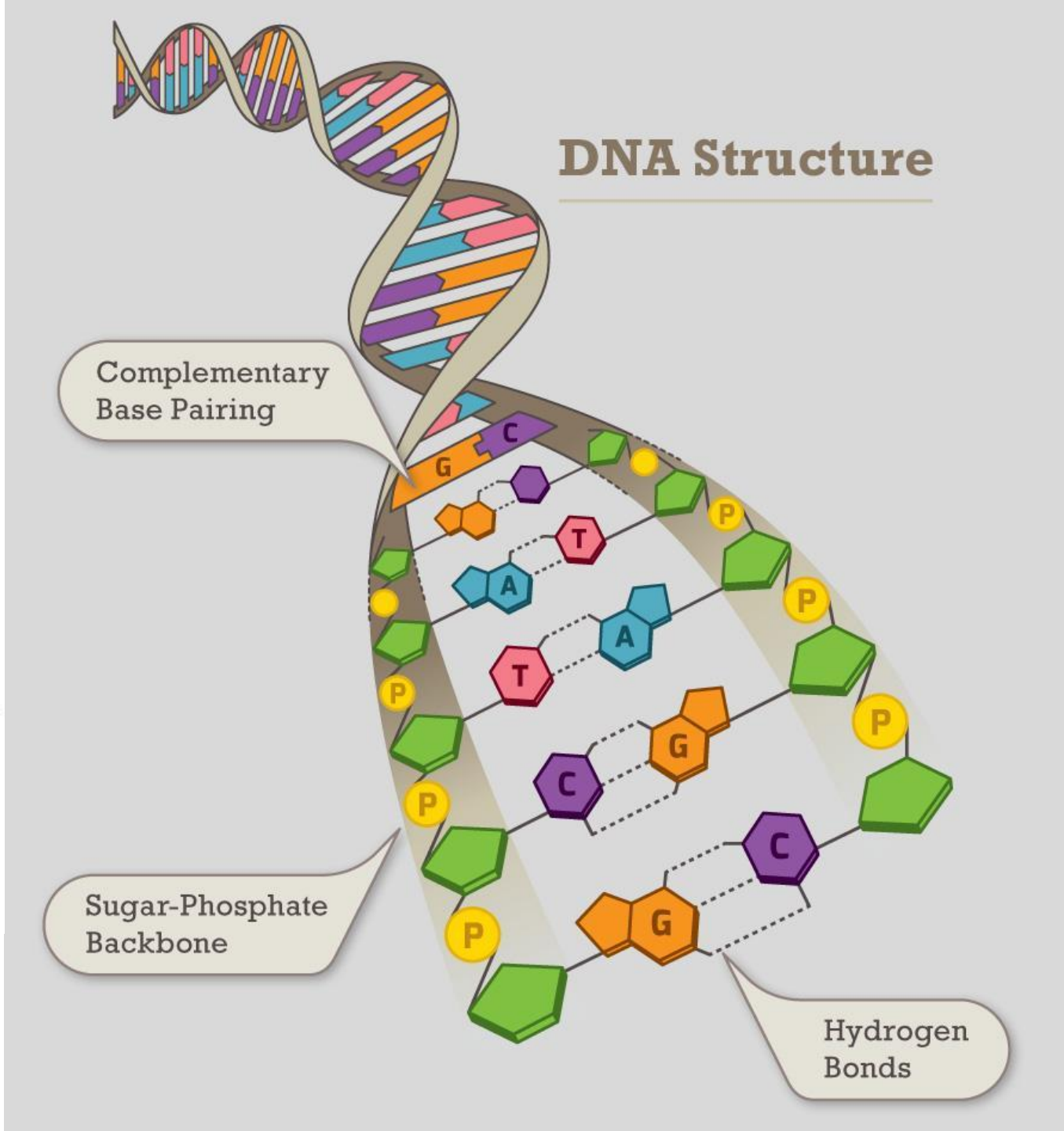
- The rules of base pairing (or nucleotide pairing) are:
  - **A with T**: the purine adenine (A) always pairs with the pyrimidine thymine (T)
  - **C with G**: the pyrimidine cytosine (C) always pairs with the purine guanine (G)

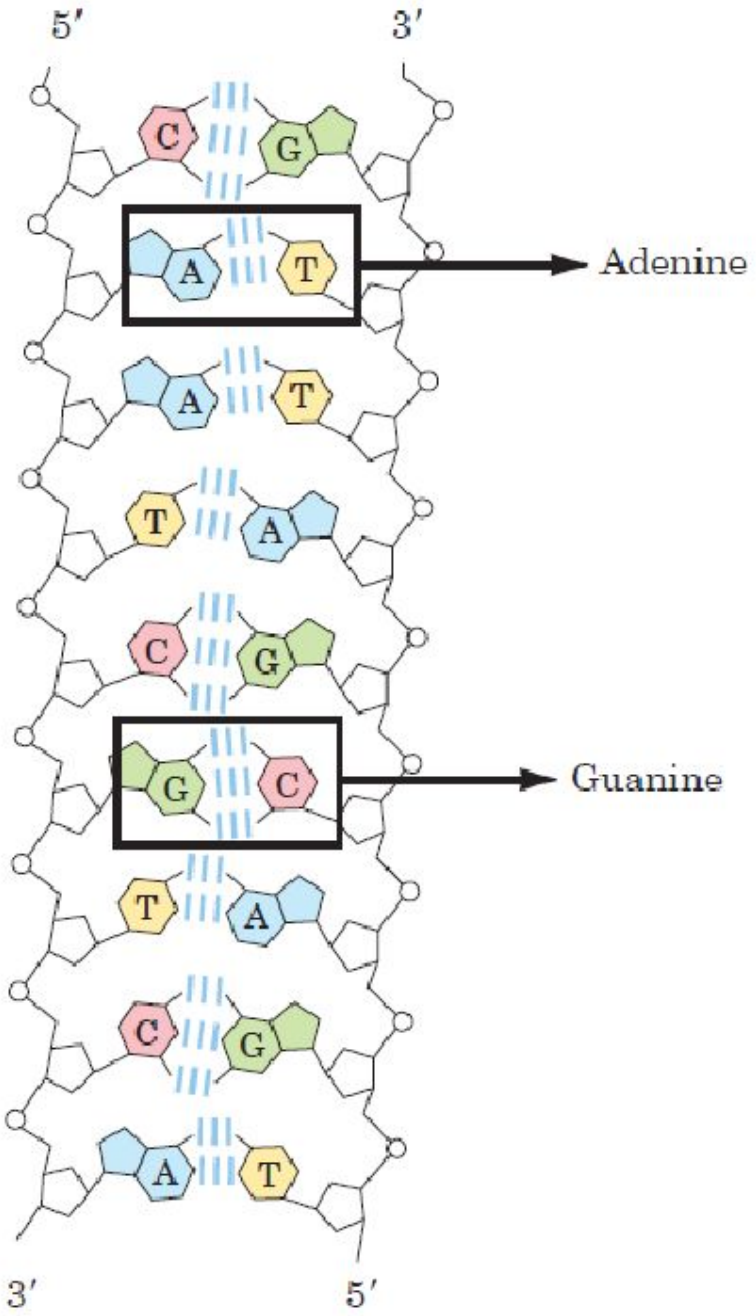






Almost always read in 5' and 3' direction





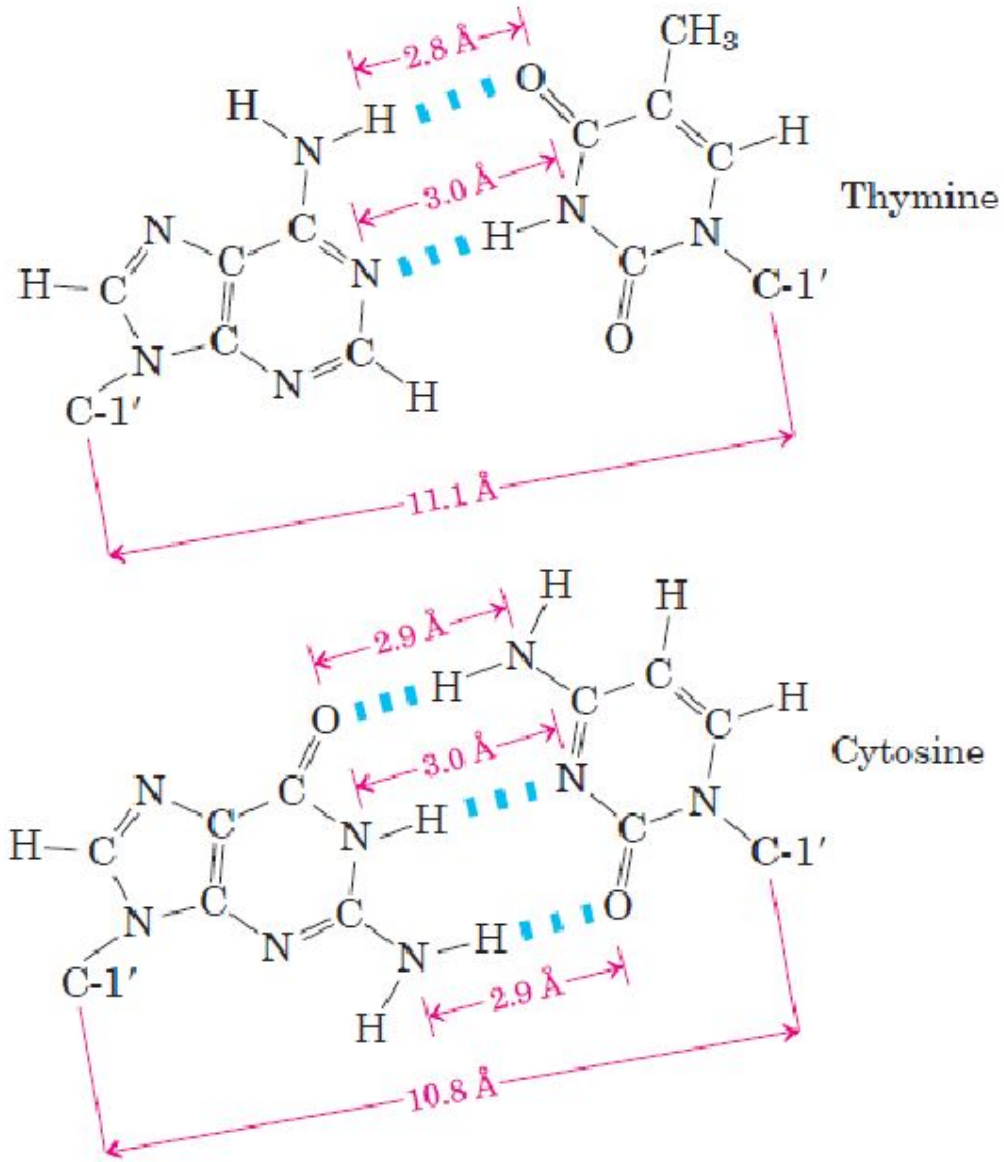
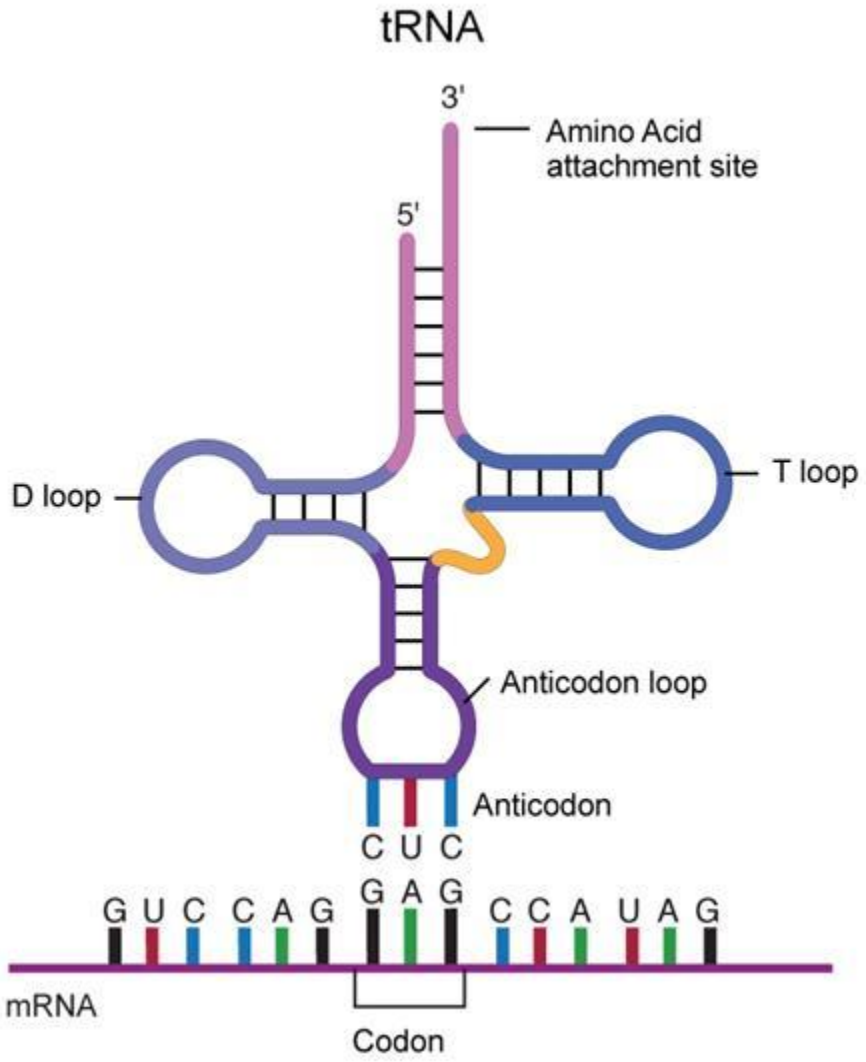
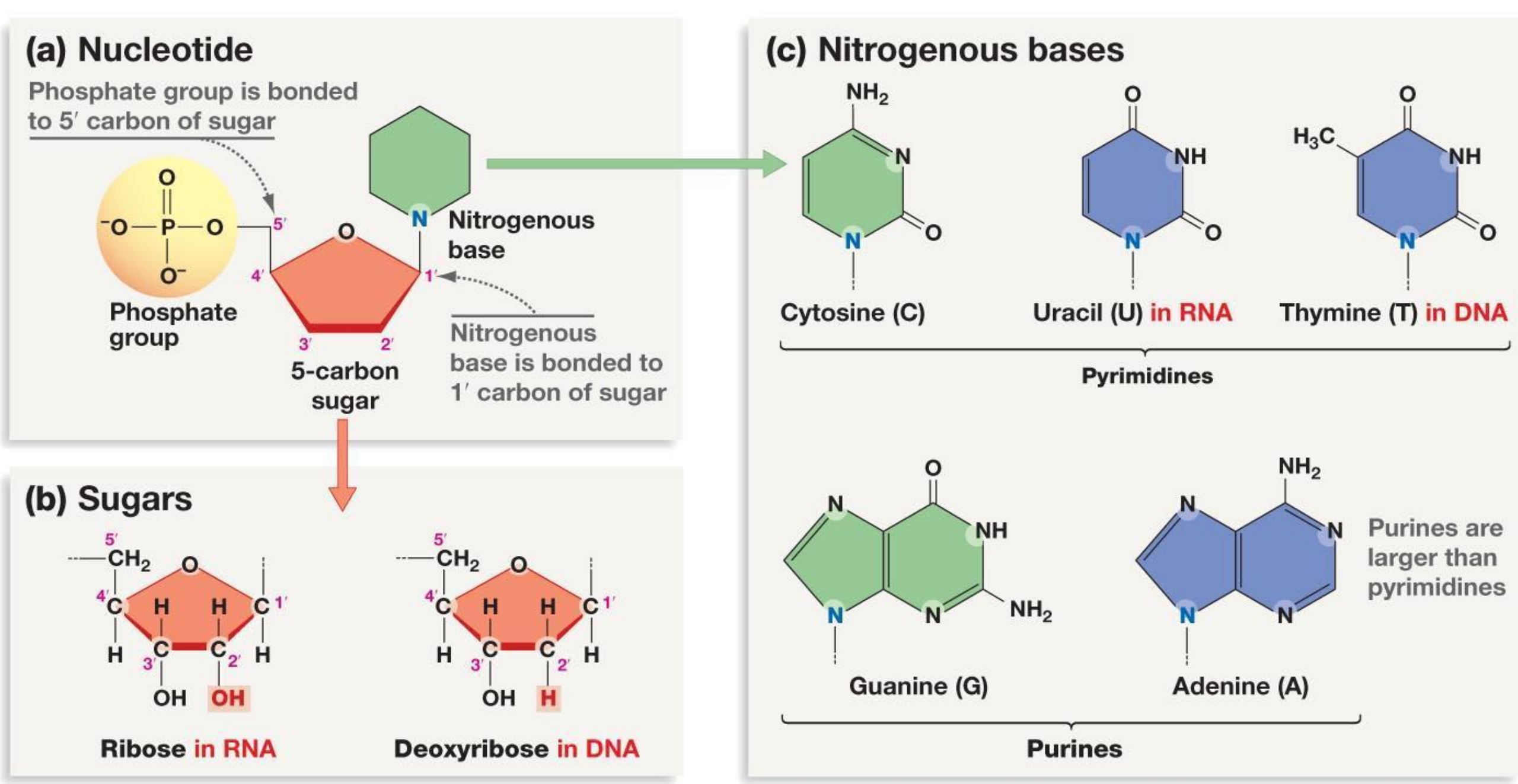


FIGURE 8–11 Hydrogen-bonding patterns in the base pairs defined by Watson and Crick. Here as elsewhere, hydrogen bonds are represented by three blue lines.

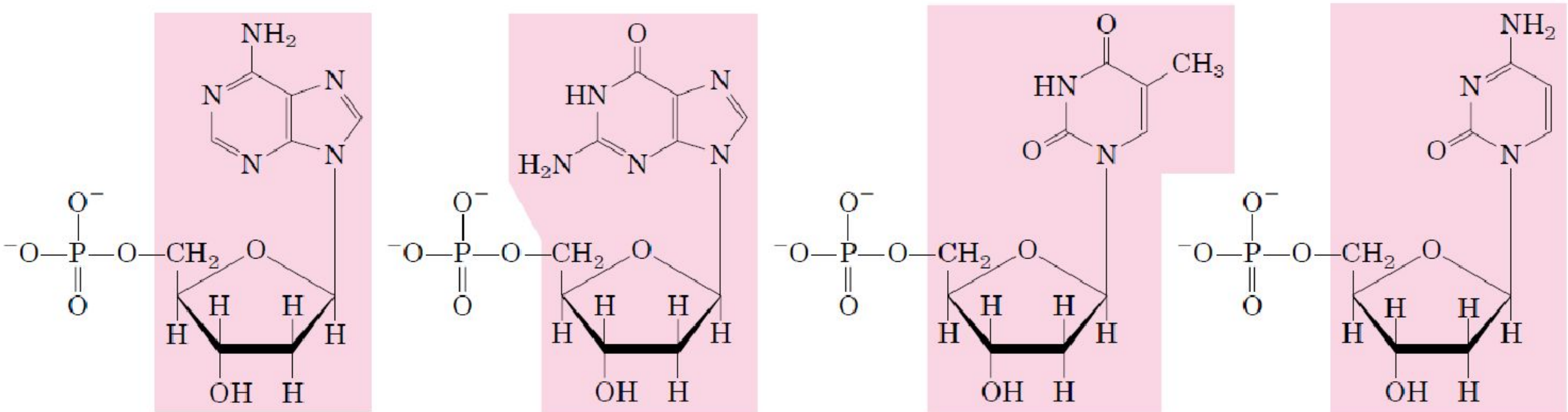
RNA stands for **R**ibo**N**ucleic**A**cid

- Uracil (U) in RNA [Thymine (T) in DNA]
- Primarily Single-stranded
- Example: mRNA, tRNA, rRNA





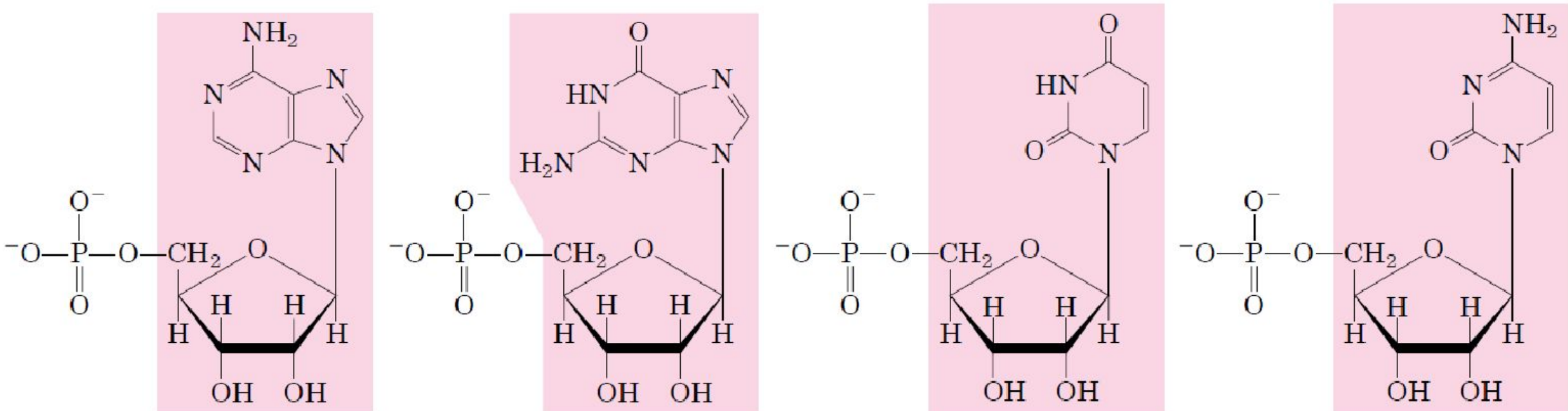
Deoxyribonucleotides



Nucleotide:	Deoxyadenylate (deoxyadenosine 5′-monophosphate)	Deoxyguanylate (deoxyguanosine 5′-monophosphate)	Deoxythymidylate (deoxythymidine 5′-monophosphate)	Deoxycytidylate (deoxycytidine 5′-monophosphate)
Symbols:	A, dA, dAMP	G, dG, dGMP	T, dT, dTMP	C, dC, dCMP
Nucleoside:	Deoxyadenosine	Deoxyguanosine	Deoxythymidine	Deoxycytidine

(a) Deoxyribonucleotides

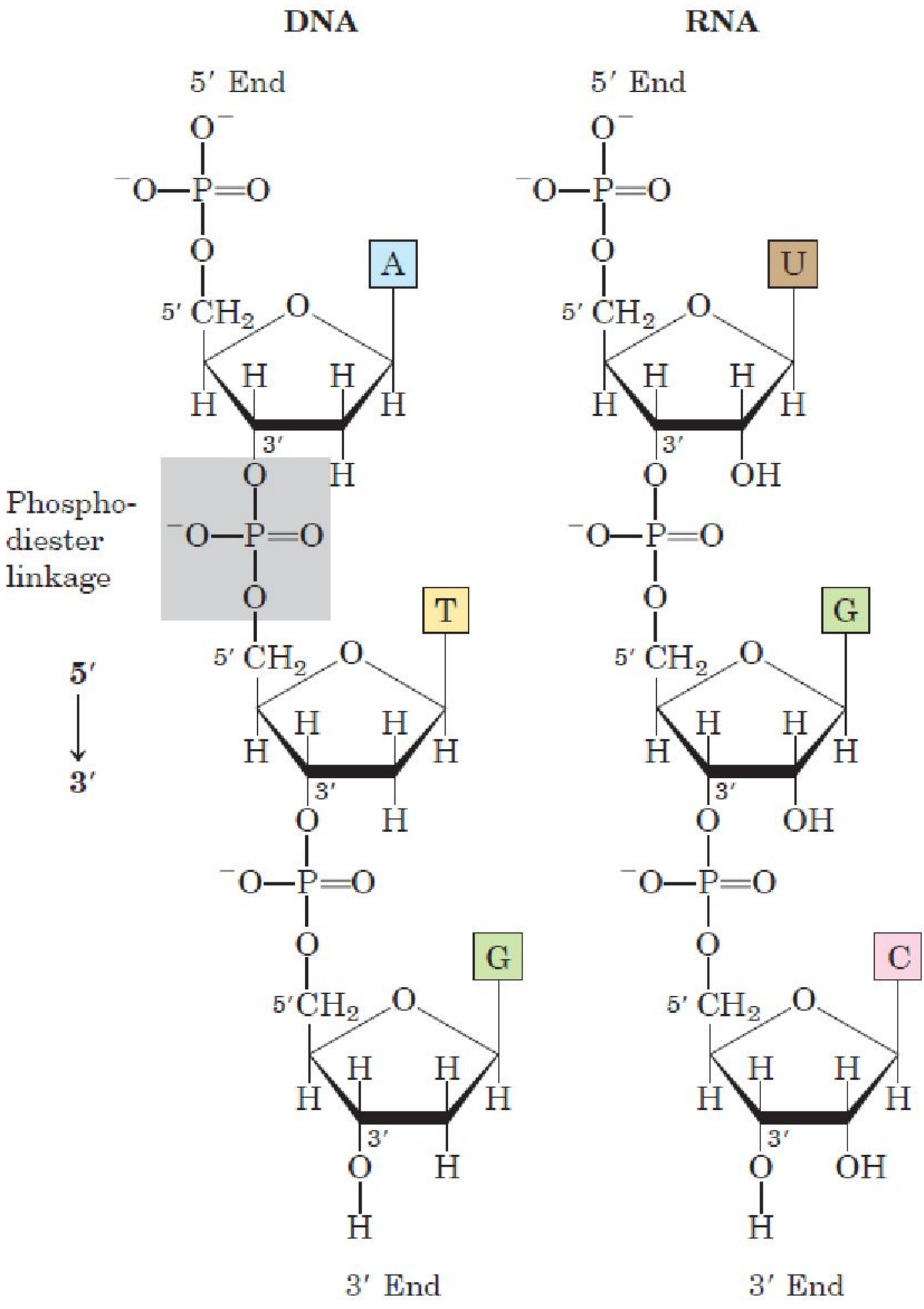
Ribonucleotides



	Adenylate (adenosine 5′-monophosphate)	Guanylate (guanosine 5′-monophosphate)	Uridylate (uridine 5′-monophosphate)	Cytidylate (cytidine 5′-monophosphate)
Nucleotide:				
Symbols:	A, AMP	G, GMP	U, UMP	C, CMP
Nucleoside:	Adenosine	Guanosine	Uridine	Cytidine

(b) Ribonucleotides

Phosphodiester linkages in the covalent backbone of DNA and RNA.



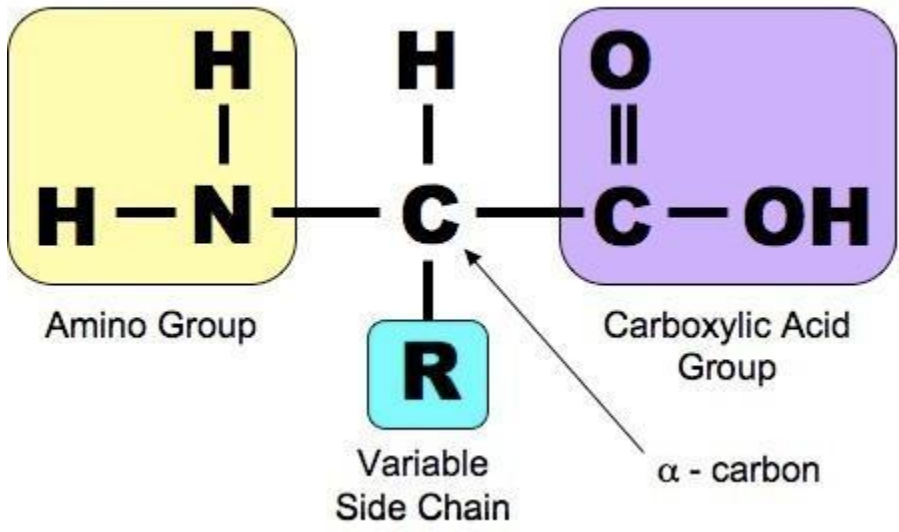
Protein

- **Proteins** are large biomolecules, or **macromolecules**, consisting of one or more long chains of **amino acid residues**
- Are found in every cell in the body
- They do most of the work in cells and are required for the structure, function, and regulation of the body's tissues and organs.
- The sequence of **amino acids** is determined by DNA

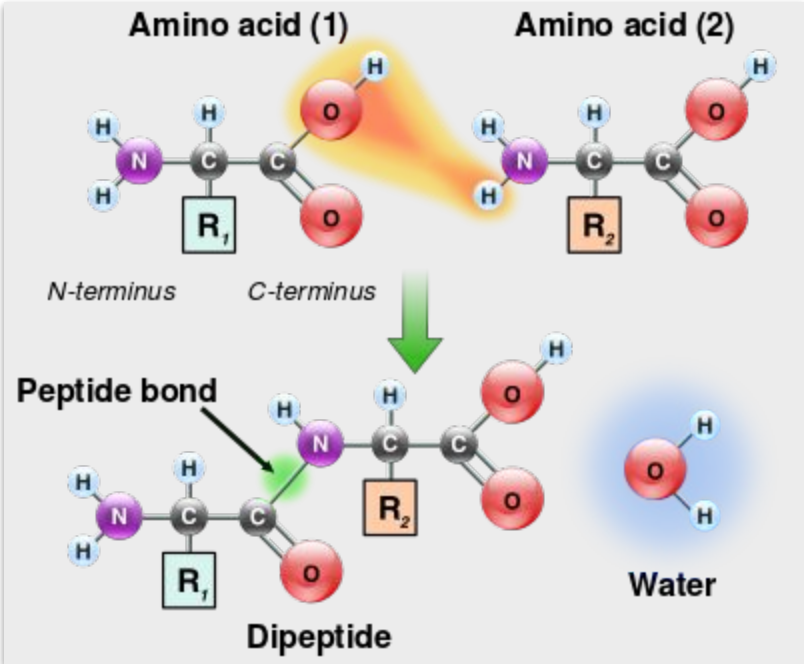
Protein Function

Function	Description	Example
Antibody	Antibodies bind to specific foreign particles, such as viruses and bacteria, to help protect the body.	Immunoglobulin G (IgG)
Enzyme	Enzymes carry out almost all of the thousands of chemical reactions that take place in cells. They also assist with the formation of new molecules by reading the genetic information stored in DNA.	Phenylalanine hydroxylase
Messenger	Messenger proteins, such as some types of hormones, transmit signals to coordinate biological processes between different cells, tissues, and organs.	Growth hormone
Structural component	These proteins provide structure and support for cells. On a larger scale, they also allow the body to move.	Actin
Transport/storage	These proteins bind and carry atoms and small molecules within cells and throughout the body.	Ferritin

Amino Acids



- Amino acids are small organic molecules that consist of an **alpha (central) carbon atom** linked to an amino group, a carboxyl group, a hydrogen atom, and a variable component called a side chain
- There are **20** different types of **amino acids** that can be combined to make a protein.
- Within a protein, multiple amino acids are linked together by **peptide bonds**, thereby forming a long chain.
- **Peptide bonds** are formed by a biochemical reaction that extracts a water molecule as it joins the amino group of one amino acid to the carboxyl group of a neighboring amino acid.
- The linear sequence of amino acids within a protein is considered the **primary structure** of the protein.



The 20 Amino acids in Protein

#	1-Letter Code	3-Letter Code	Name
1	A	Ala	Alanine
2	R	Arg	Arginine
3	N	Asn	Asparagine
4	D	Asp	Aspartic acid
5	C	Cys	Cysteine
6	Q	Gln	Glutamine
7	E	Glu	Glutamic acid
8	G	Gly	Glycine
9	H	His	Histidine
10	I	Ile	Isoleucine
11	L	Leu	Leucine
12	K	Lys	Lysine
13	M	Met	Methionine
14	F	Phe	Phenylalanine
15	P	Pro	Proline
16	S	Ser	Serine
17	T	Thr	Threonine
18	W	Trp	Tryptophan
19	Y	Tyr	Tyrosine
20	V	Val	Valine

Structure of Proteins

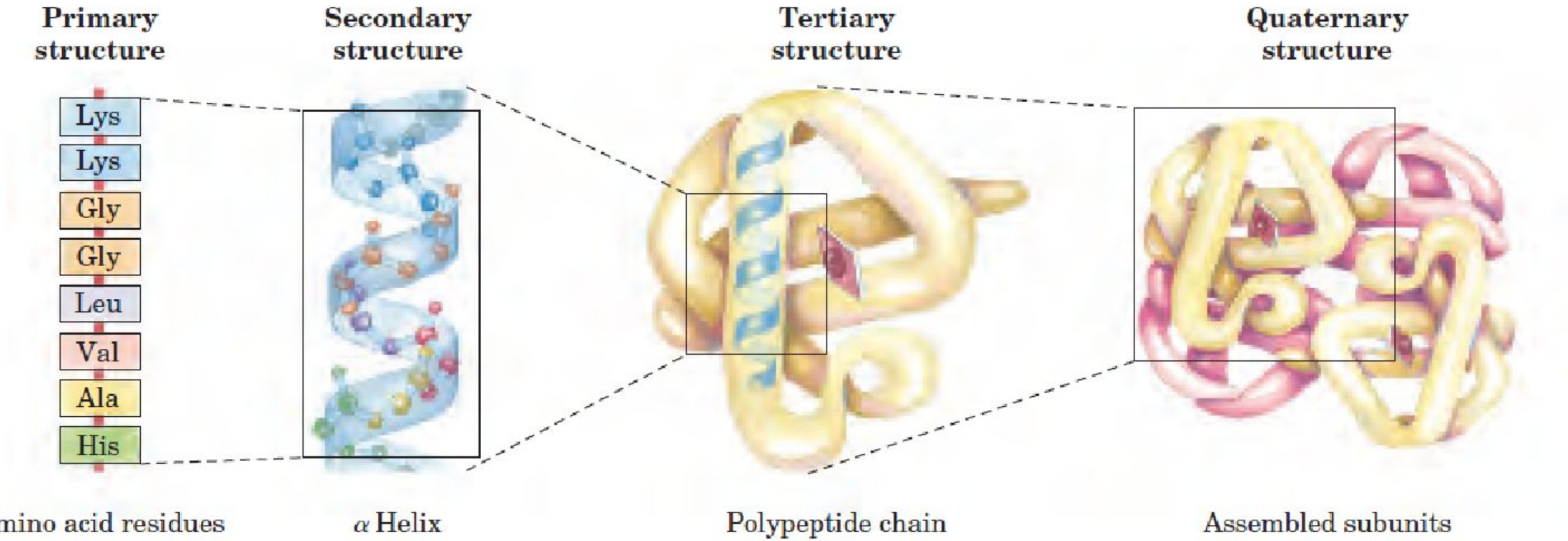
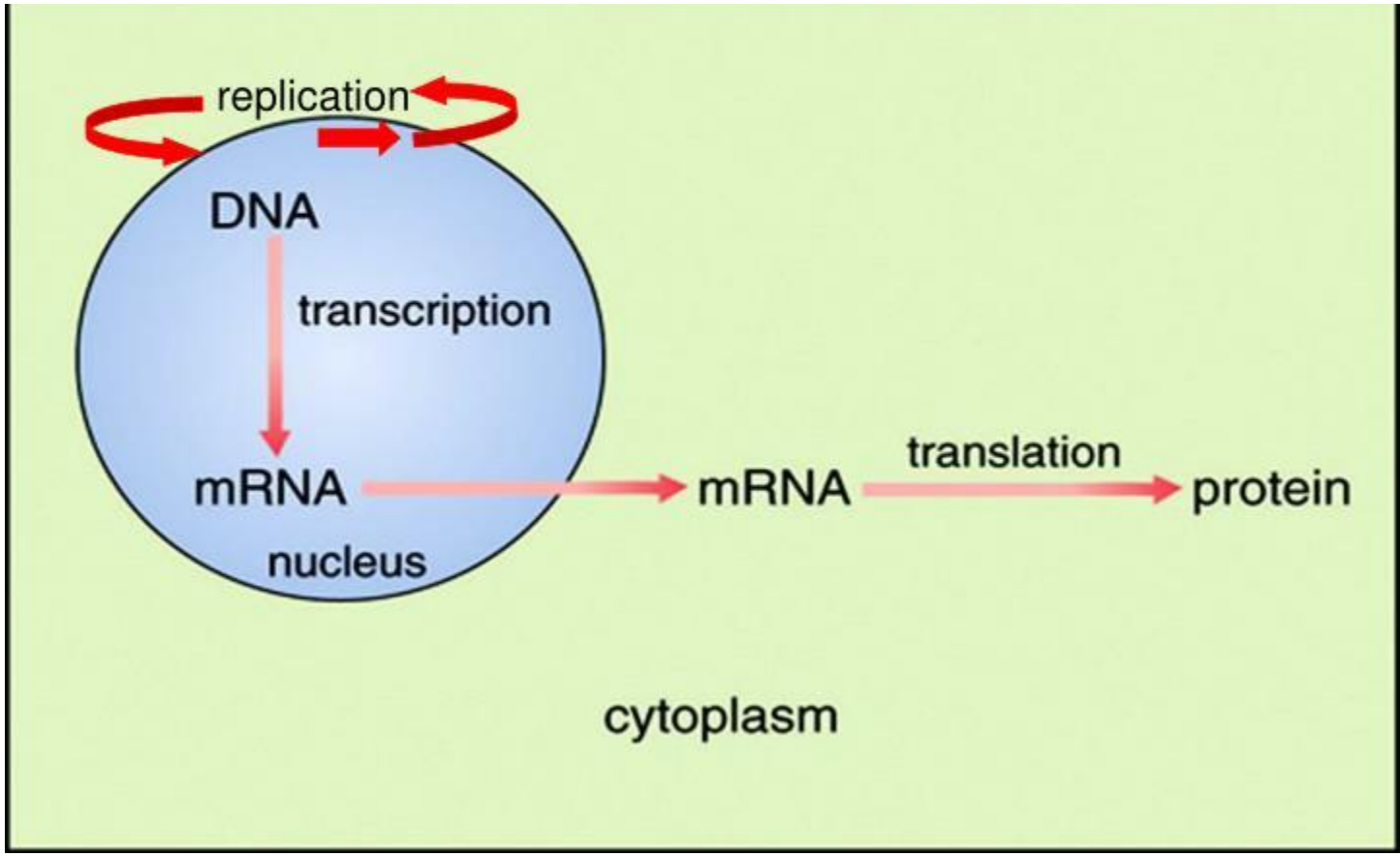


FIGURE 3-16 Levels of structure in proteins. The *primary structure* consists of a sequence of amino acids linked together by peptide bonds and includes any disulfide bonds. The resulting polypeptide can be coiled into units of *secondary structure,* such as an α helix. The helix is a part of the *tertiary structure* of the folded polypeptide, which is itself one of the subunits that make up the *quaternary structure* of the multisubunit protein, in this case hemoglobin.

Central Dogma

- The **central dogma** of molecular biology, showing the general pathways of information flow via **replication**, **transcription**, and **translation**.
- The term "**dogma**" is a misnomer. Introduced by Francis Crick at a time when little evidence supported these ideas, the dogma has become a well-established principle.
- The first is **replication,** the copying of parental DNA to form daughter DNA molecules with identical nucleotide sequences.
- The second is **transcription,** the process by which parts of the genetic message encoded in DNA are copied precisely into RNA.
- The third is **translation,** whereby the genetic message encoded in messenger RNA is translated on the ribosomes into a polypeptide with a particular sequence of amino acids.



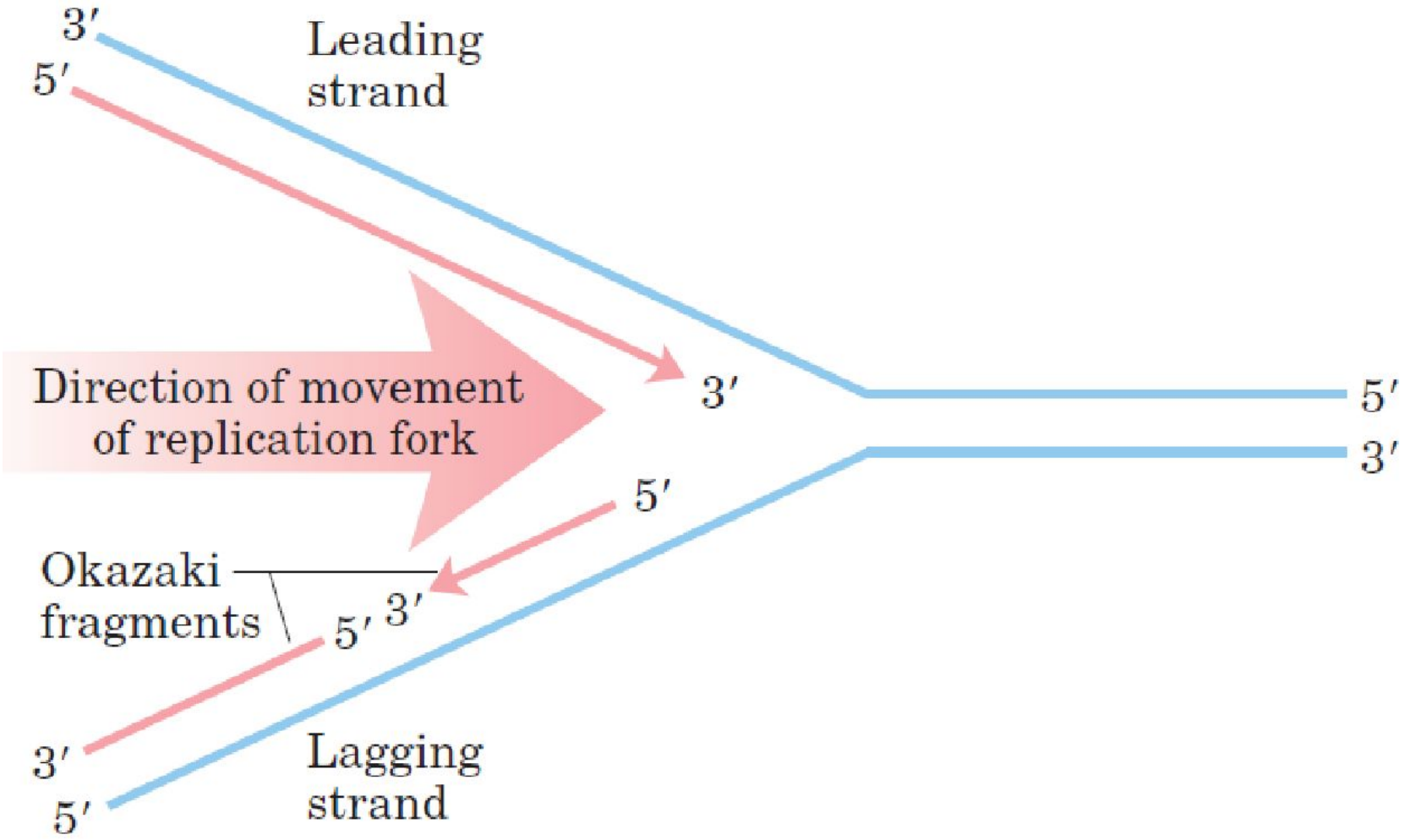
DNA Replication

- DNA replication is the process by which DNA makes a copy of itself during cell division.
- Animation Link: https://www.youtube.com/watch?v=TNKWgcFPHqw

1. The first step in DNA replication is to **'unzip'** the double helix structure of the DNA molecule.
2. This is carried out by an **enzyme** called **helicase** which breaks the hydrogen bonds holding the complementary bases of DNA together (A with T, C with G).
3. The separation of the two single strands of DNA creates a 'Y' shape called a **replication 'fork'**. The two separated strands will act as templates for making the new strands of DNA.
4. One of the strands is oriented in the 3' to 5' direction (towards the replication fork), this is the **leading strand**. The other strand is oriented in the 5' to 3' direction (away from the replication fork), this is the **lagging strand**. As a result of their different orientations, the two strands are replicated differently.

FIGURE: Defining DNA strands at the replication fork

- A **new DNA** strand (red) is always synthesized in the **5′→3′ direction**.
- The **template** is read in the opposite direction, **3′→5′**.
- The **leading** strand is **continuously** synthesized in the direction taken by the replication fork.
- The other strand, the **lagging** strand, is synthesized **discontinuously** in short pieces (Okazaki fragments) in a direction opposite to that in which the replication fork moves.
- The **Okazaki fragments** are spliced together by **DNA ligase**.
- In bacteria, Okazaki fragments are ~1,000 to 2,000 nucleotides long.
- In eukaryotic cells, they are 150 to 200 nucleotides long.



DNA Replication Cont.

Leading Strand:

1. A short piece of **RNA** called a **primer** (produced by an enzyme called **primase**) comes along and binds to the end of the leading strand. The primer acts as the **starting point** for DNA synthesis.
2. **DNA polymerase** binds to the leading strand and then 'walks' along it, adding new complementary nucleotide bases (A, C, G and T) to the strand of DNA in the **5' to 3' direction**.
3. This sort of replication is called **continuous**.

Lagging strand

1. Numerous **RNA primers** are made by the **primase enzyme** and bind at **various points** along the lagging strand.
2. Chunks of DNA, called **Okazaki fragments**, are then added to the lagging strand also in the 5' to 3' direction.
3. This type of replication is called **discontinuous** as the Okazaki fragments will need to be joined up later.

DNA Replication Cont.

8. Once all of the bases are matched up (A with T, C with G), an enzyme called **exonuclease strips** away the **primer**(s). The gaps where the primer(s) were, are then filled by yet more complementary nucleotides.
9. The new strand is **proofread** to make sure there are no mistakes in the new DNA sequence by **DNA polymerase**.
10. Finally, an **enzyme** called **DNA ligase** seals up the sequence of DNA into two continuous double strands.
11. The result of DNA replication is two DNA molecules consisting of **one new** and **one old** chain of nucleotides. This is why DNA replication is described as **semi-conservative**, half of the chain is part of the original DNA molecule, half is brand new.
12. Following replication the new DNA automatically winds up into a double helix.

Transcription: DNA to RNA

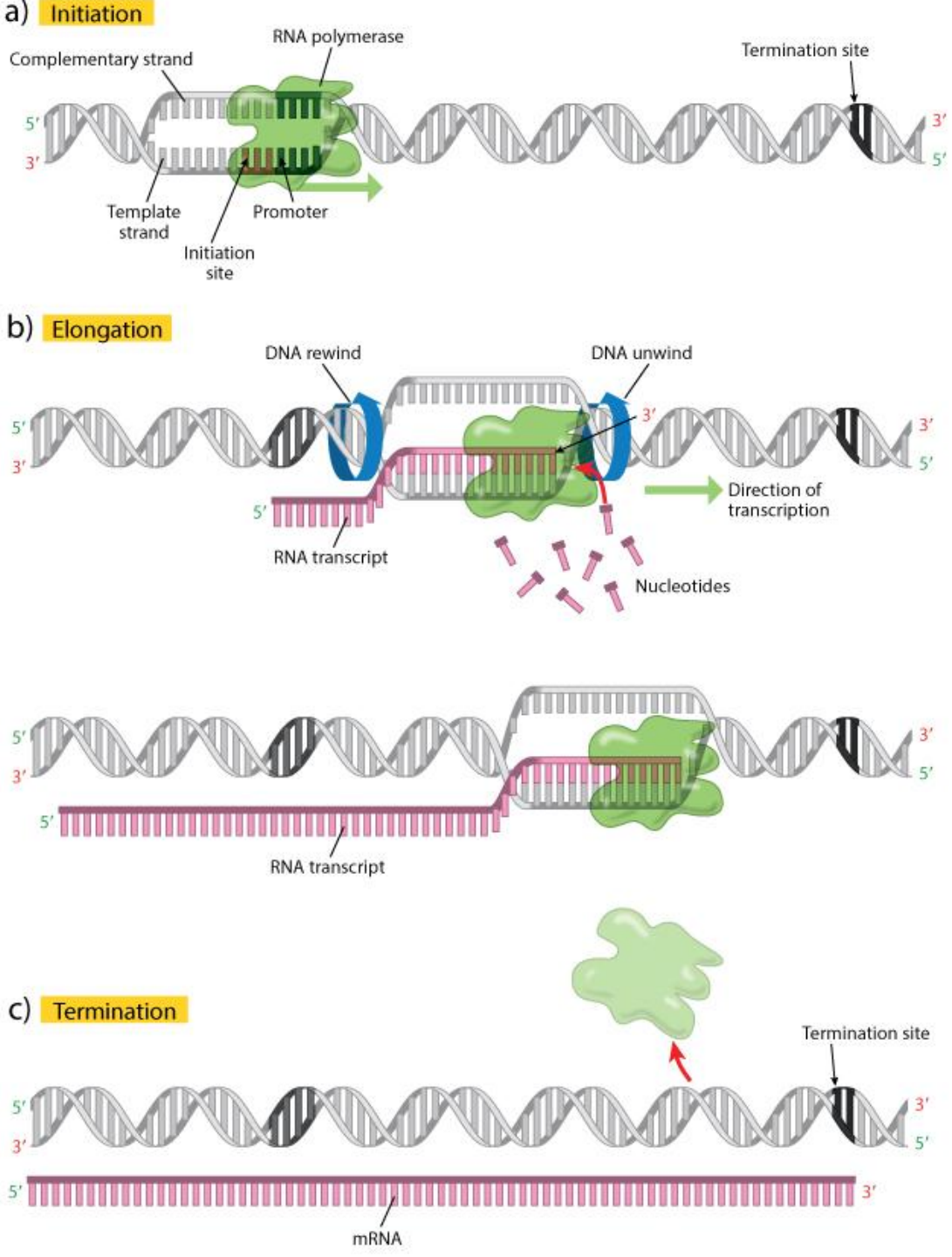
- In **transcription**, a portion of the double-stranded DNA template gives rise to a **single-stranded RNA** molecule.
- In some cases, the RNA molecule itself is a "finished product" that serves some important function within the cell.
- Often, however, transcription of an RNA molecule is followed by a translation step, which ultimately results in the production of a protein molecule.

The three stages of DNA transcription.

(A) The transcription process is initiated when the enzyme **RNA polymerase** binds to a DNA template at a **promoter sequence**.

(B) During the elongation process, the DNA double helix unwinds. RNA polymerase reads the template DNA strand and adds nucleotides to the three-prime (3’) end of a growing RNA transcript.

(C) When RNA polymerase reaches a **termination sequence** on the DNA template strand, transcription is terminated and the **mRNA transcript** and RNA polymerase are released from the complex.

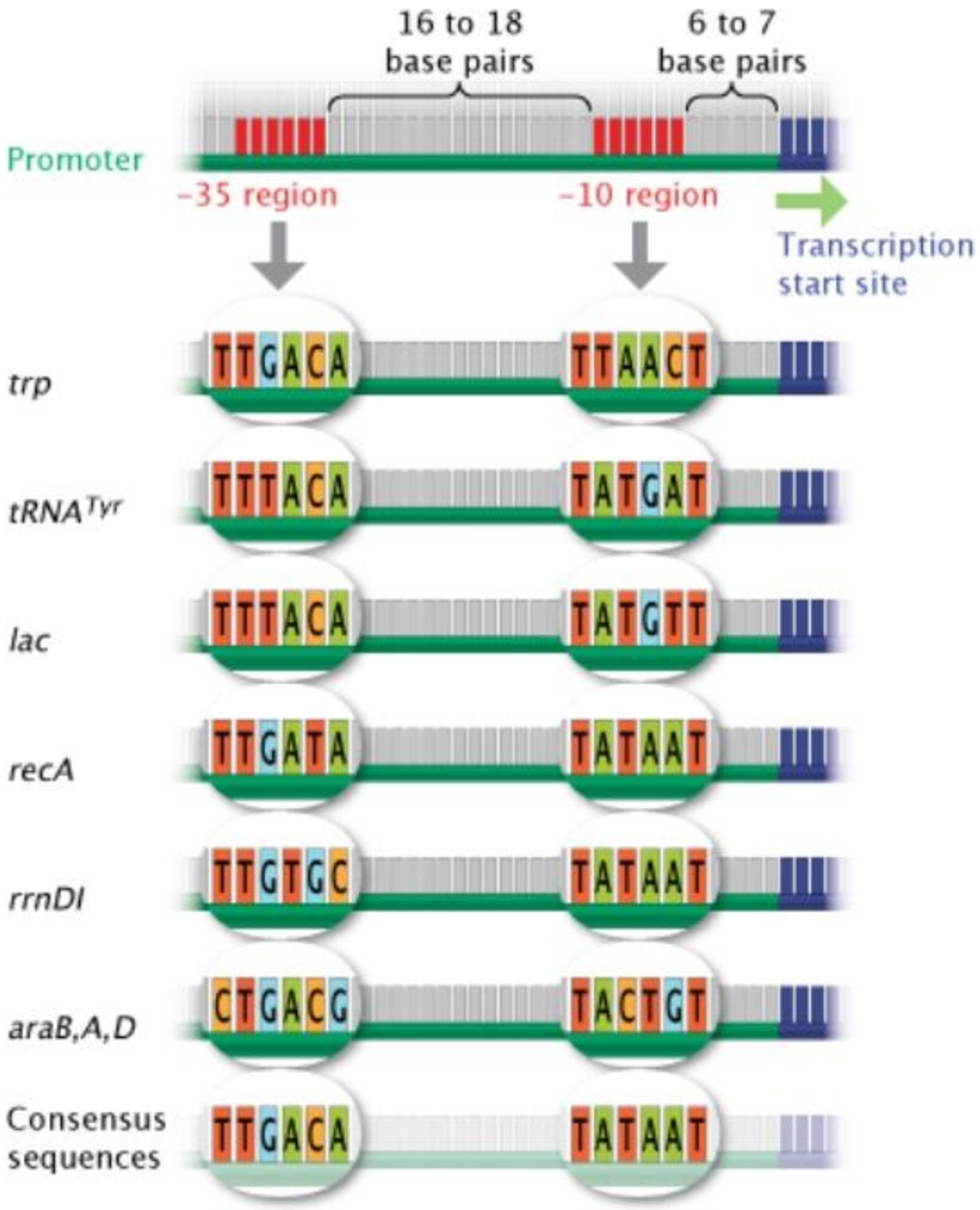


Promoter

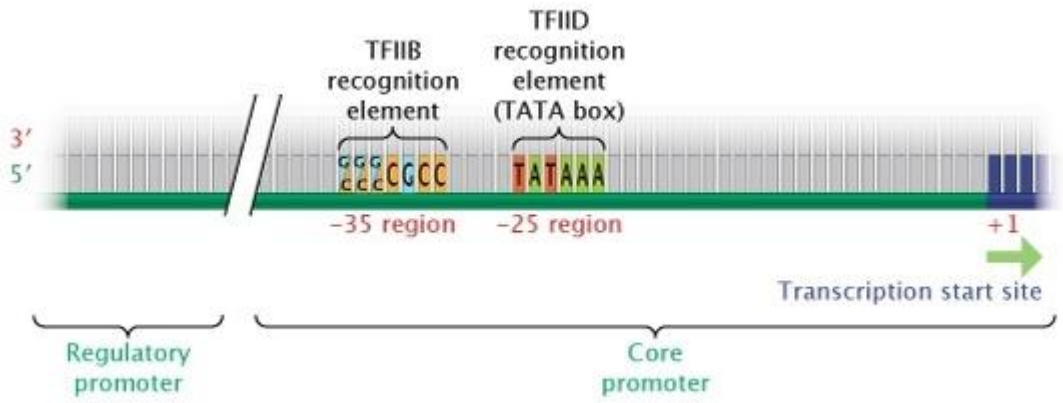
- **Promoter** sequences are DNA sequences that define where **transcription** of a **gene** by **RNA polymerase** begins. Promoter sequences are typically located directly **upstream** or at the **5' end** of the transcription initiation site. RNA polymerase and the necessary **transcription factors** bind to the promoter sequence and initiate transcription. **Promoter** sequences define the **direction** of transcription and indicate which DNA strand will be transcribed; this strand is known as the **sense strand**.
- Many eukaryotic genes have a **conserved promoter sequence** called the **TATA box**, located **25 to 35 base pairs** upstream of the transcription start site. Transcription factors bind to the TATA box and initiate the formation of the RNA polymerase transcription complex, which **promotes transcription**.

Prokaryotic transcription units.

- A prokaryotic **transcription** unit is composed of a transcription **start site** (or initiation site), a **-10 DNA** region, and a **-35 DNA** region.
- The **-10** region is located ten nucleotides **upstream** of the transcription **start site**; the **-35** region is located 35 nucleotides **upstream** of the transcription **start site**.
- Many prokaryotes share a **common**, or **similar**, sequence at their -35 and -10 regions. These shared sequences are called **consensus** sequences.

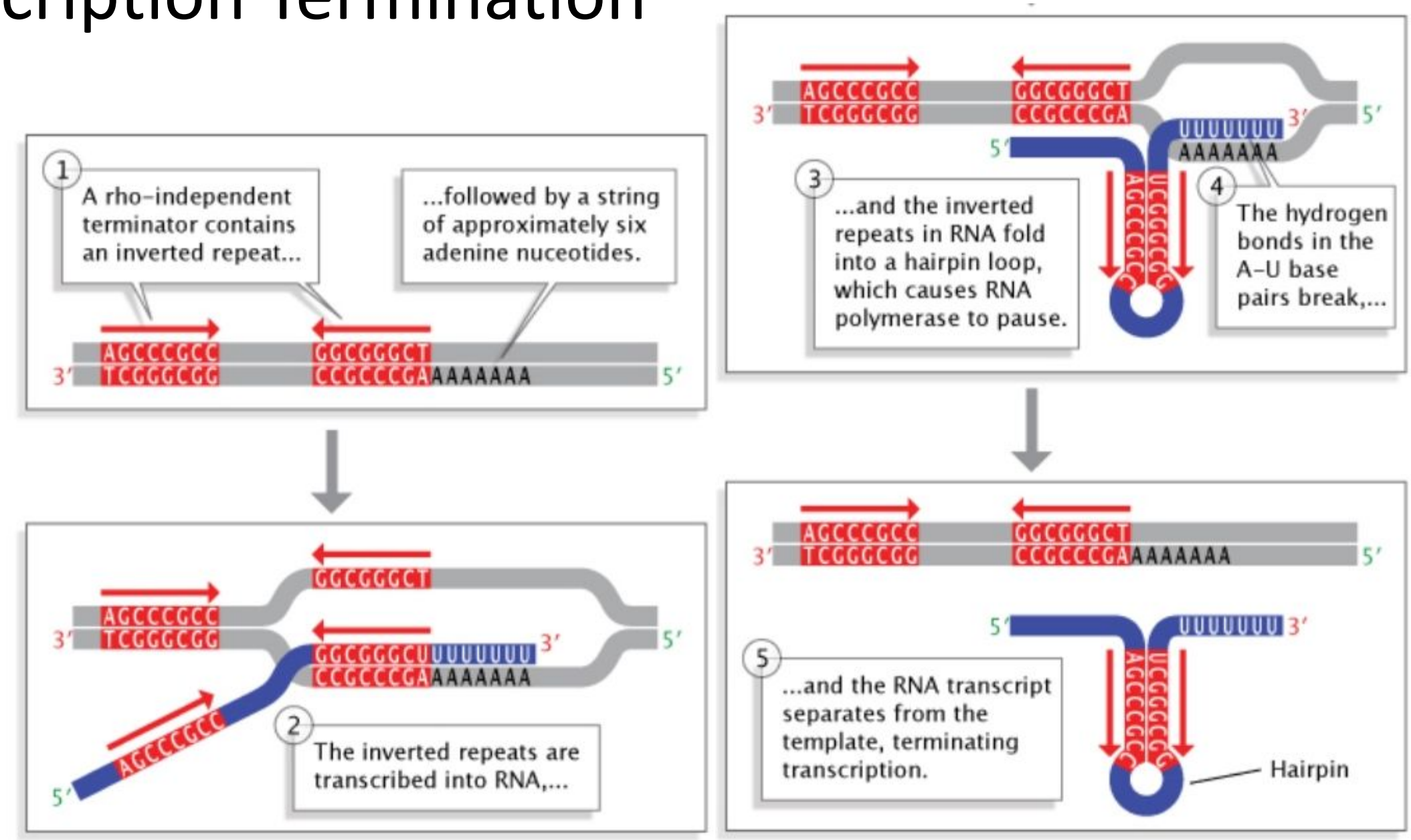


Eukaryotic core promoter region



- In eukaryotes, genes transcribed into RNA transcripts by the enzyme **RNA polymerase II** are controlled by a **core promoter**.
- A core promoter consists of a transcription start site, a **TATA box** (at the **-25 region**), and a **recognition** element (at the **-35 region**).

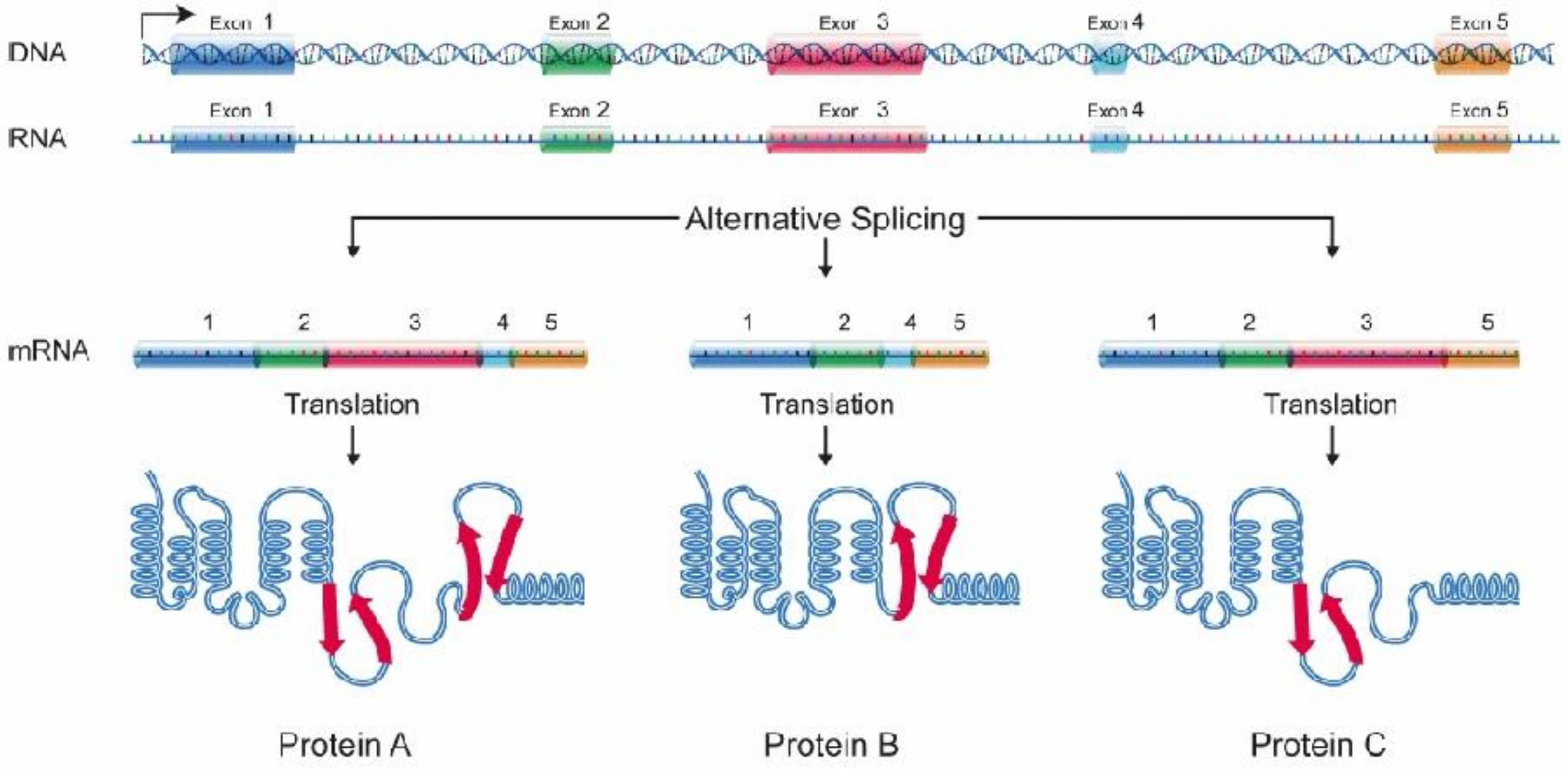
Transcription Termination



RNA Splicing: Introns, Exons

- For most eukaryotic genes (and some prokaryotic ones), the initial RNA also called **Precursor mRNA** (**pre-mRNA**) that is transcribed from a gene's DNA template must be processed before it becomes a **mature messenger RNA (mRNA)** that can direct the synthesis of protein.
- One of the steps in this processing, called **RNA splicing**, involves the removal or "splicing out" of certain sequences referred to as intervening sequences, or **introns**.
- The final mRNA thus consists of the remaining sequences, called **exons**, which are connected to one another through the splicing process.
- RNA splicing was initially discovered in the **1970s**, overturning years of thought in the field of gene expression.

Alternative Splicing



poly-A tail

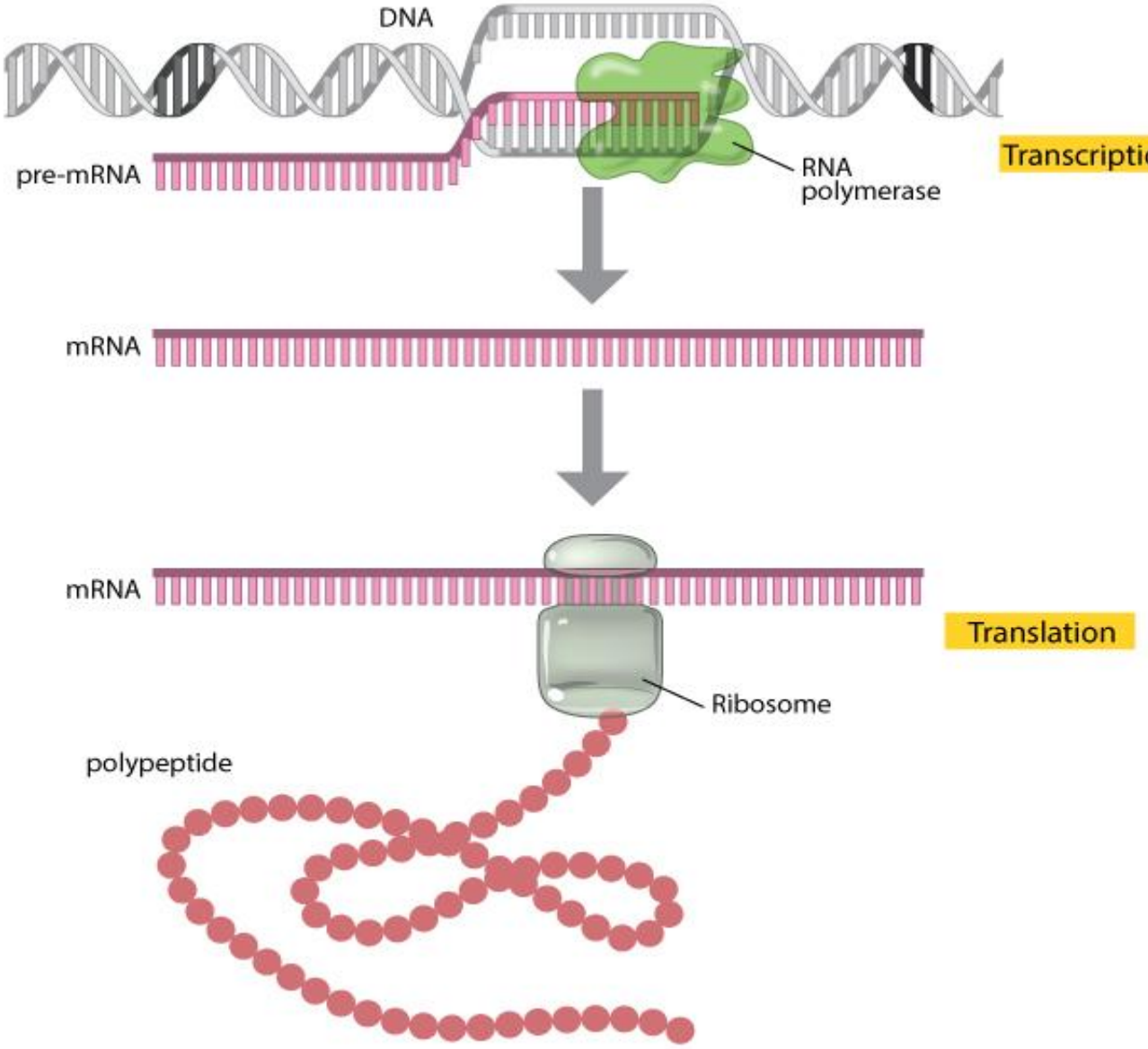
- The **poly-A tail** is a long **chain** of **adenine** nucleotides that is added to a messenger RNA (**mRNA**) molecule during RNA processing to increase the stability of the molecule.
- Immediately after a gene in a eukaryotic cell is transcribed, the new RNA molecule undergoes several modifications known as **RNA processing**. These modifications alter both ends of the primary RNA transcript to produce a mature mRNA molecule.
- The processing of the **3' end** adds a **poly-A tail** to the RNA molecule. First, the 3' end of the transcript is cleaved to free a 3' hydroxyl. Then an enzyme called **poly-A polymerase** adds a chain of adenine nucleotides to the RNA. This process, called **polyadenylation**, adds a poly-A tail that is between **100** and **250** residues long.
- The poly-A tail makes the RNA molecule more **stable** and **prevents** its **degradation**. Additionally, the poly-A tail **allows** the mature messenger RNA molecule to be **exported** from the nucleus and translated into a protein by ribosomes in the cytoplasm.

Translation: RNA to Protein

- During **translation**, which is the second major step in gene expression, the **mature mRNA** is "read" according to the **genetic code**, which relates the DNA sequence to the **amino acid** sequence in proteins (**Codon Table**).
- Each group of **three bases** in mature mRNA constitutes a **codon**, and each codon specifies a particular amino acid (hence, it is a triplet code).
- The mRNA sequence is thus used as a template to assemble—in order—the chain of amino acids that form a protein.
- Within all cells, the **translation machinery** resides within a specialized organelle called the **ribosome**. In eukaryotes, **mature mRNA** molecules must leave the nucleus and travel to the cytoplasm, where the ribosomes are located. On the other hand, in prokaryotic organisms, ribosomes can attach to mRNA while it is still being transcribed. In this situation, translation begins at the 5' end of the mRNA while the 3' end is still attached to DNA.

Translation Process

- During transcription, the enzyme **RNA polymerase** (green) uses DNA as a template to produce a **pre-mRNA** transcript (pink).
- The pre-mRNA is processed to form a **mature mRNA** molecule that can be translated to build the protein molecule (**polypeptide**) encoded by the original gene.



Translation Machinery: Ribosomes

Ribosomes are the structures where polypeptides (proteins) are built. They are made up of protein and RNA (**ribosomal RNA**, or **rRNA**). Each ribosome has **two subunits**, a large one and a small one, which come together around an mRNA—kind of like the two halves of a hamburger bun coming together around the patty.

The ribosome has **three binding sites**: an amino acid site (A), a polypeptide site (P), and an exit site (E). Which provides a set of handy slots where **tRNAs** can find their matching codons on the mRNA template and deliver their amino acids. Not only that, but the ribosome also acts as an **enzyme**, catalyzing the chemical reaction that links amino acids together to make a chain.

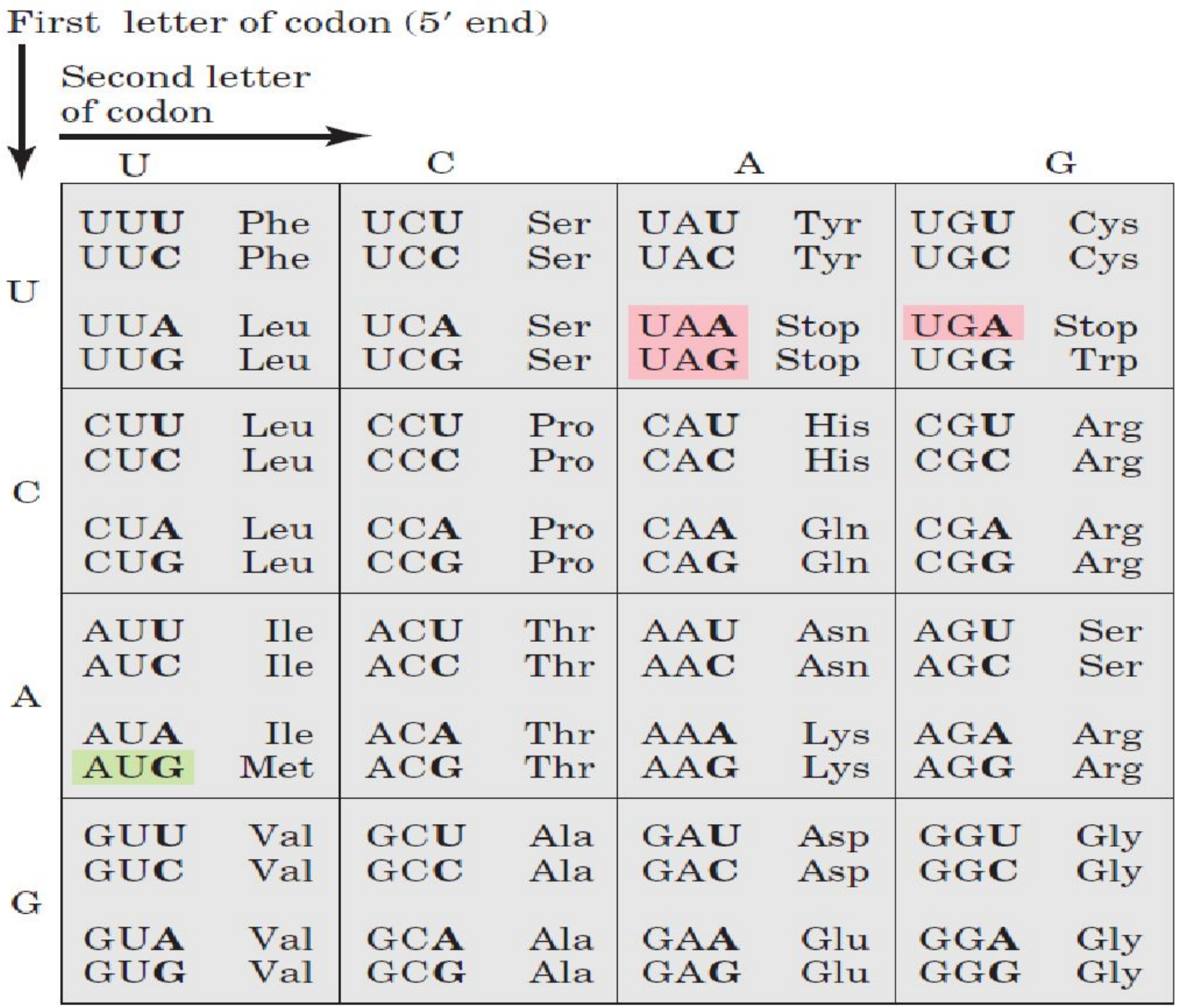
Translation Machinery: Transfer RNAs (tRNAs)

Transfer RNAs, or **tRNAs**, are molecular "bridges" that connect mRNA **codons** to the amino acids they encode. One end of each tRNA has a sequence of **three nucleotides** called an **anticodon**, which can bind to specific mRNA codons. The other end of the tRNA carries the amino acid specified by the codons.

There are many different types of tRNAs. Each type reads one or a few codons and brings the right amino acid matching those codons.

Codon Table

- The amino acids specified by each **mRNA codon**. Multiple codons can code for the same amino acid.
- The codons are written **5' to 3'**, as they appear in the mRNA.
- **AUG** is an **initiation codon**; **UAA**, **UAG**, and **UGA** are **termination** (stop) **codons**.



First letter of codon (5′ end)	Second letter of codon: U		Second letter of codon: C		Second letter of codon: A		Second letter of codon: G	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

FIGURE 27–7 "Dictionary" of amino acid code words in mRNAs.

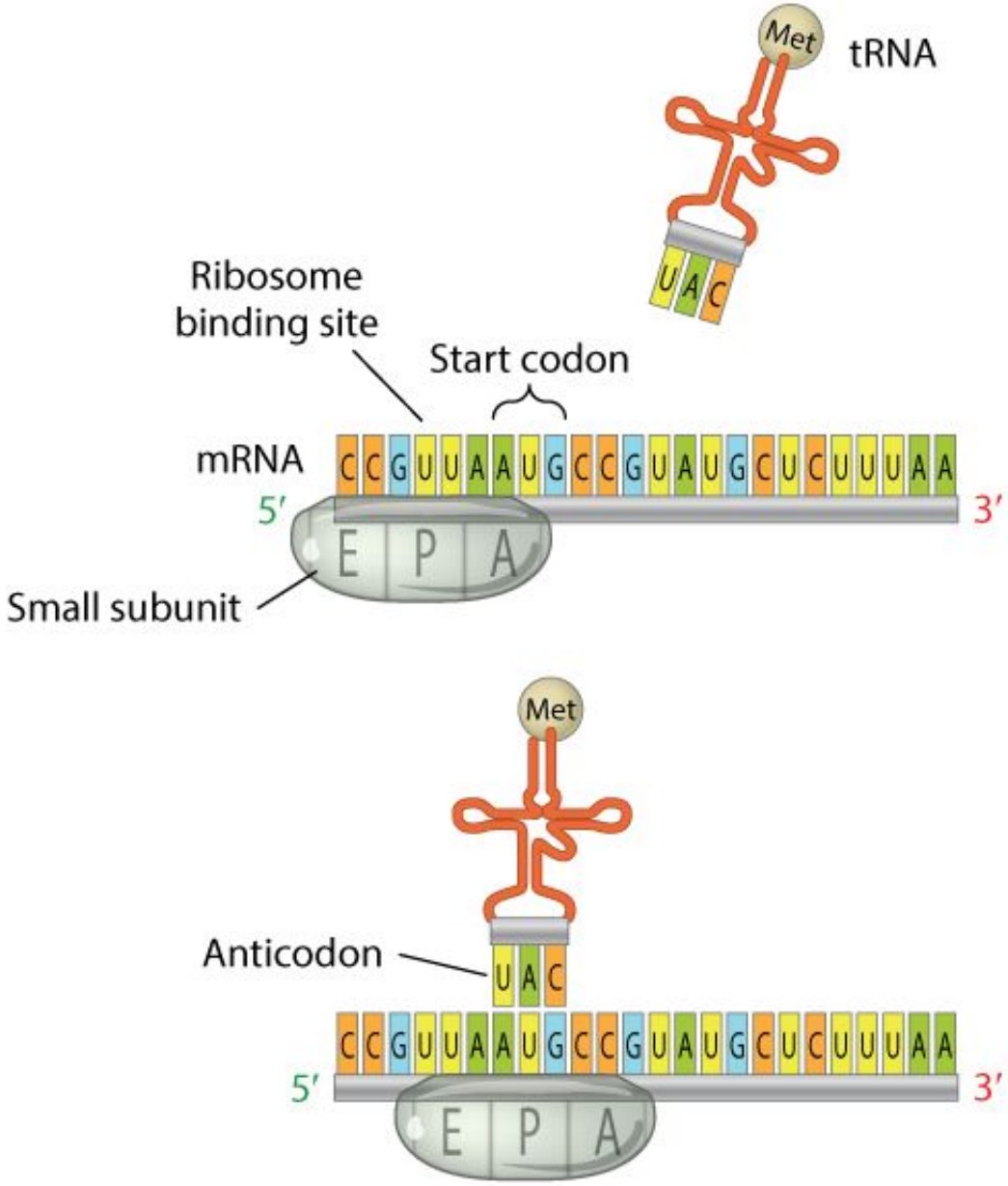
Translation Process: Three Stages

Translation process can be divided into **three** stages:

- **Initiation** (starting off): In this stage, the ribosome gets together with the mRNA and the first tRNA so translation can begin.
- **Elongation** (adding on to the protein chain): In this stage, amino acids are brought to the ribosome by tRNAs and linked together to form a chain.
- **Termination** (finishing up): In the last stage, the finished polypeptide is released to go and do its job in the cell.

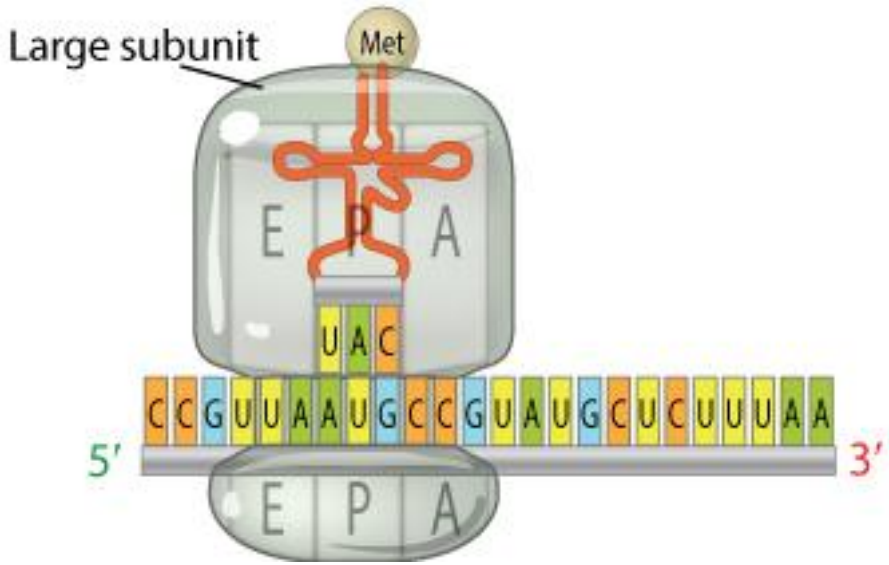
Translation Initiation Stage

- When translation begins, the small subunit of the **ribosome** and an initiator **tRNA** molecule assemble on the mRNA transcript.
- The small subunit of the ribosome has **three binding sites**: an amino acid site (A), a polypeptide site (P), and an exit site (E).
- The initiator tRNA molecule carrying the amino acid **methionine** binds to the **AUG** start codon of the mRNA transcript at the ribosome's **P site** where it will become the **first amino acid** incorporated into the growing polypeptide chain.
- Here, the initiator tRNA molecule is shown binding after the small ribosomal subunit has assembled on the mRNA; the order in which this occurs is unique to prokaryotic cells. In eukaryotes, the free initiator tRNA first binds the small ribosomal subunit to form a complex. The complex then binds the mRNA transcript, so that the tRNA and the small ribosomal subunit bind the mRNA simultaneously.



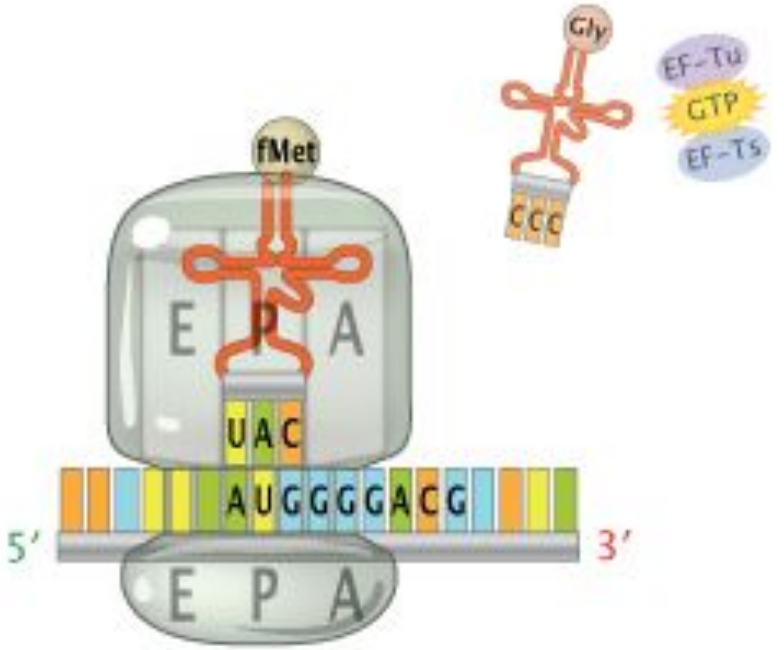
Translation Initiation Stage

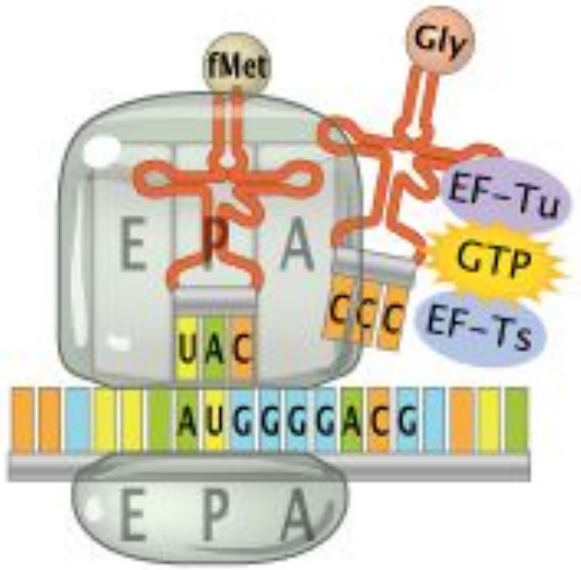
- Although methionine (**Met**) is the first amino acid incorporated into any new protein, it is **not** always the **first** amino acid in mature proteins—in many proteins, methionine is **removed** after translation.
- Once the **initiation complex** is formed on the mRNA, the large ribosomal subunit binds to this complex



Translation Elongation Stage

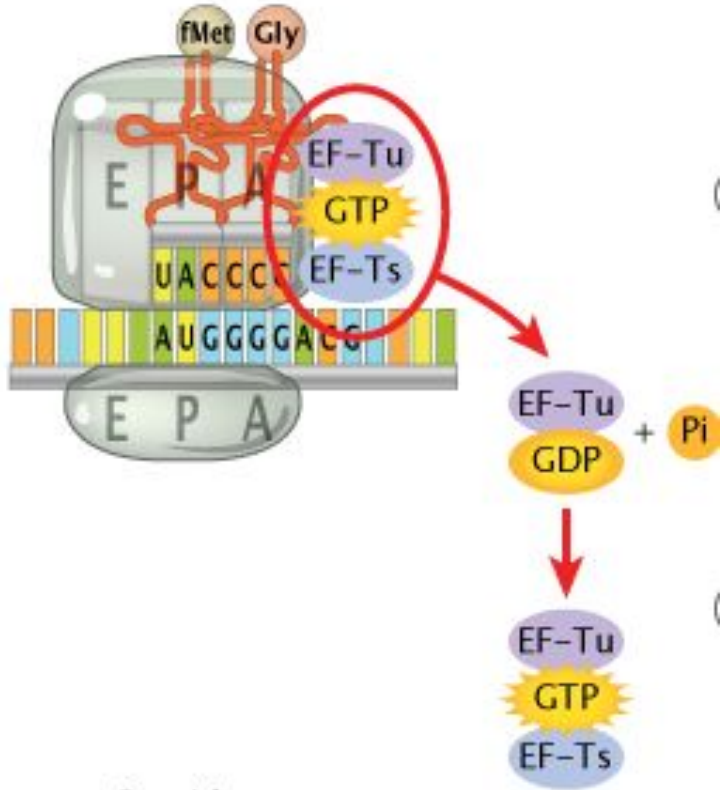
- At the beginning of elongation, an initiator tRNA molecule occupies the P site of a ribosome assembled on the mRNA transcript.
- This initiator tRNA carries the amino acid formylmethionine. The ribosome's A site is open and ready to receive a second, incoming tRNA molecule.

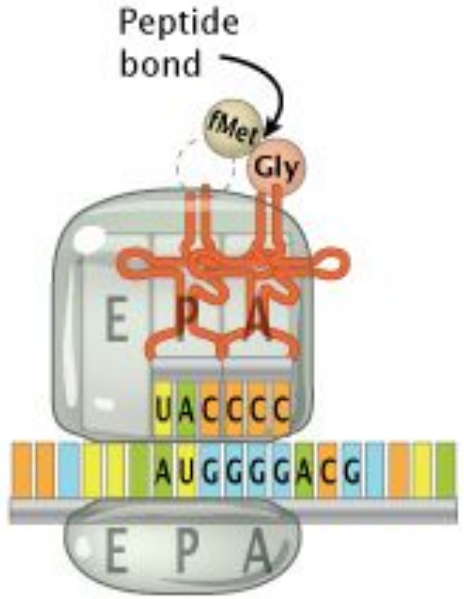




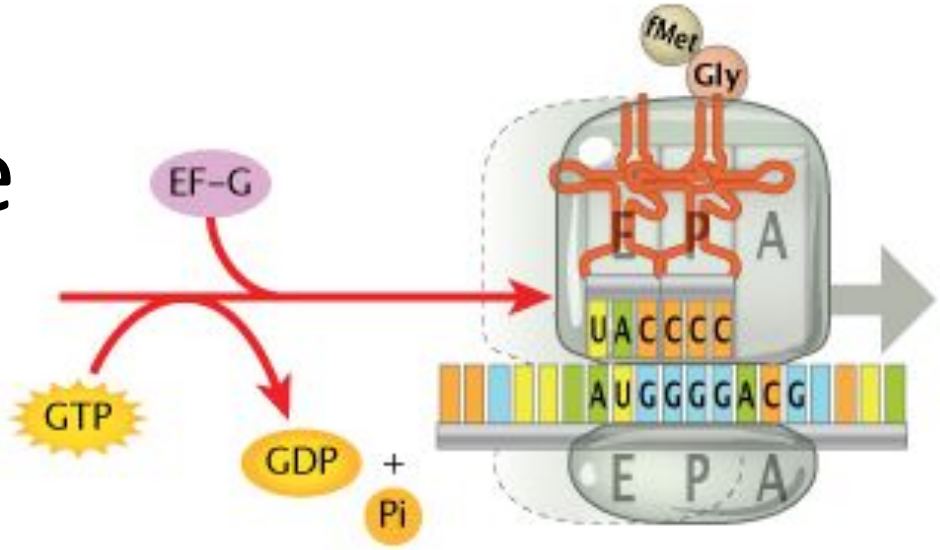
Translation Elongation Stage

- The amino acid bound to the tRNA that occupies the P site is added to the amino acid bound to the tRNA that occupies the A site, forming a growing peptide chain.
- As the ribosome moves from one codon to the next along the mRNA molecule, the tRNA molecule that occupies the A site is shifted to the P site.

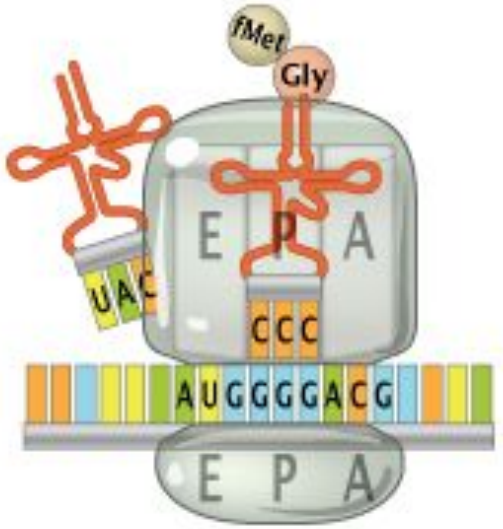




Translation Elongation Stage



- The A site therefore cycles between occupied and exposed states, and is able to receive the incoming tRNA molecule that corresponds to each sequential mRNA codon.
- The growing peptide chain is continuously transferred to the amino acid associated with the tRNA molecule located at the A site.



Translation Termination Stage

- There are three termination codons that are employed at the end of a protein-coding sequence in mRNA: **UAA**, **UAG**, and **UGA**. No tRNAs recognize these codons.
- Thus, in the place of these tRNAs, one of several proteins, called release factors, binds and facilitates release of the mRNA from the ribosome and subsequent dissociation of the ribosome.

Untranslated Region of mRNA (5' UTR & 3' UTR)

- Interestingly, not all regions of an mRNA molecule correspond to particular amino acids. In particular, there is an area near the **5' & 3 end** of the molecule that is known as the **untranslated region (UTR)**.
- Gene **expression** is finely **regulated** at the **post-transcriptional level**. Features of the untranslated regions of mRNAs that control their **translation**, **degradation** and **localization** include stem-loop structures, upstream initiation codons and open reading frames, internal ribosome entry sites and various *cis*-acting/*trans*-acting elements that are bound by RNA-binding proteins.
- So, what is the purpose of the **UTR**? It turns out that the leader sequence is important because it contains a **ribosome-binding site**.
- In bacteria, this site is known as the **Shine-Dalgarno box** (**AGGAGG**), after scientists John Shine and Lynn Dalgarno, who first characterized it.
- A similar site in vertebrates was characterized by Marilyn Kozak and is thus known as the **Kozak box**. In bacterial mRNA, the 5' UTR is normally short; in human mRNA, the median length of the 5' UTR is about 170 nucleotides. If the leader is long, it may contain regulatory sequences, including binding sites for proteins, that can affect the stability of the mRNA or the efficiency of its translation.

Some Terms

- **Coding Regions**:
  The **coding region** of a gene, also known as the **coding** sequence or CDS (from **coding** DNA sequence), is that portion of a gene's DNA or RNA, composed of exons, that codes for protein. The **region** is bounded nearer the 5' end by a start codon and nearer the 3' end with a stop codon.
- **Non-coding Regions**:
  Regions that does not codes for protein or **does not translated** to protein

Extension of Central Dogma

- The existence of **RNA replication** requires an elaboration of the central dogma (Fig. 26–28). The **enzymes** involved in RNA replication have profound implications for investigations into the nature of **self-replicating** molecules that may have existed in **prebiotic** times.
- **Reverse Transcriptase** Produces DNA from **Viral RNA**. Certain RNA viruses that infect animal cells carry within the viral particle an RNA-dependent DNA polymerase called **reverse transcriptase.** The RNA viruses that contain reverse transcriptases are known as **retroviruses**

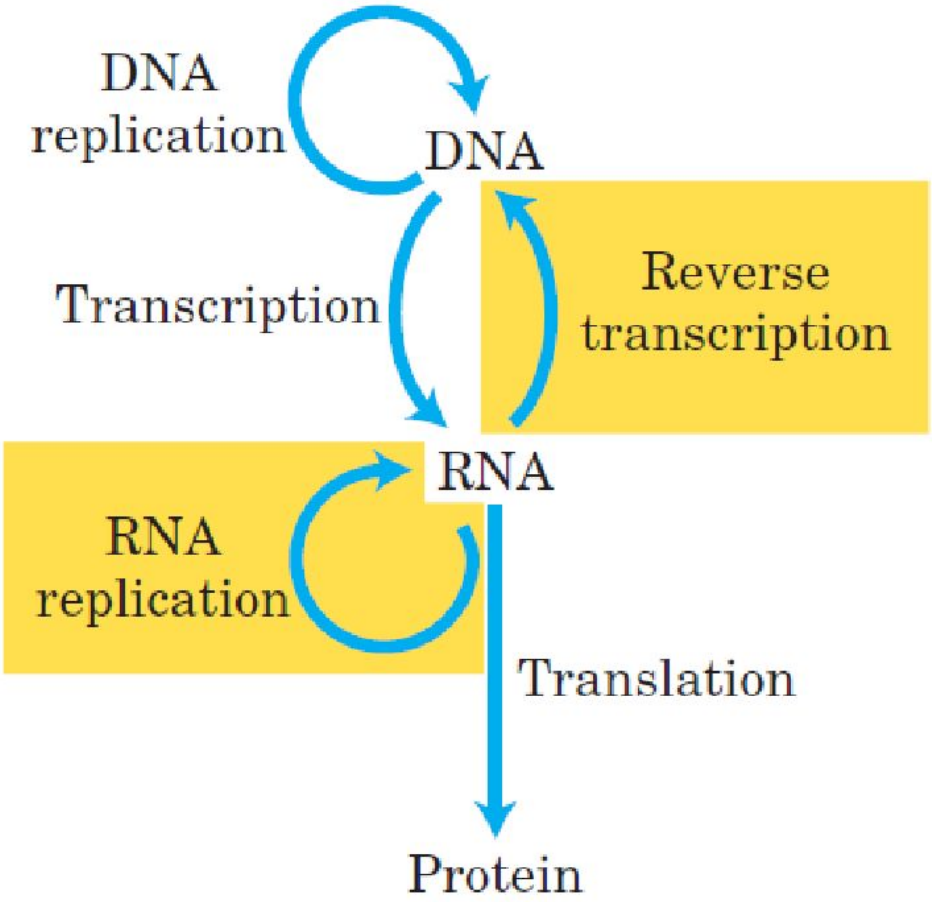


FIGURE 26–28 Extension of the central dogma to include RNA-dependent synthesis of RNA and DNA.