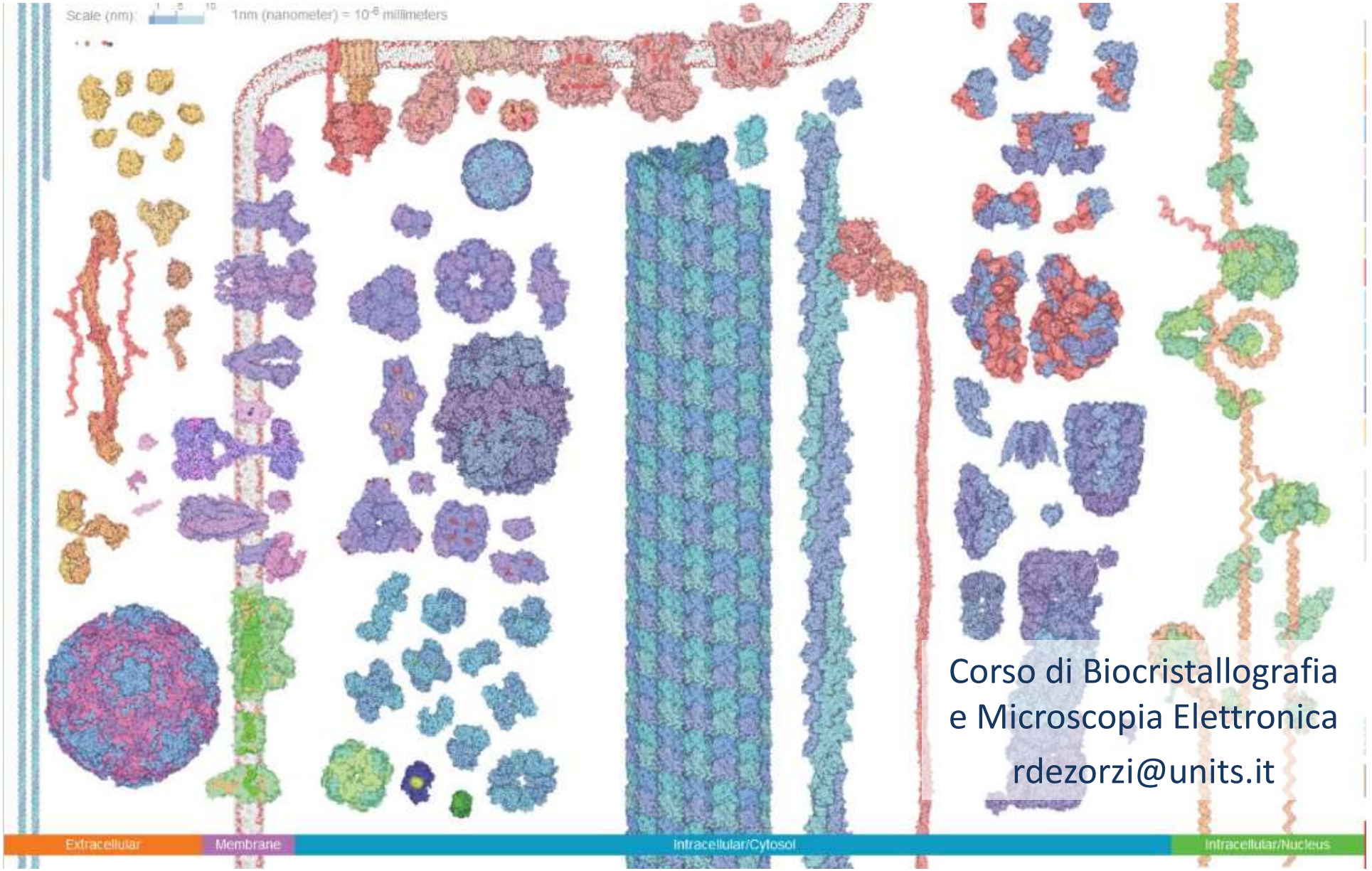
Basic elements of protein structure



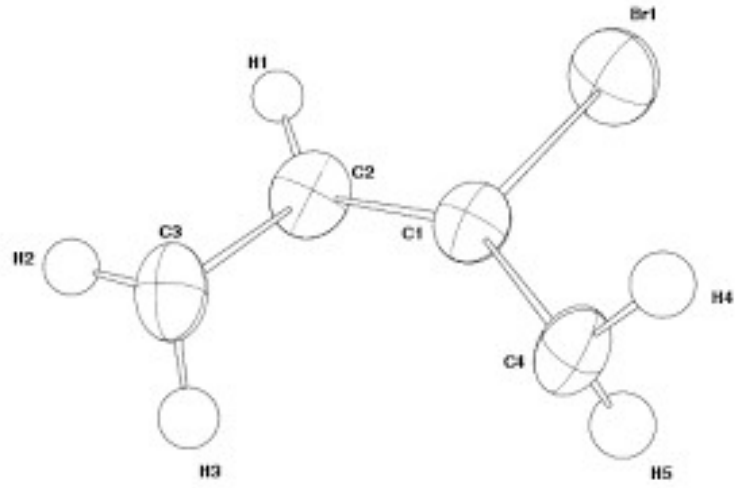
Corso di Biocristallografia
e Microscopia Elettronica
rdezorzi@units.it

Why?

Protein structures have a large number of variables, due to the number of atoms present in the protein.

For each atom:

- 3 variables for the position (x, y, z)
- 1 or 6 varibles for the thermal factor (isotropic or anisotropic)



For medium/low resolution structures, data could be insufficient to refine all parameters: a good refinement requires 8/10 data for each parameter.
Geometrical restraints are added to increase number of data.

In addition, geometrical considerations help in validation.

Proteins

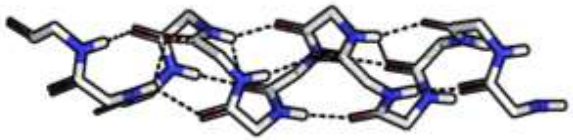
Large and diverse group of molecules, different for structure and function, divided in 60˙000 protein families (and growing!)

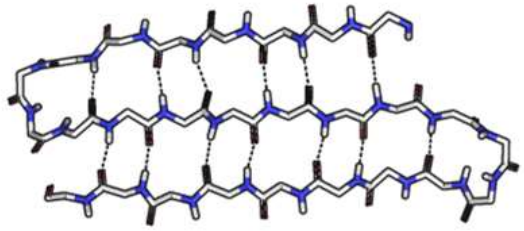
Primary structure: residue sequence

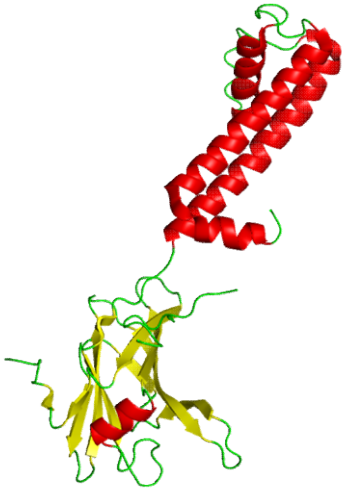
```
MET THR GLY GLY MET LYS PRO PRO
ALA ARG LYS PRO ARG ILE LEU ASN
SER ASP GLY SER SER ASN ILE THR
ARG LEU GLY LEU GLU LYS ARG GLY
TRP LEU ASP ASP HIS TYR HIS ASP
LEU LEU THR VAL SER TRP PRO VAL
PHE ILE THR LEU ILE THR GLY LEU
```

[The structure depends on the primary sequence, but not only! The folding is determined also by the action of chaperones, interactors, solution composition, protein partners...]

Secondary structure: local folding of the polypeptide (H-bonds!)

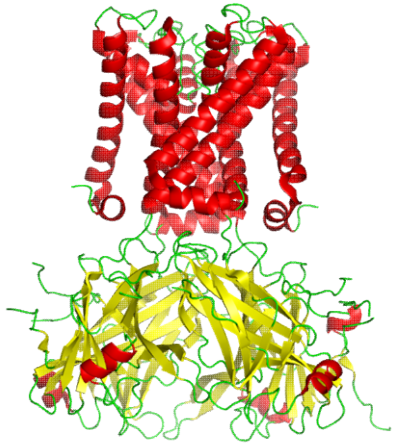






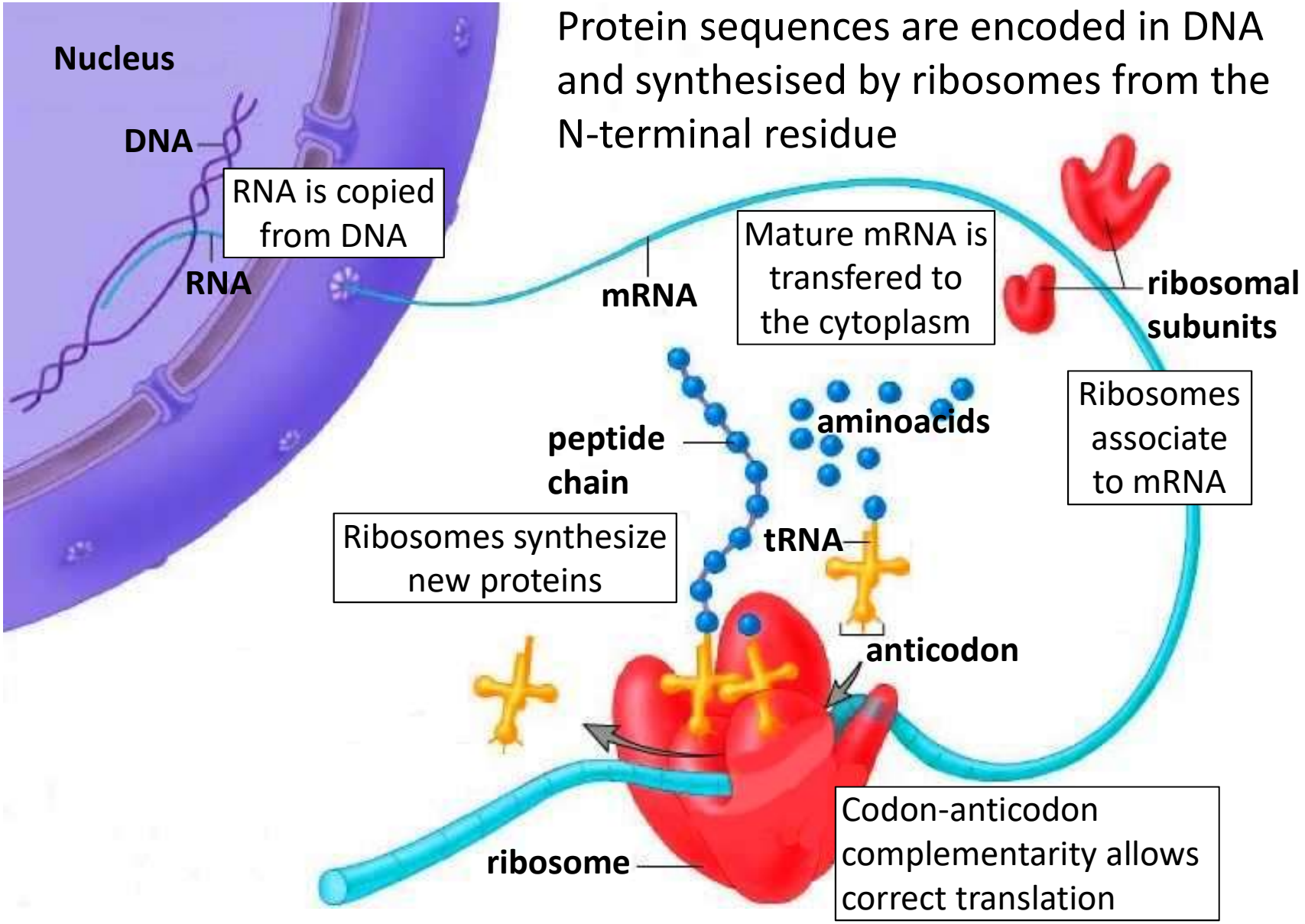
Tertiary structure: overall folding of the protein

Quaternary structure: biologically functional complex formed by more than one polypeptide chain

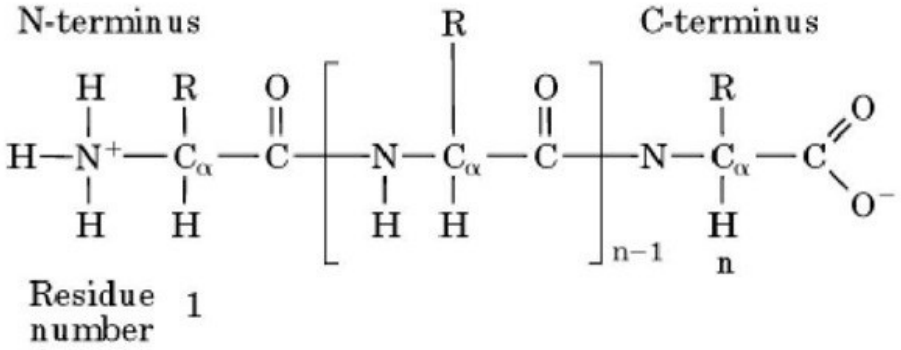


Biosynthesis of proteins

Protein sequences are encoded in DNA and synthesised by ribosomes from the N-terminal residue



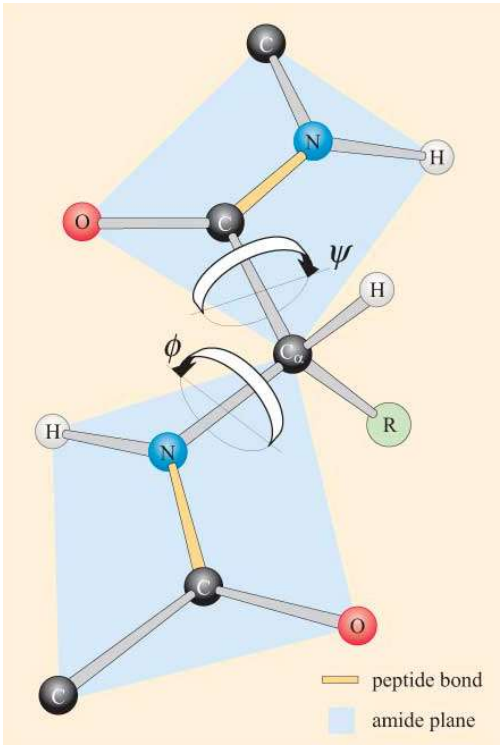
Geometry of the polypeptide chain

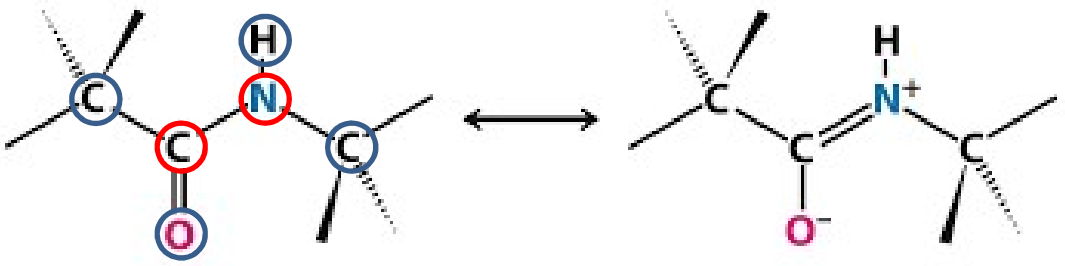


20 amino acids, linked by peptide bonds

Geometry of the polypeptide chain:

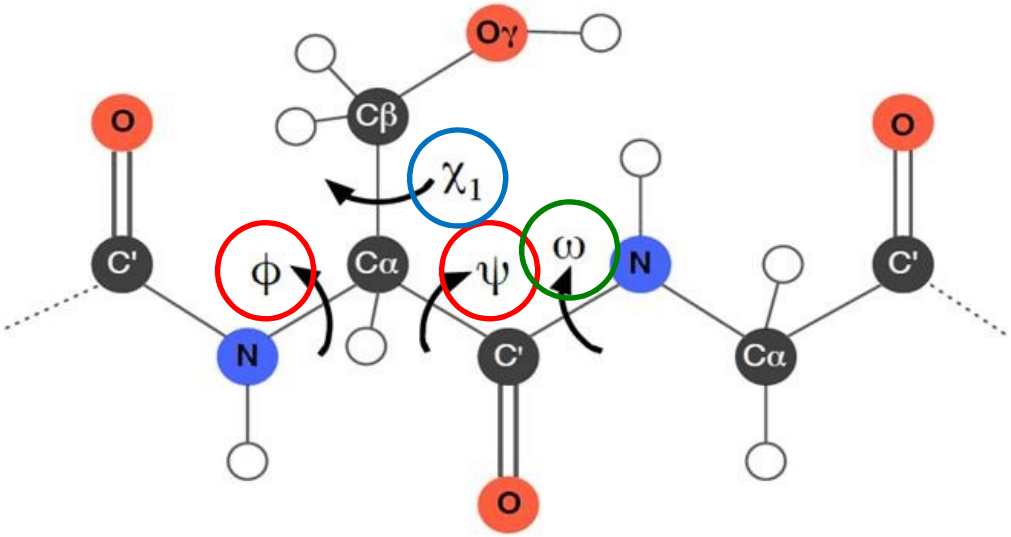
- Bond distances (defined by 2 atoms)
- Bond angles (defined by 3 atoms)
- Torsion angles (defined by 4 atoms)
- Planarity of the peptide bond
- Chirality of Cα (and Cβ in Ile and Thr)

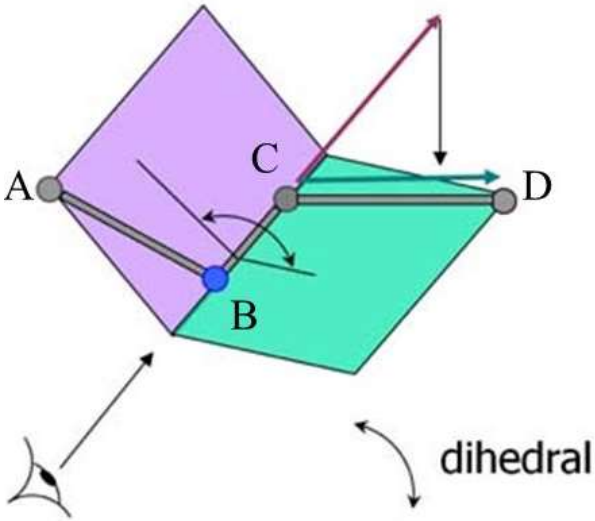




Geometry of the polypeptide chain

Torsion angles





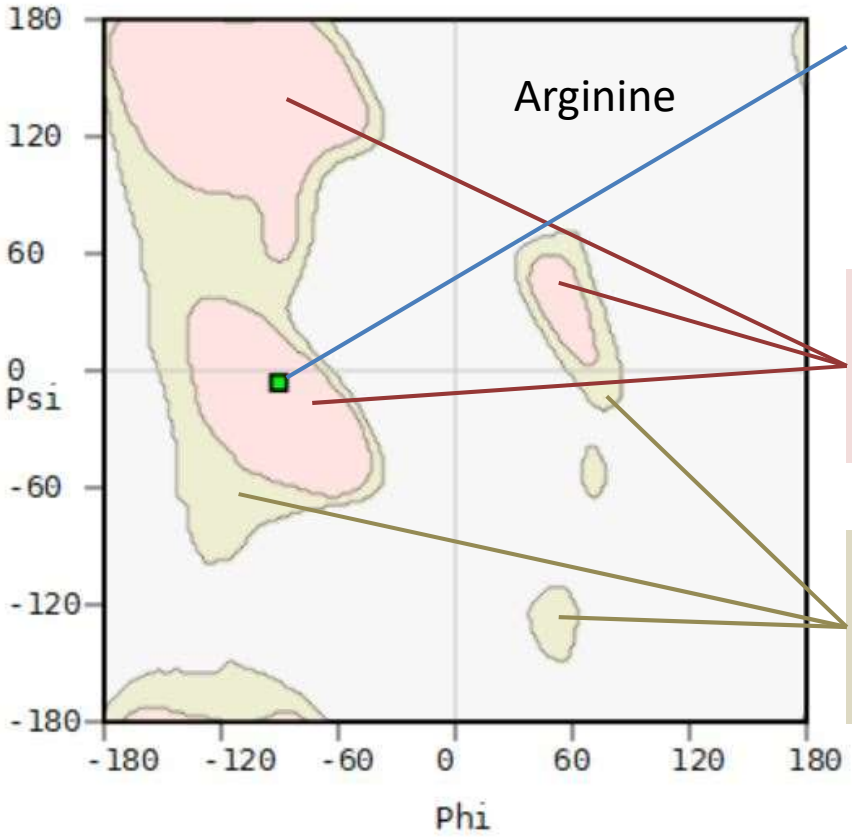
ω Torsion angle of the peptide bond: close to 180° (*trans*) or 0° (*cis* conformation, very rare! except for proline residues)

ϕ, ψ Torsion angles of the backbone: their variation shapes the conformation of the peptide chain

χ_1, χ_2, χ_3... Torsion angles of the side chain: vary according to side chains

Ramachandran Plot

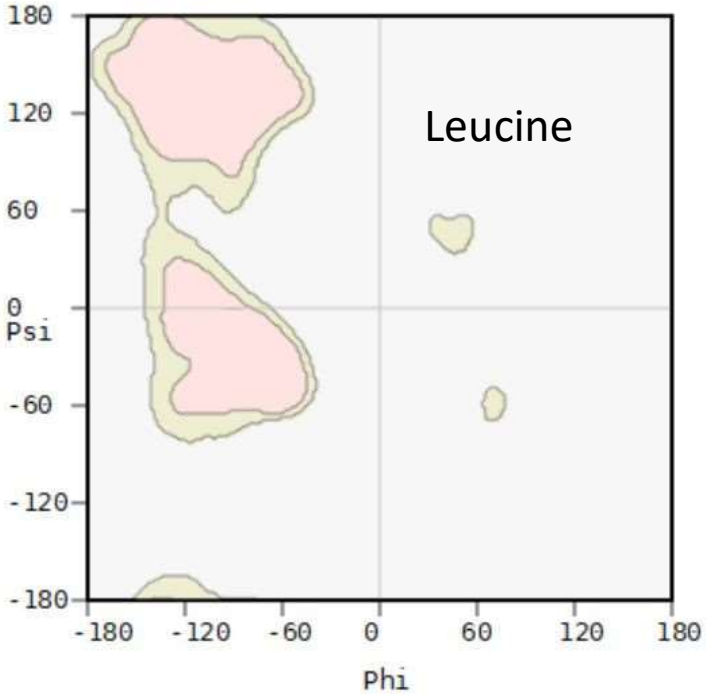
Analysis of the energetically favored conformations of the protein backbone, taking into account torsion angles ϕ and ψ.

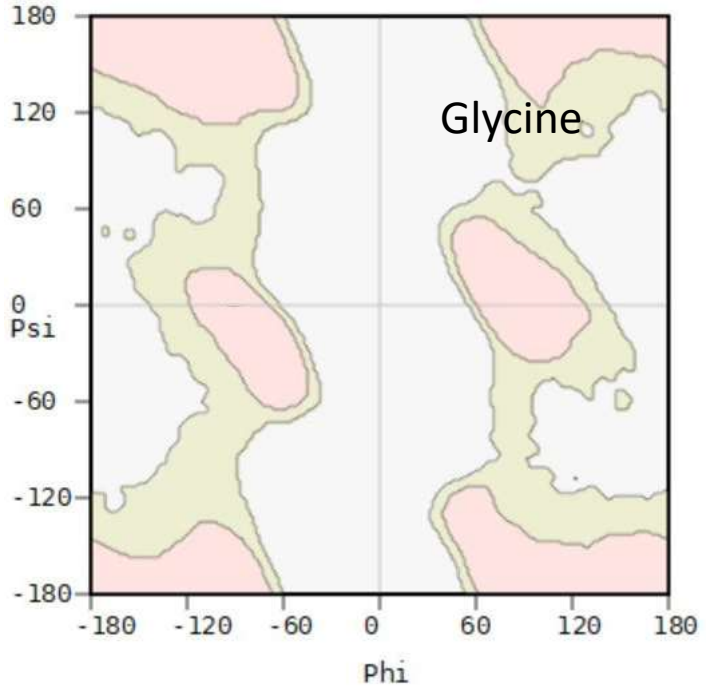


ϕ and ψ position of a single residue

Energy favoured regions

Energy allowed regions

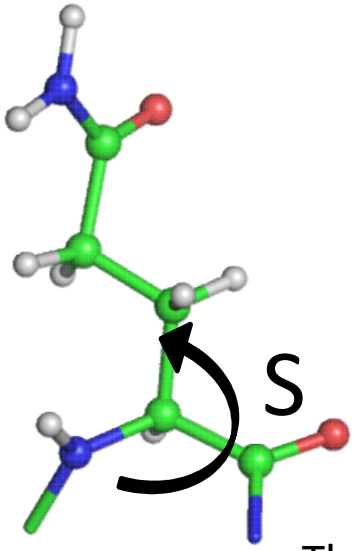




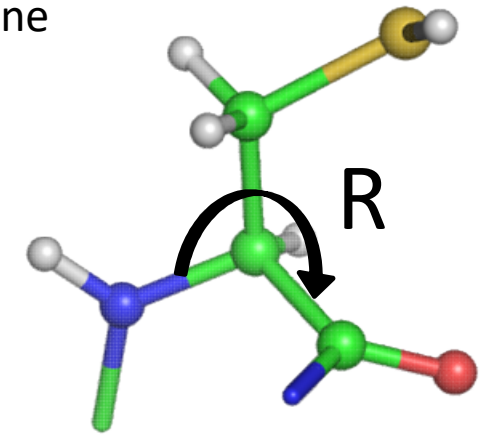
Chirality

In natural proteins, Cαs of all residues have the S configuration – except for Cys (R configuration) and Gly (non chiral).

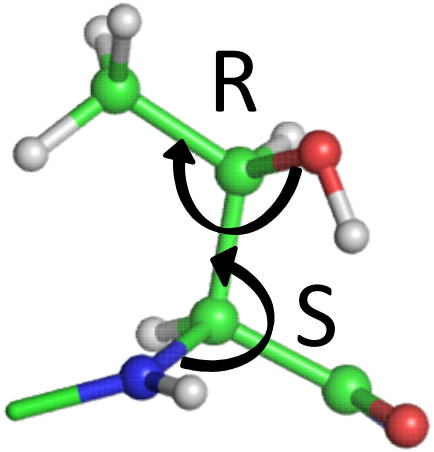
Glutamine



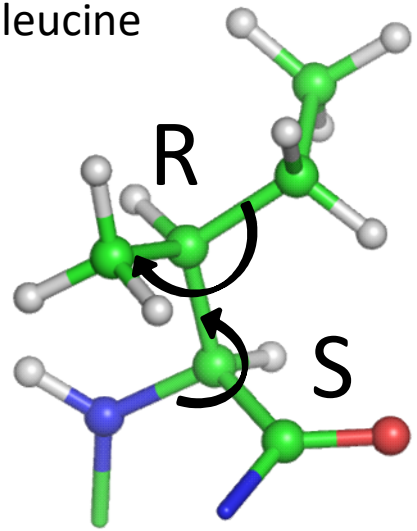
Cysteine



Threonine

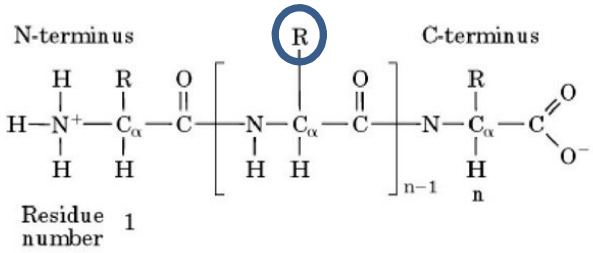


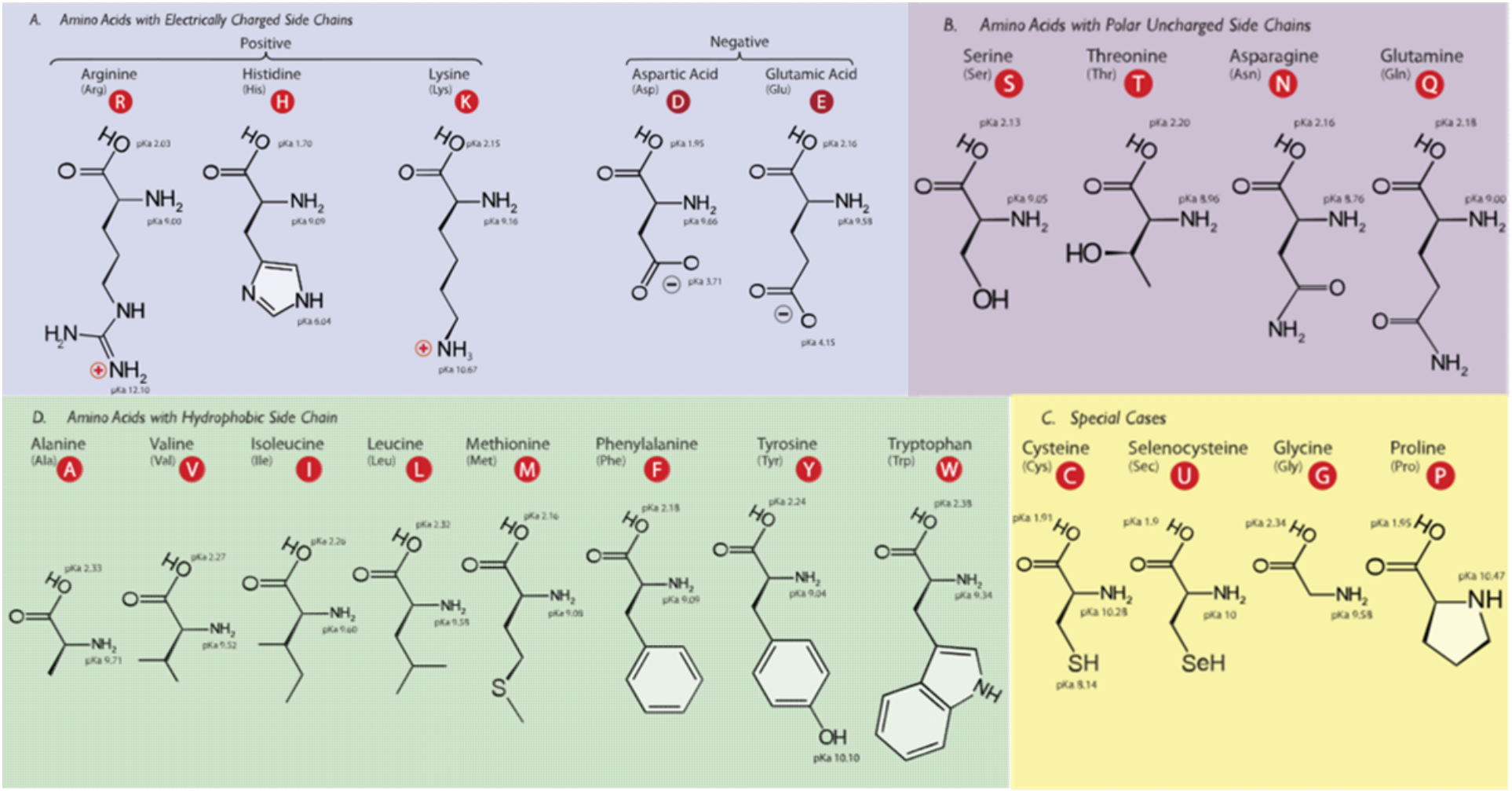
Isoleucine



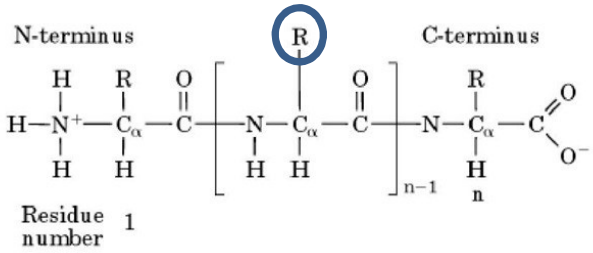
In Threonine and Isoleucine, Cβ is also chiral, with R configuration.

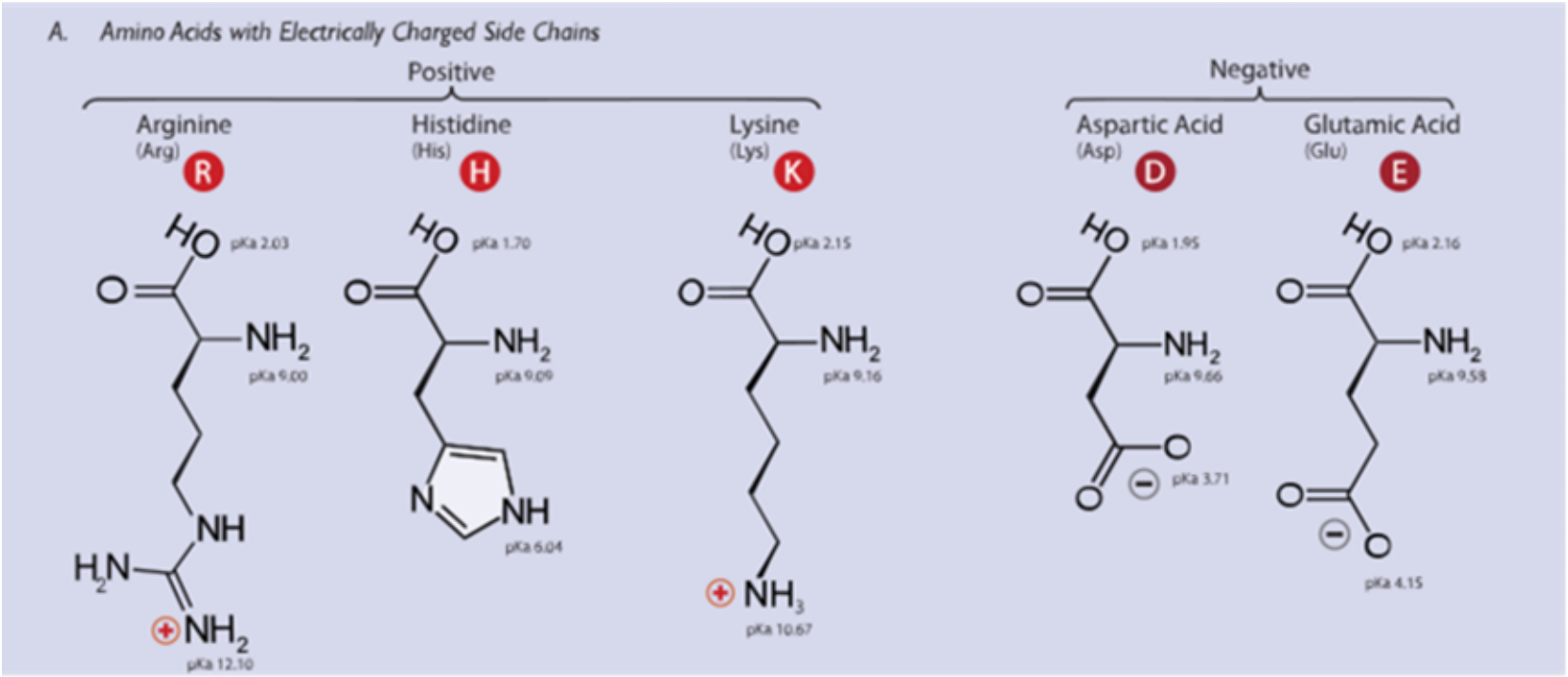
Side chains


N-terminus
R
C-terminus
Residue number 1
n−1
n


A. Amino Acids with Electrically Charged Side Chains
Positive
Arginine (Arg) R
Histidine (His) H
Lysine (Lys) K
Negative
Aspartic Acid (Asp) D
Glutamic Acid (Glu) E
B. Amino Acids with Polar Uncharged Side Chains
Serine (Ser) S
Threonine (Thr) T
Asparagine (Asn) N
Glutamine (Gln) Q
D. Amino Acids with Hydrophobic Side Chain
Alanine (Ala) A
Valine (Val) V
Isoleucine (Ile) I
Leucine (Leu) L
Methionine (Met) M
Phenylalanine (Phe) F
Tyrosine (Tyr) Y
Tryptophan (Trp) W
C. Special Cases
Cysteine (Cys) C
Selenocysteine (Sec) U
Glycine (Gly) G
Proline (Pro) P

Side chains



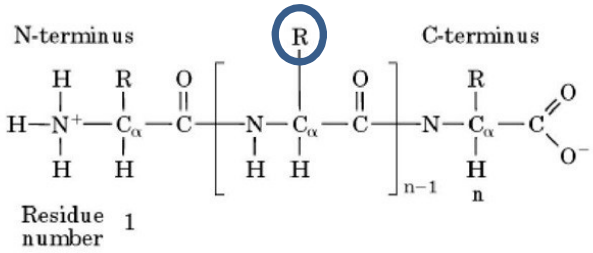


Charged residues are often present on the protein surface. pK_a of these residues depends on their surroundings in the protein folding.

Involved in salt bridges.

Both positively and negatively charged residues are often crucial for enzymatic activity and, therefore, located in active sites.

Side chains

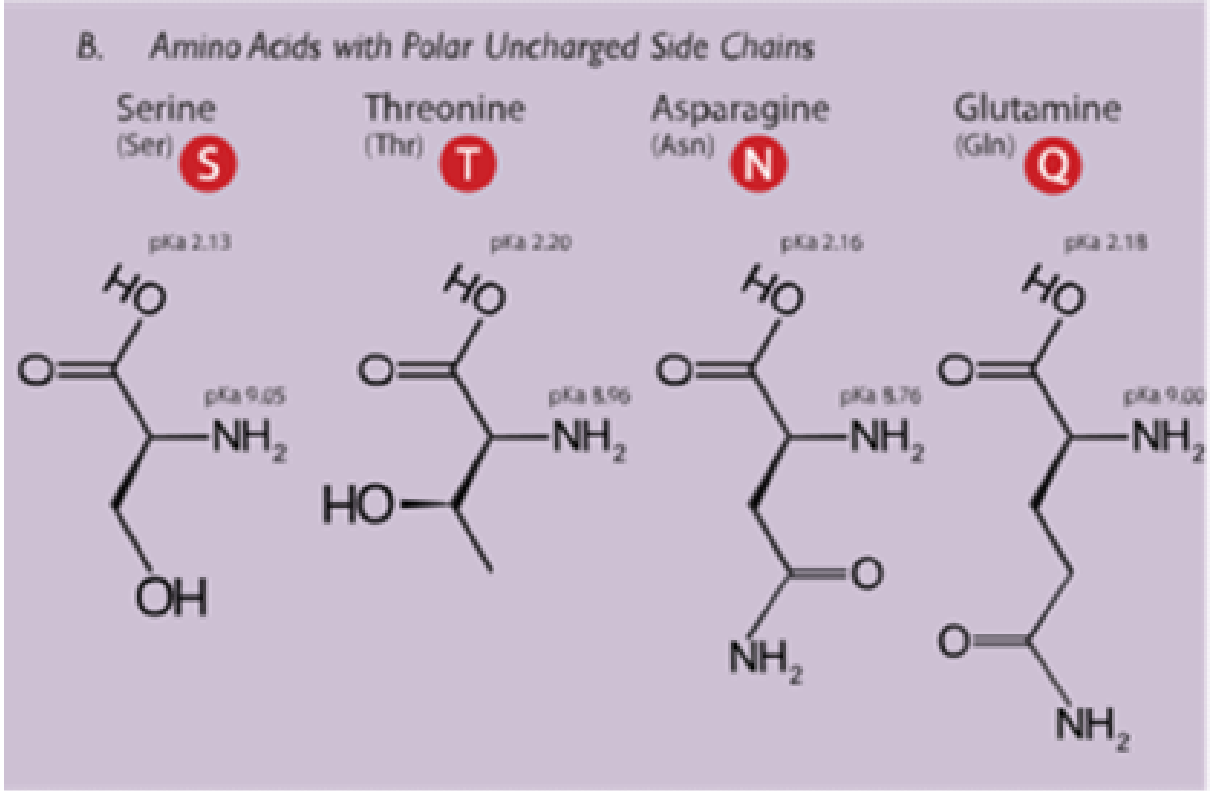


Often involved in enzyimatic reaction mechanisms.

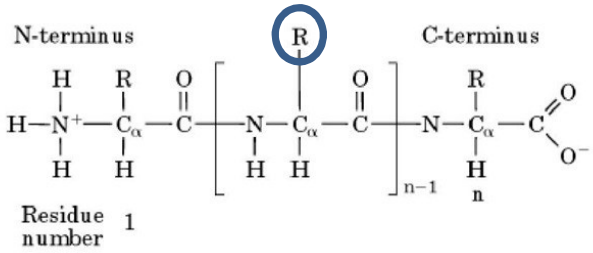
Act as hydrogen bond donors and acceptors.

Often sites of common post-translational modifications.

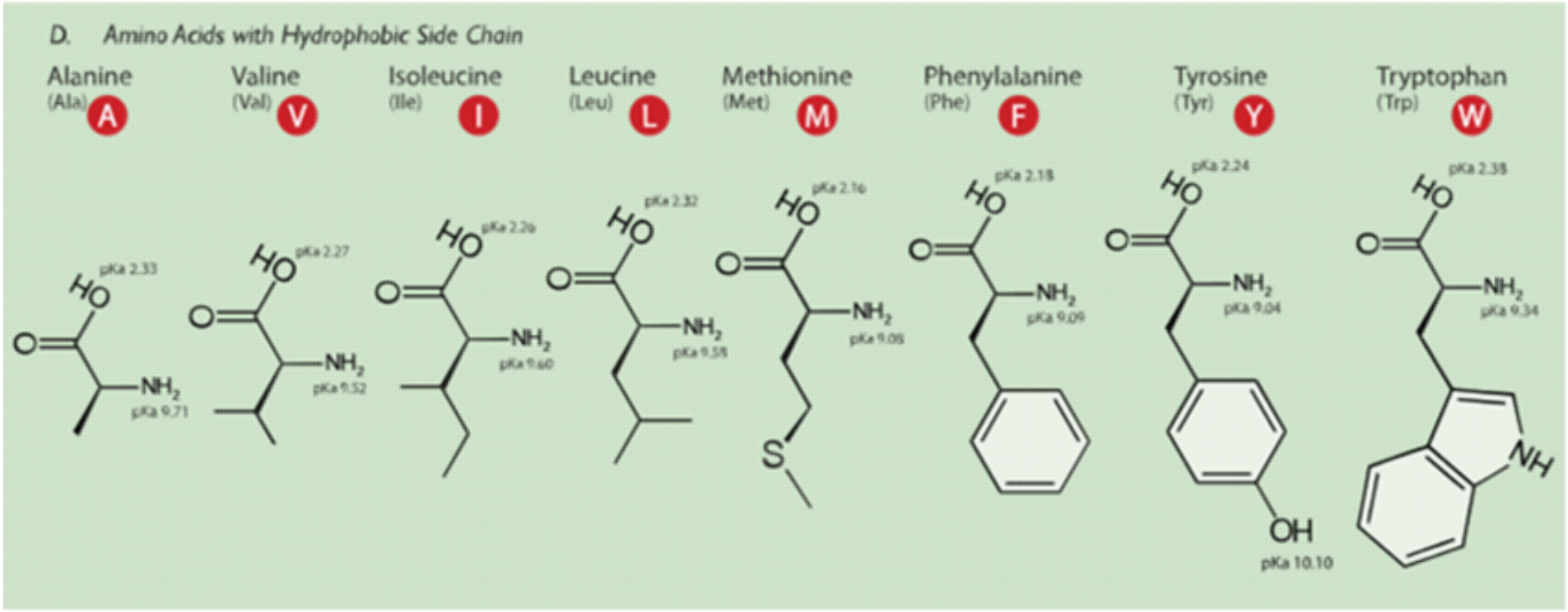
(Tyrosine residue may be included in this list, despite its large hydrophobic surface.)



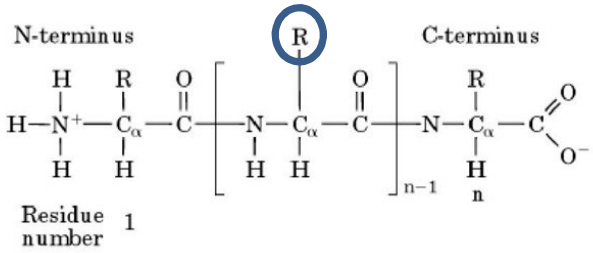
Side chains



Located in the core of the protein, their exclusion from water contact is an important driving force for protein folding. Structure predictions are also based on patterns formed by hydrophobic residues.



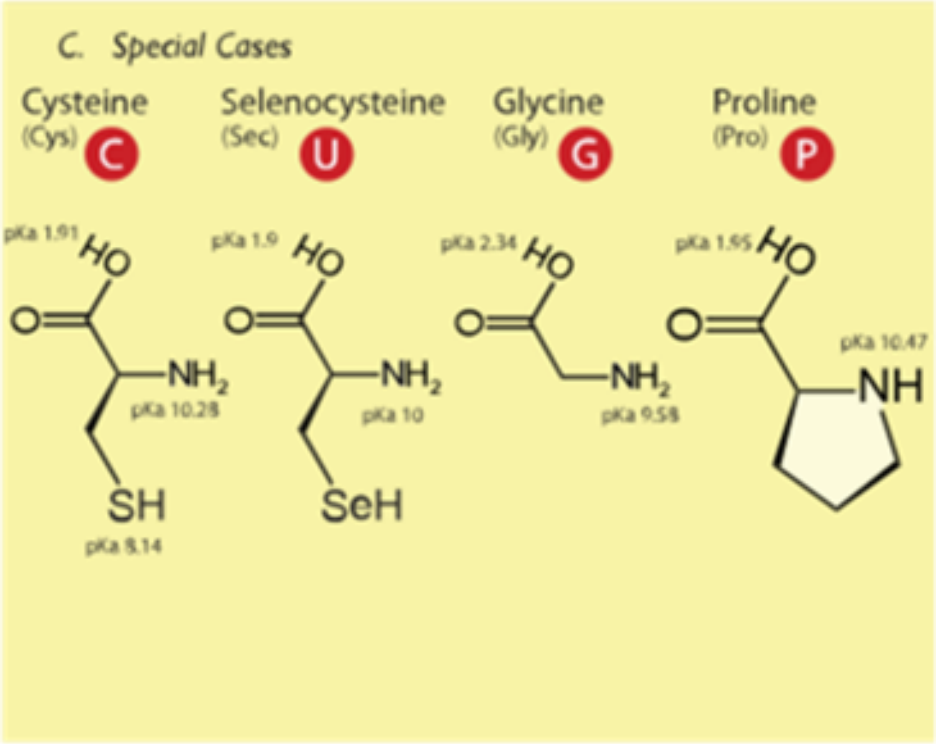
Side chains



Cysteine residues may form disulfide bridges – and are usually a serious issue for correct folding of recombinantly expressed proteins. Cysteine residues are useful for phasing as they can bind metal ions and the sulfur anomalous signal may be detected.

Glycine residues are often located in hinge regions of protein, due to their very small size. Glycine is the only non-chiral residue.

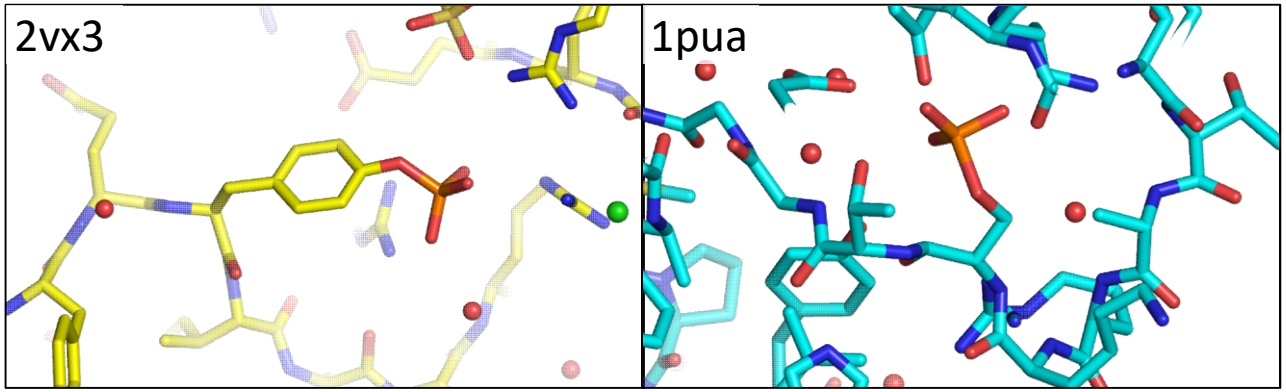
Proline is the only cyclic aminoacid. In protein folding it is often involved in breaking secondary structure elements (particularly α-helices). *Cis* and *trans* conformations have a similar energy.

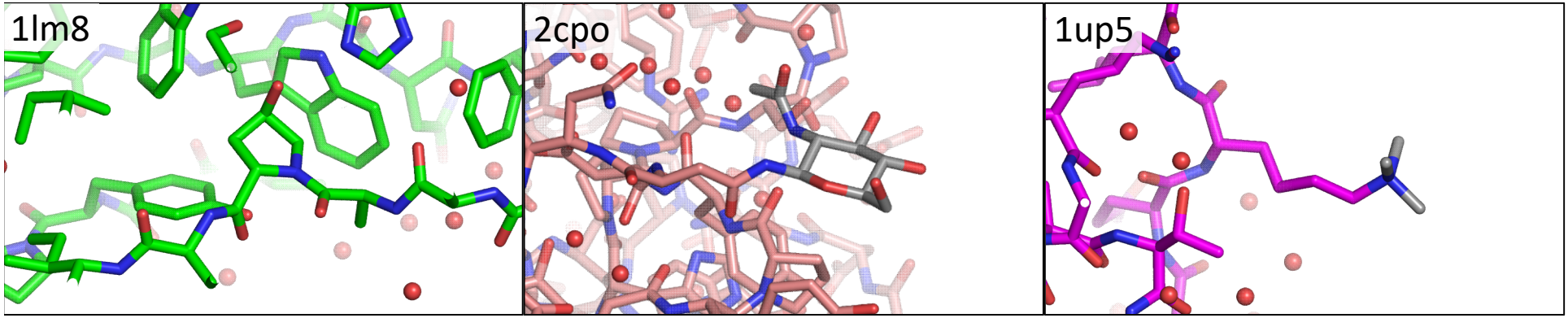


Post-translational modifications

Frequent in eukaryotic systems, include:

Phosphorylations of Ser, Thr, Tyr residues (fundamental for activation pathways of many enzymes by kinases)





Hydroxylation (e.g. Hydroxy-Pro are fundamental for collagen)

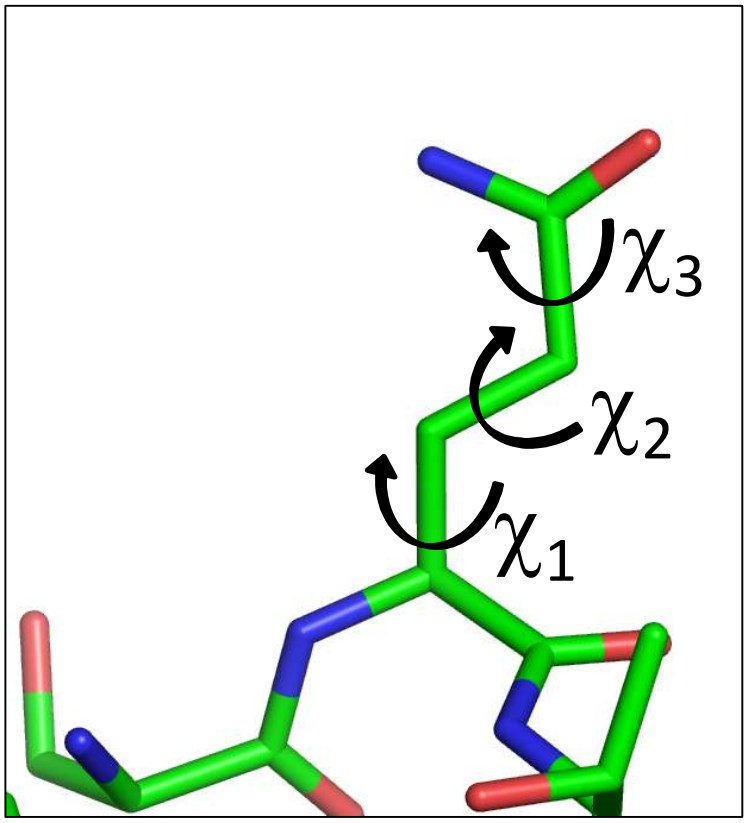
N-glycosylations on Asn and Gln, O-glycosylation on Thr and Ser

Methylation or acetylation of Lys and Arg in histones play a crucial role in gene expression

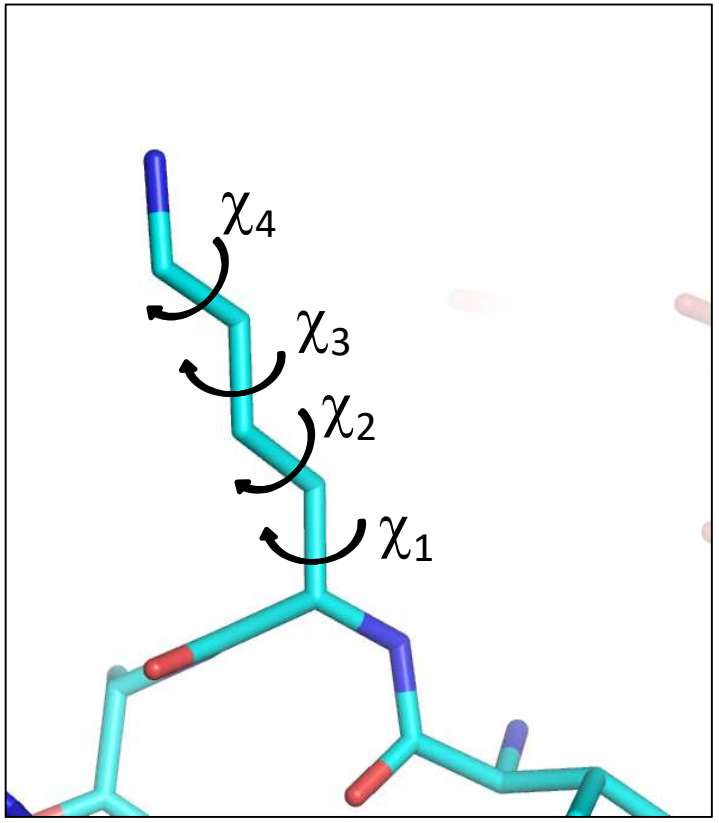
Side chain conformations

According to the different side chain, a number of different conformations (torsion angles χ_1, χ_2, χ_3, etc.) are energetically favored.

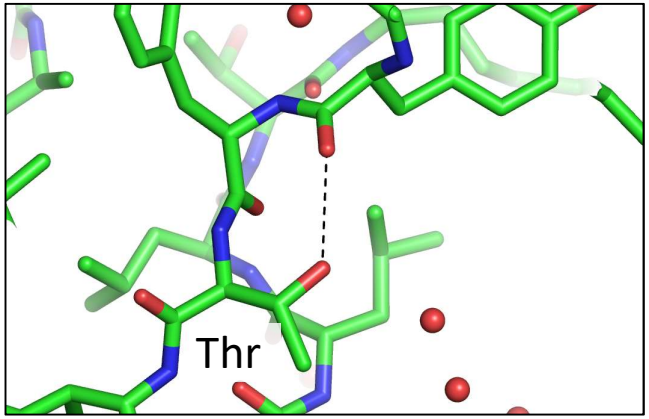
Possible conformers of Gln

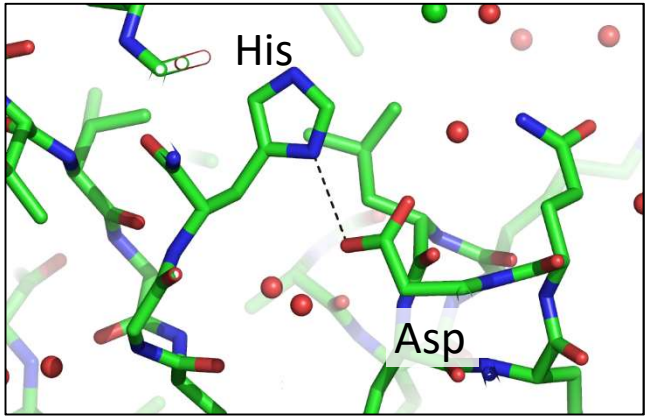


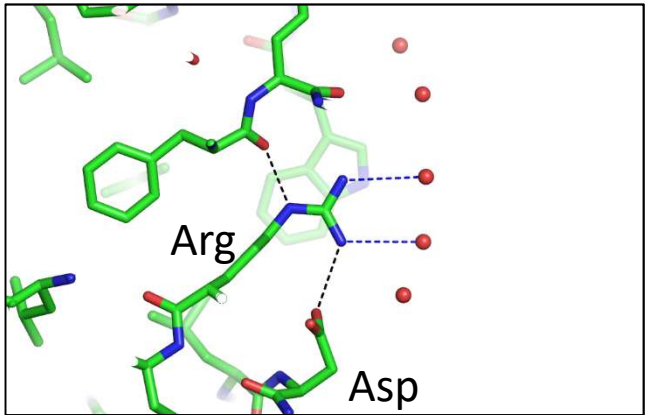
Possible conformers of Lys



Side chain interactions





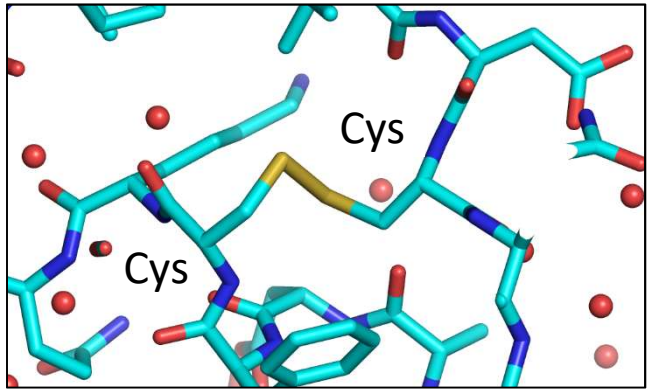


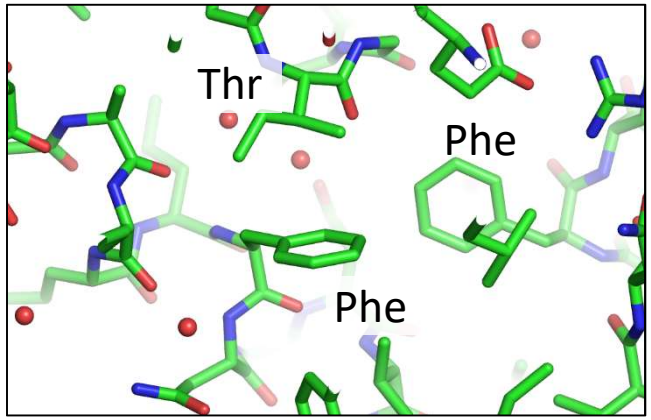
Fundamental for protein folding.

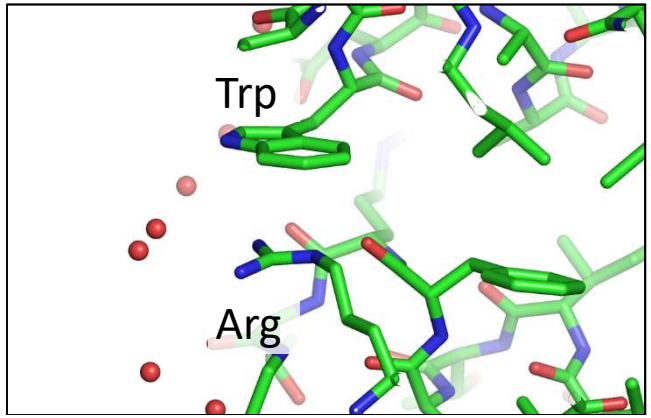
Covalent interactions: <u>disulfide bonds</u> involving Cys residues, other covalent bonds involving cofactors, inhibitors, ligands...

Hydrophilic interactions: <u>Hydrogen bonds</u> (average donor-acceptor distance 2.8 Å) involving polar groups or water, <u>salt bridges</u> involving charged groups or ions, <u>dipole-dipole</u> interactions involving side chains or solvent molecules.

Hydrophobic interactions: interactions based on <u>London dispersion forces</u>, or <u>π-π stacking</u>.



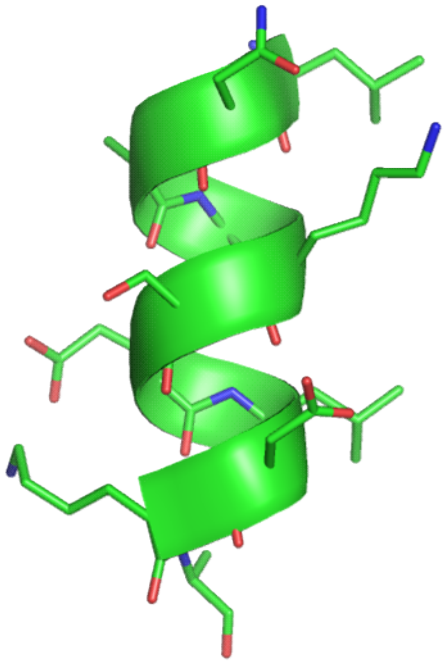




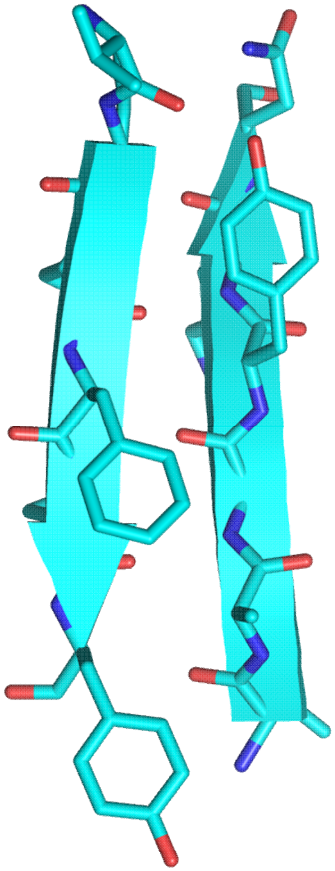
Secondary structure

Local arrangements of the peptide chain defined according to the geometry of **hydrogen bonds involving backbone** polar groups (C=O as acceptor and N-H as donor).

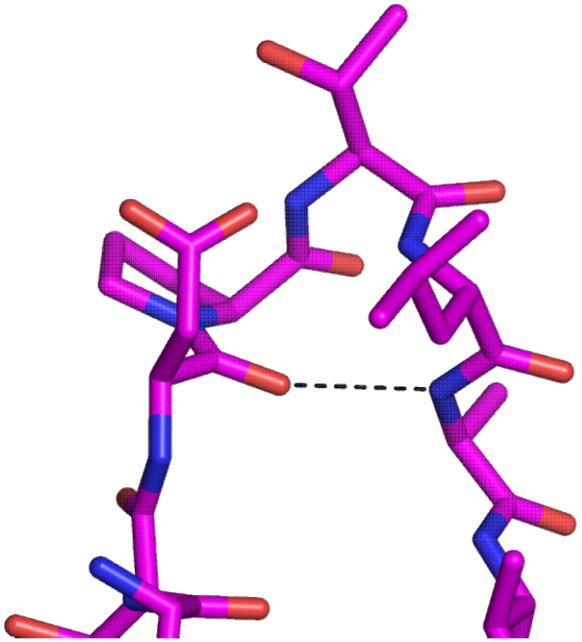
Helices



Sheets

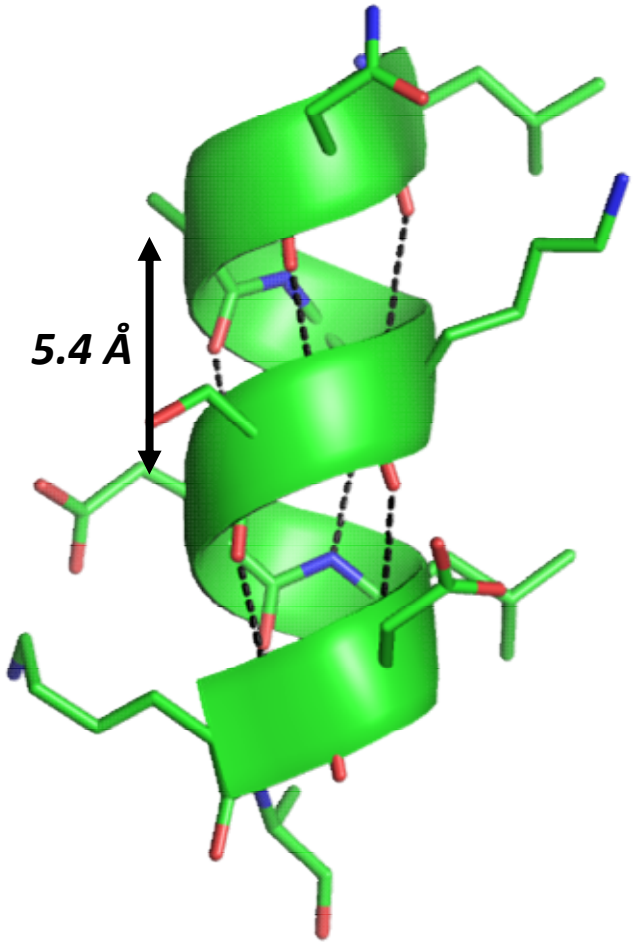


Turns

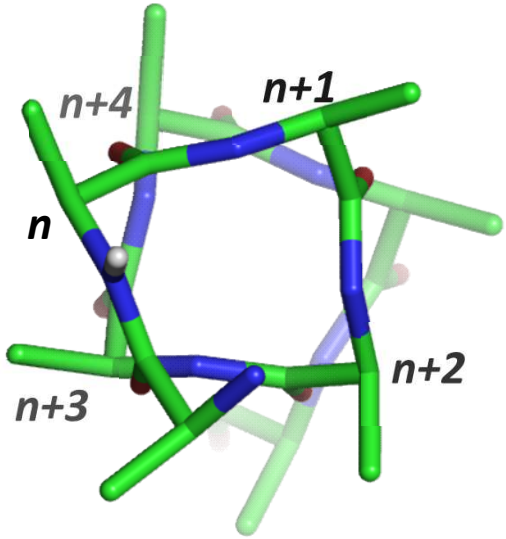


α-Helix

Clockwise helix – turns clockwise moving away from observer.



Hydrogen bonding pattern of the backbone:
C=O of residue *n* ····· H-N of residue *n+4*

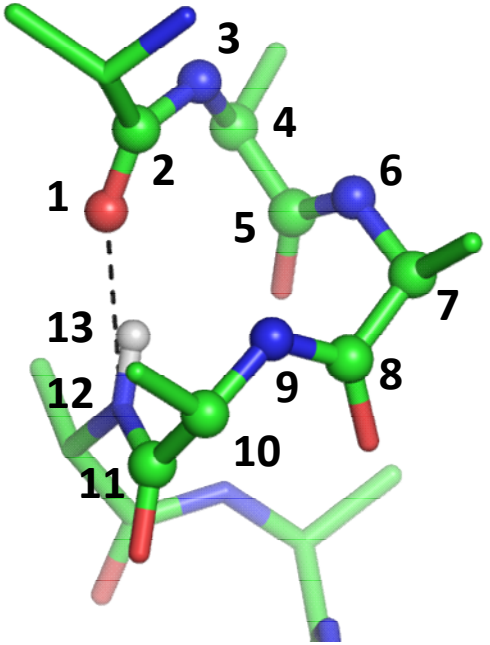


Each hydrogen bond forms a 13-atom closed circle (including H atom)

Also known as helix 3.6_{13}

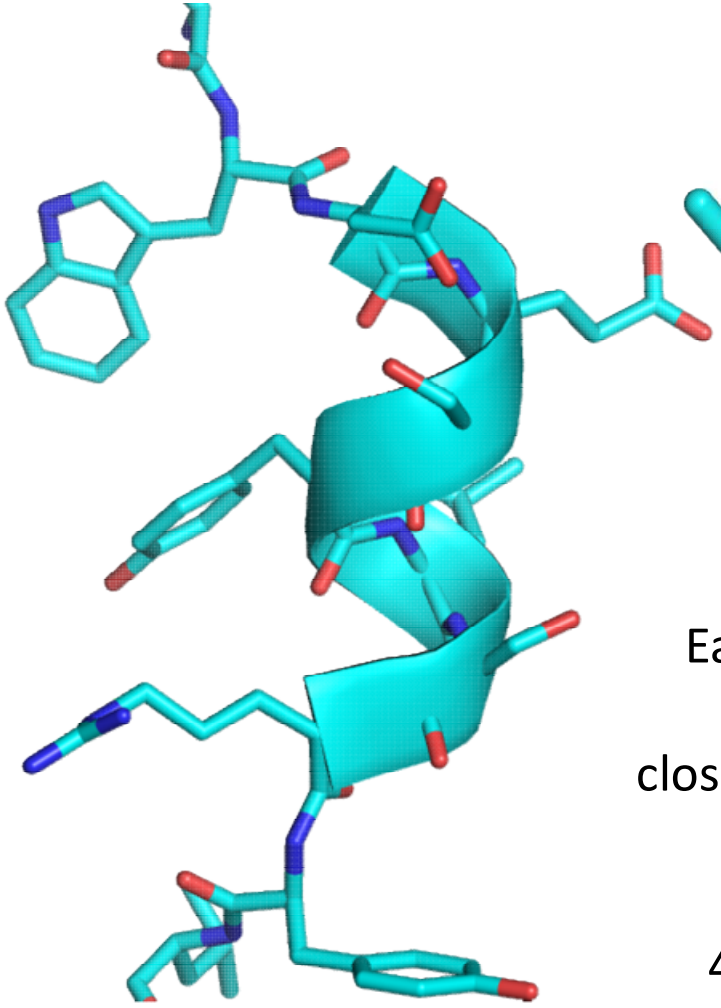
Geometry of the helix:

- 3.6 residues/turn
- rise of 1.5 Å/residue
- rise of 5.4 Å/turn

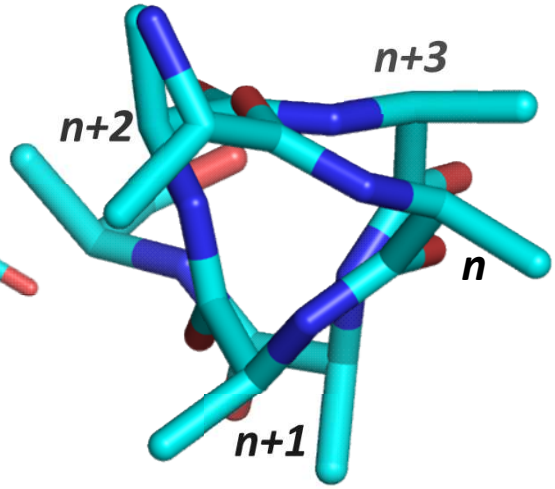


Helix 3_{10}

Clockwise helix, tightened with respect to α-helix

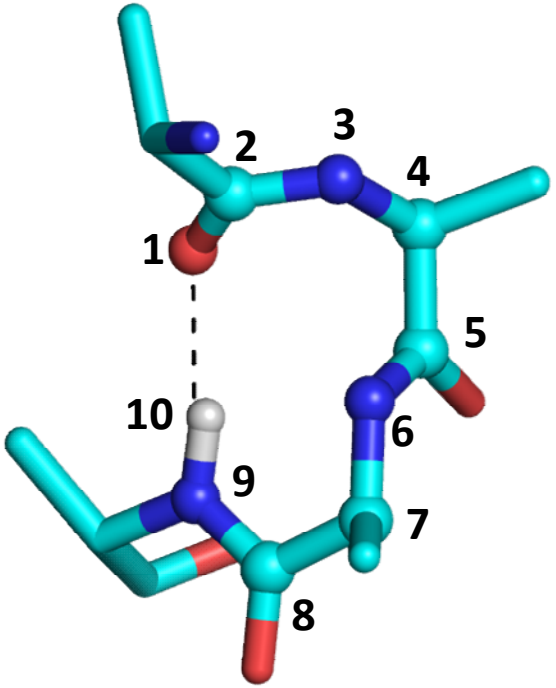


Hydrogen bonding pattern of the backbone:
C=O of residue *n* ····· H-N of residue *n+3*



3 residues/turn,
H-bonds tilted with respect to helix axis.

Usually short helices

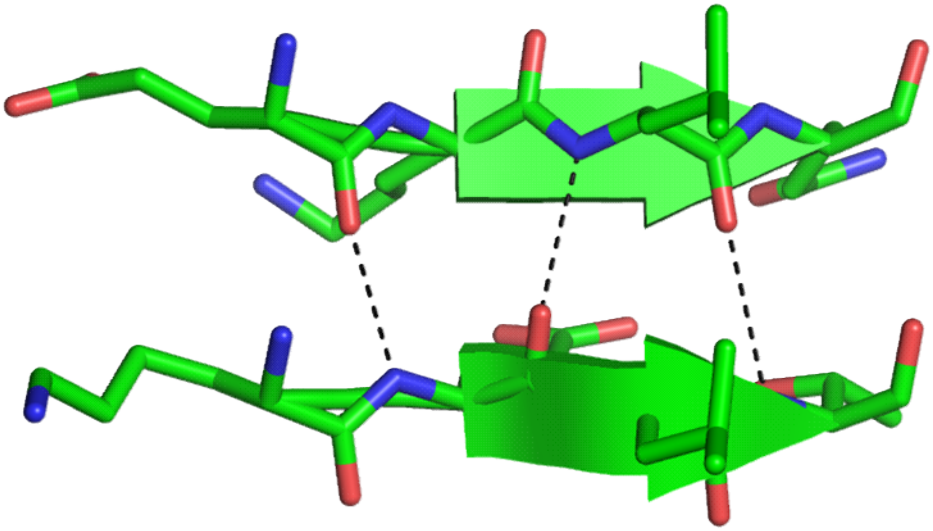


Each hydrogen bond forms a 10-atom closed circle (including H atom)

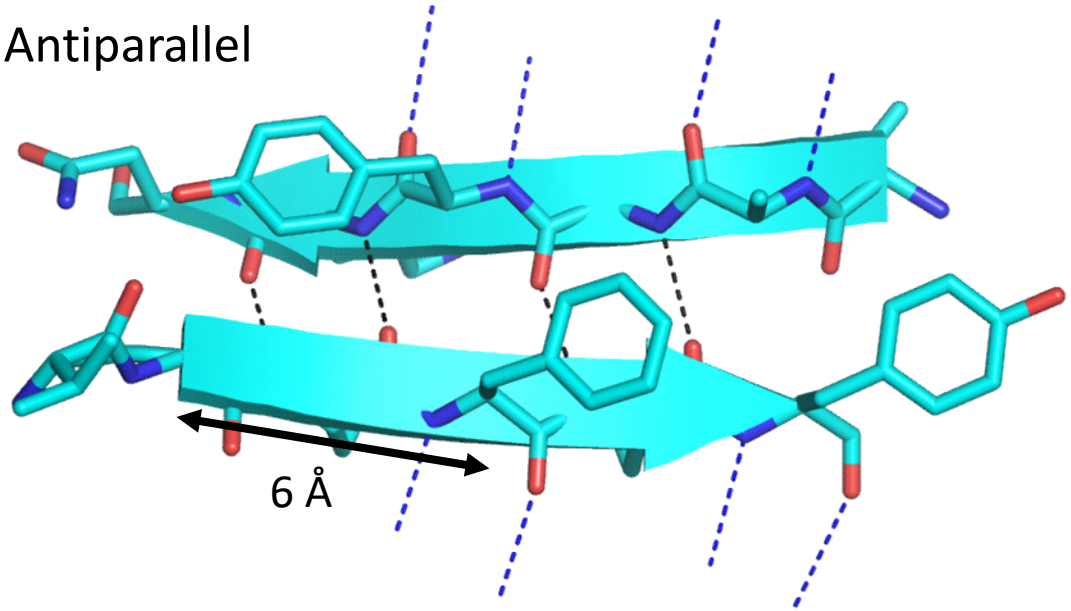
4% of all secondary structure elements

β-sheet

Parallel



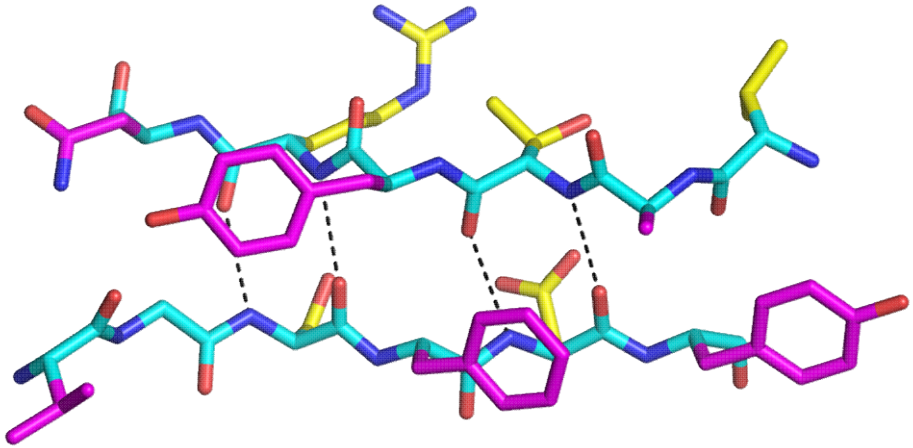
Antiparallel



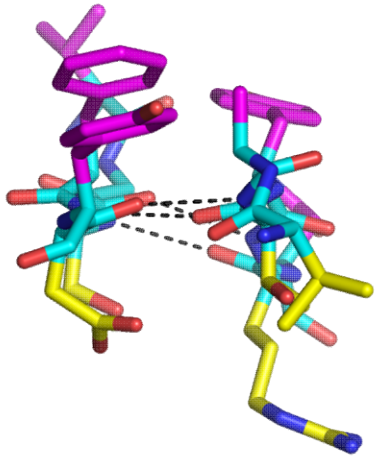
β-sheet

$C\alpha^{n}$-$C\alpha^{n+2}$ distance: 6 Å

Side chains are located above and below the plane of the hydrogen bonds.



View from side:

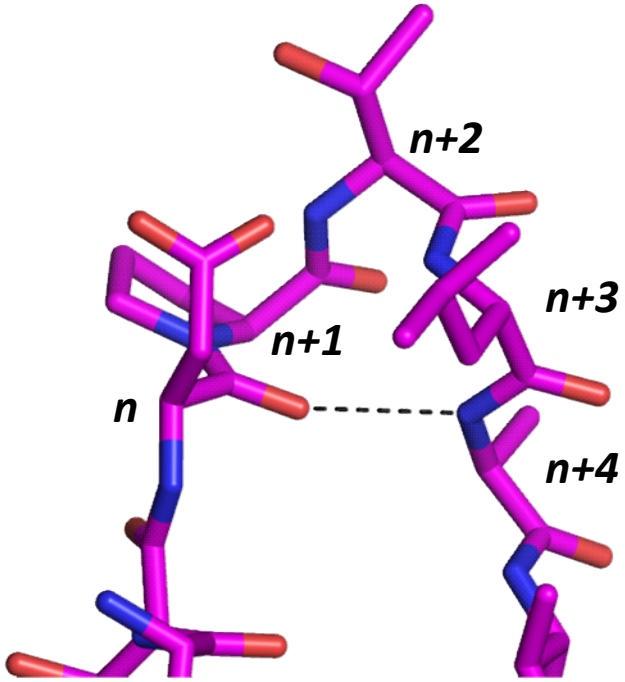


Turns e loops

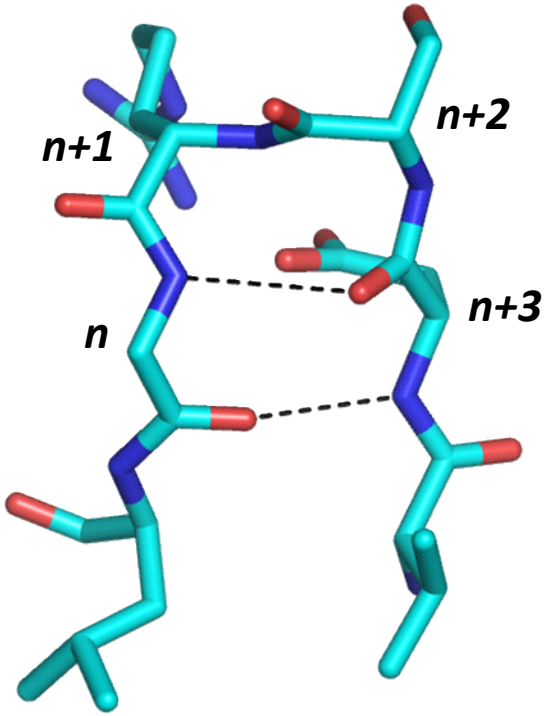
Connect other secondary structure elements. Turns can be defined based on the geometry of their H-bonds, loops are more flexible and less regular.

In biocrystallography, loops are usually the most difficult structural elements to determine, as their electron density is often poorly defined.

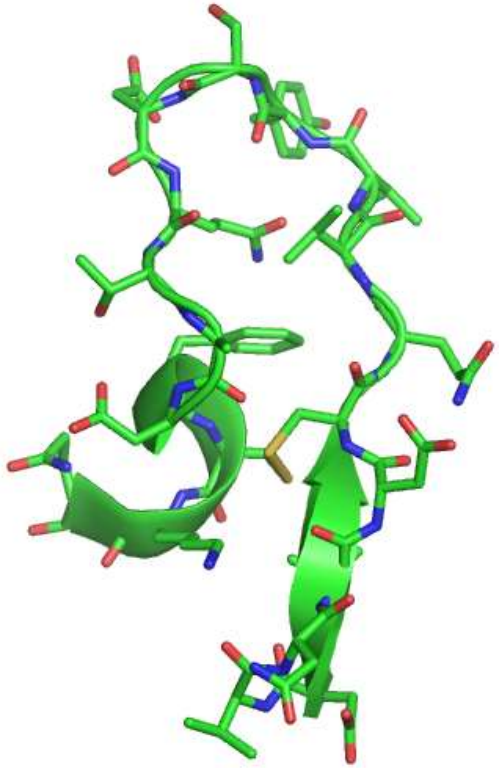
α-turn (4 residues involved)



β-turn (3 residues involved), also known as β-hairpin

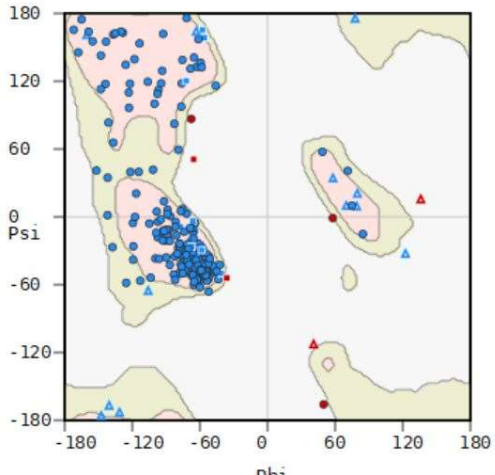


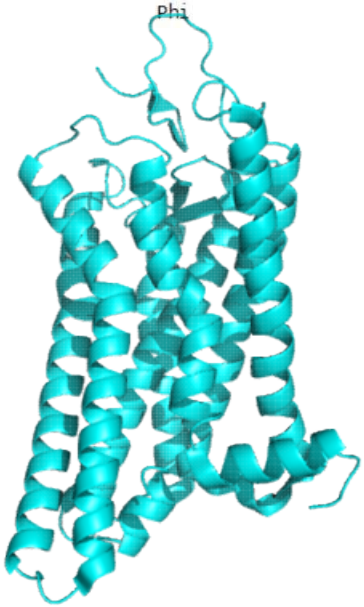
Loop



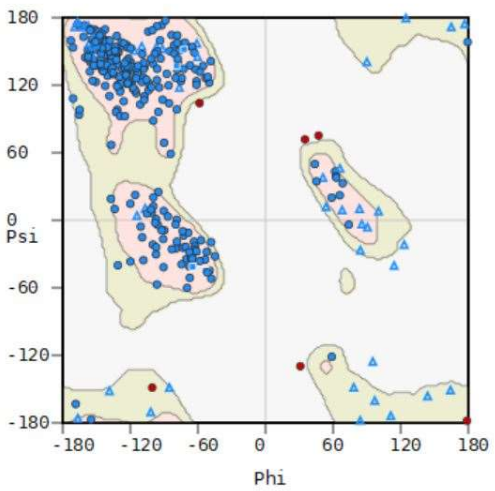
Ramachandran plot and secondary structure

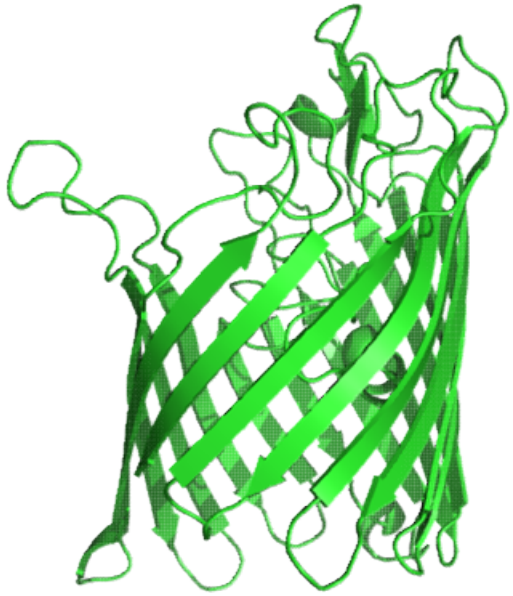
G-protein coupled receptor (4j4q): mostly α-helices



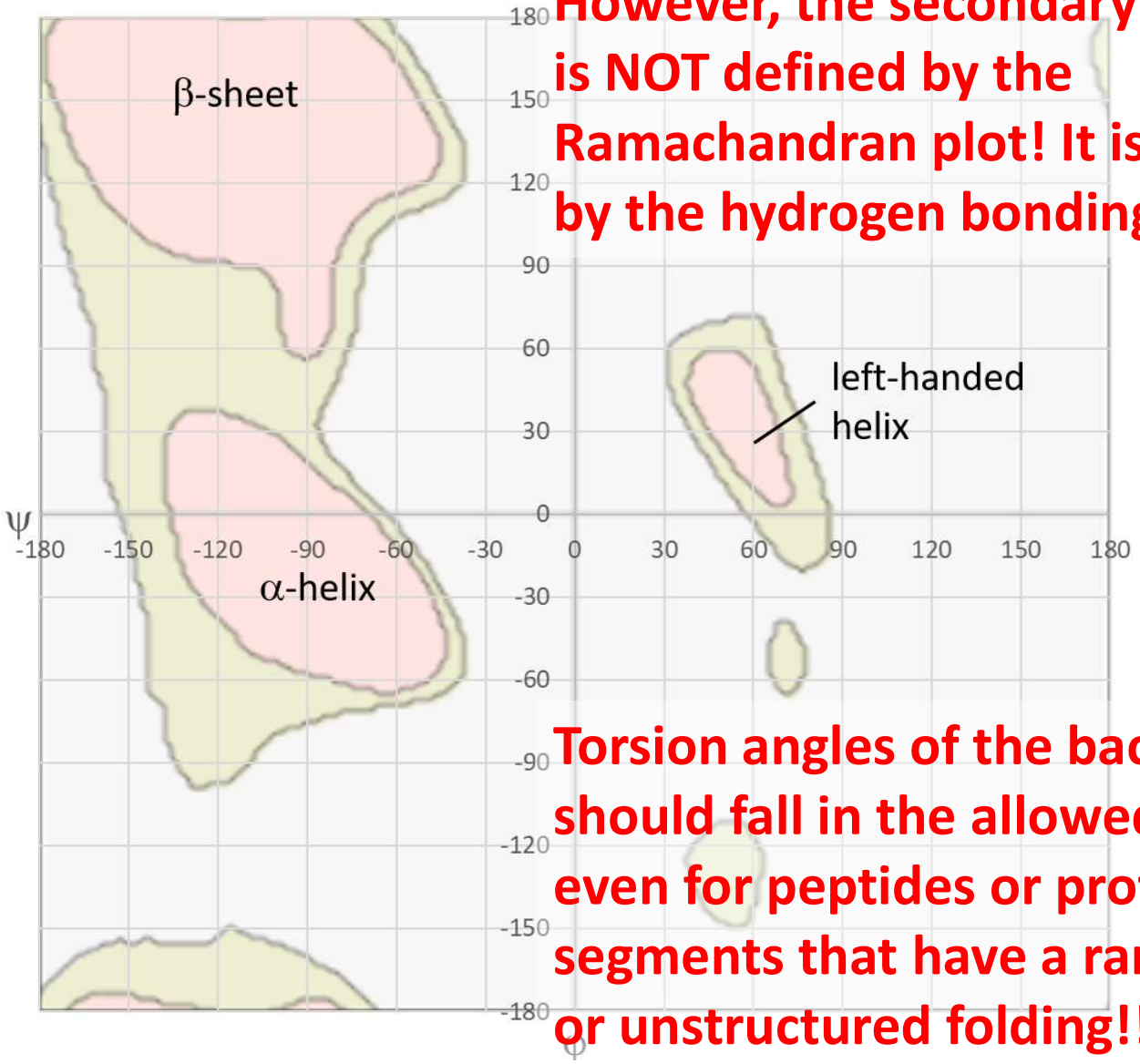


Outer membrane protein (2omp): mostly β-sheets





Ramachandran plot and secondary structure



However, the secondary structure is NOT defined by the Ramachandran plot! It is defined by the hydrogen bonding pattern!

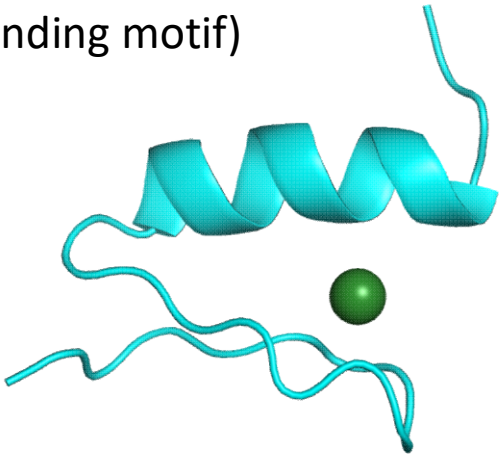
Torsion angles of the backbone should fall in the allowed regions even for peptides or protein segments that have a random coil or unstructured folding!!

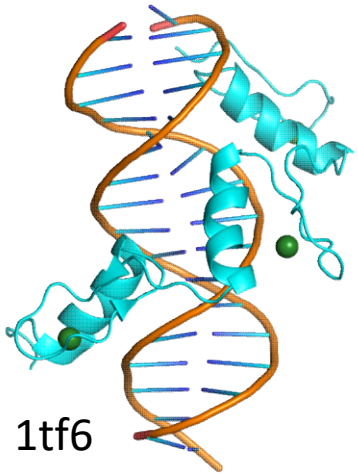
Tertiary structure: motifs and domains

Motif = combination of secondary structure elements, conserved in different structures

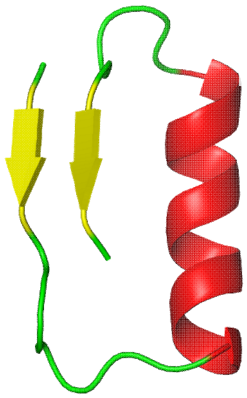
Domain = protein sequence that folds independently and usually has a specific function

Zinc finger (DNA binding motif)



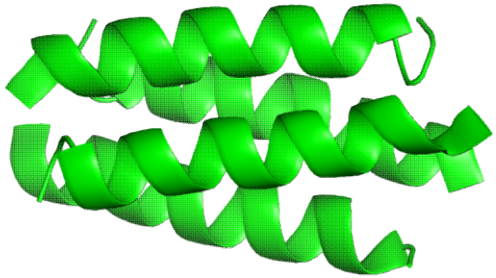


1tf6



2bnh

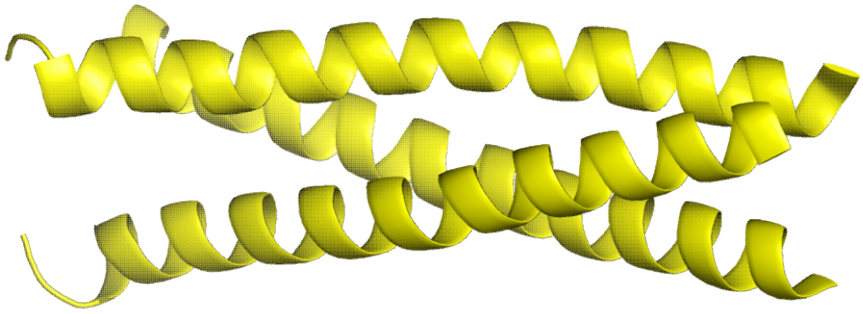
β-α-β motif



1lbi

Four- helix bundle

Leucine zipper domain



2pnv

Tertiary structure databases

Primary sequence similarity is not the only criteria to compare proteins: 3D structures may be conserved even when the sequence is not.

Common domain folding of diverse proteins suggests a similar function. To compare 3D structures of domains, structure classification databases:

- SCOP (scop.berkeley.edu): hierarchical classification of protein domains, based on the classification of all-α, all-β, α/β and α+β domains

Classes in SCOPe 2.07:

1. a: All alpha proteins [46456] (289 folds)
2. b: All beta proteins [48724] (178 folds)
3. c: Alpha and beta proteins (a/b) [51349] (148 folds)
4. d: Alpha and beta proteins (a+b) [53931] (388 folds)
5. e: Multi-domain proteins (alpha and beta) [56572] (71 folds)
6. f: Membrane and cell surface proteins and peptides [56835] (60 folds)
7. g: Small proteins [56992] (98 folds)
8. h: Coiled coil proteins [57942] (7 folds)
9. i: Low resolution protein structures [58117] (25 folds)
10. j: Peptides [58231] (148 folds)
11. k: Designed proteins [58788] (44 folds)
12. l: Artifacts [310555] (1 fold)

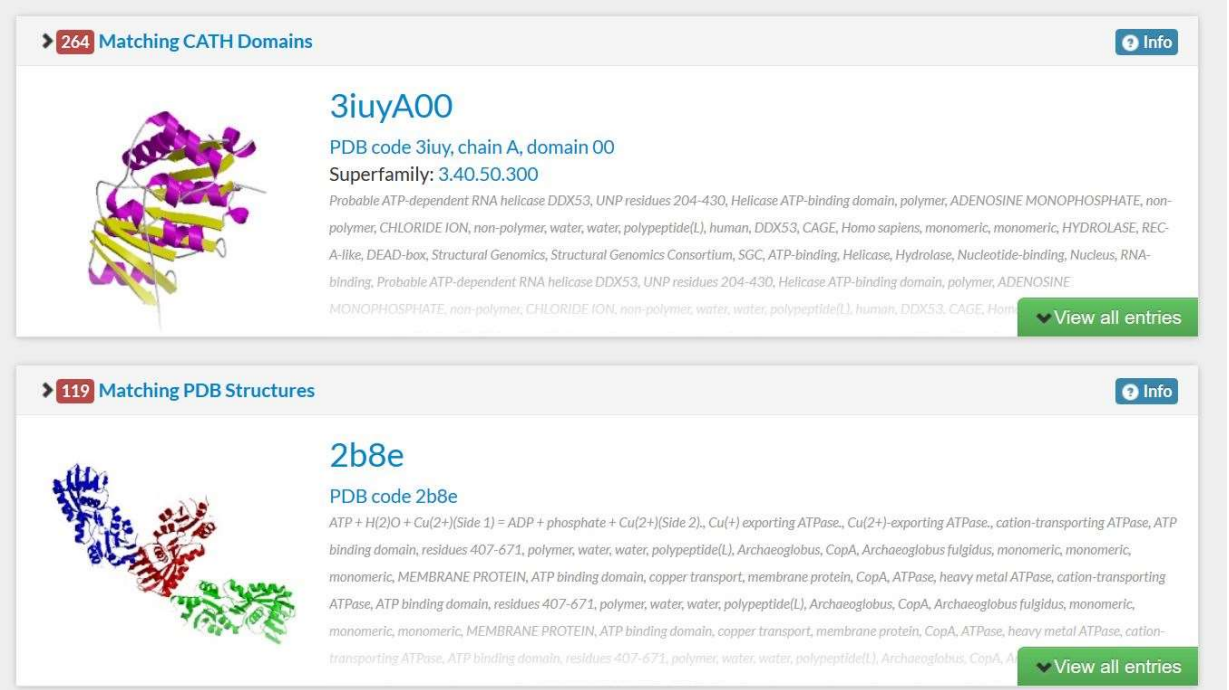
Tertiary structure databases

Primary sequence similarity is not the only criteria to compare proteins: 3D structures may be conserved even when the sequence is not.

Common domain folding of diverse proteins suggests a similar function. To compare 3D structures of domains, structure classification databases:

- SCOP (scop.berkeley.edu): hierarchical classification of protein domains, based on the classification of all-α, all-β, α/β and α+β domains
- CATH (www.cathdb.info/): classification according to structure and phylogenetics

Query: ATP binding domain

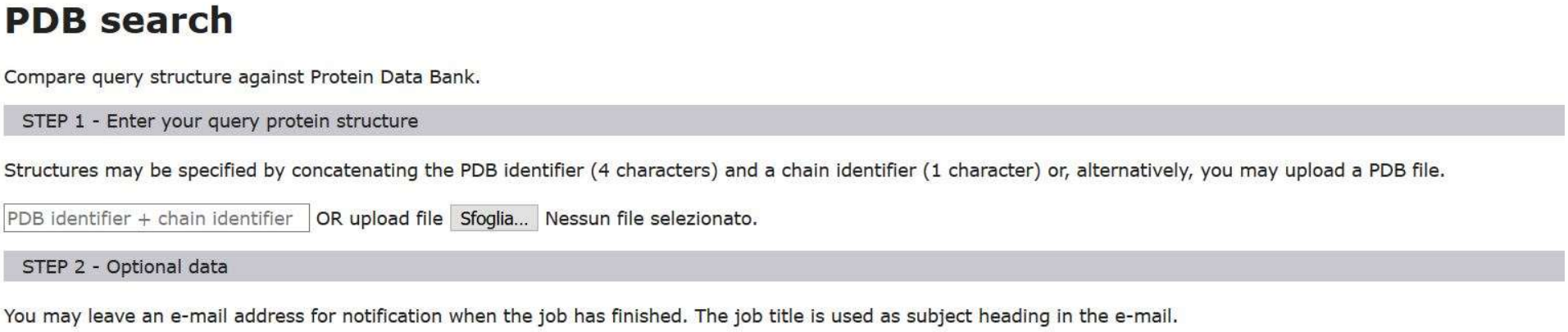


Tertiary structure databases

Primary sequence similarity is not the only criteria to compare proteins: 3D structures may be conserved even when the sequence is not.

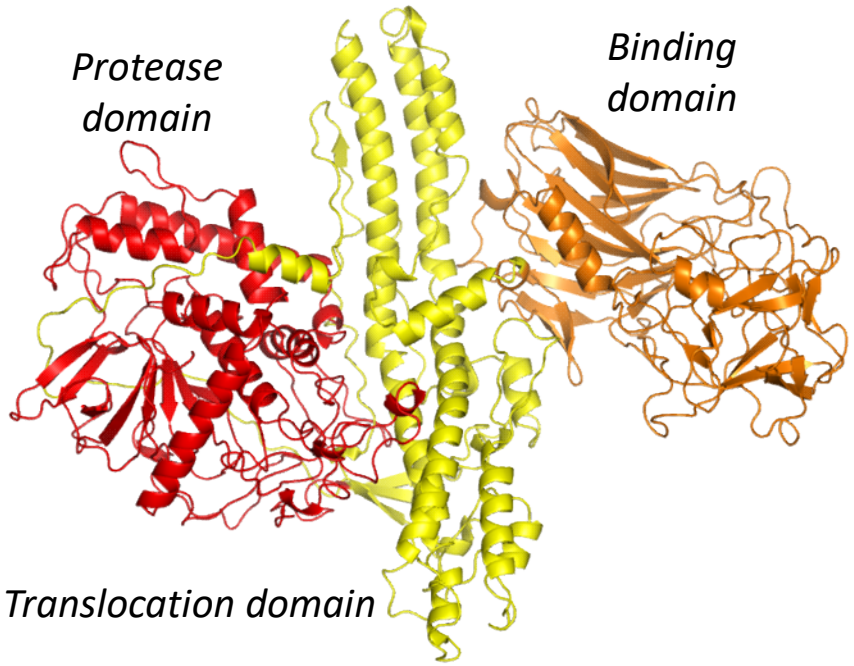
Common domain folding of diverse proteins suggests a similar function. To compare 3D structures of domains, structure classification databases:

- SCOP (scop.berkeley.edu): hierarchical classification of protein domains, based on the classification of all-α, all-β, α/β and α+β domains
- CATH (www.cathdb.info/): classification according to structure and phylogenetics
- Dali (ekhidna2.biocenter.helsinki.fi/dali): online software for tertiary structure comparison



Multidomain proteins

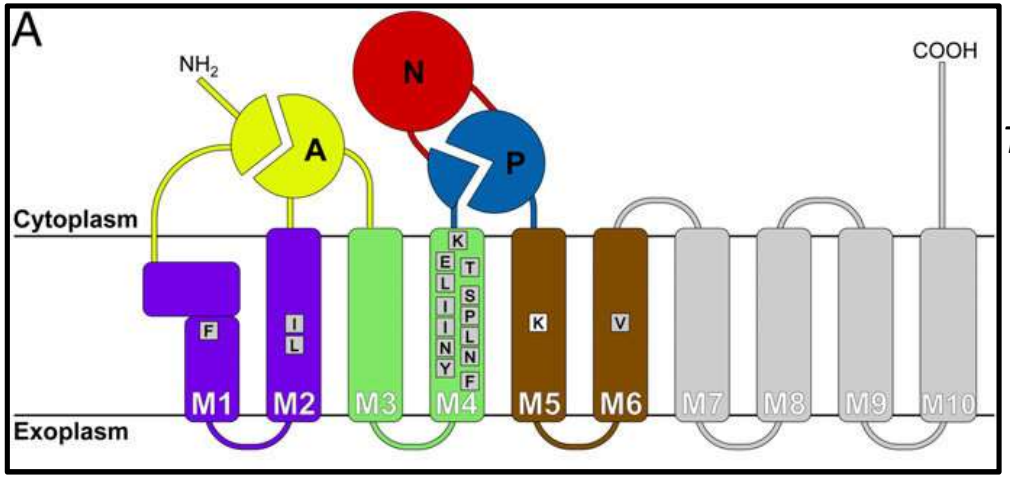
The tertiary structure of a multidomain protein includes different domains, that can be either subsequent in the protein sequence, or interdigitated. Each domain had a specific function connected to the protein activity.

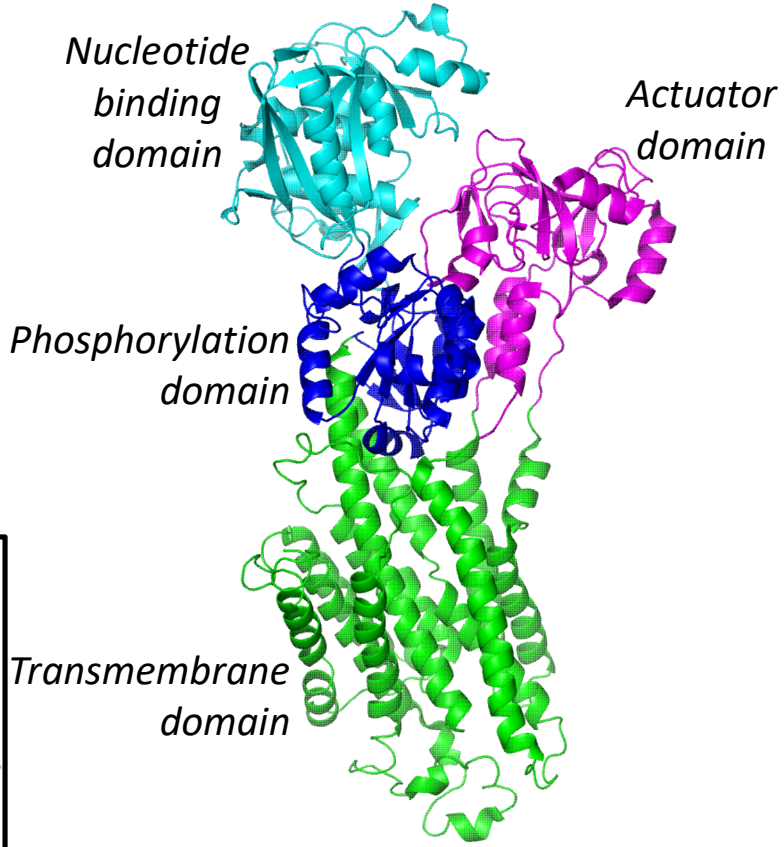


Neurotoxin from *Clostridium botulinum* (3bta)

Multidomain proteins

The tertiary structure of a multidomain protein includes different domains, that can be either subsequent in the protein sequence, or interdigitated. Each domain had a specific function connected to the protein activity.





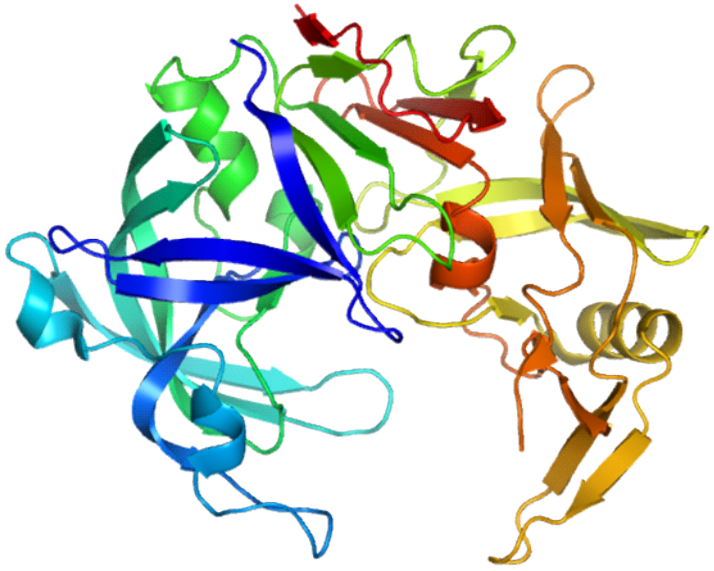
Sarcoplasmic/Endoplasmic Reticulum Ca^{2+} ATPase (4bew)

Quaternary structure

In some proteins, the functional unit is not formed by a single polypeptide chain, but by more subunits, held together by non-covalent interactions between their facing surfaces.

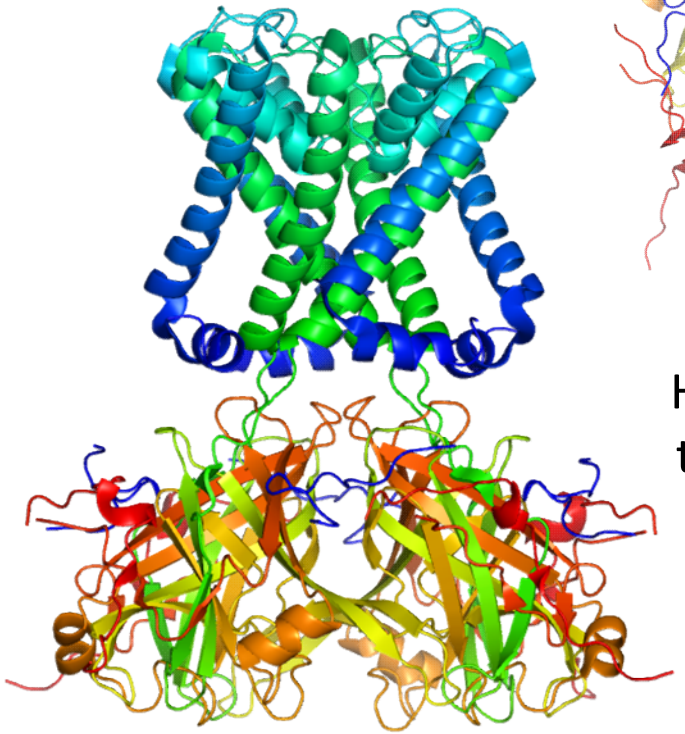
Multimeric proteins may be formed by repetitions of the same polypeptidic chain…

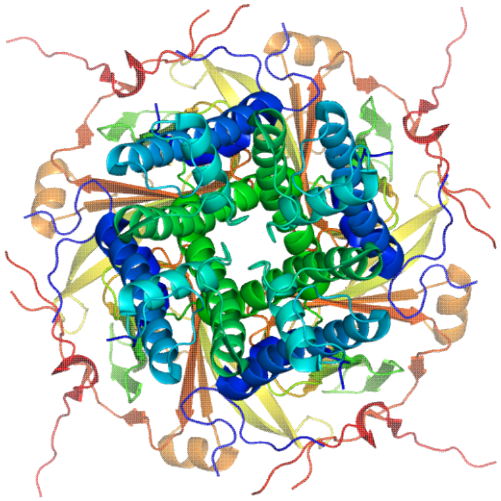
HIV protease
(3q70)



Homodimeric soluble protein

K^+ inwardly rectifying channel
(3zrs)

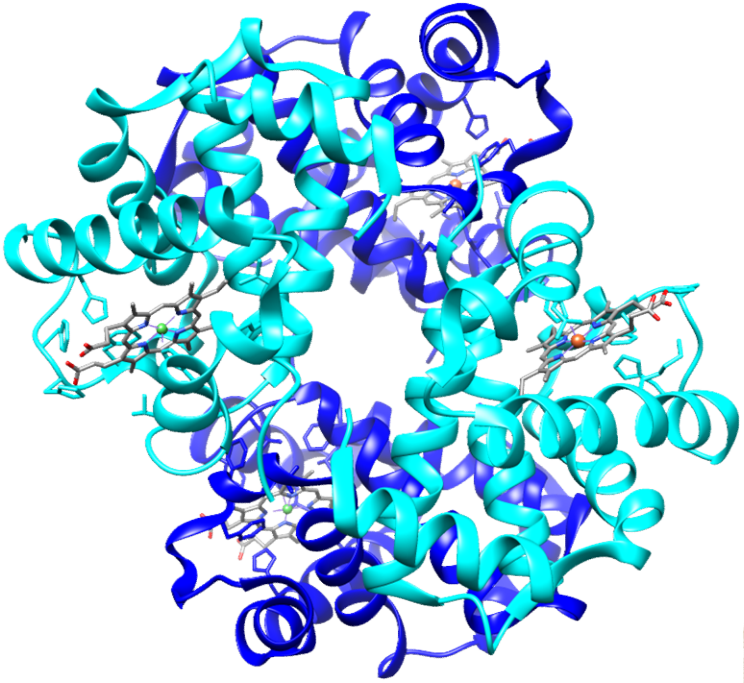




Homotetrameric transmembrane protein

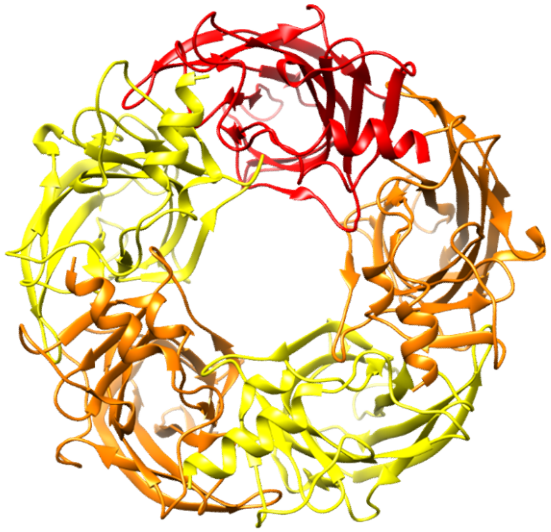
… or by different polypeptidic chains

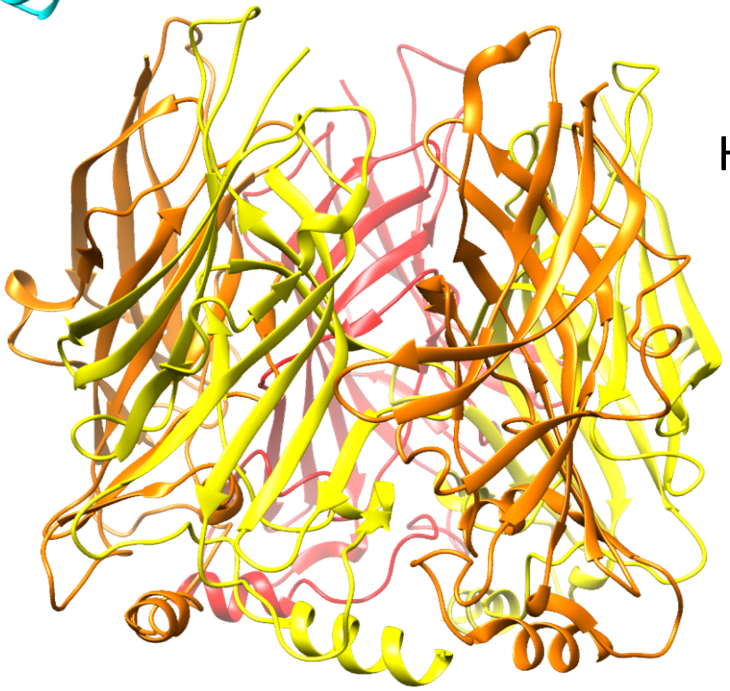
Haemoglobin (4n7o)



Heterotetrameric soluble protein

γ-aminobutyric acid receptor (GABA receptor) – neurotransmitter-gated ion channel (6dw1)



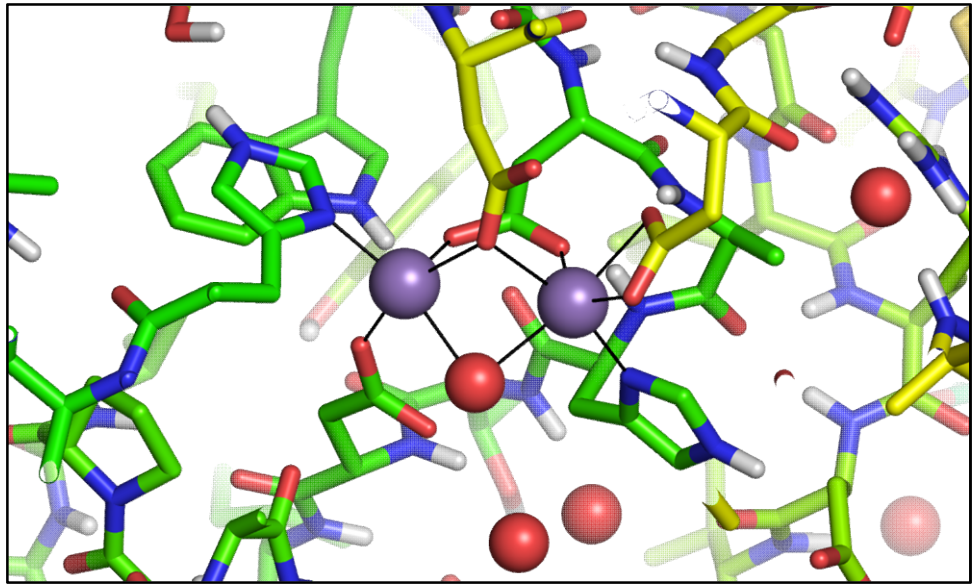


Heteropentameric transmembrane protein

Cofactors

Some proteins require the presence of cofactors, crucial for protein function.
Cofactors may be covalently bound or interacting through non-covalent interactions.
Some cofactors are inorganic metal ion complexes...

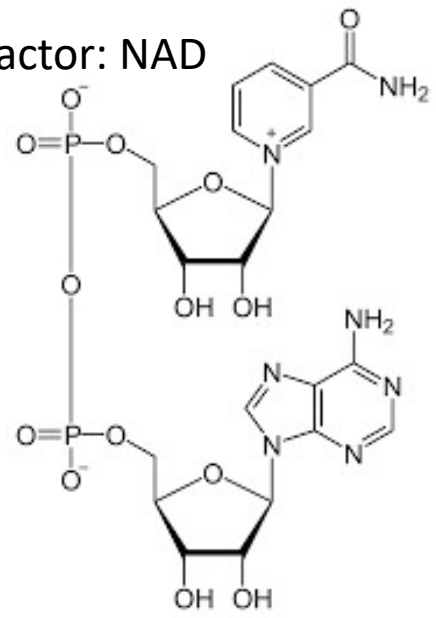
Cofactor: dinuclear manganese (II) complex

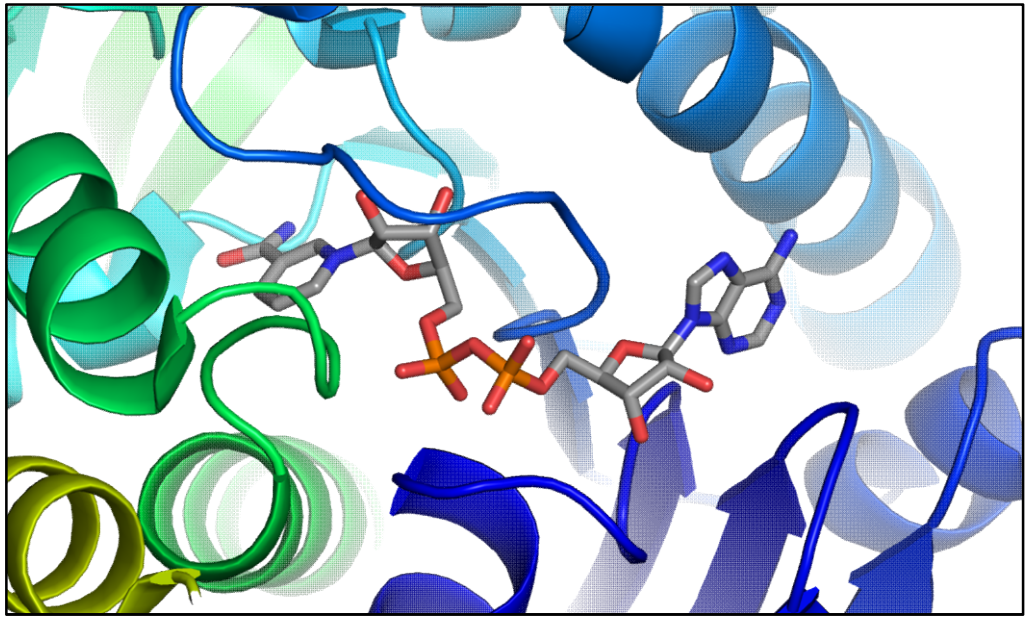


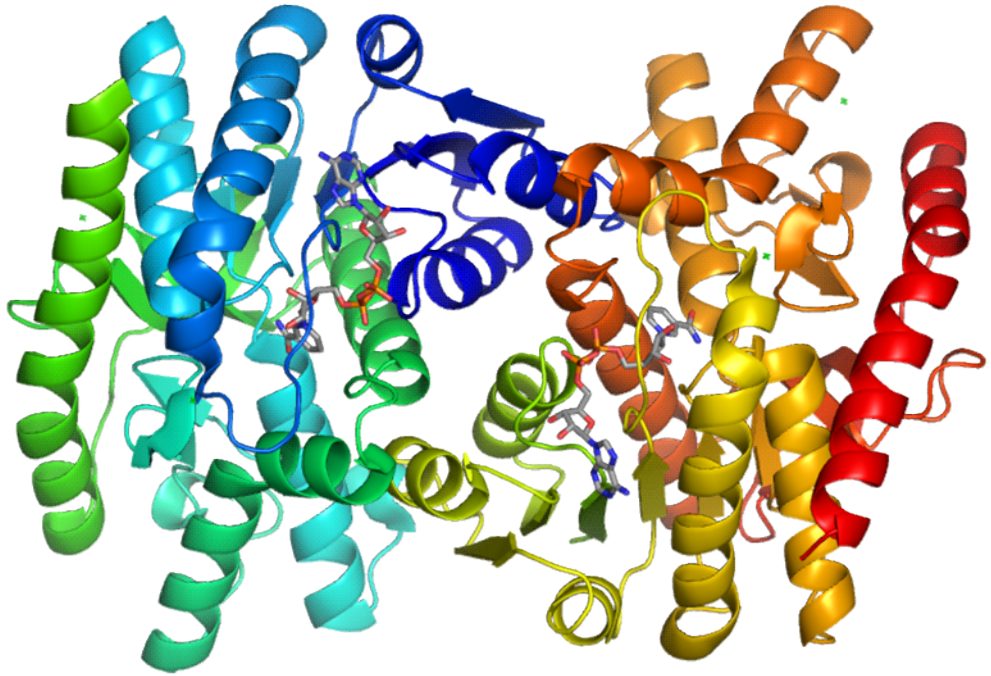
Liver Arginase
(1rla)

... other are organic molecules

Cofactor: NAD

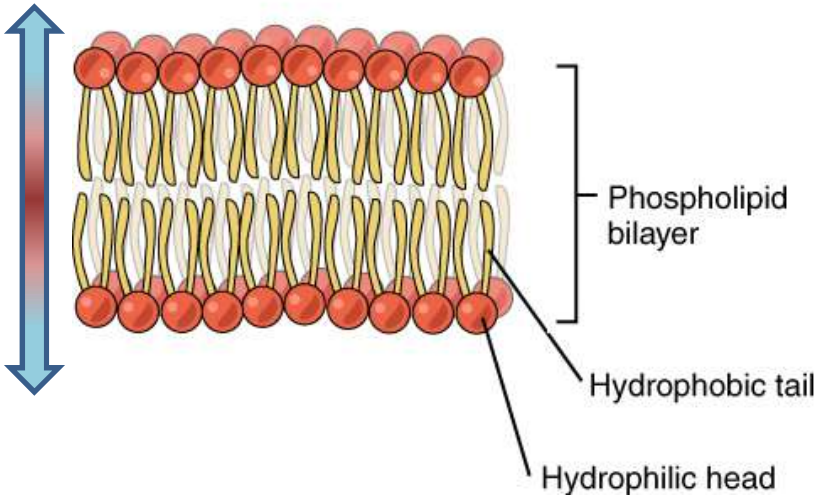






Human
Mitochondrial
Malate
Dehydrogenase
(2dfd)

Membrane proteins



30% of the proteome, with important for many physiological functions and for pharmaceutical chemistry as drug targets

External surfaces of the protein exposed to highly hydrophobic environment, crucial influence on protein folding

Only 2 tertiary structures for the transmembrane domains

β-barrel and **α-helix bundle**