أكاديمية كاوست
KAUST ACADEMY

Introduction to Bioinformatics

Day 4: Proteomics and Structural Bioinformatics

15th January 2026

KAUST

Outline Day 4

Morning (9-12 pm):

Proteomics

- **Introduction to Proteomics**
- **Protein Interaction**
- **Structural Proteomics**

Practical Session:

Basic Proteomics Exercises

Afternoon (2-5 pm):

Structural Bioinformatics:

Practical Session:

أكاديمية كاوست
KAUST ACADEMY

Day 4: Proteomics and Structural Bioinformatics

Morning Session

KAUST

أكاديمية كاوست
KAUST ACADEMY

Introduction to Proteomics

KAUST

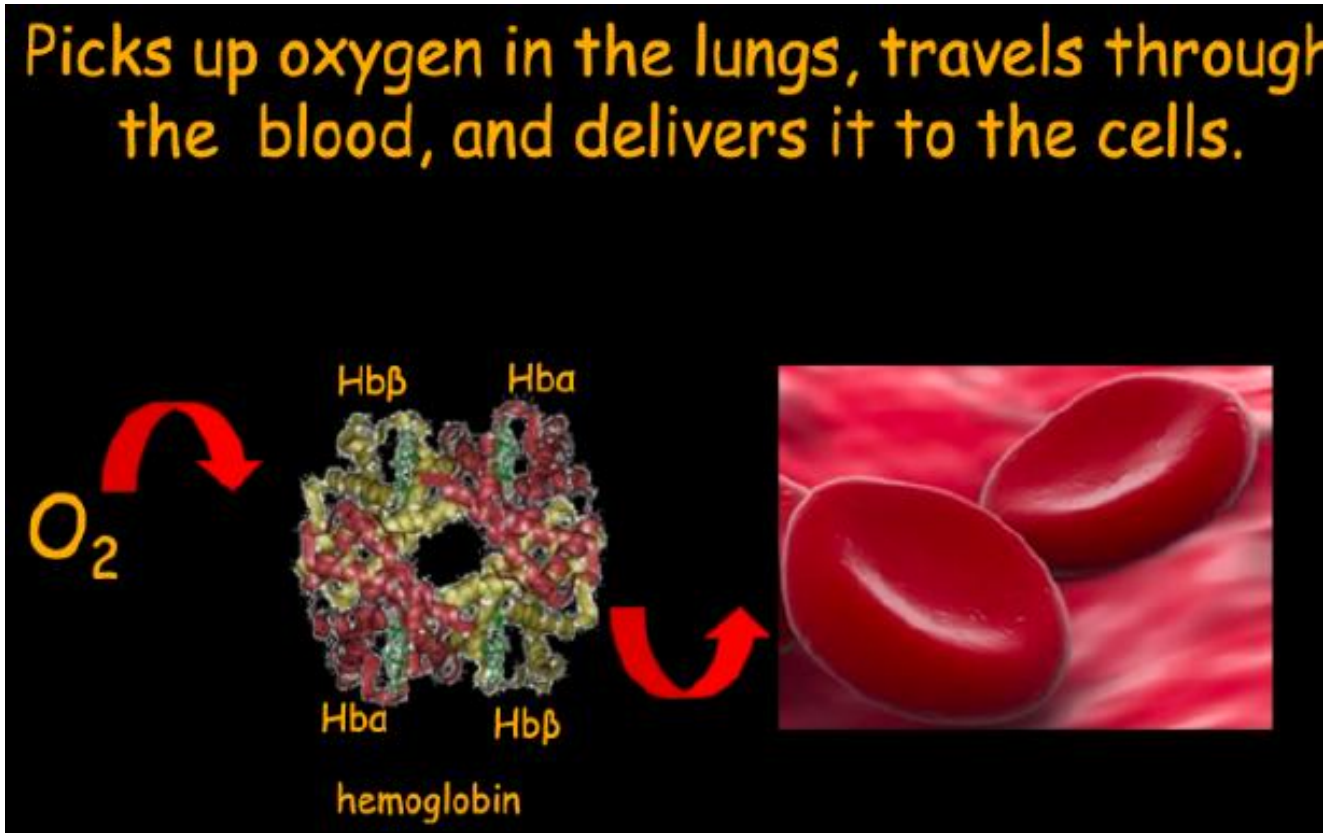
What is a Protein?

Definition:
Essential macromolecules in all living organisms, made up of chains of amino acids (called a polypeptide). A protein contains at least one long polypeptide.

Where we find proteins?
Cells and body tissues, hormones, antibodies, and enzymes.

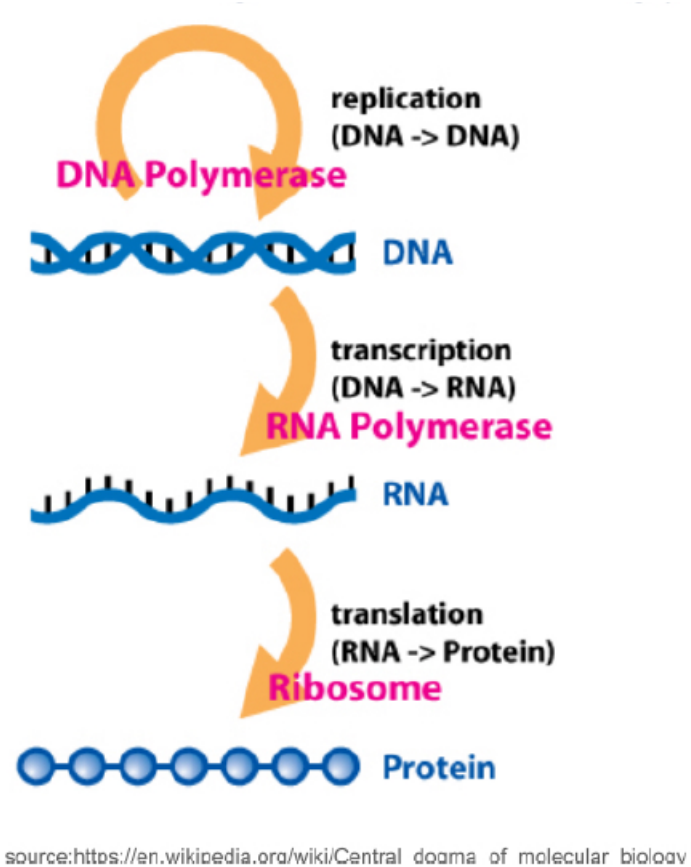
Key Functions of Proteins

Structural Support:
Collagen in connective tissues.
Enzymatic Activity:
Catalysts for biochemical reactions (e.g., amylase).
Transport:
Hemoglobin carries oxygen.
Defense:
Antibodies in the immune system.
Regulation:
Hormones like insulin.

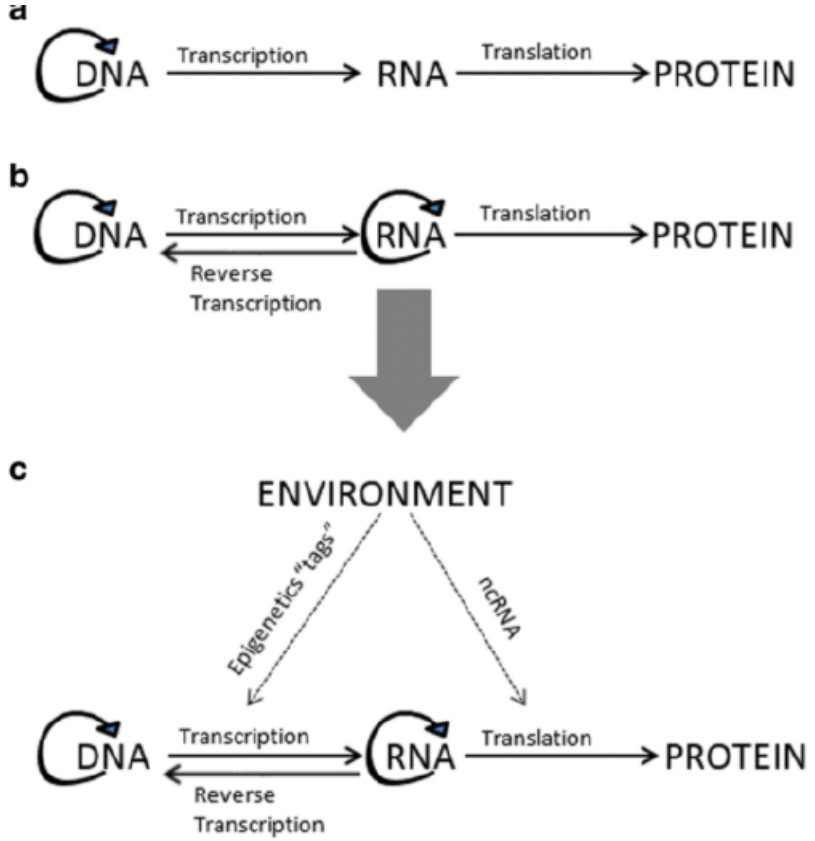


Source: https://slideplayer.com/slide/7042803/

Central Dogma of Biology



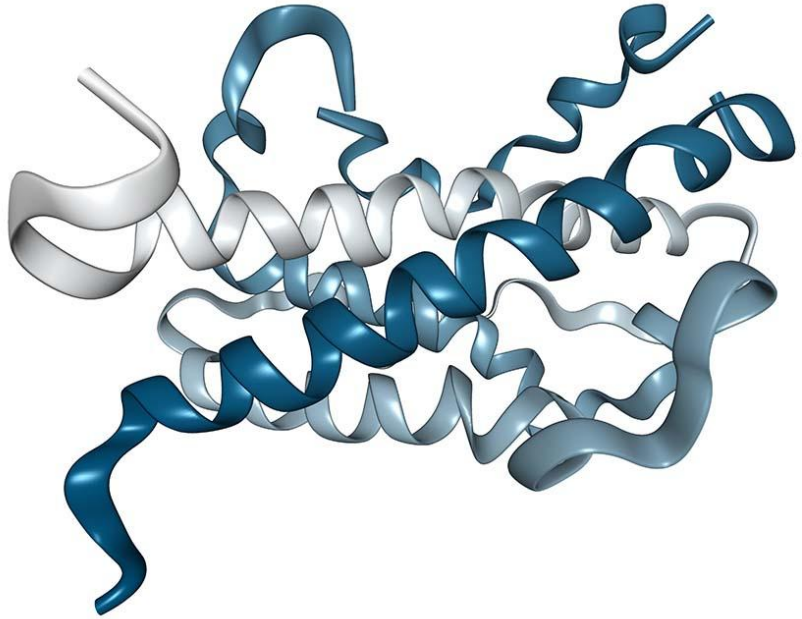
source:https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology



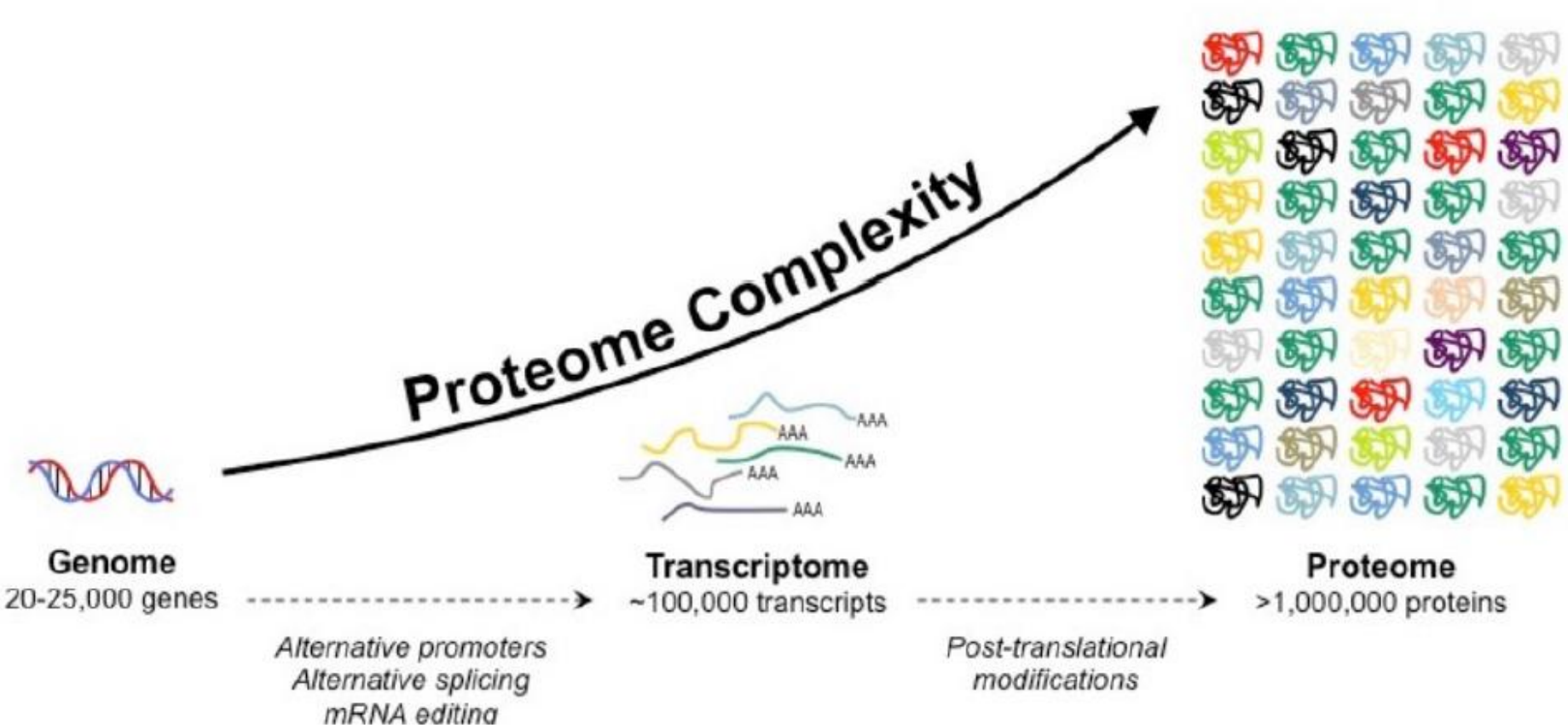
What is Proteome?

- The proteome is the entire set of proteins expressed by an organism, cell, or tissue at a specific time and under specific conditions.

- It is dynamic and influenced by:
  - Gene expression patterns.
  - Environmental conditions.
  - Developmental stages and disease states.

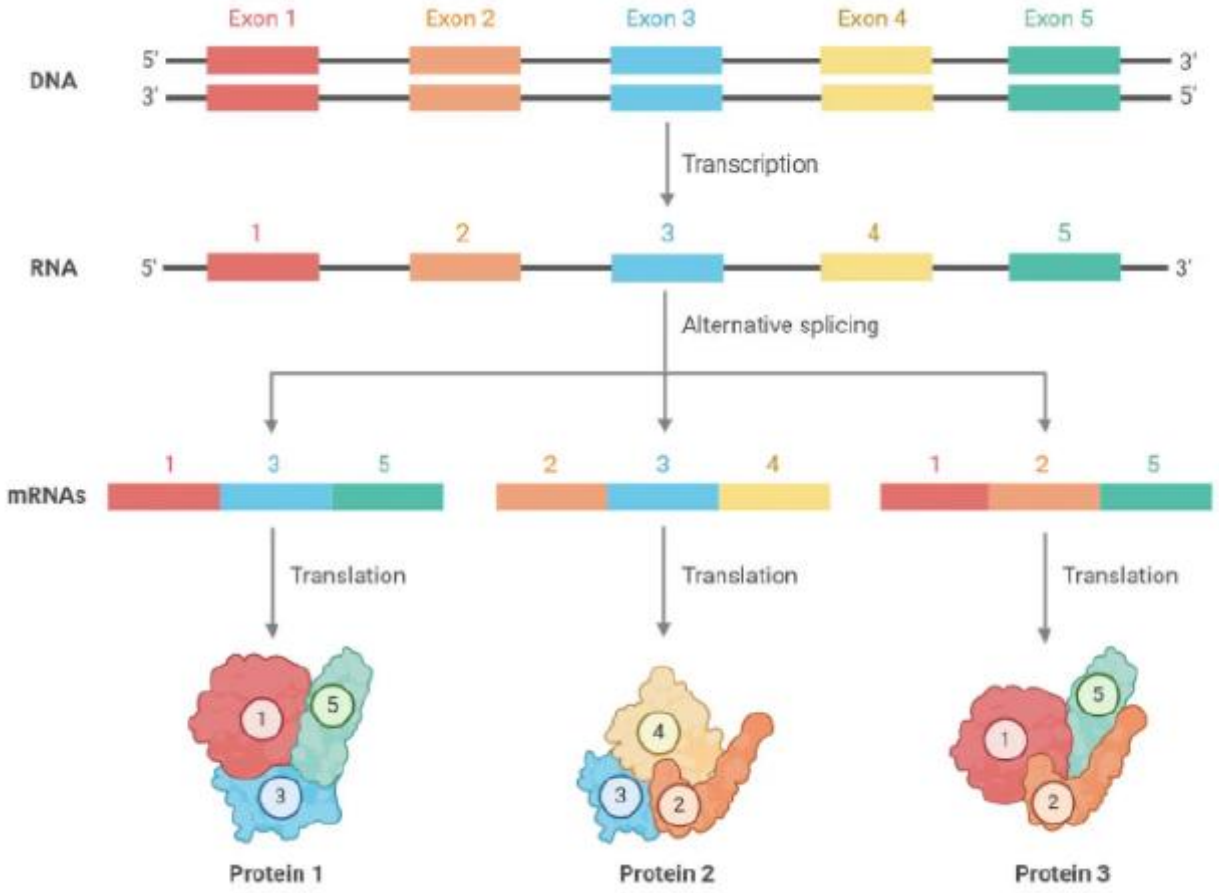


Formation of the Proteome



source:https://old-ib.bioninja.com.au/standard-level/topic-2-molecular-biology/24-proteins/proteome.ht

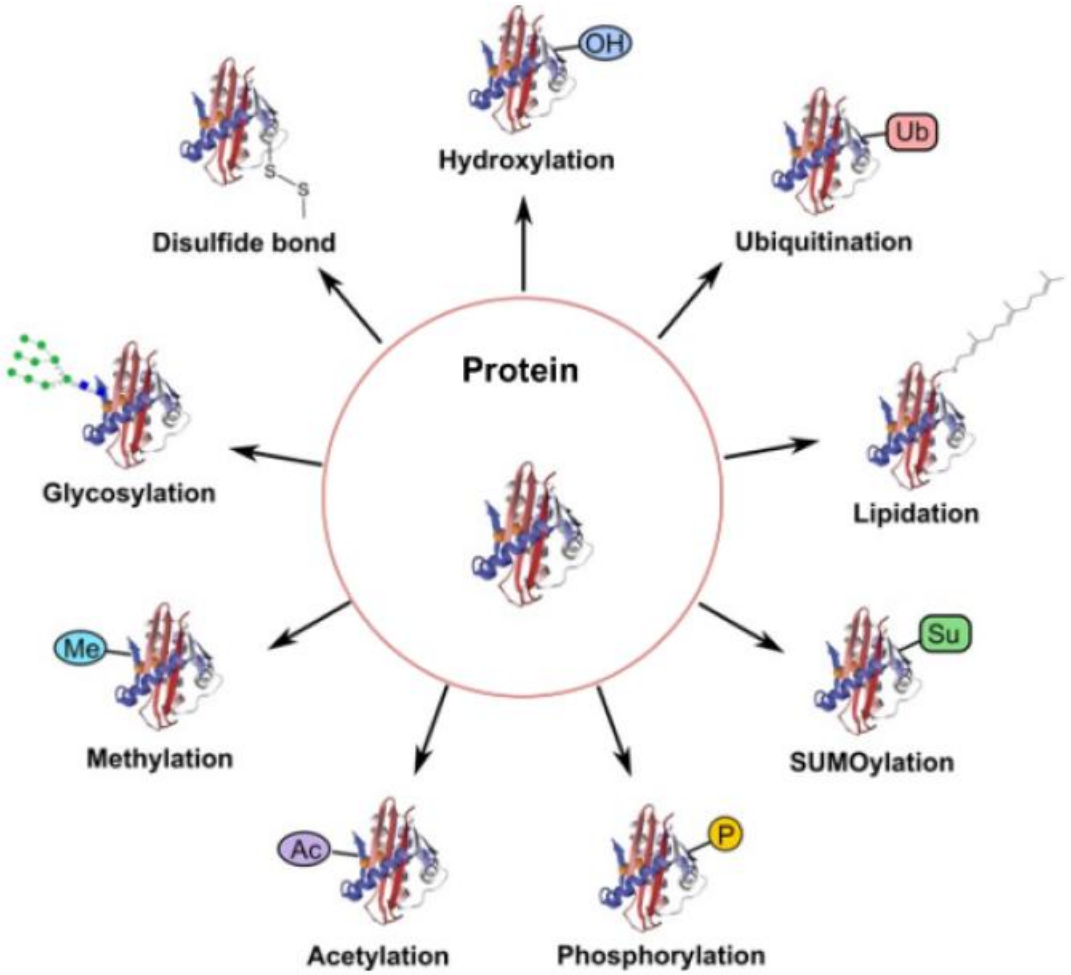
How Alternative splicing increases proteome complexity?



Source: https://microbenotes.com/alternative-splicing/

Post-Translational Modifications (PTMs)

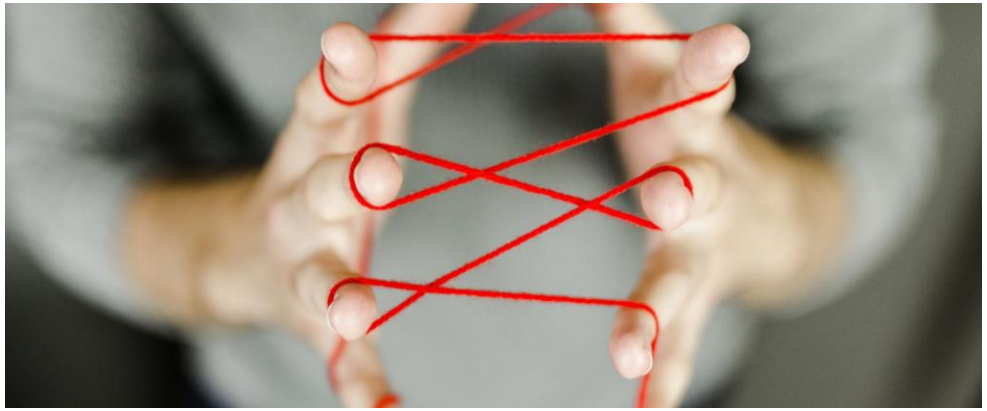
PTM refers to the modification that occurs on a protein after translation catalyzed by enzymes.



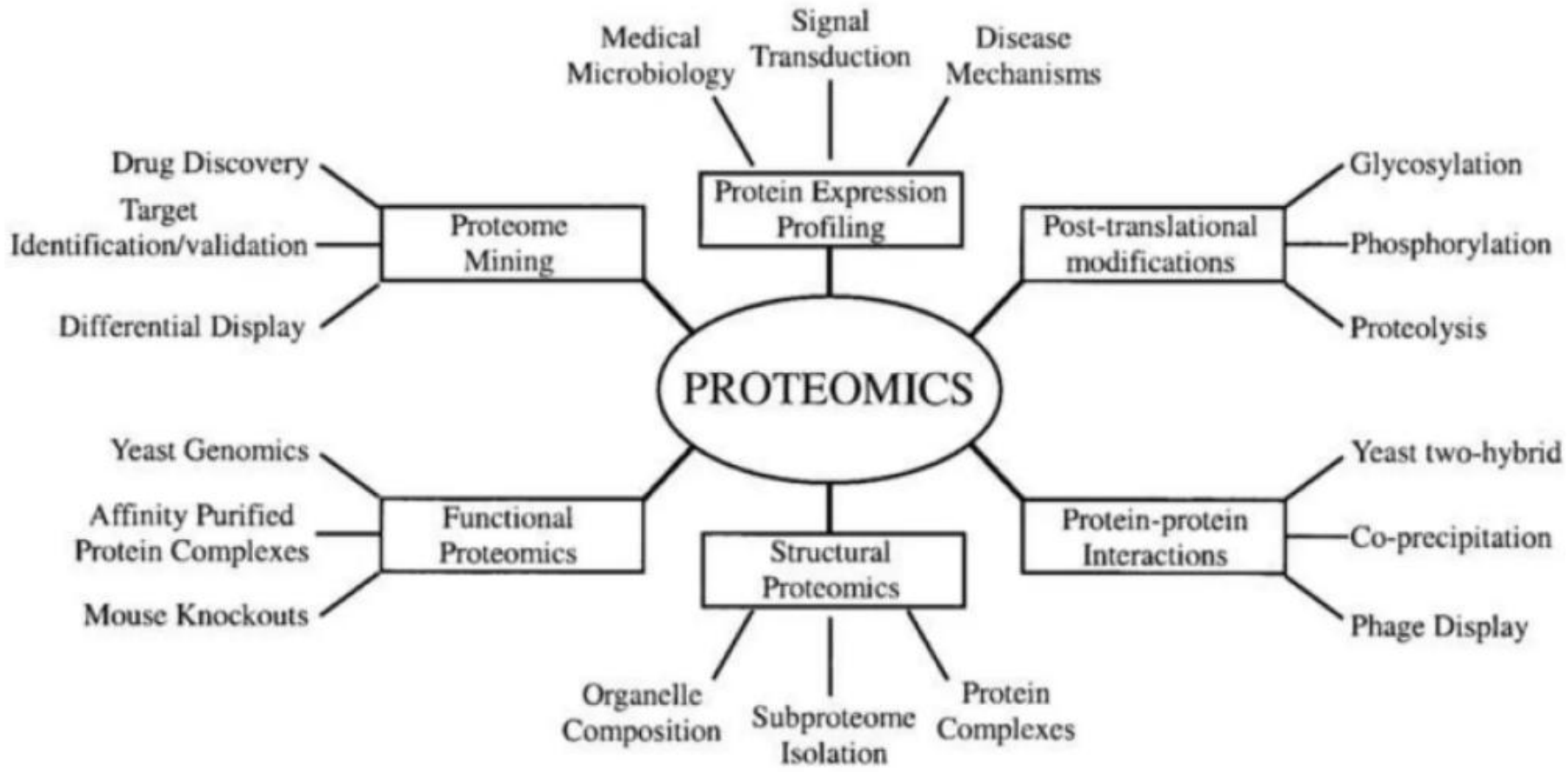
source:https://www.creative-proteomics.com/blog/i ndex.php/strategies-for-post-translational-modificati ons-ptms/

What is Proteomics?

- Proteomics is the large-scale study of proteins, focusing on:
  Structure, function, and interactions.
- It aims to:
  Identify and quantify all proteins in a system.
  Understand biological processes and pathways.
  Discover biomarkers for diseases or therapeutic targets.



Applications of Proteomics



Three Main Types of Proteomics

Expressional Proteomics

Concerned with the display, measurement and analysis of global changes in protein expressions

Functional Proteomics

Study of protein functions, interactions, and modifications within biological systems to understand their roles in cellular processes, disease mechanisms, and therapeutic applications.

Structural Proteomics

Focuses on determining the three-dimensional structures of proteins and their complexes to understand their functional roles, interactions, and contributions to biological processes.

Three Main Types of Proteomics

Expressional Proteomics

Concerned with the display, measurement and analysis of global changes in protein expressions

Functional Proteomics

Study of protein functions, interactions, and modifications within biological systems to understand their roles in cellular processes, disease mechanisms, and therapeutic applications.

Structural Proteomics

Focuses on determining the three-dimensional structures of proteins and their complexes to understand their functional roles, interactions, and contributions to biological processes.

Why Study Expressional Proteomics?

Understand Disease Mechanisms: Identify how protein expression changes in diseases like cancer or diabetes.

Biomarker Discovery: Detect proteins as diagnostic, prognostic, or therapeutic biomarkers.

Therapeutic Targeting: Uncover proteins that can serve as drug targets.

Personalized Medicine: Correlate protein expression patterns with individual responses to treatments.

Adaptation and Stress Response: Study how organisms respond to environmental stress at the protein level.

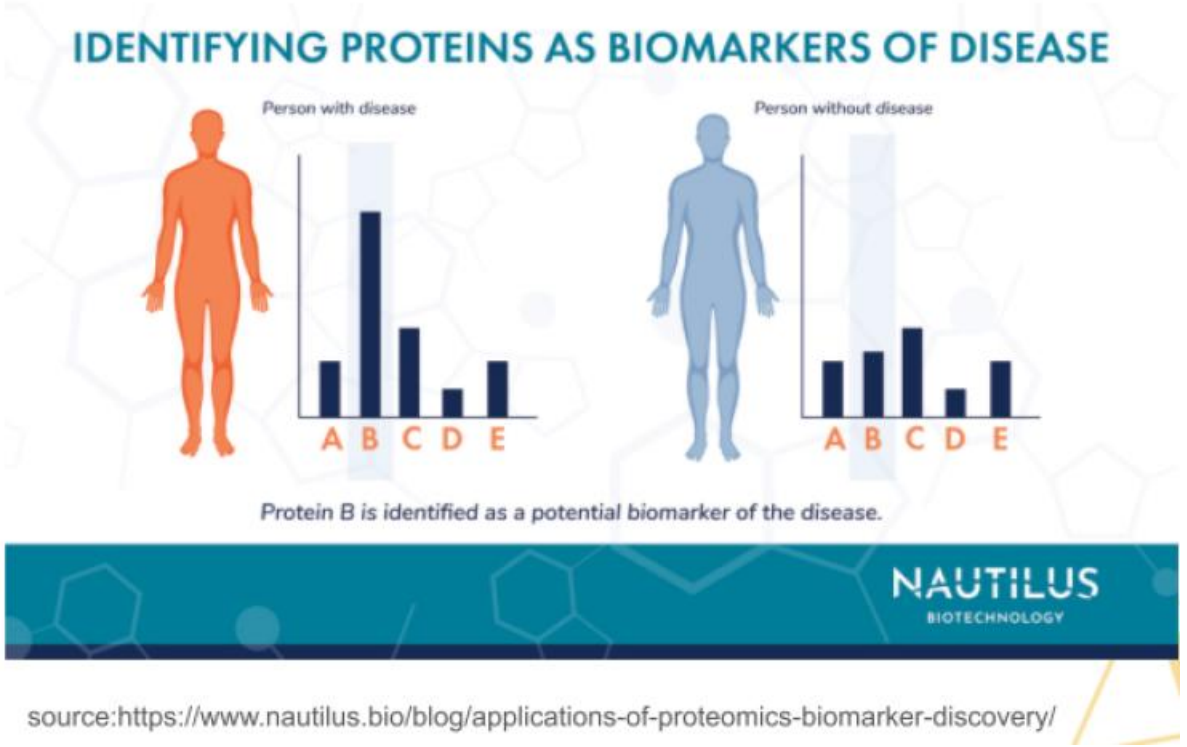
Advancing Biotechnology: Improve processes like enzyme engineering or crop improvement through protein expression insights.

Example Application of Expressional Proteomie Biomarker Discovery in Cancer

Biomarkers-proteins that indicate the presence or progression of diseases such as cancer. E.g.

- **Objective:** Identify proteins uniquely expressed in cancer cells versus normal cells.
- **Process:** Mass spectrometry and protein microarrays analyze protein expression in patient samples.
- **Outcome:** Potential biomarkers like PSA for prostate cancer or HER2 for breast cancer, which guide early diagnosis and personalized treatment.

Impact: Enables targeted therapies and improves patient outcomes through precision medicine.



source:https://www.nautilus.bio/blog/applications-of-proteomics-biomarker-discovery/

A Generic Proteomics Workflow

Step 1: Sample Preparation: Extract and purify proteins from biological samples.

Step 2: Peptide Separation:

2D-PAGE: Resolves proteins by isoelectric point and molecular weight.

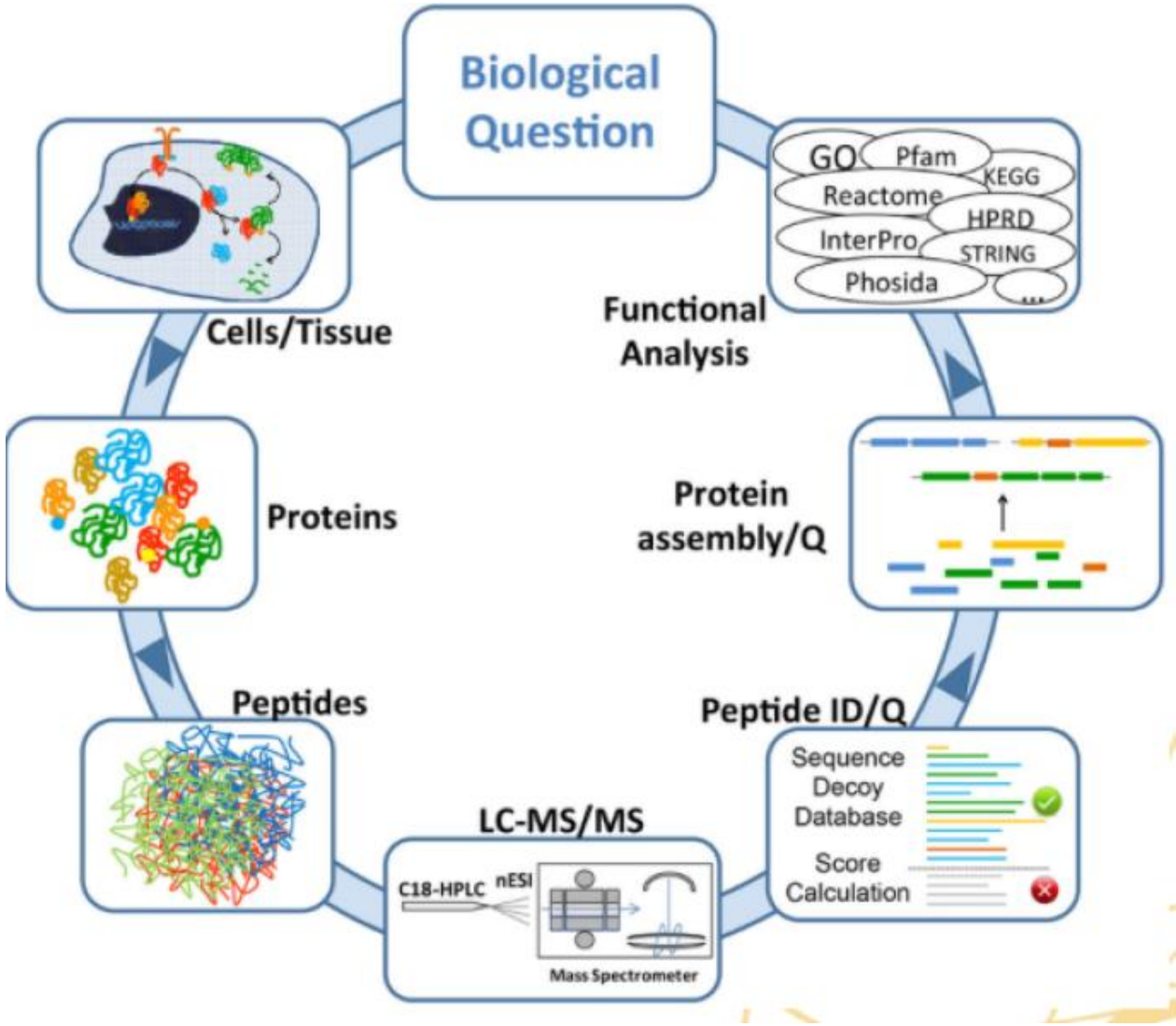
HPLC: Separates proteins/peptides based on their chemical properties.

Step 3: Mass Spectrometry (MS):

Process: Ionizes proteins, detects mass-to-charge ratios, and analyzes data.

Output: Provides protein identification and quantification.

Step 4: Bioinformatics Analysis: Utilizes computational tools to analyze protein data and map pathways, enabling functional insights.



Why Study Functional Proteomics?

Functional proteomics bridges the gap between protein expression and biological roles, enabling deeper insights into health and disease.

- **Protein Function Insight:** Understand how proteins interact and carry out biological roles.
- **Disease Mechanisms:** Study functional disruptions in diseases like cancer and Alzheimer's
- **Drug Development:** Identify and validate therapeutic targets and pathways.
- **Cellular Pathways:** Explore signal transduction and metabolic networks in living organisms.
- **Precision Medicine:** Correlate protein functions with specific patient conditions for tailored therapies.
- **Dynamic Protein Interactions:** Investigate real-time changes in response to environmental or cellular stimuli.

Why Study Functional Proteomics?

Functional proteomics bridges the gap between protein expression and biological roles, enabling deeper insights into health and disease.

- **Protein Function Insight:** Understand how proteins interact and carry out biological roles.
- **Disease Mechanisms:** Study functional disruptions in diseases like cancer and Alzheimer's
- **Drug Development:** Identify and validate therapeutic targets and pathways.
- **Cellular Pathways:** Explore signal transduction and metabolic networks in living organisms.
- **Precision Medicine:** Correlate protein functions with specific patient conditions for tailored therapies.
- **Dynamic Protein Interactions:** Investigate real-time changes in response to environmental or cellular stimuli.

Functional Proteomics Example: Drug Target Identification

- **Purpose:** To find proteins that can be targeted by new medicines.

- **How it works:**
  - Scientists study how proteins interact with each other in the body (e.g.Yeast-two-hybrid).
  - They identify specific proteins involved in diseases.
  - These proteins are tested to see if they can be affected by drugs.

- **Real-life Example:**
  - In cancer, researchers found a protein called BCR-ABL causing leukemia.
  - A drug, **Imatinib (Gleevec)**, was developed to block it and successfully treat the disease.

- **Impact:** Functional proteomics helps create targeted treatments, improving patient outcomes.

Protein Interactions

Biological Importance: Understanding PPIs and other interactions is essential for drug design, disease mechanisms, and cellular function.

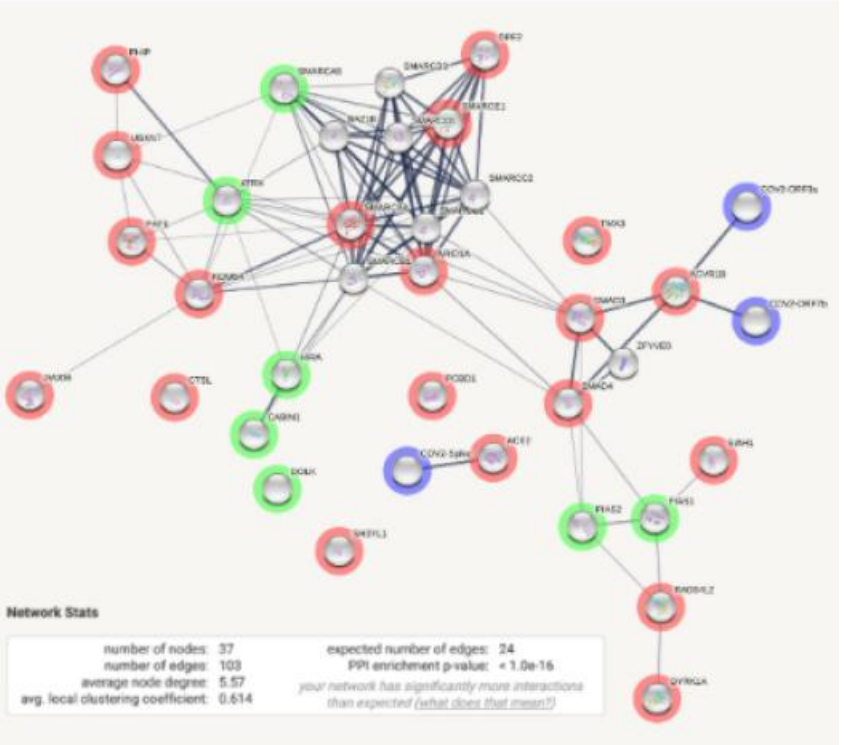
Types:

- **Protein-Protein Interactions (PPIs):** Crucial for cellular processes like immune response, and regulation (e.g., enzymes, receptors).
- **Protein-DNA/RNA Interactions:** Important in gene expression, replication, and repair (e.g., transcription factors).
- **Protein-Ligand Interactions:** Enable metabolic processes and regulatory activities (e.g., hormones binding to receptors).

How to identify PPIs?

1. **Experimental Techniques:**
   - **Yeast Two-Hybrid (Y2H):** Detects physical interactions in living cells.
   - **Co-Immunoprecipitation (Co-IP):** Pulls down interacting proteins using specific antibodies.
   - **Mass Spectrometry (MS):** Identifies interaction partners in protein complexes.

2. **Computational Methods:**
   - **Docking Algorithms:** Predicts interaction interfaces.
   - **Text Mining:** e.g. the STRING database



Why Study Structural Proteomics?

Structural proteomics is vital for deciphering the molecular mechanisms underlying biological functions and therapeutic innovations.

- **Understand 3D Protein Architecture:** Reveals how protein structures enable their function.
- **Disease Insight:** Identifies structural changes in proteins associated with diseases.
- **Drug Discovery:** Assists in designing molecules that fit specific protein targets (e.g., inhibitors, activators).
- **Protein-Protein Interactions:** Explores binding interfaces for complex formation and signal transduction.
- **Functional Correlation:** Links structure with biological roles and cellular mechanisms.
- **Evolutionary Studies:** Analyzes structural conservation and variations across species.

Structural Proteomics Example: Understanding Disease Mechanisms

- **Purpose:** To study the 3D shapes of proteins and understand how they contribute to diseases.
- **How it works:**
  - Scientists analyze protein structures to understand how they work in the body.
  - Structural information helps in discovering how proteins malfunction in diseases like Alzheimer's or cancer.
  - By understanding these structures, they can develop drugs that specifically target abnormal proteins.
- **Real-life Example:**

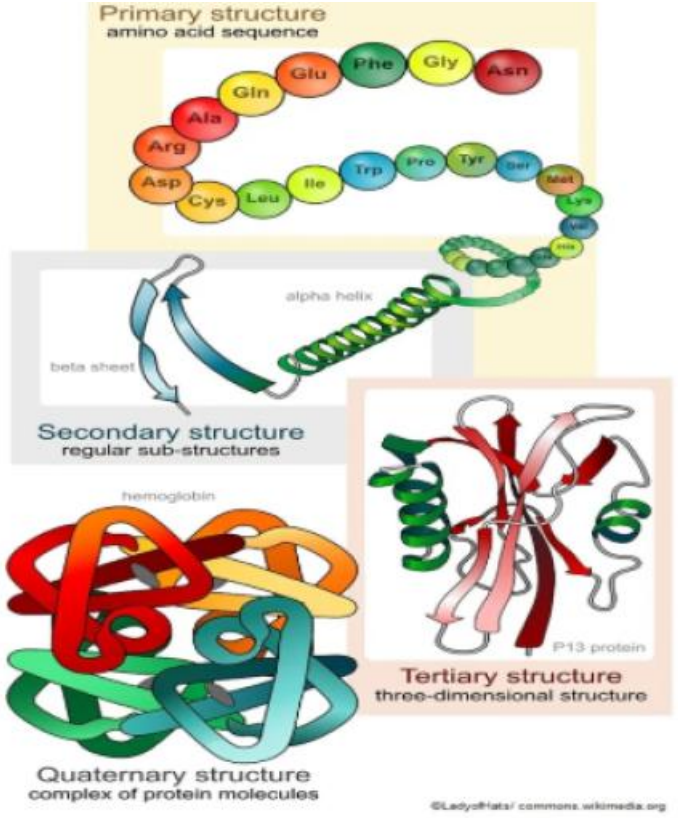
- **Alzheimer's Disease:**
  - Structural proteomics helped reveal the structure of **amyloid-beta plaques** (protein clumps in the brain).
  - This led to research on therapies aimed at preventing plaque formation, which is linked to the disease's progression.

Impact:

- Structural proteomics plays a crucial role in drug discovery by providing detailed protein structures, enabling the development of more precise treatments.

Protein Structure

- Primary: the exact ordering of amino acids forming their chains.
- Secondary: local folded structures that form within a polypeptide due to interactions between atoms of the backbone.
- Tertiary: represents overall folding of the polypeptide chains, further folding of the secondary structure.
- Quaternary: the spatial arrangement of various tertiary structures gives rise to the quaternary structure.



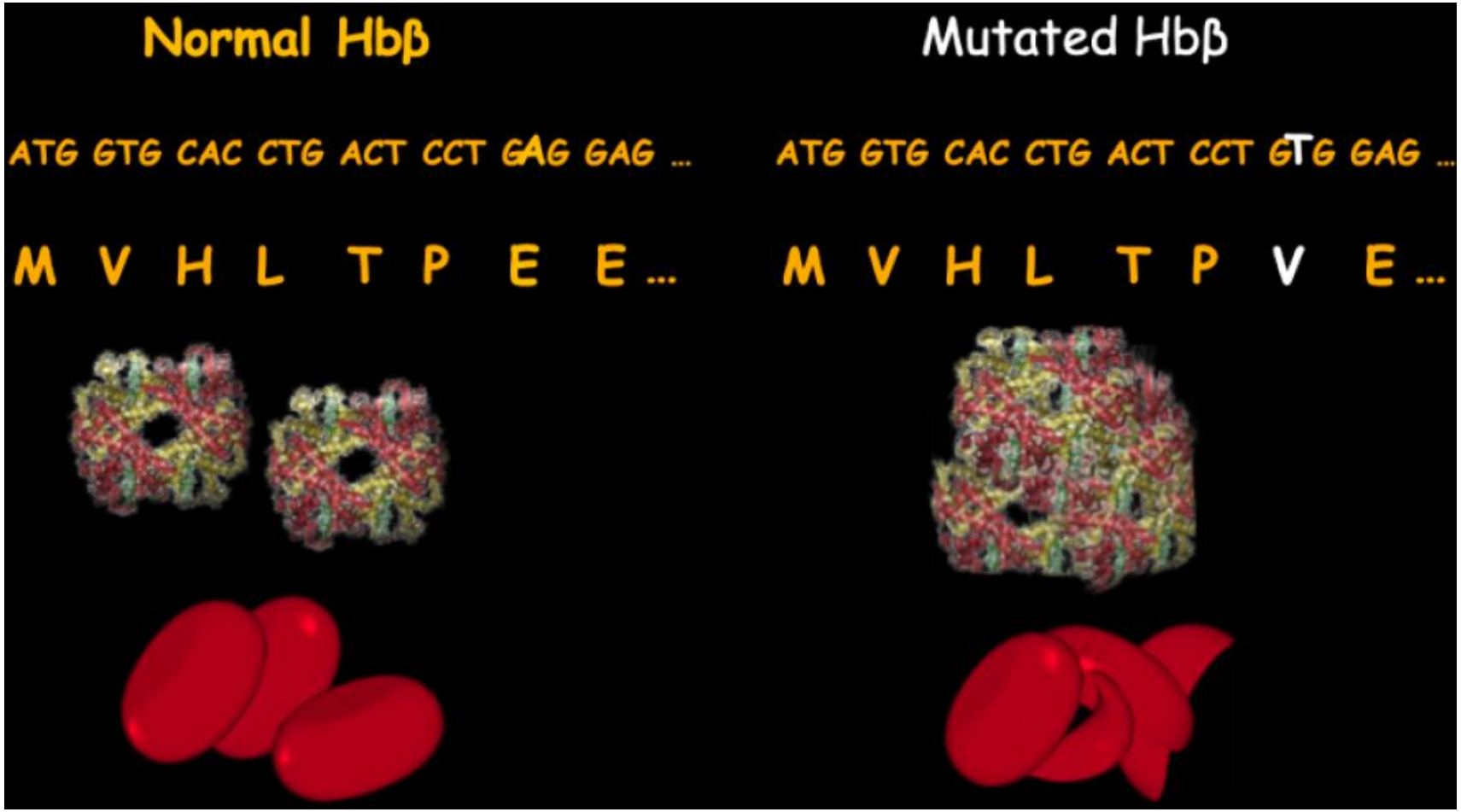
source:https://www.news-medical.neUlife-sciences/Protein-Structl.Jre-and-Function.aspx

Protein structure determines function

- Each protein has a unique sequence of amino acids, and the interactions among these amino acids determine its specific shape.

- This shape is crucial for the protein's function, whether it's digesting food in the stomach or transporting oxygen in the blood.

Sickle Cell Disease is caused by single amino acid change



Source: https://slideplayer.com/slide/7042803/

How to predict protein function?

أكاديمية كاوست
KAUST ACADEMY

- Understanding protein function is key to drug development, disease research, and biotechnology.

- **Experimental Methods**:
  - **Mutagenesis Studies:** Alter genes to observe changes in protein function.
  - **Biochemical Assays:** Test enzymatic activities or binding properties.
  - **Localization Studies:** Use techniques like fluorescence microscopy to determine where the protein works.
- **Bioinformatics Approaches:**
  - **Sequence Analysis:** Compare to known proteins using BLAST.
  - **Structure-Function Relationship:** Predict function based on 30 structure.
  - **Databases**: Use UniProt or Pfam to access functional annotations.
- **Omics Studies**:
  - Transcriptomics and proteomics link expression levels to biological roles.

Protein data

Protein Structure - Protein DataBank

- A global database for storing 3D structural data of biological molecules like proteins, DNA, and RNA.
- Provides atomic-level details for research in biology, drug design, and bioinformatics.
- Helps visualize molecular structures and their interactions.
- Contains experimentally validated entries

RCSB PDB
PROTEIN DATA BANK

https://www.rcsb.org/

Protein sequences and functions: UniProt Knowledge Base

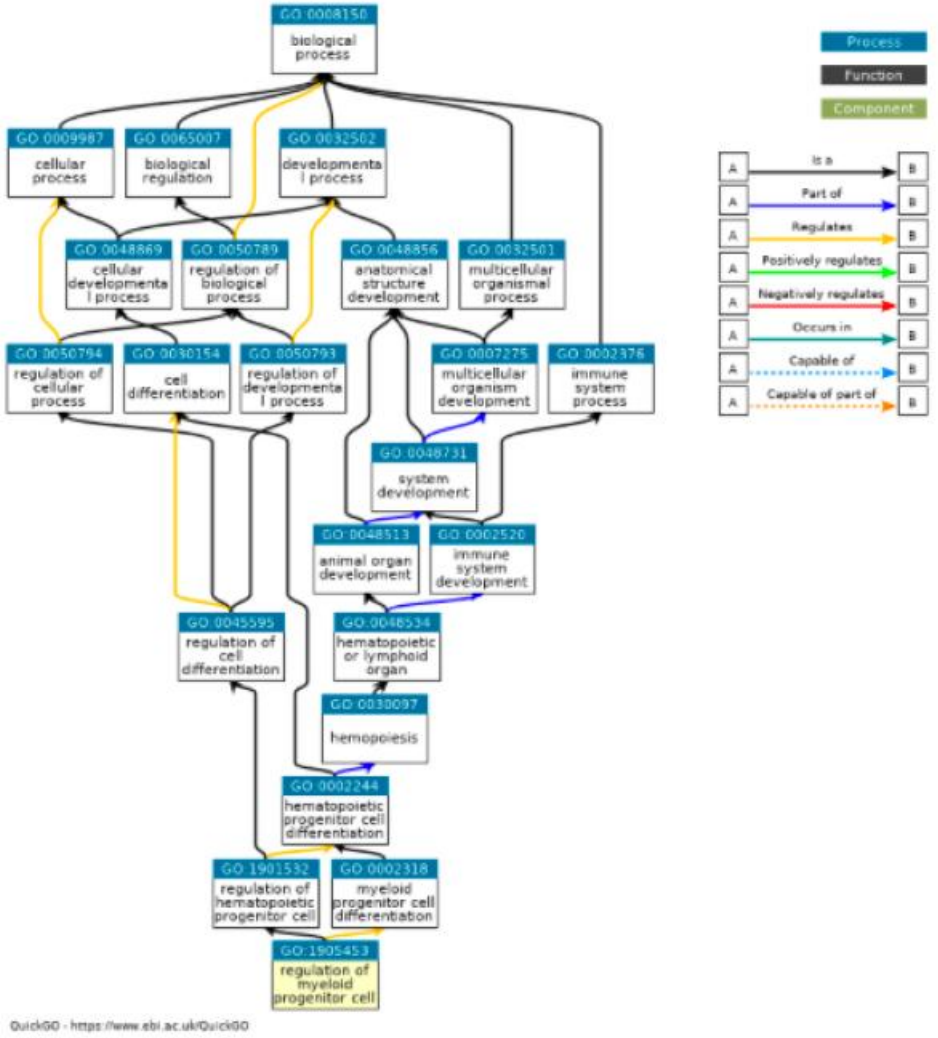
UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information.

- Around 250 Million protein sequences
- More than 550 thousand reviewed and manually annotated proteins

https://www.uniprot.org/

Classification of Protein Functions - Gene Ontology

- Biological Process - processes to which gene or gene product contributes
- Molecular Function - biochemical activity of a gene product
- Cellular Component - location of a gene product where it is active



InterPro: Protein Families and Domains

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites.

www.ebi.ac.uk/interpro

Protein families and functions: Pfam

- Focuses on **functional annotation** of protein sequences by identifying conserved domains and families.
- Uses computational methods for high-accuracy predictions.
- Essential for understanding protein roles and evolutionary relationships

pfam.xfam.org

PPIs: STRING, BioGRID, IntAct

- **STRING**:
  - Integrates known and predicted protein-protein interactions.
  - Includes direct (physical) and indirect (functional) associations.
  - string-db.org

- **BioGRIO**:
  - A curated database of physical and genetic interactions.
  - Focuses on experimentally verified data.
  - thebiogrid.org

- **IntAct:**
  - Provides detailed interaction data, including experimental conditions.
  - Allows network visualization and analysis.
  - ebi.ac.uk/intact

أكاديمية كاوست
KAUST ACADEMY

Day 4: Proteomics and Structural Bioinformatics

Practical Morning Session

KAUST

أكاديمية كاوست
KAUST ACADEMY

Hands-on: Basic Proteomics Exercises

KAUST

أكاديمية كاوست
KAUST ACADEMY

Day 4: Proteomics and Structural Bioinformatics

Afternoon Session

KAUST

أكاديمية كاوست
KAUST ACADEMY

Introduction to Structural Bioinformatics

KAUST

What is Structural Bioinformatics?

Computational methods to study the 3D structures of proteins and nucleic acids, combining biology, chemistry, physics, and computer science to predict and analyze molecular structures

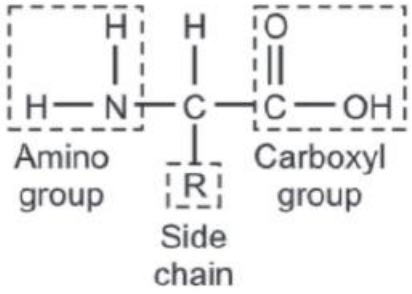
Applications of Structural Bioinformatics

Accelerates precision medicine, improves biomolecular design, and advances our understanding of life at the molecular level.

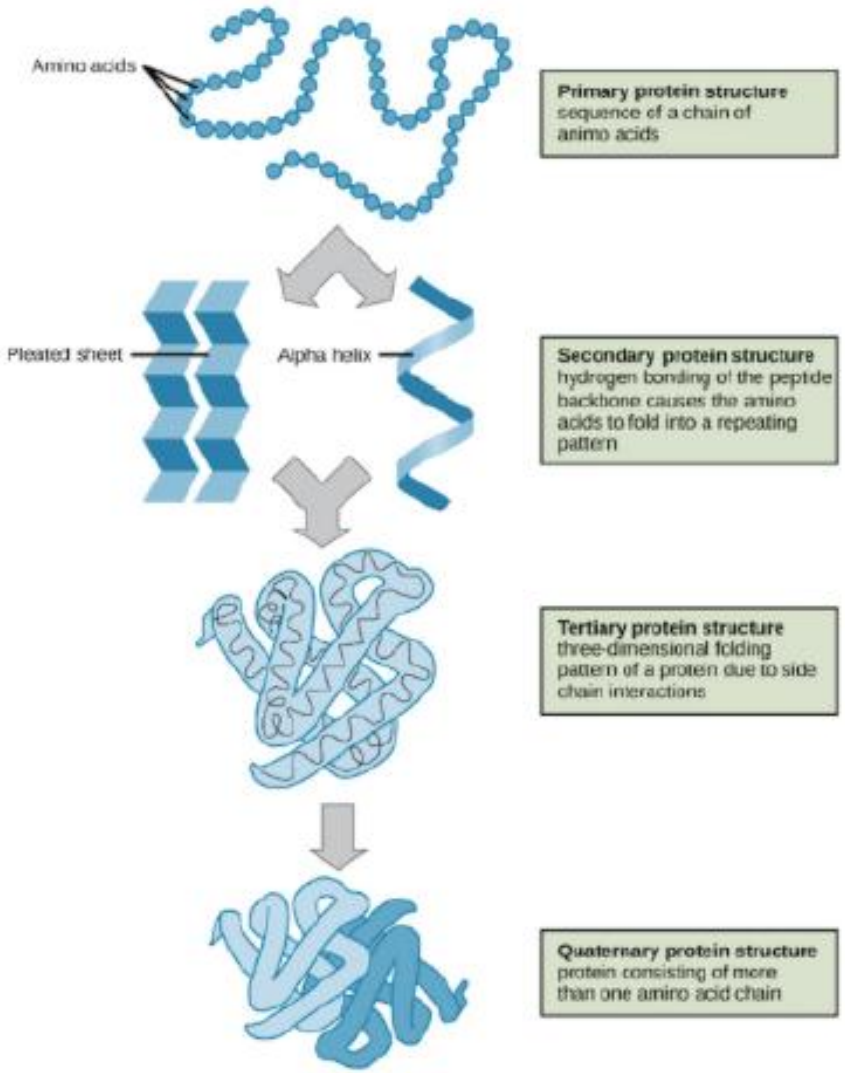
- **Drug Discovery:** Design and optimization of small molecules, vaccines, and antibody therapeutics.
- **Protein Engineering:** Designing enzymes, improving stability, and creating novel functions.
- **Disease Mechanisms:** Understanding molecular basis of diseases and identifying potential therapeutic targets.
- **Functional Annotation:** Predicting functions of uncharacterized proteins using structural information.
- **Evolutionary Studies:** Exploring protein evolution and phylogenetics.

What is protein structure?

- **Protein structure** refers to the three-dimensional arrangement of amino acids in a protein molecule.
- **Functionality:** A protein's structure determines its role in biological processes, such as enzyme activity or signaling.
- **Specificity:** Small changes in structure (e.g., misfolding) can lead to diseases like Alzheimer's or cystic fibrosis.



Structural levels of proteins (Hierarchical)



source:https://biologyease.com/protein-structur1;1s/#google_vignette

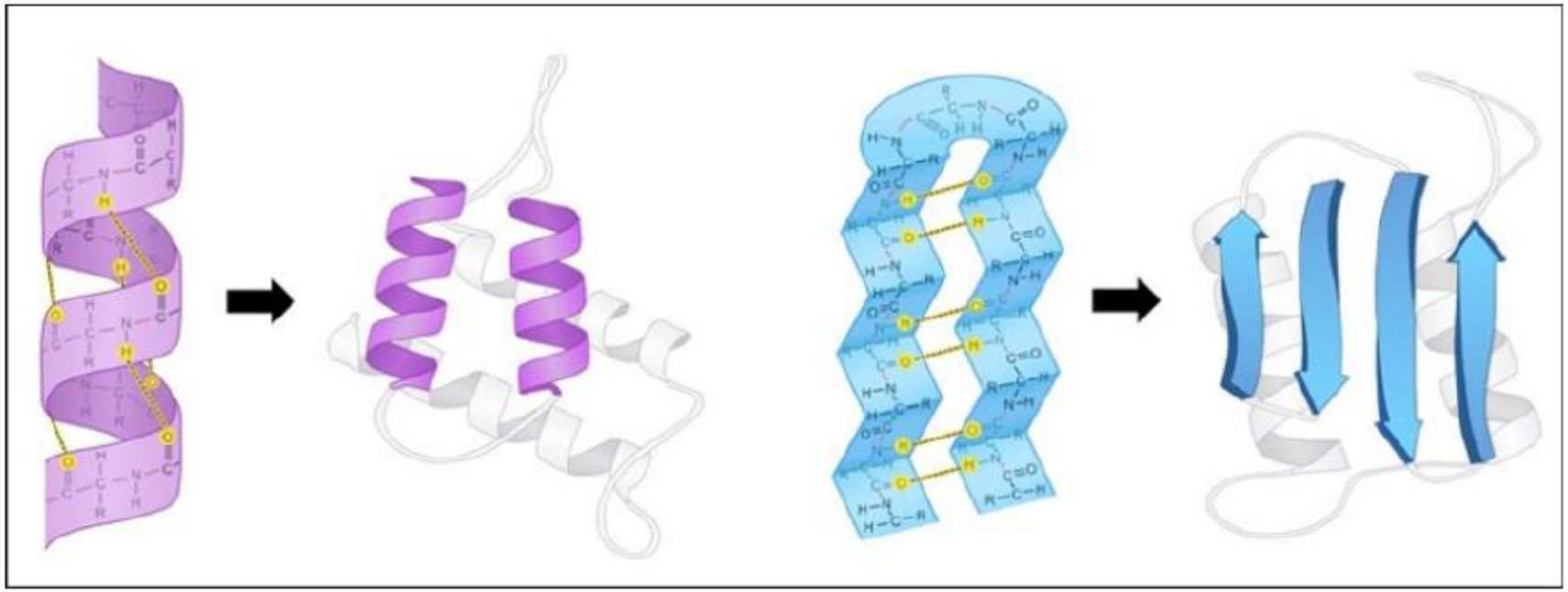
Levels of protein structure: Primary Structure

A protein's primary structure refers to the amino acid sequence of its polypeptide chain.

MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGINGTKRFDNPVLP
FNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEY
VSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSS
SGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGE
VFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKL
PDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRV
VVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGV
SVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS
PRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNR
ALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLI
CAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQ
DSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVT
QRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVA
KNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT

Levels of protein structure: Secondary Structure

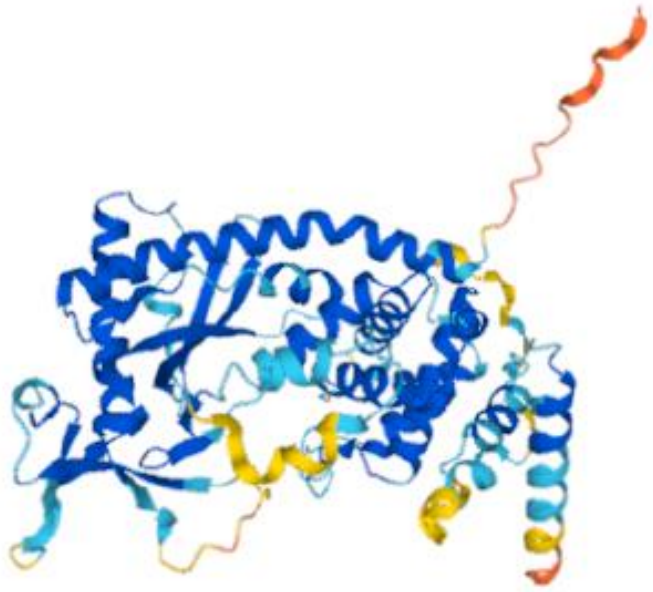
A secondary structure is a repeating substructure that forms as a substructure of the overall folded protein.



https://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/73-translation/protein-structure.html

Levels of protein structure: Tertiary Structure (3D)

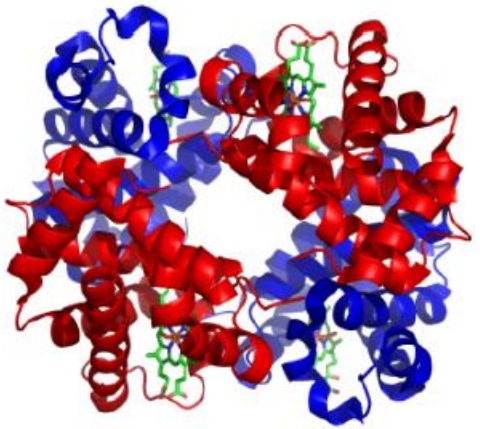
A protein's tertiary structure describes its final 3D shape after the polypeptide chain has folded and is chemically stable. This is what we most commonly refer to as the "structure" of a protein.



Levels of protein structure: Quaternary Structure

Some proteins have a quaternary structure, which describes the protein's interaction with other copies of itself to form a single functional unit, or a multimer.

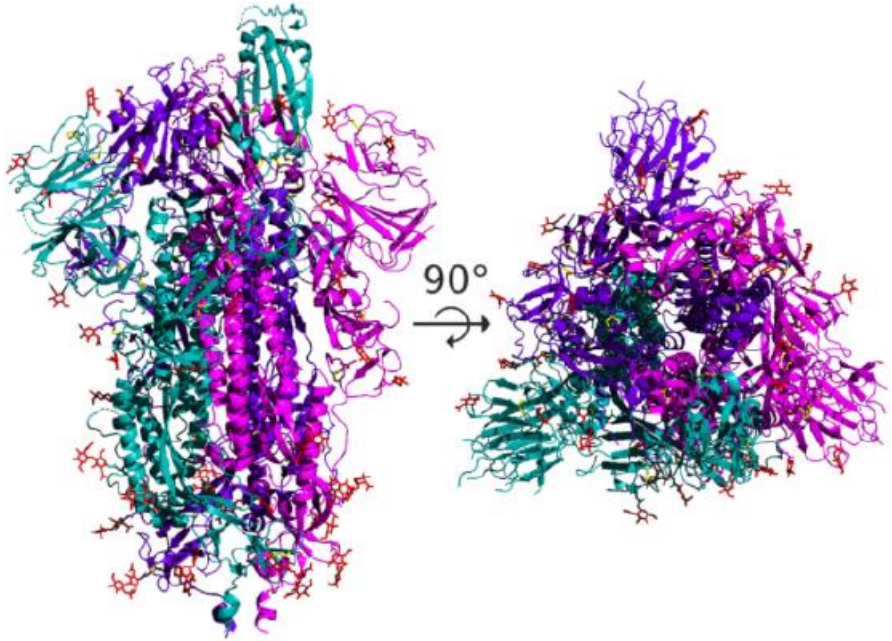
Hemoglobin is a multimer consisting of two alpha subunits and two beta subunits.



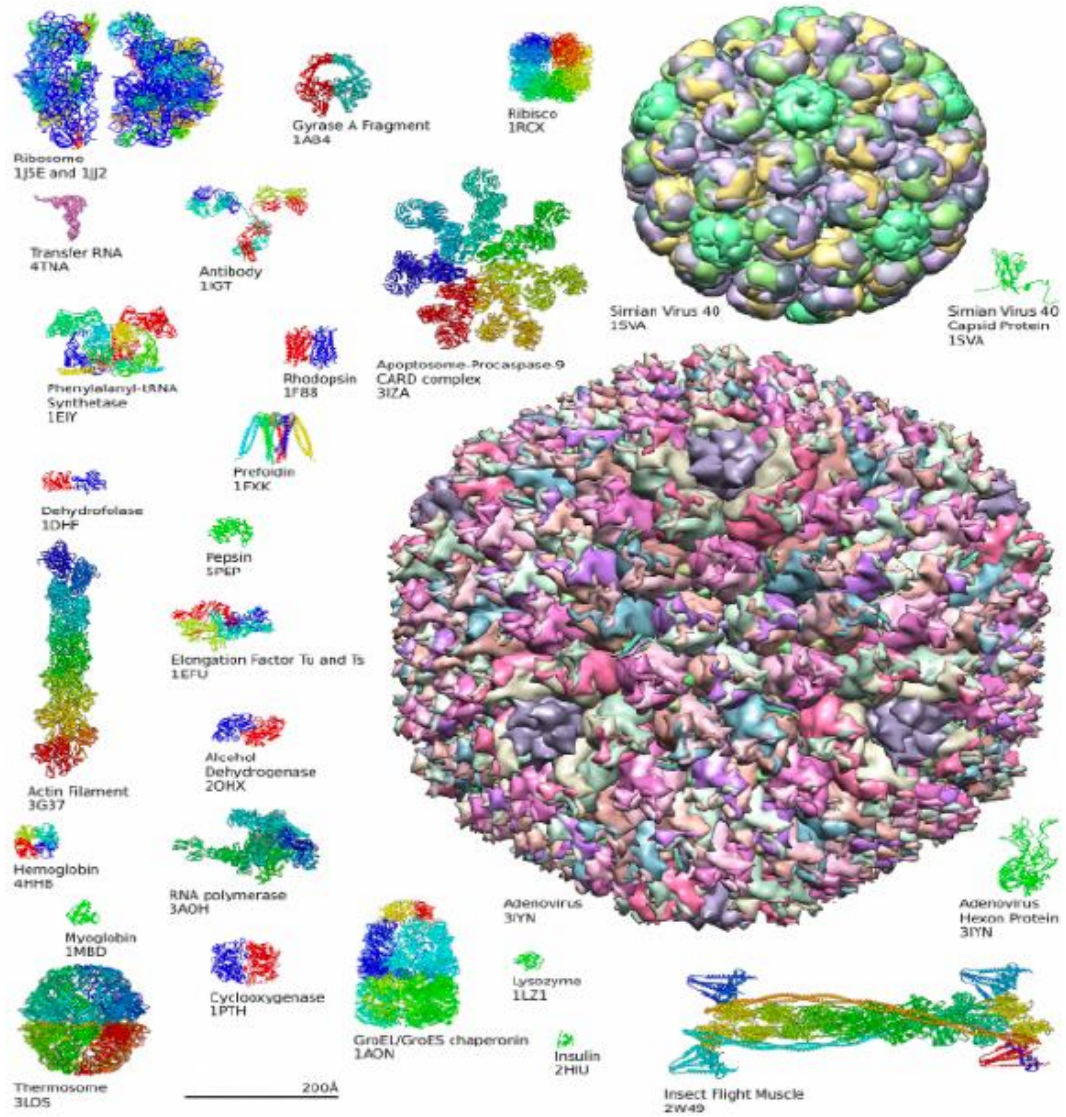
https://commons.wikimedia.org/wiki/File:1GZX_Haemoglob

Example: Spike protein

The spike protein is a homotrimer, formed of three essentially identical units called chains, each one translated from the same genome region.



Proteins come in all different shapes



https://en.wikipedia.org/wiki/Protein_structure

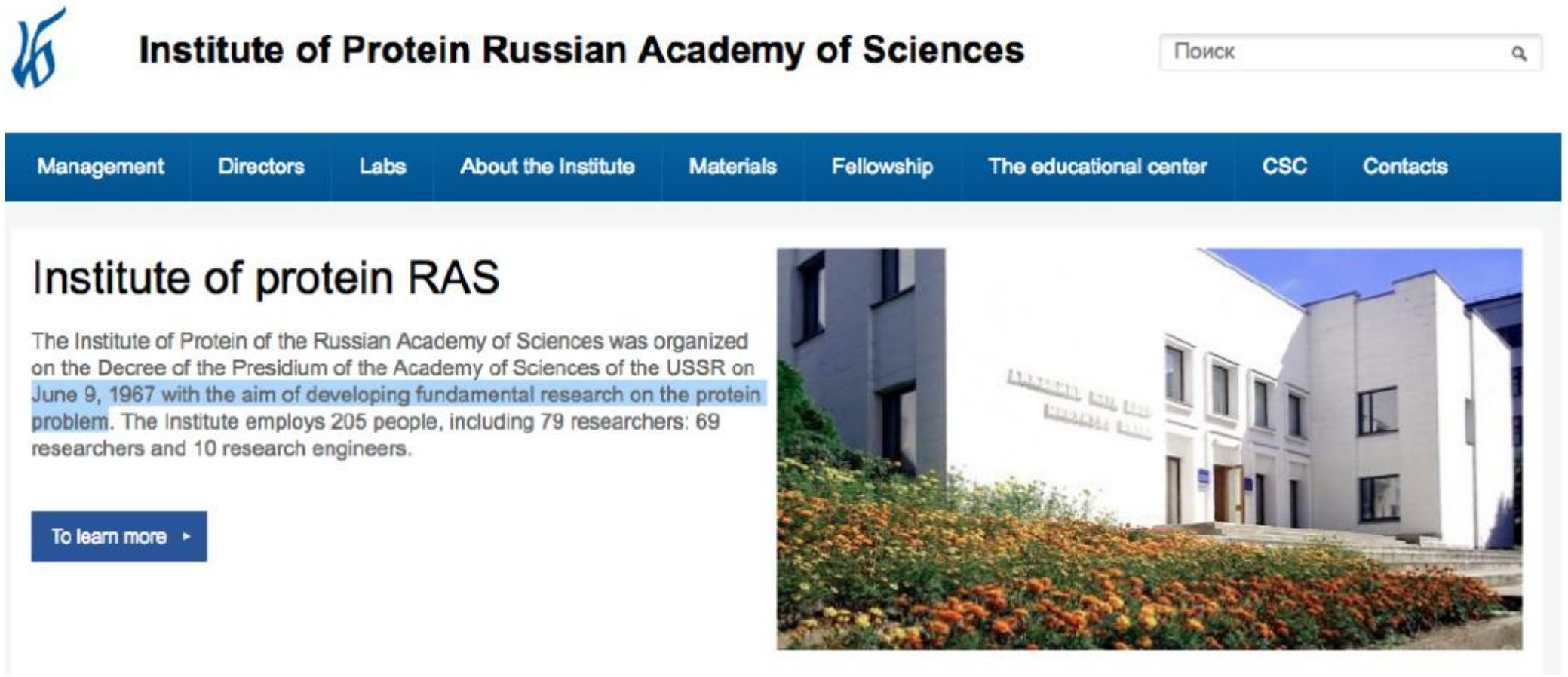
How do we determine protein structure?

أكاديمية كاوست
KAUST ACADEMY

- **Experimental Methods:**
  - **X-ray Crystallography:** Determines atomic structure via X-ray diffraction from protein crystals.
  - **NMR Spectroscopy:** Uses magnetic fields to study proteins in solution.
  - **Cryo-Electron Microscopy:** Visualizes proteins at near-atomic resolution without crystallization.

- **Computational Methods:**
  - **Homology Modeling:** Predicts structure based on known templates.
  - **Molecular Dynamics:** Simulates atomic interactions.
  - **Deep Learning Models:** Tools like **AlphaFold** predict structures using AI by learning patterns from large datasets of protein structures.

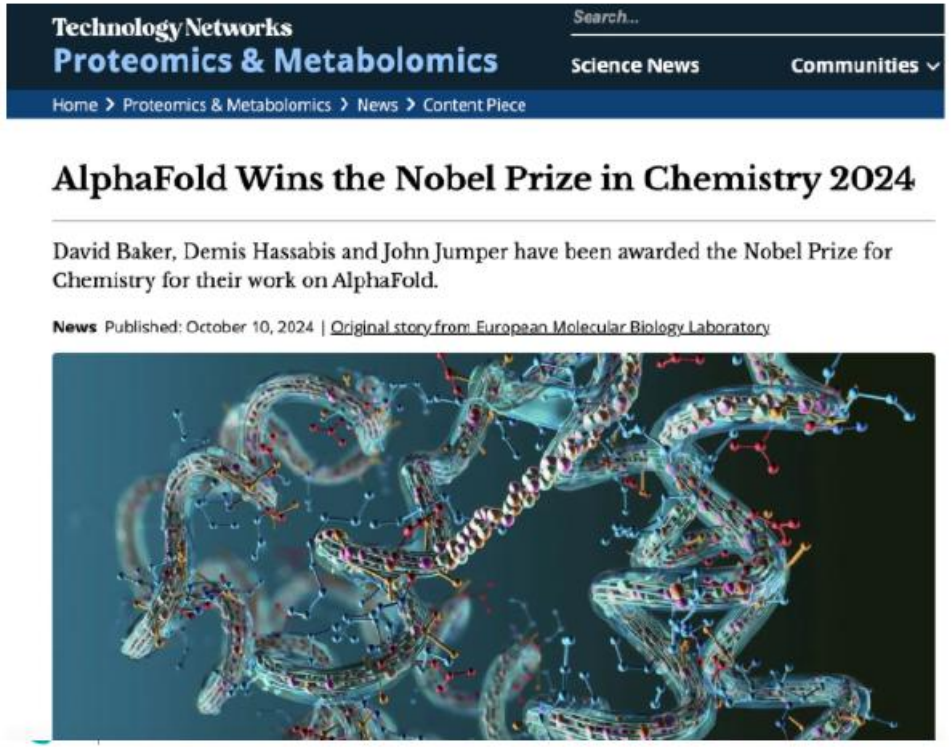
The Russian Academy of Sciences' Protein Institute has been tackling this challenge for over 50 years!



Can we discover the structure of all proteins experimentally?

- The required electron microscope costs over $SM and is expensive to operate.
- Humans alone have between 600,000 and 6 million isoforms!
- Small changes in the primary sequence can significantly alter the structure.
- Determining the structure of all proteins experimentally is impossible!

How do we determine protein structure computationally?



AlphaFold3 is an advanced AI model by DeepMind that predicts the 3D structures of proteins, nucleic acids, small molecules, and post-translational modifications with high accuracy.

It also supports multi-chain predictions, making it suitable for cofolding and docking applications.

Experimental vs Computational Methods

- Experimental methods are highly accurate but time-consuming
- Computational methods like AlphaFold offer faster, scalable predictions for unknown structures.

أكاديمية كاوست
KAUST ACADEMY

Hands on

Predict protein structure with AlphaFold2:

KAUST

أكاديمية كاوست
KAUST ACADEMY

Practice for Exam

KAUST

Question 1

1. What is the correct sequence of information flow in the Central Dogma?

A) RNA → DNA → Protein

B) DNA → RNA → Protein

C) Protein → RNA → DNA

D) RNA → Protein → DNA

Question 2

2. Which process synthesizes RNA from a DNA template?

A) Translation

B) Transcription

C) Replication

D) Reverse Transcription

Question 3

3. What is a codon?

A) A sequence of three nucleotides in mRNA

B) An enzyme that catalyzes transcription

C) A site where tRNA binds to mRNA

D) A protein involved in DNA replication

Question 4

4. What process is described by the synthesis of proteins from an mRNA template?

A) Replication

B) Transcription

C) Translation

D) Splicing

Question 5

5. What is the purpose of a FASTA file in bioinformatics?

A) Storing protein 3D structures

B) Representing DNA or RNA sequences

C) Visualizing phylogenetic trees

D) Performing sequence alignment

Question 6

6. Which bioinformatics tool is often used to compare DNA or protein sequences?

A) BLAST

B) Ribosome

C) PCR

D) Reverse Transcription

Question 7

7. What does the .transcribe() method in Biopython do?

A) Converts RNA back into DNA

B) Converts DNA into RNA

C) Aligns sequences

D) Finds codon usage

Question 8

8. In Biopython, what is the purpose of the .translate() method?

A) Converts DNA to RNA

B) Translates a nucleotide sequence to a protein sequence

C) Returns the complementary DNA strand

D) Reads a FASTA file

Question 9

9. What is the role of the Bio.Entrez module in Biopython?

A) Performing sequence alignments

B) Accessing NCBI's online databases

C) Visualizing protein structures

D) Calculating sequence statistics

Question 10

10. What would the RNA sequence be after transcribing the DNA sequence ATGCA?

A) ATGCA

B) UACGU

C) TACGT

D) UGCAU

Question 11

11. What is the complementary DNA strand for the sequence ATGCA?

A) TACGT

B) TGACG

C) TGCAT

D) AGTCA

Question 12

12. What is the reverse complement of the DNA sequence ATGCA?

A) TGCAT

B) ATGCA

C) TACGT

D) ACGTA

Question 13

13. What is the primary purpose of the Seq object in Biopython?

A) To align sequences

B) To represent biological sequences (DNA, RNA, or proteins)

C) To visualize phylogenetic trees

D) To access online biological databases

Question 14

14. What happens if you try to modify a Seq object directly (e.g., change a base in a DNA sequence)?

A) The sequence will update without error

B) An error will occur because Seq objects are immutable

C) The sequence will reverse itself

D) The sequence will automatically convert to lowercase

Question 15

15. Which database specializes in visualizing 3D protein structures?

A) UniProt

B) GenBank

C) PDB

D) KEGG

Question 16

16. What is the pp submodule in Scanpy responsible for?

A) preprocessing

B) plotting

C) utilities

D) statistical testing

Question 17

17. As covered in the course material ___ is used for pairwise alignment, while ___ is used for multiple alignment

A) BLAST, ClustalOmega

B) ClustalOmega, BLAST

C) Uniprot, NCBI

D) PDB, EMBL

Question 18

18. Each read in a fastq file consists of how many lines?

A) 1

B) 2

C) 3

D) 4

Question 19

19. What does using ! at the start of a line do in google colab?

A) Executes the line as a shell command

B) Comments out the line

C) Negates the output of the line

D) Marks the line as a header

Question 20

20. in which diagram of the ones listed below does branch length matter?

A) phylograms

B) cladograms

C) Topograms

D) Taxograms

Question 21

21. What is the primary difference between genomics and genetics?

A) Genomics studies specific genes, while genetics studies the entire genome.

B) Genomics studies the entire genome, while genetics focuses on individual genes.

C) Genomics deals with the study of RNA molecules, while genetics deals with DNA only.

D) Genomics is exclusively about studying genetic diseases, while genetics is about all traits.

Question 22

22. Proteomics aims to?

A) Identify and quantify all proteins in a system

B) Characterize the complete genome

C) Identify and annotate cells

D) All of the above

Question 23

23. What is the key difference between FASTA and FASTQ file formats?

A) FASTA contains sequence data with quality scores, while FASTQ contains sequence data only.

B) FASTA contains sequence data without quality scores, while FASTQ includes both sequence data and quality scores.

C) FASTA is used exclusively for DNA sequences, while FASTQ is used for RNA sequences.

D) FASTA stores metadata, while FASTQ focuses on sequence alignment.