

# **IC251 – Basics of Bioinformatics (4 Credits)**

# IC251 – Basics of Bioinformatics (4 Credits)

## Recommended course texts

- ✓ Introduction to Bioinformatics by Teresa Attwood
- ✓ Essential Bioinformatics by Jin Xiong
- ✓ Bioinformatics: Sequence and Genome Analysis by David Mount
- ✓ Molecular Modelling: Principles and Applications by Andrew R. Leach
- ✓ Specific references cited in the slides.

# Evaluation method

COMPONENT	WEIGHTAGE
Tierce Exam	60 (2 exams)
Quizzes	30 (2×15)
Class Assessment (incl. Tutorials, Attendance)	10
Total	100%

- **Unmute and ask** – Any student can unmute in-between the class and ask questions.
- **Chat box** – Students having doubts can also post their questions in the chat box.
- **Assessment** – Any student may be asked question randomly in between the class.

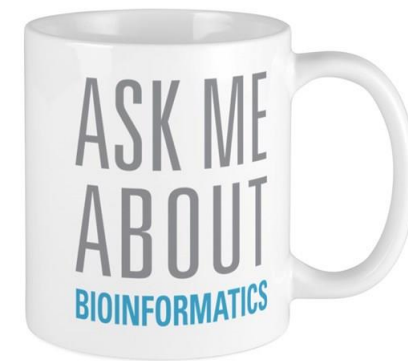
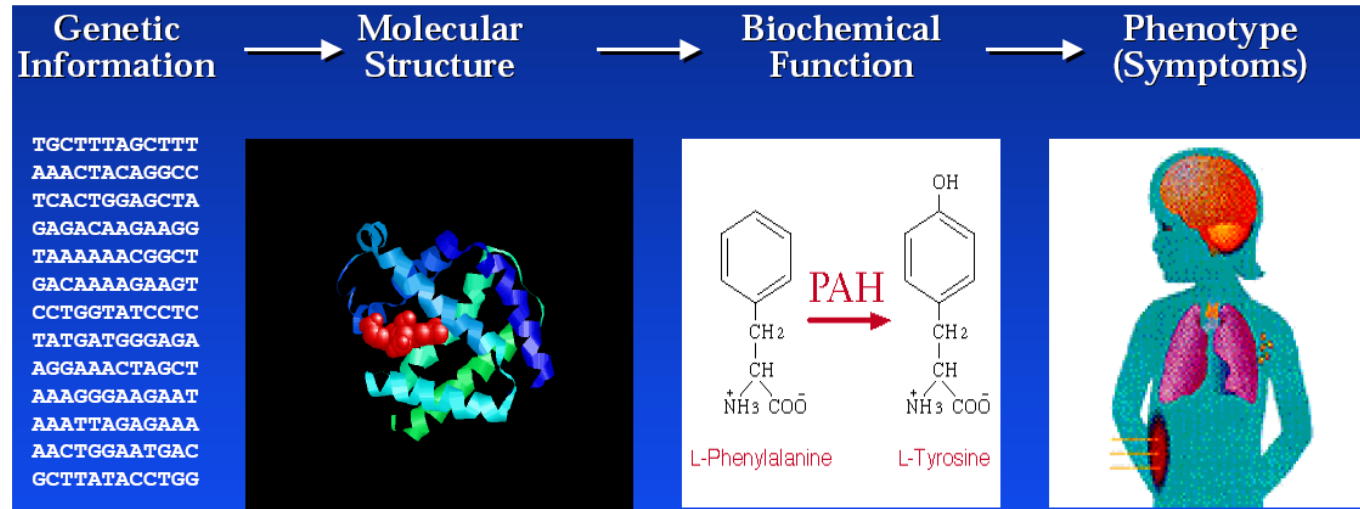
# Lecture

## Bioinformatics – General Introduction

Biology

Bioinformatics

Computer Science



# Lecture

## Bioinformatics – General Introduction

### Why required?

#### DNA

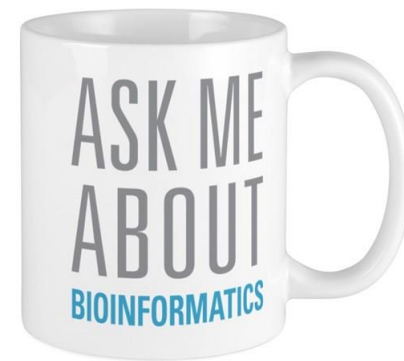
- ✓ a language over a four character alphabet, {A, C, G, T}

#### Protein

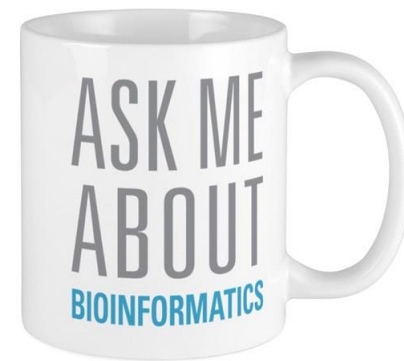
- ✓ a language over a twenty character alphabets
- ✓ 20 standard amino acids

#### Data Explosion

- ✓ Rapid growth in genetic data
- ✓ Rapid growth in structural data (protein, DNA and RNA)



# Lecture



# Biomolecules

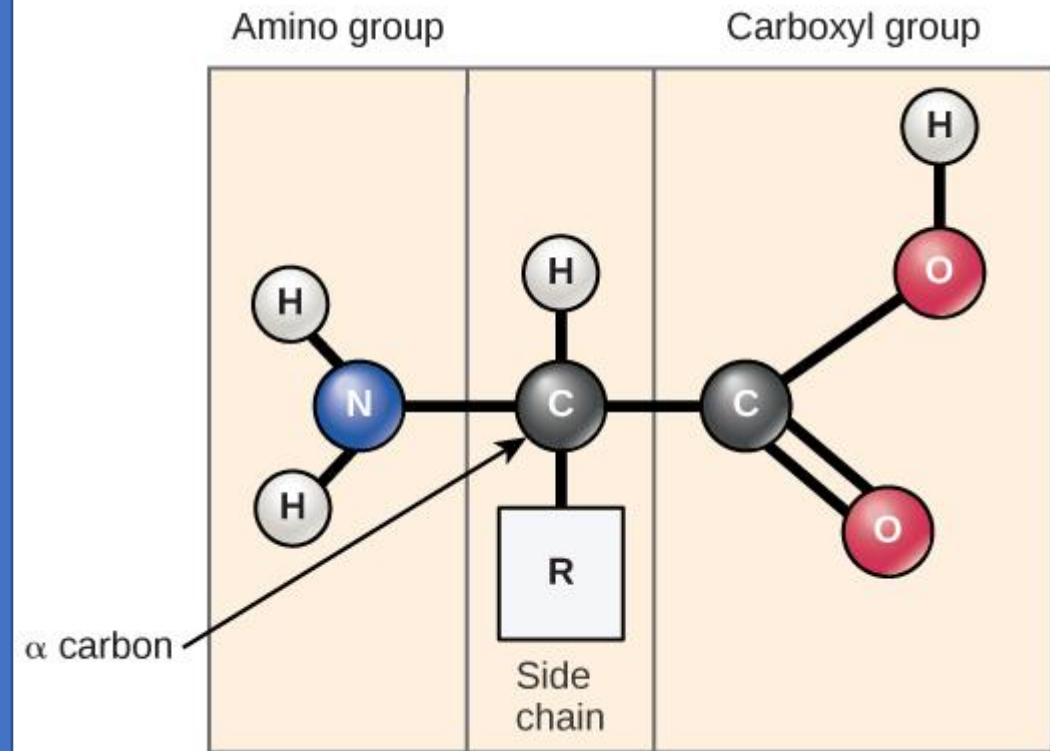
- Biomolecules, **range from small molecules**, such as **metabolites**, to **large molecules**, such as **protein** and carbohydrates, which are chemical compounds produced by living organisms.
- These **biomolecules** are **fundamental building blocks** of living organisms, and therefore, the presence and appropriate concentrations of biomolecules **are vital** for the **structure and proper function of living cells**.

- Biomolecule, also called biological molecule, any of numerous substances that are **produced by cells** and **living organisms**.
- Biomolecules have a wide range of sizes and structures and perform a **vast array of functions**.
- The four major types of biomolecules are **carbohydrates, lipids, nucleic acids, and proteins**.

# Amino acids

## Structure

- Amino acids are the **monomers** that make up proteins.
- Each amino acid has the **same fundamental structure**, which consists of a **central carbon** atom, also known as the alpha ( $\alpha$ ) **carbon**, bonded to an amino group (**NH<sub>2</sub>**), a carboxyl group (**COOH**), and to a **hydrogen** atom.
- In the **aqueous environment** of the cell, both the amino group and the carboxyl group are **ionized under physiological** conditions (pH 7), and so have the structures - **NH<sub>3</sub><sup>+</sup>** and -**COO<sup>-</sup>**, respectively.
- Every amino acid also has another atom or group of atoms bonded to the central atom known as the **R group**.
- This **R group, or side chain**, gives each amino acid proteins **specific characteristics**, including size, polarity, and pH.





# Amino acids

Types ↓

## Structure

- The name “**amino acid**” is derived from the amino group and carboxyl-acid-group in their basic structure.
- There are **20 common amino acids** present in proteins, each with a specific R group or side chain.
- Ten of these are considered essential** amino acids in humans because the human body cannot produce them and they must be obtained from the diet.
- All **organisms have different essential** amino acids based on their physiology.

## Note:

The amino acids **vary in the composition of the side chain R group** that determines its chemical nature.

Source: <https://courses.lumenlearning.com/introchem/chapter/amino-acids/>

AMINO ACID		
Nonpolar, aliphatic R groups		Glycine
		Alanine
		Valine
Nonpolar, aliphatic R groups		Leucine
		Methionine
		Isoleucine
Polar, uncharged R groups		Serine
		Threonine
		Cysteine
Polar, uncharged R groups		Proline
		Asparagine
		Glutamine
AMINO ACID		
Positively charged R groups		Lysine
		Arginine
		Histidine
Negatively charged R groups		Aspartate
		Glutamate
Nonpolar, aromatic R groups		Phenylalanine
		Tyrosine
		Tryptophan

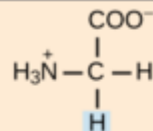
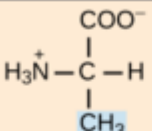
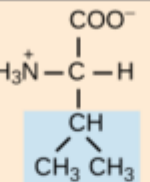
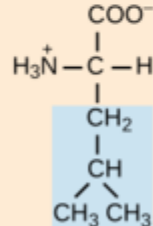
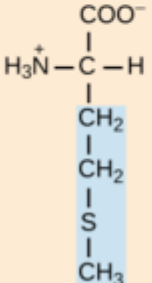
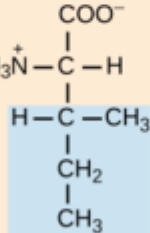
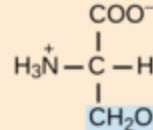
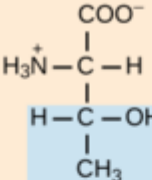
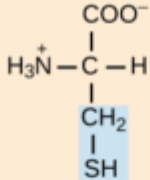
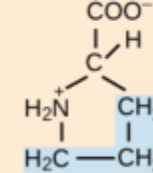
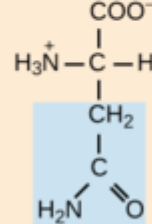
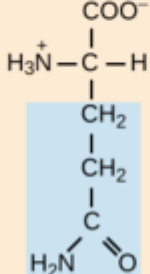
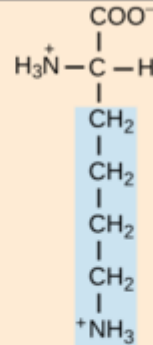
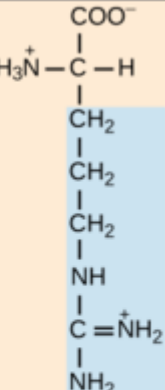
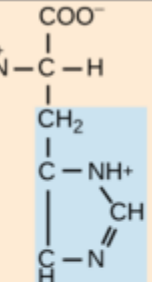
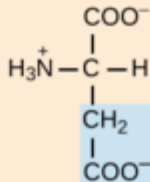
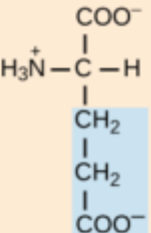
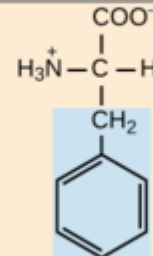
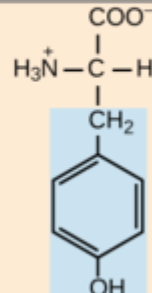
# Amino acids

## Types

### Features

- The chemical composition of the **side chain** **determines** the characteristics of the amino acid.
- Amino acids such as valine, methionine, and alanine are **nonpolar (hydrophobic)**.
- While amino acids such as serine, threonine, and cysteine are **polar (hydrophilic)**.
- The side chains of **lysine** and **arginine** are **positively charged** so these amino acids are also known as **basic (high pH)** amino acids.
- Proline** is an **exception** to the standard structure of an amino acid because its **R group is linked to the amino group**, forming a ring-like structure.
- Amino acids are represented by a **single upper case letter** or a **three-letter abbreviation**.
- For example, **valine** is known by the letter **V** or the three-letter symbol **VAL**.

Source: <https://courses.lumenlearning.com/introchem/chapter/amino-acids/>

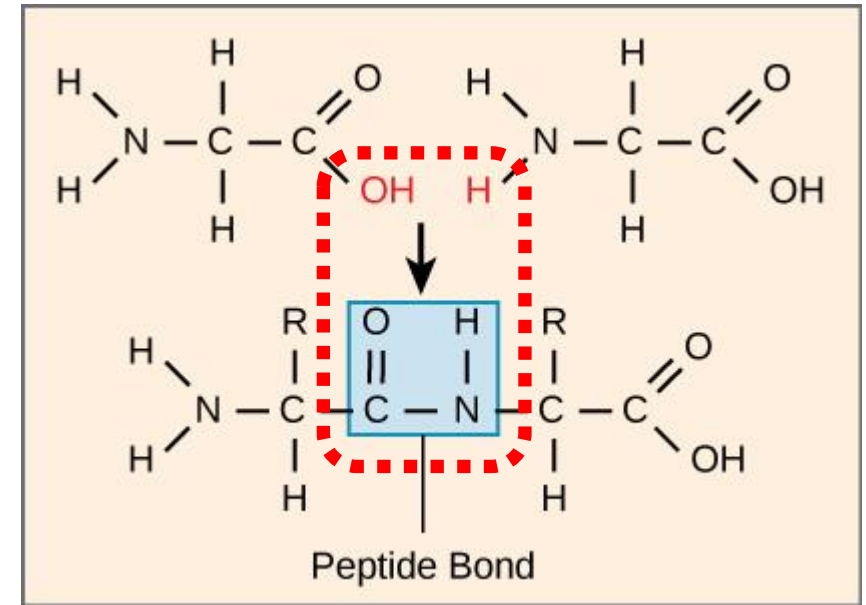
AMINO ACID				
Nonpolar, aliphatic R groups	 Glycine	 Alanine	 Valine	
	 Leucine	 Methionine	 Isoleucine	
	 Serine	 Threonine	 Cysteine	
Polar, uncharged R groups	 Proline	 Asparagine	 Glutamine	
	Positively charged R groups	 Lysine	 Arginine	 Histidine
		Negatively charged R groups	 Aspartate	 Glutamate
Nonpolar, aromatic R groups			 Phenylalanine	 Tyrosine

# Amino acids

## Structure

- The **sequence and the number of amino acids** ultimately determine the **protein's shape, size, and function**.
- Each amino acid is attached to another amino acid by a **covalent bond, known as a peptide bond**.
- When two amino acids are covalently attached by a peptide bond, the **carboxyl group of one amino acid and the amino group of the incoming amino acid combine** and release a molecule of water.
- Any reaction that combines two monomers in a reaction that **generates  $H_2O$**  as one of the products is known as a dehydration reaction, so peptide bond formation is an example of a **dehydration reaction**.

## Peptide bond



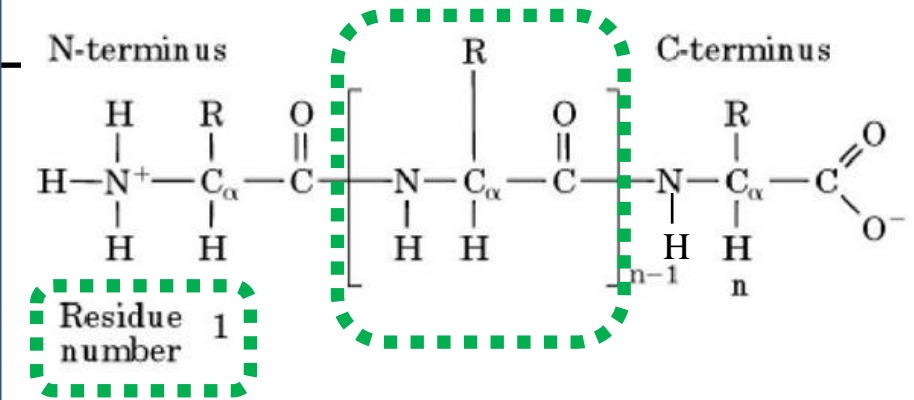
# Amino acids

# Polypeptide →

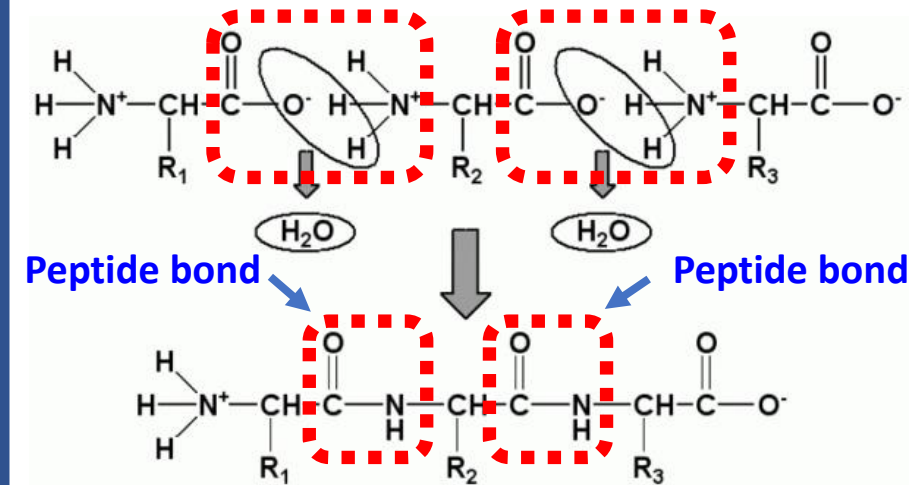
## Polypeptide chain

- The **chain of amino acids** is called a polypeptide chain.
- Each polypeptide has a **free amino group at one end**. This end is called the **N terminal**, or the amino terminal, and the other end has a **free carboxyl group**, also known as the C or **carboxyl terminal**.
- When reading or reporting the amino acid sequence of a protein or polypeptide, the **convention is to use the N-to-C direction**.
- That is, the **first amino acid** in the sequence is assumed to be the one at **the N terminal** and the **last amino acid** is assumed to be the one at the **C terminal**.
- Although the terms polypeptide and protein are sometimes used interchangeably, a **polypeptide is technically any polymer of amino acids**, whereas the **term protein** is used for a polypeptide or polypeptides that are **folded properly**, combined with any additional components needed **for proper functioning**, and is now functional.

Source: <https://courses.lumenlearning.com/introchem/chapter/amino-acids/>



Source: <https://what-when-how.com/molecular-biology/polypeptide-chain-molecular-biology/>



Source: <https://www.ncbi.nlm.nih.gov/books/NBK6824/figure/A149/>



# Amino acids

## Important notes

- ✓ **Polypeptide:** Any **polymer of** (same or different) **amino acids** joined via peptide bonds.
- ✓ **R group:** The R group is a **side chain specific to each amino acid** that confers particular chemical properties to that amino acid.
- ✓ **Amino acid:** Any of 20 naturally occurring  $\alpha$ -amino acids (having the **amino, and carboxylic acid groups** on the same carbon atom), and a variety of side chains, that combine, via peptide bonds, to form proteins.

## Takeaway

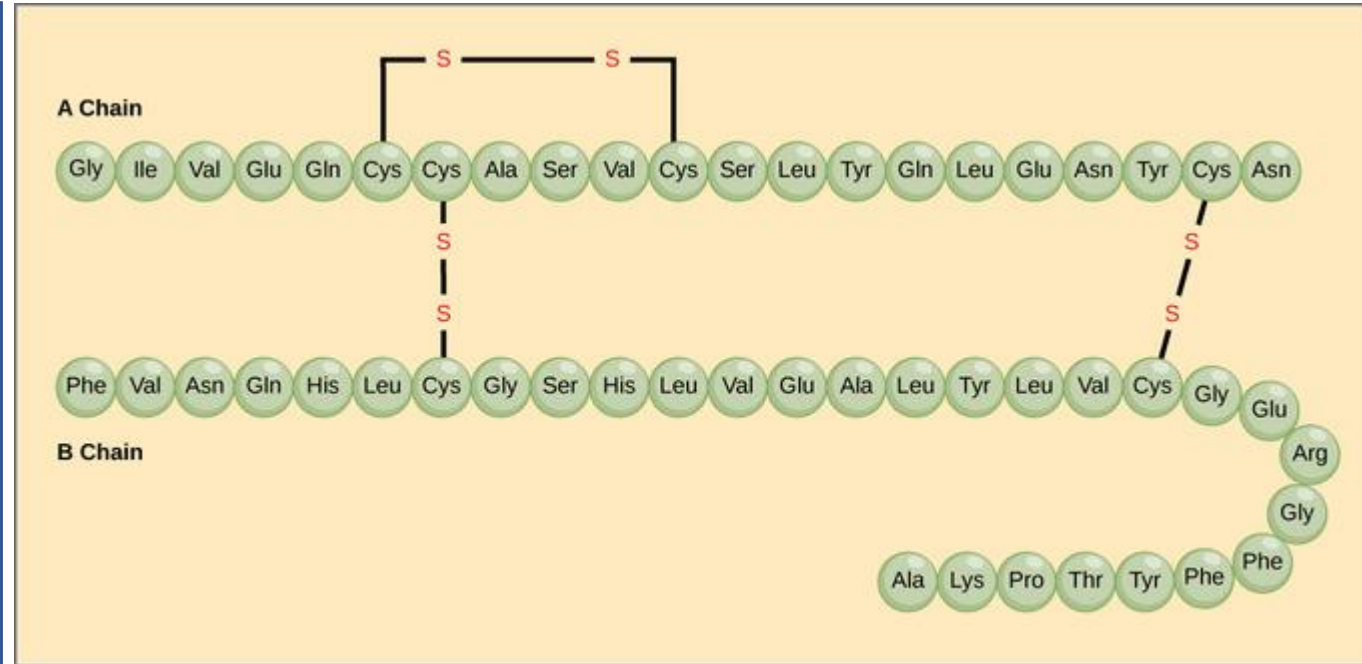
- ✓ Amino acids are the **building blocks** of protein.
- ✓ Each amino acid contains a central C atom, an amino group (**NH<sub>2</sub>**), a carboxyl group (**COOH**), and a specific **R group**.
- ✓ The **R group determines the characteristics** (size, polarity, and pH) for each type of amino acid.
- ✓ **Peptide bonds** form between the carboxyl group of one amino acid and the amino group of another through dehydration synthesis.
- ✓ A chain of amino acids is a **polypeptide**.



# Proteins

## Primary Structure

- A protein's primary structure is the **unique sequence of amino acids** in each polypeptide chain that makes up the protein.
- Really, this is just a list of which amino acids appear in which **order in a polypeptide chain**, **not really a structure**.
- But, because the final protein structure ultimately **depends on this sequence**, this was called the primary structure of the polypeptide chain.
- For example, the **pancreatic hormone insulin** has two polypeptide chains, A and B.



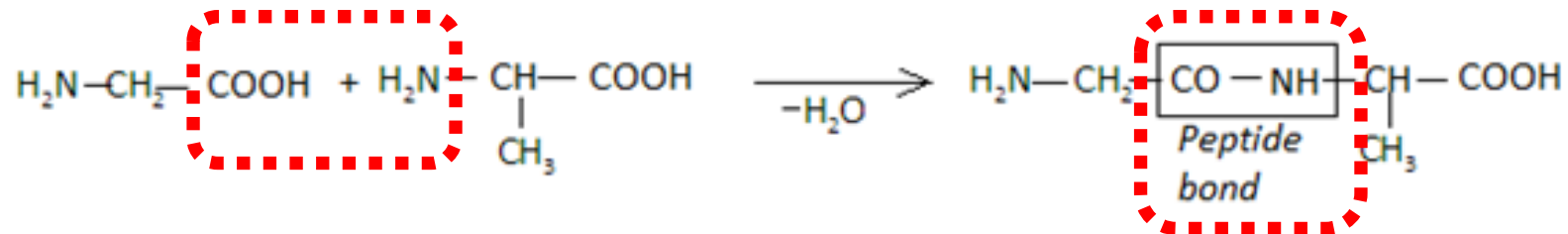
**Primary structure:** The **A chain of insulin is 21** amino acids long and the **B chain is 30** amino acids long, **and each sequence is unique** to the insulin protein.

# Proteins

## Primary Structure

Contd...

### Peptide bond formation

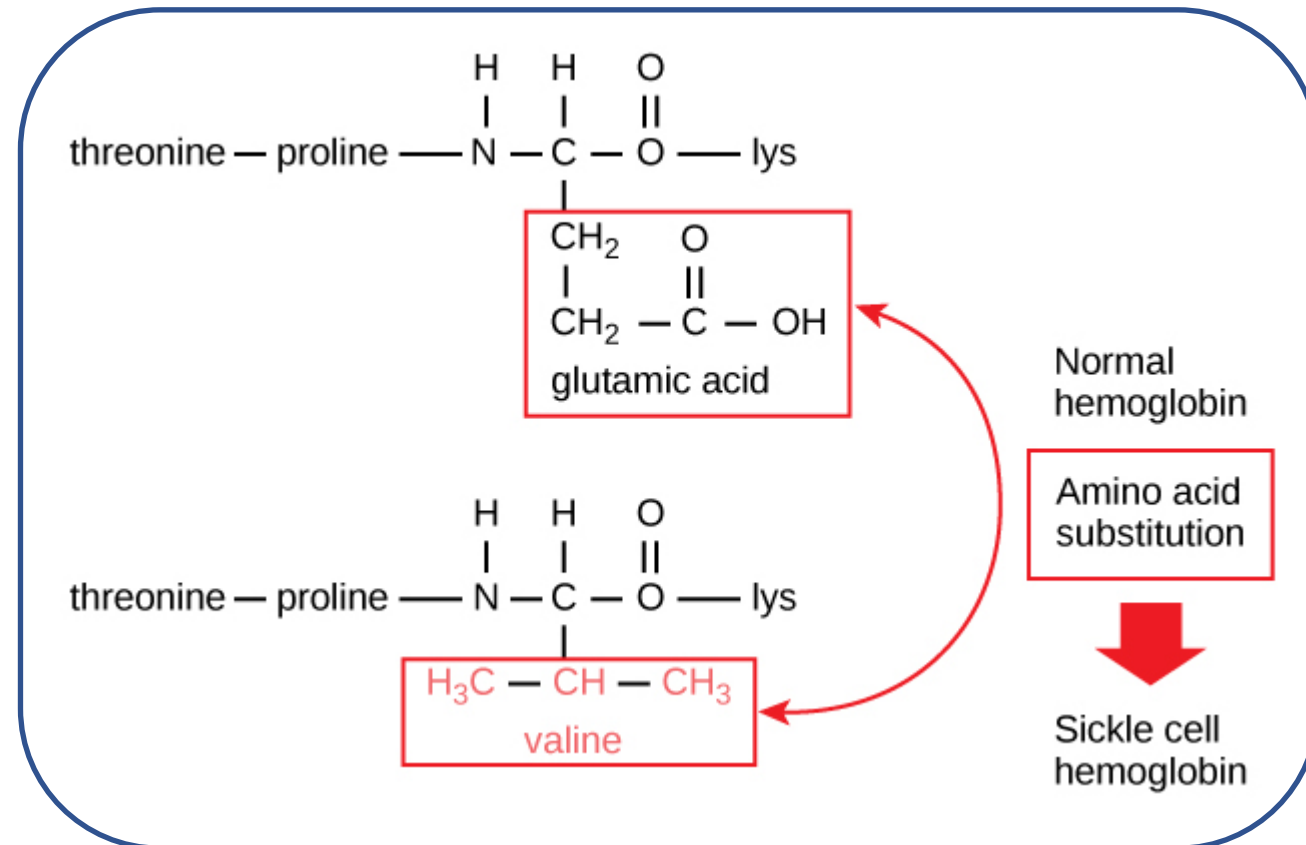


# Proteins

## Primary Structure

- The **gene, or sequence of DNA**, ultimately determines the unique sequence of amino acids in each peptide chain.
- A **change in nucleotide sequence** of the gene's coding region may lead to a **different amino acid** being added to the growing polypeptide chain, causing a **change in protein structure** and therefore function.
- Even **changing just one amino acid** in a protein's sequence can affect the protein's overall structure and function.
- For instance, a **single amino acid change** is associated with **sickle cell anemia**, an inherited disease that affects red blood cells.

Contd...



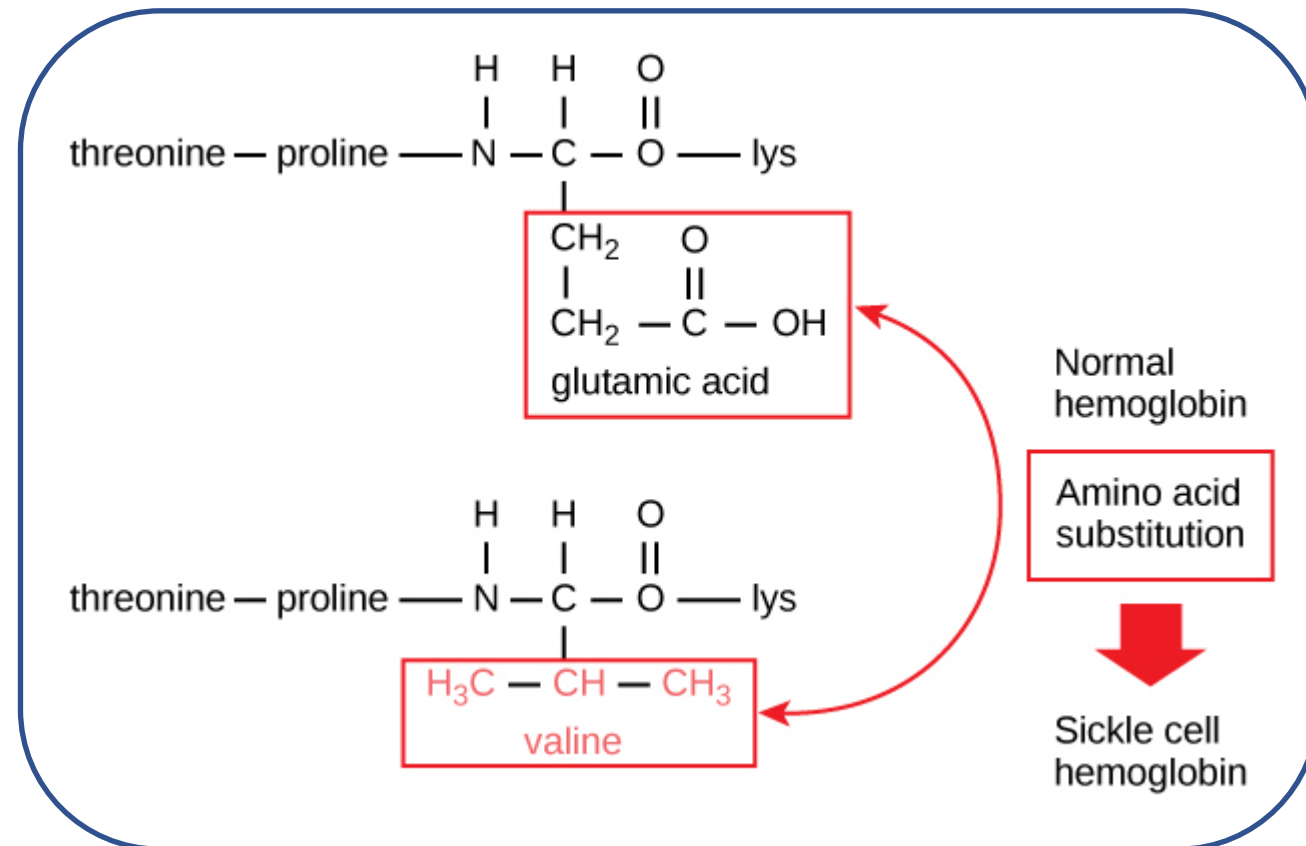


# Proteins

## Primary Structure

- In sickle cell anemia, one of the polypeptide chains that make up hemoglobin, the protein that carries oxygen in the blood, has a **slight sequence change**.
- The **glutamic acid** that is normally the sixth amino acid of the hemoglobin **β chain** (one of two types of protein chains that make up hemoglobin) is **replaced by a valine**.
- This substitution is shown for a fragment of the β chain in the diagram (right side).

Contd...

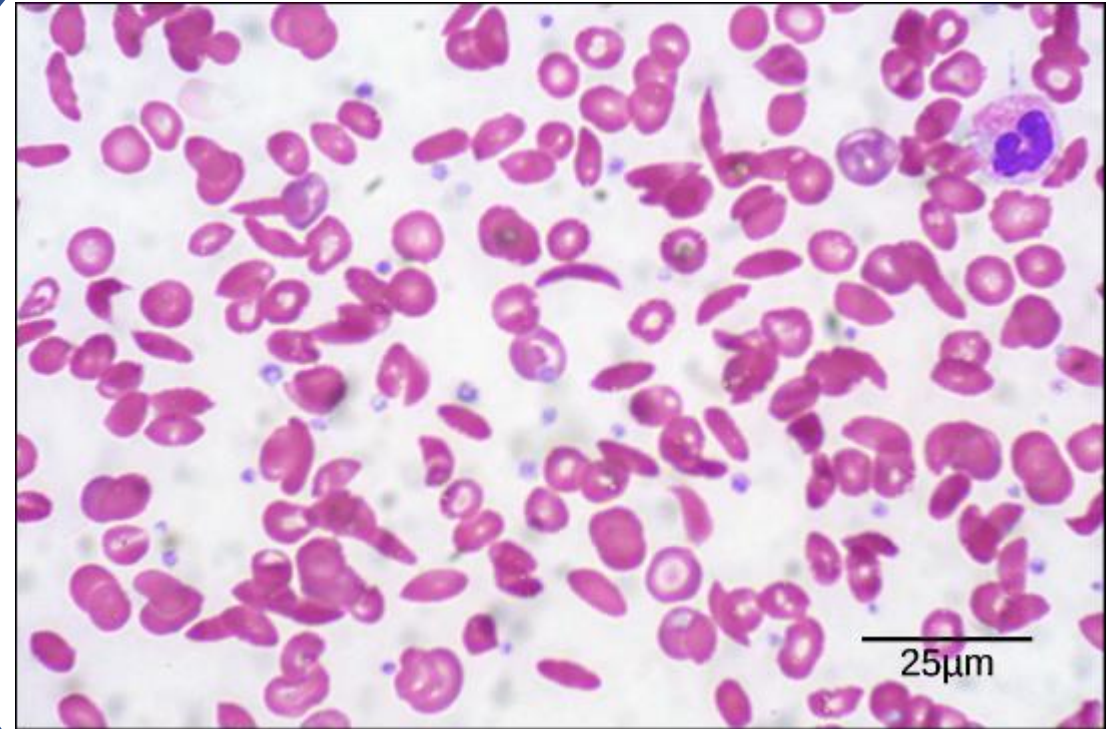


# Proteins

## Primary Structure

- The most remarkable thing is that a hemoglobin molecule is made up of **two  $\alpha$  chains and two  $\beta$  chains**, each consisting of about **150 amino acids**, for a total of about **600 amino acids** in the whole protein.
- The **difference** between a normal hemoglobin molecule and a sickle cell molecule is just **2 amino acids out of the approximately 600**.
- A person whose body makes only sickle cell hemoglobin will suffer symptoms of **sickle cell anemia**.
- These occur because the **glutamic acid-to-valine** amino acid change makes the hemoglobin molecules **assemble into long fibers**.
- The fibers **distort disc-shaped** red blood cells **into crescent shapes**.

Contd...



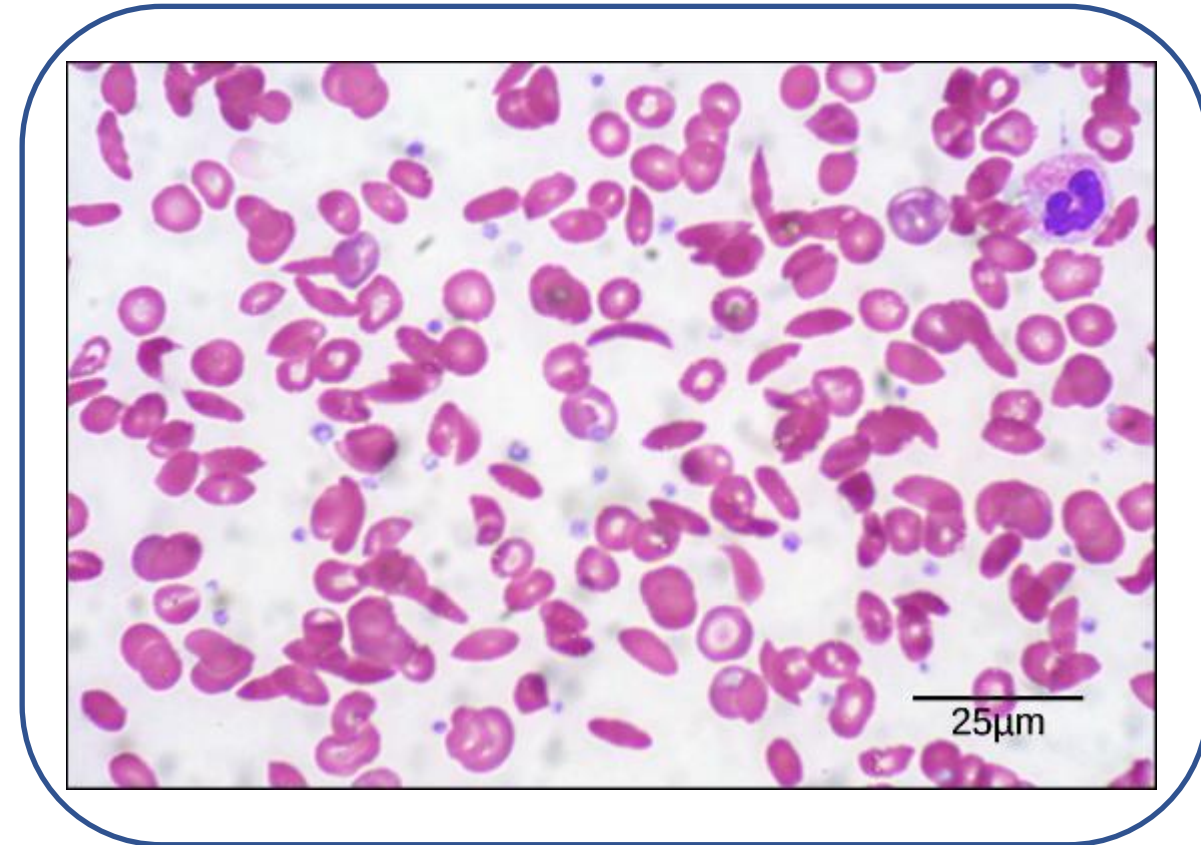
Examples of “sickled” cells can be seen mixed with normal, disc-like cells in the blood sample.

# Proteins

## Primary Structure

- The **sickled cells get stuck** as they try to pass through blood vessels.
- The stuck cells **impair blood flow** and can cause serious health problems for people with sickle cell anemia, including **breathlessness, dizziness, headaches, and abdominal pain.**

Contd...



Examples of “sickled” cells can be seen mixed with normal, disc-like cells in the blood sample.

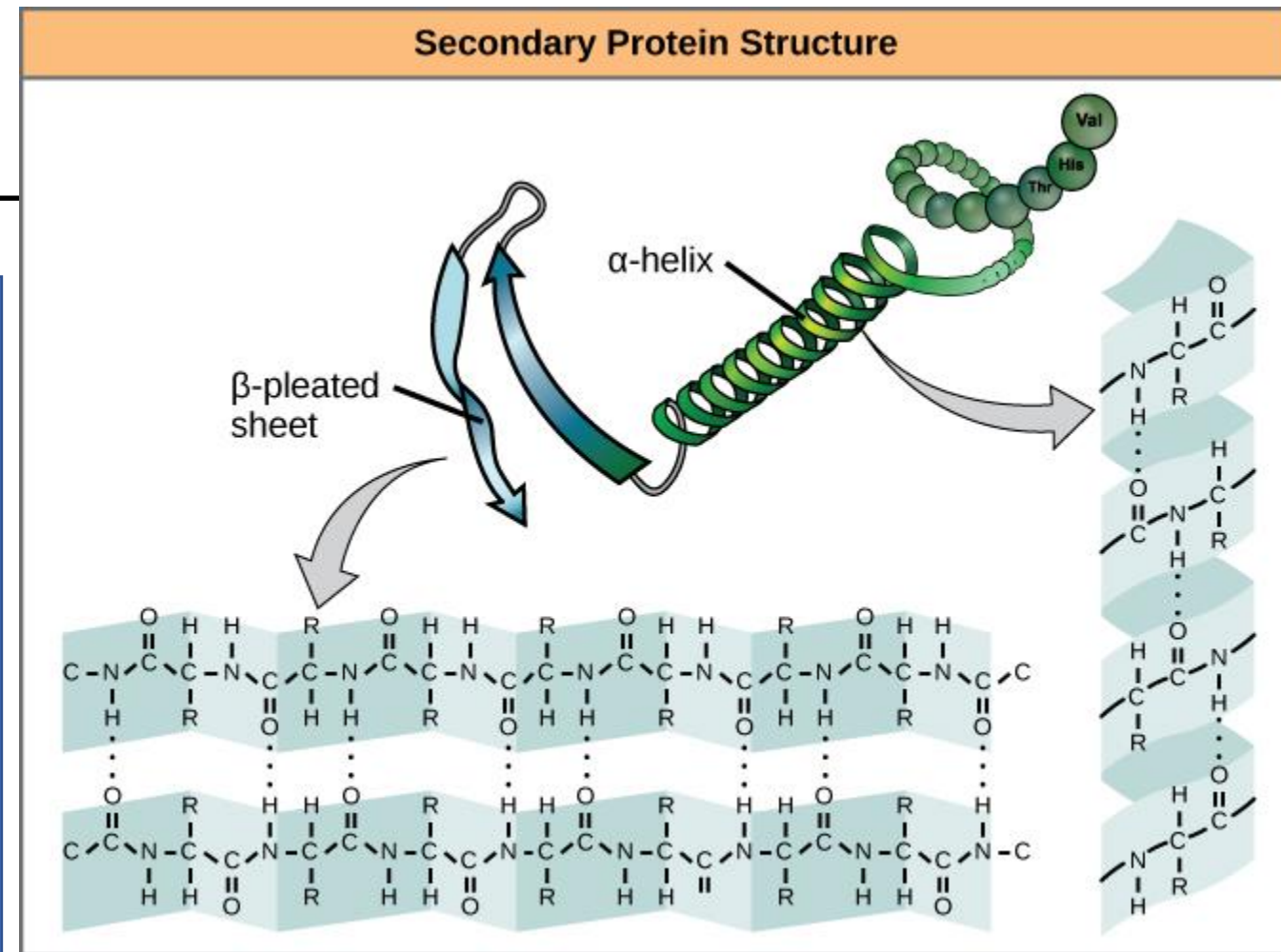
29-09-2021



# Proteins

## Secondary Structure

- A protein's secondary structure is whatever regular structures **arise from interactions between neighboring or near-by amino acids** as the polypeptide starts to **fold** into its functional three-dimensional form.
- Secondary structures **arise as H bonds** form between local groups of amino acids in a region of the polypeptide chain.
- **Rarely does a single secondary structure extend** throughout the polypeptide chain.
- It is usually just in a section of the chain.
- The most common forms of secondary structure are  **$\alpha$ -helix** and  **$\beta$ -pleated sheet** and they play an important structural role in most globular and fibrous proteins.



- The  $\alpha$ -helix and  $\beta$ -pleated sheet form because of **hydrogen bonding between carbonyl and amino groups in the peptide backbone**.
- Certain amino acids have a **propensity** to form an  **$\alpha$ -helix**, while others have a propensity to form a  **$\beta$ -pleated sheet**.

Source:

<https://courses.lumenlearning.com/introchem/chapter/protein-structure/#:~:text=Primary%20structure%20is%20the%20amino,by%20interactions%20between%20R%20groups.>

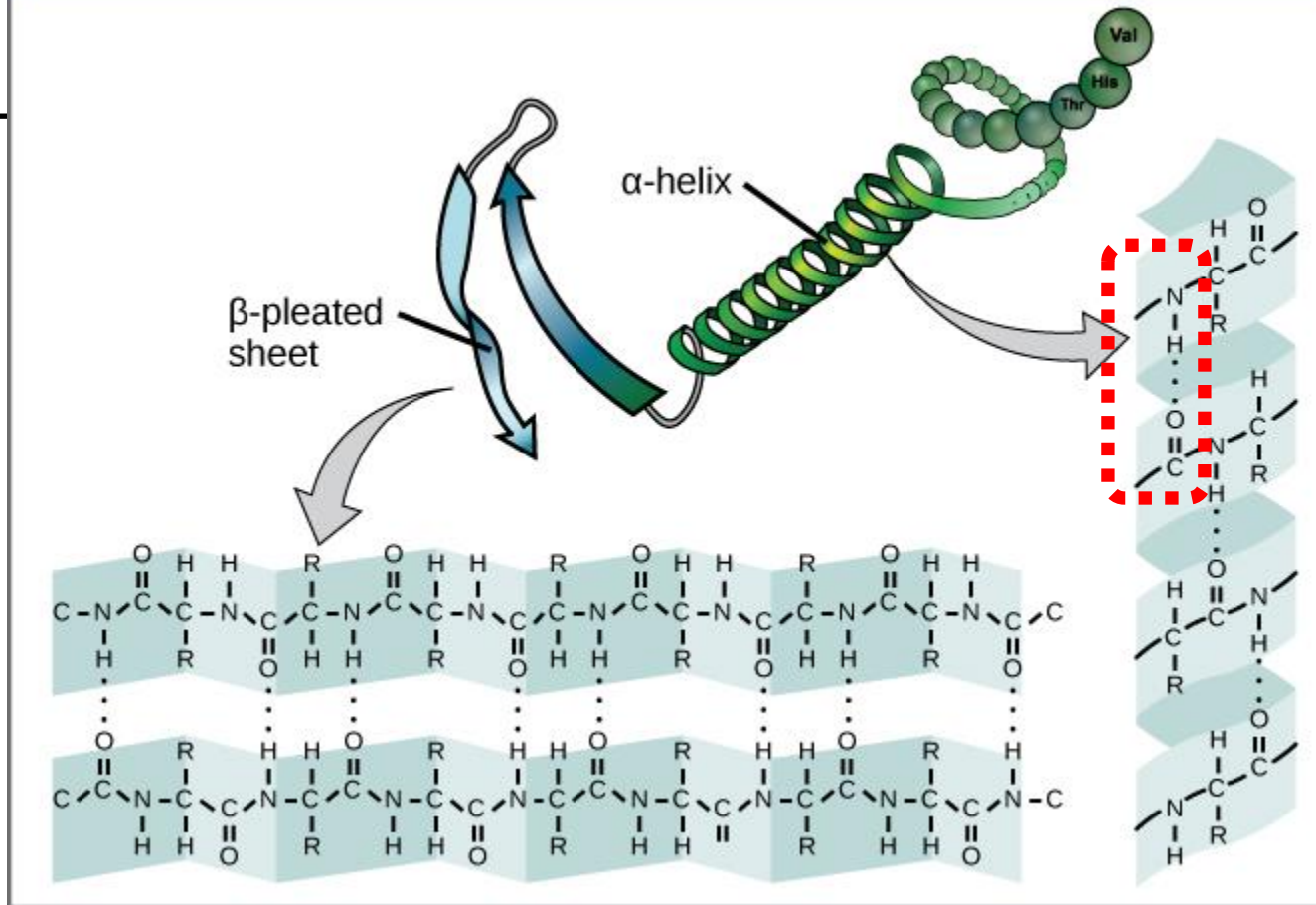
# Proteins

Contd...

## Secondary Structure $\alpha$ -helix chain

- In the  $\alpha$ -helix chain, the **hydrogen bond** forms between the **oxygen atom** in the polypeptide backbone **carbonyl group in one** amino acid and the **hydrogen atom** in the polypeptide backbone **amino group of another** amino acid that is **four amino acids farther** along the chain.
- This holds the stretch of amino acids in a **right-handed coil**.
- Every helical turn in an alpha helix has **3.6 amino acid** residues.
- The **R groups** (the side chains) of the polypeptide **protrude out** from the  $\alpha$ -helix chain and are **not involved in the H bonds** that maintain the  $\alpha$ -helix structure.

## Secondary Protein Structure



- The  $\alpha$ -helix and  $\beta$ -pleated sheet form because of **hydrogen bonding between carbonyl and amino groups in the peptide backbone**.
- Certain amino acids have a **propensity** to form an  **$\alpha$ -helix**, while others have a propensity to form a  **$\beta$ -pleated sheet**.

Source:

<https://courses.lumenlearning.com/introchem/chapter/protein-structure/#:~:text=Primary%20structure%20is%20the%20amino,by%20interactions%20between%20R%20groups.>

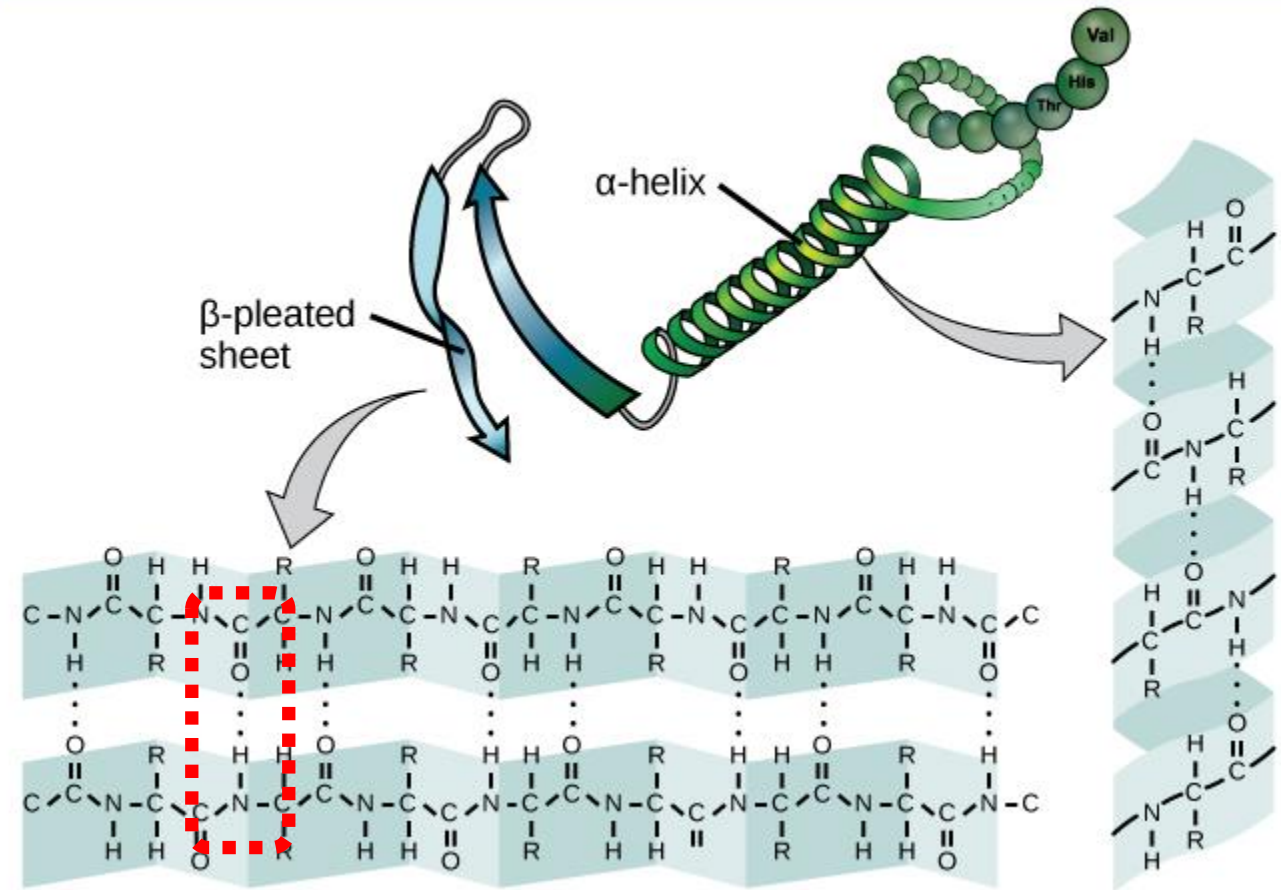
# Proteins

Contd...

## Secondary Structure $\beta$ -pleated sheets

- In  $\beta$ -pleated sheets, stretches of amino acids are held in an almost **fully-extended conformation** that “pleats” or zig-zags due to the non-linear nature of single C-C and C-N covalent bonds.
- $\beta$ -pleated sheets **never occur alone**. They are held in place by other  $\beta$ -sheets.
- The stretches of amino acids in  $\beta$ -pleated sheets are held in their pleated sheet structure because **hydrogen bonds** form between the **oxygen** atom in a polypeptide backbone **carbonyl group** of one  $\beta$ -pleated sheet and the **hydrogen** atom in a polypeptide backbone **amino group** of another  $\beta$ -pleated sheet.
- The  $\beta$ -pleated sheets which hold each other together align **parallel or antiparallel** to each other.

## Secondary Protein Structure



- ✓ The **R groups** of the amino acids in a  $\beta$ -pleated sheet **point out perpendicular to the hydrogen bonds** holding the sheets together, and are **not involved** in maintaining the  $\beta$ -sheet structure.

Source:

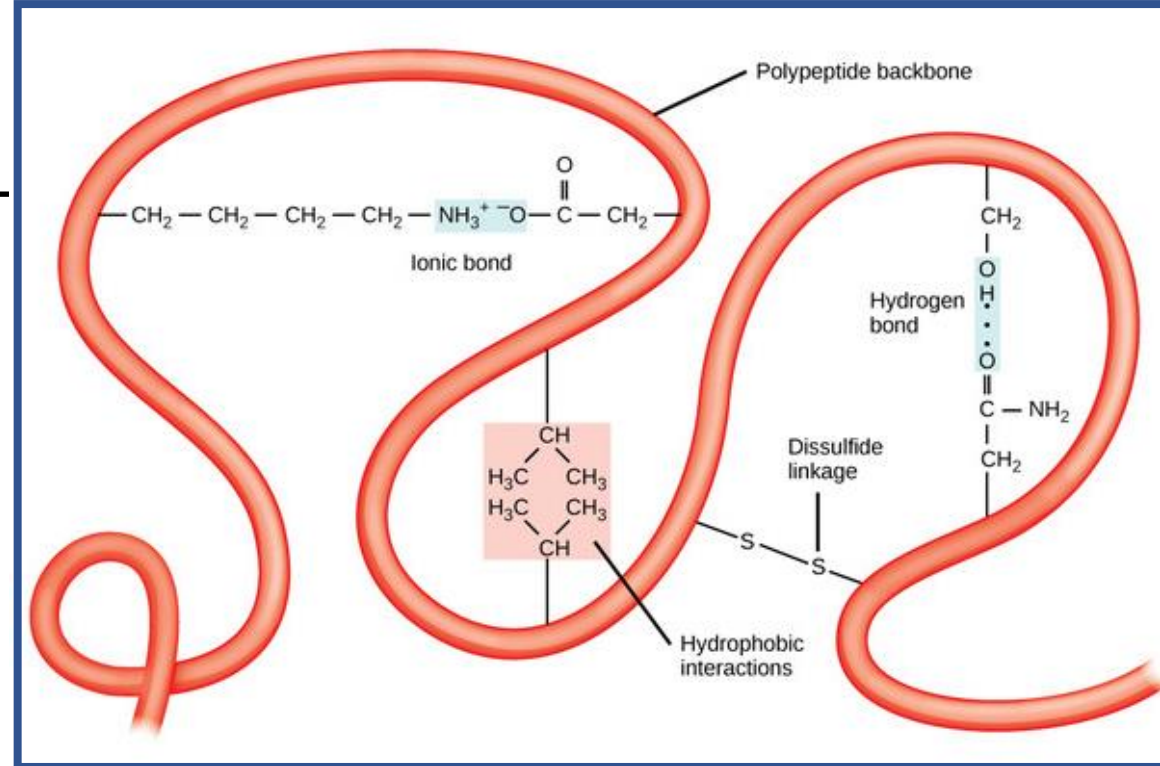
<https://courses.lumenlearning.com/introchem/chapter/protein-structure/#:~:text=Primary%20structure%20is%20the%20amino,by%20interactions%20between%20R%20groups.>



# Proteins

## Tertiary Structure

- The tertiary structure of a polypeptide chain is its **overall three-dimensional shape**, once all the secondary structure elements have folded together among each other.
- Interactions** between polar, nonpolar, acidic, and basic R group within the polypeptide chain create the complex three-dimensional tertiary structure of a protein.
- When protein folding takes place in the aqueous environment of the body, the **hydrophobic R groups of nonpolar** amino acids mostly lie in the **interior of the protein**, while the **hydrophilic R groups** lie mostly **on the outside**.
- Cysteine side chains form **disulfide linkages** in the presence of oxygen, the **only covalent bond forming** during protein folding.



**Tertiary structure:** The tertiary structure of proteins is determined by hydrophobic interactions, ionic bonding, hydrogen bonding, and disulfide linkages.

- All of **these interactions**, weak and strong, **determine the final three-dimensional shape** of the protein.
- When **a protein loses its three-dimensional shape**, it will **no longer be functional**.

Source:

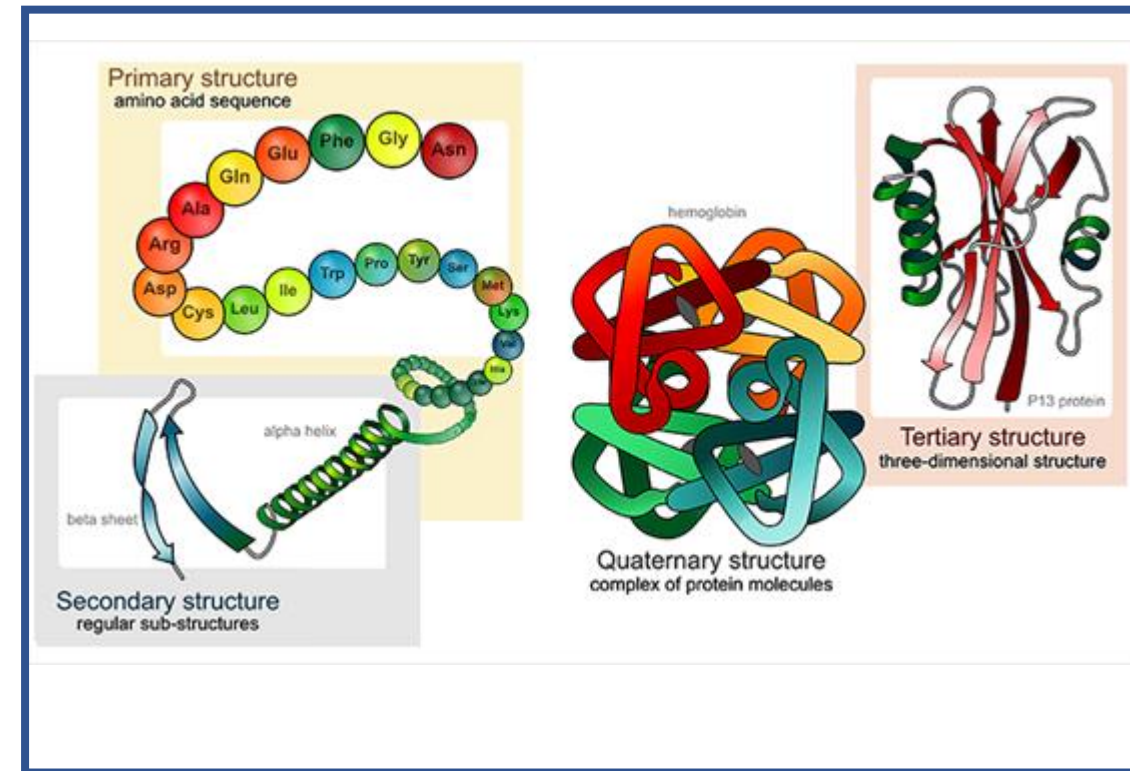
<https://courses.lumenlearning.com/introchem/chapter/protein-structure/#:~:text=Primary%20structure%20is%20the%20amino,by%20interactions%20between%20R%20groups.>



# Proteins

## Quaternary Structure

- The quaternary structure of a protein is **how its subunits are oriented and arranged** with respect to one another.
- As a result, quaternary structure **only applies to multi-subunit proteins**; that is, proteins made from more than one polypeptide chain.
- Proteins made from a **single polypeptide will not have** a quaternary structure.
- In proteins with more than one subunit, **weak interactions between the subunits** help to stabilize the overall structure.
- Enzymes often play key roles** in bonding subunits to form the final, functioning protein.
- For example, **insulin** is a ball-shaped, globular protein that contains **both hydrogen bonds and disulfide bonds** that **hold its two** polypeptide chains together.
- Silk** is a fibrous protein that results from **hydrogen bonding** between different  **$\beta$ -pleated chains**.



Source:

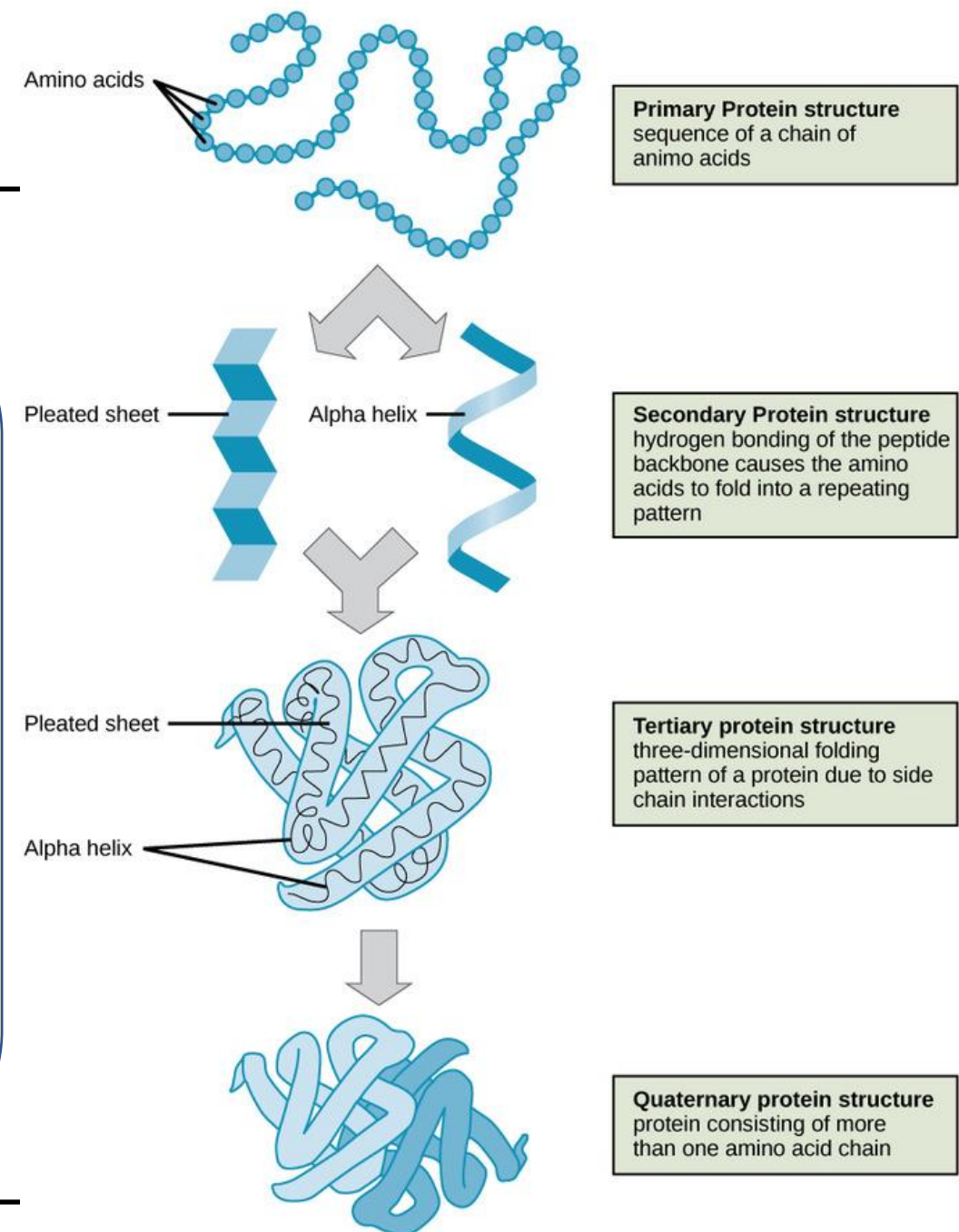
<https://byjus.com/chemistry/protein-structure-and-levels-of-protein/>

<https://courses.lumenlearning.com/introchem/chapter/protein-structure/#:~:text=Primary%20structure%20is%20the%20amino,by%20interactions%20between%20R%20groups>

# Proteins

## Four levels of protein structure:

- ✓ **Primary structure:** The hierarchy's **basic level**, and is the particular linear sequence of amino acids comprising one polypeptide chain.
- ✓ **Secondary structure:** The next level up from the primary structure, and is the **regular folding of regions into specific structural patterns** within one polypeptide chain. Hydrogen bonds between the carbonyl oxygen and the peptide bond amide hydrogen are normally held together by secondary structures.
- ✓ **Tertiary structure:** The next level up from the secondary structure, and is the particular **three-dimensional arrangement** of all the amino acids in a single polypeptide chain. This structure is usually conformational, **native, and active**, and is held together by multiple **noncovalent** interactions.
- ✓ **Quaternary structure:** The next 'step up' **between two or more polypeptide chains** from the tertiary structure and is the specific spatial arrangement and interactions.

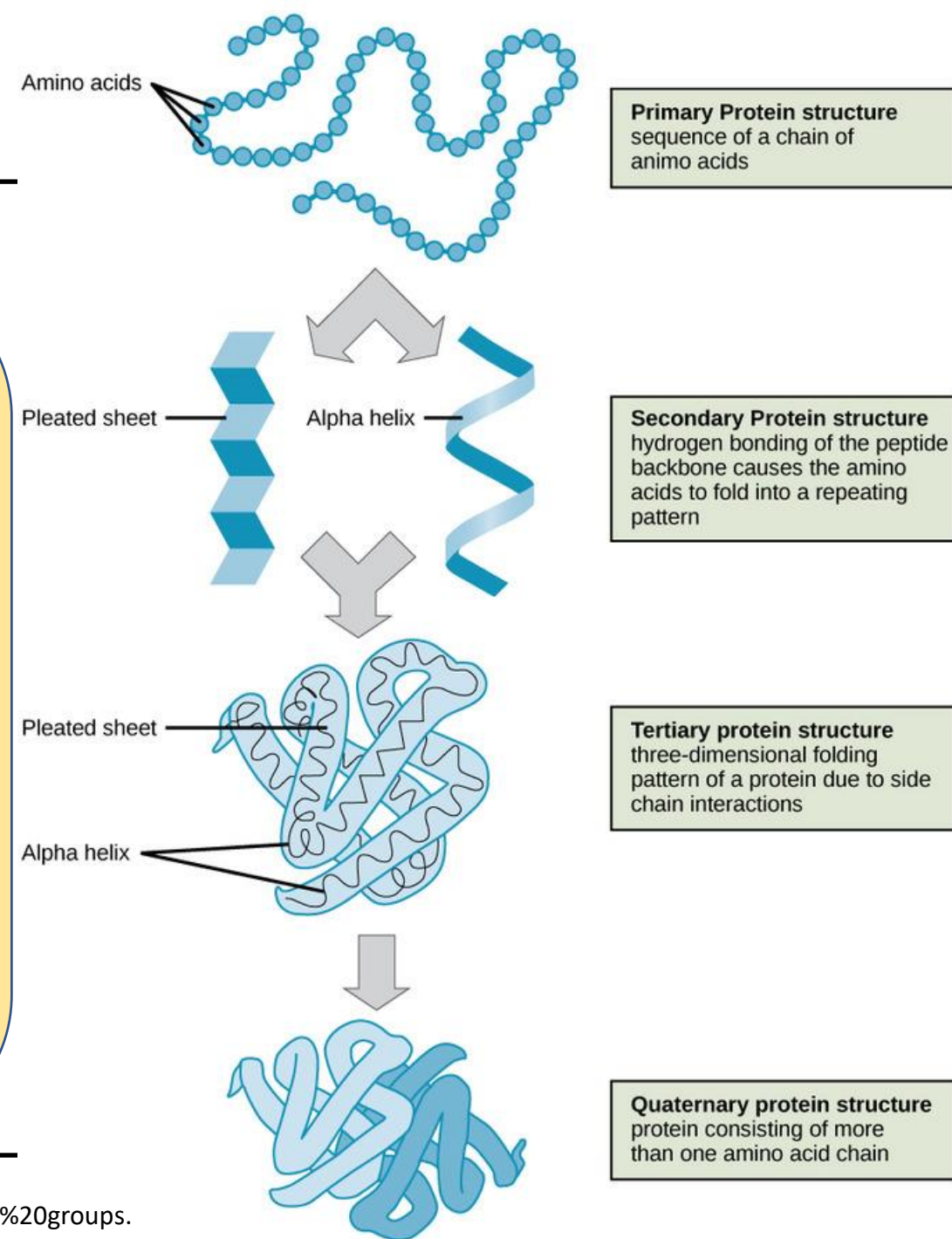


# Proteins



## Important notes

- ✓  **$\alpha$ -helix:** where every backbone **N-H** creates a hydrogen bond with the **C=O** group of the amino acid **four residues earlier** in the same helix.
- ✓  **$\beta$ -sheet:** where **N-H groups** in the backbone of one fully-extended strand establish hydrogen bonds with **C=O groups** in the backbone of an **adjacent** fully-extended strand.
- ✓ **Anti-parallel  $\beta$ -sheet:** The nature of the **opposite orientations of the two beta strands** that comprise a protein's secondary structure.
- ✓ **disulfide bond:** A bond, consisting of a **covalent bond between two sulfur atoms**, formed by the reaction of two **thiol groups**, especially between the thiol groups of two proteins.
- ✓ **Shape:** The shape of a protein is critical to its function.



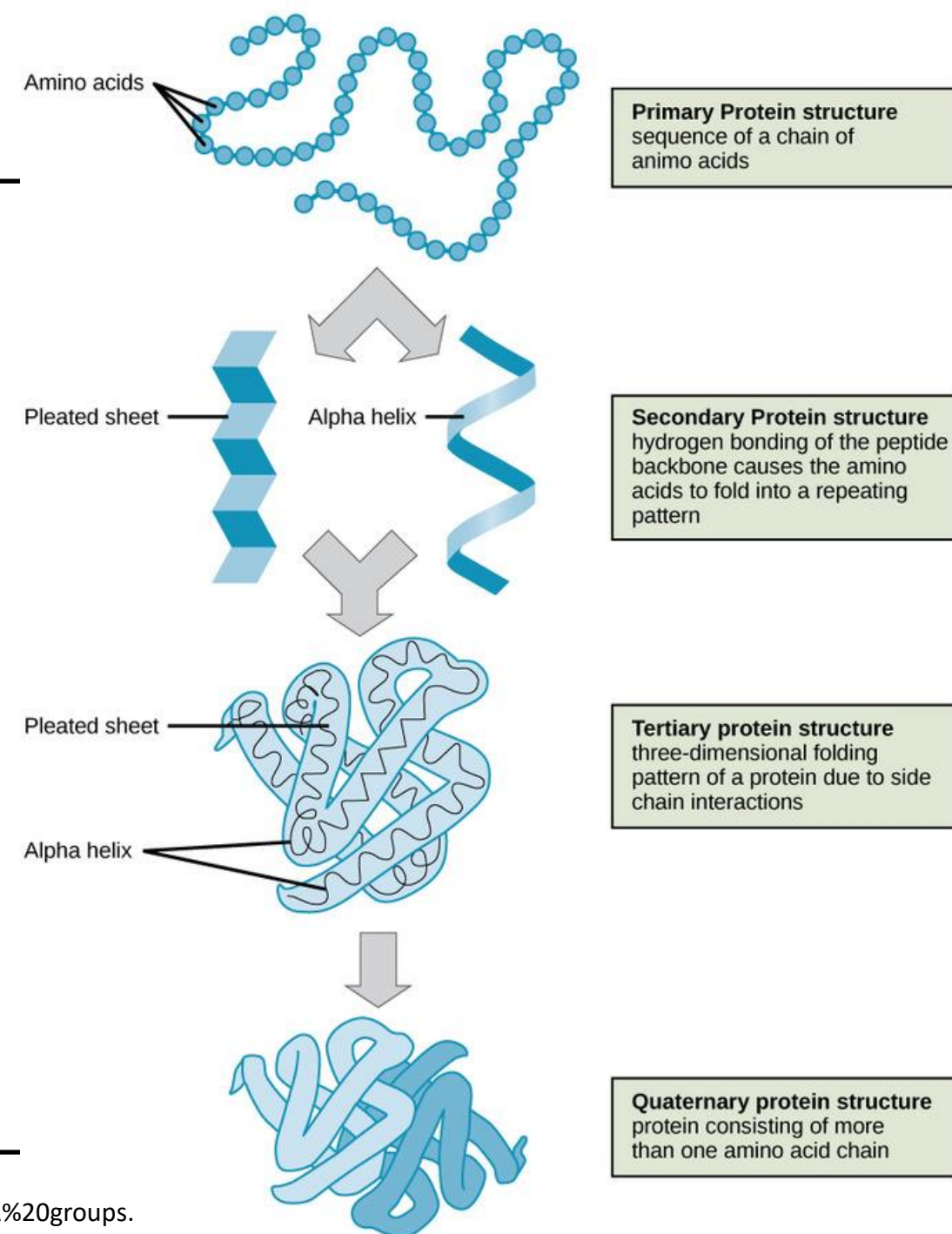


# Proteins



## Takeaway

- ✓ **Protein structure depends on its amino acid sequence** and local, low-energy chemical bonds between atoms in both the polypeptide backbone and in amino acid side chains.
- ✓ **Protein structure plays a key role in its function**; if a protein loses its shape at any structural level, it may no longer be functional.
- ✓ **Primary structure** is the amino acid sequence.
- ✓ **Secondary structure** is local interactions between stretches of a polypeptide chain and includes  **$\alpha$ -helix** and  **$\beta$ -pleated sheet** structures.
- ✓ **Tertiary structure** is the overall the **three-dimension folding** driven largely by interactions between **R groups**.
- ✓ **Quaternary structures** is the **orientation and arrangement of subunits** in a multi-subunit protein.



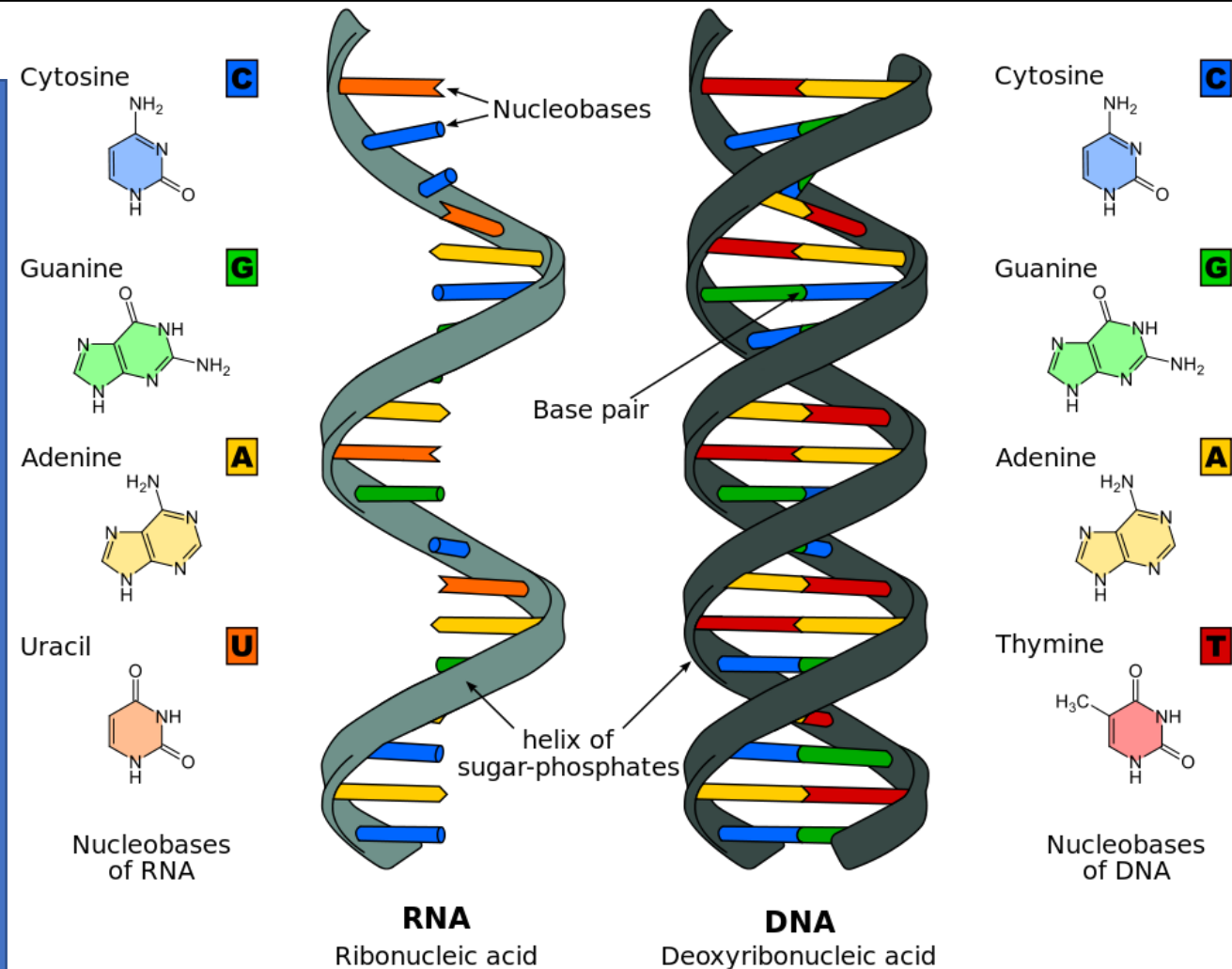
01-10-2021

# Lecture

# Nucleic acids – DNA and RNA

## Nucleic acids

- The **two main types** of nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA).
- DNA is the genetic material** found in living organisms, ranging from single-celled bacteria to multicellular mammals.
- It is found in the **nucleus of eukaryotes** and in the **chloroplasts and mitochondria**.
- In **prokaryotes**, the **DNA is not enclosed** in a membranous envelope, but rather free-floating within the cytoplasm.
- The **entire genetic content** of a cell is known as its **genome** and the study of genomes is genomics.



# Nucleic acids – DNA and RNA

## Nucleic acids

- Deoxyribonucleic acid, more commonly known as DNA, is a complex molecule that **contains all of the information necessary to build and maintain an organism.**
- **All living things have DNA** within their cells. **In fact, nearly every cell in a multicellular organism** possesses the full set of DNA required for that organism.
- However, DNA does more than specify the structure and function of living things — it also serves as the **primary unit of heredity** in organisms of all types.
- In other words, whenever organisms **reproduce, a portion of their DNA is passed along to their offspring.**
- This transmission of all or part of an organism's DNA helps **ensure a certain level of continuity** from one generation to the next, while still allowing for slight changes that contribute to the diversity of life.

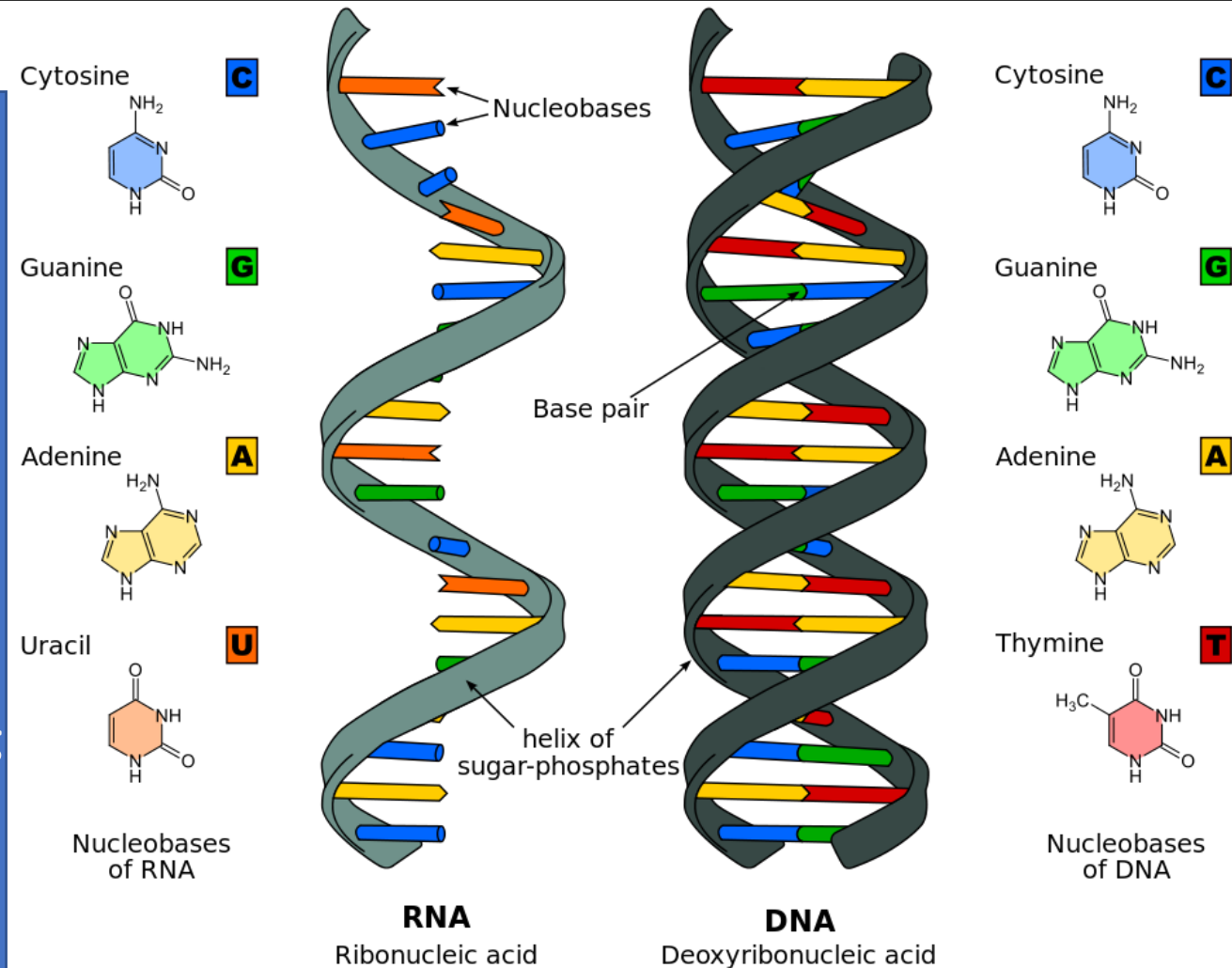




# Nucleic acids – DNA and RNA

## Nucleic acids

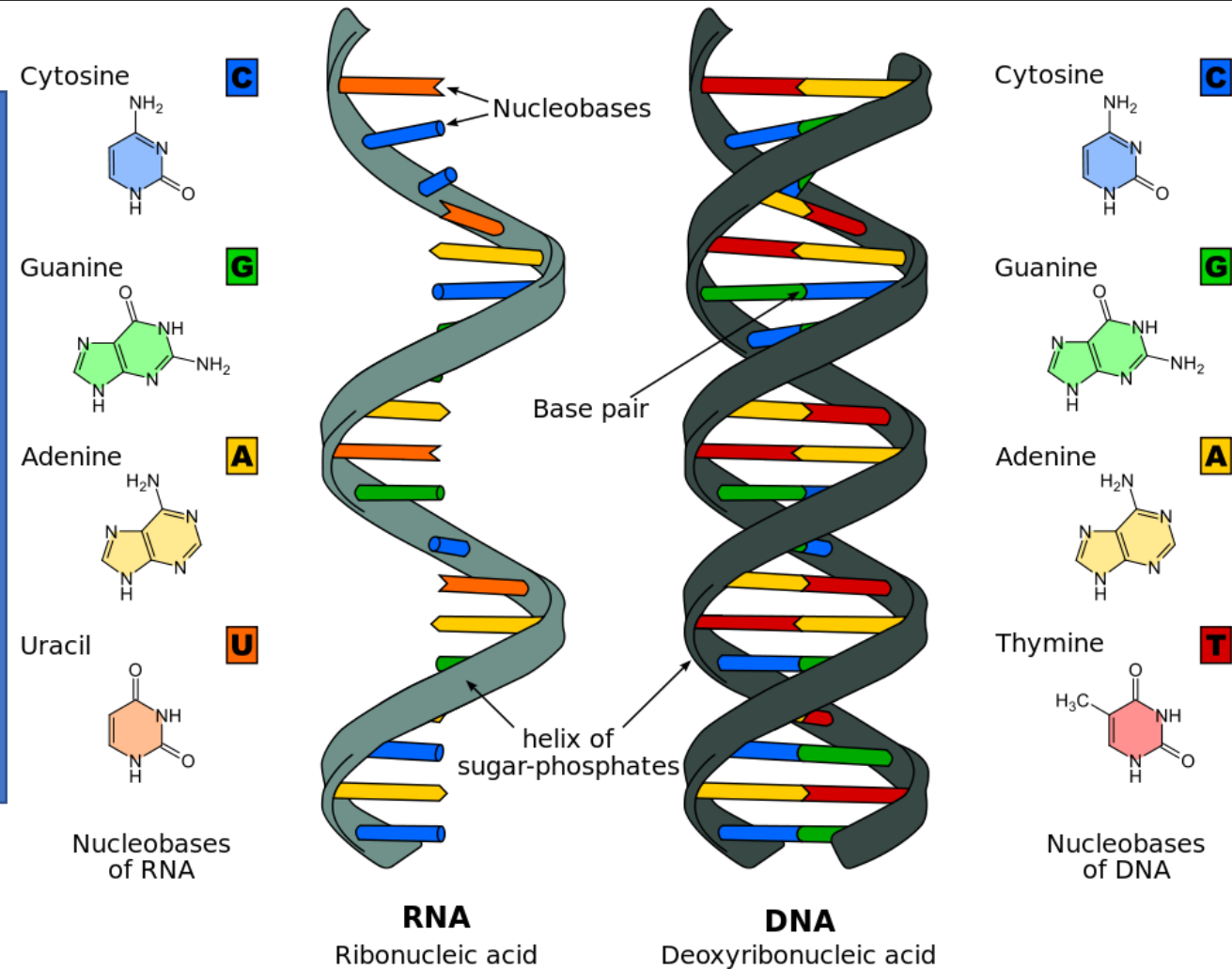
- In **eukaryotic cells**, but not in prokaryotes, DNA forms a **complex with histone** proteins to form **chromatin**, the substance of eukaryotic chromosomes.
- A **chromosome** may contain **tens of thousands** of **genes**.
- Many **genes** contain the information to make **protein products**; other genes code for RNA products.
- DNA controls all of the cellular activities** by turning the genes “on” or “off.”
- The other type of **nucleic acid**, **RNA**, is mostly involved in protein synthesis.



# Nucleic acids – DNA and RNA

## Nucleic acids

- In eukaryotes, the **DNA molecules never leave the nucleus** but instead **use an intermediary** to communicate with the rest of the cell.
- This intermediary is the **messenger RNA (mRNA)**.
- Other types of RNA—like **rRNA, tRNA, and microRNA**—are **involved in protein synthesis and its regulation**.

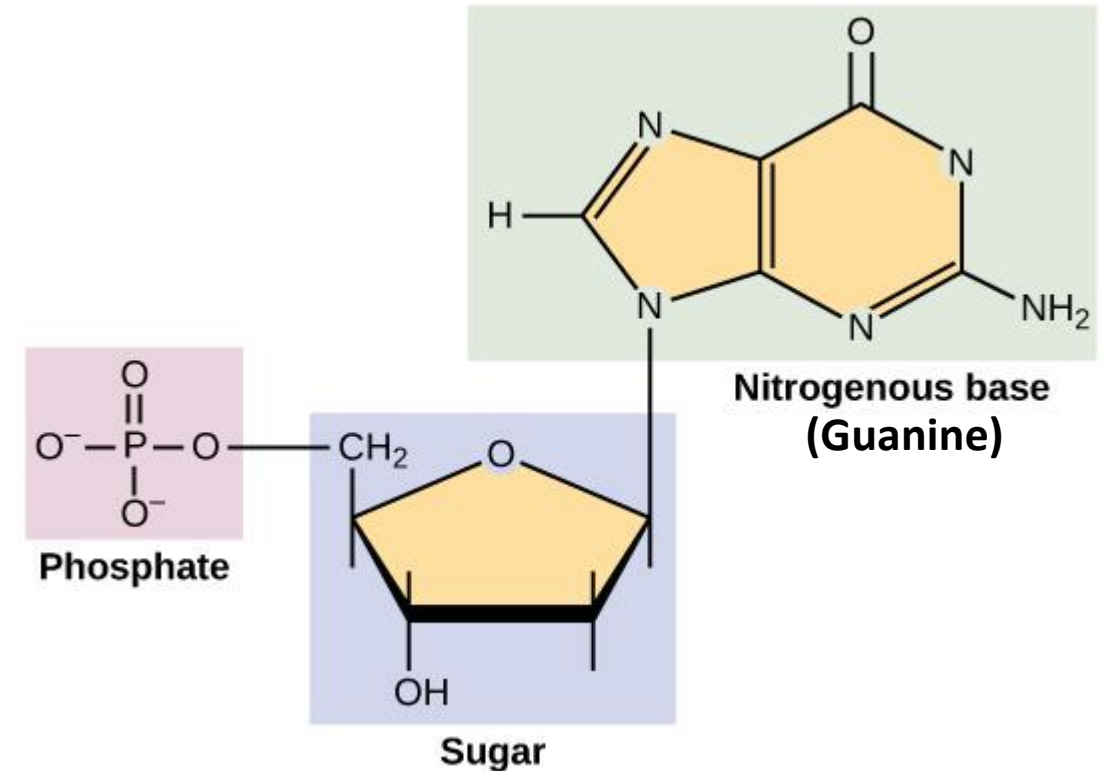


# Nucleic acids – DNA and RNA

## Nucleotides

- DNA and RNA are made up of **monomers** known as nucleotides.
- The nucleotides combine with each other to form a **polynucleotide: DNA or RNA**.
- Each nucleotide is made up of **three components**:

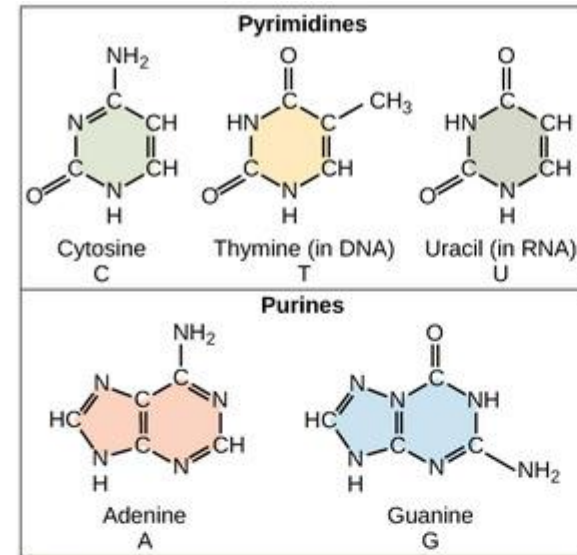
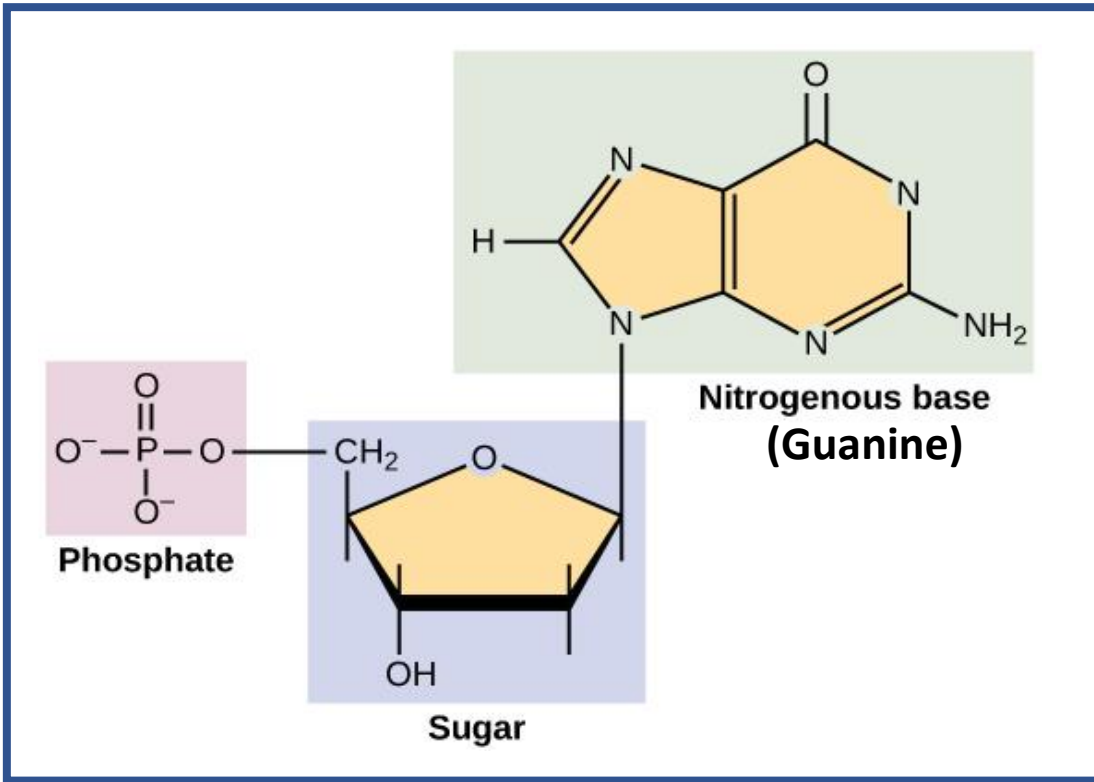
- ✓ a nitrogenous base
- ✓ a pentose (five-carbon) sugar
- ✓ a phosphate group



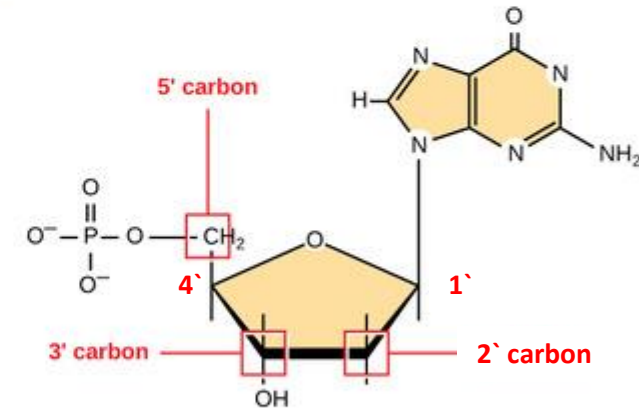
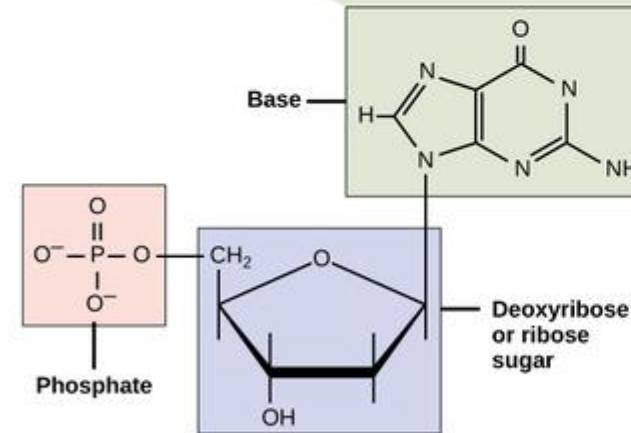
Each nitrogenous base in a nucleotide is attached to a sugar molecule, which is attached to one or more phosphate groups.

# Nucleic acids – DNA and RNA

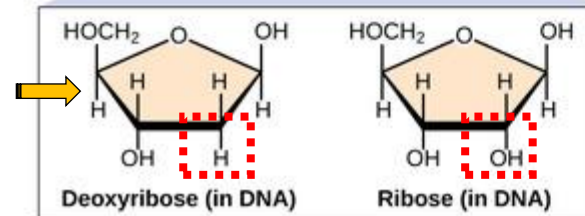
## Nucleotides



## Nitrogenous bases



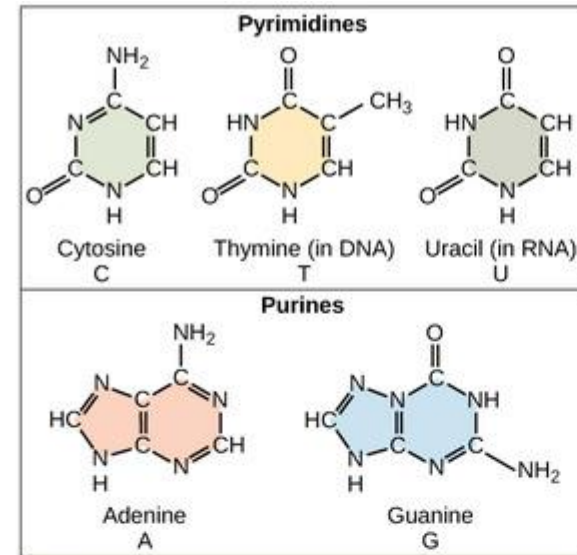
## 2-Deoxyribose



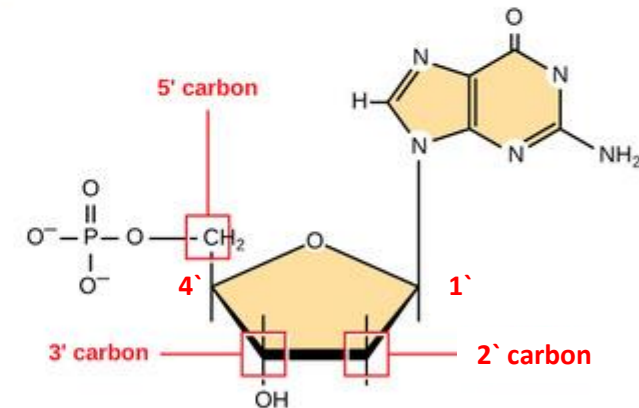
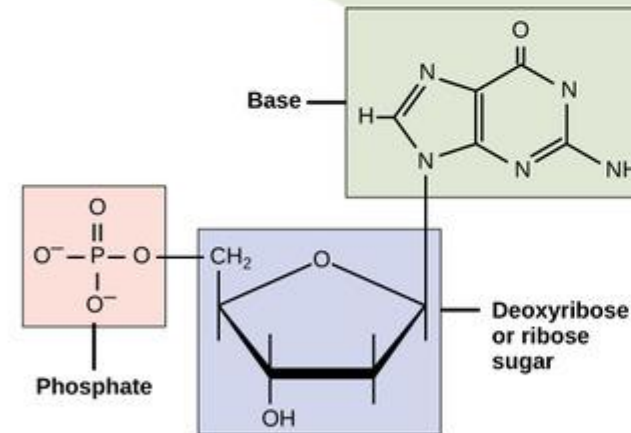
# Nucleic acids – DNA and RNA

## Nucleotides

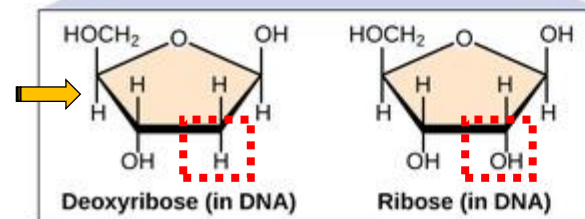
- The **base** is attached to the **1'** position of the ribose, and the **phosphate** is attached to the **5'** position.
- When a **polynucleotide** is formed, the **5' phosphate** of the incoming nucleotide attaches to the **3' hydroxyl group** at the end of the growing chain.
- Two types of **pentose** are found in nucleotides, **deoxyribose** (in DNA) and **ribose** (found in RNA).
- Deoxyribose** is similar in structure to ribose, but it has an **H instead of an OH** at the 2' position.
- Bases are of **two types**: purines and pyrimidines.
- Purines** have a **double ring structure**, and **pyrimidines** have a **single ring**.



Nitrogenous bases



## 2-Deoxyribose





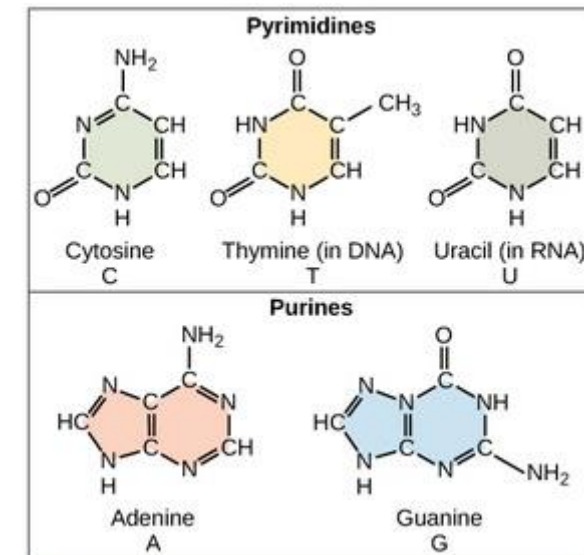
# Nucleic acids – DNA and RNA

## Nucleotides

## Nitrogenous bases

- The **nitrogenous bases** are organic molecules and are so named because they **contain carbon and nitrogen**.
- They are **bases** because they contain an **amino group** that has the **potential of binding an extra hydrogen**, and thus, **decreasing the hydrogen ion concentration** in its environment, making it **more basic**.
- Each nucleotide in **DNA** contains one of four possible nitrogenous bases: adenine (**A**), guanine (**G**), cytosine (**C**), and thymine (**T**; and **Uracil (U)** in RNA).
- A and G are **purines** and primarily consists of **two carbon-nitrogen rings**.
- C, T and U are **pyrimidines** which have a **single carbon-nitrogen ring**.
- Each of these basic carbon-nitrogen rings has **different functional groups attached** to it.

## Nitrogenous bases ↓



In molecular biology and **bioinformatics** shorthand

↓  
**Nitrogenous bases are represented by their symbols A, T, G, C, and U.**

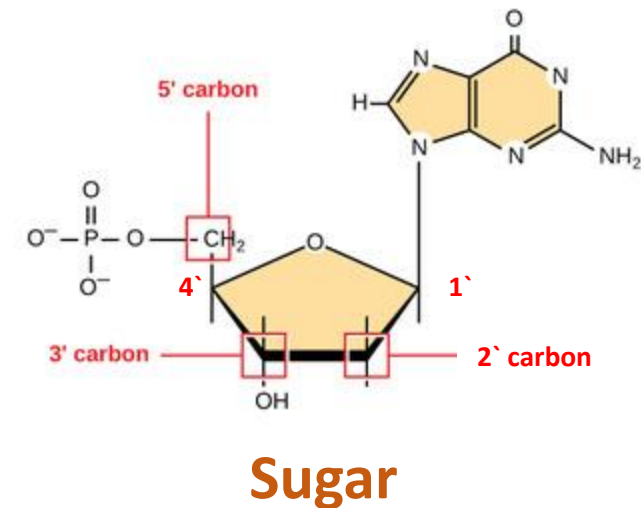
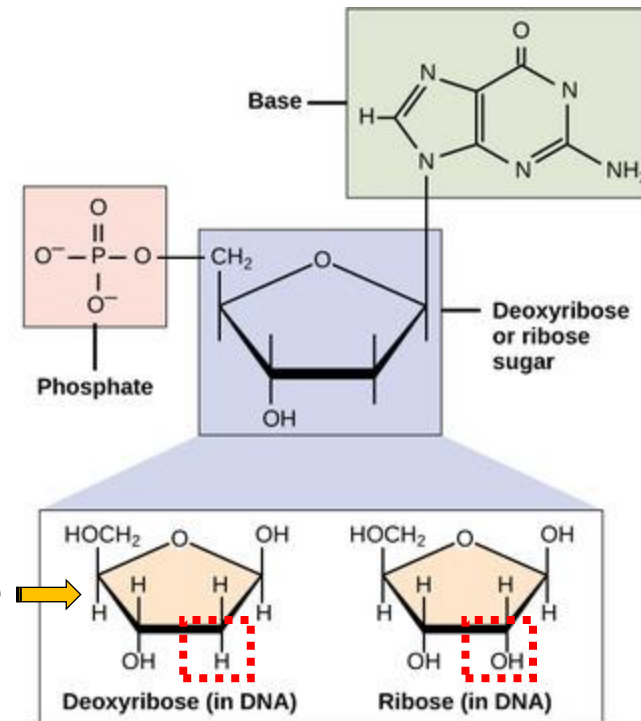
**Note:** DNA contains A, T, G, and C whereas RNA contains A, U, G, and C.

# Nucleic acids – DNA and RNA

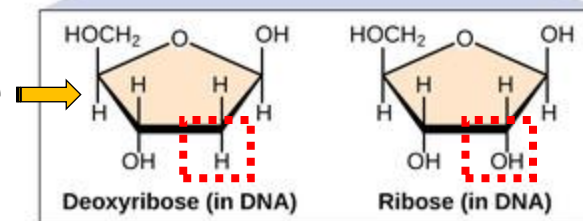
## Nucleotides

## Ribose sugar

- The pentose sugar in DNA is **deoxyribose** and in RNA it is **ribose**.
- The difference between the sugars is the presence of the **hydroxyl group on the second carbon of the ribose** and **hydrogen on the second carbon of the deoxyribose**.
- The **carbon** atoms of the sugar molecule are **numbered as 1', 2', 3', 4', and 5'** (1' is read as "one prime").



2-Deoxyribose →

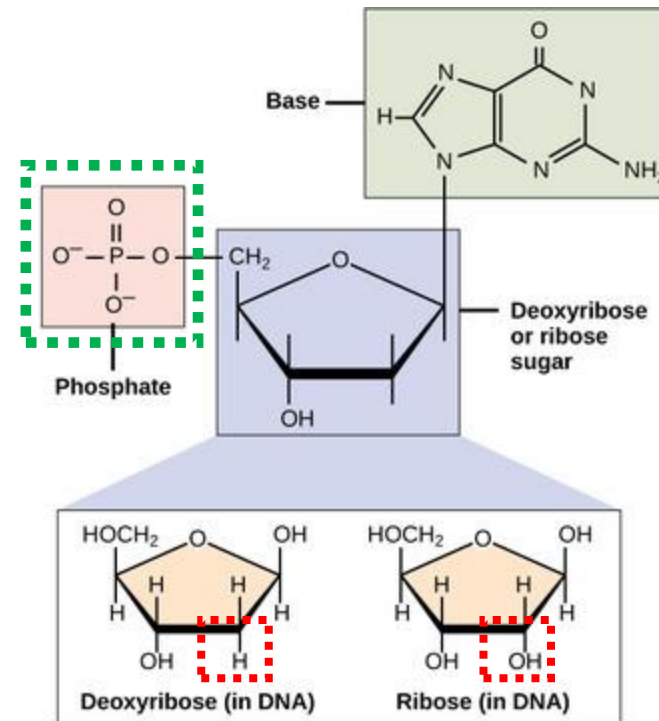


# Nucleic acids – DNA and RNA

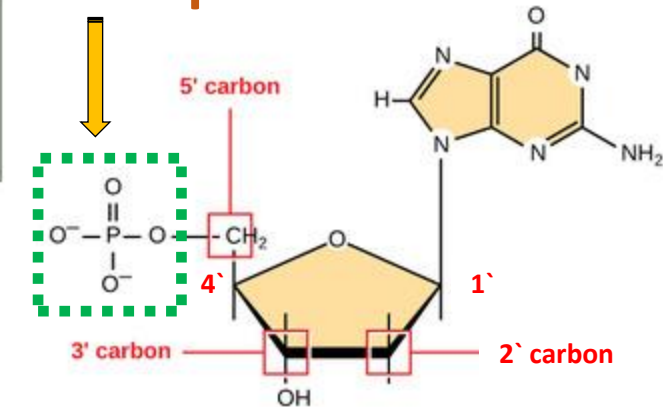
## Nucleotides

### Phosphate group

- The **phosphate** residue is attached to the **hydroxyl group of the 5' carbon of one sugar** and **the hydroxyl group of the 3' carbon of the sugar of the next nucleotide**, which forms a **5'3' phosphodiester linkage**.
- The phosphodiester linkage is not formed by simple dehydration reaction like the other linkages connecting monomers in macromolecules: its formation involves the removal of two phosphate groups.
- A polynucleotide may have thousands of such phosphodiester linkages.



### Phosphate



### Sugar



# Nucleic acids – DNA and RNA

## DNA base pairing

## Double helical structure

- The **phosphate** residue is attached to the **hydroxyl group** of the **5' carbon** of one sugar and the **hydroxyl group** of the **3' carbon** of the sugar of the next nucleotide, which forms a **5'3' phosphodiester linkage**.
- In a double stranded DNA molecule, **the two strands run antiparallel to one another** so one is upside down compared to the other.
- The **phosphate backbone** is located on the **outside**, and the **bases are in the middle**.
- Adenine forms hydrogen bonds (or base pairs) with thymine.
- Guanine base pairs with cytosine.

## Hydrogen bonding pattern of base pairs

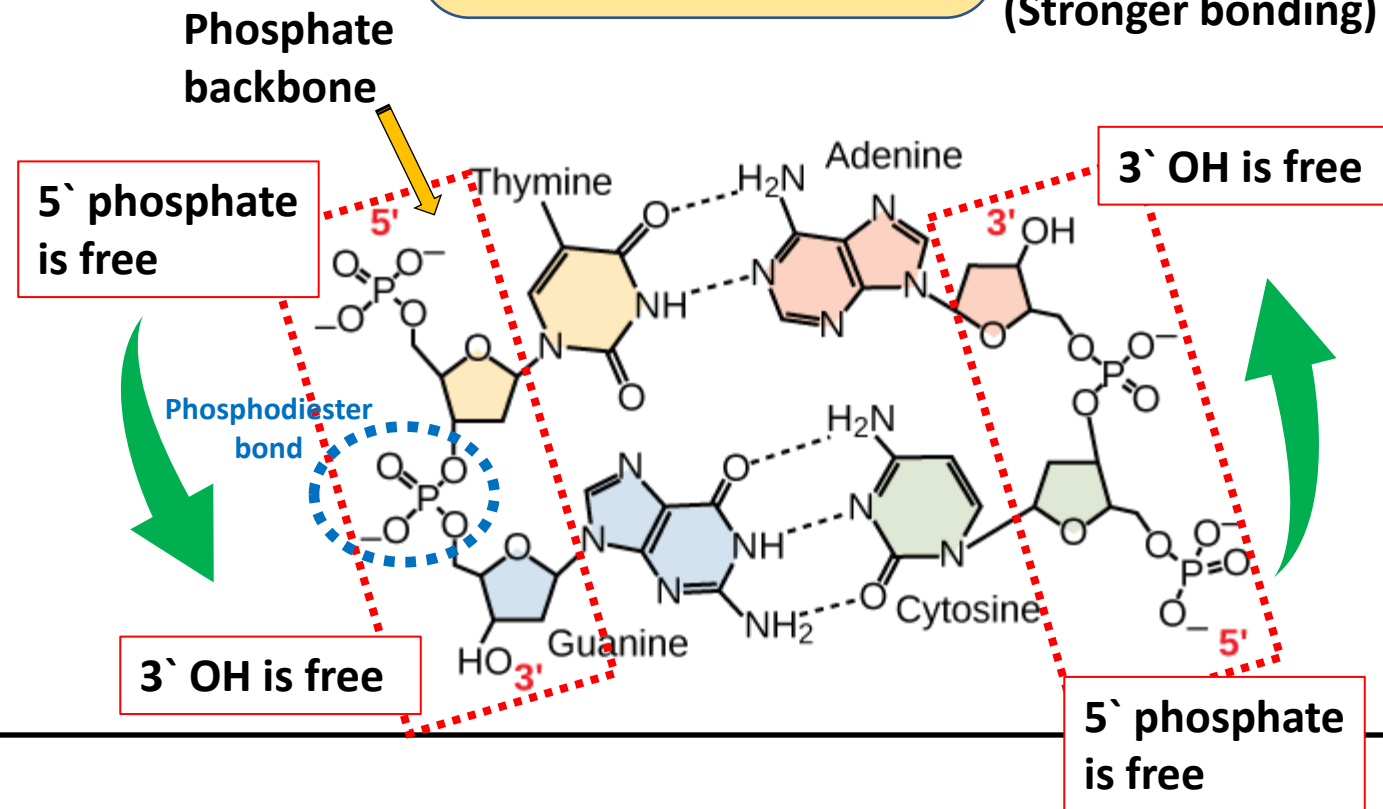
### Purines Pyrimidines

A ===== T

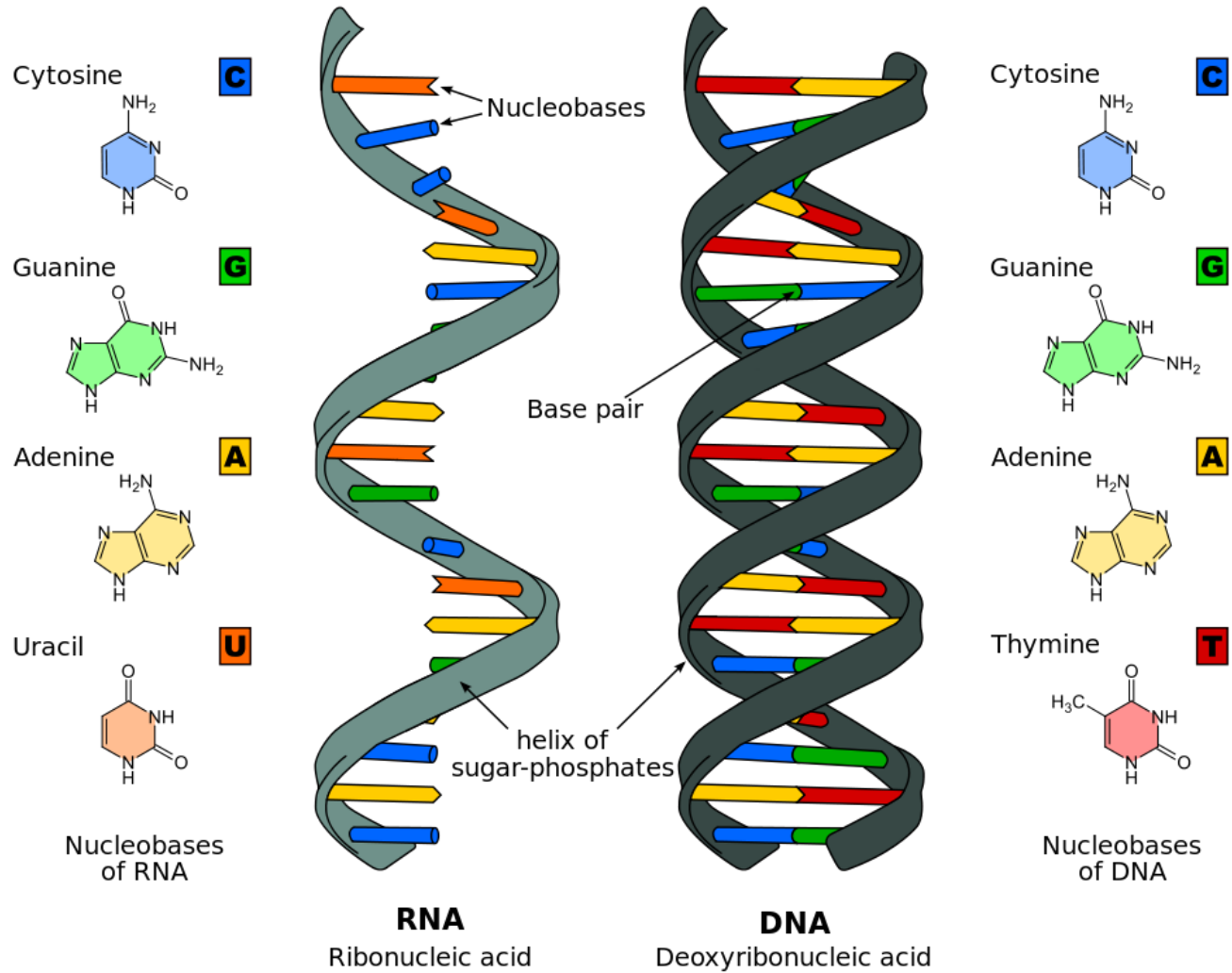
→ 2 H-bonds

G ===== C

→ 3 H-bonds  
(Stronger bonding)



# Nucleic acids – DNA and RNA



# Nucleic acids – DNA and RNA

## Important notes

- ✓ **Nucleotide:** the **monomer comprising DNA or RNA** molecules; consists of a nitrogenous heterocyclic base that can be a purine or pyrimidine, a five-carbon pentose sugar, and a phosphate group.
- ✓ **Genome:** the **cell's complete genetic information** packaged as a double-stranded DNA molecule.
- ✓ **Monomer:** A **relatively small molecule** which can be covalently bonded to other monomers to form a polymer.
- ✓ **Mutation:** any error in base pairing during the replication of DNA
- ✓ **Sugar-phosphate backbone:** The **outer support of the ladder**, forming strong covalent bonds between monomers of DNA.
- ✓ **Base pairing:** The **specific way in which bases of DNA line up** and bond to one another; A always with T and G always with C.

# Nucleic acids – DNA and RNA

## Takeaway

- ✓ The **two main types** of nucleic acids are DNA and RNA.
- ✓ Both DNA and RNA are **made from nucleotides**, each containing a five-carbon sugar backbone, a phosphate group, and a nitrogen base.
- ✓ **DNA provides the code** for the cell 's activities, **while RNA converts that code into proteins** to carry out cellular functions.
- ✓ The sequence of nitrogen bases **(A, T, C, G) in DNA** is what forms an organism's traits.
- ✓ The nitrogen bases **A and T (or U in RNA)** always go together and **C and G always go together**, forming the 5'-3' phosphodiester linkage found in the nucleic acid molecules.
- ✓ The **structure of DNA** is called a **double helix**, which looks like a twisted staircase.
- ✓ The **sugar and phosphate make up the backbone**, while the nitrogen bases are found in the center and hold the two strands together.
- ✓ The **nitrogen bases can only pair in a certain way**: A pairing with T and C pairing with G. This is called base pairing.
- ✓ Due to the base pairing, the **DNA strands are complementary to each other**, run in opposite directions, and are called antiparallel strands.

# Lecture

# Bioinformatics

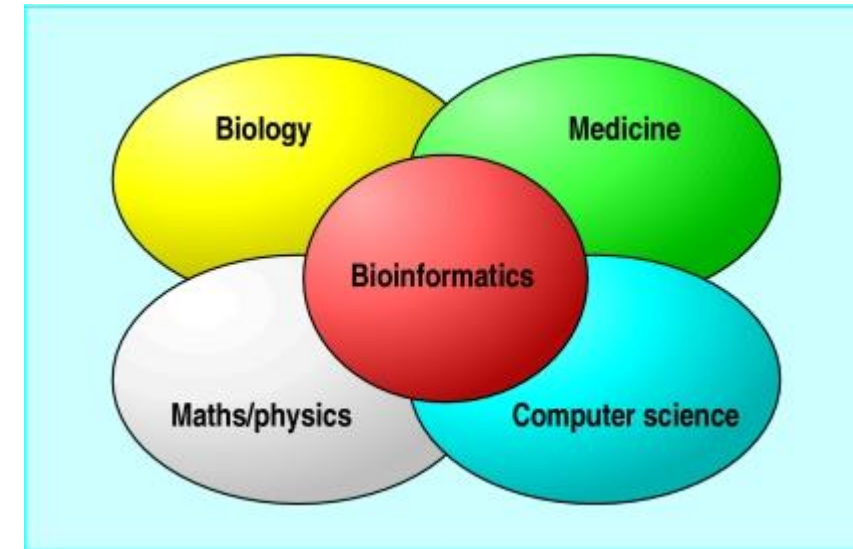
## Several viewpoints

Definitions vary widely

**Bio + informatics** ➡ Processing of biological data using information technology.

- Bioinformatics is an **interdisciplinary** research area at the interface between computer science and biological science.
- A **variety of definitions** exist in the literature and on the world wide web; some are more inclusive than others.
- Bioinformatics is a **union of biology and informatics**: bioinformatics involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as **DNA, RNA, and proteins**.

**Bioinformatics is the application of computational technology to handle the rapidly growing repository of information related to molecular biology.**



Source: BMJ. 2002 Apr 27; 324(7344): 1018–1022.  
doi: 10.1136/bmj.324.7344.1018



# Bioinformatics

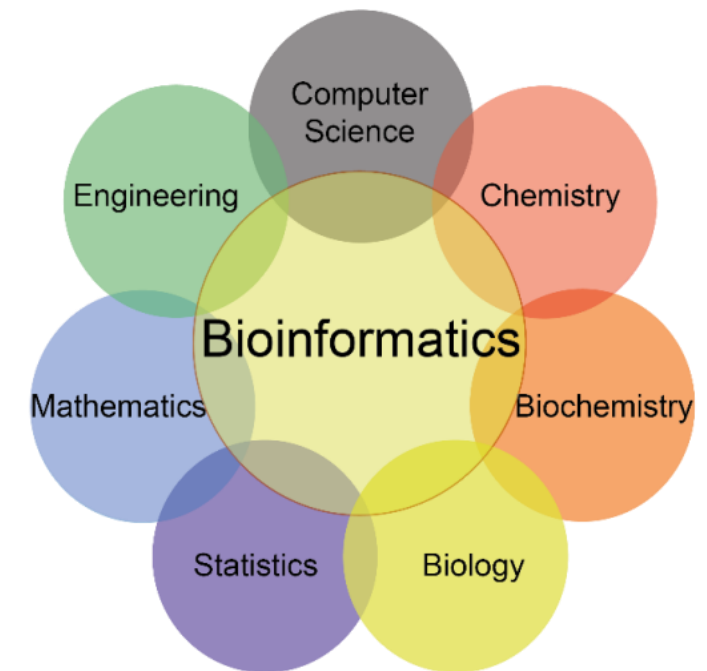
## Several viewpoints

**Bio + informatics** ➡ Processing of biological data using information technology.

- Bioinformatics **combines** different fields of study, including **computer sciences, molecular biology, biotechnology, statistics** and **engineering**.
- It is particularly **useful for managing and analyzing** large sets of data, such as those generated by the fields of **genomics and proteomics**.

## Takeaways

- ✓ Bioinformatics employs **computers and information technology** to large molecular biology data sets.
- ✓ Bioinformatics is seen as a **cutting-edge branch of the biotechnology sector, used for novel drug discovery and personalized medicines**.
- ✓ The field closely **combines computer science and artificial intelligence** with **microbiology and genomics**.



Source: <https://www.cleanpng.com/png-bioinformatics-computer-science-computational-biol-2602217/preview.html>

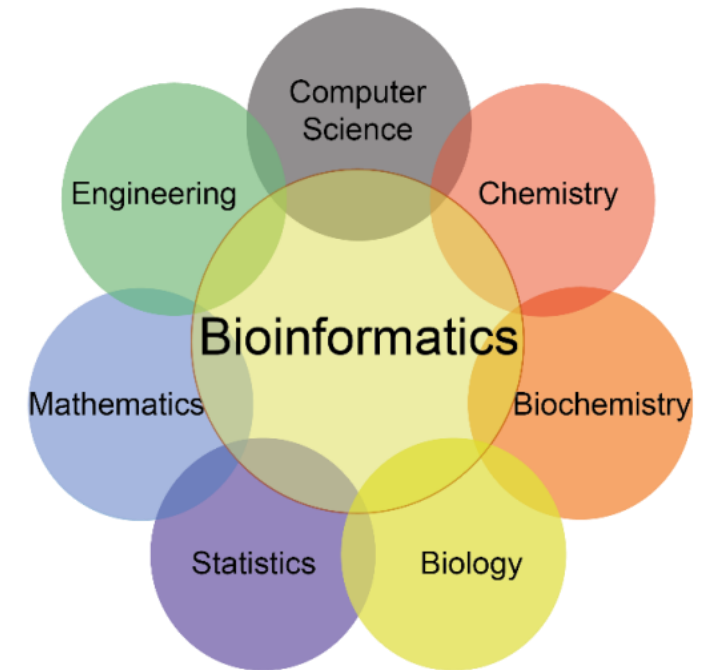
# Bioinformatics

## Several viewpoints

**Understanding Bioinformatics** ➡ **Processing of biological data using information technology.**

- While the field of bioinformatics has existed for decades, the catalyst for its rapid growth in the current millennium came from the **Human Genome Project**, a landmark international scientific research project completed in April 2003 that made available for the first time the complete genetic blueprint of a human being.

- ✓ Bioinformatics finds application in a growing number of areas, such as gene sequencing, gene expression studies and drug discovery.
- ✓ For example, in medicine, bioinformatics can be used to identify links between specific diseases and the gene sequences that cause them.
- ✓ The field of pharmacogenomics uses bioinformatics data to tailor medical treatments to the patients based on their DNA.



Source: <https://www.cleanpng.com/png-bioinformatics-computer-science-computational-biol-2602217/preview.html>

# Bioinformatics

## Bioinformatics versus Computational Biology

### Bioinformatics

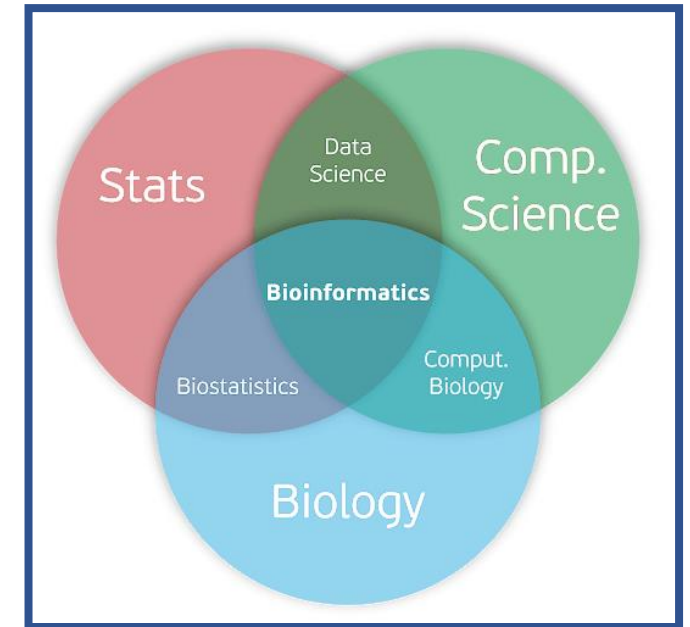


Mainly involve macromolecules

### Computational Biology



All computational studies using biological data.



Source:

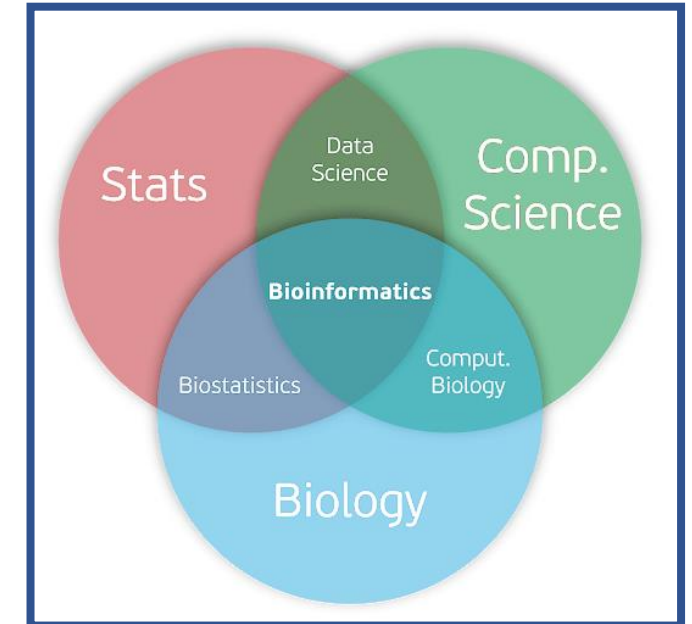
[https://www.researchgate.net/post/What\\_are\\_the\\_differences\\_between\\_Bioinformatics\\_and\\_Computational\\_Biology](https://www.researchgate.net/post/What_are_the_differences_between_Bioinformatics_and_Computational_Biology)

# Bioinformatics

## Bioinformatics versus Computational Biology



- Bioinformatics **overlaps** with other areas of research that are designated computational biology.
- Bioinformatics is involved in **organizing biological data related to genomes, proteomes** etc. with a view to implement this information to **agriculture, pharmacology, medicine** and other commercial applications.
- **Bioinformatics** involve **sequence, structural, and functional analysis** of genes and genomes and their corresponding products and is often considered **computational molecular biology**.
- However, **computational biology** encompasses **all biological areas that involve computation**. For example, **mathematical modeling of ecosystems, population dynamics**, application of the game theory in **behavioral studies**, and **phylogenetic construction using fossil records** all employ computational tools, but **do not necessarily involve biological macromolecules**.



Source:

[https://www.researchgate.net/post/What\\_are\\_the\\_differences\\_between\\_Bioinformatics\\_and\\_Computational\\_Biology](https://www.researchgate.net/post/What_are_the_differences_between_Bioinformatics_and_Computational_Biology)

# Bioinformatics

## Broad Objectives of Bioinformatics



Processing of biological data using information technology.

- ✓ To **organize vast amount of molecular biology** data in an efficient manner
- ✓ To **develop tools** that aid in the analysis of such data
- ✓ To **interpret the data** accurately and meaningfully



Source: <https://mangalmay.org/blog/what-is-bioinformatics/>

- The advent and **rapid rise of bioinformatics** have been **due to the massive increases in computing power and laboratory technology**.
- These advances have made it **possible to process and analyze** the digital information—**DNA, genes, proteins**—at the heart of life itself.
- As bioinformatics can be used in any system where information can be represented digitally, it can be **applied across the entire spectrum of living organisms**, from single cells to complex ecosystems.

# Bioinformatics

## Brief History of Bioinformatics



- Modern bioinformatics **emerged recently to assist next-generation** sequencing data analysis.
- However, the very beginnings of bioinformatics occurred **more than 50 years ago**, when desktop computers were still a hypothesis and DNA could not yet be sequenced.
- The **foundations** of bioinformatics were laid in the **early 1960s** with the application of computational methods to **protein sequence analysis** (notably, de novo sequence assembly, biological sequence databases and substitution models).
- Later on, **DNA analysis** also emerged due to parallel advances in (i) **molecular biology methods**, which allowed easier manipulation of DNA, as well as its sequencing, and (ii) **computer science**, which saw the rise of increasingly miniaturized and more powerful computers, as well as novel software better suited to handle bioinformatics tasks.
- In the **1990s through the 2000s**, major **improvements in sequencing** technology, along with reduced costs, gave rise to an exponential increase of data.
- The arrival of '**Big Data**' has laid out new challenges in terms of data mining and management, calling for more expertise from computer science into the field.
- Coupled with an ever-increasing amount of **bioinformatics tools**, biological Big Data had (and continues to have) profound implications on the predictive power and reproducibility of bioinformatics results.

### Source:

Gauthier et al., Briefings in Bioinformatics, Volume 20, Issue 6, November 2019,  
Pages 1981–1996, <https://doi.org/10.1093/bib/bby063>



# Bioinformatics

## Brief History of Bioinformatics

## Major events in Bioinformatics



1950–1970: The origins

It did not start with DNA analysis

Protein analysis was the starting point

Dayhoff: the first bioinformatician

A mathematical framework for amino acid substitutions



In 1978, Dayhoff et al developed the first probabilistic model of amino acid substitutions - point accepted mutations (PAMs).



**Margaret Dayhoff**  
(1925-1983)

- ✓ Dayhoff was an American physical chemist who pioneered the application of computational methods to the field of biochemistry.
- ✓ Dayhoff's contribution to this field is so important that David J. Lipman, former director of the National Center for Biotechnology Information (NCBI), called her 'the mother and father of bioinformatics'

### Source:

Gauthier et al., Briefings in Bioinformatics, Volume 20, Issue 6, November 2019, Pages 1981–1996, <https://doi.org/10.1093/bib/bby063>



# Bioinformatics

## Brief History of Bioinformatics

### Major events in Bioinformatics



1965	Margaret Dayhoff's Atlas of Protein Sequences
1970	Needleman-Wunsch algorithm
1977	DNA sequencing and software to analyze it (Staden)
1981	Smith-Waterman algorithm developed
1981	The concept of a sequence motif (Doolittle)
1982	GenBank Release 3 made public
1982	Phage lambda genome sequenced
1983	Sequence database searching algorithm (Wilbur-Lipman)
1985	FASTP/FASTN: fast sequence similarity searching
	National Center for Biotechnology Information (NCBI) created at NIH/NLM
1988	EMBL network for database distribution
1990	BLAST: fast sequence similarity searching
1991	EST: expressed sequence tag sequencing
1993	Sanger Centre, Hinxton, UK
1994	EMBL European Bioinformatics Institute, Hinxton, UK
1995	First bacterial genomes completely sequenced
1996	Yeast genome completely sequenced
1997	PSI-BLAST
1998	Worm (multicellular) genome completely sequenced
1999	Fly genome completely sequenced
	Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. Nature 2000 Oct 5;407(6804):651-4, PubMed
2000	The genome for Pseudomonas aeruginosa (6.3 Mbp) is published.
2000	The A. thaliana genome (100 Mb) is sequenced.
2001	The human genome (3 Giga base pairs) is published.

Source: [https://www.roseindia.net/bioinformatics/history\\_of\\_bioinformatics.shtml](https://www.roseindia.net/bioinformatics/history_of_bioinformatics.shtml)

[https://chagall.med.cornell.edu/BioinfoCourse/presentations2010/Lecture1\\_2010.pdf](https://chagall.med.cornell.edu/BioinfoCourse/presentations2010/Lecture1_2010.pdf)

# Bioinformatics

## Goals

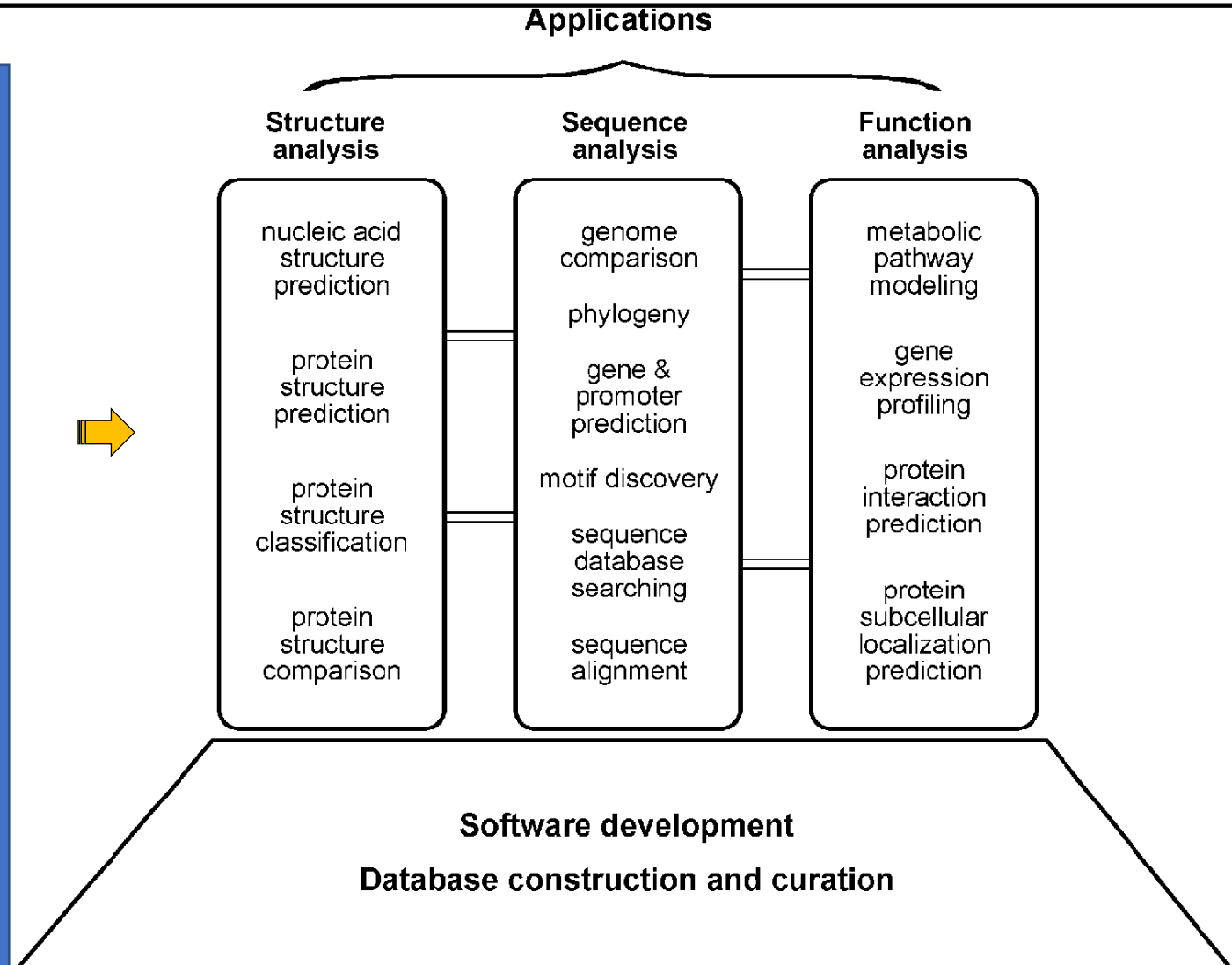


- The ultimate goal of bioinformatics is to better **understand a living cell** and **how it functions** at the molecular level.
- By analyzing **raw molecular sequence and structural data**, bioinformatics research can generate new insights and provide a **“global” perspective** of the cell.
- The reason that the **functions of a cell** can be better understood **by analyzing sequence** data is ultimately because the flow of genetic information is dictated by the **“central dogma”** of biology in which DNA is transcribed to RNA, which is translated to proteins.
- Cellular **functions** are mainly performed **by proteins** whose capabilities are ultimately **determined by their sequences**.
- Therefore, **solving functional problems** using **sequence** and sometimes **structural** approaches has proved to be a **fruitful** endeavor.

# Bioinformatics

## Scope

- Bioinformatics consists of **two subfields**: the **development** of computational tools and databases and the **application** of these tools and databases in generating biological knowledge to better understand living systems.
- These two subfields are **complementary to each other**.
- The tool development includes **writing software** for sequence, structural, and functional analysis, as well as the **construction** and curating of biological **databases**.
- These tools are **used in** three areas of genomic and molecular biological research: molecular **sequence analysis**, molecular **structural analysis**, and molecular **functional analysis**.
- The analyses of biological data often generate new problems and challenges that in turn spur the development of new and better computational tools.



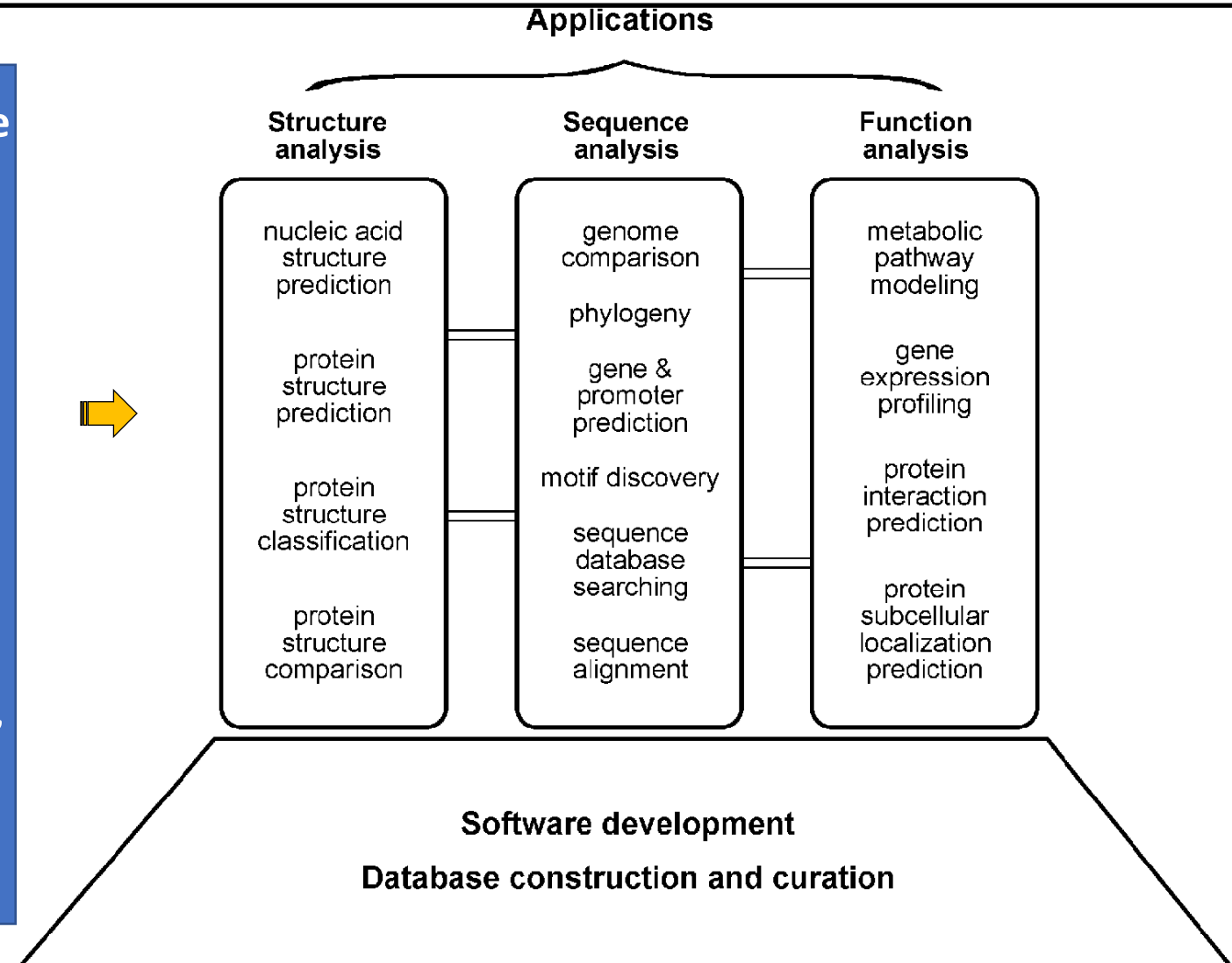
Source:

Essential Bioinformatics by Jin Xiong

# Bioinformatics

## Scope

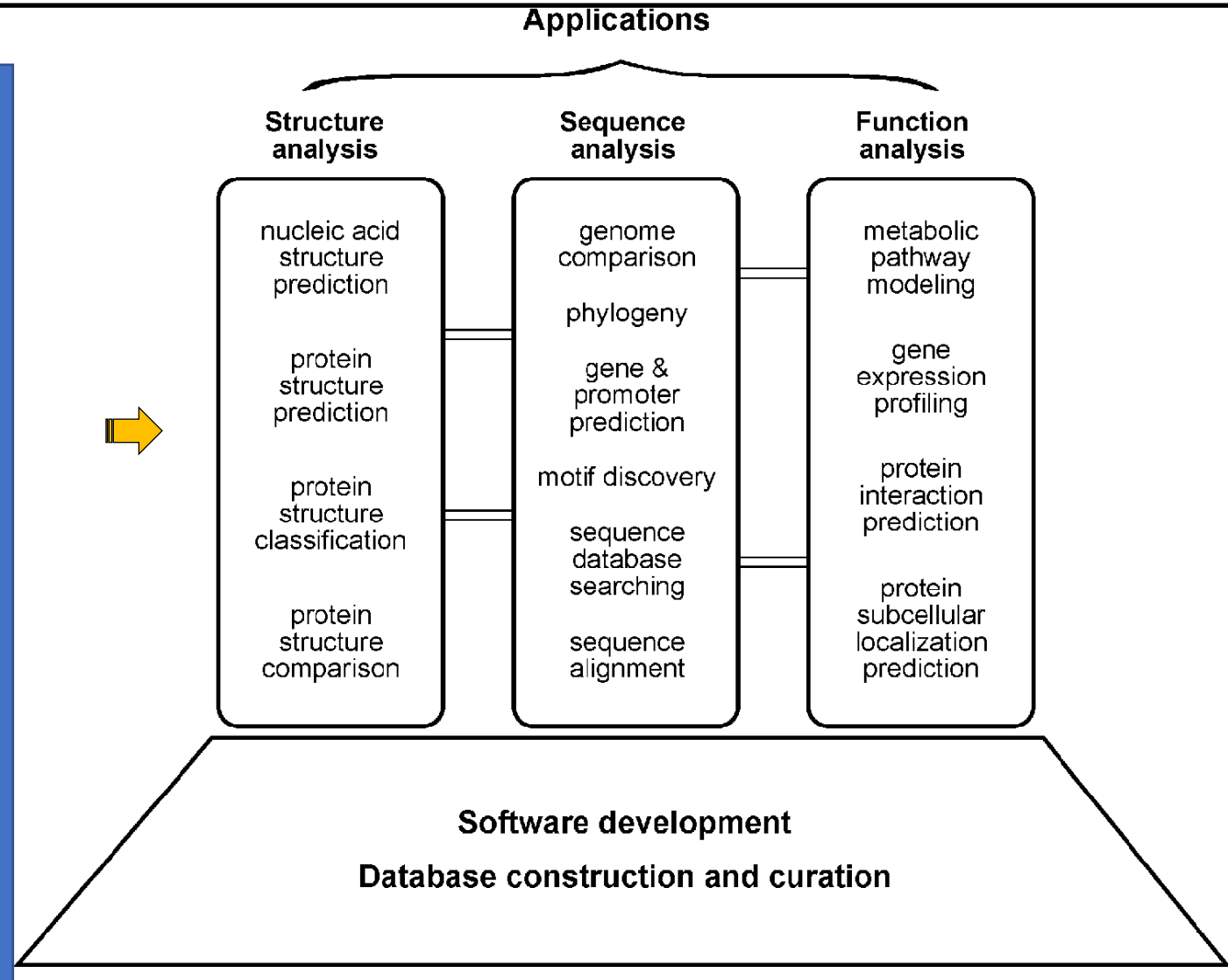
- The areas of **sequence analysis** include sequence alignment, sequence database searching, motif and pattern discovery, gene and promoter finding, reconstruction of evolutionary relationships, and genome assembly and comparison.
- Structural analyses** include protein and nucleic acid structure analysis, comparison, classification, and prediction.
- The **functional analyses** include gene expression profiling, protein–protein interaction prediction, protein subcellular localization prediction, metabolic pathway reconstruction, and simulation.



# Bioinformatics

## Scope

- The **three aspects of bioinformatics analysis are not isolated** but often interact to produce integrated results.
- For example, **protein structure prediction depends on sequence alignment** data.
- **Clustering of gene expression profiles requires the use of phylogenetic tree** construction methods derived in sequence analysis.
- **Sequence-based promoter prediction is related to functional analysis of co-expressed genes.**
- **Gene annotation** involves a number of activities, which include **distinction between coding and noncoding sequences**, identification of **translated protein sequences**, and determination of the **gene's evolutionary relationship** with other known genes; prediction of its cellular functions employs tools from all three groups of the analyses.



Source:

Essential Bioinformatics by Jin Xiong



