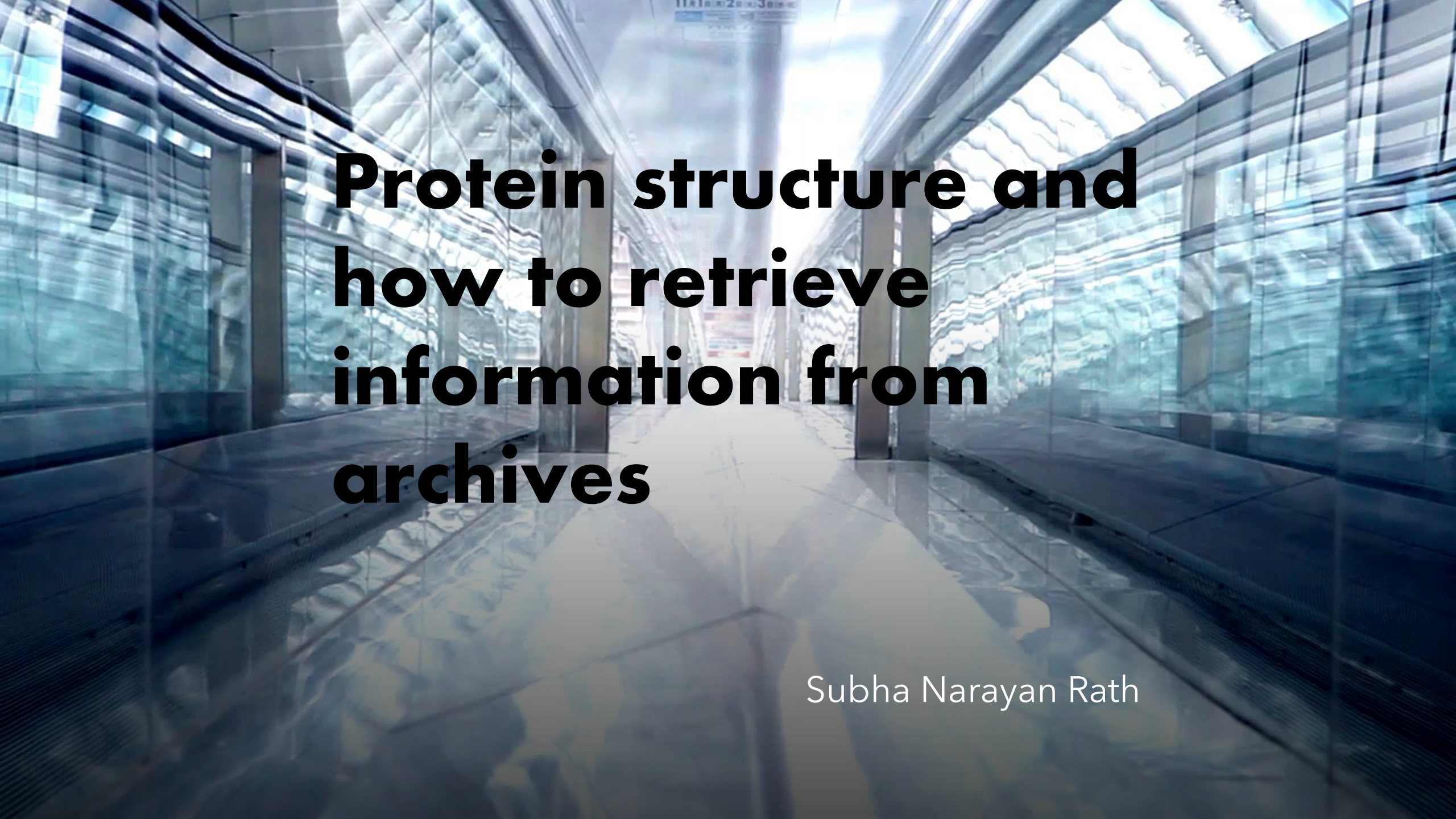


Protein structure and how to retrieve information from archives

Subha Narayan Rath

BLASTP is for:

- Exploring protein function
- Initial discovery for conserved domains
- However, can we predict protein structure and function based on amino acid sequence??
- Each protein folds to a unique 3D structure based on its amino acid sequence
- Protein structure is closely related to its function
- Protein structure prediction is a grand challenge of computational biology

Protein sequence database

- 1. Protein information resource (PIR) and associated database
  - A. PIRSF: Sequence family classification
  - B. iProClass: integrated protein knowledgebase
  - C. iProLINK: integrated protein literature, information and knowledge
- 2. SWISS-PROT (from Swiss institute of bioinformatics, Geneva, Switzerland)
  - Also contains bioinformatics tools and links called Expert Protein Analysis System (ExPASy….www.expasy.org)
  - PROSITE is a set of signature patterns characteristic of protein familiies
- 3. Trembl

In 2002, all three of them coordinated to form: UniProtKB consortium

- Amino acid sequence is not inferable from the gene sequence, because
- A. ambiguity in splicing
- B. information about ligands, disulphide bridges, subunit associations, post-translational modifications, effects of mRNA editing etc. can't be known from gene sequence
- E.g. https://www.uniprot.org/

Database of protein families

- Patterns of conservation identity features that nature has found to retain (PROSITE signatures are examples)
- R. Doolittle suggested: two full length protein sequences (>100 residues) that have 25% or more identical residues in an optical alignment are likely to be related
- < 15%; doubtful similarity
- 18-25%: twilight zone where there might be similarity like appearance of PROSITE consensus patterns
- Sequence oriented databases are: interPro, Pfam, COG
- Structure-oriented databases are: SCOP, CATH

DATABASE OF PROTEIN STRUCTURE

- They archive, annotate, distribute a set of atomic coordinates
- Worldwide Protein DataBank (wwPDB)
- Others: protein data bank Europe (EBI at UK) and protein data bank Japan (based at Osaka university)
- www.wwpdb.org
- It contains also structures of nucleic acids, carbohydrates in addition to proteins
- CCDC: Cambridge crystallographic Data Centre archives the structure of small molecules
- BioMagResBank at University of Wisconsin: archives protein structures determined by NMR
- wwPDB keeps the data from X-ray structure determinations

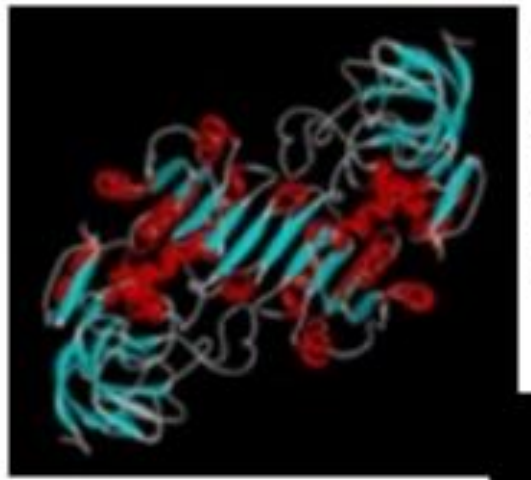
A protein databank structure contains:

- What protein and from which species
- Who solved the structure with reference
- Experimental details to solve the structure such as resolution of xray structure detemination
- Amino acid sequence
- Atomic coordinates (starting with ATOM)
- Additional molecules including cofactors, inhibitors, water molecules (keyword HETATM identifies coordinates of these)
- Assignment of secondary structure: helices and sheets
- Disulphide bridges

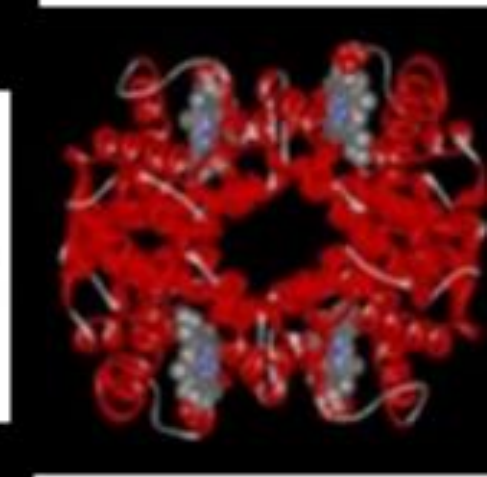
Problem 4.1

- Protein details can be found in RCSB homepage, a part of PDB
- https://www.rcsb.org/
- 1TRZ and different parts of that

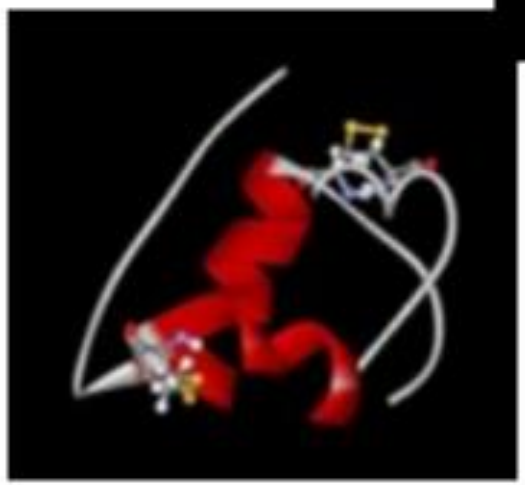
- Three examples of protein functions
  - Catalysis:
    Almost all chemical reactions in a living cell are catalyzed by protein enzymes.
  - Transport:
    Some proteins transports various substances, such as oxygen, ions, and so on.
  - Information transfer:
    For example, hormones.



Alcohol dehydrogenase oxidizes alcohol to aldehydes or ketones



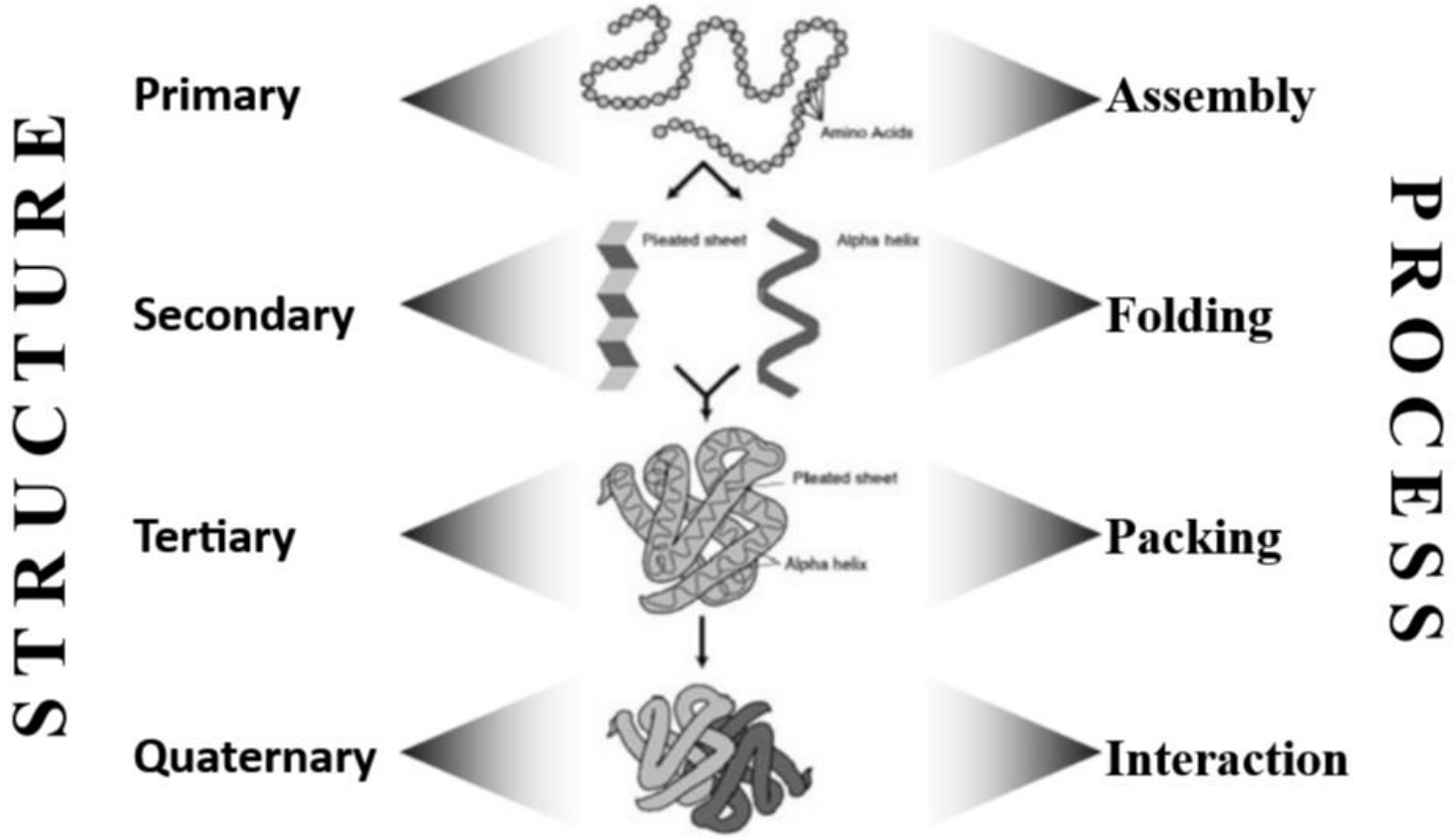
Haemoglobin carries oxygen



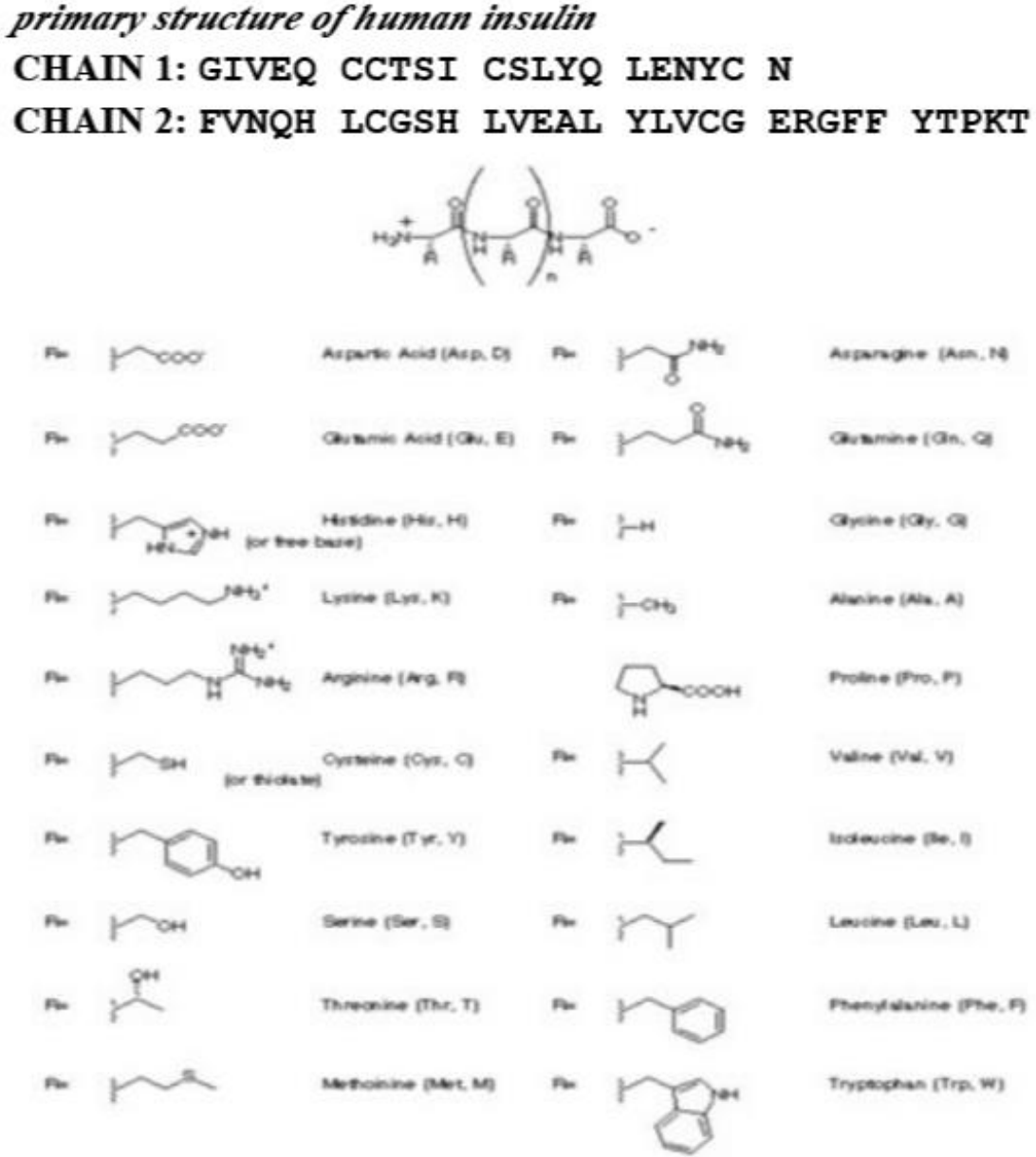
Insulin controls the amount of sugar in the blood

Basic unit: amino acids

- Depending on R sub-unit of amino acids the properties vary
- 20 amino acids and their properties
- Polymerization of amino acids while synthesis, causes the primary structure

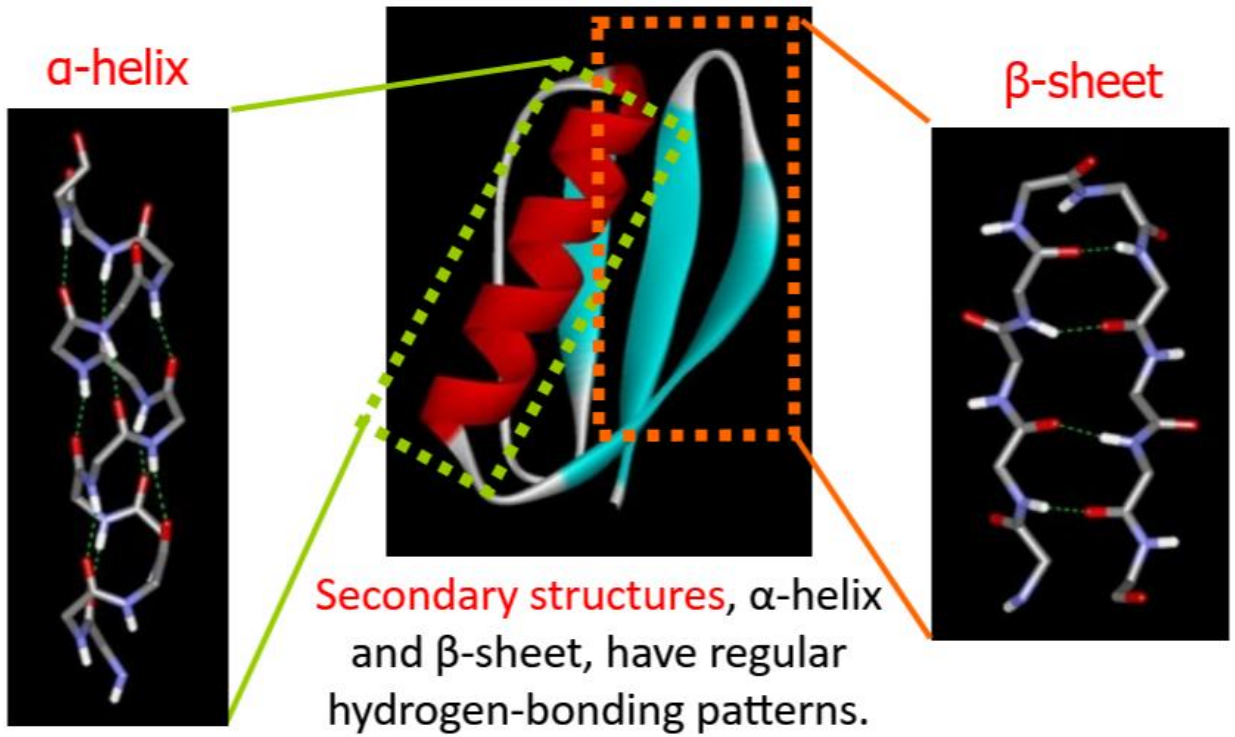


Primary structure



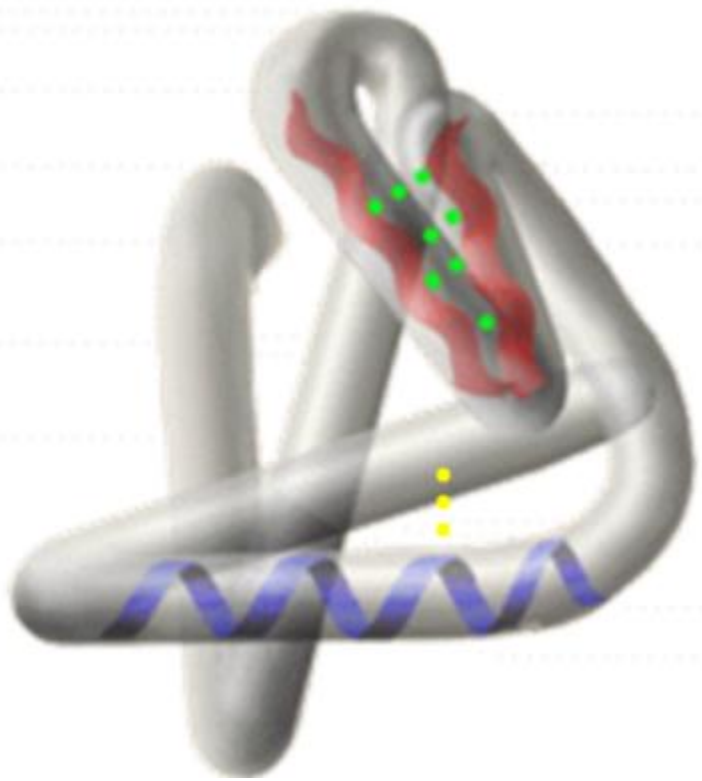
- Linear and ordered
- 1D
- Sequence of amino acids
- Written from amino end to carboxyl end by convention
- This linear structure is neither functional nor stable; so it has to do folding

Secondary structure due to protein folding



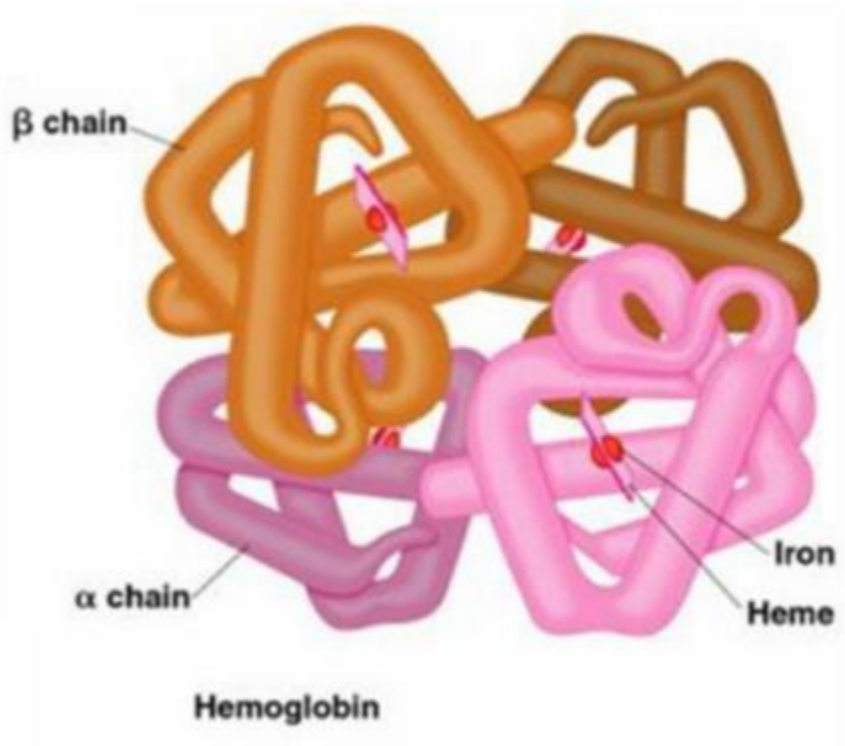
- Non-linear
- 3D
- Localized to regions of amino acid chain
- Formed and stabilized by
  - Hydrogen bonding
  - Van-der-walls interactions
  - Electrostatic forces

Tertiary structure due to protein packing



- Occurs in cytosol (upto 60% of bulk water or 40% of water of hydration)
- Due to interaction of 2 structure and solvents
- It’s non-linear and 3D

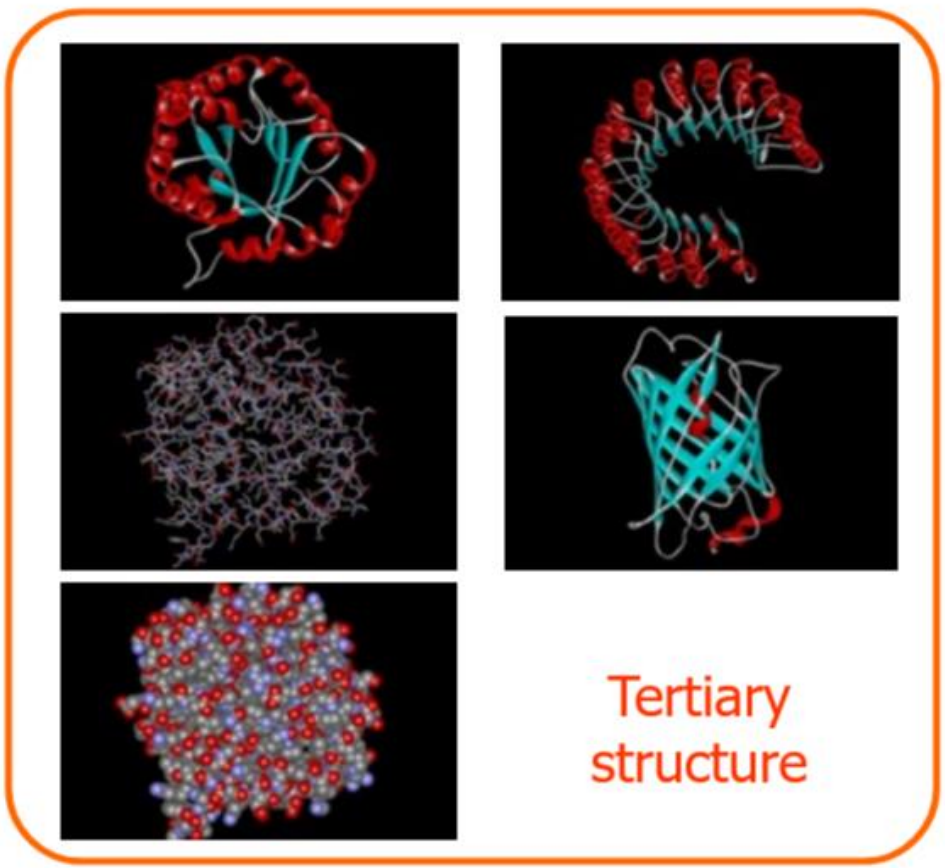
Quaternary structure due to protein interactions

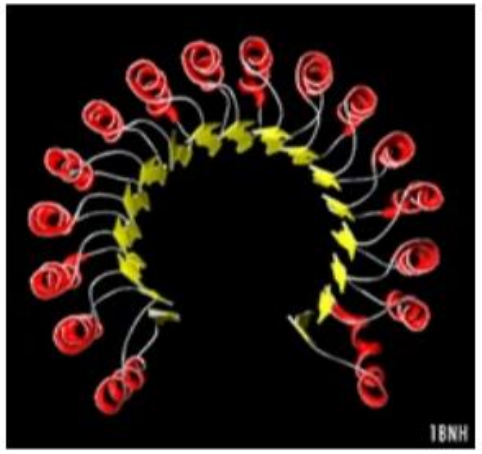


Hemoglobin

- In cytosol due to close proximity to other folded and packed proteins
- Involve interaction of tertiary structure elements of separate protein molecules
- Non-linear and 3D

3d structure of proteins





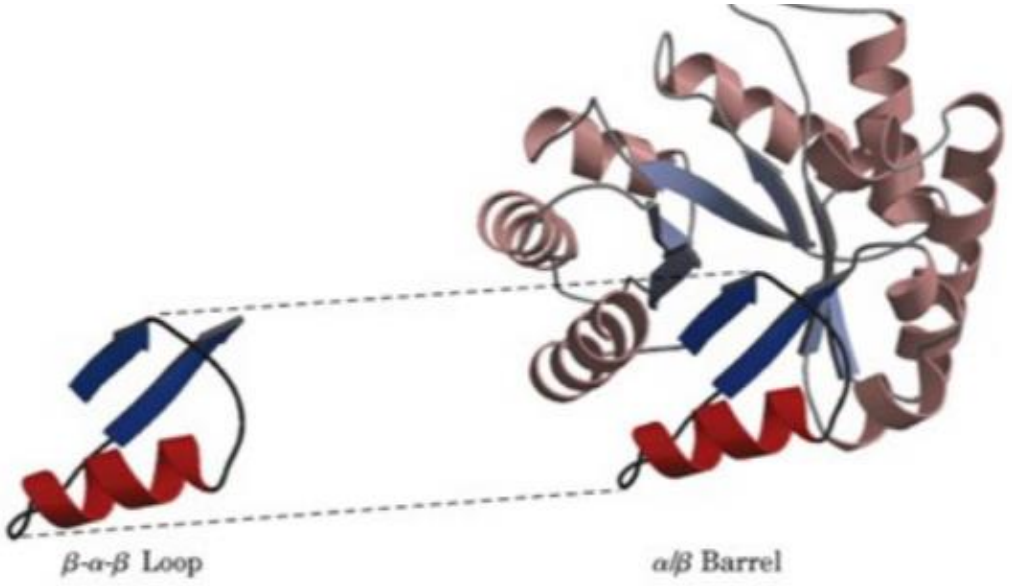
α/β

β-α-β Loop

Class/motif

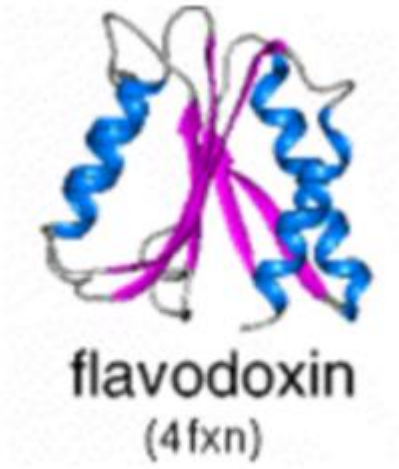
- Class = secondary structure composition
- Could be all α, all β, α/β, α+β
- Motif = small specific combination of secondary structure elements e.g. β –α – β loop

Fold



- Fold is architecture, the overall shape and orientation of secondary structures, ignoring connectivity between the structures
- E.g. alpha/beta barrel

Fold families or superfamilies



CLASS: α+β
FOLD: sandwich
FOLD FAMILY: flavodoxin

- Fold families: catergorizations that take into account topology and previous subsets as well as empirical or biological properties
- Superfamilies: above plus it includes evolutionary and ancestral properteis

Hierarchical protein structure

- Primary (amino acid sequence)
- Secondary (alpha helix and beta sheet)
- Tertiary (3D structure formed by assembly of secondary structures)
- Quaternary (more than one polypeptide chains)

Protein structure and function