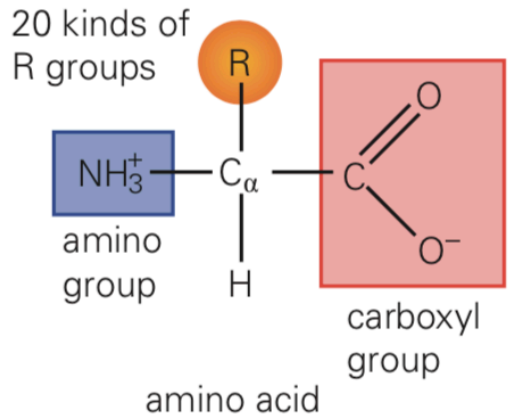
BIOINFORMATICS-II

Proteins are polymers of amino acids

- All proteins with few exceptions are polymers of same 20 different amino acids
- Amino acids contains
  - An amino group ($-NH_3^+$)
  - A carboxylate group ($-COO^-$)
  - A hydrogen atom (-H)
  - A side chain (variable, denoted by R)
  - All are attached to a CARBON atom

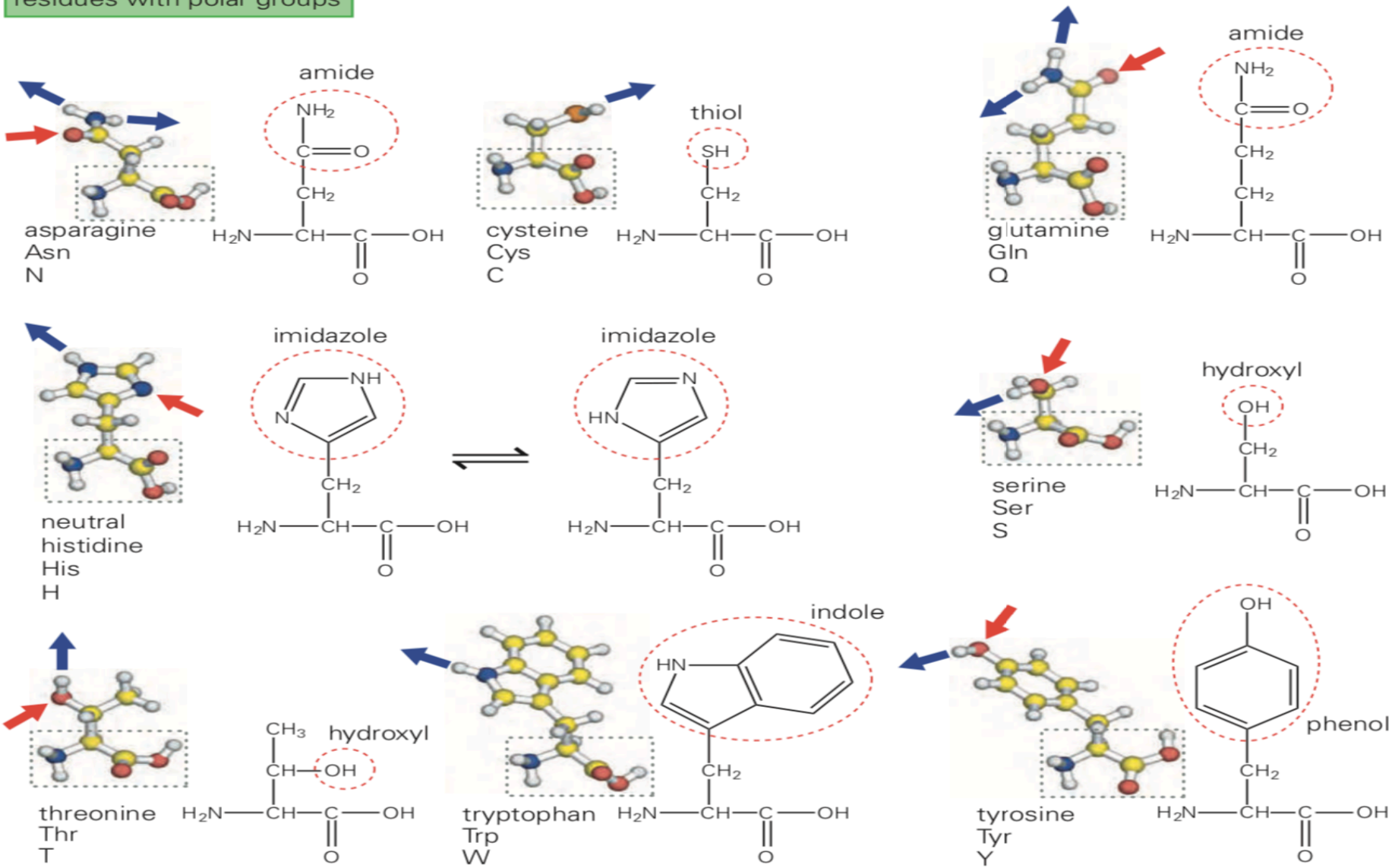


- At physiological pH (pH ~7) both carboxylate group and the amino group are almost completely ionized, thus forming a zwitterion - a dipolar molecule that contains charged groups but is electrically neutral overall

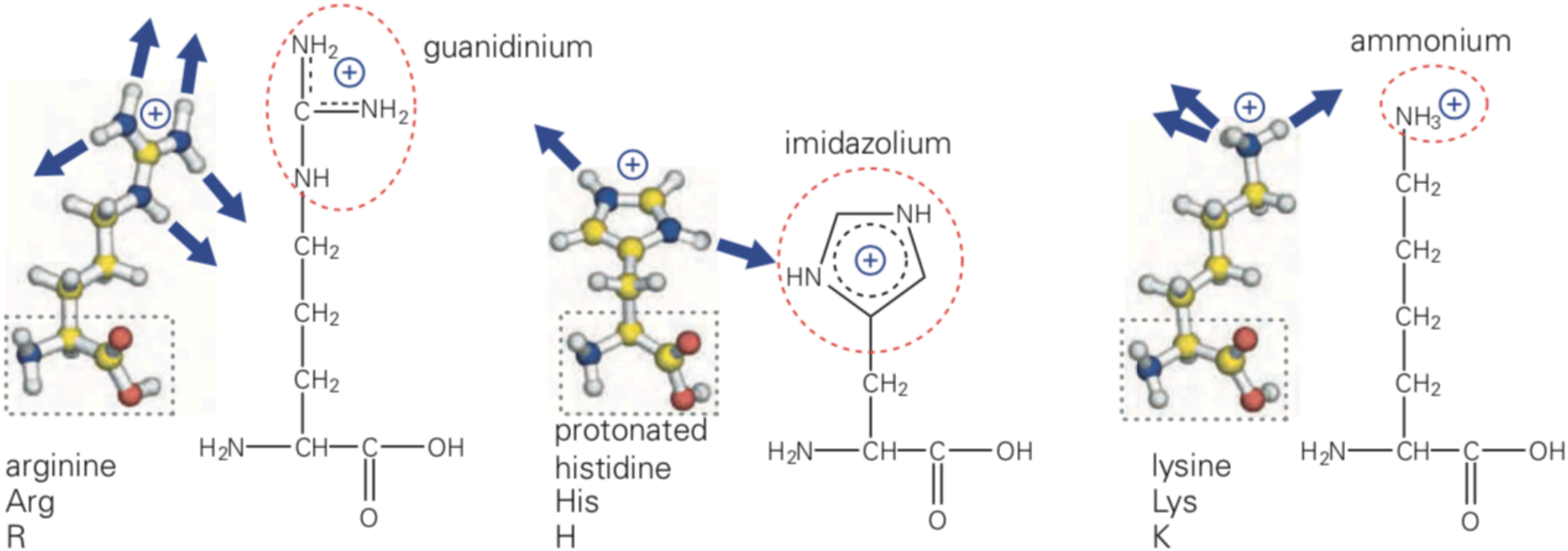
- Nineteen of the 20 amino acids are primary amines (that is, they contain a $–NH_3^+$group) and differ only in the nature of the side chain
  - The exception is proline, which is secondary amine ($-NH_2^+$) because its nitrogen and alpha-carbon atoms are part of a five-membered pyrrolidine ring

- **TWENTY AMINO ACIDS**

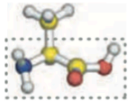
residues with polar groups



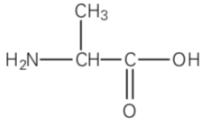
positively charged, hydrophilic residues

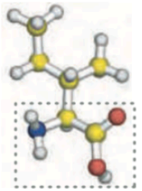


nonpolar, hydrophobic residues

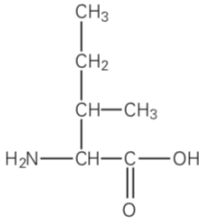


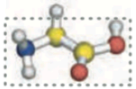
alanine
Ala
A



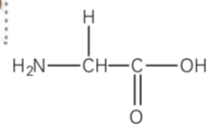


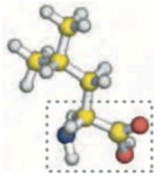
isoleucine
Ile
I



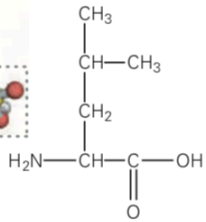


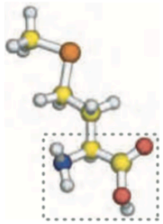
glycine
Gly
G



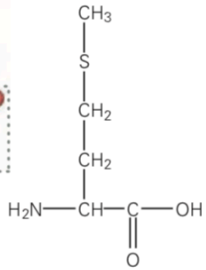


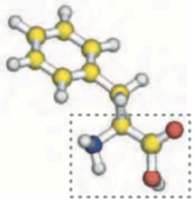
leucine
Leu
L



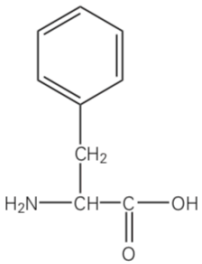


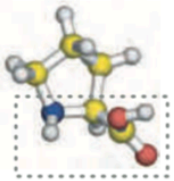
methionine
Met
M



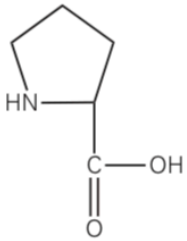


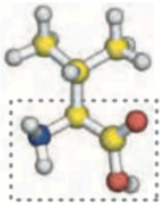
phenylalanine
Phe
F



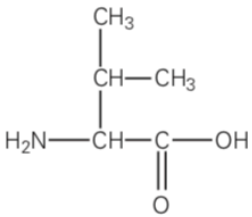


proline
Pro
P

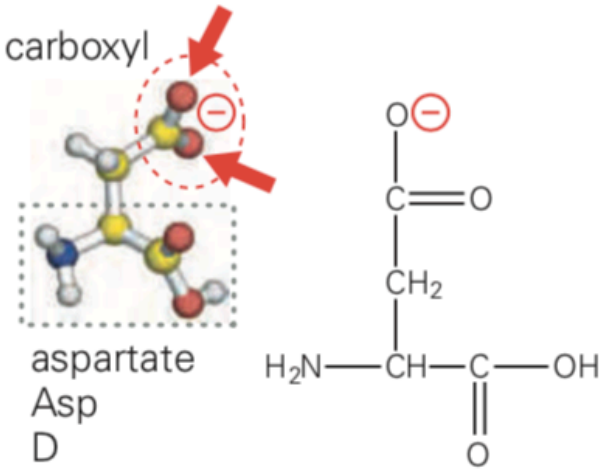


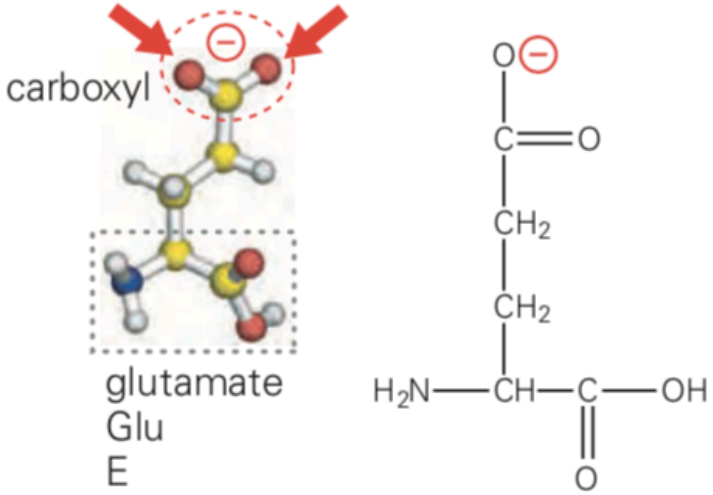


valine
Val
V



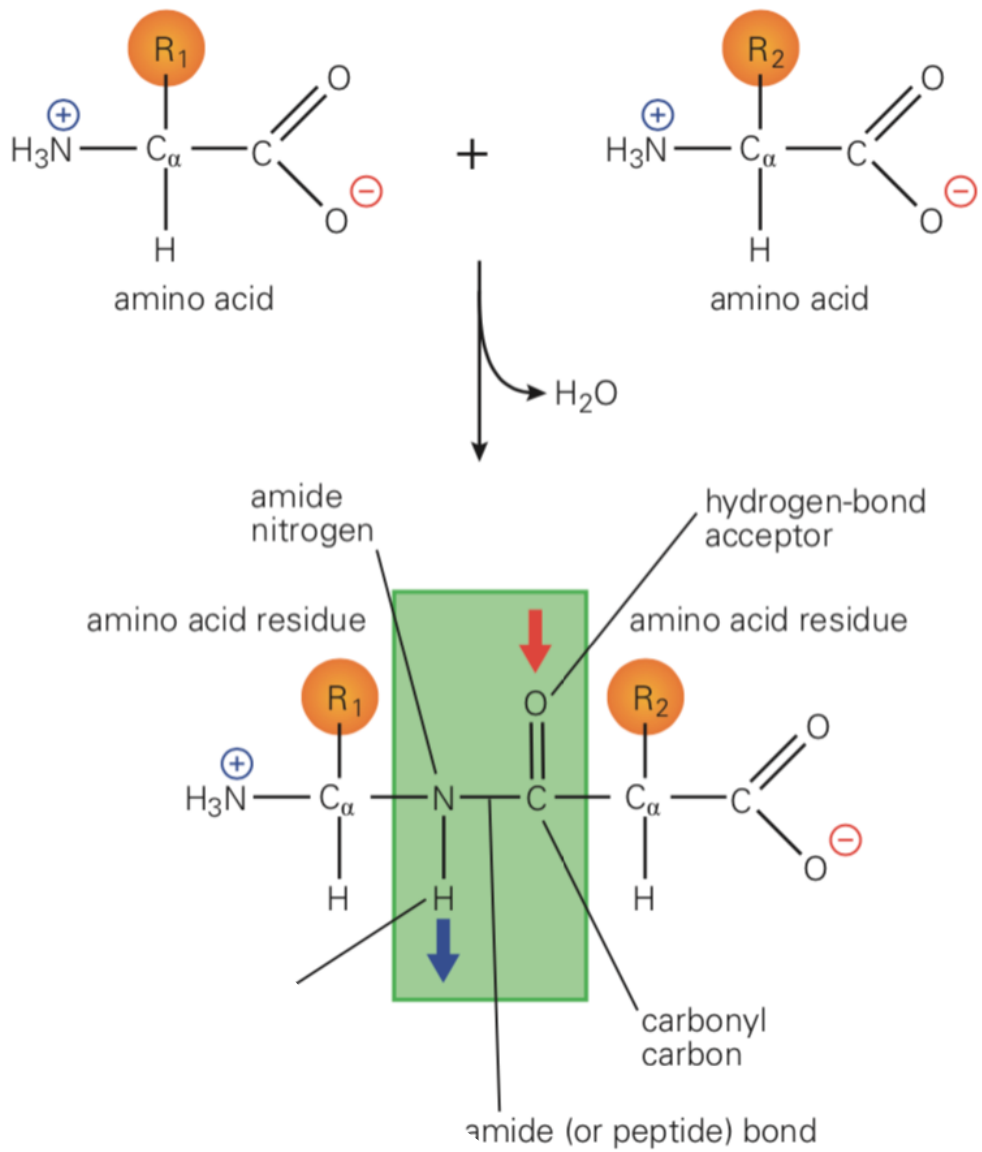
negatively charged, hydrophilic residues





Peptide Bond

- Condensation reaction with the elimination of water molecule
- Amino acids in a peptide bonds are known as amino acid residues as amino and carboxyl groups are not free then
- When the chain contains only a small number of residues (less than 50), it is referred to as a peptide or a polypeptide, however protein term is reserved usually for longer chains
- The repetitive sequence of –NH-CH-CO- atoms that runs the length of the polymer is called the peptide backbone
- Sequence of a protein is written from N-terminus to C-terminus



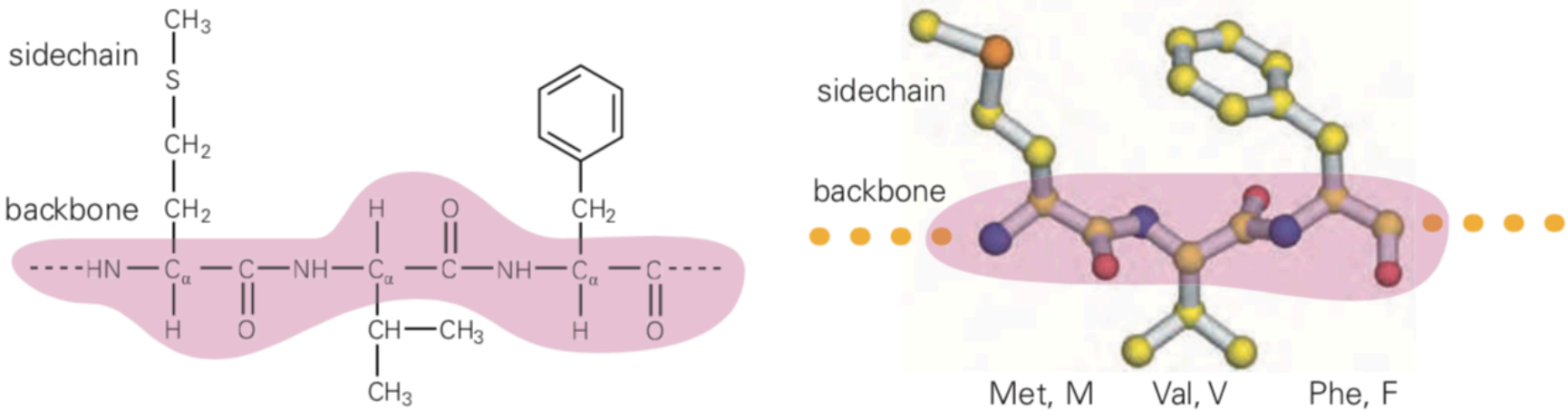
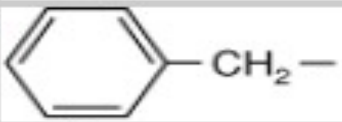
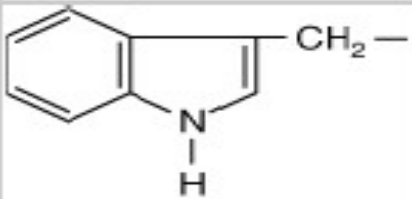
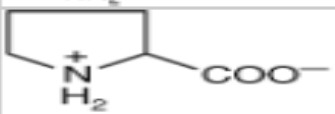
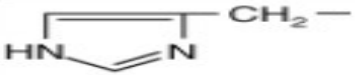


Figure 1.31 The structure of a peptide. The chemical and three-dimensional structure of a short peptide sequence (Met-Val-Phe) are shown here. Hydrogen atoms are not shown in the three-dimensional drawing on the right. The peptide backbone is highlighted in *purple*.

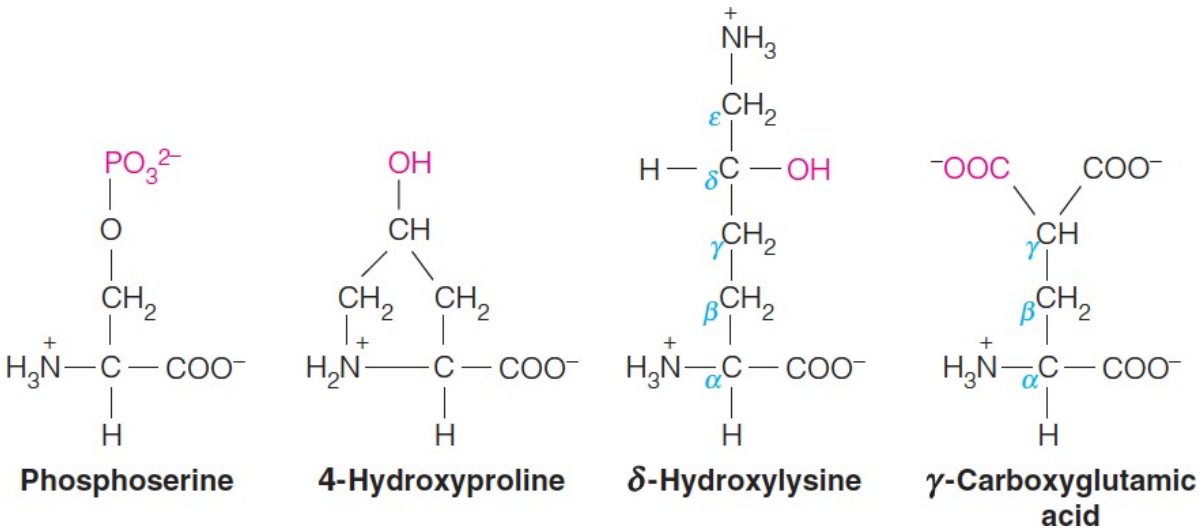
Table 1-1. Amino Acids—R-Group Classifications

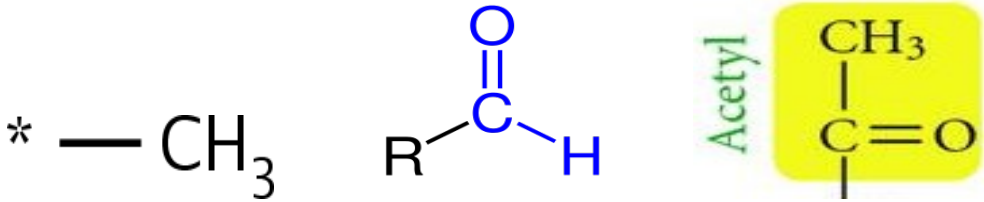
	Name	R-Group	Notes
Hydrophobic	Glycine	H—	Smallest R-group
	Alanine	CH_3—	Methyl R-group that normally folds easily within a protein
	Valine	H_3C \ CH— / H_3C	Bulky structure can impact folding of protein
	Leucine	H_3C \ CH — CH_2 — / H_3C	Bulky structure can impact folding of protein
	Isoleucine	CH_3 \ CH_2 \ CH— / CH_3	Bulky structure can impact folding of protein
	Phenylalanine		Aromatic ring
	Tryptophan		Indole ring

Hydrophilic	Serine	CH_2- | OH	Hydroxyl (OH) group with partial negative (–) charge (not shown); may be phosphorylated
	Threonine	CH_3-CH- | OH	Hydroxyl (OH) group with partial negative (–) charge; may be phosphorylated
	Asparagine	$H_2N-C-CH_2-$ ‖ O	Amino (NH_2) group with partial positive (+) charge (not shown)
	Glutamine	$H_2N-C-CH_2-CH-$ ‖ O	Amino (NH_2) group with partial positive (+) charge (not shown)
Charged	Tyrosine		Aromatic ring with hydroxyl group, giving partial or full negative (–) charge (not shown); may be phosphorylated
	Aspartic acid	$^-OOC-CH_2-$	Negative (–) charge from COO^-
	Glutamic acid	$^-OOC-CH_2-CH_2-$	Negative (–) charge from COO^-
	Lysine	$CH_2-CH_2-CH_2-CH_2-$ | NH_3^+	Positive (+) charge from NH_3^+
	Arginine	$H-N-CH_2-CH_2-CH_2-$ | $C=NH_2^+$ | NH_2	Positive (+) charge from NH_2^+
Special	Proline		β-turns
	Cysteine	CH_2- | SH	Disulfide bonds
	Methionine	CH_2-CH_2- | $S-CH_3$	Sulfur atom
	Histidine		Partial or full positive (+) charge from NH^+ (not shown)

Properties of the Amino Acid Side Chains

Amino acids can undergo *post-translational modification* resulting in modified amino acids with unique properties

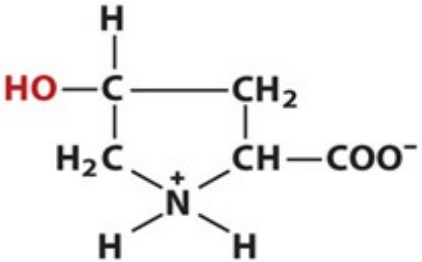




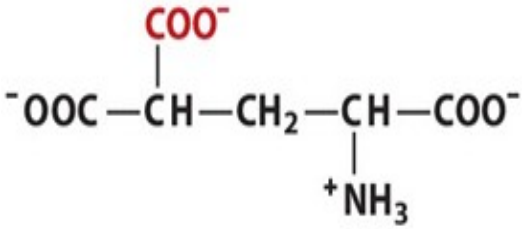
- **Example: Methylation, formylation, acetylation,** and **phosphorylation,....etc.** of certain residues.

- These modifications extend the biologic diversity of proteins by altering their solubility, stability, and interaction with other proteins

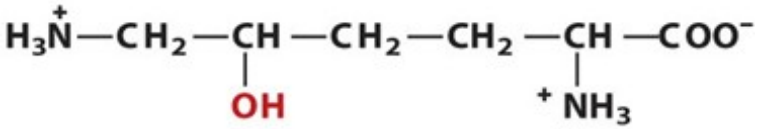
- Conversion of **proline to 4-hydroxyproline** *(major componenet of collagen)*
- Conversion of lysine to and **5-hydroxylysine** *(major componenet of collagen)*
- conversion of **glutamate** to **γ-carboxyglutamate**



4-Hydroxyproline

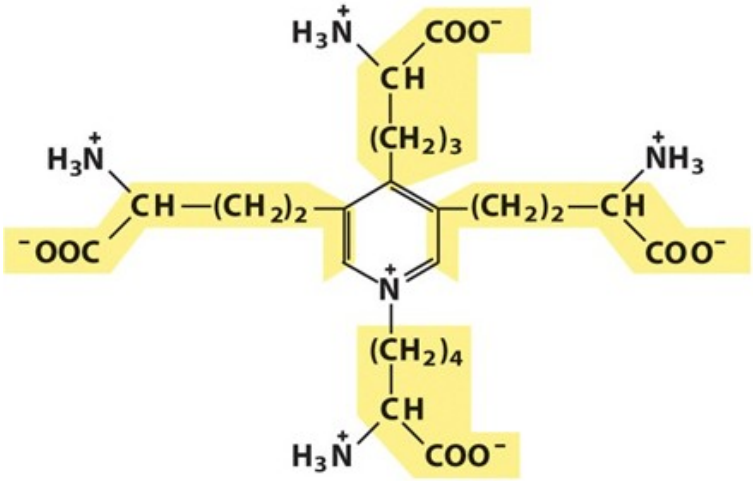


γ-Carboxyglutamate



5-Hydroxylysine

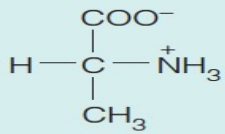
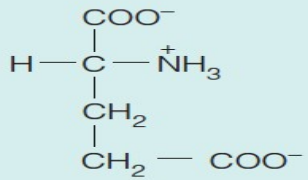
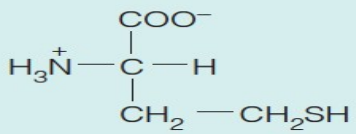
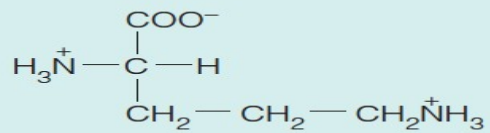
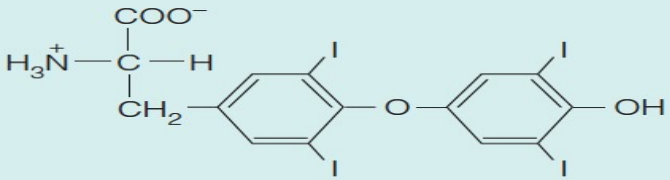
Is found in clotting factors and other proteins of the coagulation cascade. This modification introduces an affinity for calcium ions



Desmosine

Desmosine: Derived from four lysine residues, and found only in elastin. It allows elastin to stretch in all directions, a component of connective tissues.

TABLE 5.2 Some biologically important amino acids not typically found in proteins

Name	Formula	Biochemical Source, Function
β-Alanine	$H_3\overset{+}{N}-CH_2-CH_2-COO^-$	Found in the vitamin pantothenic acid and in some important natural peptides
D-Alanine		In polypeptides in some bacterial cell walls
γ-Aminobutyric acid	$H_3\overset{+}{N}-CH_2-CH_2-CH_2-COO^-$	Brain, other animal tissues; functions as neurotransmitter
D-Glutamic acid		In polypeptides in some bacterial cell walls
L-Homocysteine		Many tissues; precursor for methionine biosynthesis
L-Ornithine		Many tissues; an intermediate in arginine synthesis
Sarcosine	$CH_3-\underset{\underset{H}{\vert}}{N}-CH_2-COO^-$	Many tissues; intermediate in amino acid synthesis
L-Thyroxine		Thyroid gland; is thyroid hormone (I = iodine)

4 Levels of Protein Structure

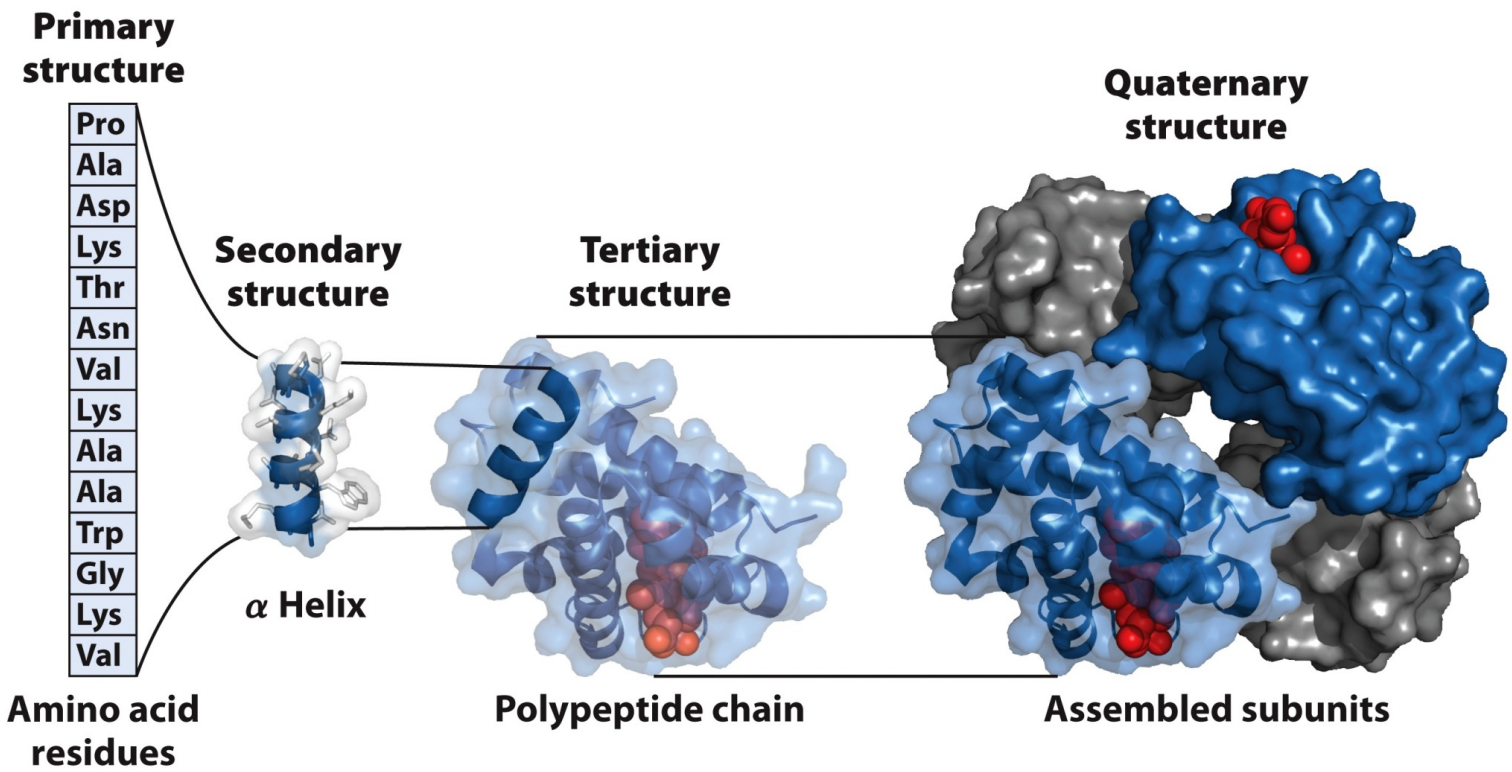


Figure 3-23
Lehninger Principles of Biochemistry, Sixth Edition

- Primary (Low complexity)
- Secondary
- Tertiary
- Quaternary (Highly complexity)

Primary Structure of proteins

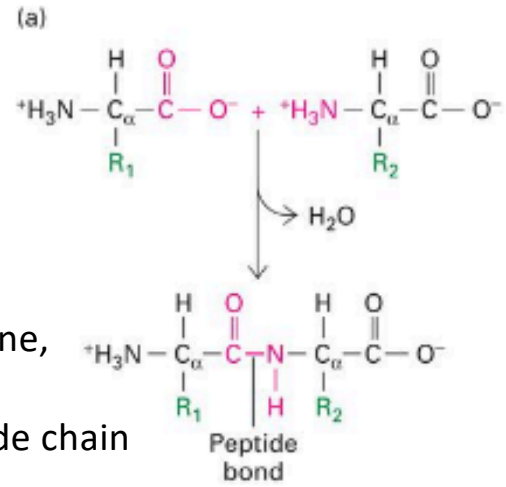
- Simple level of complexity
  - Specific sequence of the protein, it ultimately determine the final three dimensional structure of the proteins
- 5 amino acids means 4 peptide bonds
  - If we have n = amino acids, the peptide bonds will be = n-1
- Under physiological condition, the polypeptide will have polarity which means
  - On one hand , we will have a positive charge (NH^+) while on the other hand we will have a negative charge(COO^-)
-

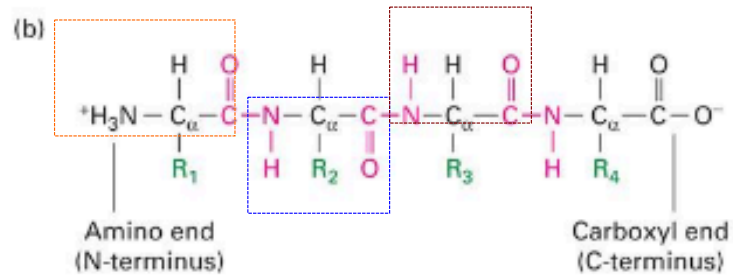
If we go along a polypeptide chain, we will see a Repeating Units of 3 atoms like

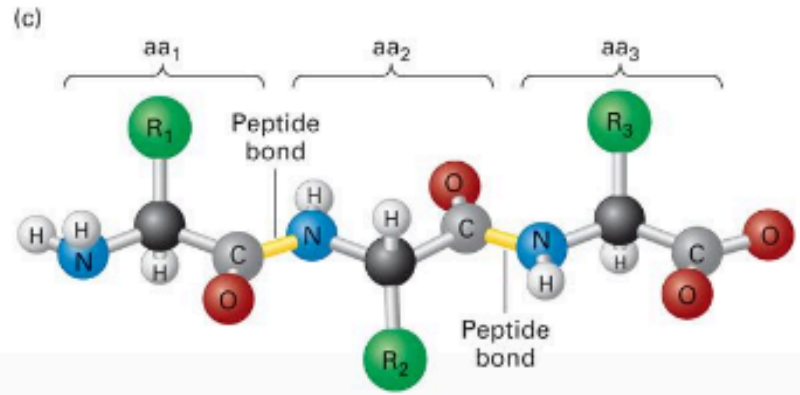
- Nitrogen-Carbon-Carbon
- This repeating section is known as a backbone, as it is repeated over and over

-However **the variable section** in our polypeptide chain Is the side chains like:

-R_1, R_2, R_3, R_4







Secondary Structure of Proteins

- The term refers to the local conformation of some part of a polypeptide.
- Four different types of regular patterns
  - α-helix
  - β-pleated sheet
  - B-turn
  - Ω-loop (Omega loops)
- The most common are the α-helix and β-pleated sheet

Tertiary Structure

- Refers to the spatial arrangements of amino acids in a specified/ compact structure found faraway from one another along the polypeptide chain.
- Number of interactions are involved in maintaining tertiary structure
  - Hydrophobic interactions
  - Van der Waal's interactions
  - Di-sulphide bridges
  - Hydrogen bonds
  - Ionic interactions

Quaternary Structure

- A protein is quaternary if it consists of two or more individual chains
- A simple two polypeptides chains structure is a dimer
  - Subunit may be 2 , 3 or many
- Two major categories of proteins (Discussed in detail in earlier slides)
  - Structural
    - Forms long fibers and play a structural roles. Keratin and collagen are the two examples
  - Globular
    - Relatively spherical shape
    - Hemoglobin (quaternary) vs. myoglobin (tertiary)

Forces that keep the different protein structures together

Level of protein structure	Interactions that stabilize the structure
Primary	Covalent bond (amide/peptide bond)
Secondary	Hydrogen bonds
Tertiary	Ionic bonds, disulfide bonds, hydrophobic interactions, hydrogen bonding
Quaternary	Ionic bonds, disulfide bonds, hydrophobic interactions, hydrogen bonding

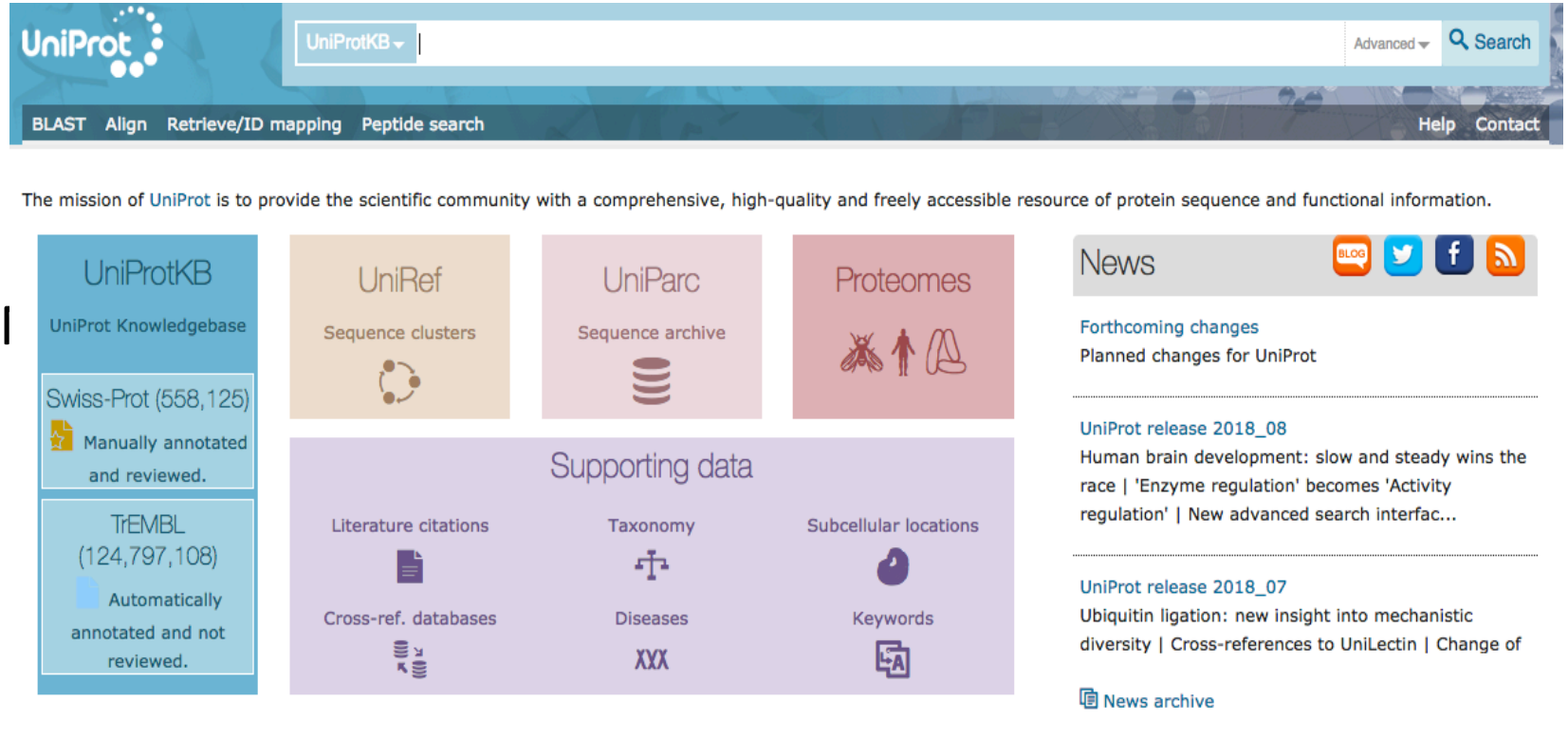
PROTEIN DATABASES

- **Protein Database**
  - Store protein sequences
  - Motifs : specific patterns of amino acids making up motifs
  - Structures
  - Structural Alignments
- First sequences to be collected were proteins using Sanger and Tupy's methods (1951) where common proteins families like cytochromes were sequenced

- Those protein sequences (mainly cytochromes) was assembled into atlas under leadership of Margret Dayhoff and her collaborators at National Biomedical Research Foundation (NBRF) in 1960s.
- The collection of Dayhoff and co became PIR (Protein Information Resource), which is now collaboration of NBRF, Munich Centre for Protein Sequences (MIPS) and Japan International Protein Information Database (JIPID)

- Swiss-Prot (Protein Sequences)
  - Is a collaboration between SIB (Swiss Institute of Bioinformatics) and EBI (European Bioinformatics Institute)
  - Mainly controlled by ExPASy (?) in Geneva
- International Partnership between PIR, EBI and SIB created:
  - UniProt, by unifying PIR-PSD, Swiss-Prot and TrEMBL databses

- UI

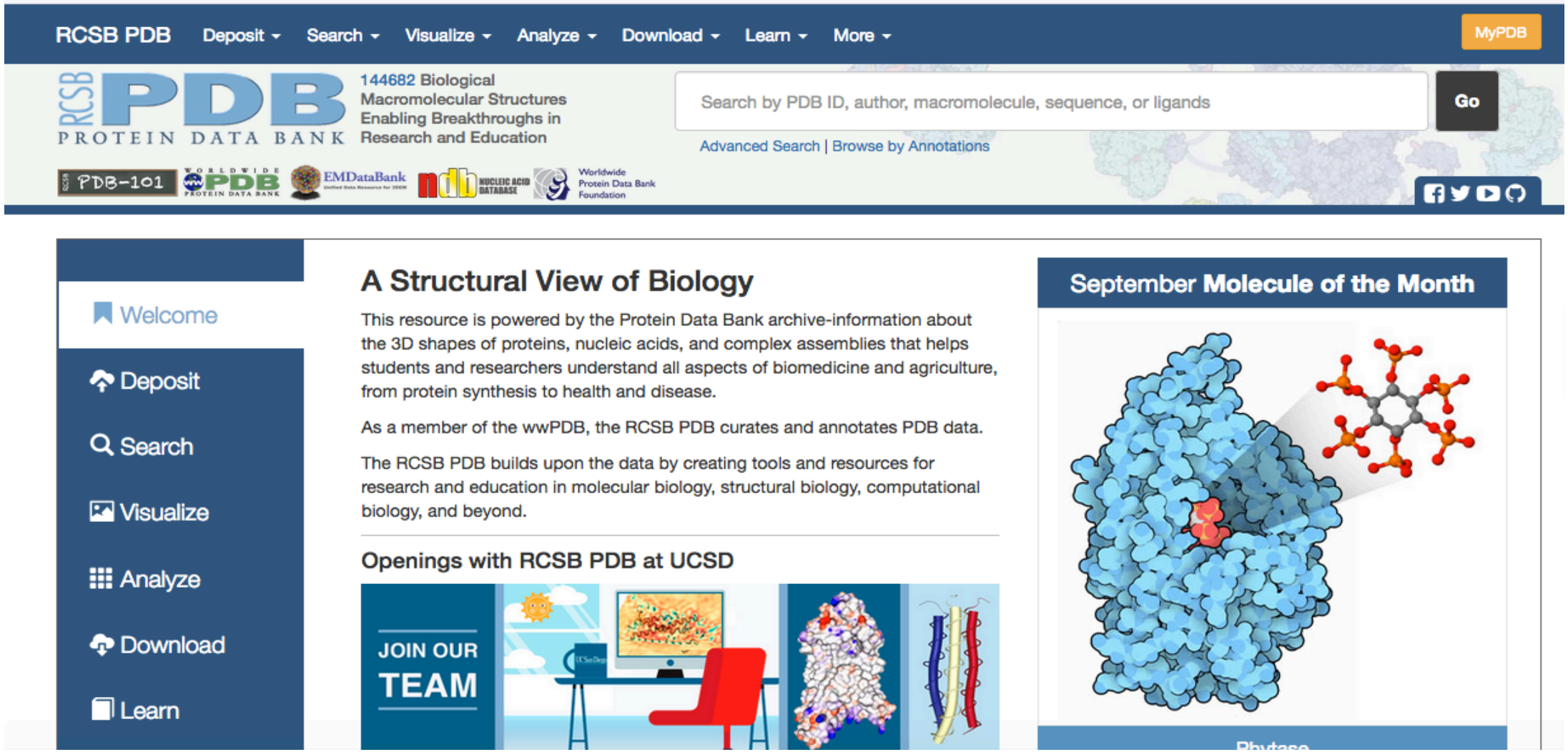


UniRef: The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records. This hides redundant sequences and obtains complete coverage of the sequence space at three resolutions:

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world. Proteins may exist in different source databases and in multiple copies in the same database. UniParc avoids such redundancy by storing each unique sequence only once and giving it a stable and unique identifier (UPI).

Proteome: provides total expressed proteins in a fully sequenced specie.

- **Details through practical work by taking an example of your assigned protein.**
  - **?**

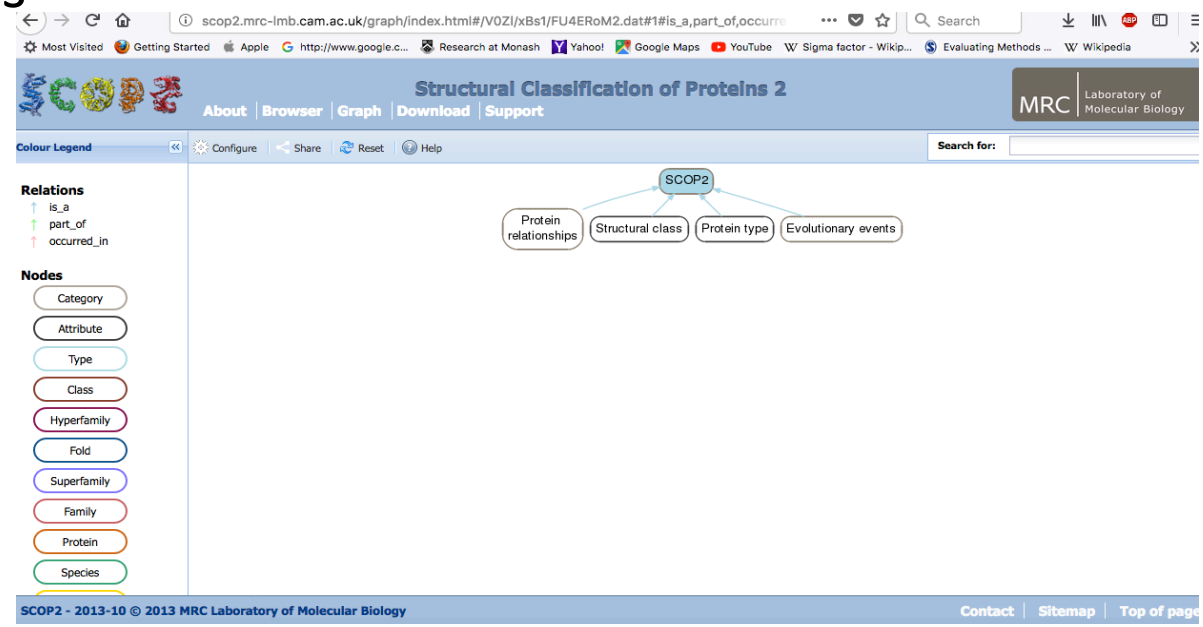


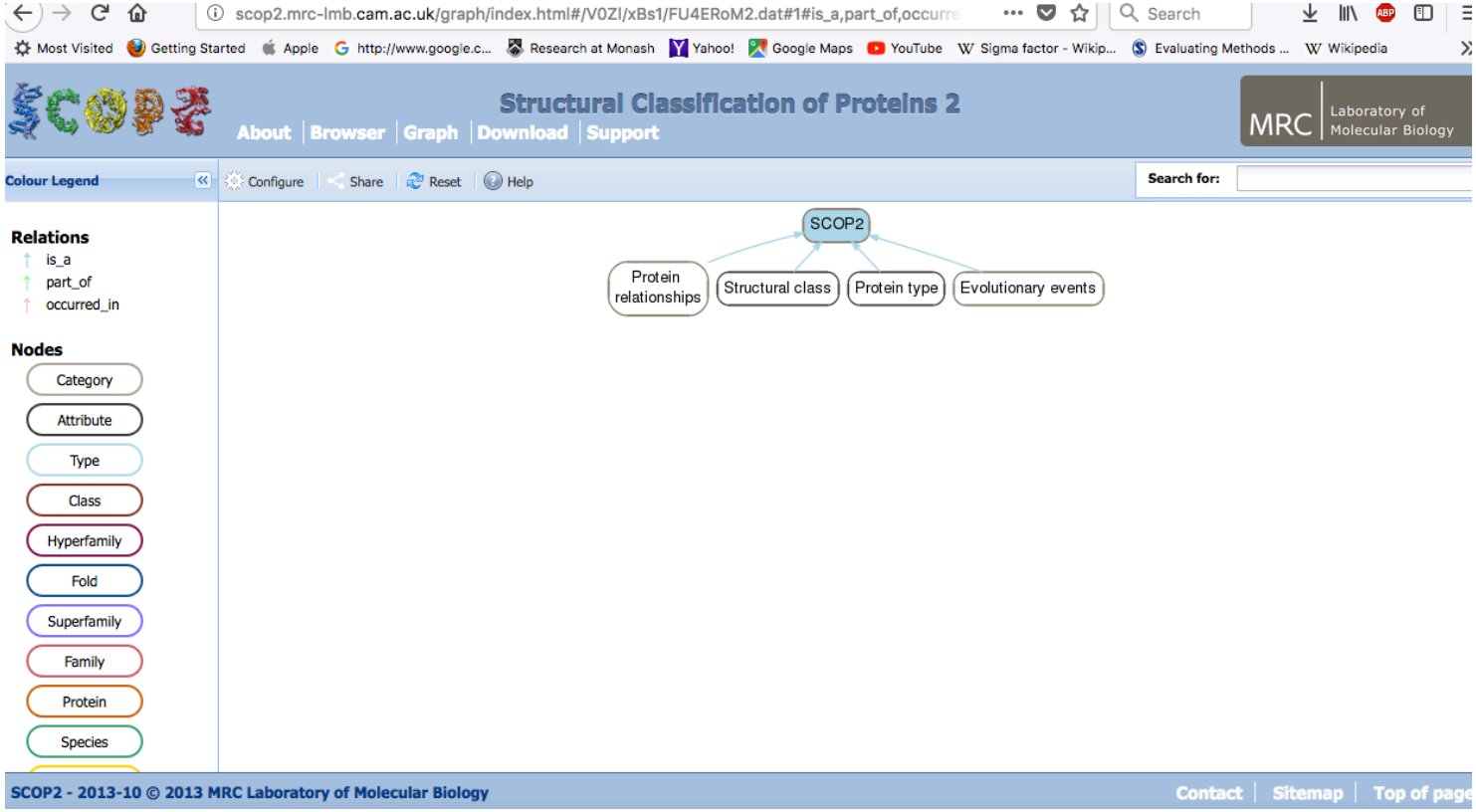
Mainly for 3D shapes of proteins..(structures generated through X-Ray and NMR)

Details through practical work by taking an example of your assigned protein.

SCOP2-Structural Analysis of Proteins

- SCOP2 is a successor of SCOP
- Proteins are organized according to their structural and evolutionary relationships




scop2.mrc-lmb.cam.ac.uk/graph/index.html#/V0ZI/xBs1/FU4ERoM2.dat#1#is_a,part_of,occurre
Search
Most Visited
Getting Started
Apple
http://www.google.c...
Research at Monash
Yahoo!
Google Maps
YouTube
Sigma factor - Wikip...
Evaluating Methods ...
Wikipedia
Structural Classification of Proteins 2
About
Browser
Graph
Download
Support
MRC
Laboratory of Molecular Biology
Colour Legend
Configure
Share
Reset
Help
Search for:
Relations
is_a
part_of
occurred_in
Nodes
Category
Attribute
Type
Class
Hyperfamily
Fold
Superfamily
Family
Protein
Species
SCOP2
Protein relationships
Structural class
Protein type
Evolutionary events
SCOP2 - 2013-10 © 2013 MRC Laboratory of Molecular Biology
Contact
Sitemap
Top of page

- **Protein relationships**
  - Further subdivided into three sub-categories
    - Structural
    - Evolutionary
    - Other relationships
  - Evolutionary Relationship
    - Evolutionary levels are retained like
      - Species
      - Protein
      - Family
      - Superfamily

- **Species:**
  - Corresponds to the individual gene product and is represented by its full length sequence
- **Protein**
  - Groups together orthologous proteins
- **Family**
  - Corresponds to the conserved sequence region shared by closely related proteins
- **Superfamily**
  - Is represented by the common structural region shared by different protein families

- **Structural relationships**
  - In SCOP2 the structural and evolutionary relationships are presented in separate branches to ensure more consistent classification of evolutionary related but structurally distinct proteins
- **Other relationships**
  - This aims to define and annotate relationships such as internal structural repeats, common motifs and sub-folds that have not been a subject of classification in the SCOP database.

- **Protein Types**
  - Soluble
  - Membrane
  - Fibrous
  - Intrinsically disordered
- Each type to a large extent correlates with characteristics sequence and structural features

- **Evolutionary events**
  - This section facilitates the annotation of various structural rearrangements and peculiarities that have been observed amongst related proteins and which have given rise to substantial structural differences
- **Structural classes**
  - the Structural classes, organizes protein folds according to their secondary structural

Homology Modeling

- It is a technique which is used to construct an unknown atomic-resolution model of a **“target protein”** from its primary structure and the of a related hom**experimental 3D structure** ologous protein which is known as **template**
- It is a predicted structure which is based on the similarity of the template protein and the template protein is obtained from experimental work like :
  - NMR
  - X-Ray crystallography

- NMR:
  - Low resolution
  - Multiple frames, which means that the PDB uploaded structure shows moment
  - Hydrogen atoms are present
- X-ray method
  - High resolution
  - Single frame
  - No hydrogen atoms present. Their position has to be guessed by using the topology information of the residues

- In homology modeling, more than ~40 % sequence identity will usually generate a useful model.
- In 1969, David Phillips,, Brown and co-workers published the first paper regarding homology modeling.
- They modeled alpha-lactalbumin based on the structure of hen egg white lysozyme. The sequence identity between these two proteins was 39 %.

Steps

- Identify related structures (Templates)
  - Sequence Identity
    - Shows % identity
  - Query coverage
    - How much our query sequence is covered in comparison with the template
      - If our sequence is 200 aa, and if some chunks (may be 50, 60 or 100 aa) are covered then we don't select that as a template, as otherwise the uncovered regions will make loops which will disrupt the overall structure
- Select particular template
  - Align target sequence with template structure
- Build a model for the target (using information from template structure)

Steps

- Evaluate the model
  - RMSD (Root Mean Square Deviation)
  - Ramachandran Plot
- Analysis
  - Yes
    - Or
  - No
    - If no, then selection of template is repeated

RMSD (Root Mean Square Deviation)

- Is the most commonly used quantitative measure of the similarity between two superimposed atomic coordinates.
- RMSD values are represented in Å (Angstrom, 1Å= 10^{-10}m)
  - Measuring very small distances
- RMSD is calculated by the squared difference between two sets of atomic coordinates after superposition.
- The RMSD values are also used in model quality evaluation where lower RMSD values indicates a lesser deviation between template and model structure, and eventually it then shows that the model has more nearer native-like fold and also helps in identifying the dissimilarity between them.
- RMSD is always non-negative, and a value of 0 (almost never achieved in practice) would indicate a perfect fit to the data
  - Values (0.1, 0.2, 0.5, 0.6, 1, 2, 4, 5, etc)

- **RMSD Values**
  - 0.0-0.5 Å Essentially Identical
  - <1.5 Å Very good fit
  - <5.0 Å Moderately
  - >7.0 Å Dubious
  - >12.0 Å Completely unrelated

Ramachandran Plot

- It is plot to visualize energetically allowed regions for a polypeptide backbone, torsion angles phi and psi of amino acids residues presnt in a proteins structure
- Used to analyse the structure of a protein, the conformation of amino acids present in the protein and the close contacts between the atoms

19.1.3 Ramachandran Plot

Since the peptide units are effectively rigid groups that are linked into a chain by covalent bonds at the C_α atoms, the only degrees of freedom they have are rotations around these bonds. Each unit can rotate around two such bonds: the C_α- C' and the N-C_α bonds. By convention, the angle of rotation around the N-C_α bond is called **phi** (ϕ) and the angle around the C_α-C' bond from the same C_α atoms is called **psi** (ψ). In this way, the conformation of the whole main chain is completely determined when the ϕ and ψ angles for each amino acids are defined.

Most combinations of ϕ and ψ angles of an amino acids are not allowed because of steric collisions between the side chains and main chain. The angle pairs ϕ and ψ are usually plotted against each other in a diagram called a **Ramachandran plot** after the indian biophysicist G.N.Ramachandran who first made calculations of sterically allowed regions. Figure 19.6 shows the results of such calculation for all amino acids except glycline from a number of accurately determined protein structures. The major allowed regions in Figure 19.6 are the right-handed α -helical cluster (Figure 19.7) in the lower left quadrant; the broad region of extended β strands (Figure 19.7) of both parallel and antiparallel β structures in the upper left quadrant; and the small, sparsely popluated left-handed α-helical region in the upper right quadrant.

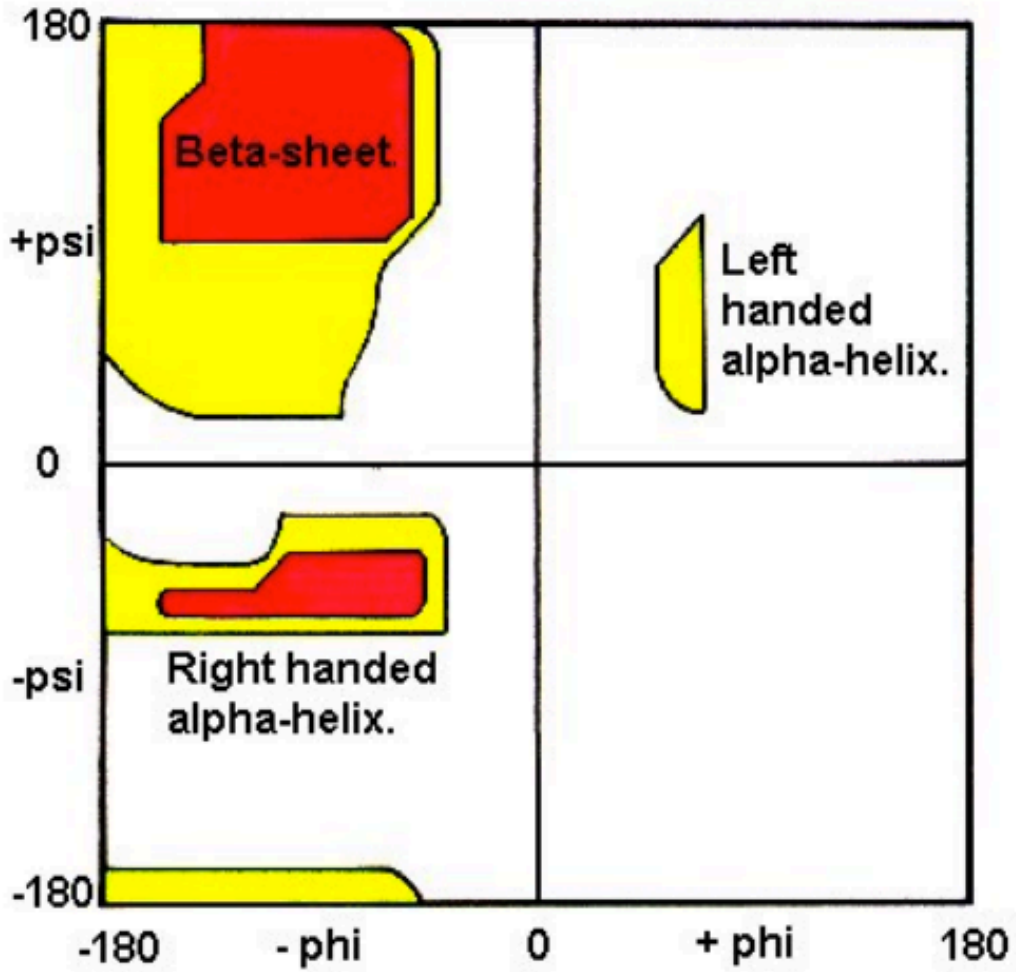


Figure 19.6: Ramachandran plot[4].

Visualization tools

- PyMOL
- VMD
- RasMOL
- Chimera
- ….