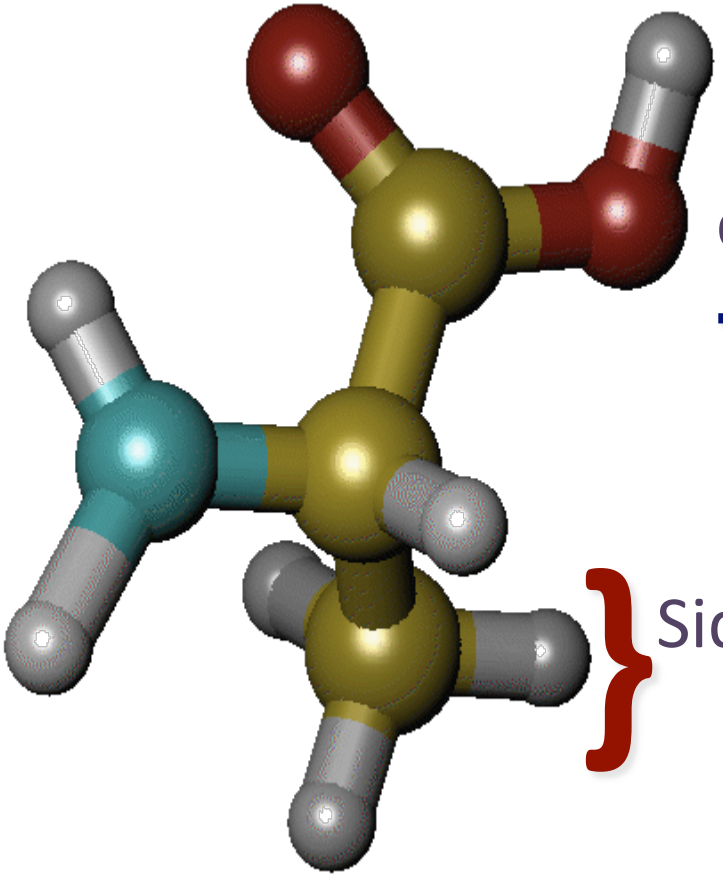
Principles of Protein Structure and Conformational Space

Principles of Protein structure

1. Introduction
    1. Amino acids
    2. Peptide bond
    3. Native conformation
    4. Levels of structure (1,2,3)
2. Energy
    1. Bonding energy terms
    2. Non-bonding terms
3. Secondary Structure
    1. Strands
    2. Helix
    3. Ramachandran Plot
    4. Aa propensities

1. Introduction

1. Amino Acids



amino
$-NH_2$

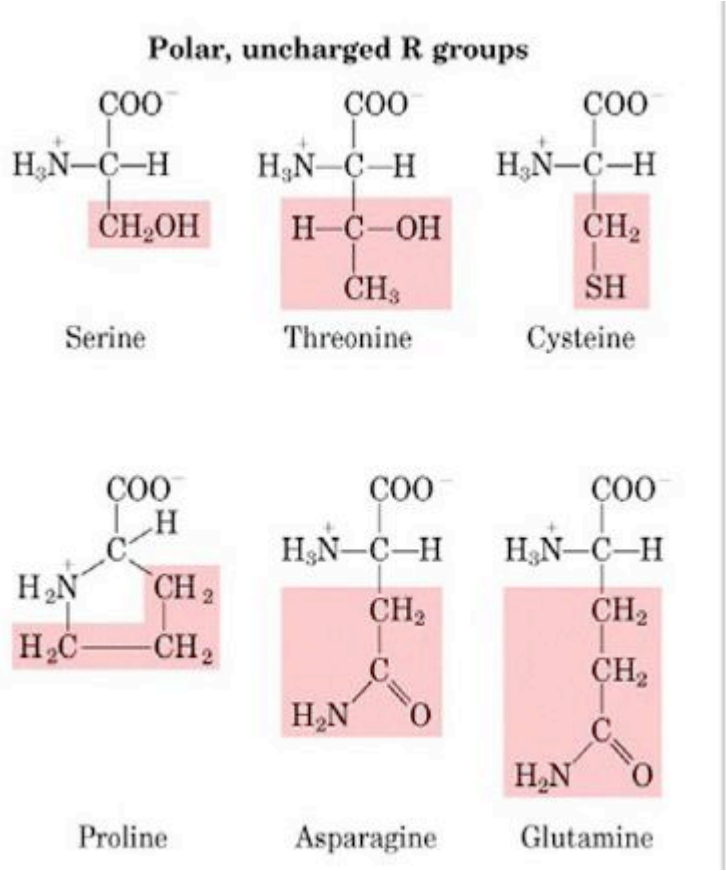
carboxyl
-COOH

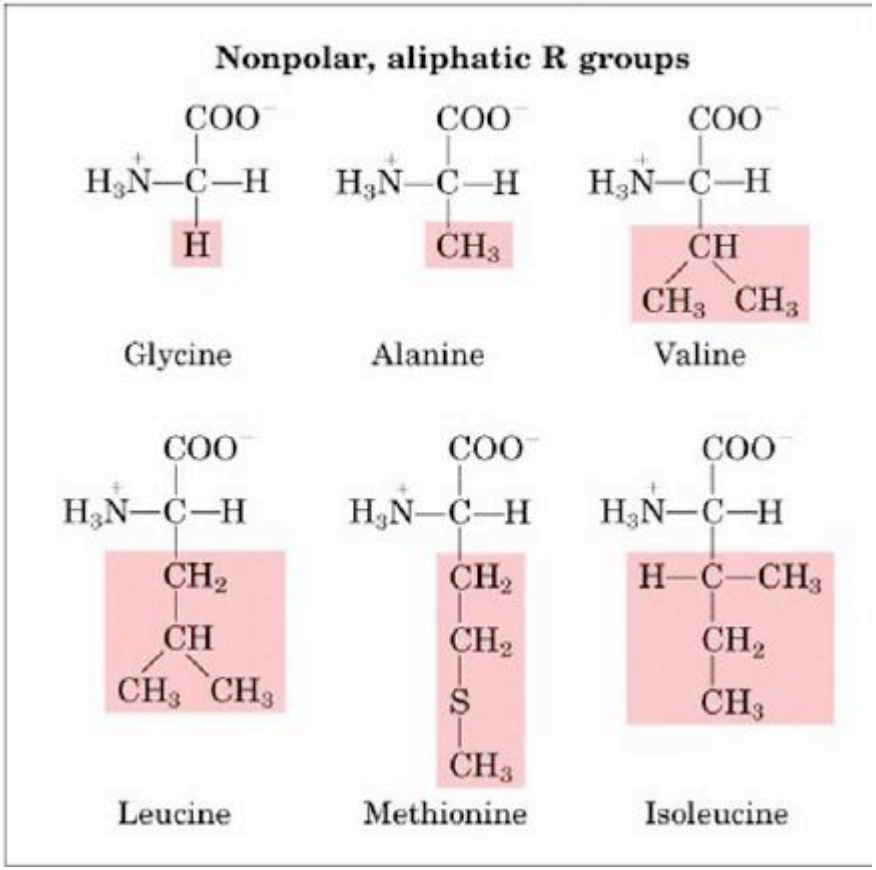
Side-chain **($-CH_3$)**

Estereoisomer L

1. Introduction

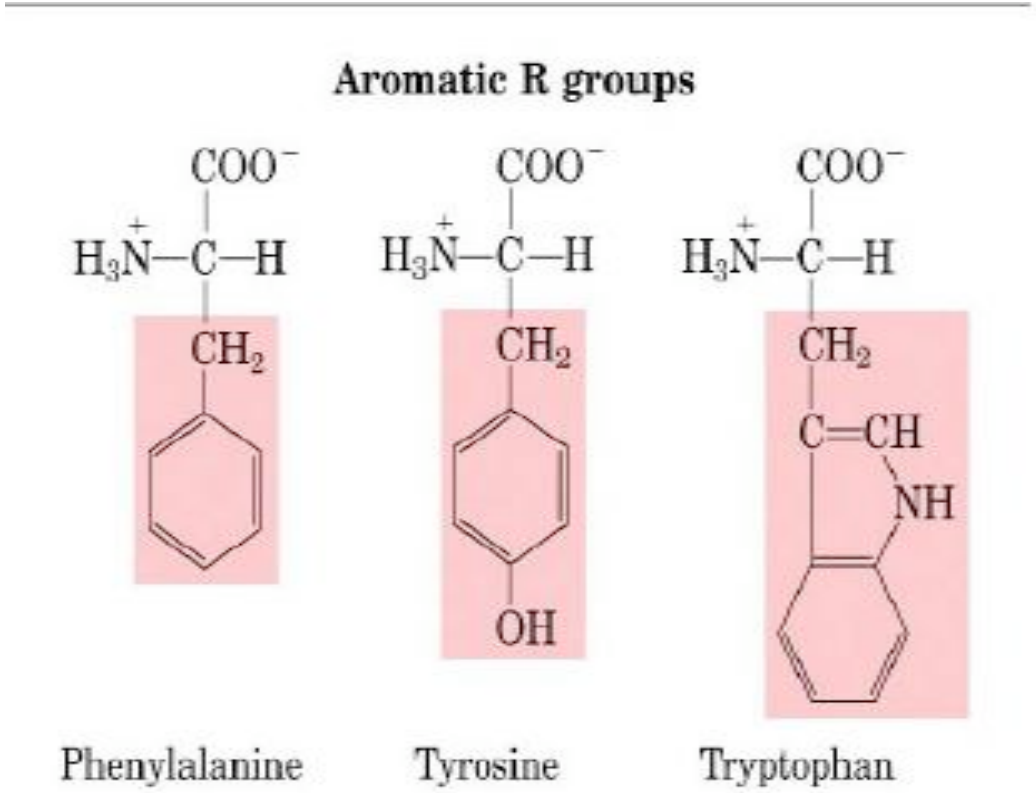
1. Amino Acids





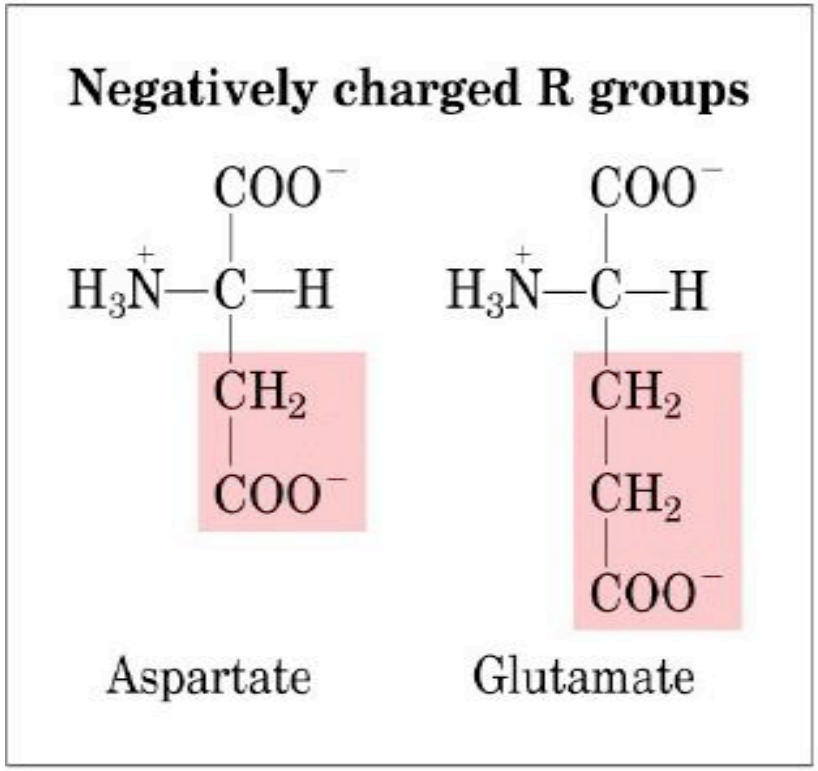
1. Introduction

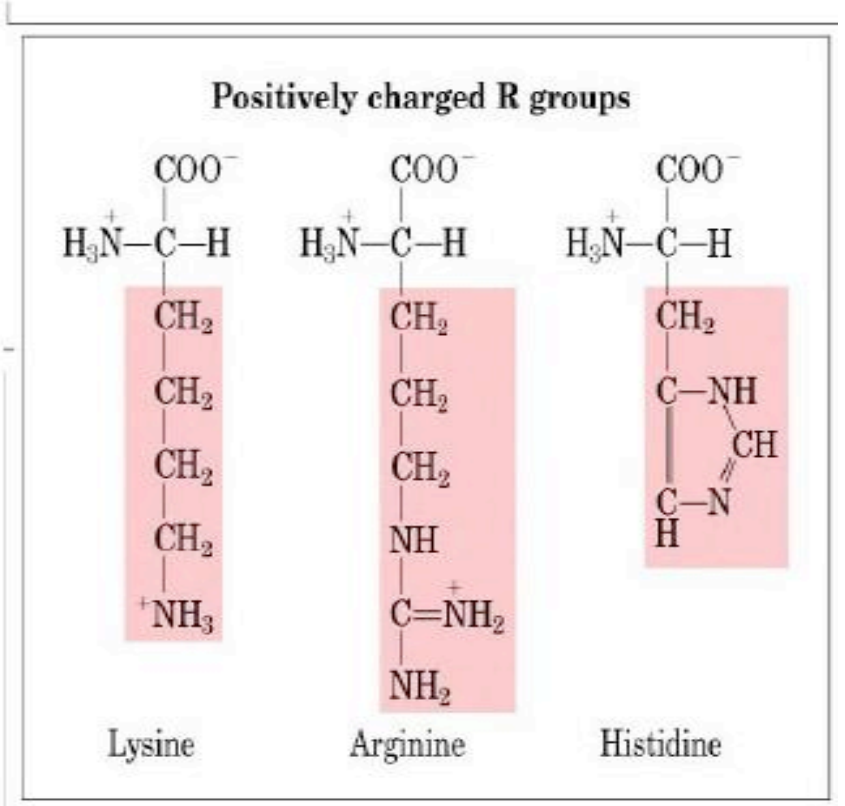
1. Amino Acids



1. Introduction

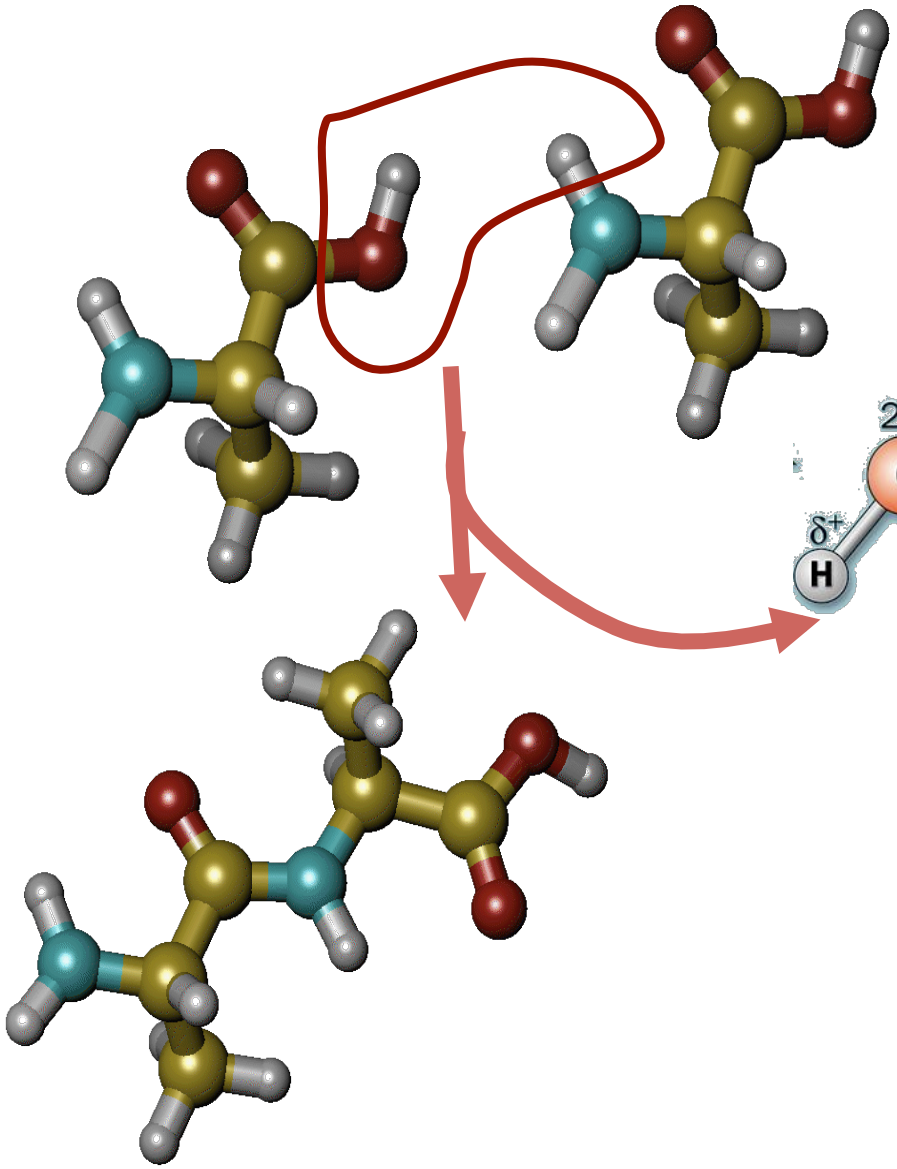
1. Amino Acids

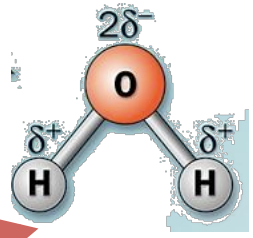




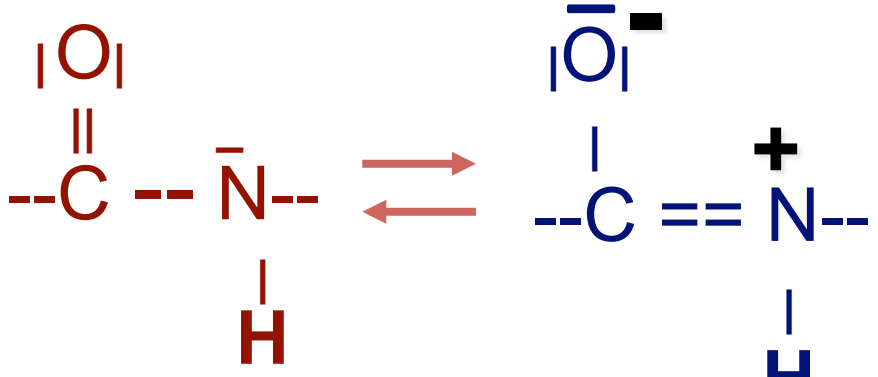
1. Introduction

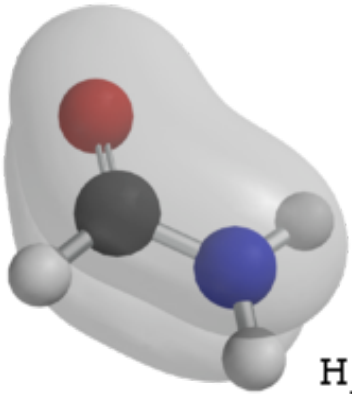
2. Peptide Bond

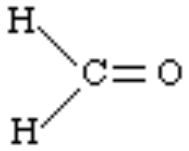


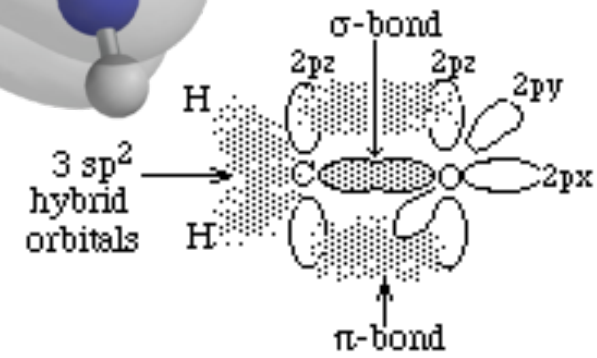


PLANE -CO-NH-

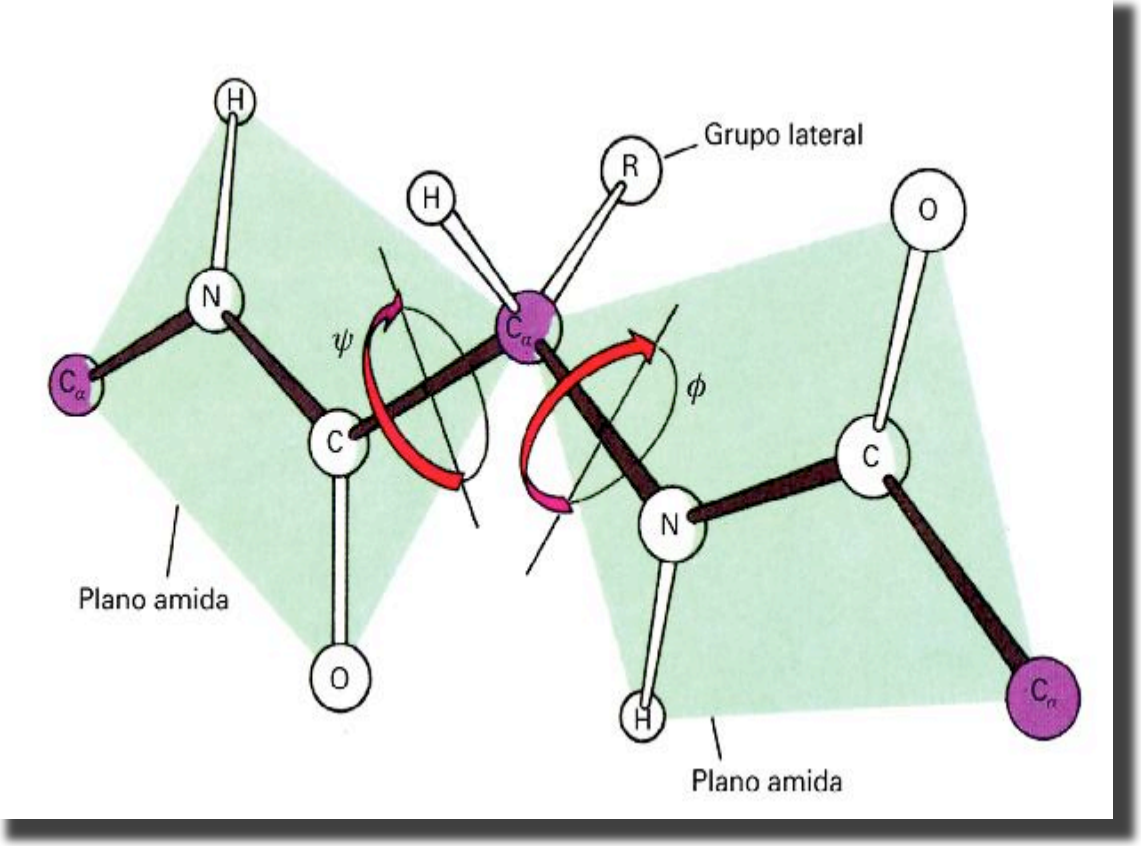








Angles ϕ & ψ



ψ = [-180º, +180º]

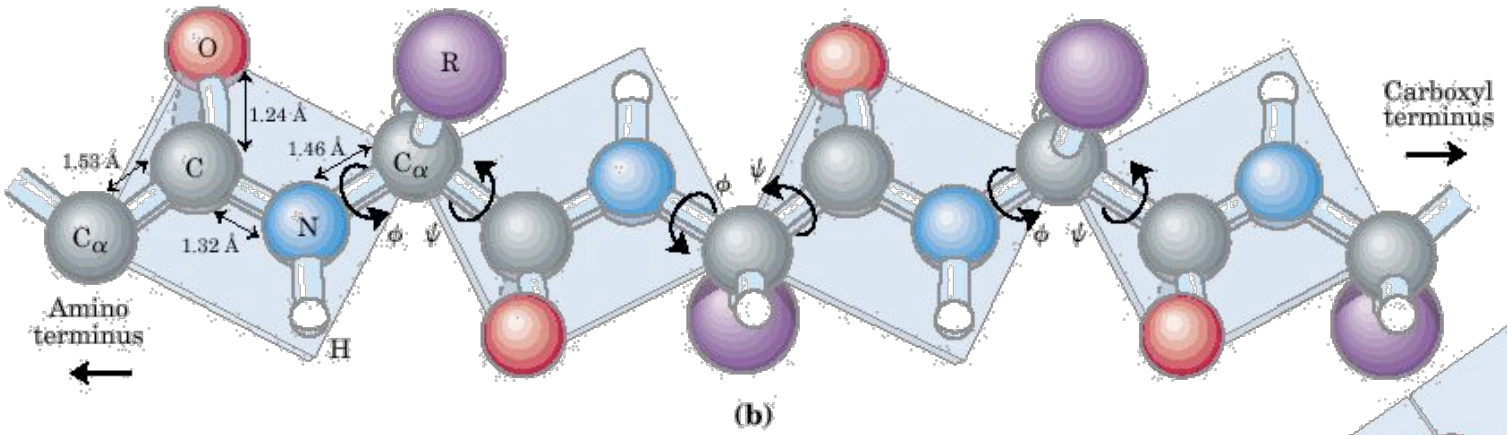
ϕ = [-180º, +180º]

1. Introduction

3. Native Conformation

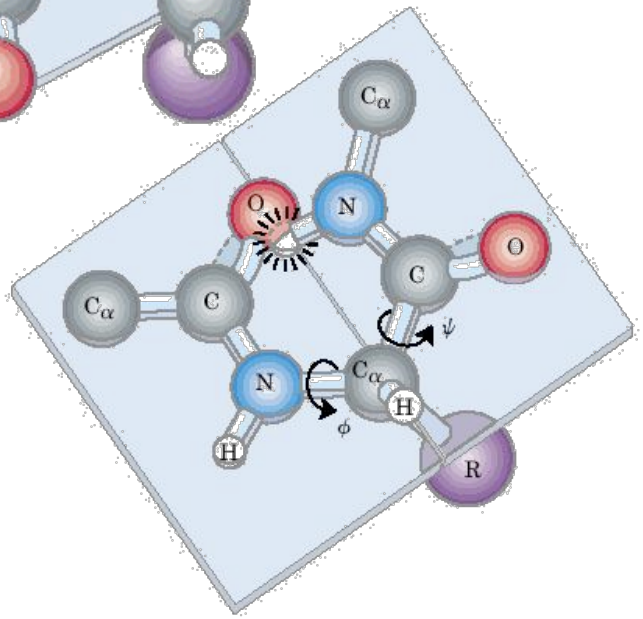
Native conformation:

1) Polipeptide + environment
2) Function & stability



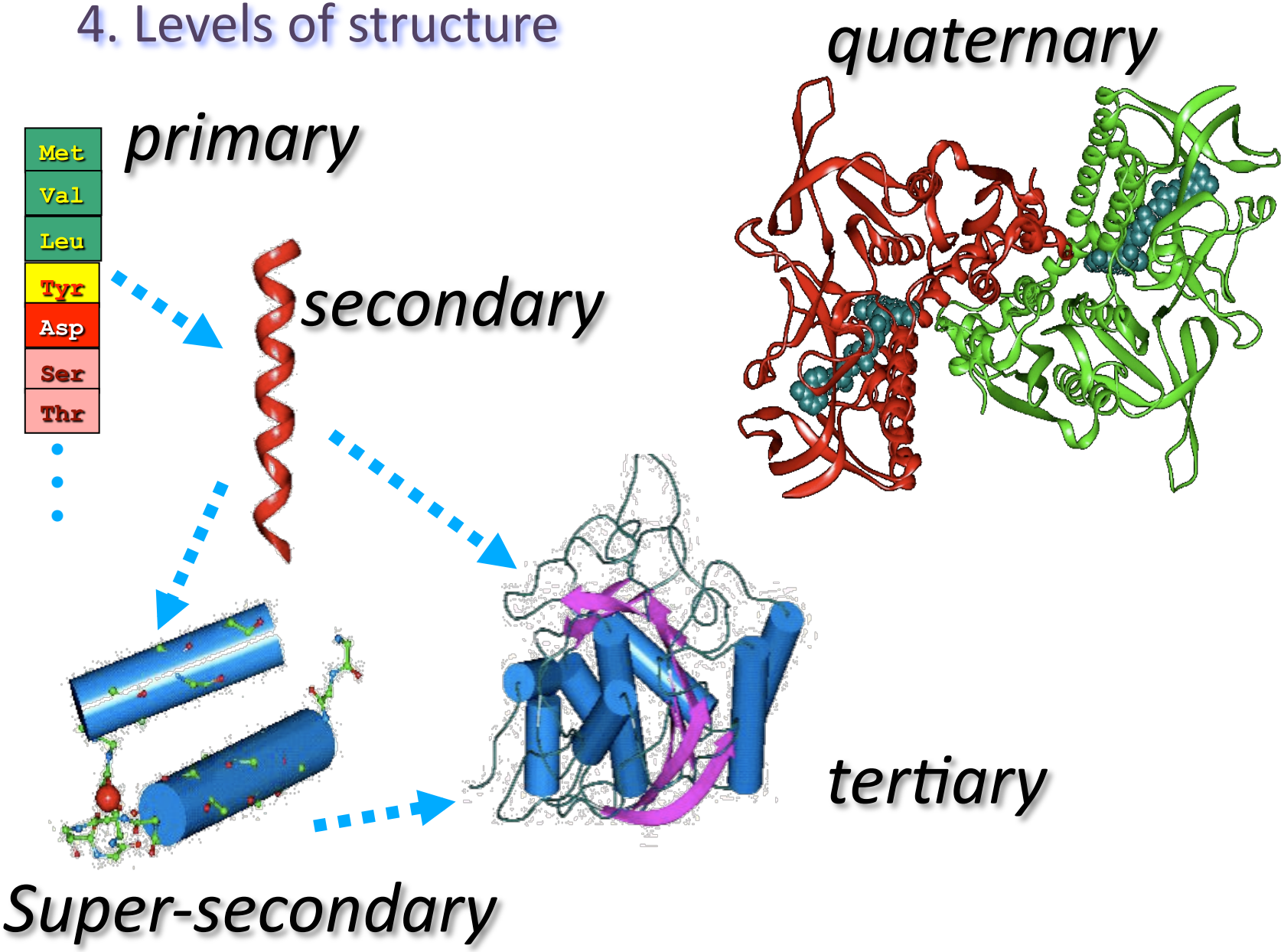
Conformational restraints:

1) Peptide bond planarity
2) The total space is not 360 x 360



1. Introduction

4. Levels of structure



2. Energy

1. Bonding energy terms

$$\vec{F} = m * \vec{a} = \sum_{i=1}^{N} \vec{F}_i$$

$$W = \int_{path} \vec{F} * d\vec{r}$$

2. Energy

1. Bonding energy terms

$$\vec{a} = \frac{\partial}{\partial t}\left(\frac{\partial \vec{r}}{\partial t}\right) = \frac{\partial \vec{v}}{\partial t}$$

$$d\vec{r} = \vec{v} * dt$$

$$\frac{\partial\left(v^2\right)}{\partial t} = \frac{\partial(\vec{v} * \vec{v})}{\partial t} = 2\left(\vec{v} * \frac{\partial \vec{v}}{\partial t}\right)$$

$$\vec{F} * d\vec{r} = \left(m * \frac{\partial \vec{v}}{\partial t}\right) * \left(\vec{v} * dt\right) = \frac{1}{2} m \frac{\partial\left(v^2\right)}{\partial t} dt$$

2. Energy

1. Bonding energy terms

$$W = \int_{Path} \vec{F} * d\vec{r} = \int_{Time} \frac{1}{2} m \frac{\partial\left(v^2\right)}{\partial t} dt = \frac{1}{2} m v^2 + C$$

$$Potential = E_p = -\int_{Path} \vec{F} * d\vec{r}$$

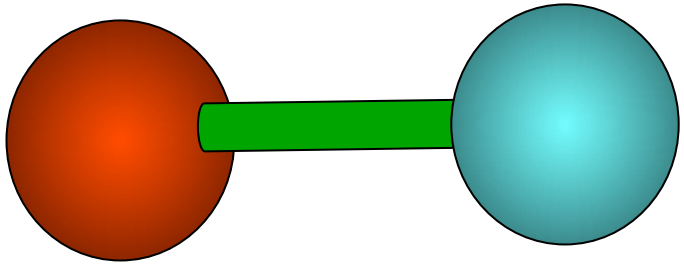
$$Kinetic = E_k = \frac{1}{2} m v^2$$

$$Total = U = E_p + E_k$$

2. Energy

1. Bonding energy terms

Bond



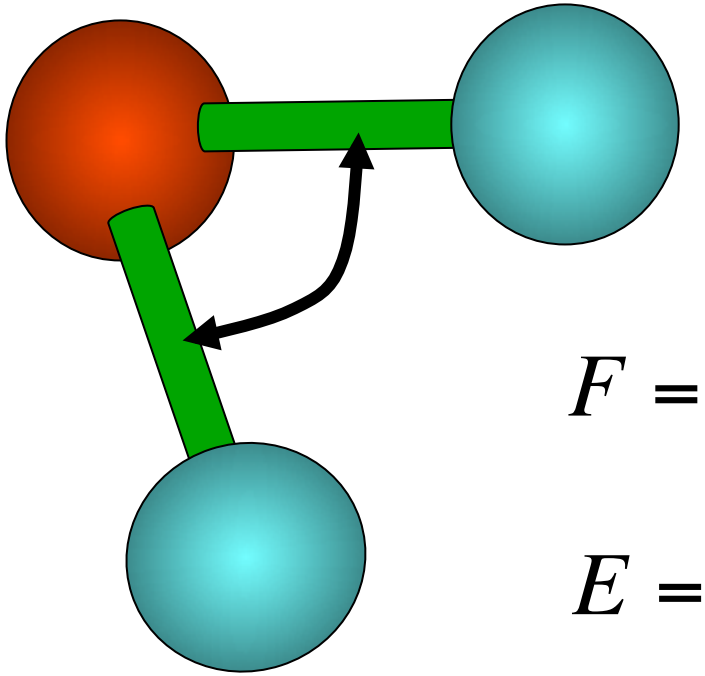
$$F = -k(d - d_0)$$

$$E = \frac{1}{2}k(d - d_0)^2$$

2. Energy

1. Bonding energy terms

Angle



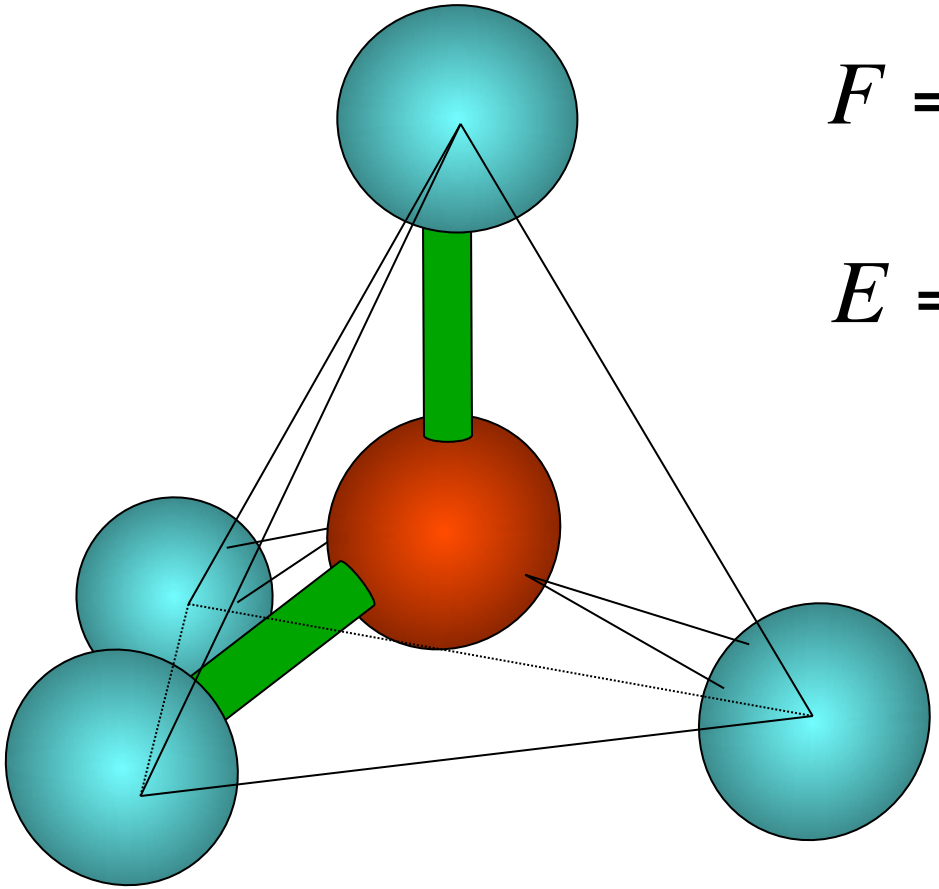
$$F = -k(\alpha - \alpha_0)$$

$$E = \frac{1}{2}k(\alpha - \alpha_0)^2$$

2. Energy

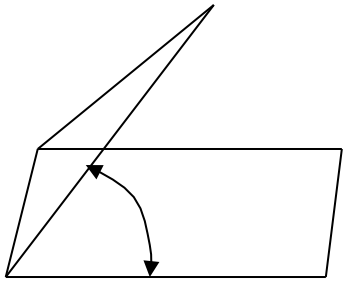
1. Bonding energy terms

Improper dihedral



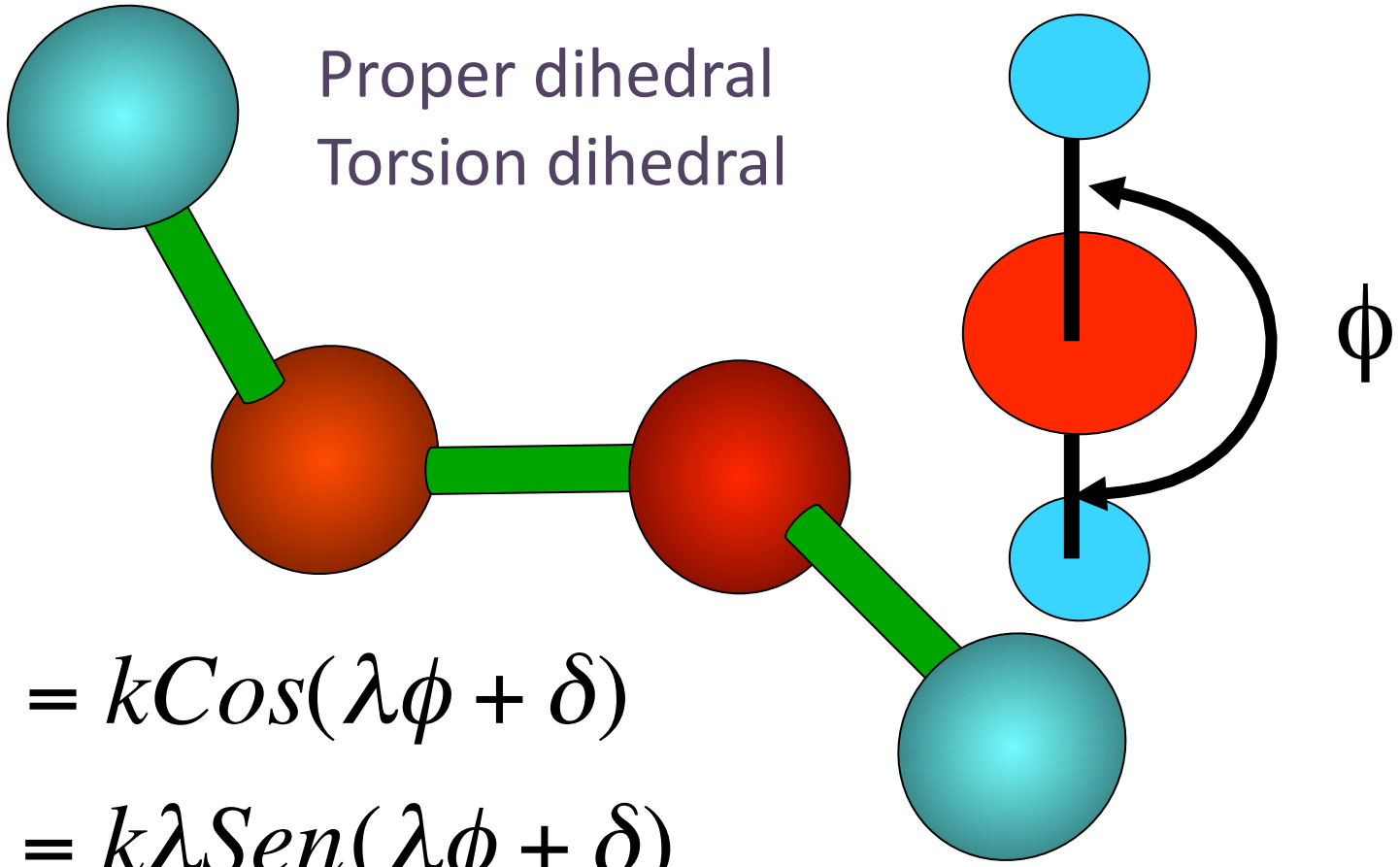
$$F = -k\left(\varpi - \varpi_0\right)$$

$$E = \frac{1}{2}k\left(\varpi - \varpi_0\right)^2$$



2. Energy

1. Bonding energy terms



$$E = kCos(\lambda\phi + \delta)$$

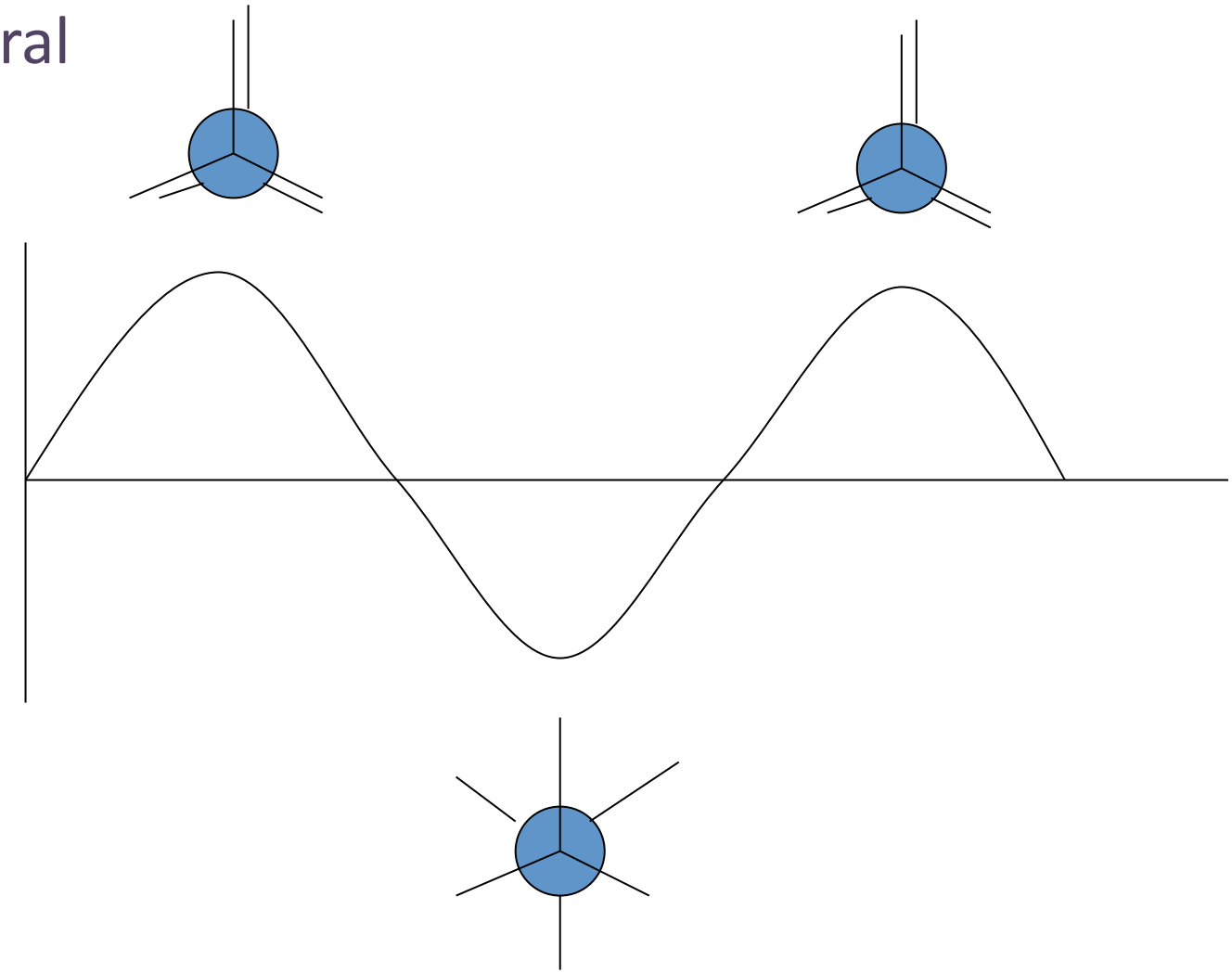
$$F = k\lambda Sen(\lambda\phi + \delta)$$

2. Energy

1. Bonding energy terms

Proper dihedral

Torsion dihedral



2. Energy

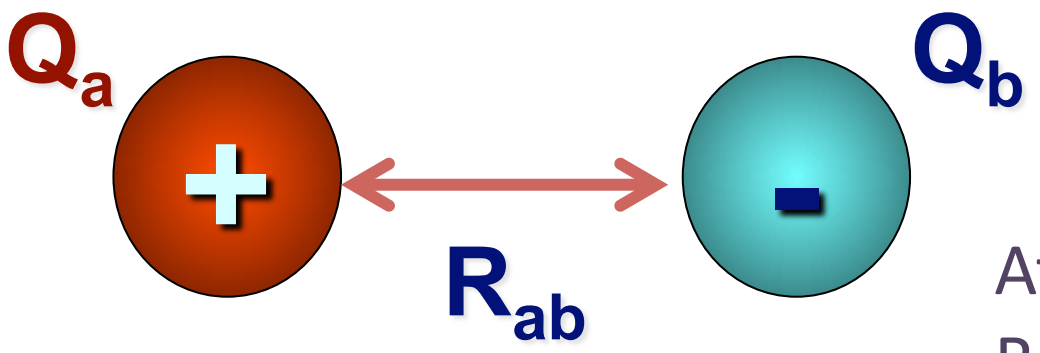
2. Non-Bonding terms

Electrostatic energy

Long range interactions
Involve in folding

$$\vec{F} = -\frac{1}{4\pi\varepsilon_0}\frac{Q_a Q_b}{r_{ab}^2}\left(\frac{\vec{r}_{ab}}{r_{ab}}\right)$$

$$E = \frac{1}{4\pi\varepsilon_0}\frac{Q_a Q_b}{r_{ab}}$$



Atraction: opposite sign
Repulsion: same sign
Approx. 150 kcal/mol

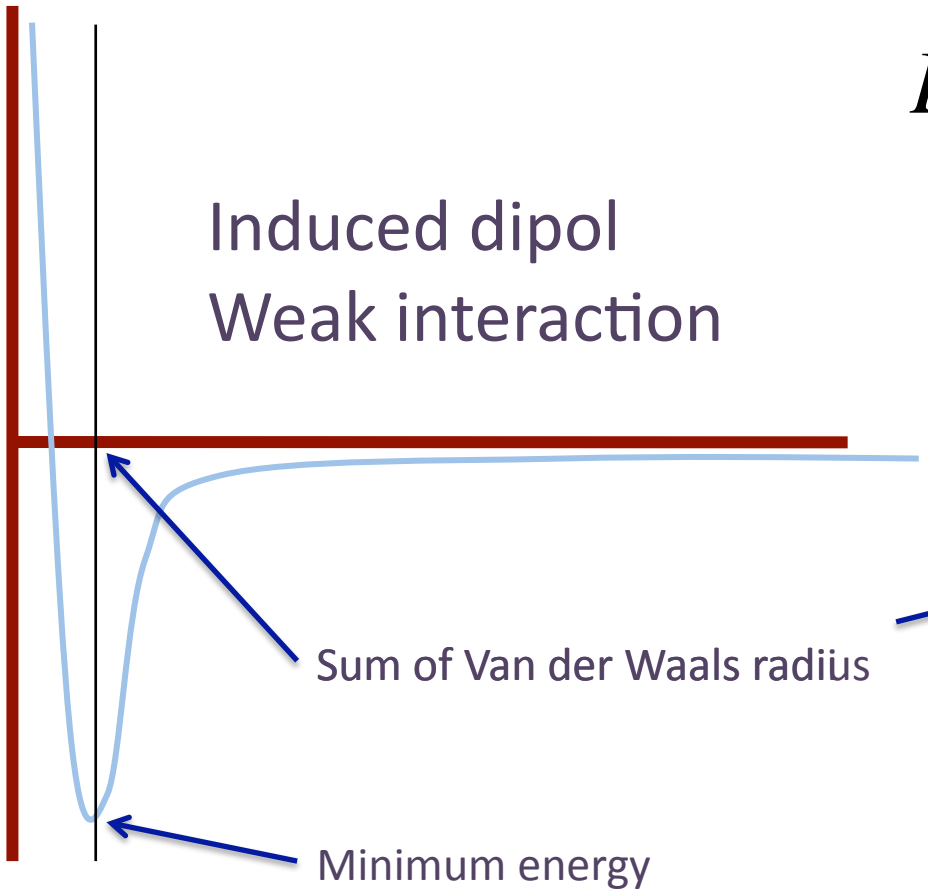
2. Energy

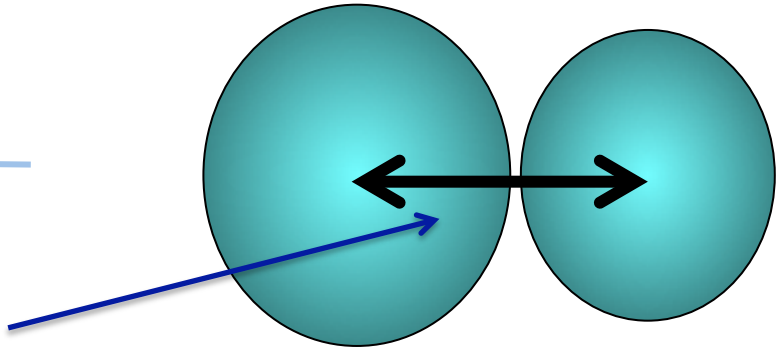
2. Non-Bonding terms

Van der Waals energy

Lennard-Jones equation

$$E = \frac{C_6(a,b)}{r_{ab}^6} - \frac{C_{12}(a,b)}{r_{ab}^{12}}$$



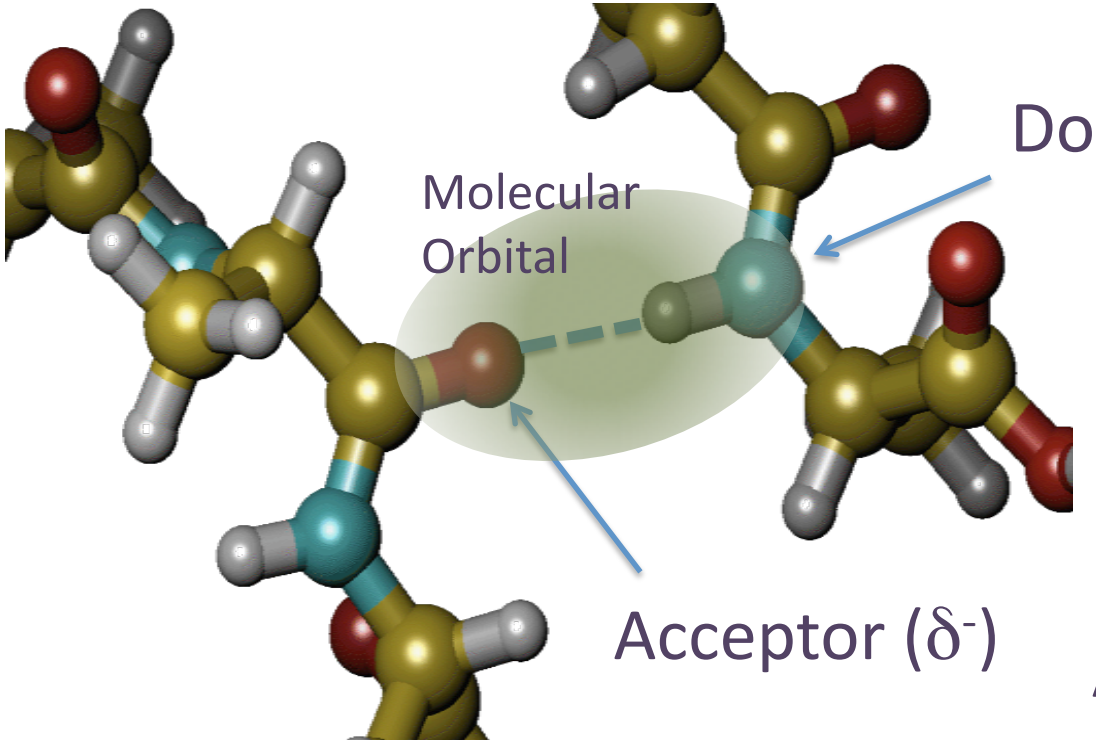


Avoids atoms collision

Approx. 2 kcal/mol

2. Energy

3. Hydrogen Bonds



Donor (H^+)

Acceptor (δ^-)

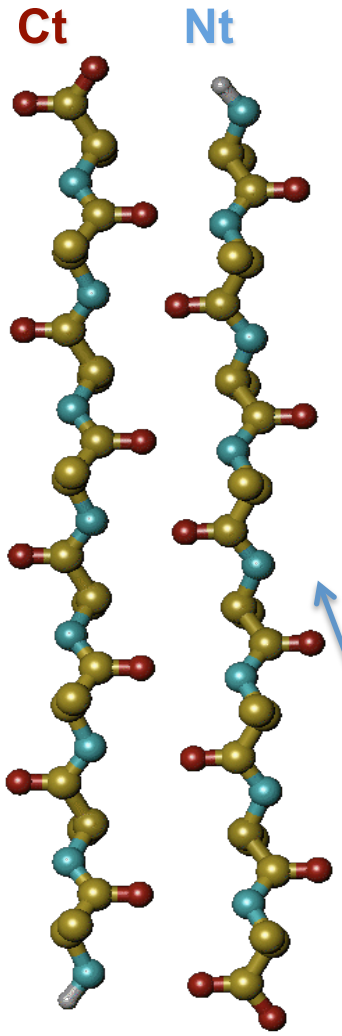
85% ionic bond
15% covalent
Approx. 4 kcal/mol

Restriction:
Angle OHN = 180^0+/-35^0
Distance ON = 2.8A

Inter-chain (long distance/short distance in sequence)
Intra-chain

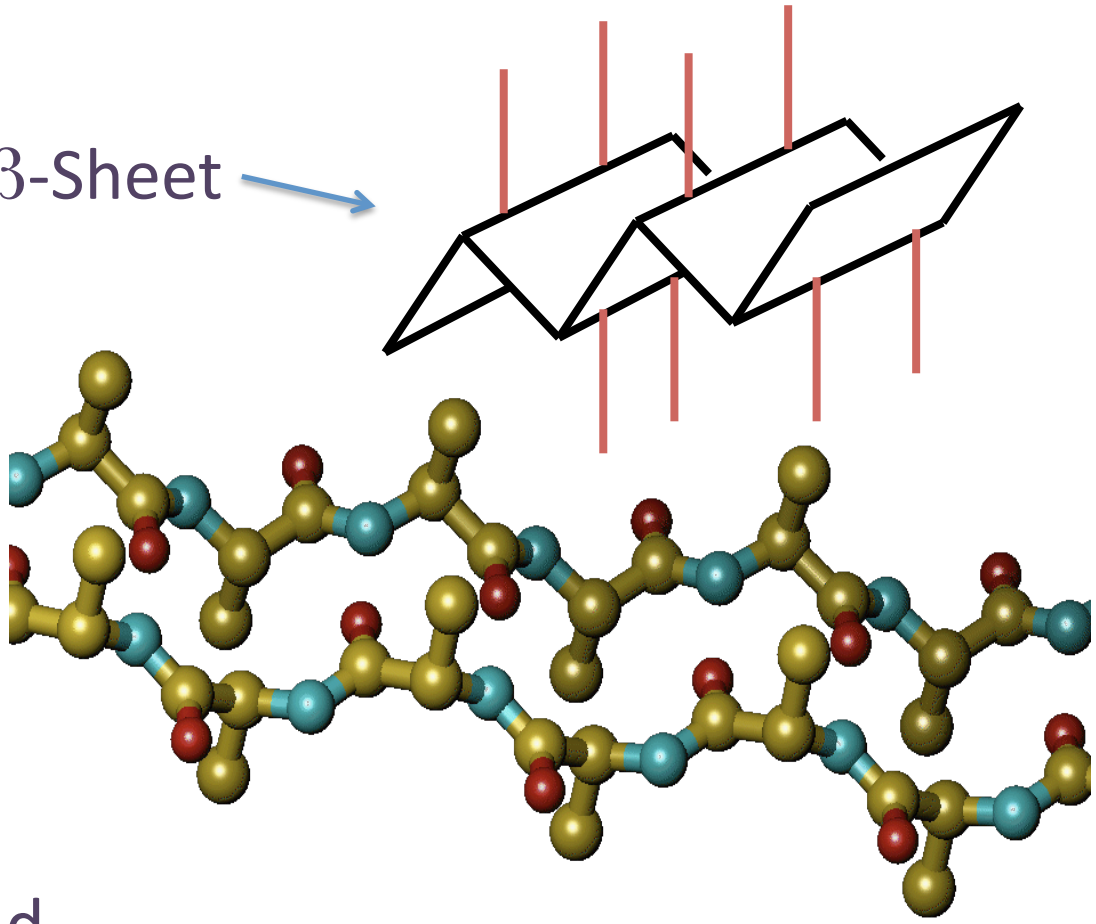
3. Secondary Structure

1. Strands



β-ladder

β-Sheet



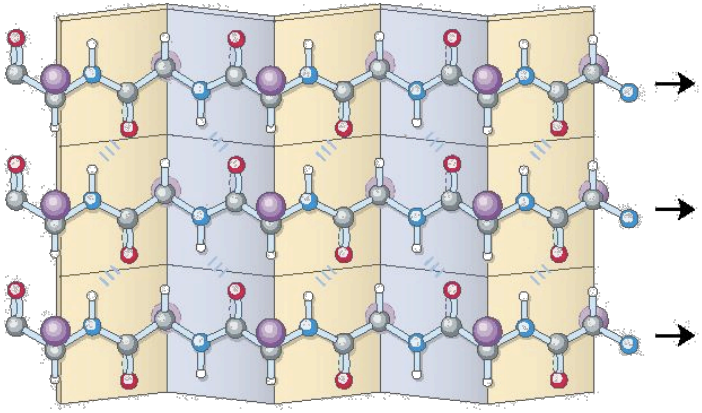
β-strand

- Hydrogen Bonds (HB) are long distance in sequence
- Side-chains are at both sides

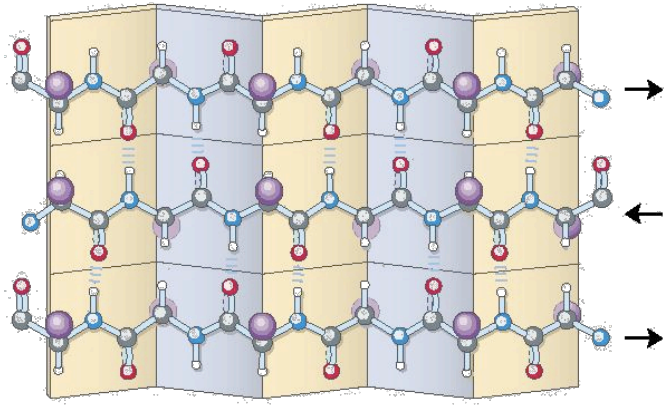
3. Secondary Structure

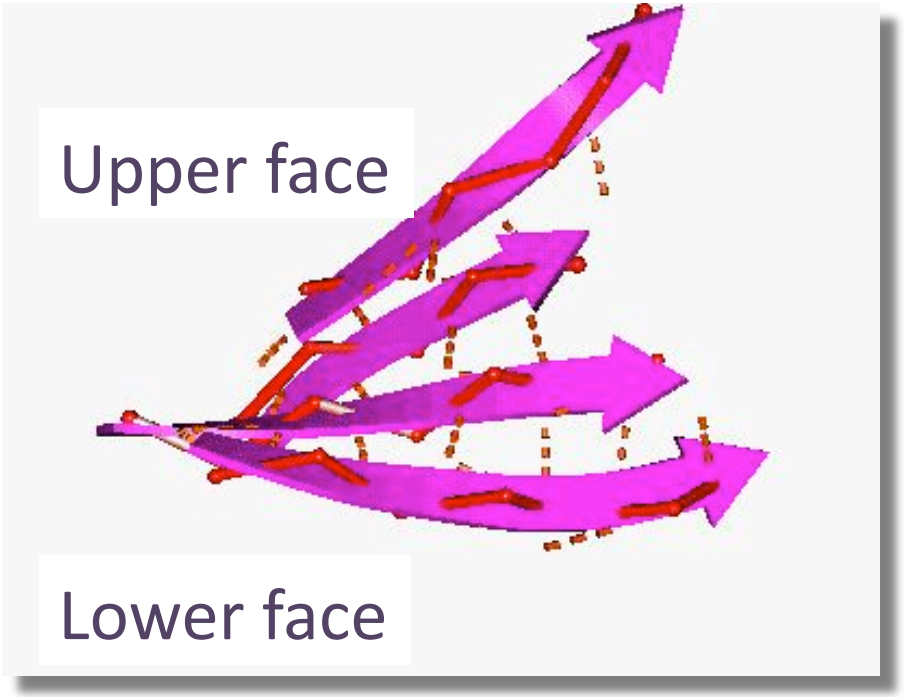
1. Strands

Parallel β−sheet



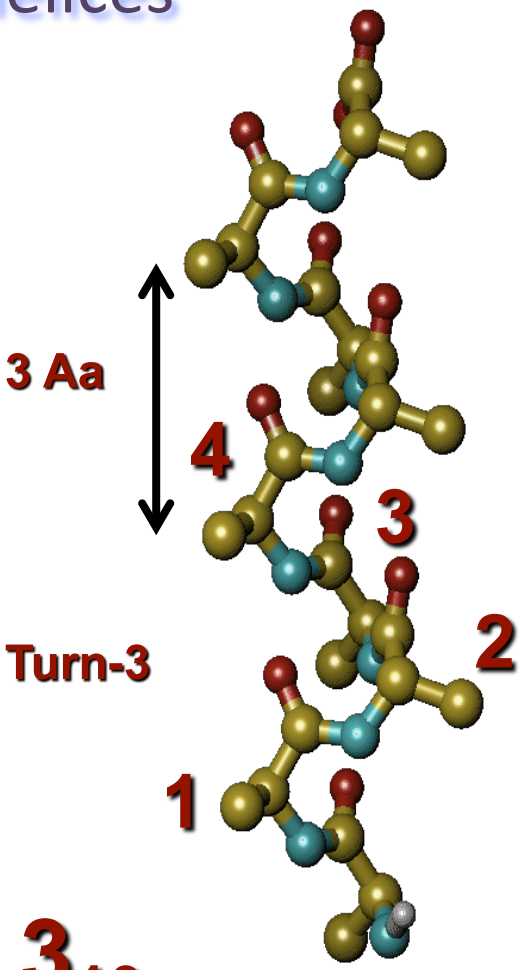
Anti-parallel β−sheet



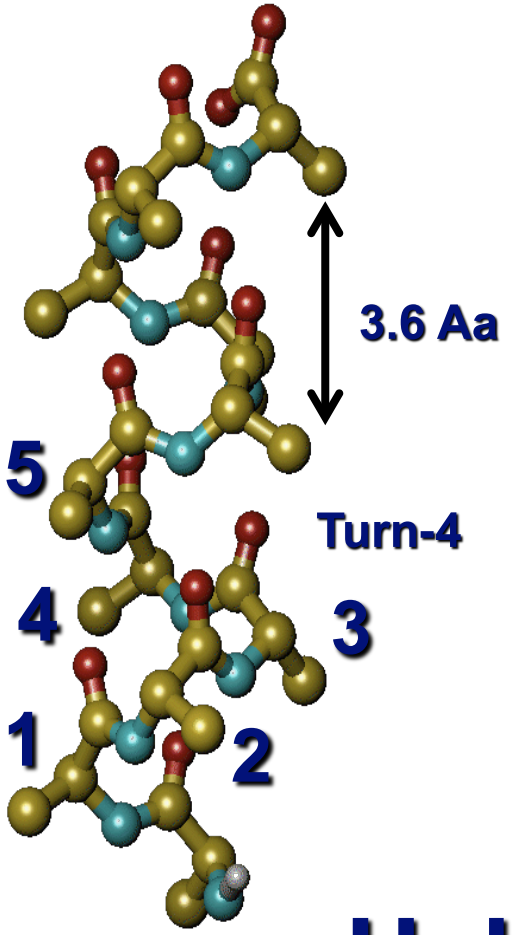


3. Secondary Structure

2. Helices



Helix 3_{10}

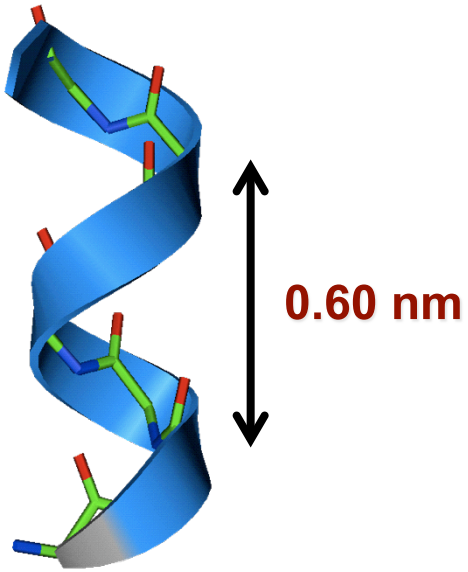


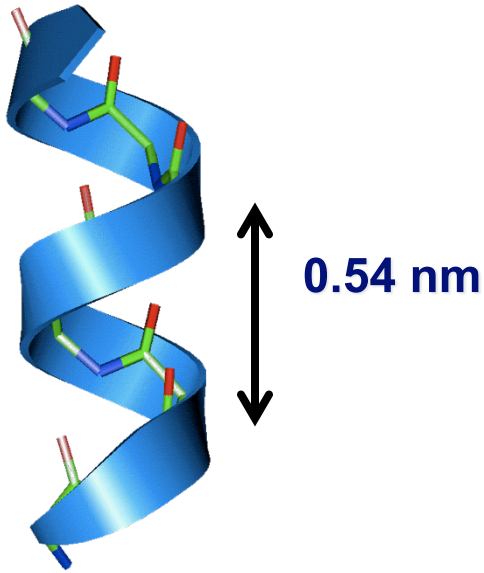
Helix α

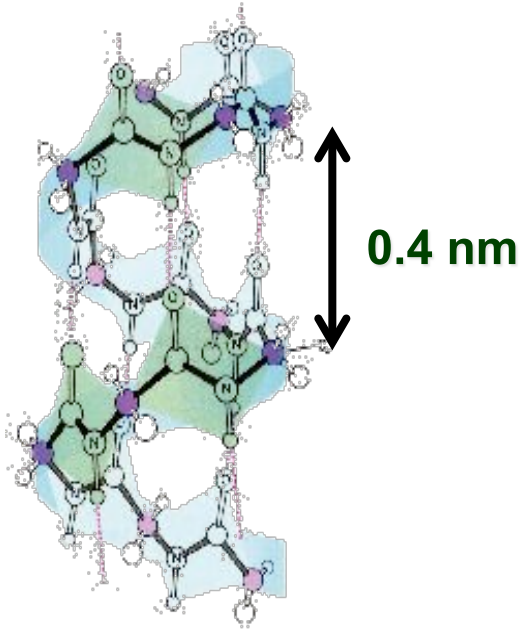
- HB are short distance in sequence
- Side-chains protrude out of the helix

3. Secondary Structure

2. Helices







Helix 3_{10}

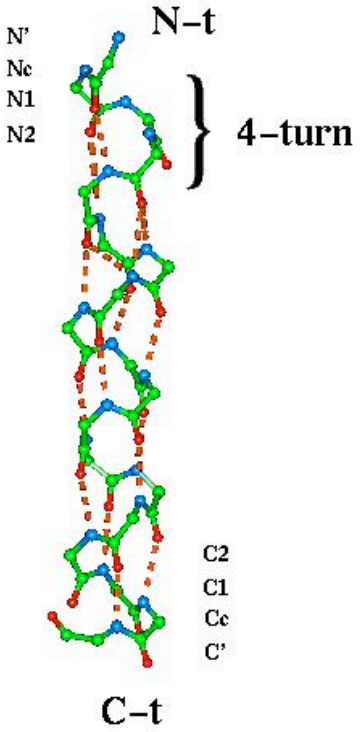
Helix α

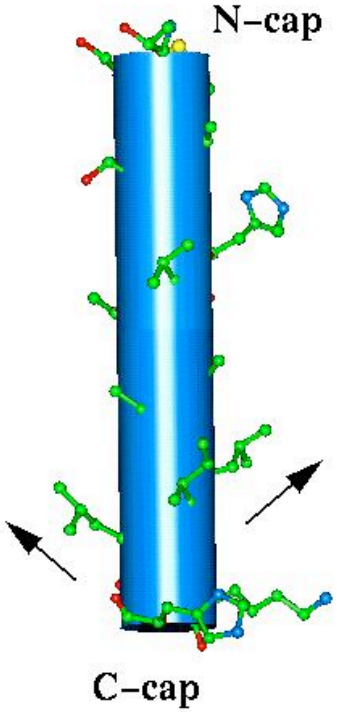
Helix π

UNSTABLE

3. Secondary Structure

2. Helices

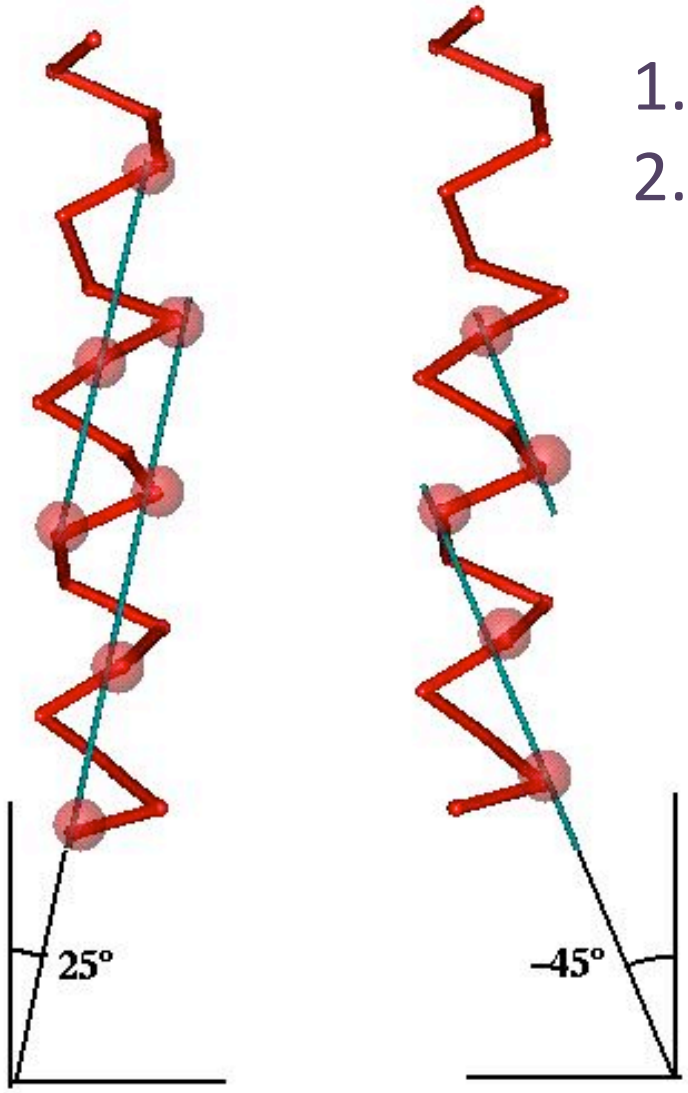




	Hbond	Hidrofobic
N-cap box	N_c-N_3	N'-N_4
Big box	N_c-N_3	N''-N_4
Shellman	C''-C_3 C'-C_2	C''-C_3
α_L	C'-C_3	C''-C_3

3. Secondary Structure

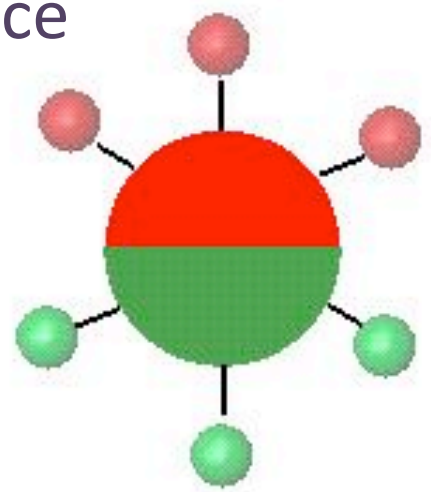
2. Helices



Side-chains location in the helix:

1. Groove formation as in a screw
2. Amphipathic helix: two faces with different solvation properties

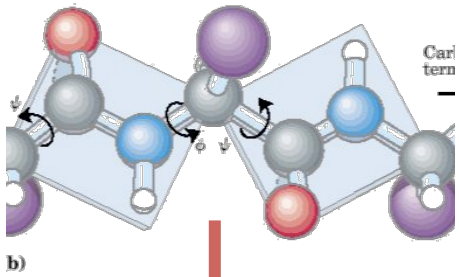
Polar face



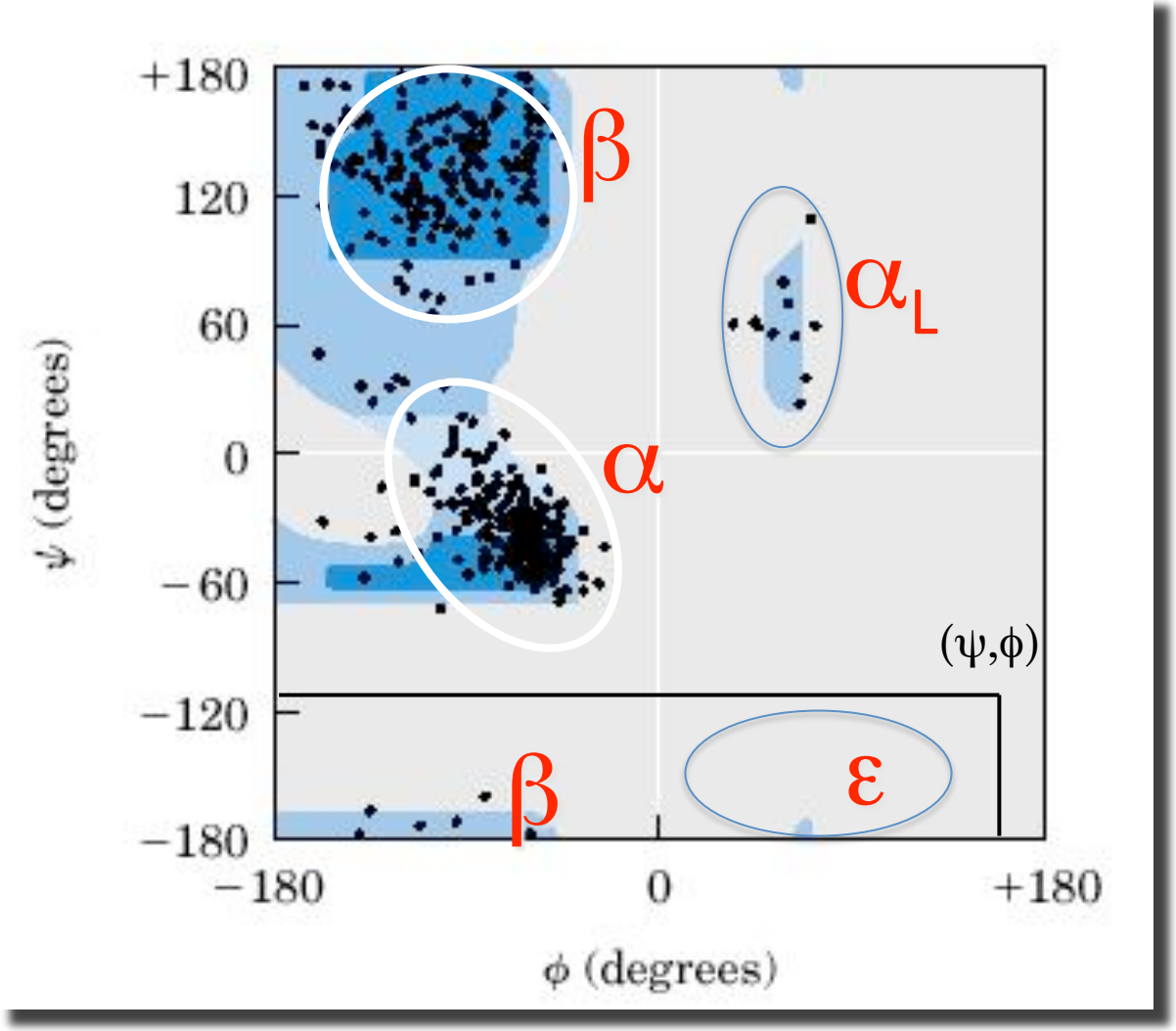
Non-polar face (hydrophobic)

3. Secondary Structure

3. Ramachandran Plot

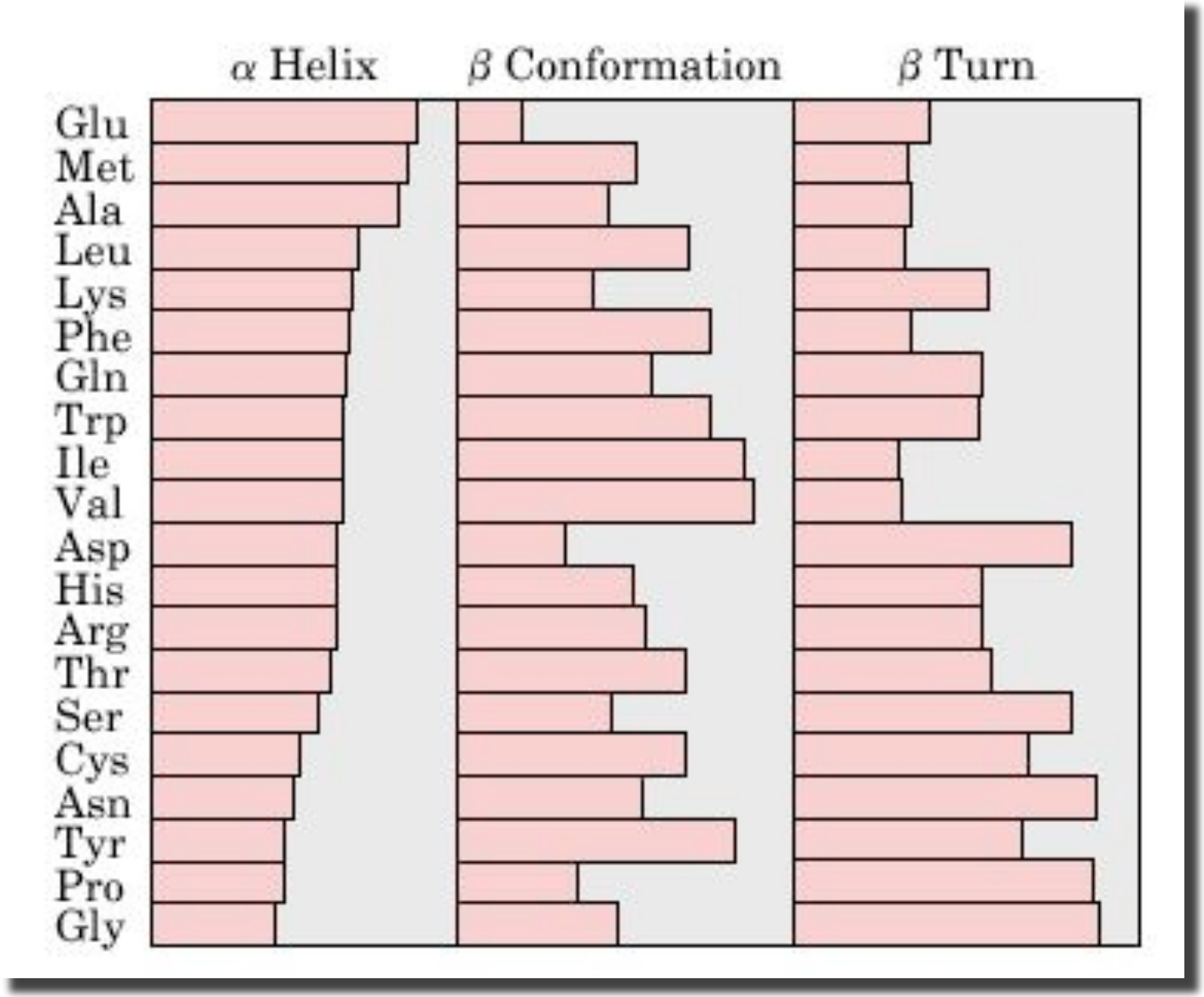


(ψ,ϕ)



3. Secondary Structure

4. Aa propensity



Protein structure I

1. Escleroproteins
    1. α-keratin
    2. Fibroine
    3. Collagen
2. Principles of thermodynamics
    1. State functions
    2. Statistical thermodynamics
    3. Entropy and Boltzman equation
3. Soluble and globular proteins
    1. Protein folding
        1. Hydrophobic effect
        2. Disulphide bridges
    2. Protein Folds

Protein structure II

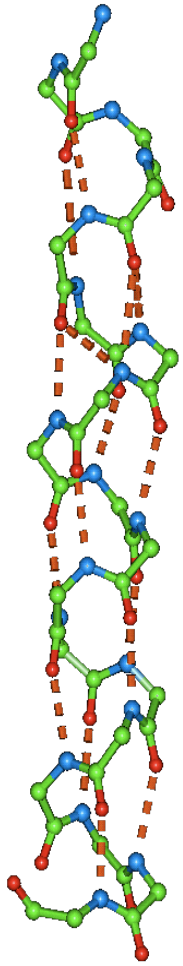
3. Soluble and globular proteins
    3. Supersecondary structure
        1. Definition
        2. α−α
        3. β−β
        4. β−α & α−β
        5. β−α−β
    4. Domains
        1. Definition
        2. Classification
        3. Function-Sequence-Structure
        4. SCOP
        5. CATH

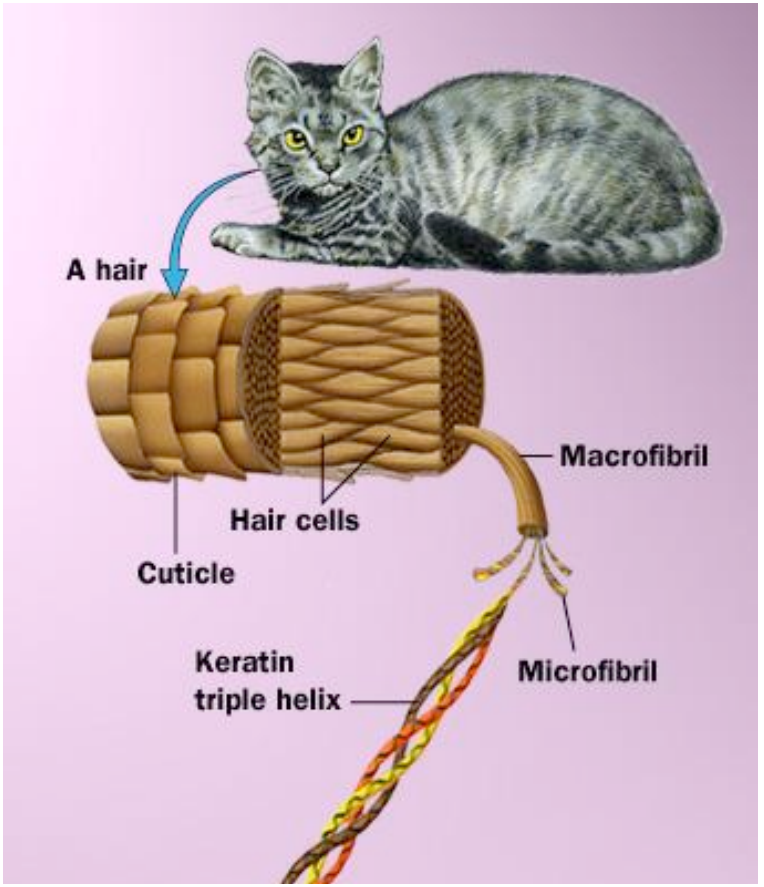
1. Escleroproteins

1. α-keratin

Fiber formed by α-helices. Found in nails and hair

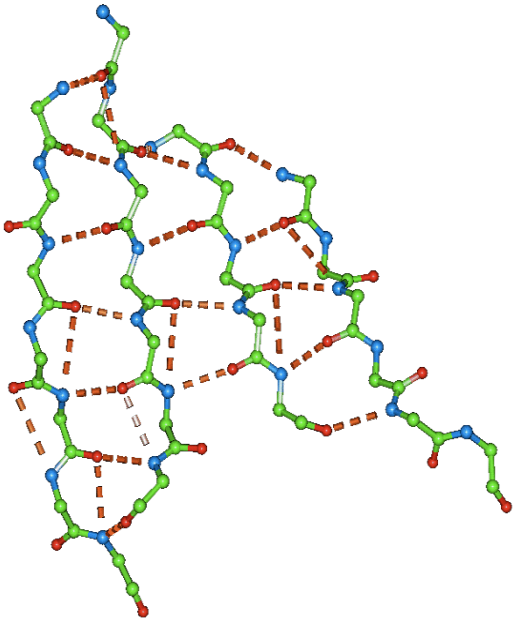
Protein structures from fibers can be solved by X-ray, because they produce repetitive patterns and these produce diffraction.



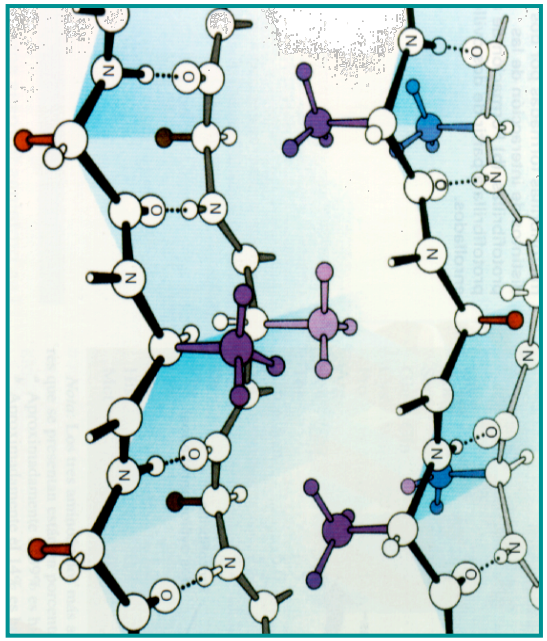


1. Escleroproteins
2. Fibroine

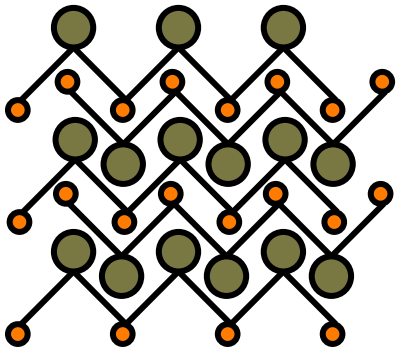
Fiber found in spider web, silk and elastin.
Highly resistant. Formed by β-sheets



sheet β



Sheet 1 Sheet 2



Fiber produced by packing between sheets.

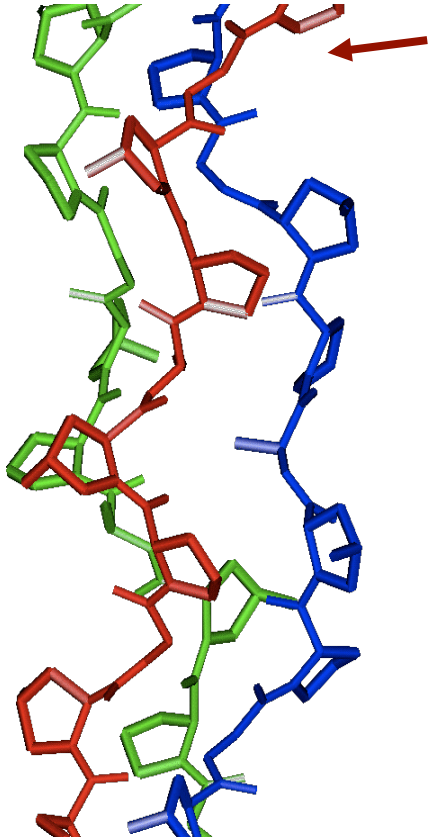
Packing due to Aa composition:

Gly-Ala-Gly-Ala-Gly-Ser-Gly-Ala-Ala-Gly-(Ser-Gly-Ala-Gly-Ala-Gly)$_8$

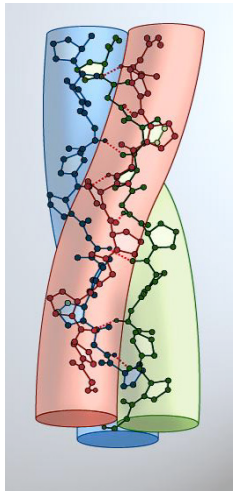
1. Escleroproteins

3. Collagen

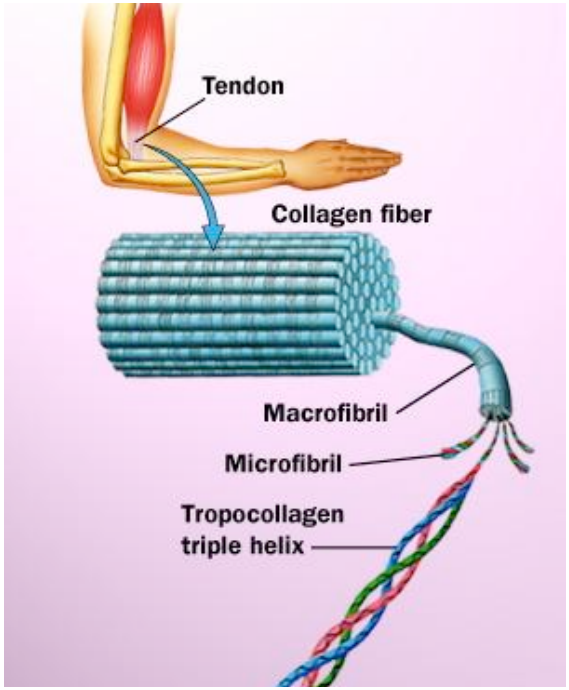
Collagen is a fiber formed by grouping macrofibrils and microfibrils of tropocollagen. Collagen is the main component of the connective tissue



Procollagen



Tropocollagen



$\{ X\text{-}Pro\text{-}Gly \}_n$

X is often Lys. Lys is modified by hydroxylation and further oxydation into alfdehyde. Cross-linking between oxydated Lys produces links of tropocollagens

2. Principles of thermodynamics

1. State functions

The state of a system is described by means of macroscopic variables: N, p, T, V, U (total energy).

These variables are known as thermodynamic variables. If the variables are independent, then they are known as state variables.

Any function described by means of thermodynamic variables are known as state functions.

The difference of a state function between two states is independent of the path driving one state to the next.

2. Principles of thermodynamics

1. State functions

Second law of thermodynamics:

There is a state function such that it gets always the maximum of the system in the absence of restraints. This function is named entropy. It follows that the entropy of an isolated macroscopic system never decreases. The second law states that spontaneous natural processes increase entropy overall

A system is in thermodynamic equilibrium when the macroscopic variables describing the system are independent of time.

Examples of state functions:

$H = U + pV$

$G = H - TS$

$F = U - TS$

2. Principles of thermodynamics

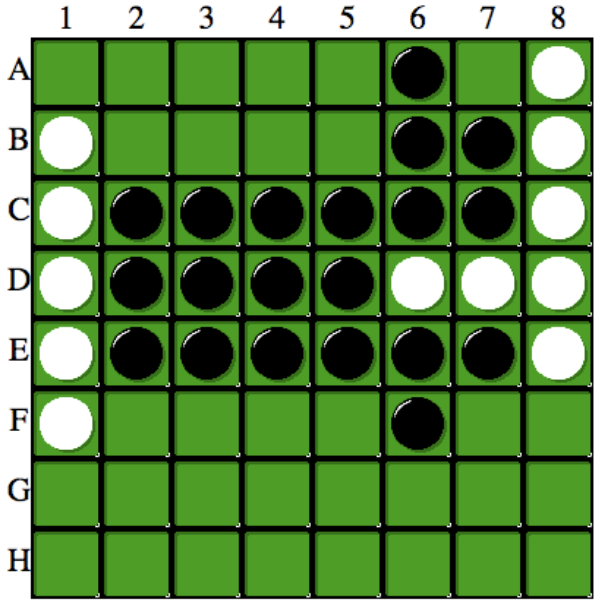
2. Statistical thermodynamics

Microstate: State of a system specified by means of the coordinates and velocities of each particle.
Macrostate: State of a system specified by means of thermodynamic variables
Hypothesis: all microstates of a system are equally accessible with the same probability

The thermodynamic probability (W) is defined as the total number of microstates accessible by the particles of an ensemble in the same macrostate

Boltzmann hypothesis: Entropy (S) can be defined as a function of the thermodynamic probability (W), this is
$S = f(W)$

2. Principles of thermodynamics
2. Statistical thermodynamics



Othello Game: White and black discs filling the board

Microstate: specify the disc color in each cross of the board.
Macrostate: Color of the board
Hypothesis: all microstates of a system are equally accessible with the same probability

Macrostate with maximum entropy is grey.

2. Principles of thermodynamics

3. Entropy and Boltzman equation

Theorem:

$$S_b - S_a = k \log (W_b/W_a)$$

Demonstration:

Let be two systems “a” and “b”. When joining both we have a new system “c” which is the sum of “a” and “b”. The total number of states in “a” and “b” are W_a and W_b, the entropy of “a” and “b” are S_a and S_b. Then:

$S = S_c = S_a + S_b$ and $W = W_c = W_a * W_b$ with $S = f(W) = f(W_a * W_b)$

$$\frac{\partial f(W)}{\partial W_a} = \frac{\partial f(W)}{\partial W} * W_b$$

$$\frac{\partial f(W)}{\partial W_b} = \frac{\partial f(W)}{\partial W} * W_a$$

$$\frac{1}{Wa}\frac{\partial f(W)}{\partial Wb} = \frac{1}{Wb}\frac{\partial f(W)}{\partial Wa}$$

$$Wb\frac{\partial f(W)}{\partial Wb} = Wa\frac{\partial f(W)}{\partial Wa} = k$$

$$k = W\frac{\partial f(W)}{\partial W}$$

$$\int k\frac{1}{W}\mathrm{dW} = \int \frac{\partial f(W)}{\partial W}\mathrm{dW} = S_b - S_a$$

$$S_b - S_a = k\ln\left(\frac{W_b}{W_a}\right)$$

2. Principles of thermodynamics

3. Entropy and Boltzman equation

Theorem

The probability of a microstate with energy E_i is

$$P_i = \frac{e^{-\beta E_i}}{Z} \qquad \text{and} \qquad S = -R \sum_i P_i \times \ln(P_i)$$

where

$$\beta = \frac{1}{kT} \quad ; \quad Z = \sum_{i=1}^{\infty} e^{-\beta E_i} \quad ; \quad k = R/Na$$

2. Principles of thermodynamics

3. Entropy and Boltzman equation

Theorem

In the isothermal-isobaric ensemble these are the equations of state functions

$$Z = \sum_V \sum_j e^{-\beta E_{Vj}} e^{-\beta pV}$$

$$G = -kNT \ln Z$$

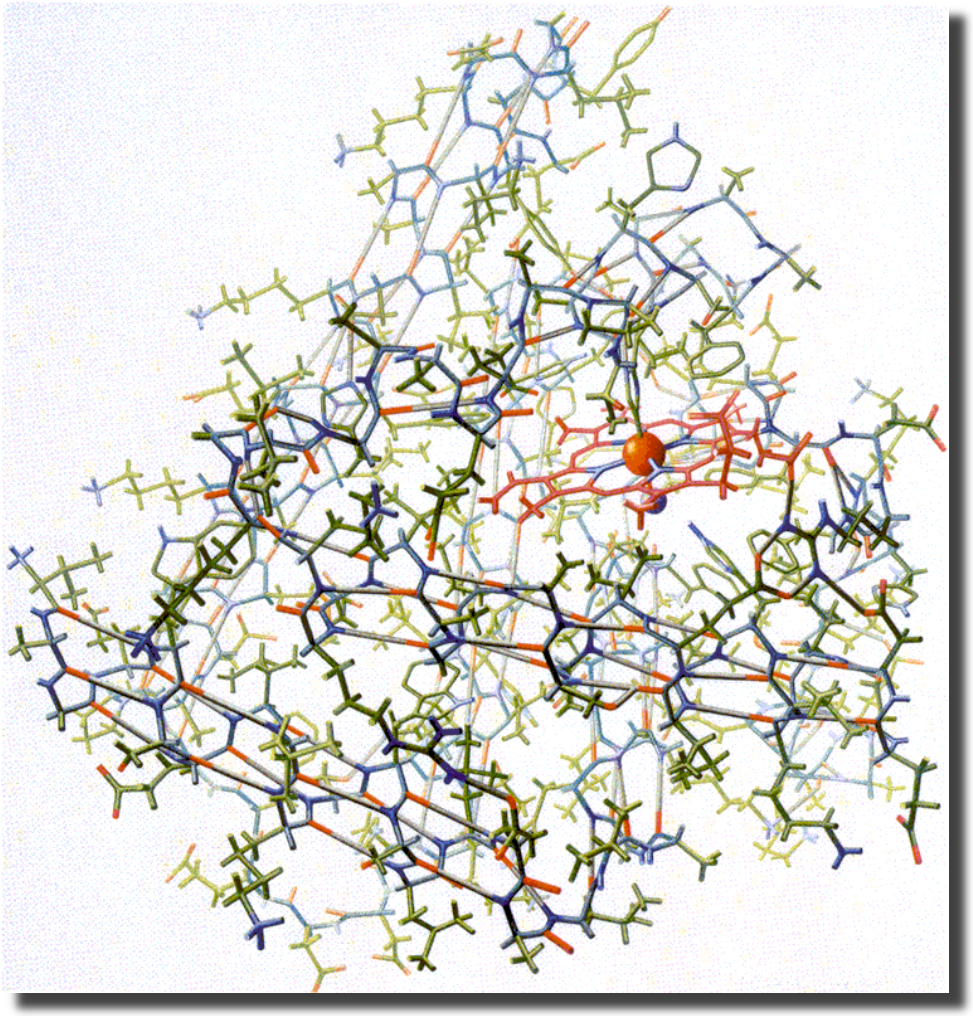
$$S = kN \ln Z + kN \left(\frac{\partial \ln Z}{\partial T} \right)_{N,p}$$

$$H = kNT^2 \left(\frac{\partial \ln Z}{\partial p} \right)_{T,N}$$

3. Soluble and globular proteins

1. Protein folding

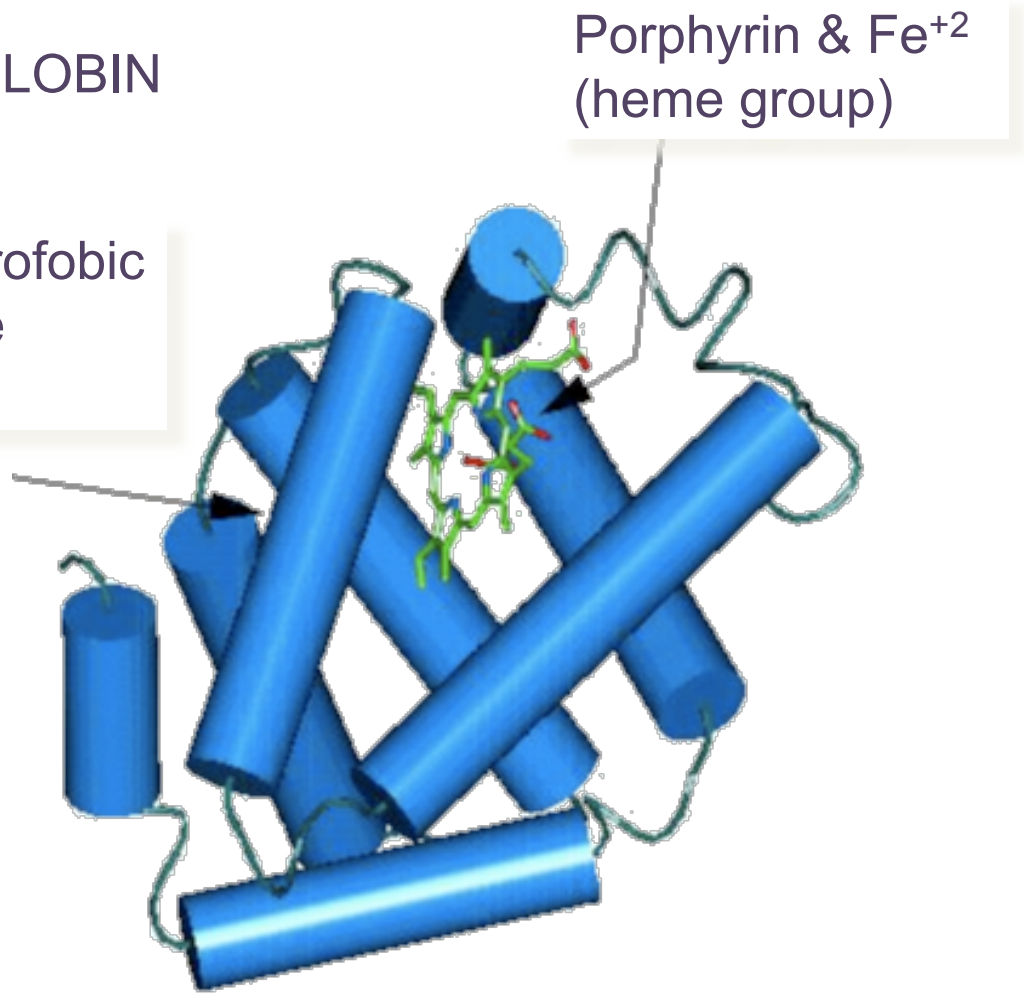
Example: MYOGLOBIN



3. Soluble and globular proteins

1. Protein folding

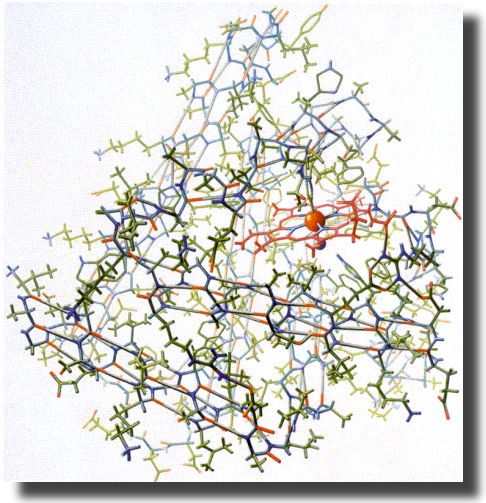
Example: MYOGLOBIN

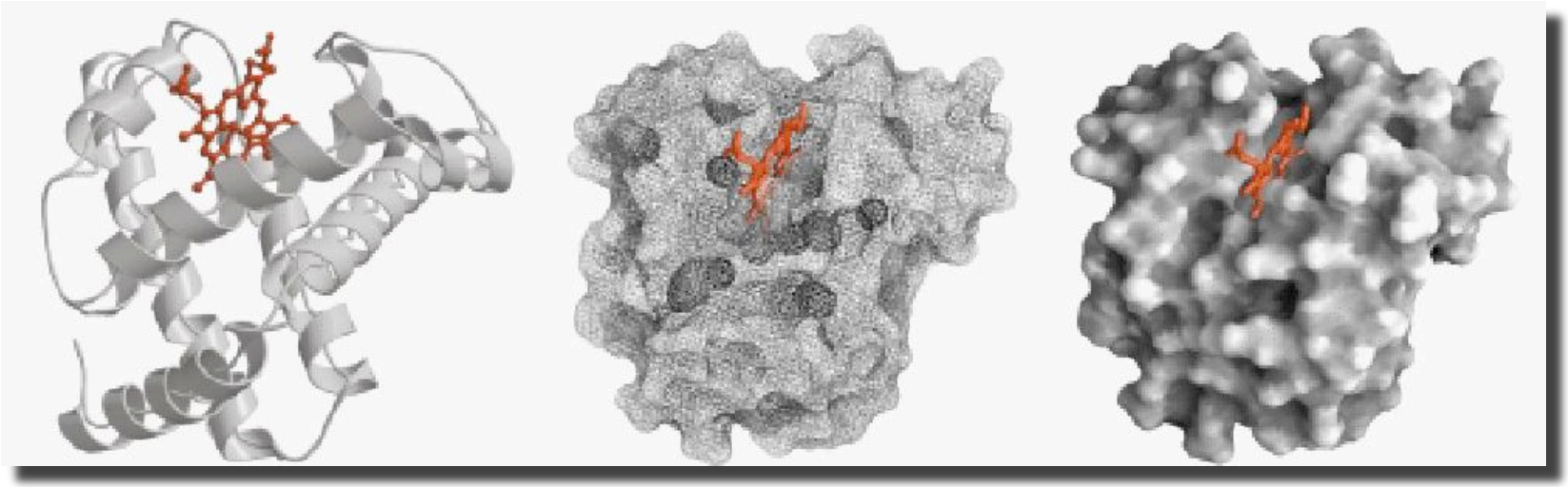


3. Soluble and globular proteins

1. Protein folding

Example: MYOGLOBIN

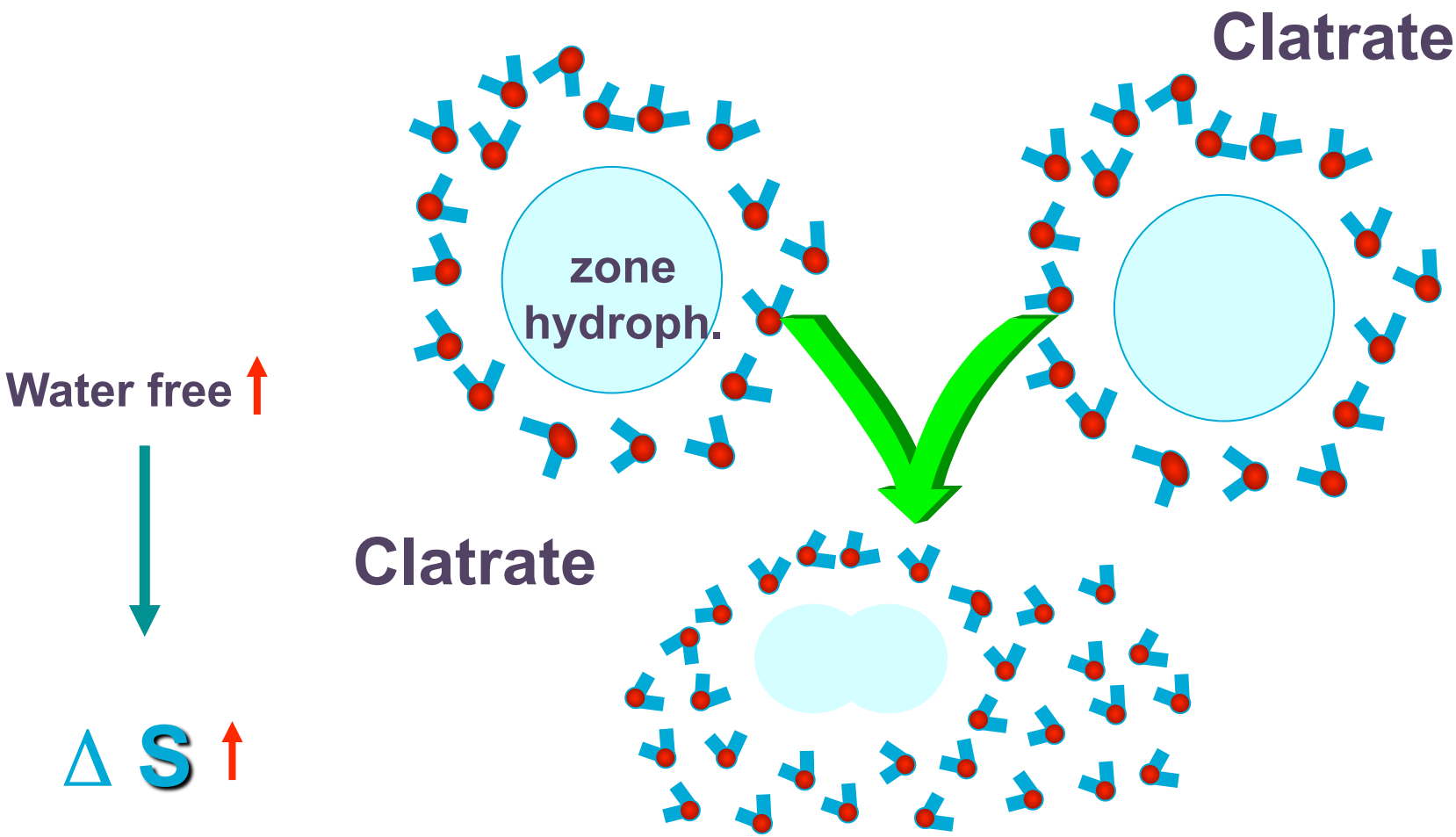




3. Soluble and globular proteins

1. Protein folding

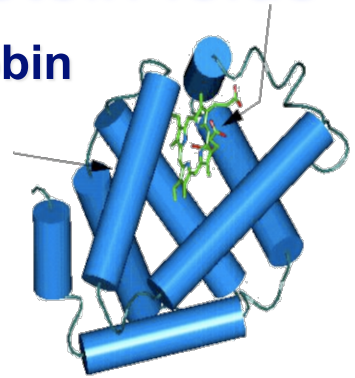
Hydrophobic effect



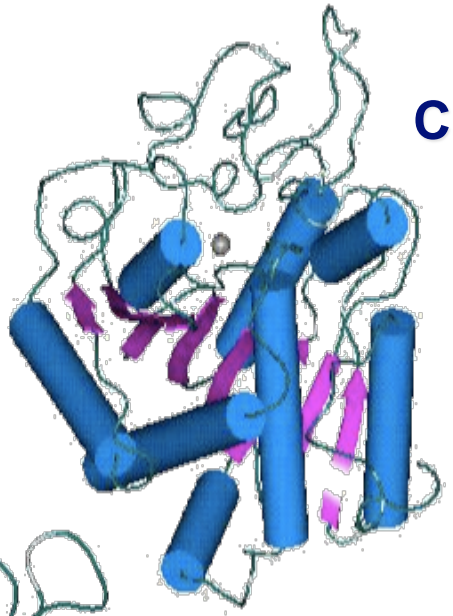
3. Soluble and globular proteins

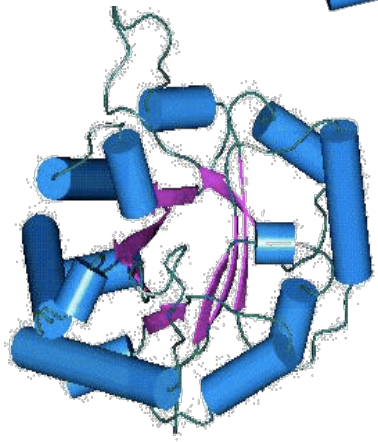
2. Protein folds

Myoglobin

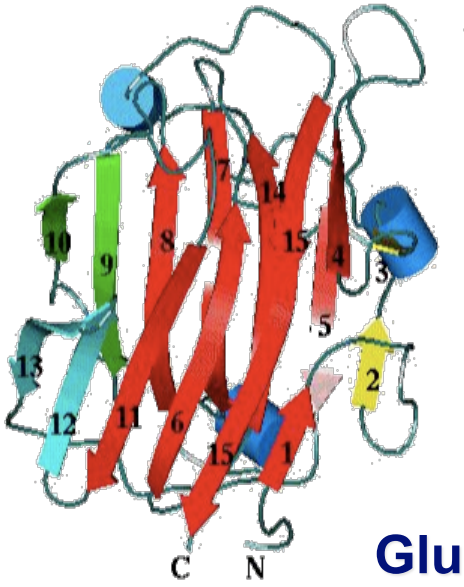


Carboxypeptidase

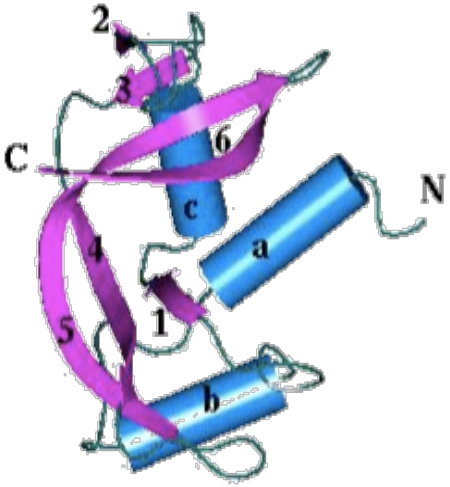




Triose Phosphate Isomerase



Glucanase



Ribonuclease

3. Soluble and globular proteins

3. Supersecondary structure

1. Definition

MOTIF or Super-Secondary Structure:
This is defined as a cluster of 2-4 regular secondary structures, usually involved in a particular function.

- Regular secondary structures are connected by loops.
- Loops are highly flexible and produce change of the polypeptide chain orientation.
- Length of loops varies from 1 to more than 40 residues
- Some super-secondary structures are stabilized with an hydrophobic core

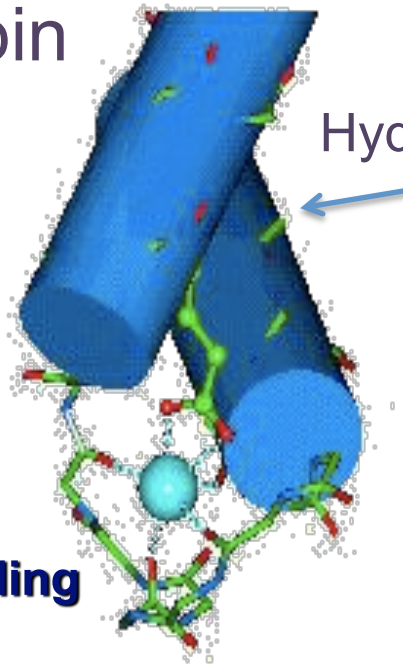
3. Soluble and globular proteins

3. Supersecondary structure

2. α−α

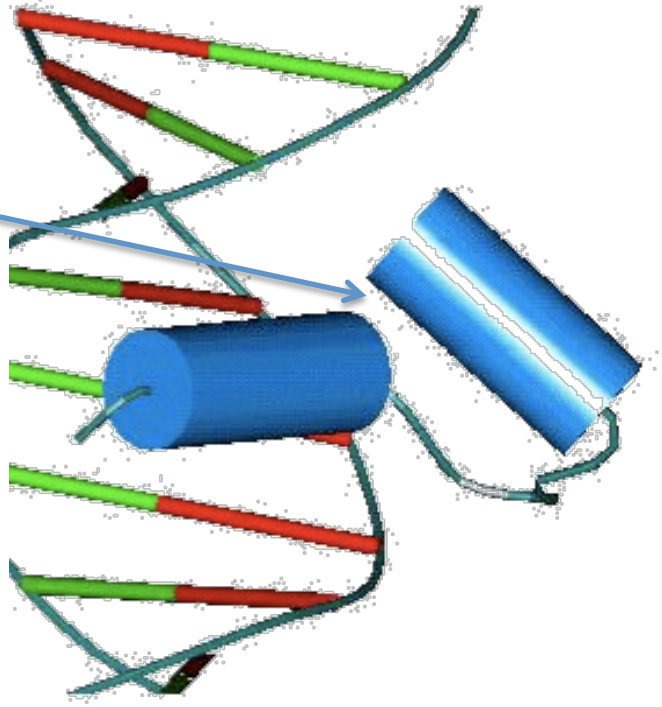
Ca^{+2} EF hand

α hairpin



Hydrophobic core

DNA binding motif
(Cro repressor)

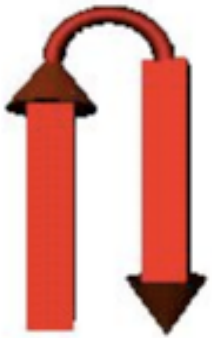


3. Soluble and globular proteins

3. Supersecondary structure

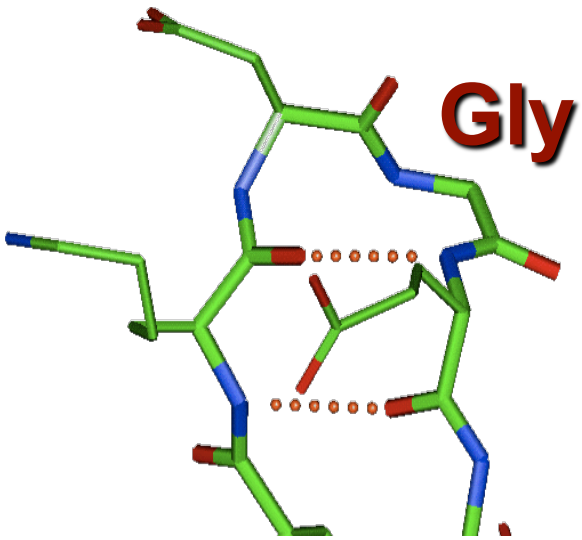
3. β−β

β−β hairpin



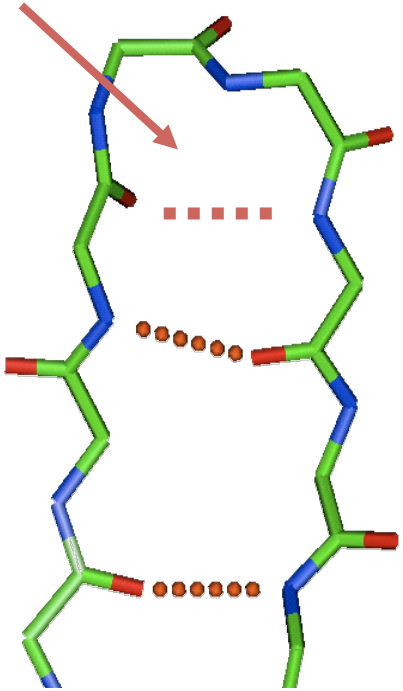
Anti-parallel
β-ladder

Turn β = 2 Aa



Type I’

H.B. lost



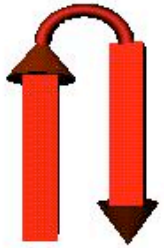
Type II

3. Soluble and globular proteins

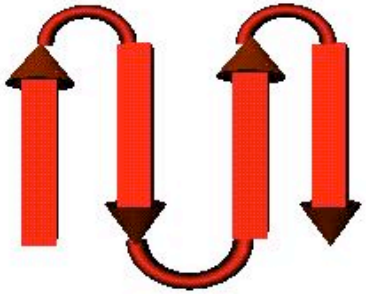
3. Supersecondary structure

3. β−β

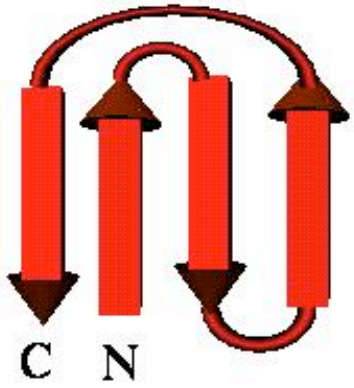
β hairpin



β meander



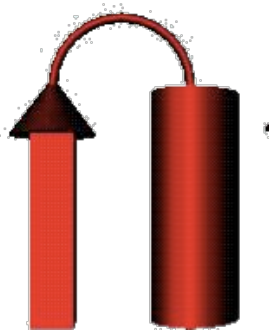
Greek Key

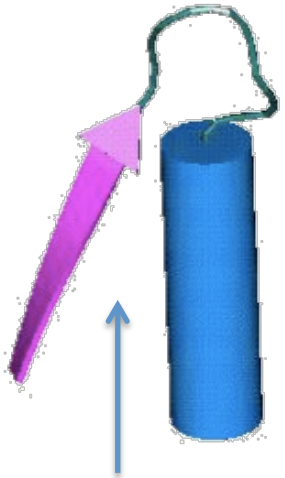


3. Soluble and globular proteins

3. Supersecondary structure

4. β−α & α−β

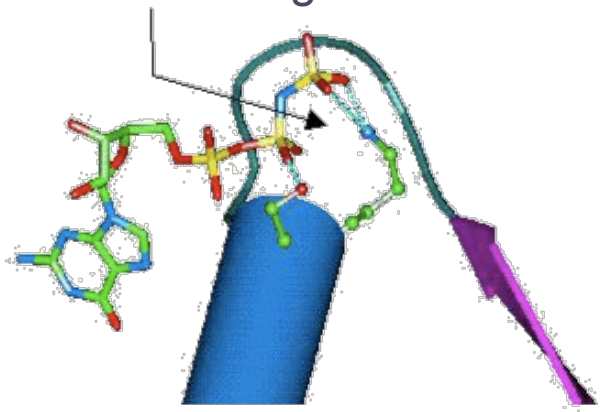




Hydrophobic core

P-loop

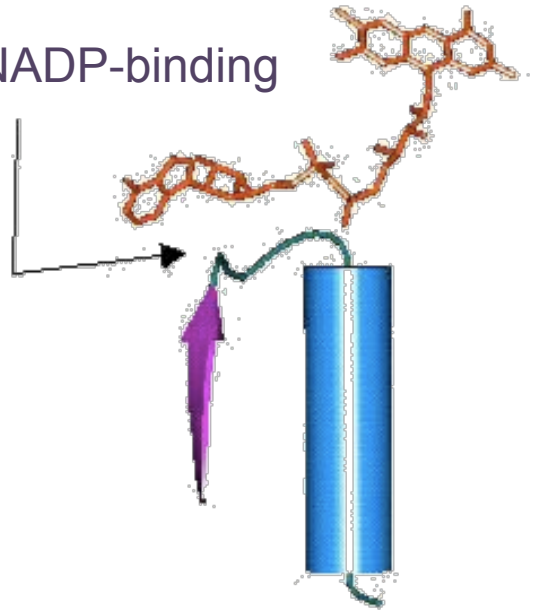
GDP-binding



Ras P21

FAD/NADP binding loop

FAD/NADP-binding

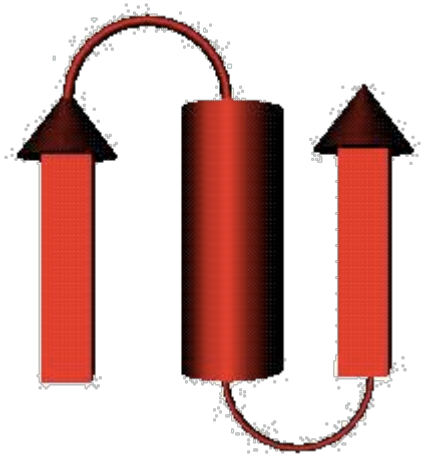


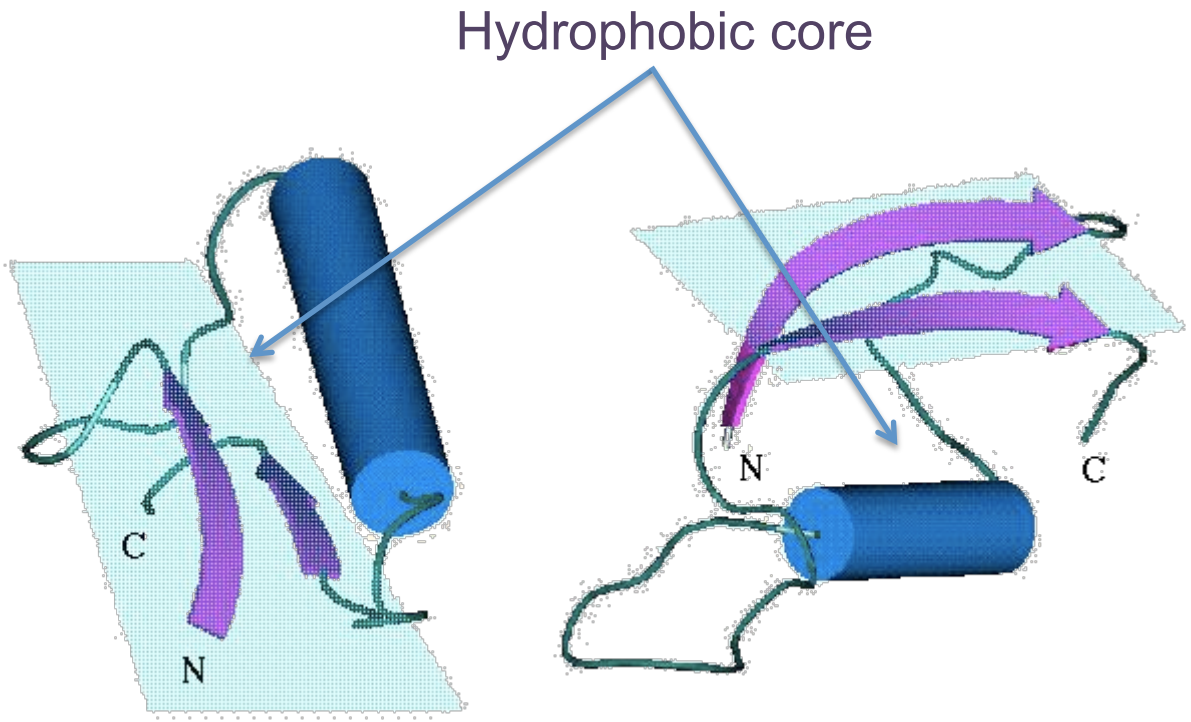
P-Hydroxybenzoate-Hydroxylase

3. Soluble and globular proteins

3. Supersecondary structure

5. β−α−β



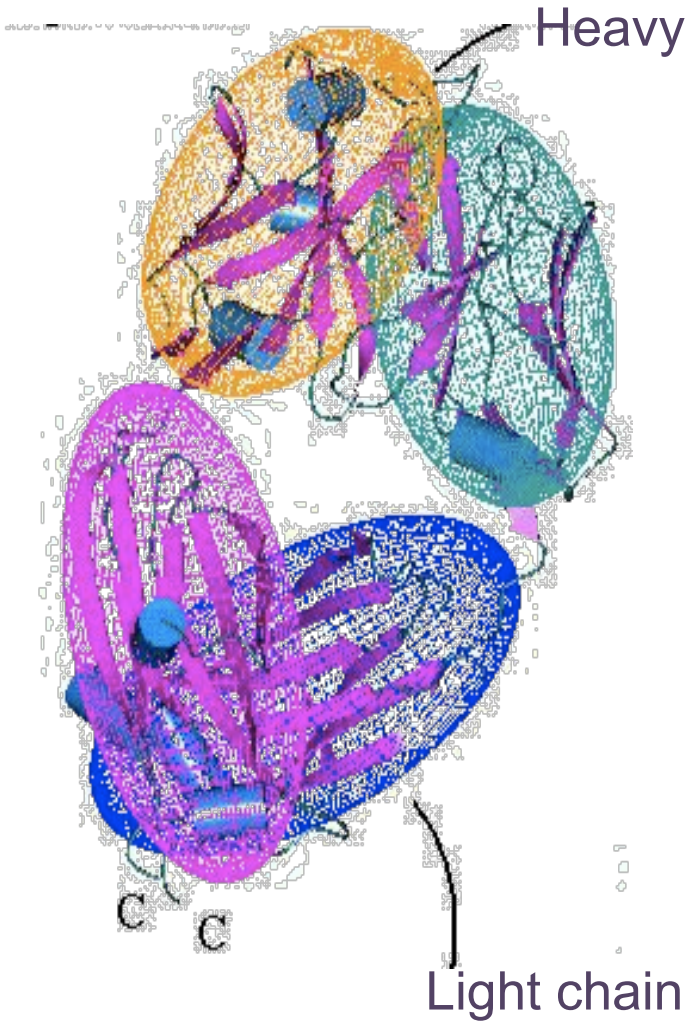


Right handed **Left handed**

3. Soluble and globular proteins

4. Domains

1. Definition

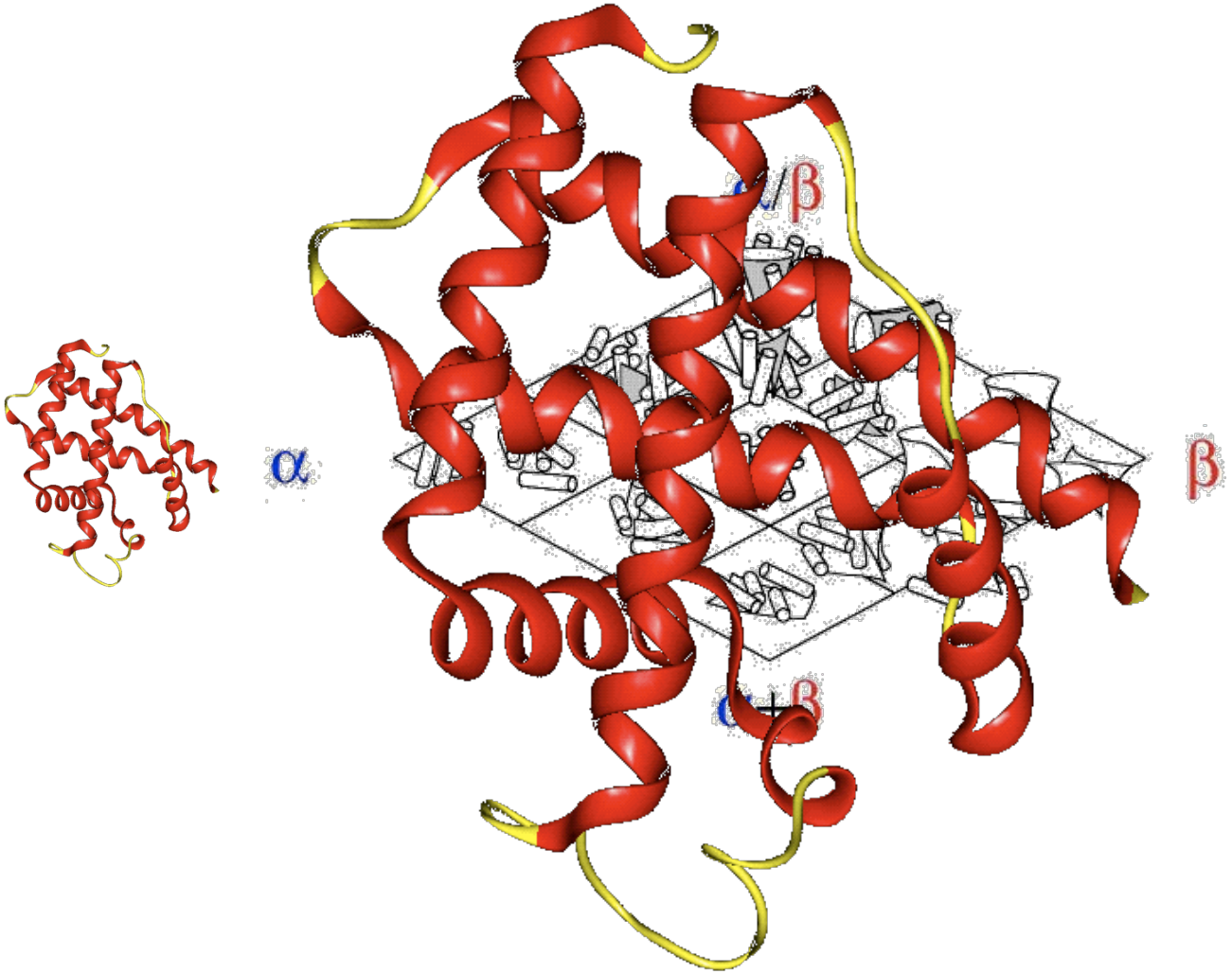


The protein domain is defined as the fundamental unit of 3D structure, able to fold by itself in the right conditions with independence of the rest of the protein.
It often corresponds to functional local and compact units of a protein

3. Soluble and globular proteins

4. Domains

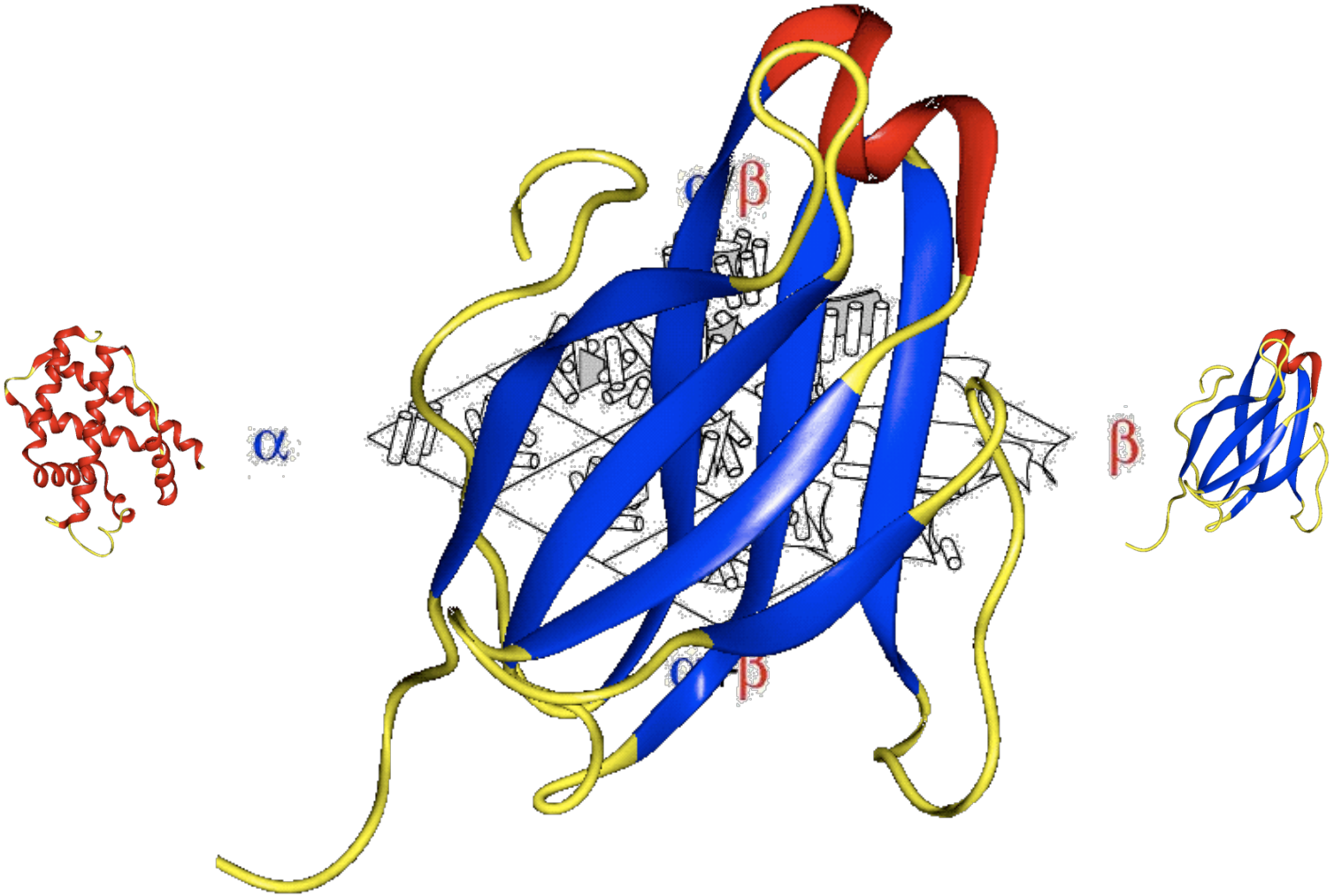
2. Classification



3. Soluble and globular proteins

4. Domains

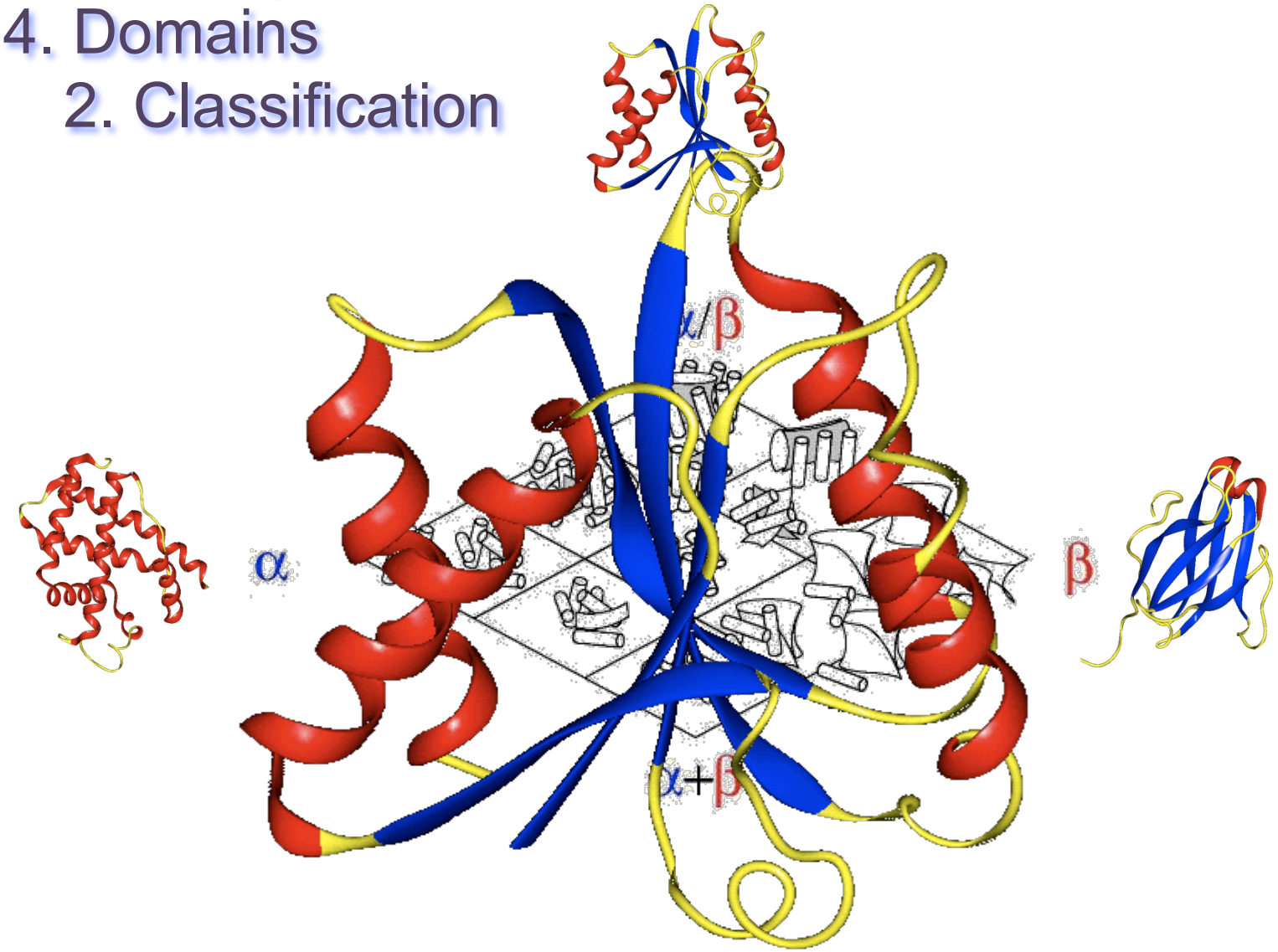
2. Classification



3. Soluble and globular proteins

4. Domains

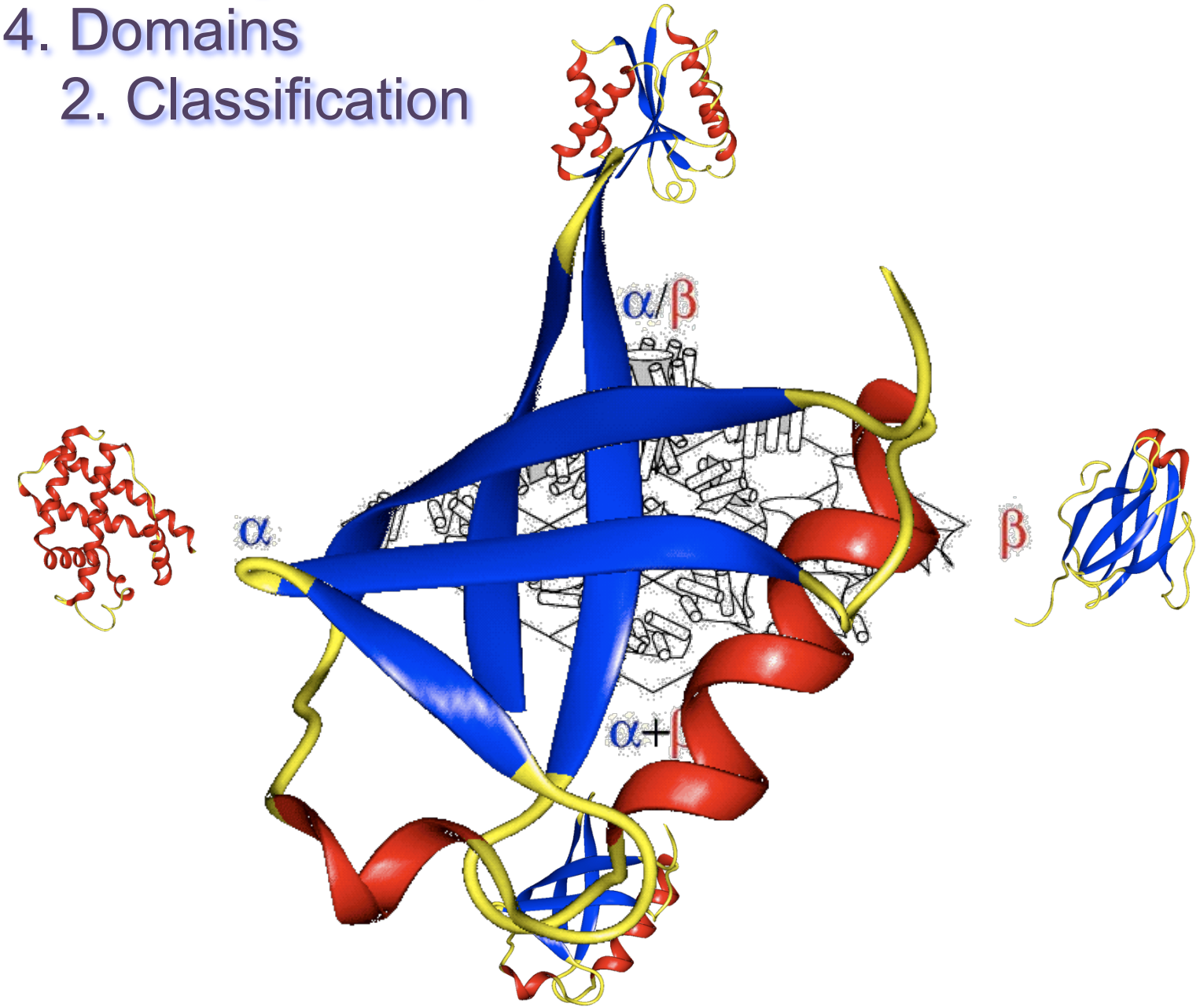
2. Classification



3. Soluble and globular proteins

4. Domains

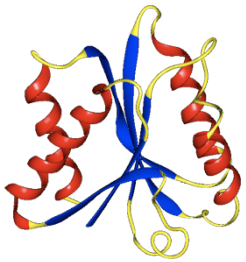
2. Classification



3. Soluble and globular proteins

4. Domains

2. Classification



6th class

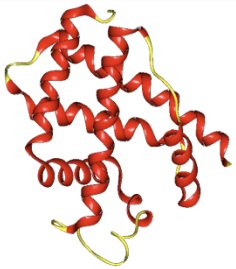
Membrane and cell surface proteins and peptides.

Does not include proteins in the immune system

5th class

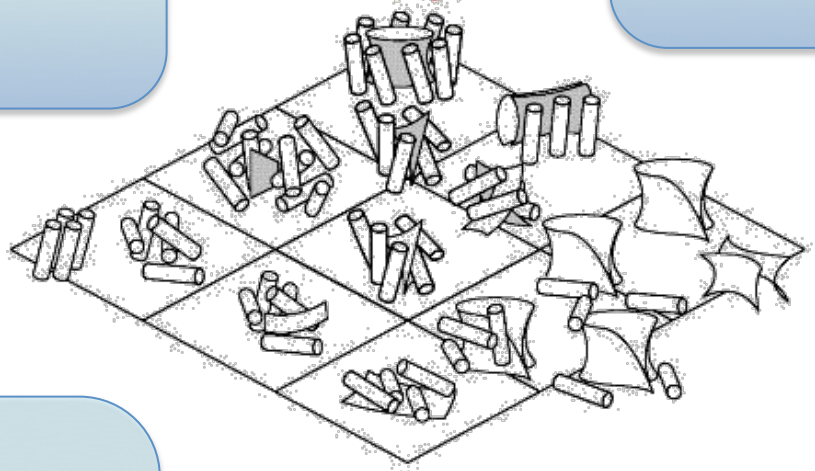
Multi-domain proteins α and β

Folds consisting of two or more domains belonging to different classes



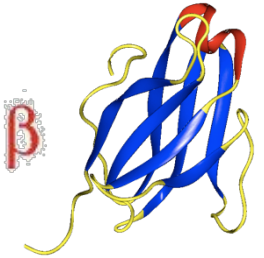
α

α/β



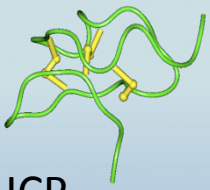
α+β

β

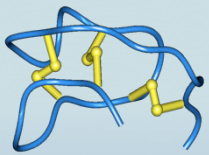


7th class

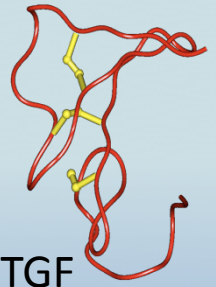
Small proteins (disulphide-rich, metal rich)



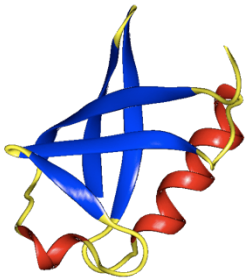
ICP



BPTI



TGF

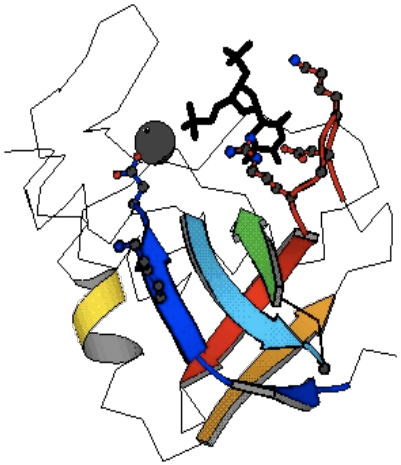


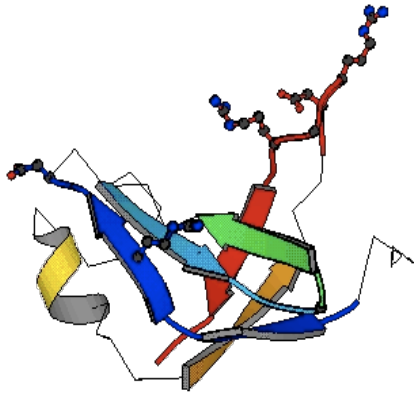
3. Soluble and globular proteins

4. Domains

3. Function-Sequence-Structure

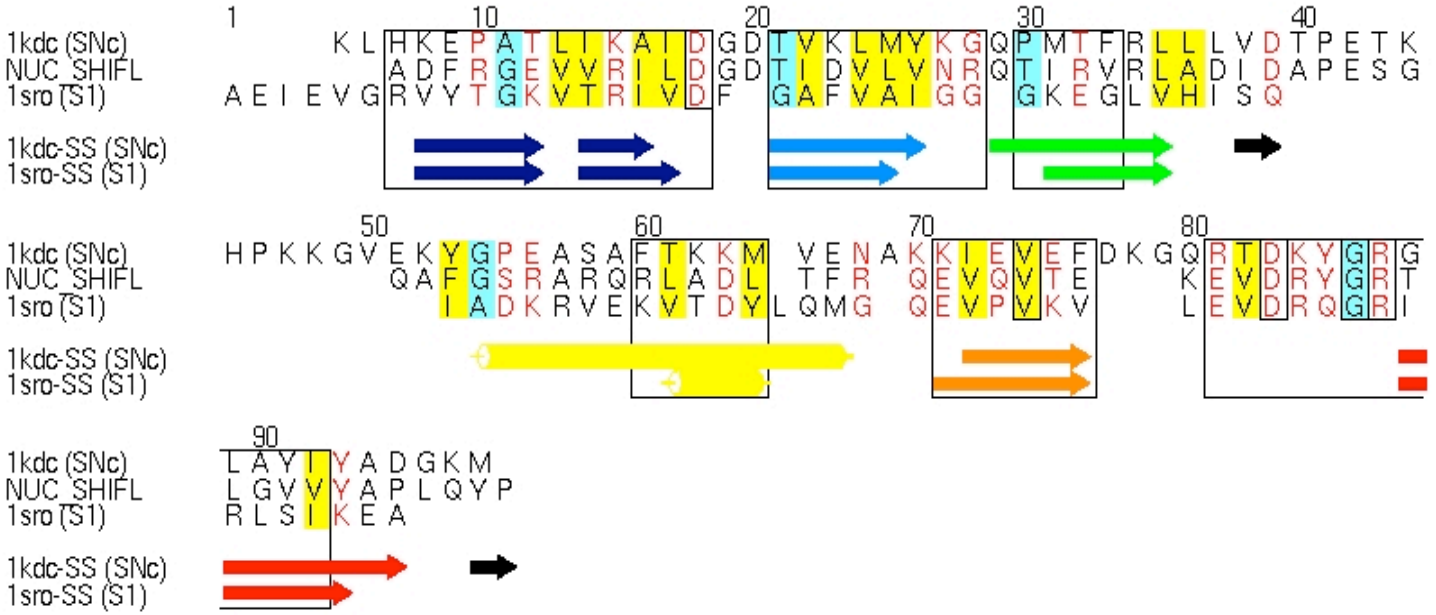
Stapphococcyl nuclease





S1 RNA binding domain

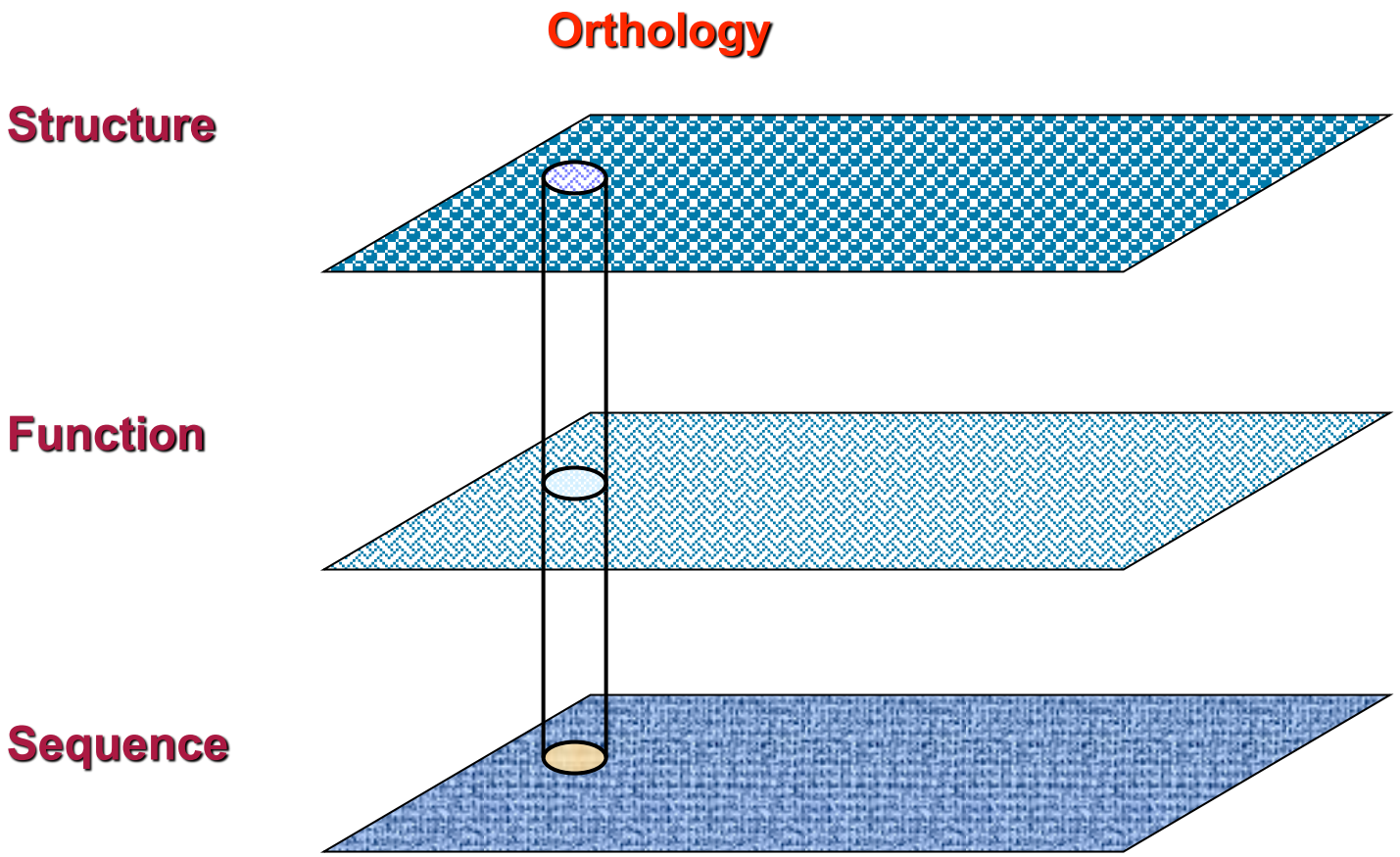
The structural alignment shows the best conservation of residues and a possible relationship in evolution



3. Soluble and globular proteins

4. Domains

3. Function-Sequence-Structure

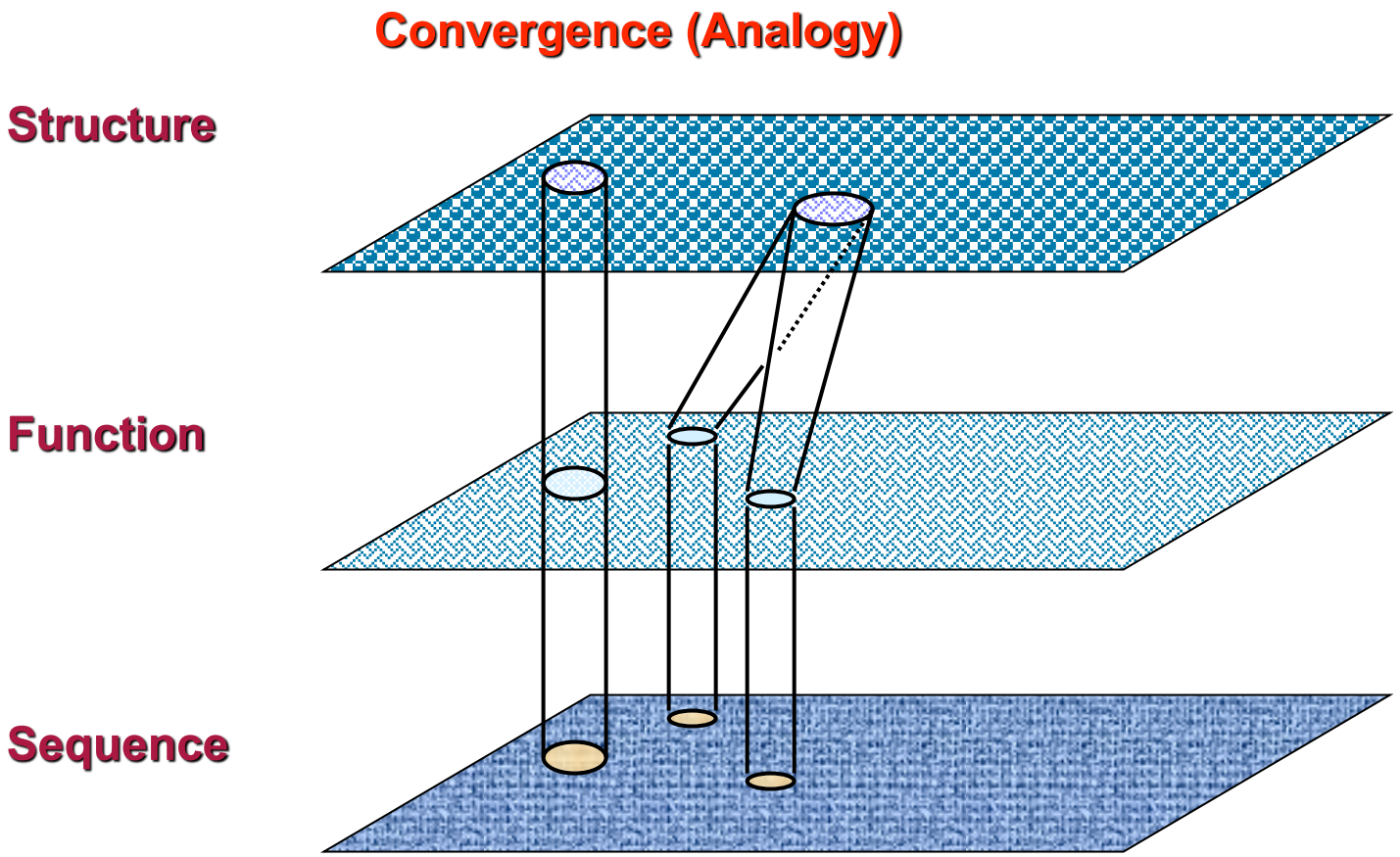


Homology: two proteins are homologous if they are the products of genes that evolved from the same ancestor

3. Soluble and globular proteins

4. Domains

3. Function-Sequence-Structure

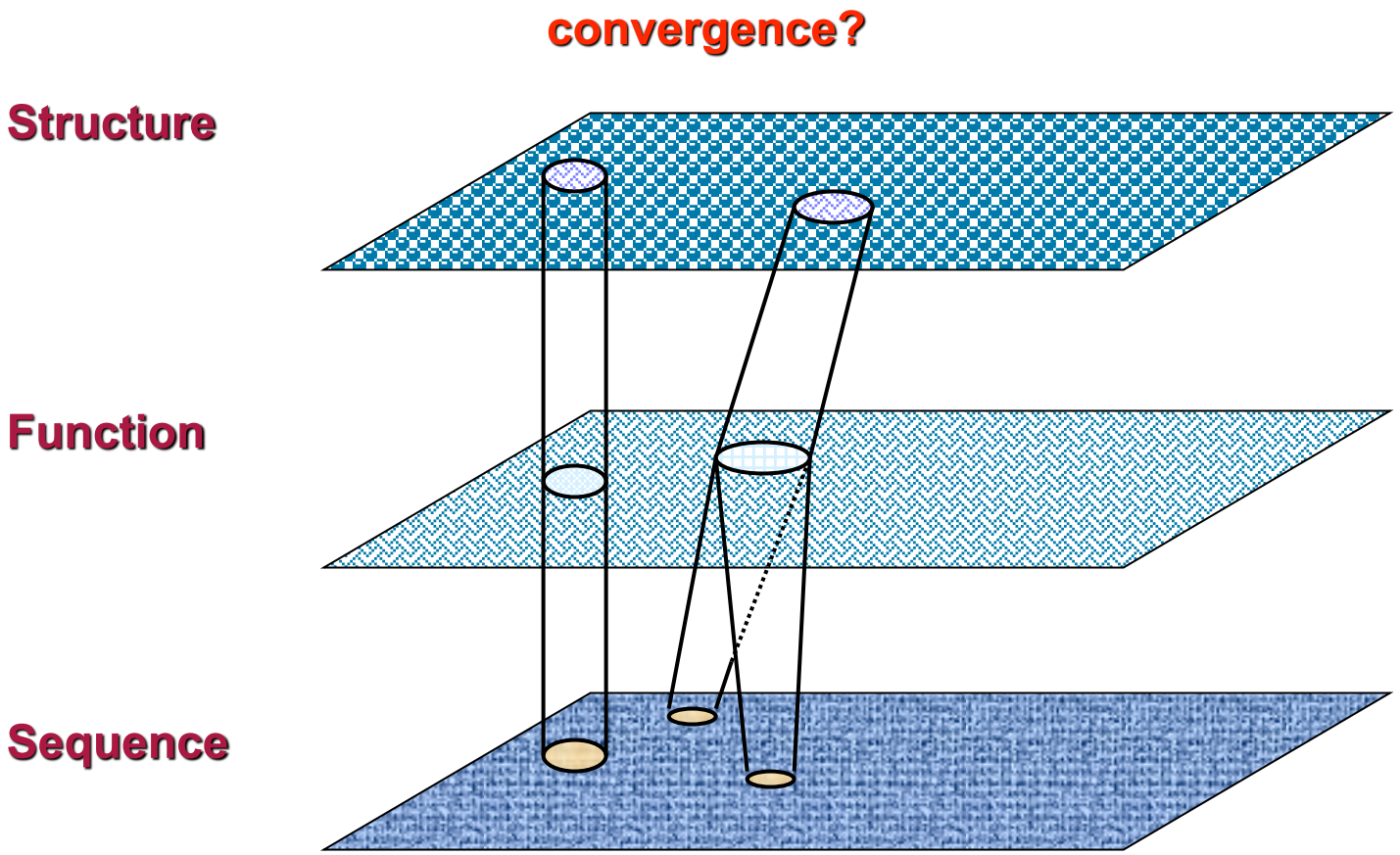


Homology: two proteins are homologous if they are the products of genes that evolved from the same ancestor

3. Soluble and globular proteins

4. Domains

3. Function-Sequence-Structure

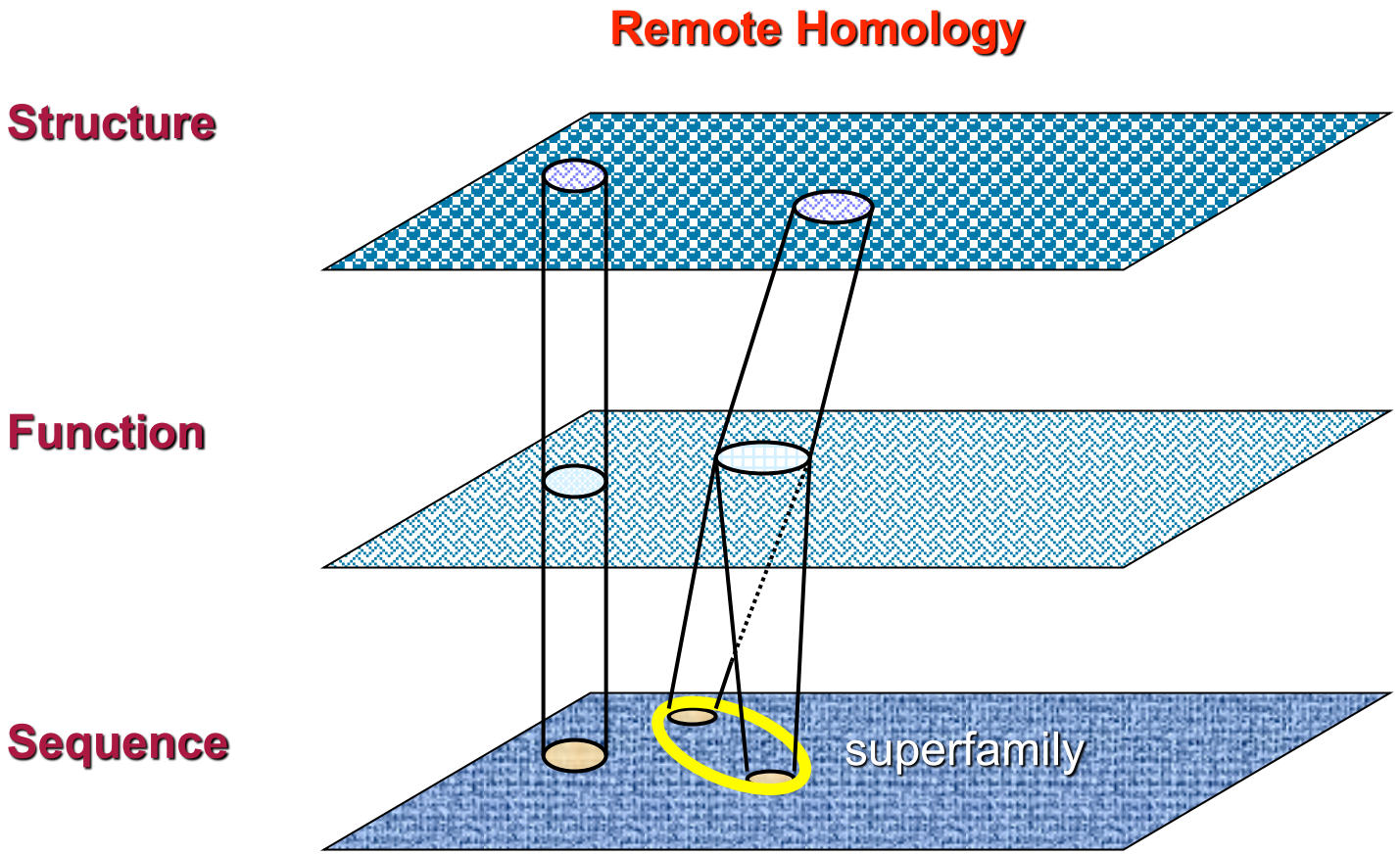


Homology: two proteins are homologous if they are the products of genes that evolved from the same ancestor

3. Soluble and globular proteins

4. Domains

3. Function-Sequence-Structure

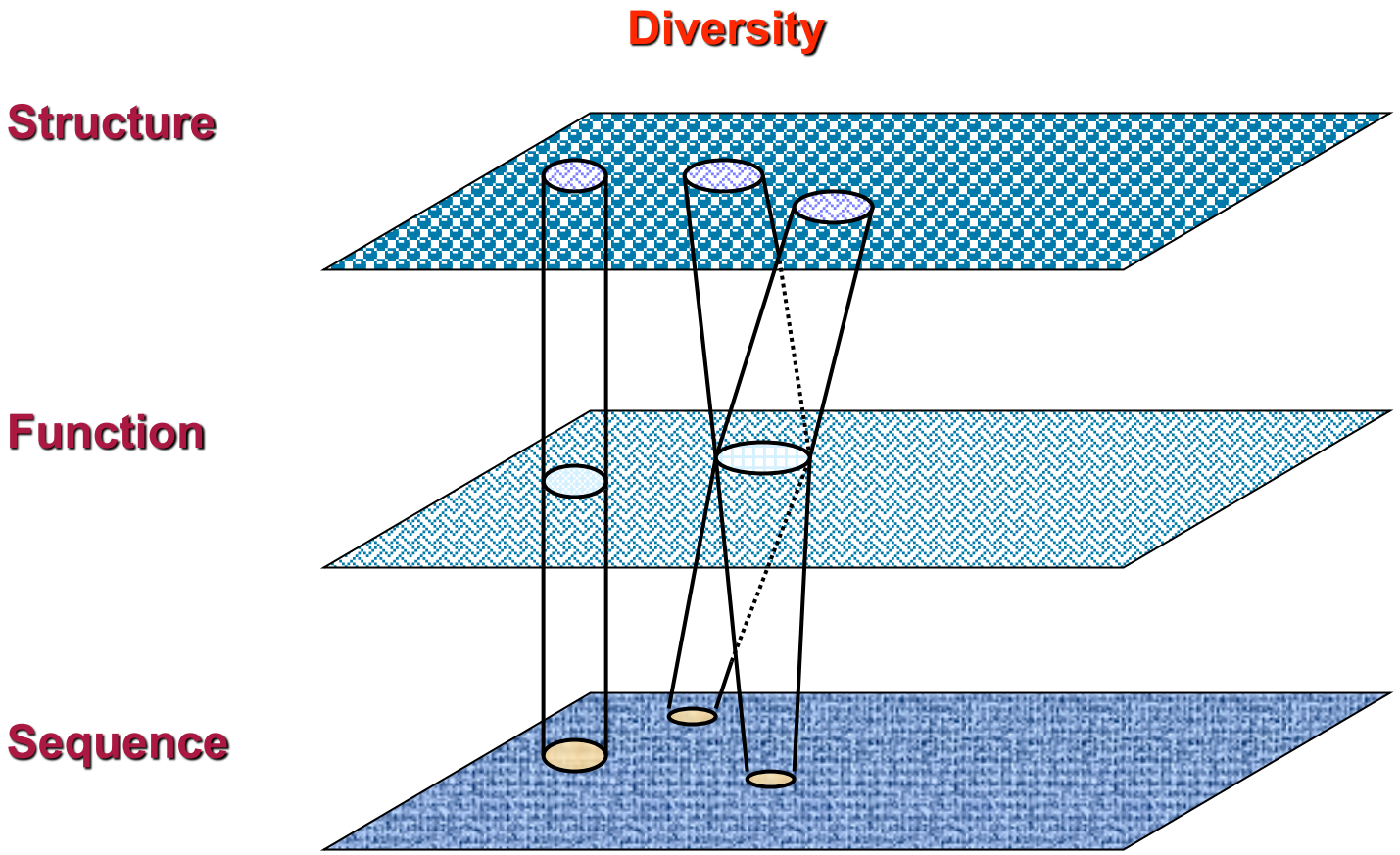


Homology: two proteins are homologous if they are the products of genes that evolved from the same ancestor

3. Soluble and globular proteins

4. Domains

3. Function-Sequence-Structure

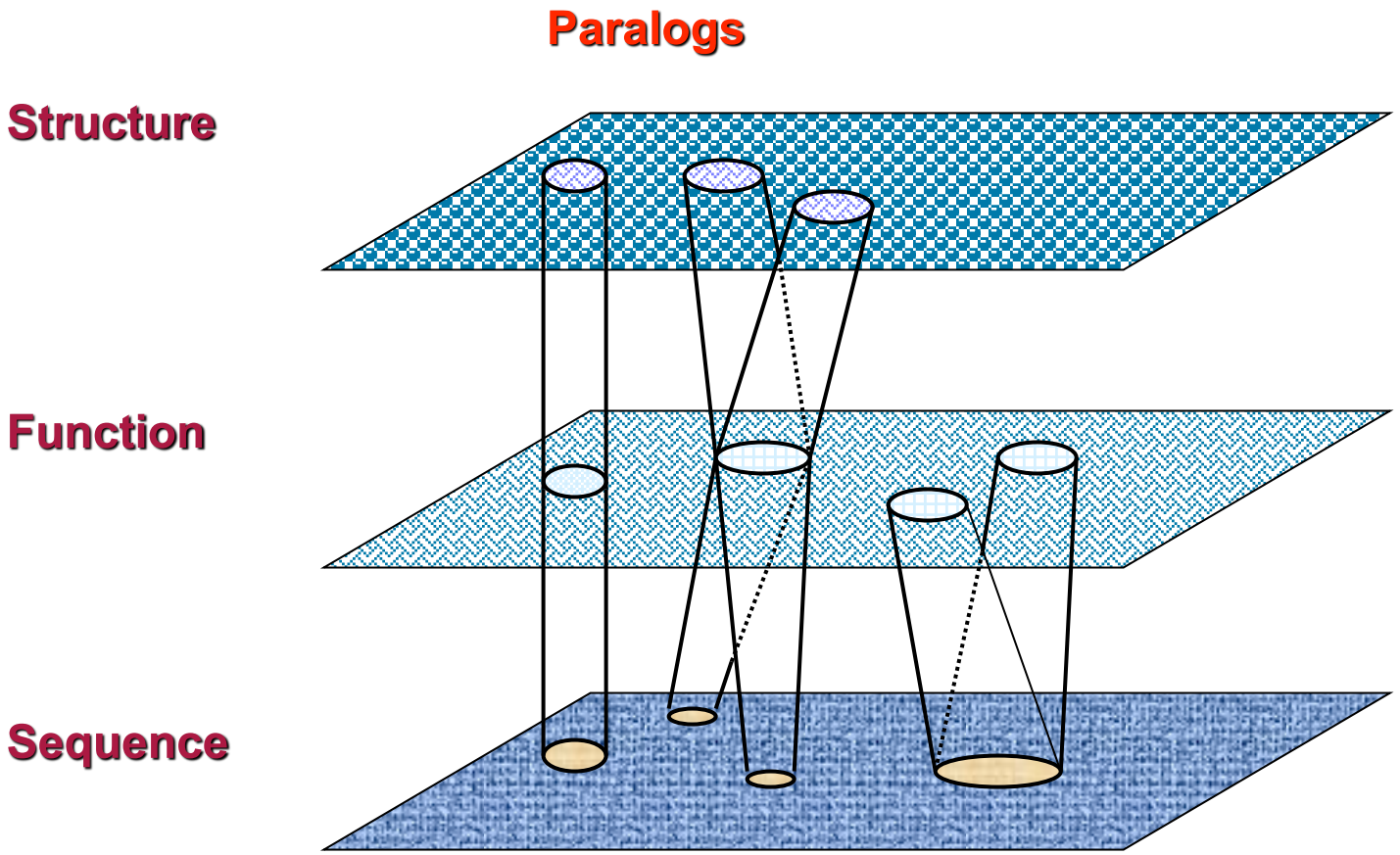


Homology: two proteins are homologous if they are the products of genes that evolved from the same ancestor

3. Soluble and globular proteins

4. Domains

3. Function-Sequence-Structure

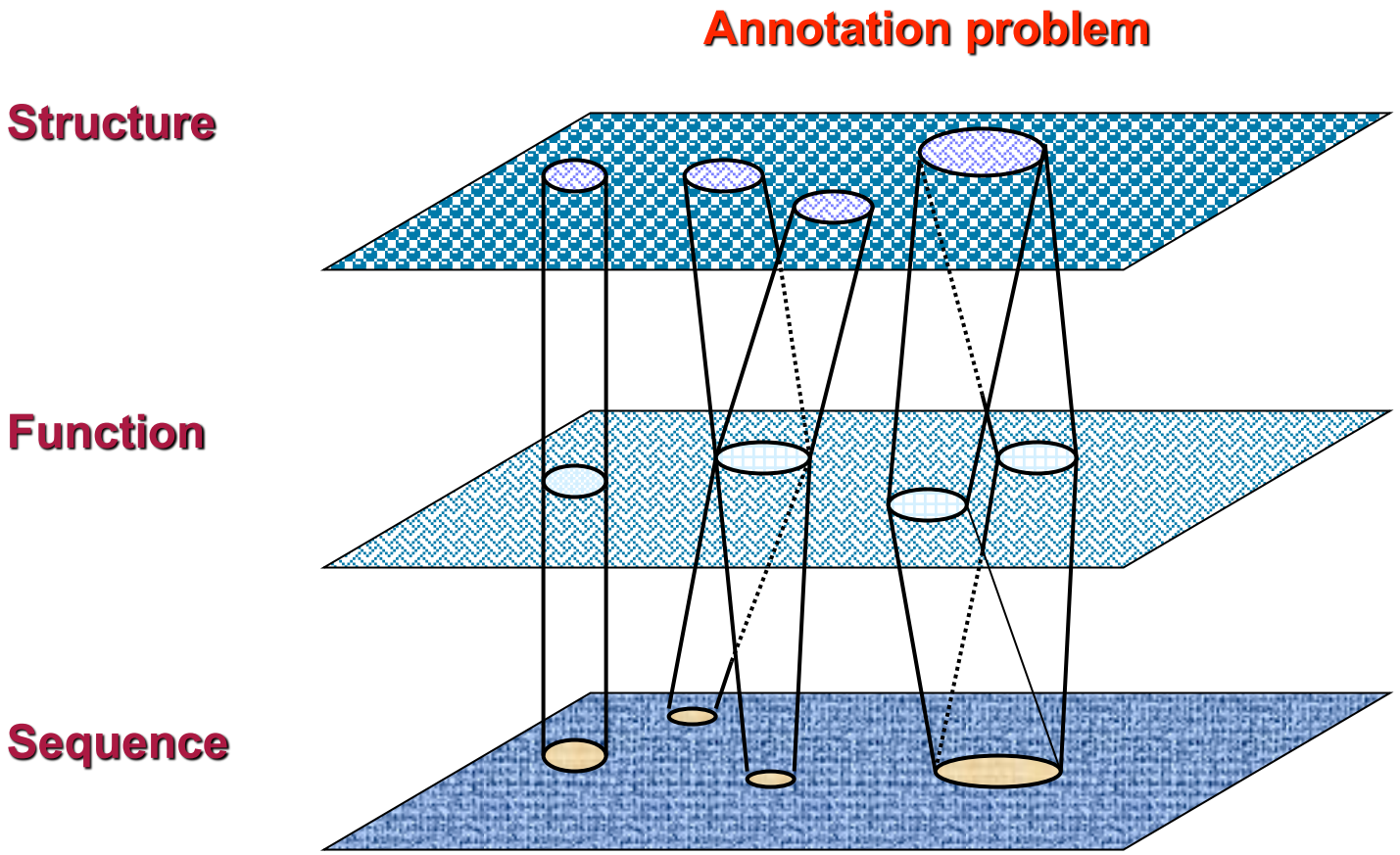


Homology: two proteins are homologous if they are the products of genes that evolved from the same ancestor

3. Soluble and globular proteins

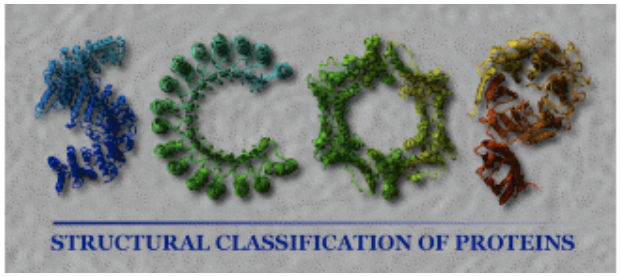
4. Domains

3. Function-Sequence-Structure



Homology: two proteins are homologous if they are the products of genes that evolved from the same ancestor

3. Soluble and globular proteins
4. Domains
4. SCOP



Structural Classification Of Proteins (SCOP)

Class: It groups the folds according to the percentage and 3D disposition of the regular secondary structures.

Family: It groups proteins clearly related by homology. In general we assume they are homologs if the alignment has >30% ID, they have the similar structure and similar function

Superfamily: Proteins which sequences align with very few %ID but showing similar structural patterns and similar function.

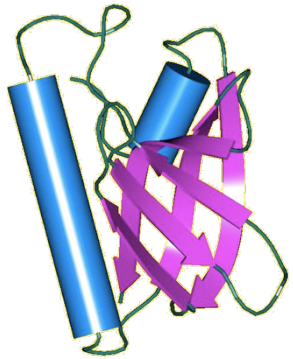
Fold: Proteins with very similar disposition of the regular secondary structures

3. Soluble and globular proteins

4. Domains

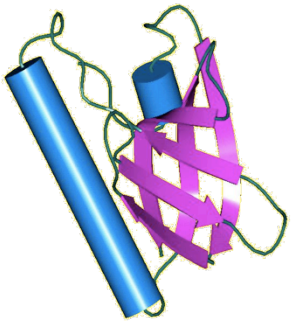
4. SCOP

- **Family**
- **Superfamily**
- **Fold**
- **Class**



Enterotoxin

Homology



Cholera toxin

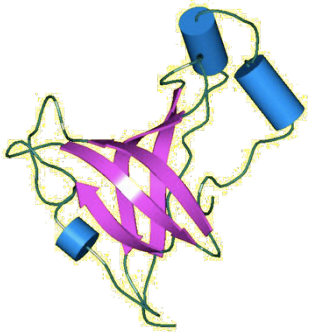
80% Id

Remote Homology

TSS toxin

8.8% Id

Analogs



Aminoacyl tRNA synthetase

4.4% Id

3. Soluble and globular proteins

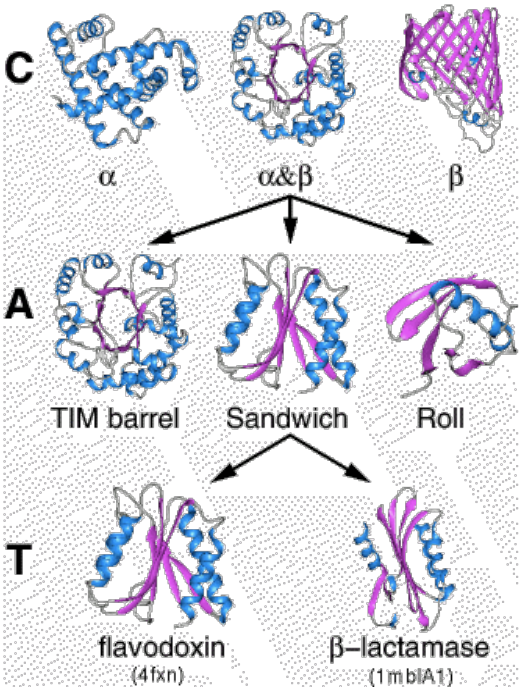
4. Domains

5. CATH

Class 1
Mainly α

Class 3
Mixed
α & β

Class 2
Mainly β

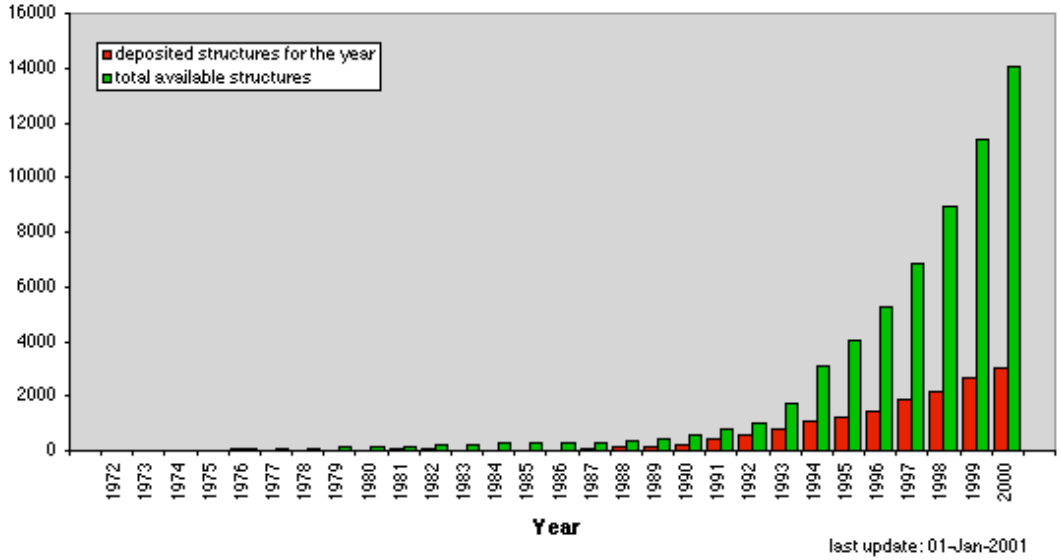


- **Class(C)**
  Class is determined according to the secondary structure composition and packing within the structure. It is assigned automatically
- **Architecture(A)**
  This describes the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures. It is currently assigned manually
- **Topology(T)**
  Structures are grouped into fold groups at this level depending on both the overall shape and connectivity of the secondary structures. This is done using an automated structure comparison algorithm.
- **Homologous superfamily (H)**.
  This level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous. Similarities are automatically identified either by high sequence identity or structure comparison.
- **Sequence Family Levels: (S,O,L,I, D)**
  Domains within each H-level are subclustered into sequence families using multi-linkage clustering S(35%), O(60%, L (90%), I (100%)

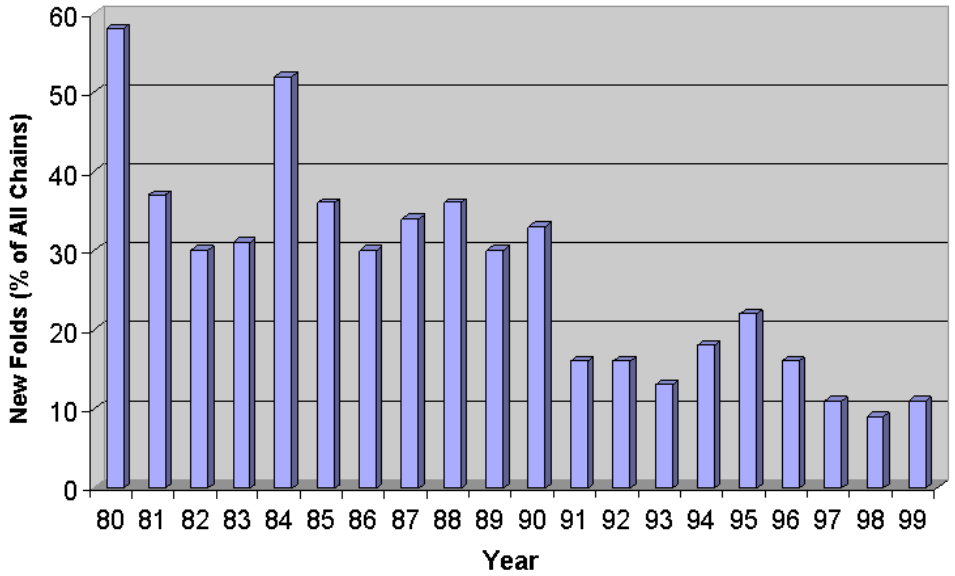
Protein structure II

3. Soluble and globular proteins
    5. Principles observed in folds
    6. Classical folds (TIM & Rossman)
    7. All α
    8. All β
    9. α/β
    10. $\alpha+\beta$
4. Globular membrane proteins
    1. Definition
    2. Classification
        1. Integral
        2. Periferal

3. Soluble and globular proteins
5. Principles observed in folds



Geometric increase of the number of structures in PDB



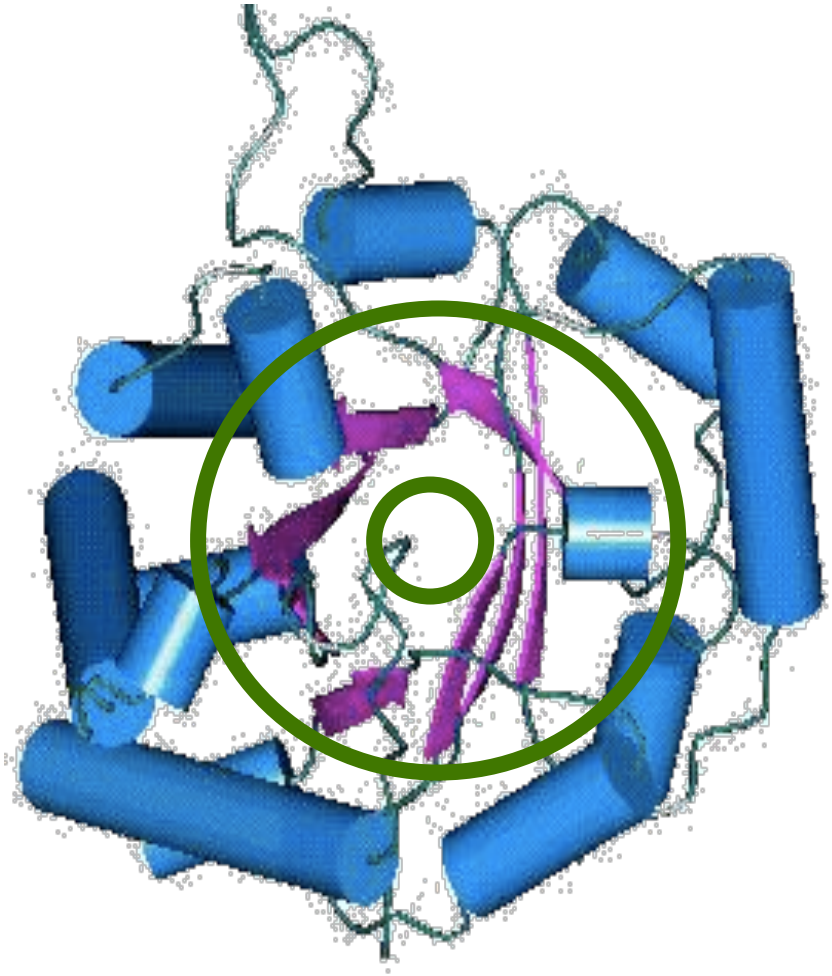
The % of new folds has reached a plateau and the tendency will be to decrease until no more new folds are found.

3. Soluble and globular proteins
5. Principles observed in folds

- Folds have a compact structure formed by the organization of regular secondary structures
- Folds have an hydrophobic core that stabilizes its conformation
- The hydrophobic core is formed by non-polar residues, while polar residues are mostly in the surface and active sites
- Structure is conserved among proteins with similar sequence
- The total number of folds is finite
- The structure of a protein an be build by joining several folds (this can be produced by gene fusion).

3. Soluble and globular proteins
6. Classical folds

Triose Phosphate Isomerase (TIM Barrel)



- Double hydrophobic core
- Highly stable
- Found in many proteins, mostly enzymes.
- It was recently proved that all of them have the same ancestor

3. Soluble and globular proteins

6. Classical folds

TIM Barrel Evolution

- A fusion gene is a hybrid gene formed from two previously separate genes. It can occur as the result of a translocation, interstitial deletion, or chromosomal inversion.

- Gene duplication plays an important role in enzyme evolution. It has been estimated that 50% of all genes in microorganisms are the result of duplication events, which are followed by diversification of the twin genes (Fani et al 1998, Lynch & Conery 2000).

- Gene duplication and fusion events that multiply and link functional protein domains are crucial mechanisms of enzyme evolution.

- The analysis of amino acid sequences and three-dimensional structures suggested that the (αβ)8-barrel, has evolved by the duplication, fusion, and mixing of (αβ)4-half-barrel domains (Höcker et al 2004).

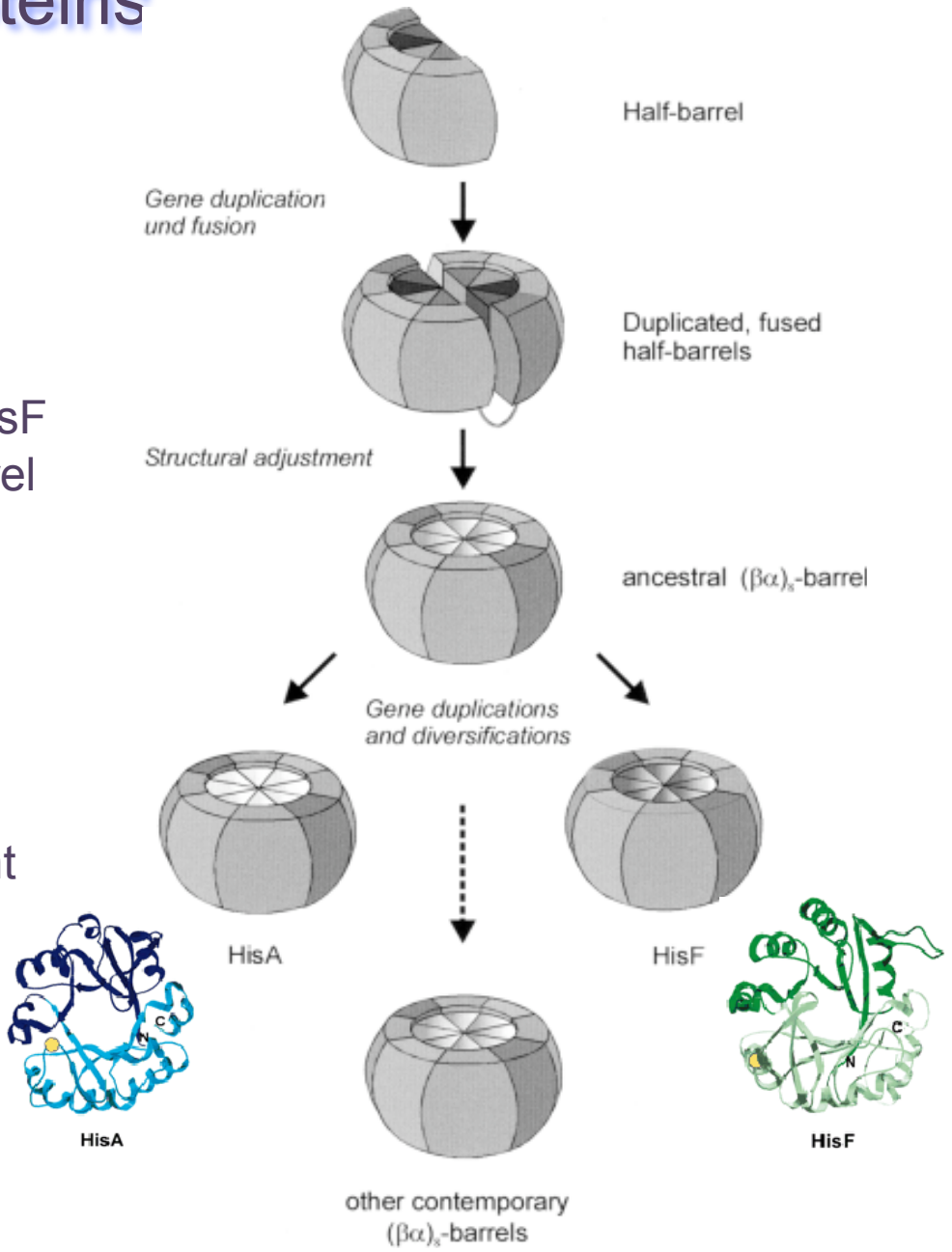
3. Soluble and globular proteins

6. Classical folds

TIM Barrel Evolution

It has been postulated that HisA and HisF evolved from a common (βα)4-half-barrel by a series of gene duplication and diversification events (Höcker et al 2004).

(βα)8-barrels were derived from 'half-barrels' will motivate the search for ancestral domains within other apparent single-domain protein folds (Petsko, 2000; Gerlt and Babbitt, 2001, Jürgens al 2000).



3. Soluble and globular proteins

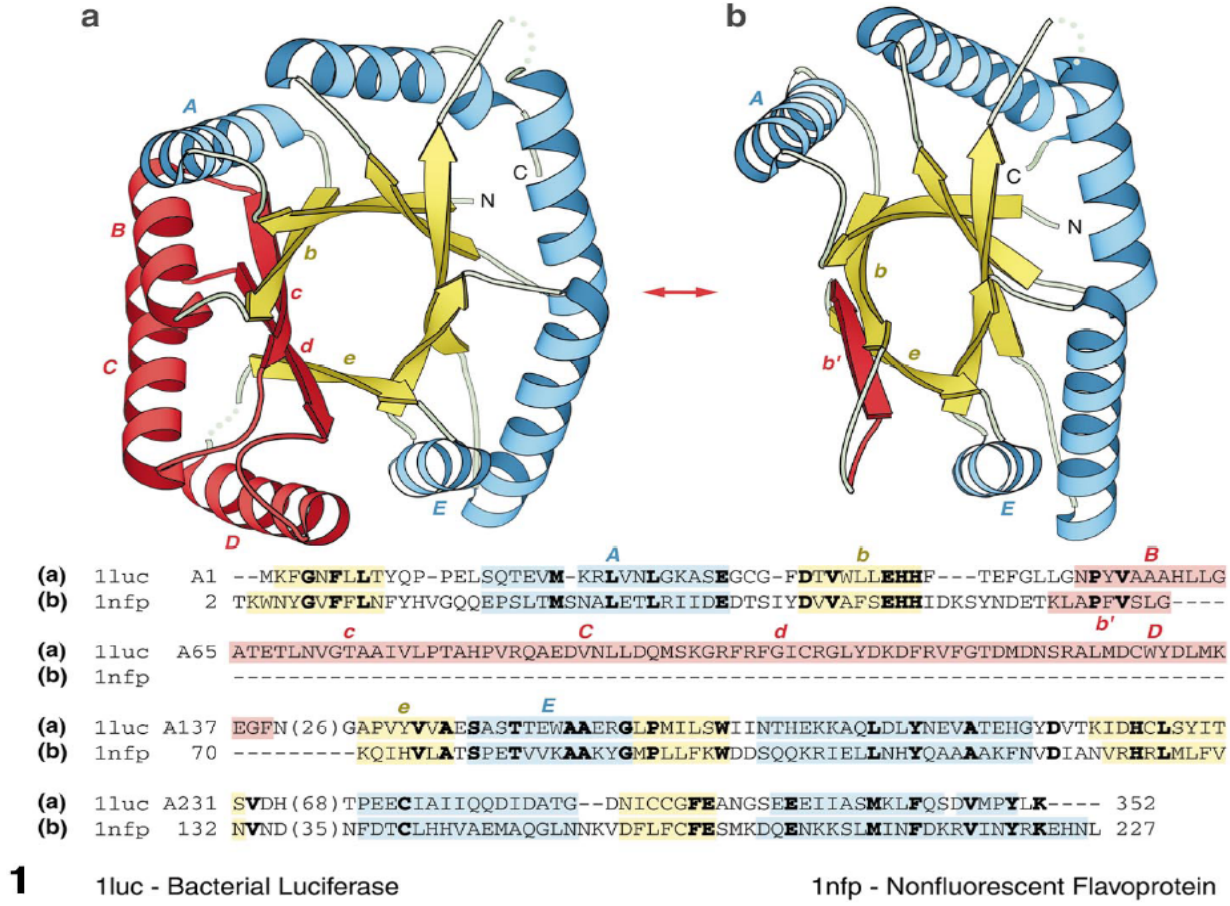
6. Classical folds

TIM Barrel Evolution

Insertion, deletion and substitution

Luciferase and NFP

- NFP deletes a 90-residue section αβαβα
- Missing segment is compensated by a single anti-parallel strand
- Shared regions are 30% identical



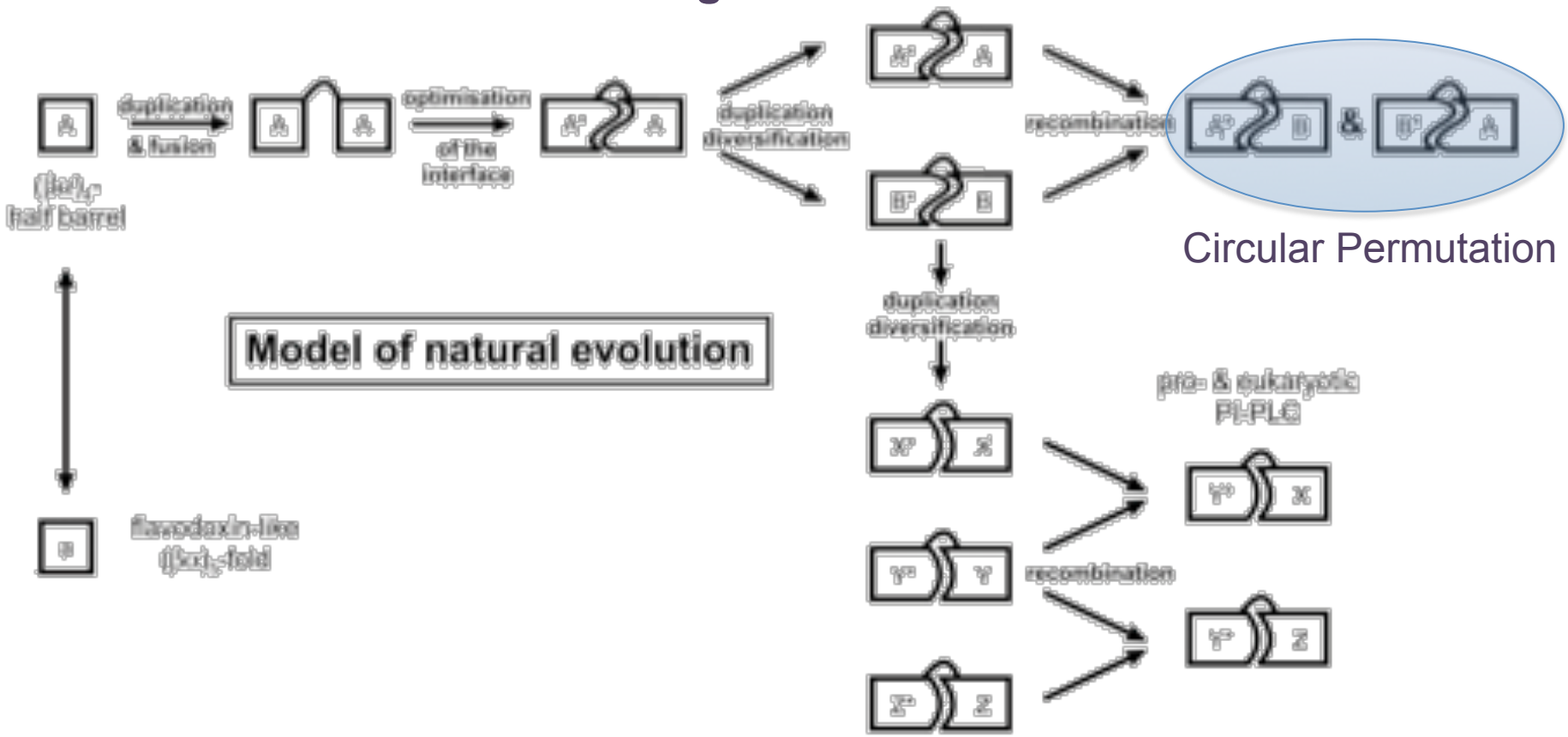
3. Soluble and globular proteins

6. Classical folds

TIM Barrel Evolution

Circular Permutation

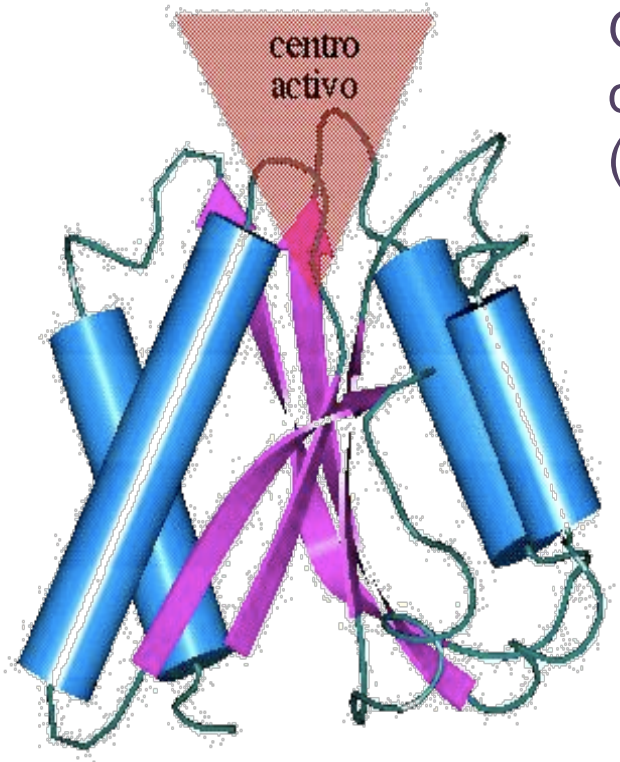
- Circular permutations can be detected when the N- and C-termini are near
- Similar structures are found but the alignment in the sequence is shifted.
- Example: between proteins of the super families **FAD-linked oxidoreductase** and **PLP-binding barrel**.



3. Soluble and globular proteins

6. Classical folds

Rossmann Fold



Rossman Fold of Flavodoxin

Often found as binding domain of mono-nucleotides (NAD, FAD, FMN, etc.).

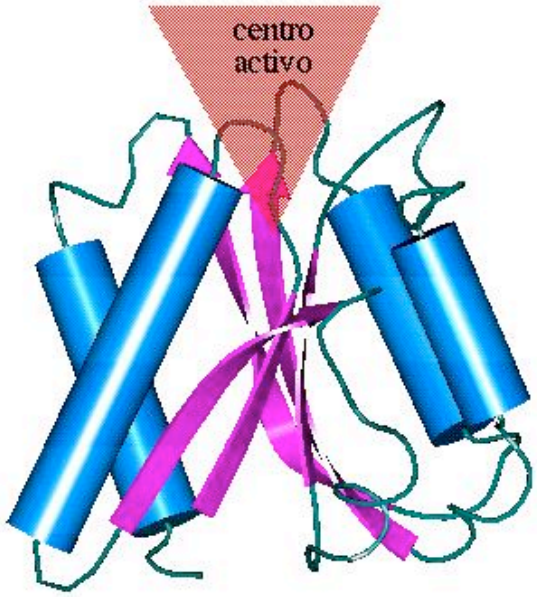
- Found in 1974 by Rossmann
- This domain appears in many proteins
- **Hypothesis**: many proteins have incorporated this domain by gene fusion, changing the activity by its interaction with mononucleotides

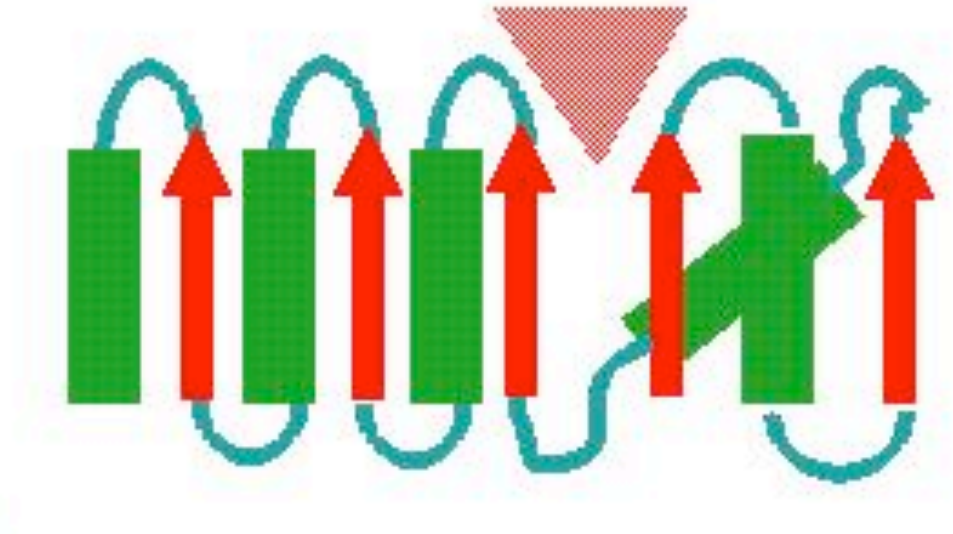
3. Soluble and globular proteins

6. Classical folds

Rossmann Fold

- The active site is located in a βα motif of the type NADP/FAD binding loop
- The location of the loop was often found in the change of orientation of the topological diagram. This is named **topological switch point**



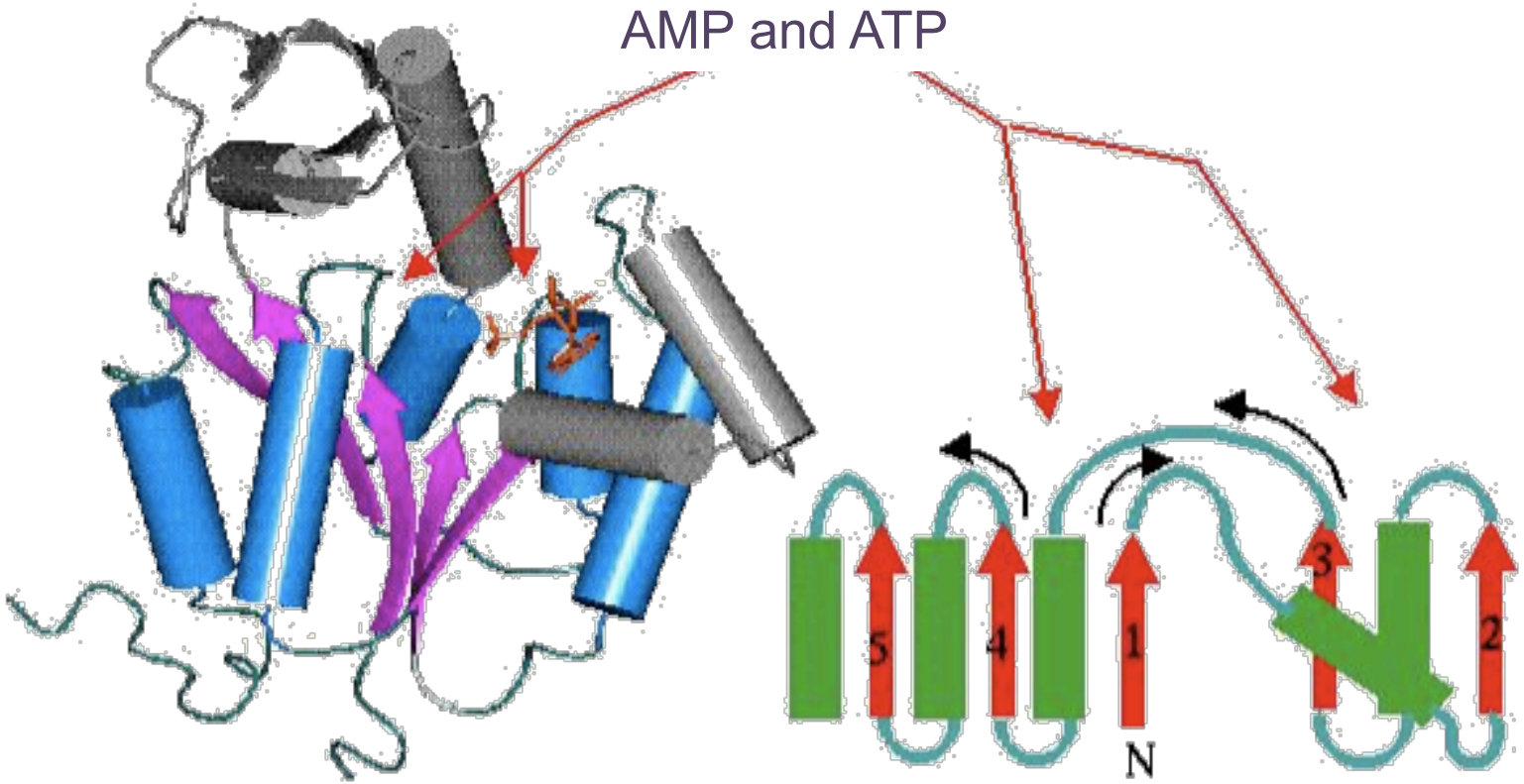


3. Soluble and globular proteins

6. Classical folds

Rossmann Fold

Active site has two binding sites for AMP and ATP

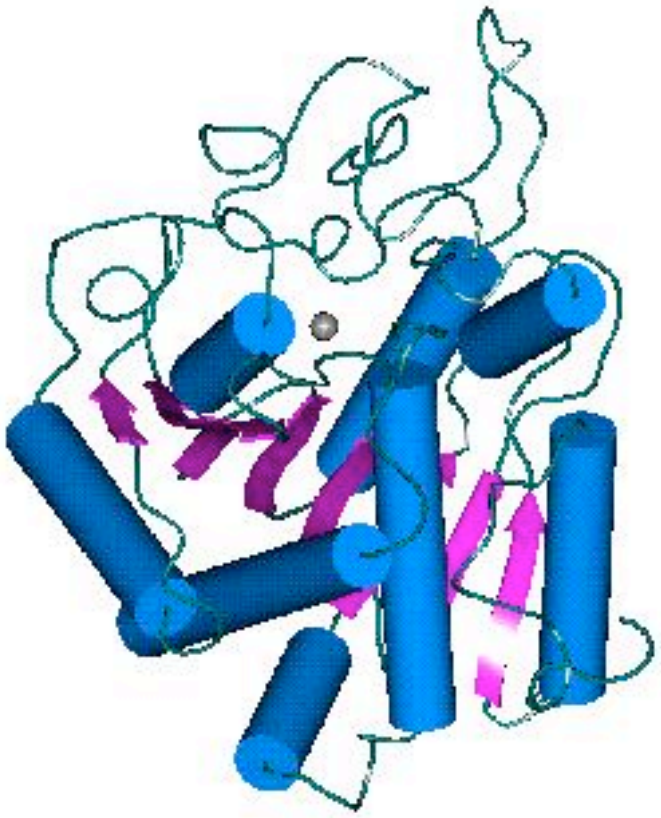


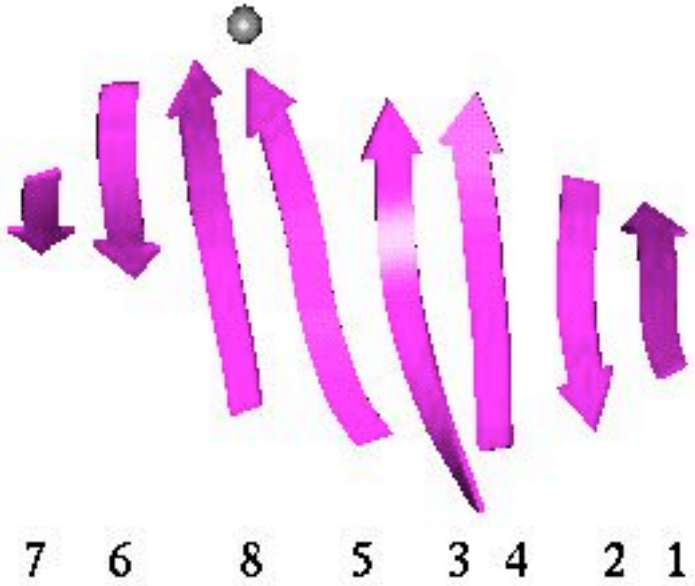
Adenilate Kinase

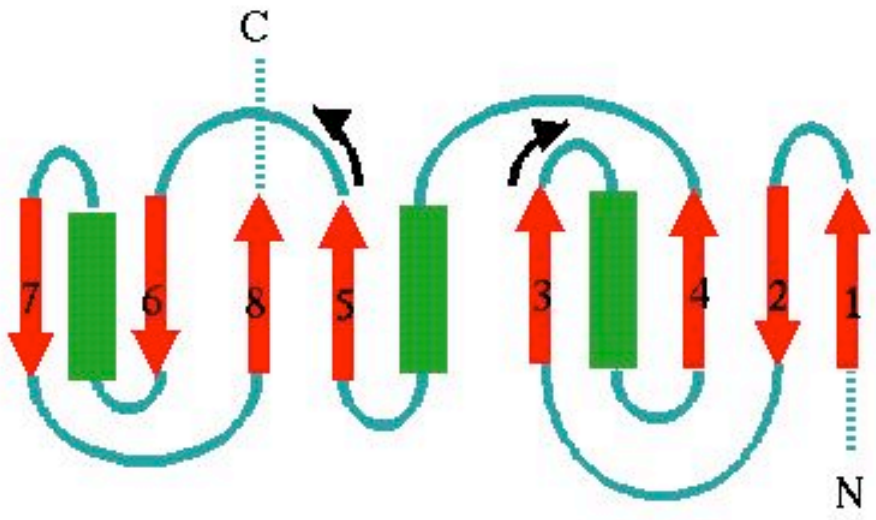
3. Soluble and globular proteins

6. Classical folds

Rossmann Fold





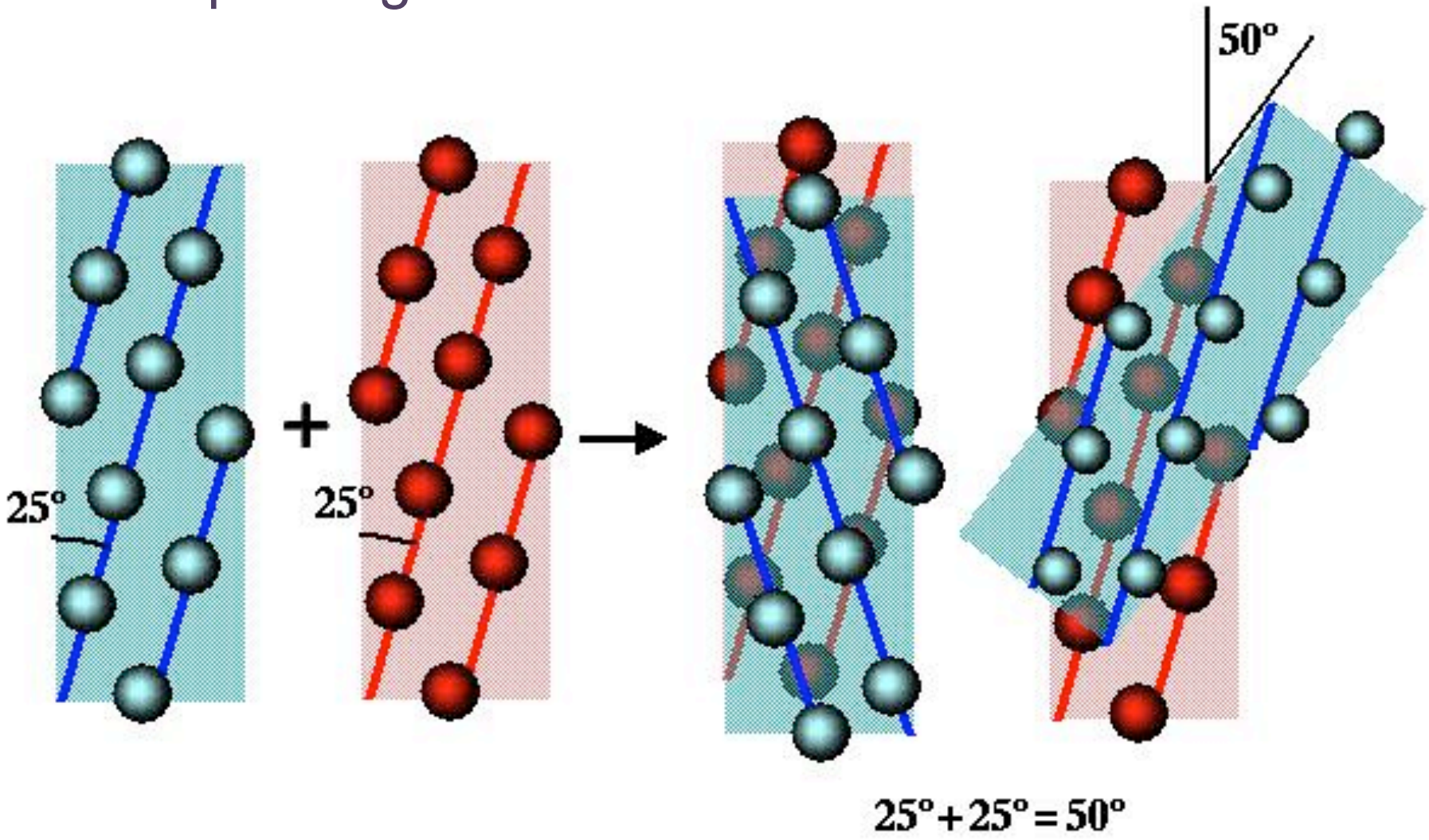


Carboxypeptidase

3. Soluble and globular proteins

7. All α

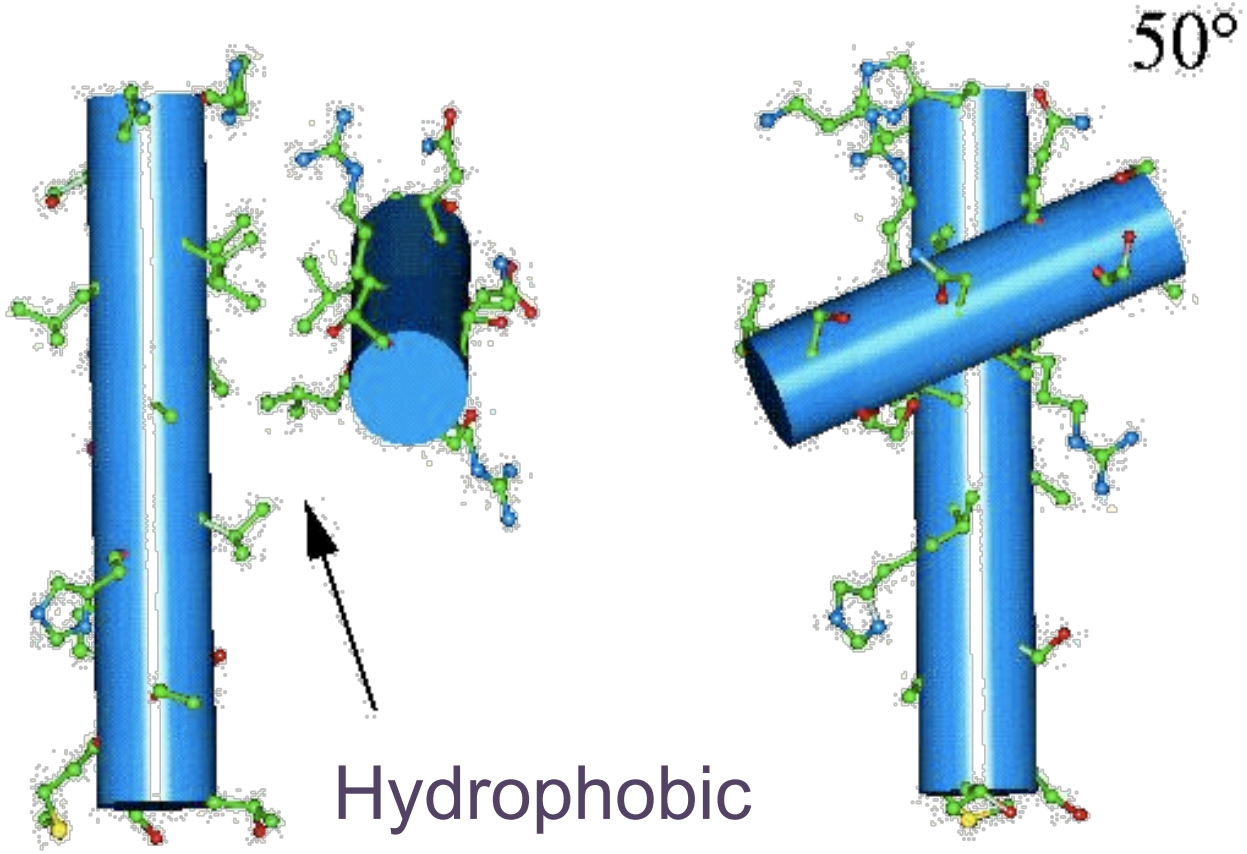
α-helix packing of two helices



3. Soluble and globular proteins

7. All α

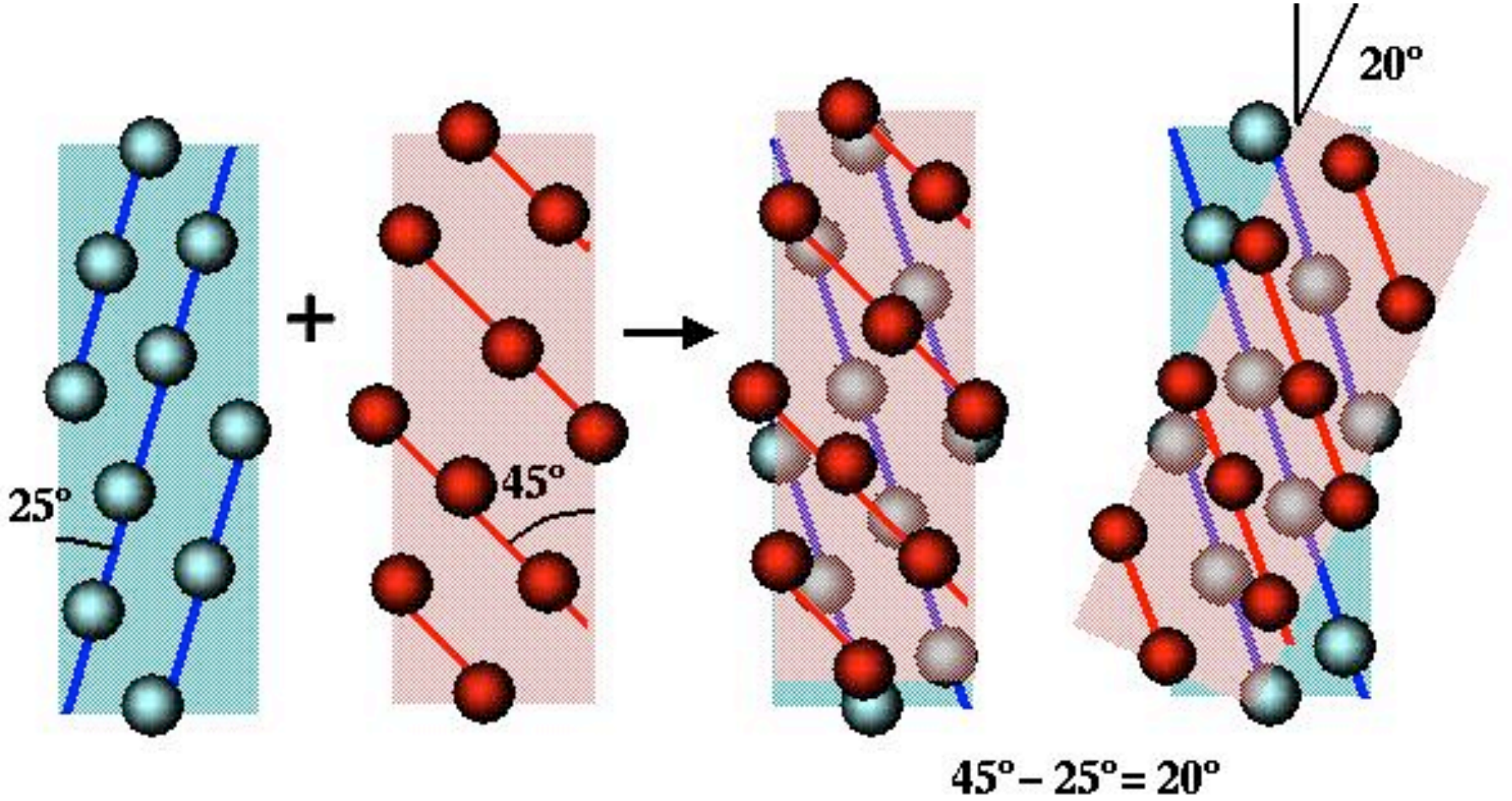
α-helix packing of two helices



3. Soluble and globular proteins

7. All α

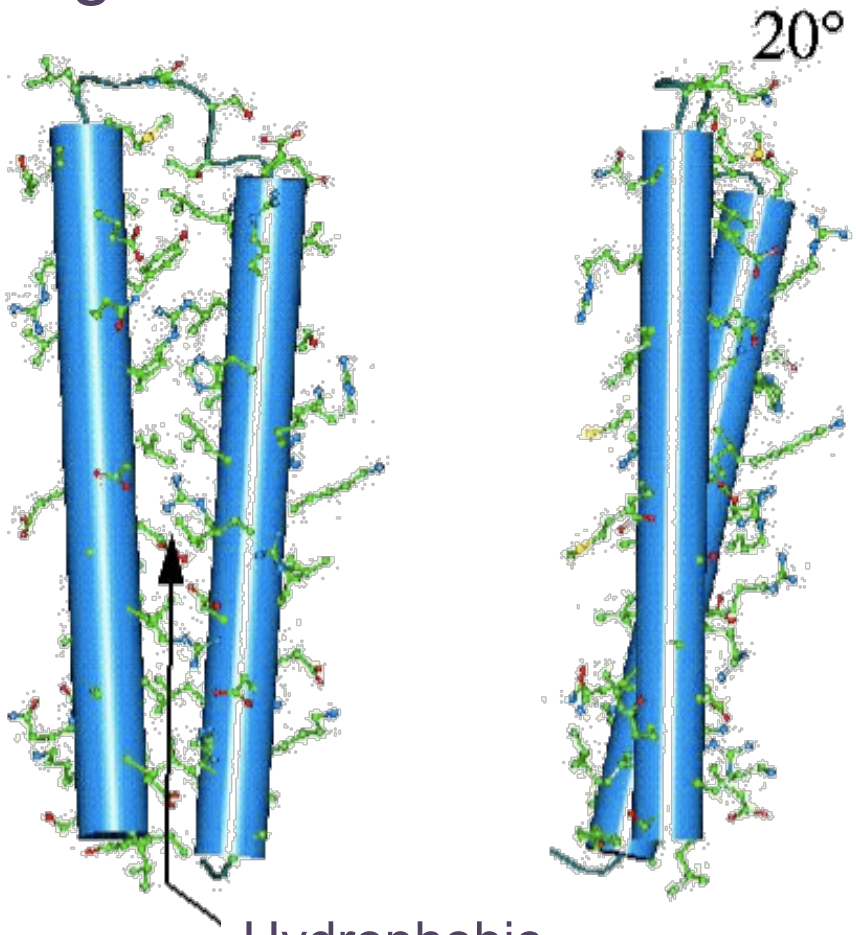
α-helix packing of two helices



3. Soluble and globular proteins

7. All α

α-helix packing of two helices



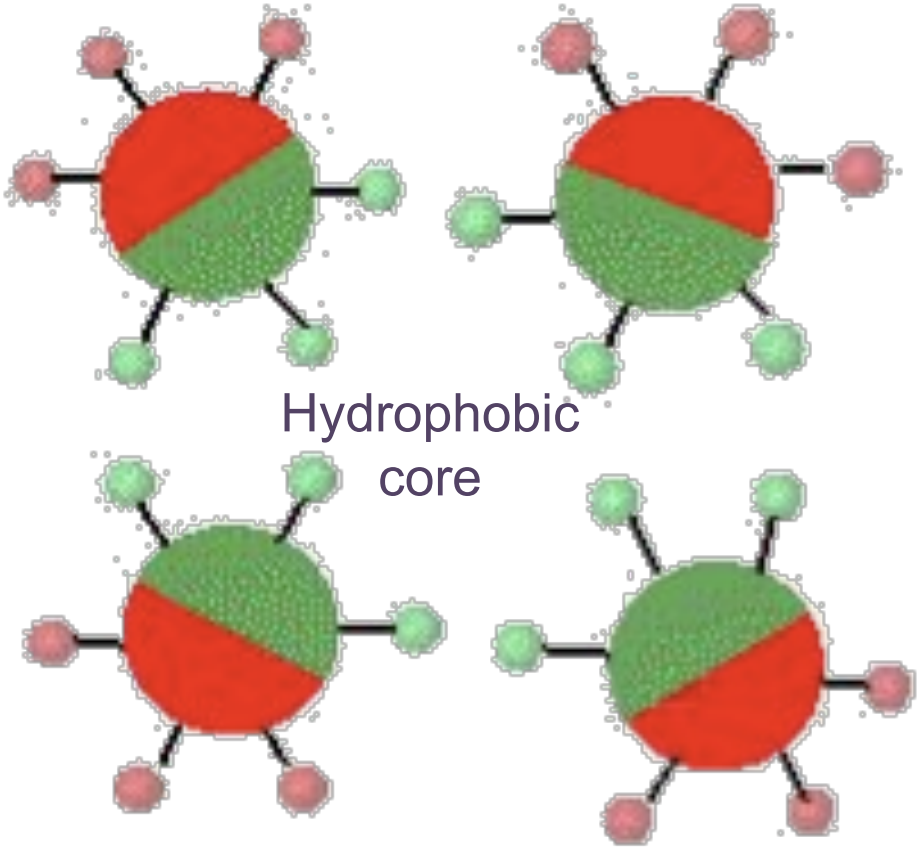
3. Soluble and globular proteins

7. All α

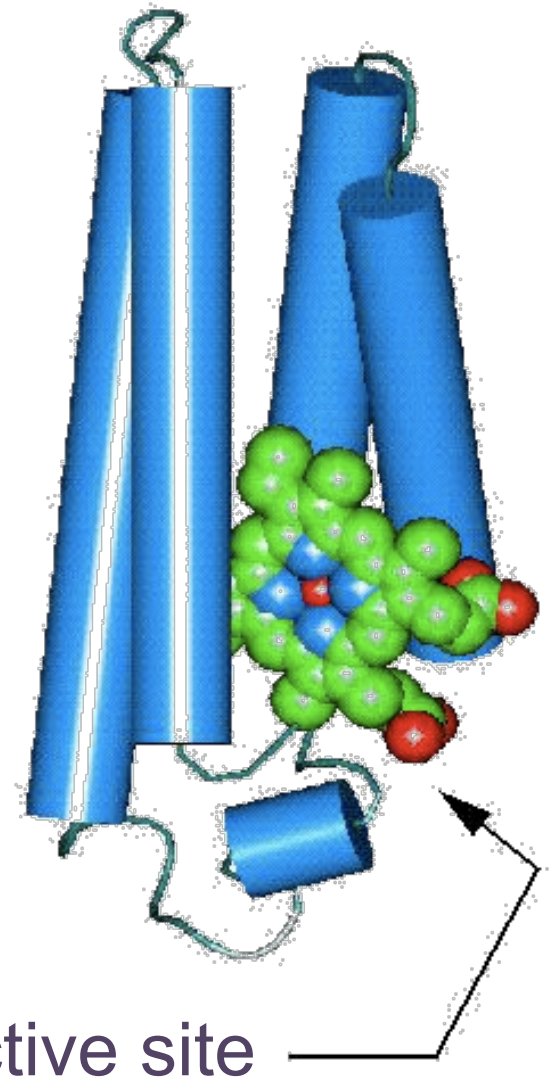
4 helix bundle

Polar residues

Non-polar residues



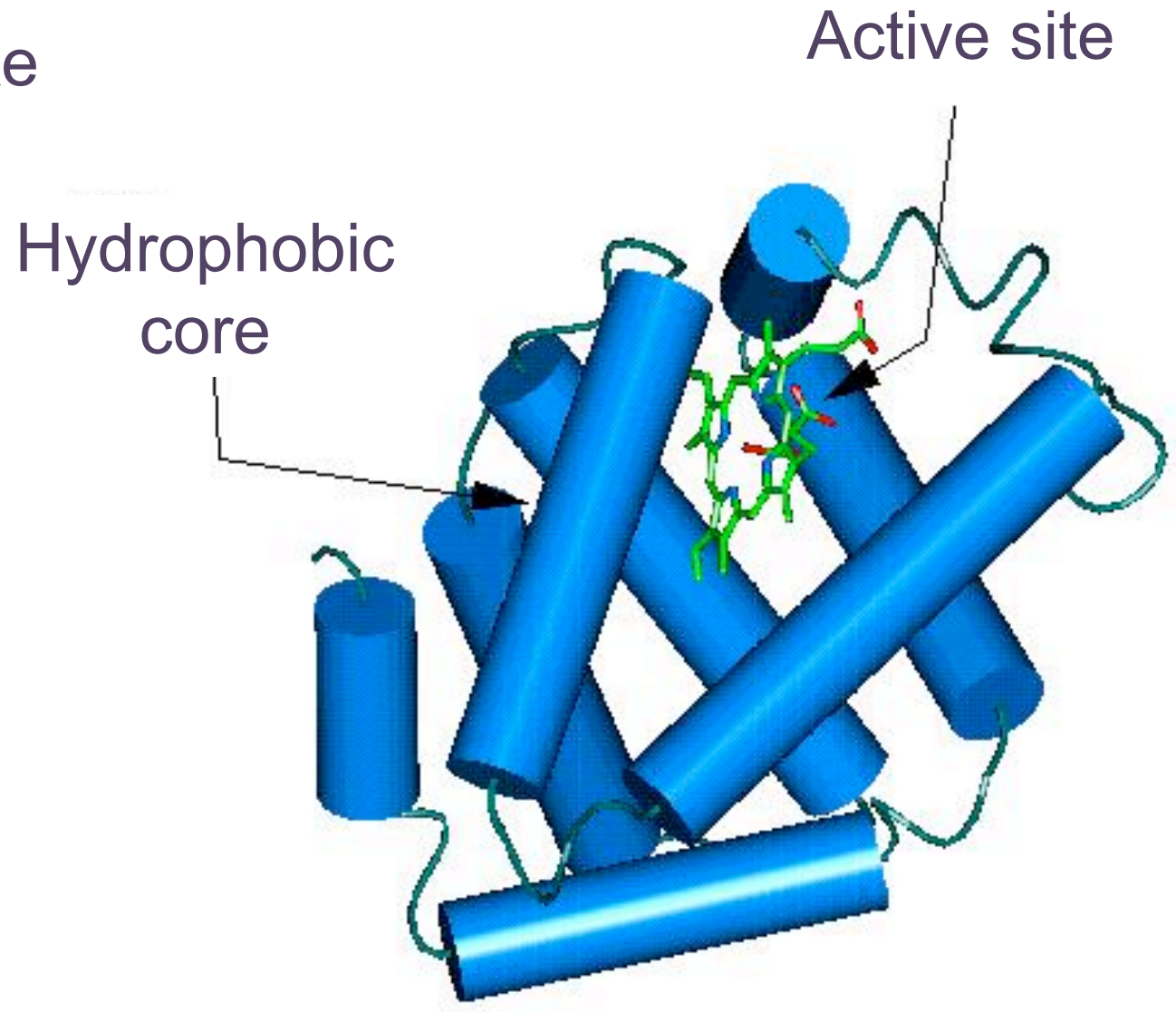
Cytochrome b_{562}



3. Soluble and globular proteins

7. All α

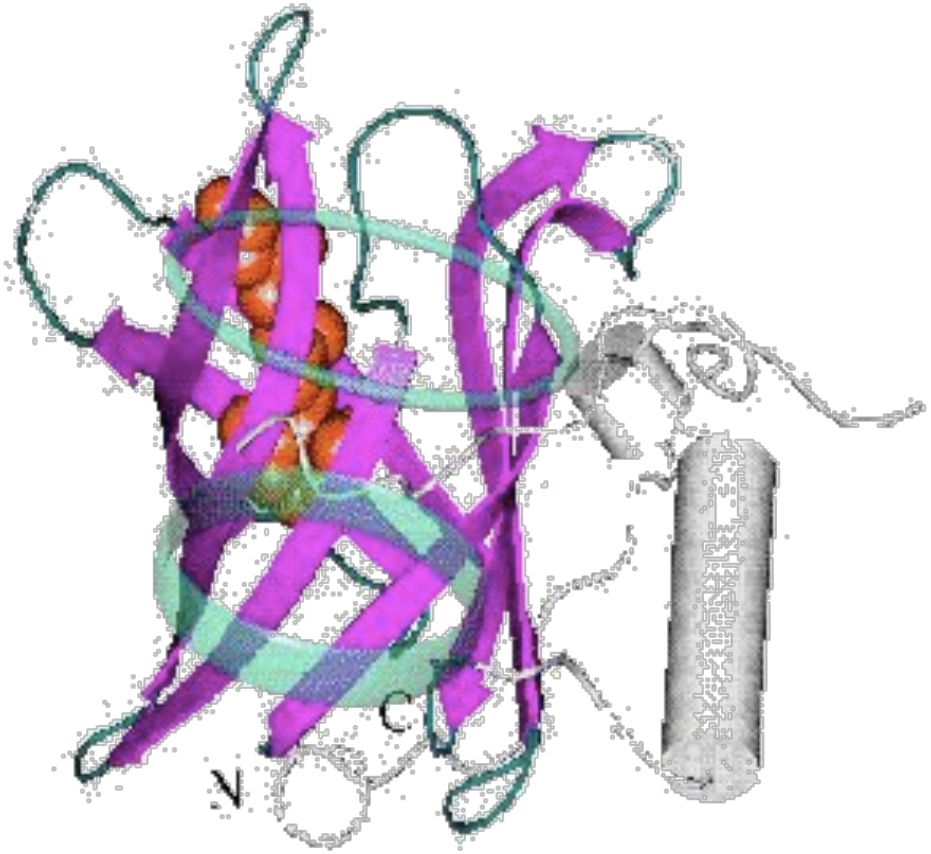
Globin-like



3. Soluble and globular proteins

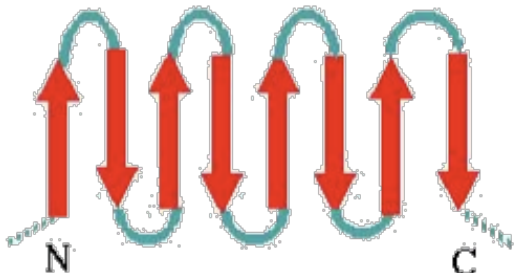
8. All β

Up & Down β−barrel

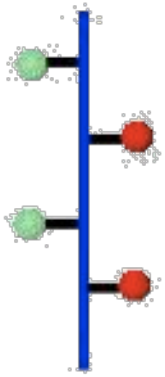


Retinol binding protein is a carrier of retinol (hydrophobic)

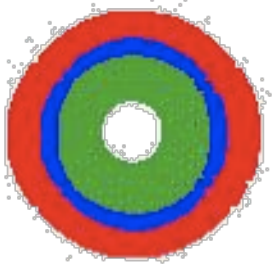
Topologic diagram

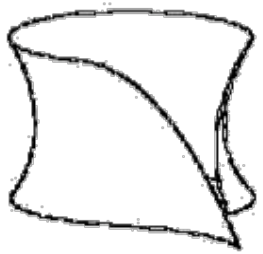


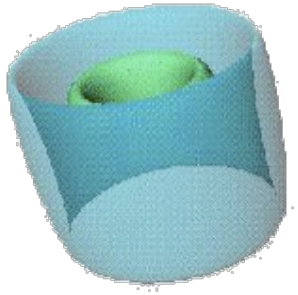
β-meanders are the main super-secondary structure



Hydrophobic core



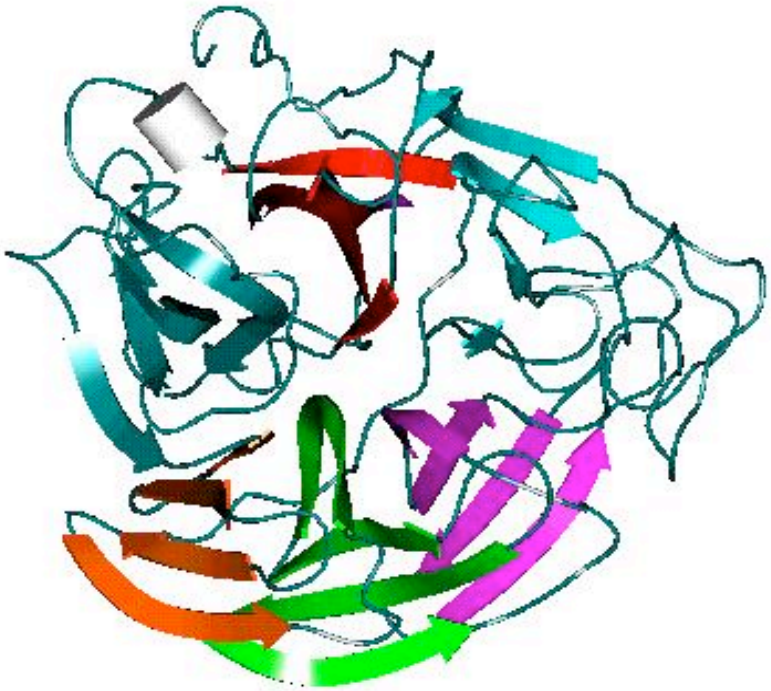




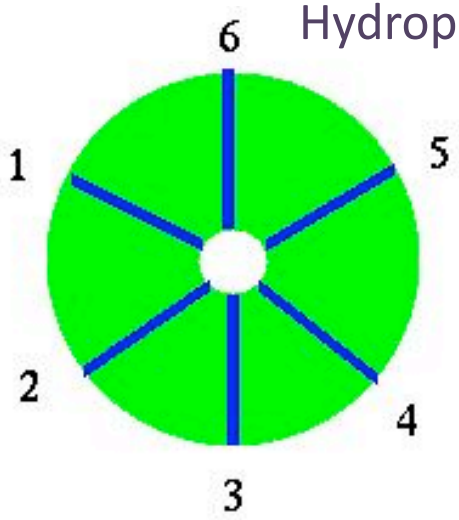
3. Soluble and globular proteins

8. All β

β−propeller (super barrel)



formed by N β-meanders placed as blades



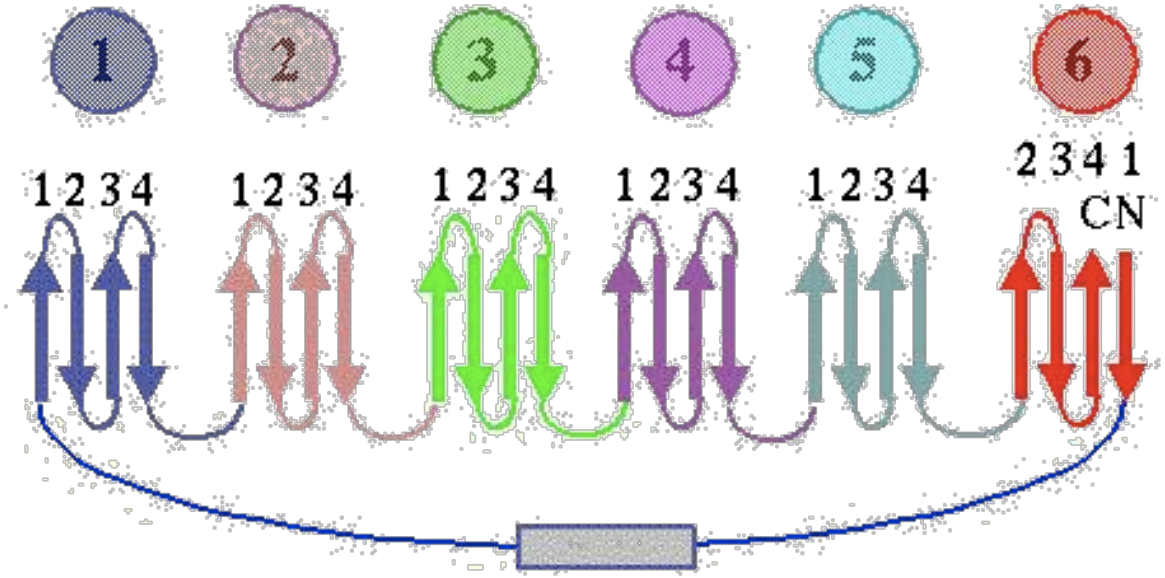
Hydrophobic core

3. Soluble and globular proteins

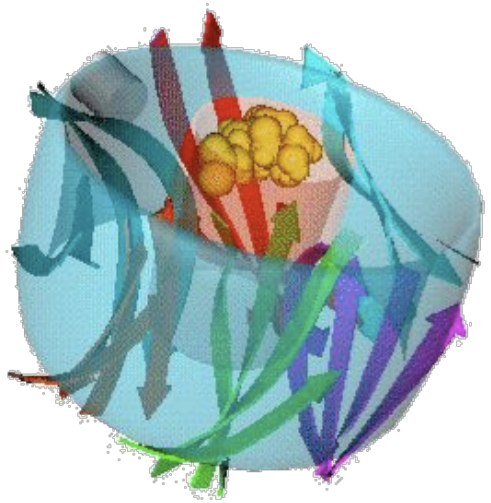
8. All β

β−propeller (super barrel)

Neuraminidase is a 6 β−meander β−propeller



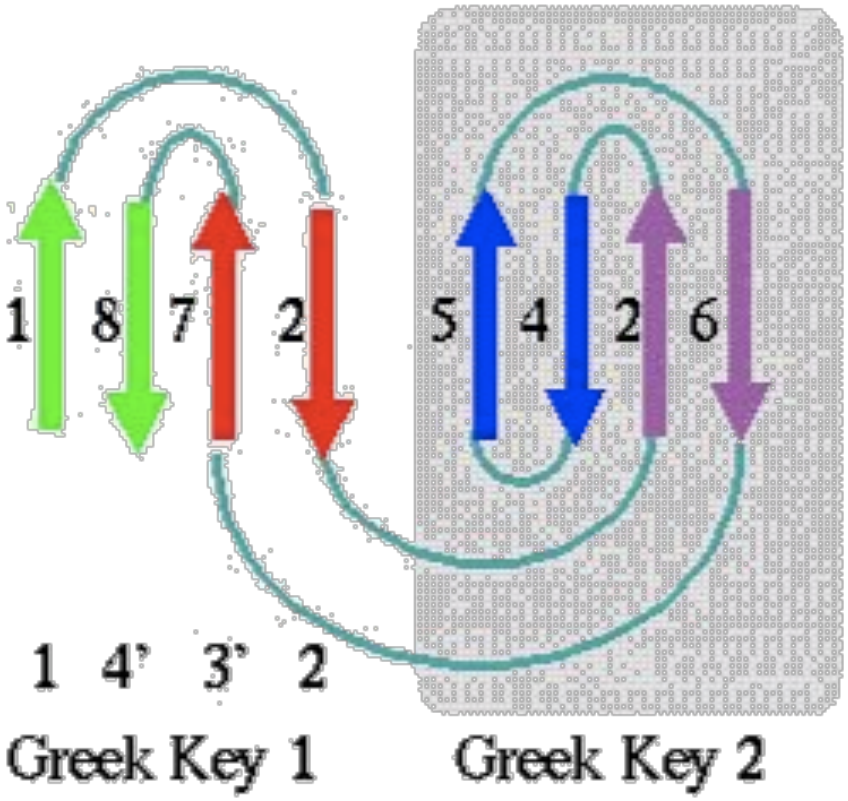
Active site

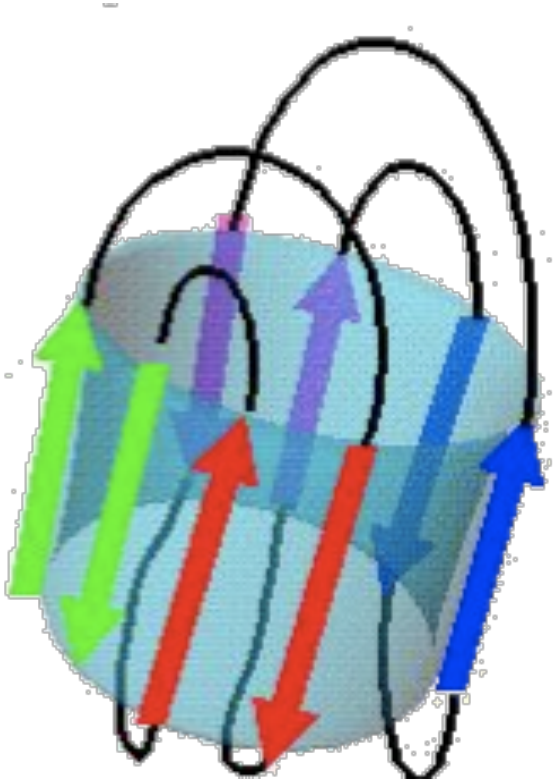


3. Soluble and globular proteins

8. All β

Greek key Jelly-Roll



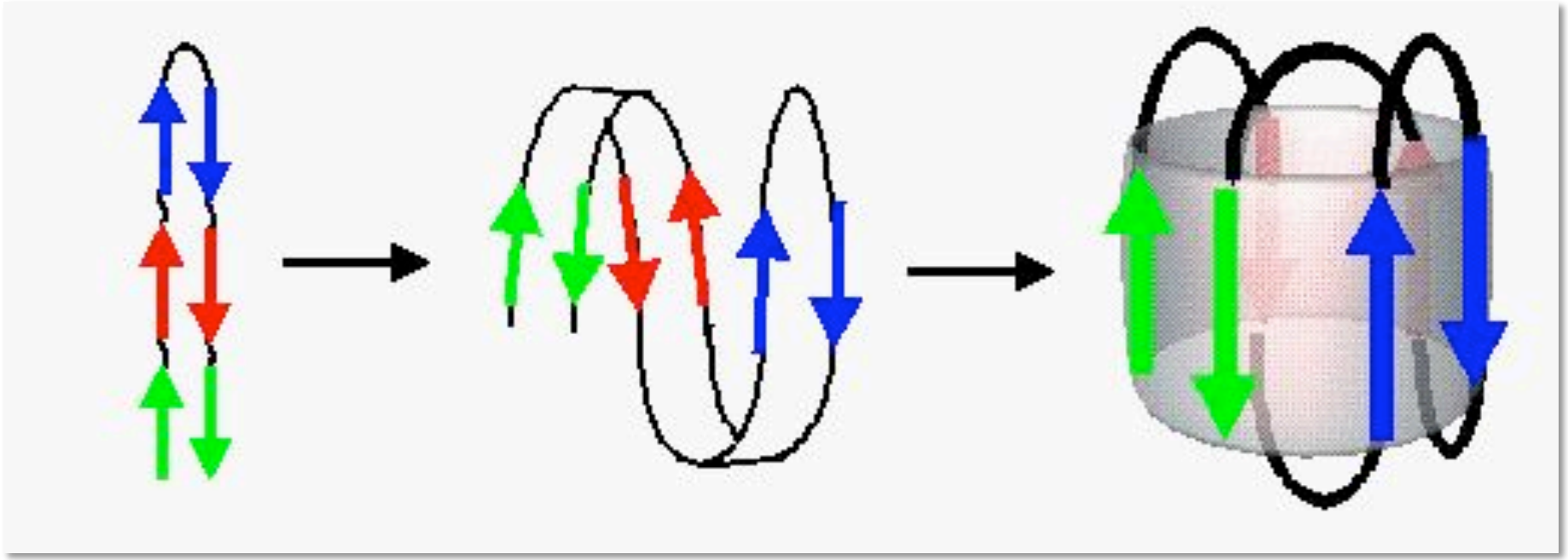


3. Soluble and globular proteins

8. All β

Greek key Jelly-Roll

Proposed mechanism of folding



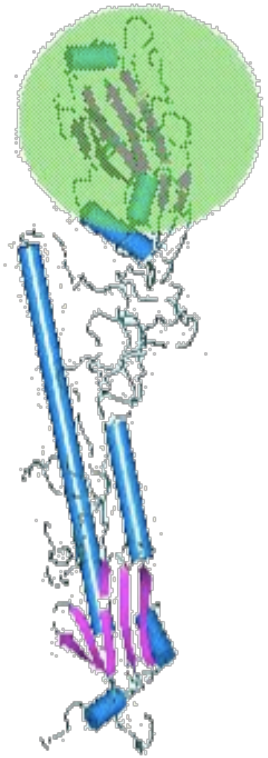
3. Soluble and globular proteins

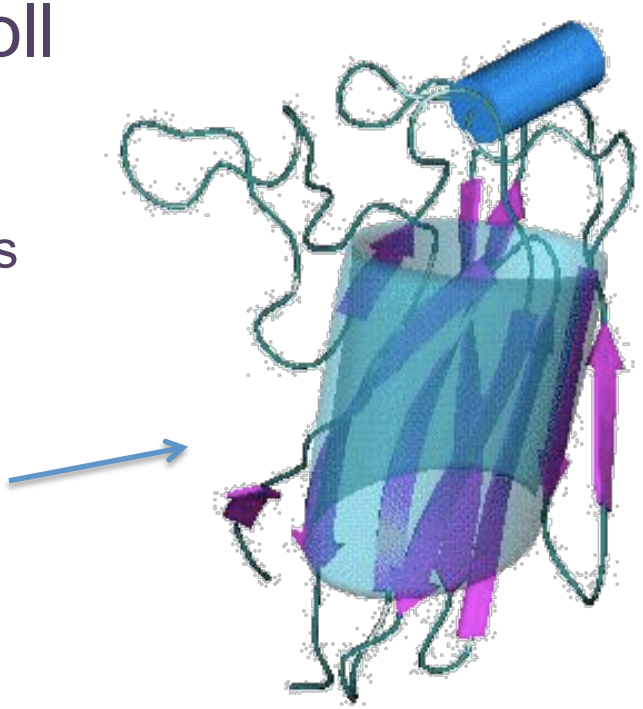
8. All β

Greek key Jelly-Roll

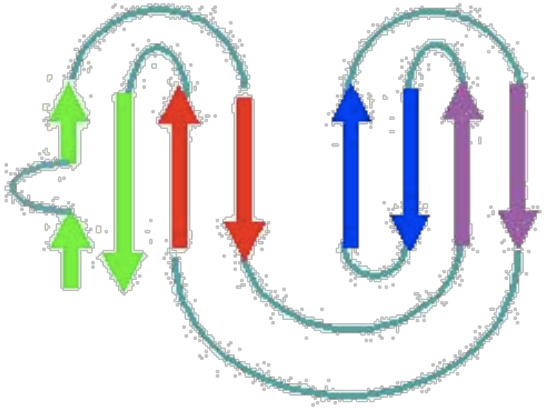
Hemaglutinin

Trimer of the capsid of the virus of flu. Membrane fusion





Jelly-roll barrel

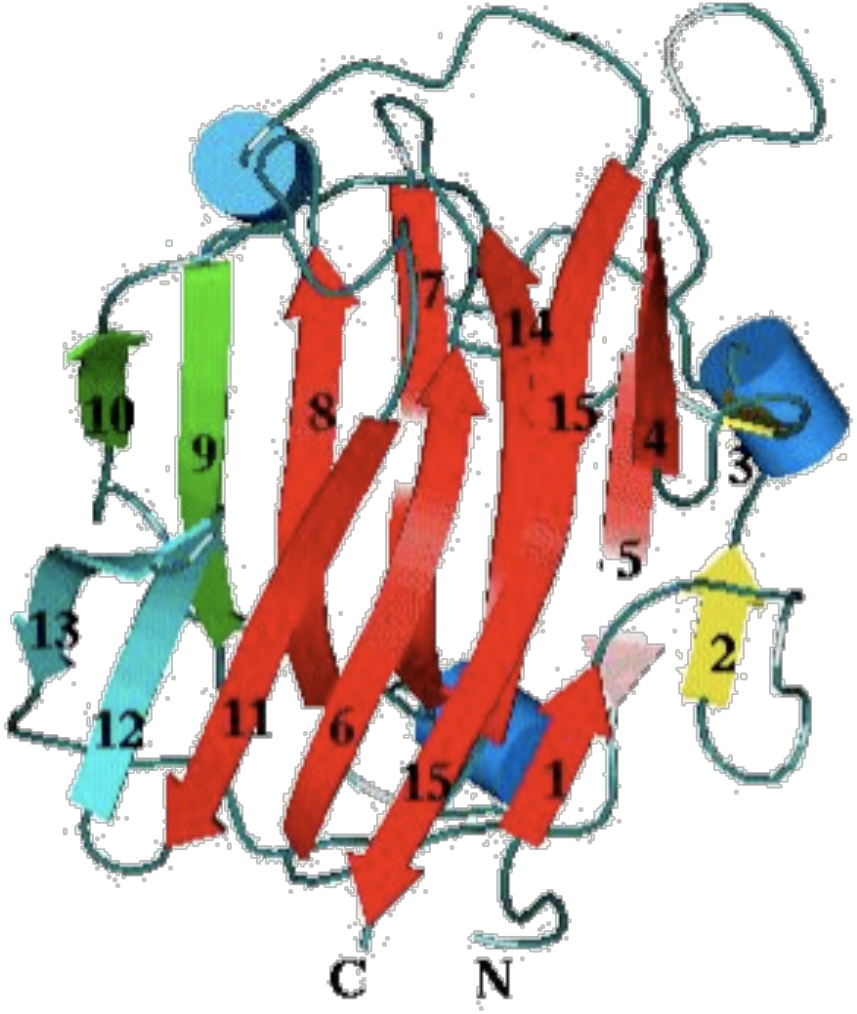


Topologic diagram

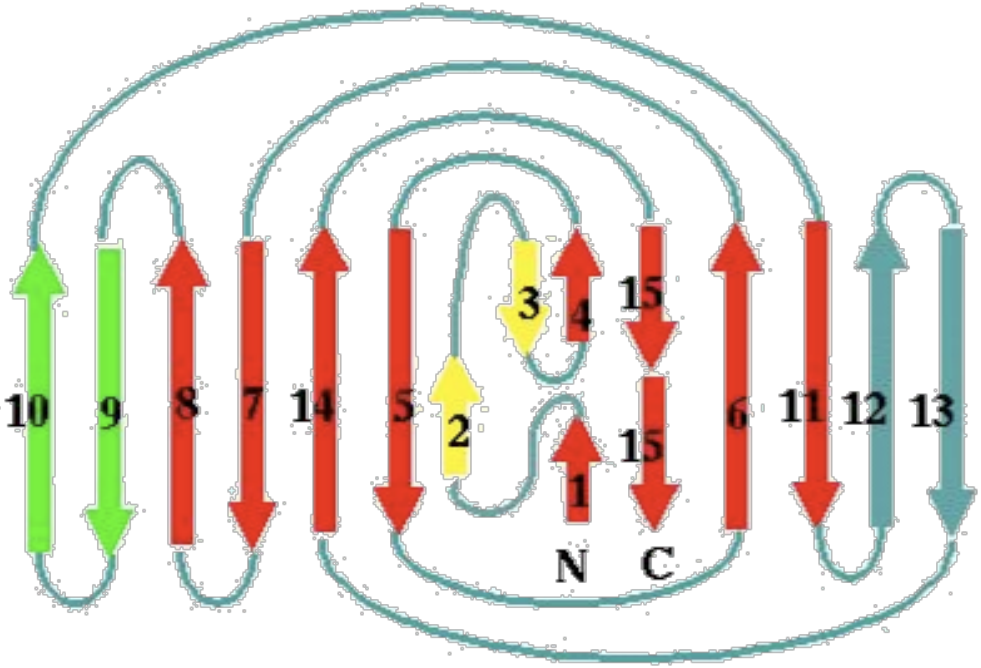
3. Soluble and globular proteins

8. All β

Greek key Jelly-Roll



Glucanase

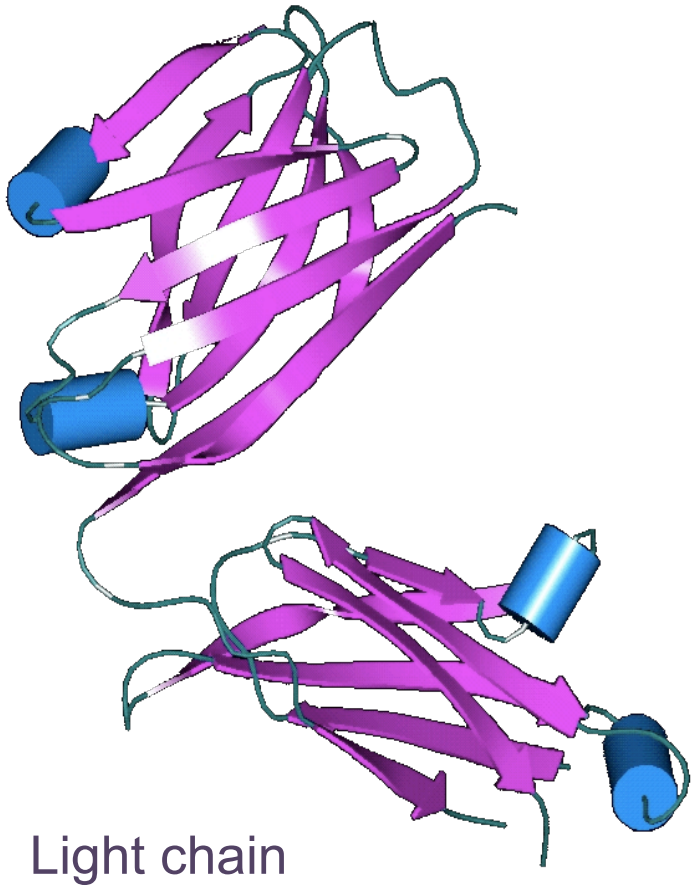


Topologic diagram

3. Soluble and globular proteins

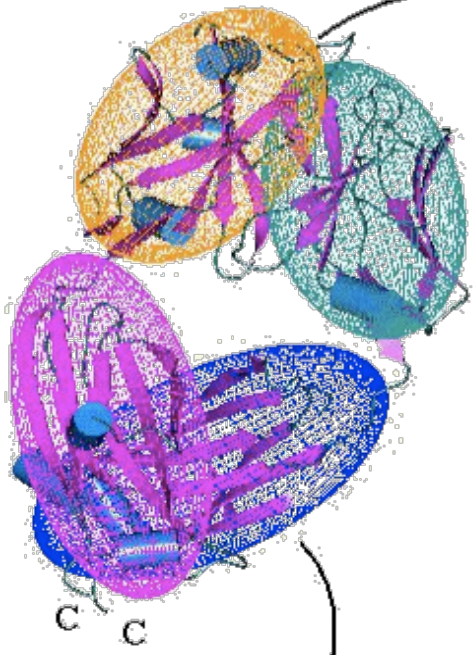
8. All β

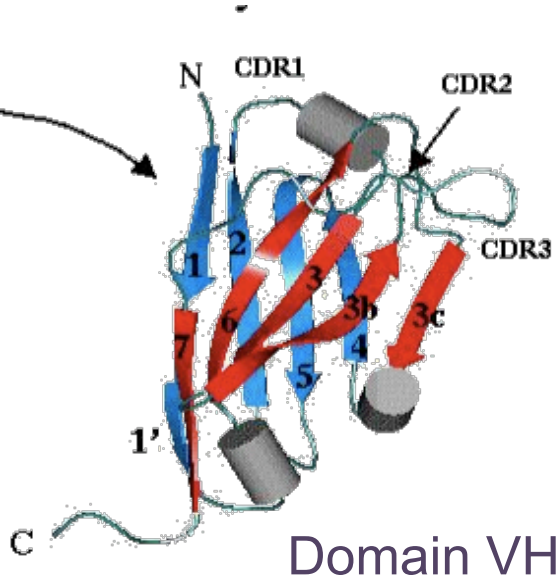
β-sandwich
Immunoglobulin-like



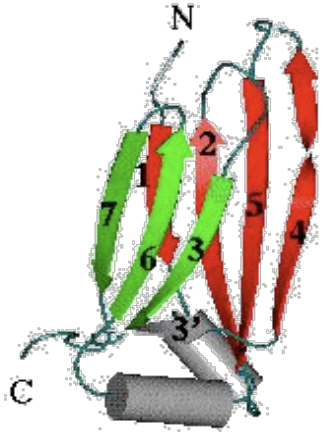
Light chain

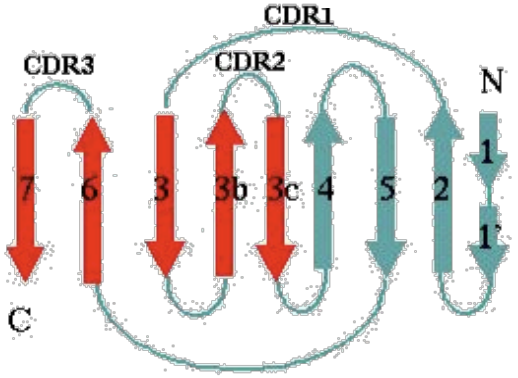
Fragment Fab

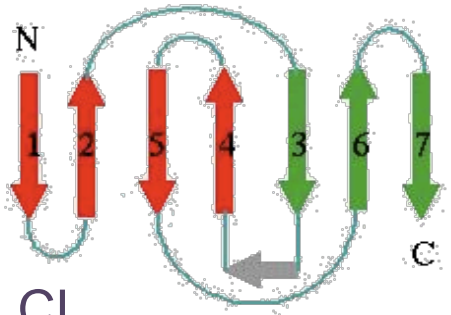




Domain VH





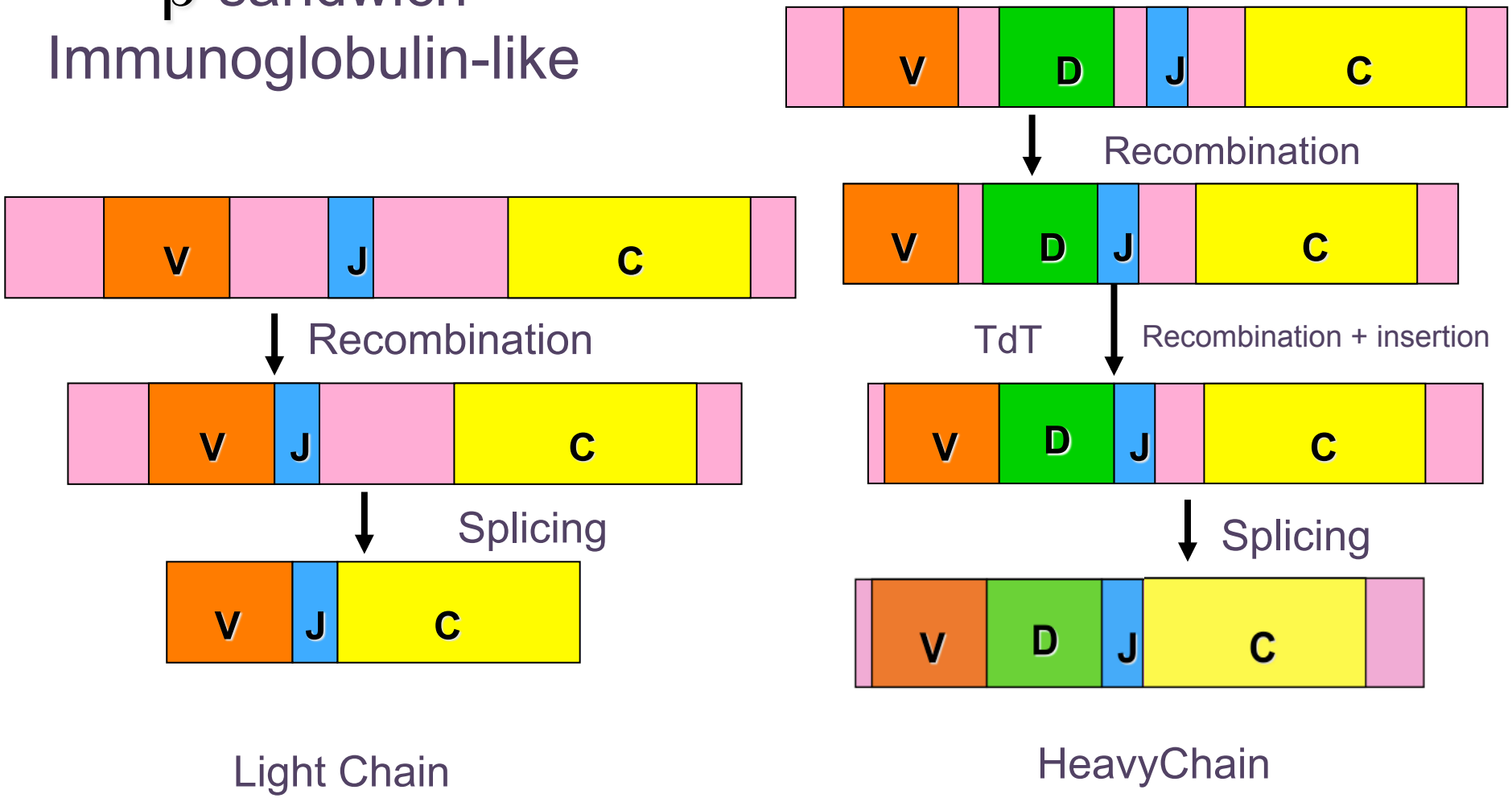


Domain CL

3. Soluble and globular proteins

8. All β

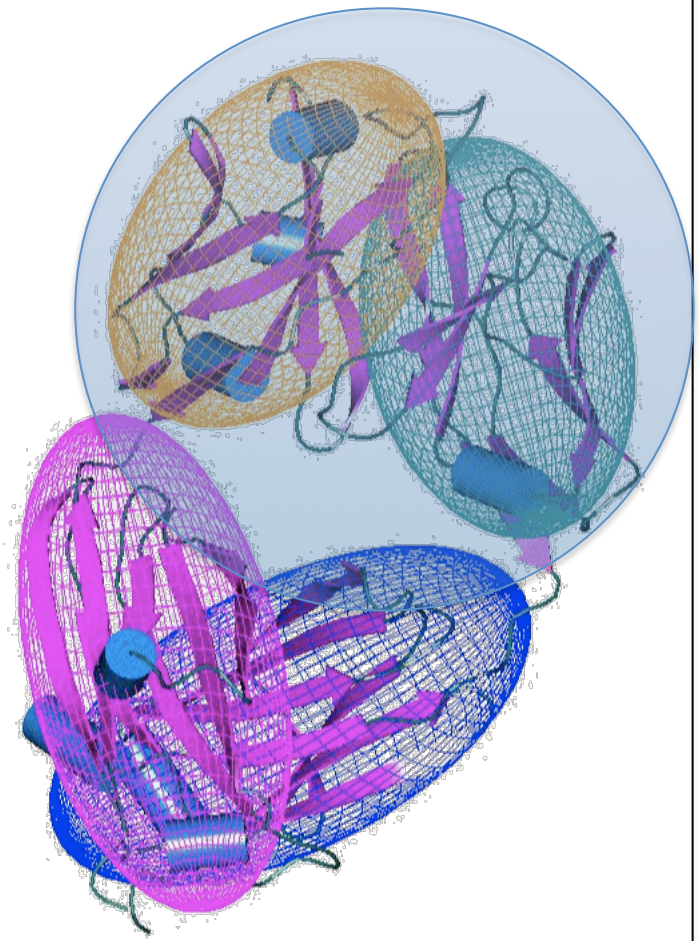
β-sandwich
Immunoglobulin-like



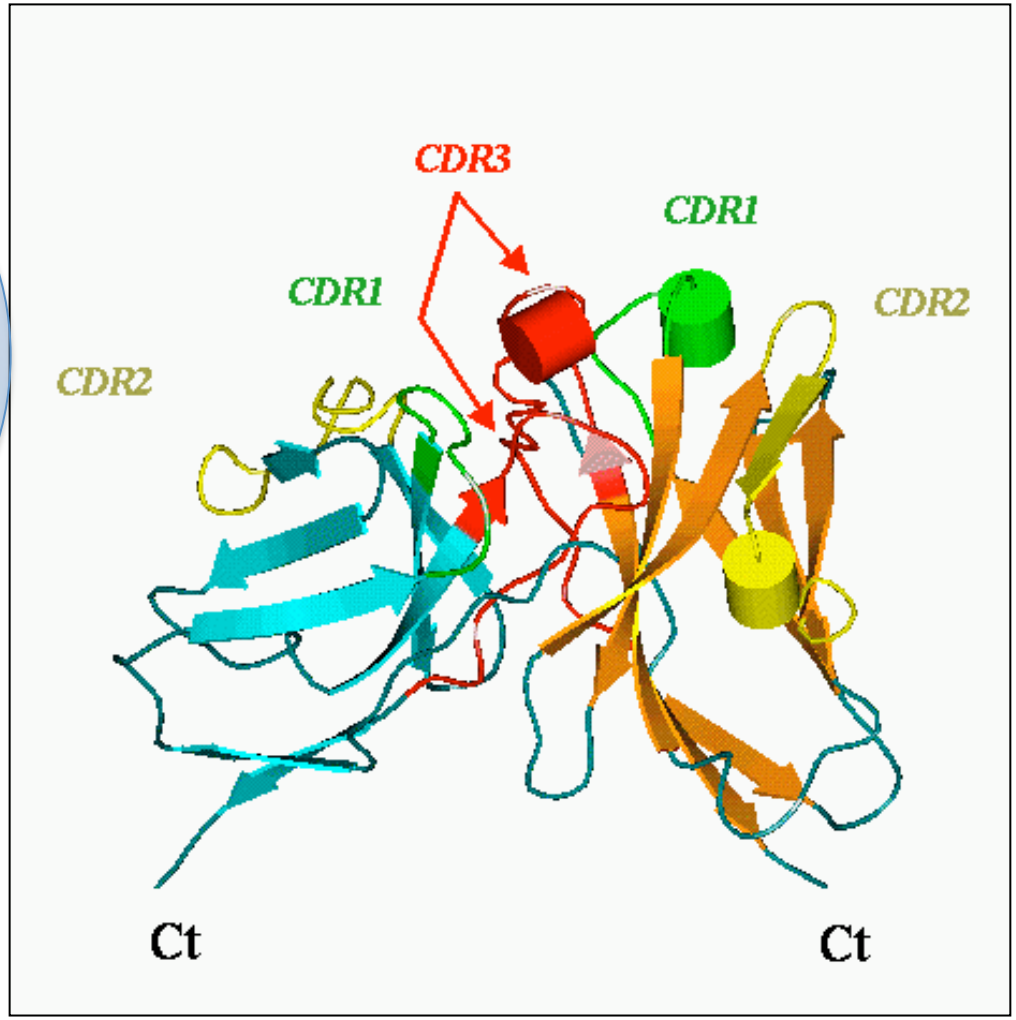
3. Soluble and globular proteins

8. All β

β-sandwich
Immunoglobulin-like



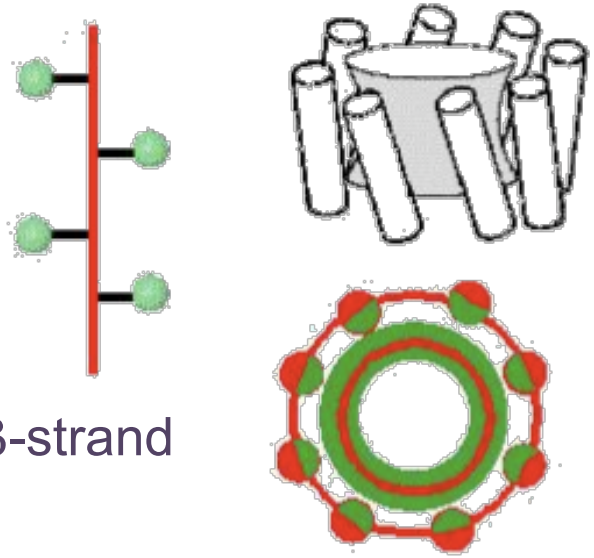
CDR loops



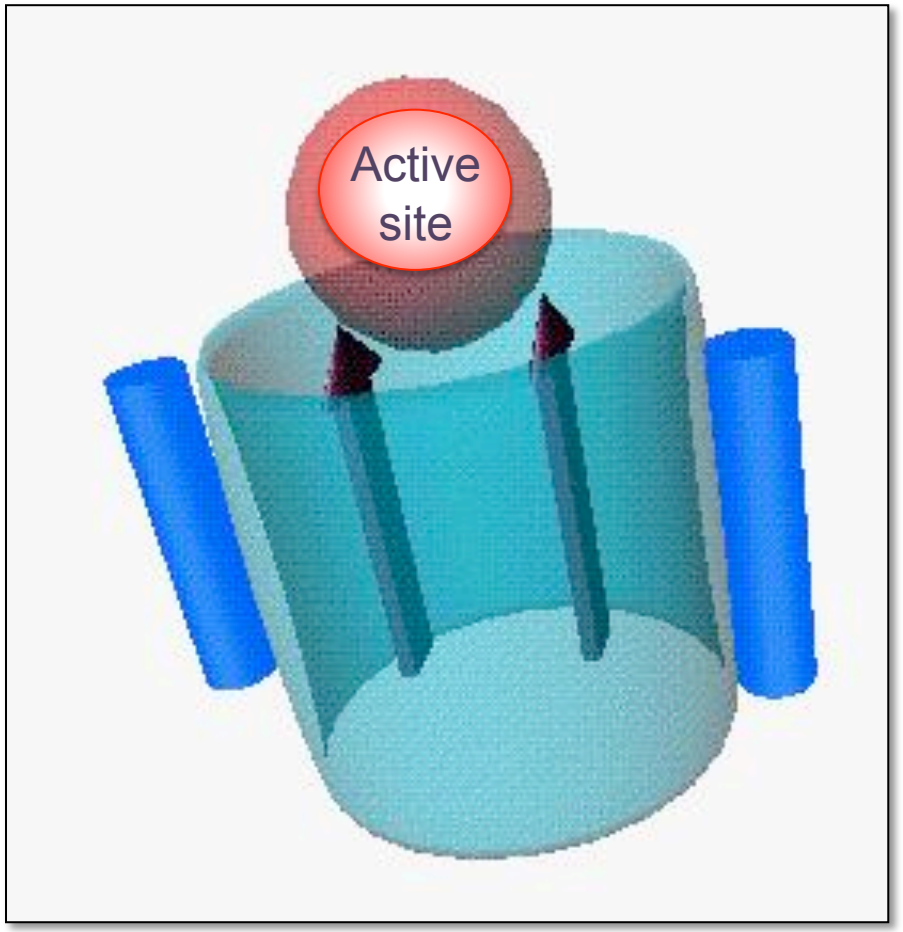
3. Soluble and globular proteins

9. α/β

TIM Barrel



β-strand

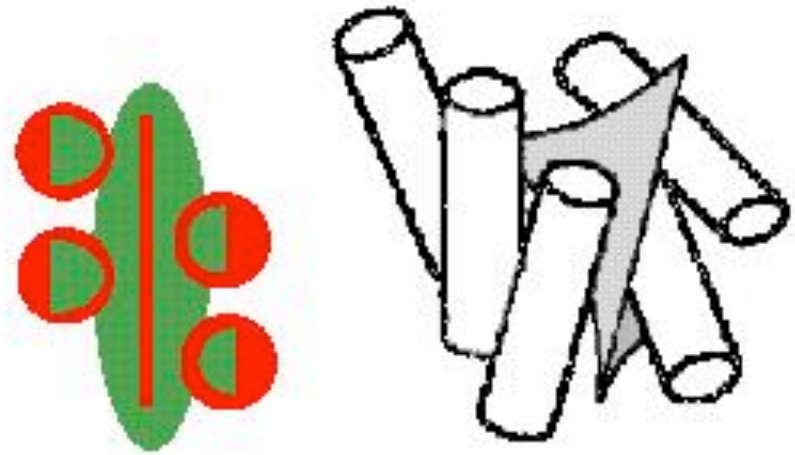


Highly stable structure due to the double hydrophobic core

3. Soluble and globular proteins

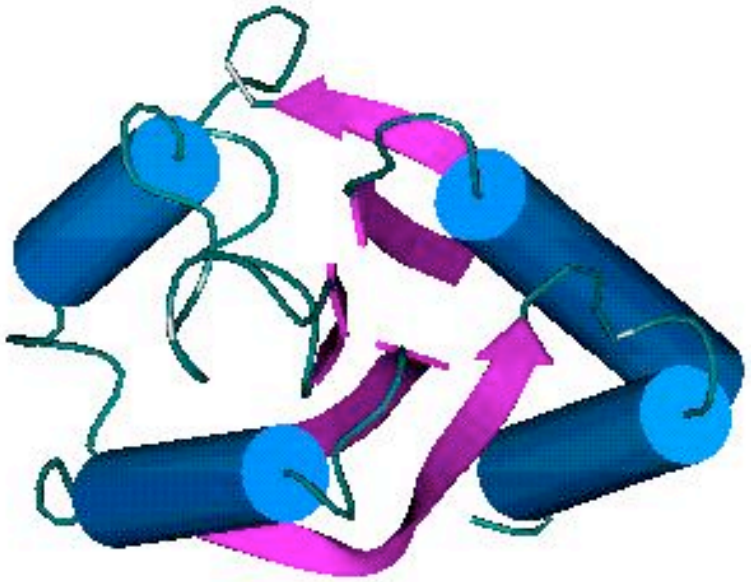
9. α/β

Rossmann Fold



Open sheet

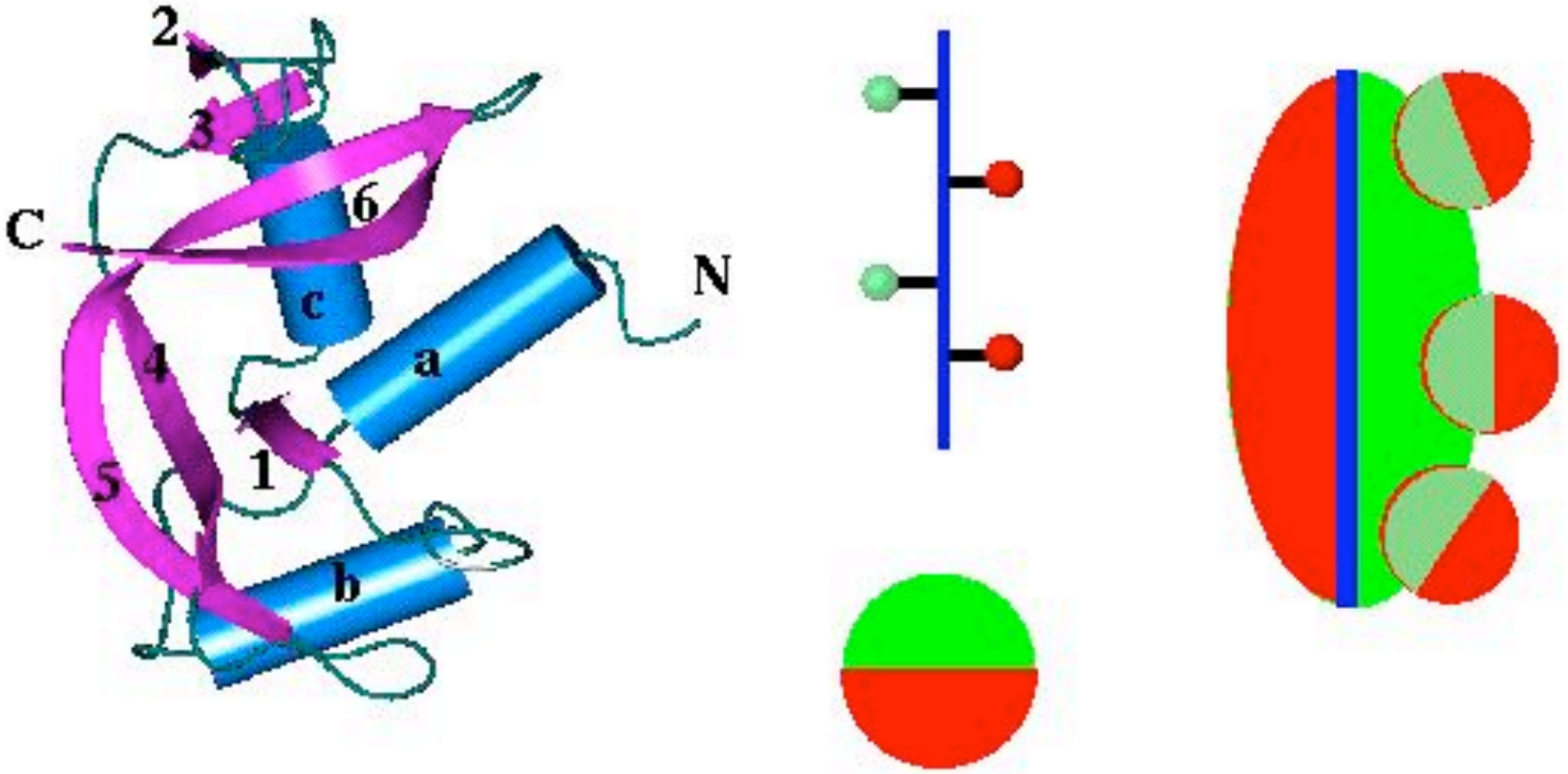
Flavodoxin



3. Soluble and globular proteins

10. $\alpha+\beta$

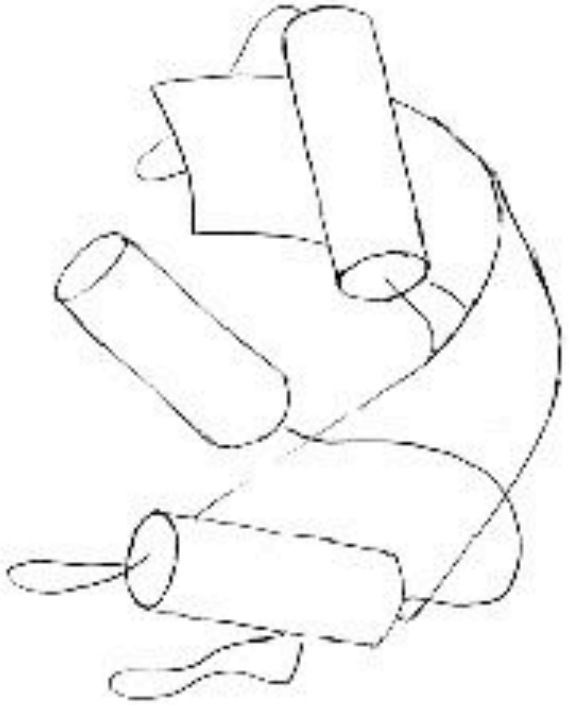
Ribonuclease



3. Soluble and globular proteins

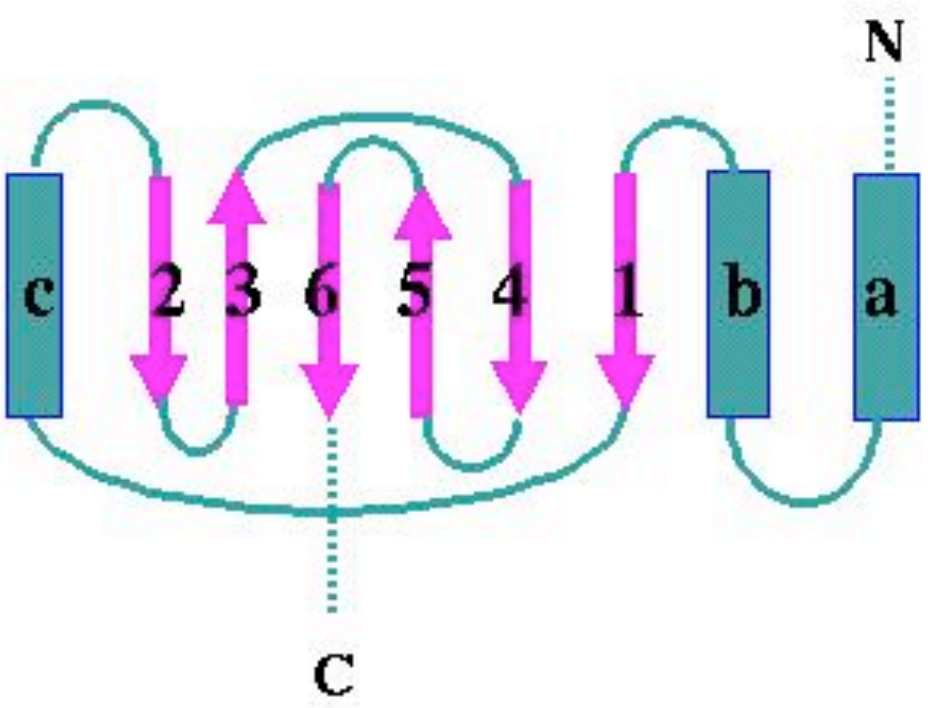
10. α+β

Ribonuclease



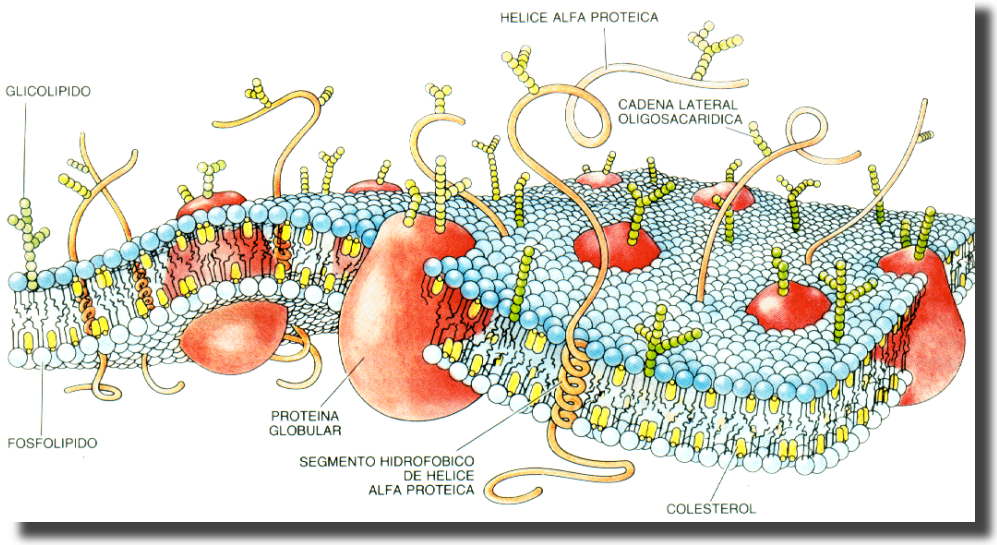
RNase A-like fold

α+β class segregates α and β regular secondary structures in 3D and in the topologic diagram



4. Globular membrane proteins

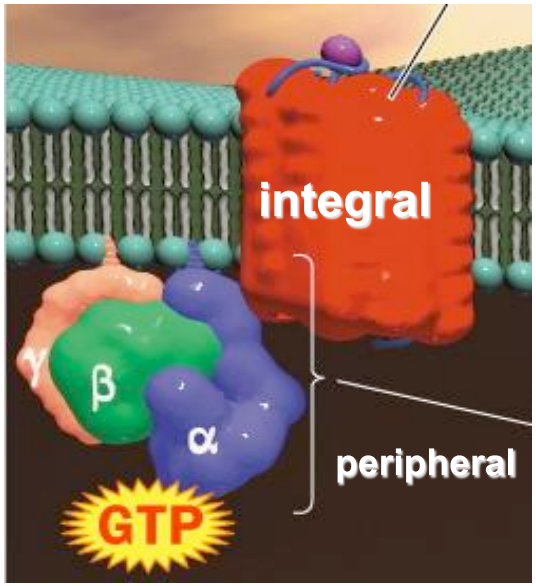
1. Definition

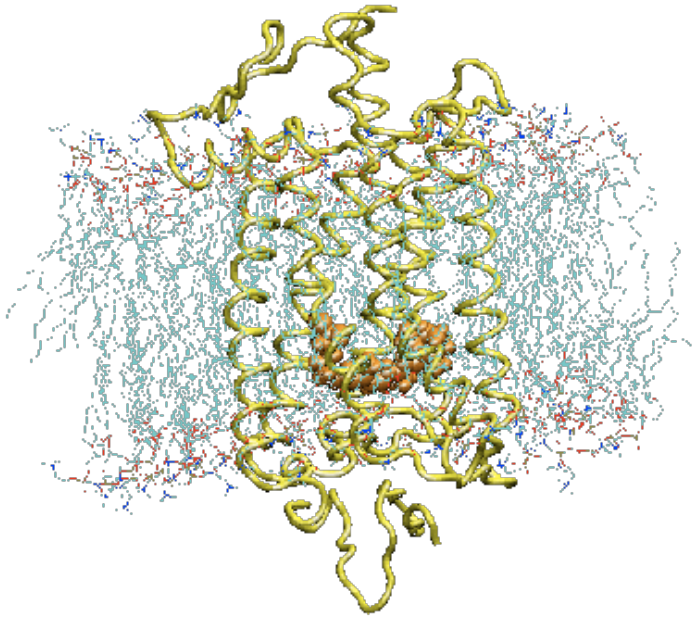


- Membrane proteins are protons bound to the membrane
- Membrane Proteins can be of two types: peripheral and integral
- Integral membrane proteins are inserted in the membrane, where the main bulk of the domain is buried inside the membrane
- Peripheral membrane proteins are only linked by few residues to the membrane
- The residues of the integral proteins are mostly non-polar, changing the rules of hydrophobicity patterns described for soluble domains.

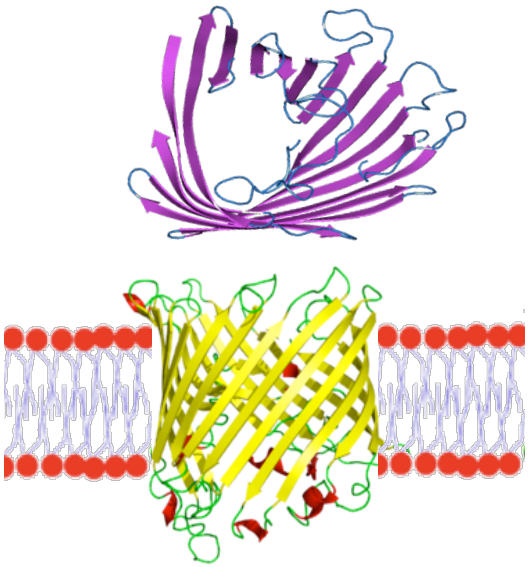
4. Globular membrane proteins
2. Classification
1. Integral

Two types of integral proteins:
α-helix transmembrane
β-sheet transmembrane





α-helix transmembrane
Rhodopsine



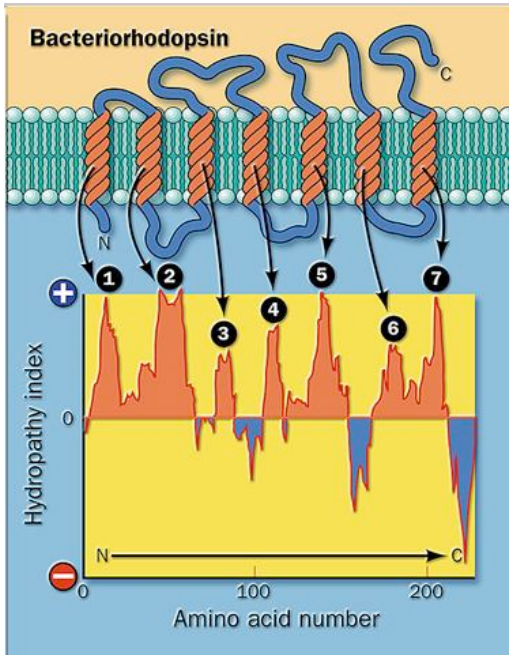
β-sheet transmembrane
Porin

4. Globular membrane proteins

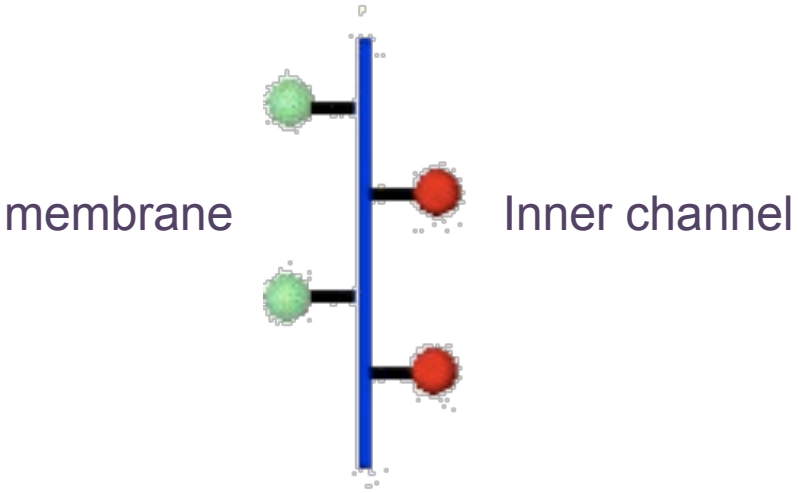
2. Classification

1. Integral

Hydrophobic properties



α-helix transmembrane
Rhodopsine



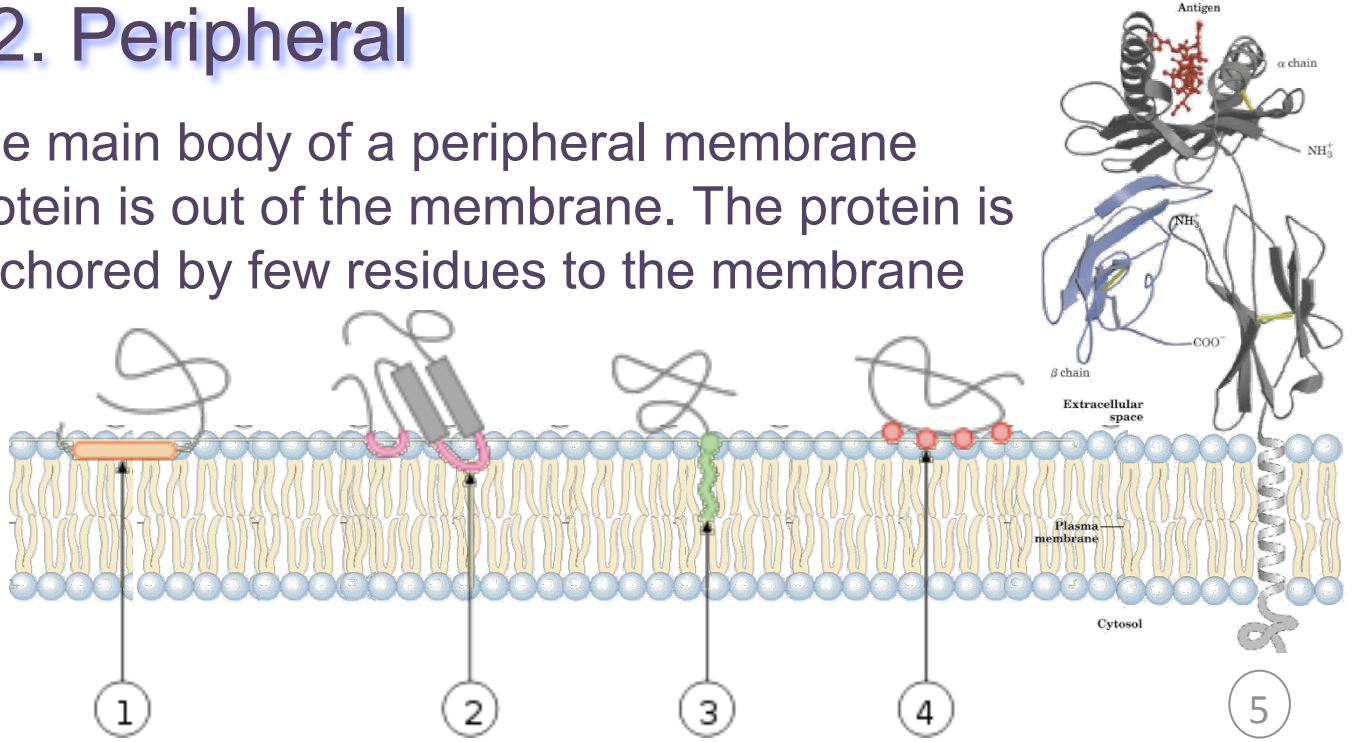
β-sheet transmembrane
Porin

4. Globular membrane proteins

2. Classification

2. Peripheral

The main body of a peripheral membrane protein is out of the membrane. The protein is anchored by few residues to the membrane



Example of different types of interaction between membrane proteins and the cell membrane: 1. interaction by an amphipathic helix parallel to the membrane plane 2. interaction by a hydrophobic loop 3. interaction by a covalently bound membrane lipid (lipidation, i.e. myristilation) 4. electrostatic or ioinic interaction with membrane lipids (e.g. through a calcium ion) 5. hydrophobic helix transmembrane.

4. Globular membrane proteins

2. Classification

2. Peripheral

Some proteins have two main bodies one at each side of the membrane, an in general its function is to transfer a biochemical signal from one side to the other