# Principles of Protein Structure and Conformational Space (part 2)

# Protein structure

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

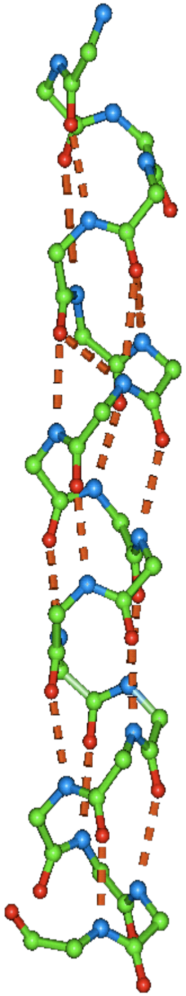
3. Soluble and globular proteins
    3. Supersecondary structure
        1. Definition
        2. α−α
        3. β−β
        4. β−α & α−β
        5. β−α−β
    4. Domains
        1. Definition
        2. Classification
        3. Function-Sequence-Structure
        4. SCOP
        5. CATH

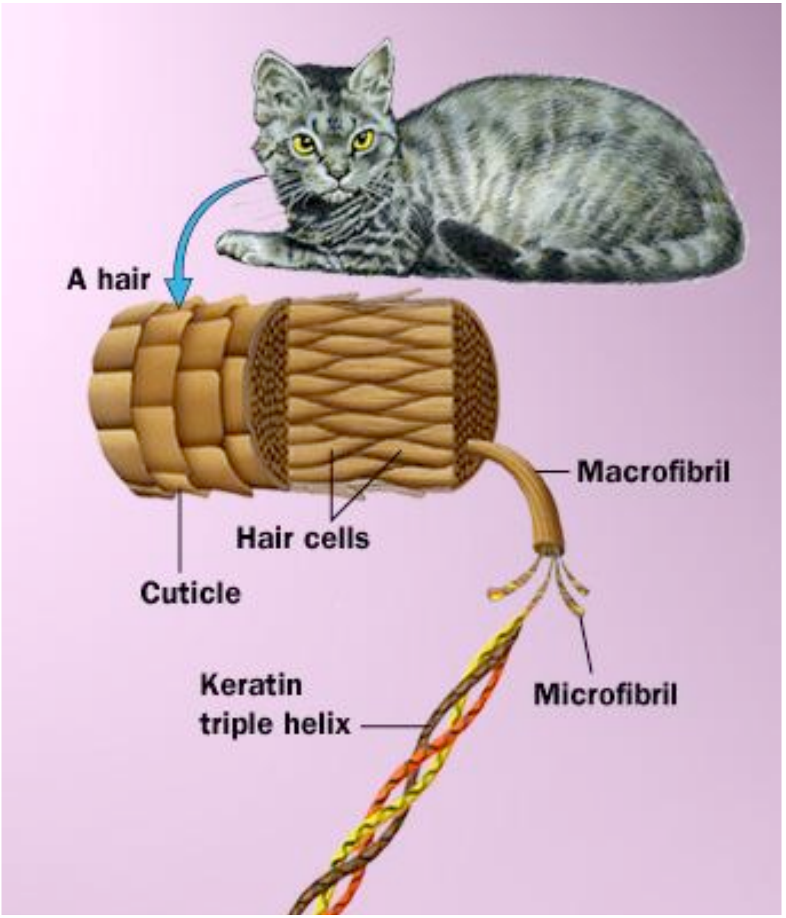
# 1. Escleroproteins

## 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.

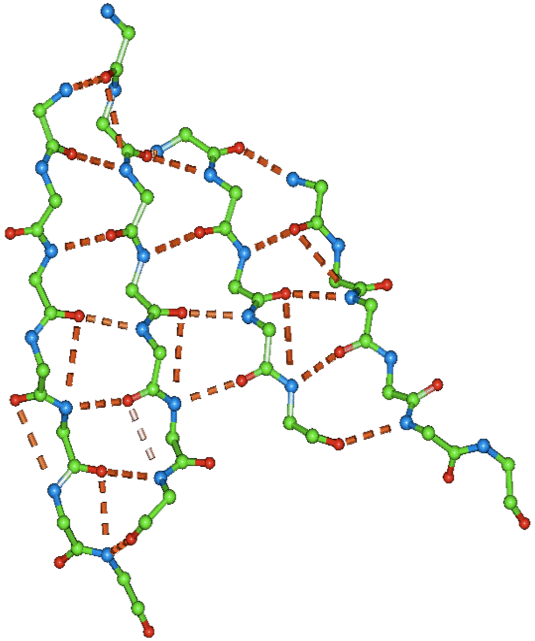




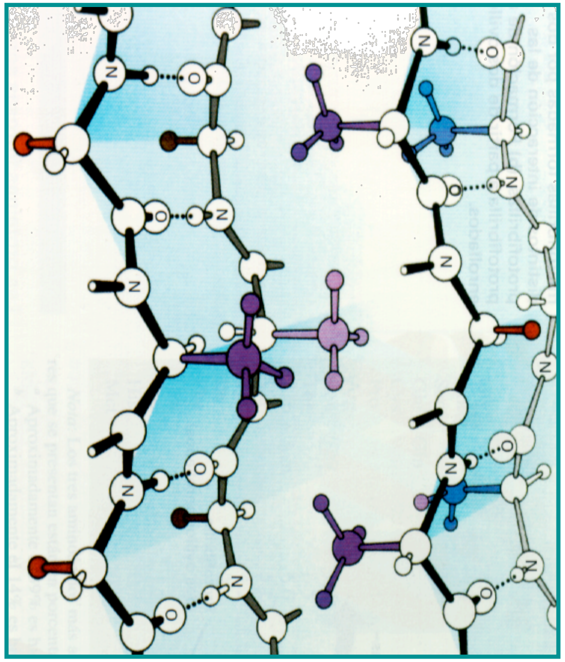
# 1. Escleroproteins

## 2. Fibroine

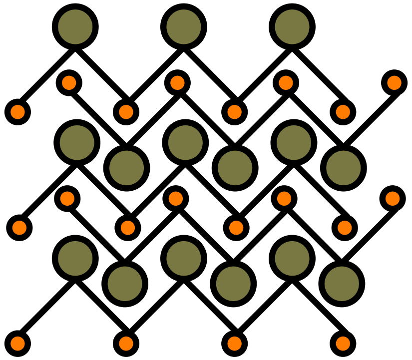
Fiber found in spider web, silk and elastin. Highly resistant. Formed by β-sheets



**sheet β**



**Sheet 1** **Sheet 2**



Fiber produced by packing between sheets.

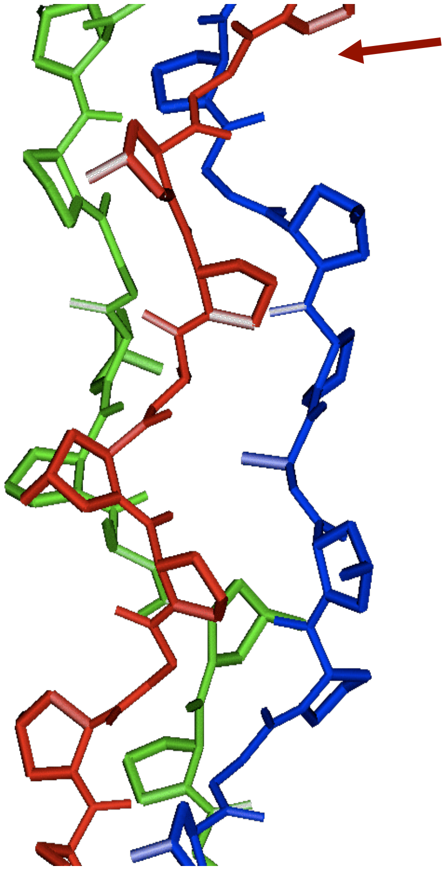
Packing due to Aa composition:

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

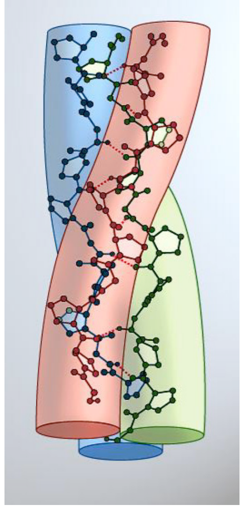
# 1. Escleroproteins

## 3. Collagen

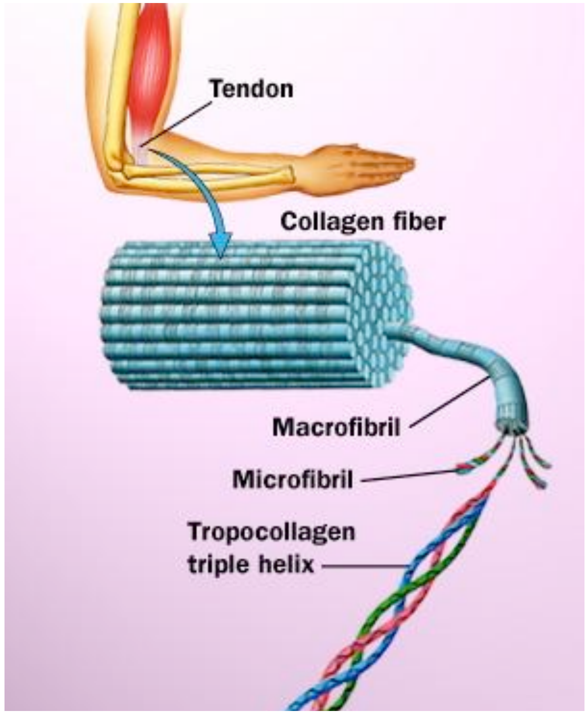
Collagen is a fiber formed by grouping macrofibrils and microfibrils of tropocollagen. Collagen is the main component of the connective tissue



**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

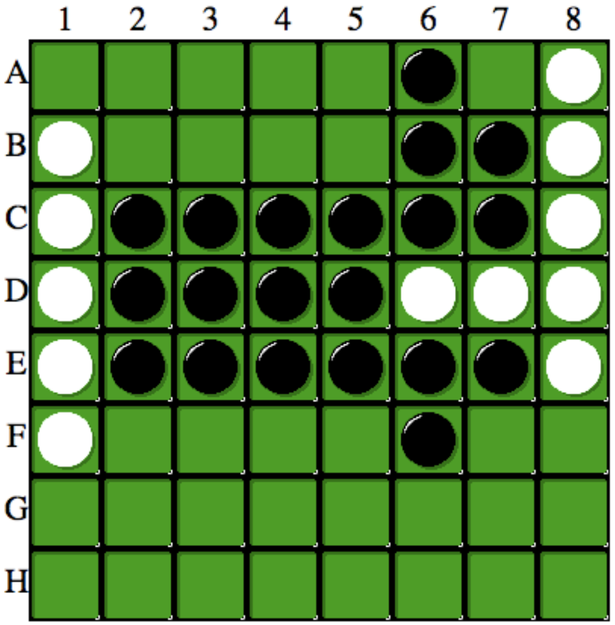
**Microstate**: State of a system specified by means of the coordinates and velocities of each particle.
**Macrostate**: State of a system specified by means of thermodynamic variables
**Hypothesis**: all microstates of a system are equally accessible with the same probability

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 $S=f(W)=f(W_a*W_b)$

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

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

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

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

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

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

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

# 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

$$Z = \sum_{V} \sum_{j} e^{-\beta E_{Vj}} e^{-\beta pV}$$

$$G = -kNT \ln Z$$

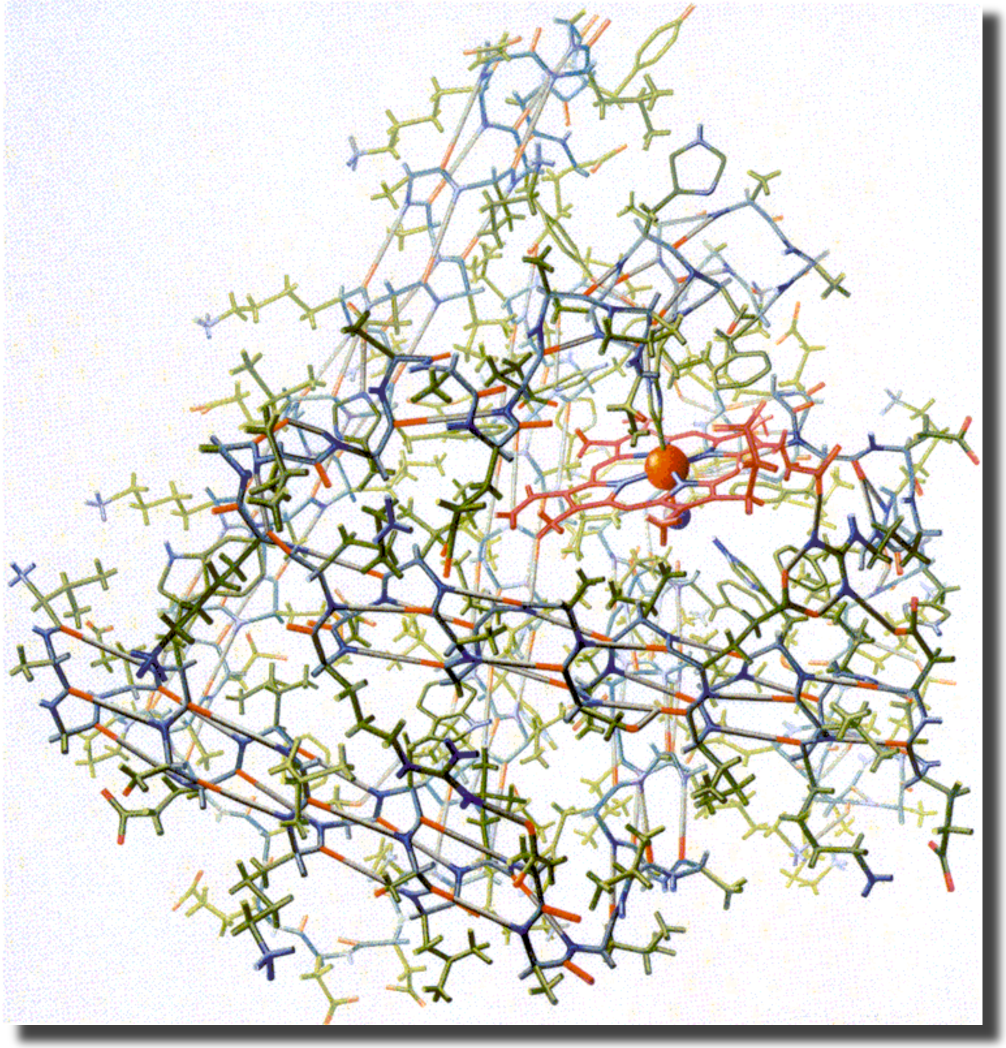
$$S = kN \ln Z + kN \left( \frac{\partial \ln Z}{\partial T} \right)_{N,p}$$

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

# 3. Soluble and globular proteins

## 1. Protein folding

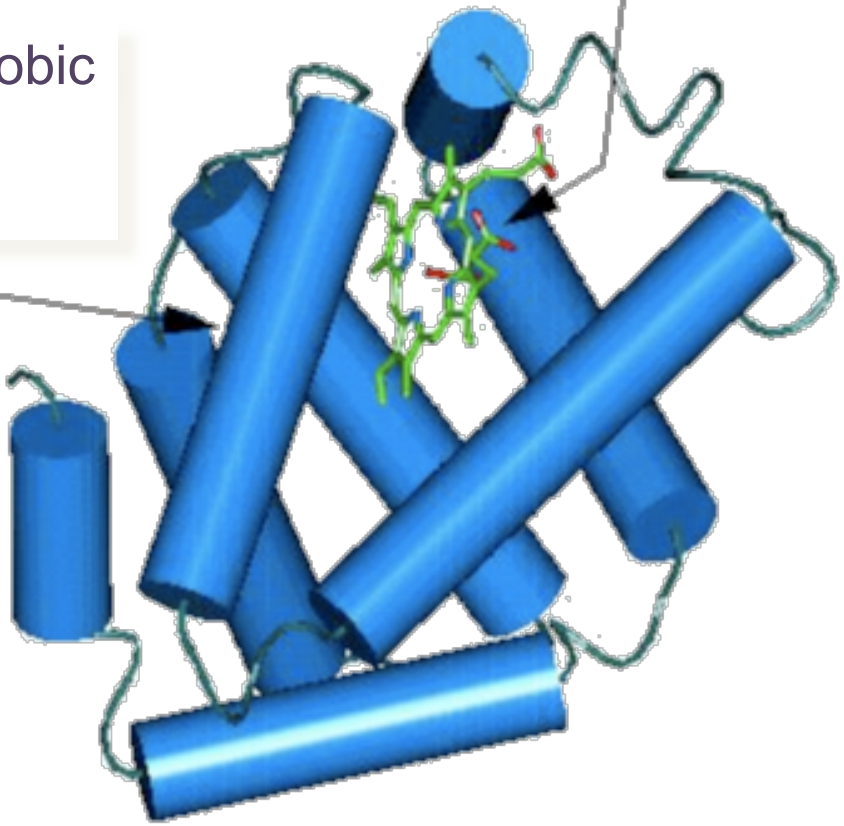
Example: MYOGLOBIN



# 3. Soluble and globular proteins

## 1. Protein folding

Example: MYOGLOBIN



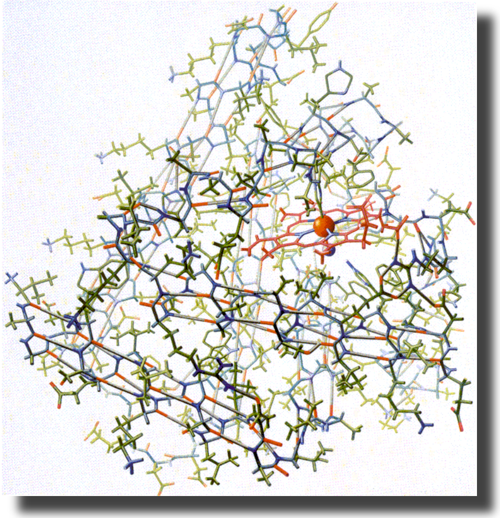
Hydrofobic Core

Porphyrin & $Fe^{+2}$ (heme group)

# 3. Soluble and globular proteins

## 1. Protein folding

Example: MYOGLOBIN



# 3. Soluble and globular proteins

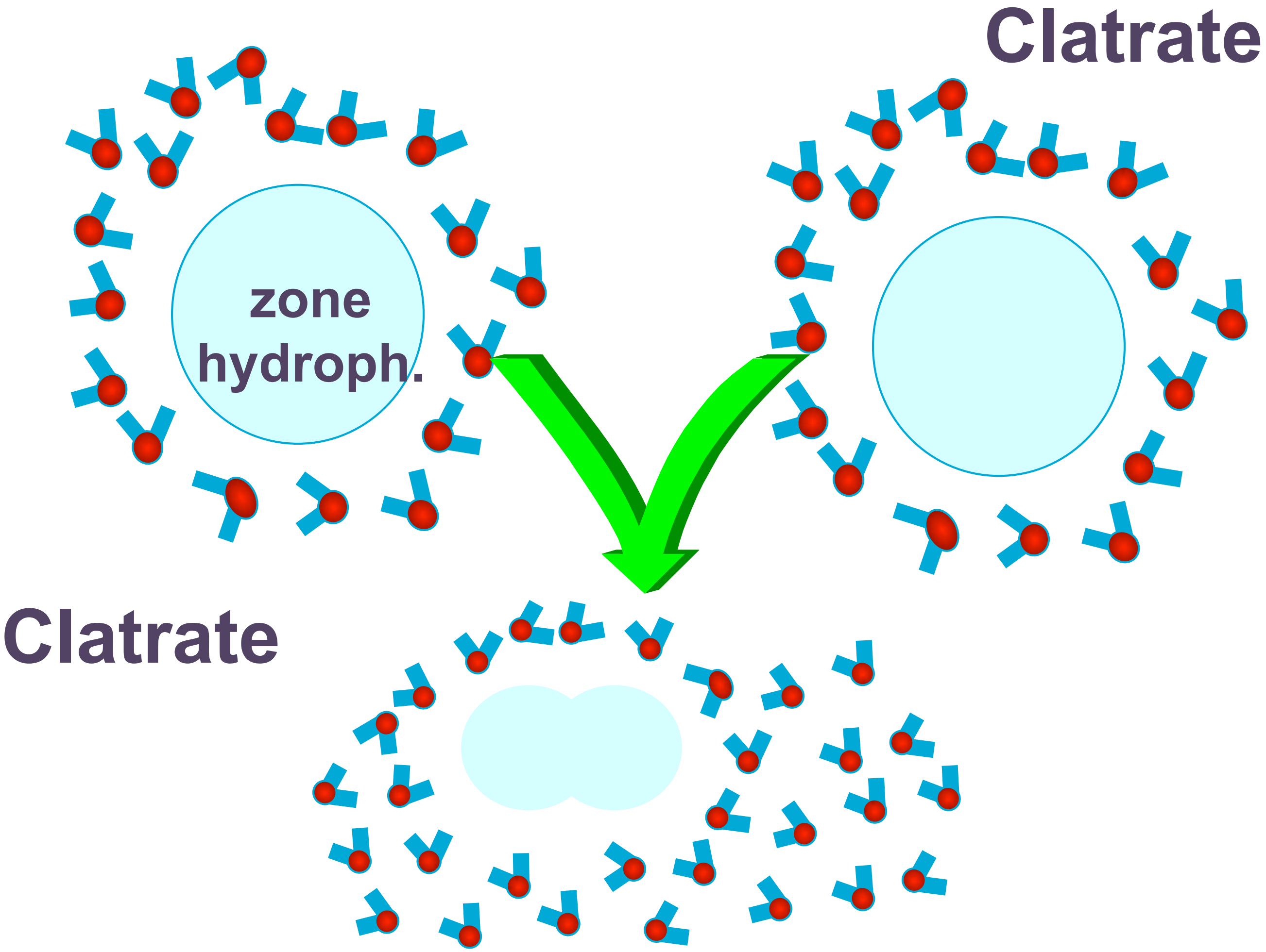
## 1. Protein folding

### Hydrophobic effect

Water free ↑

↓

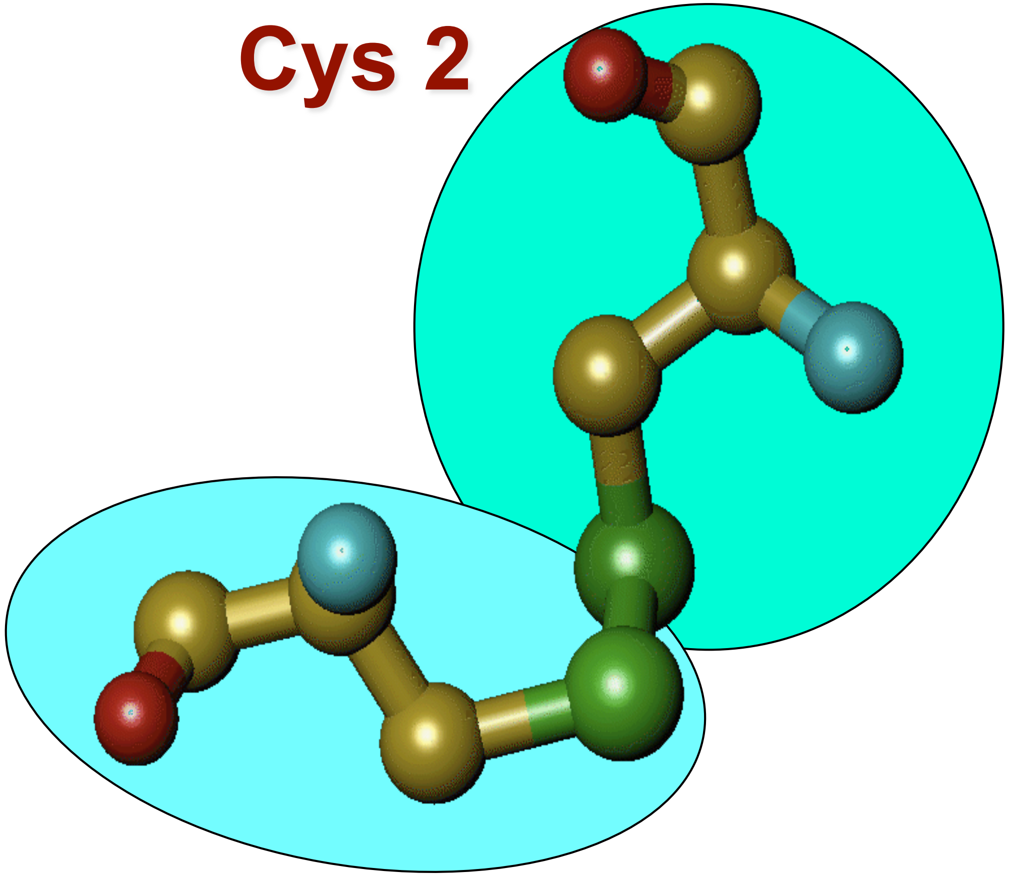
Δ **S** ↑



# 3. Soluble and globular proteins

## 1. Protein folding

### Disulphide bridges



Stabilization of the fold.

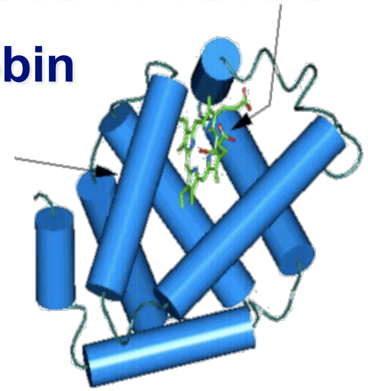
Produced by oxydation of close thiol groups (-SH).

They may play a relevant role in small proteins (peptides)

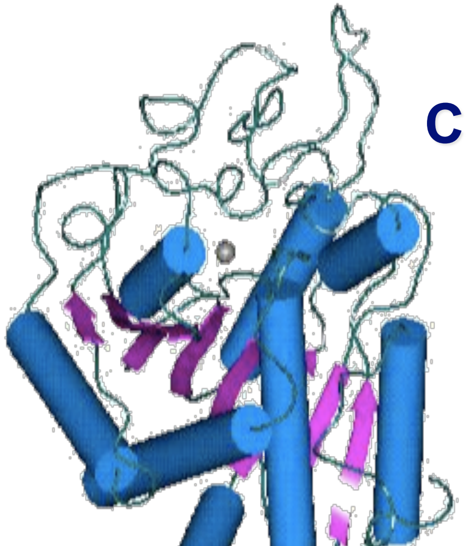
# 3. Soluble and globular proteins

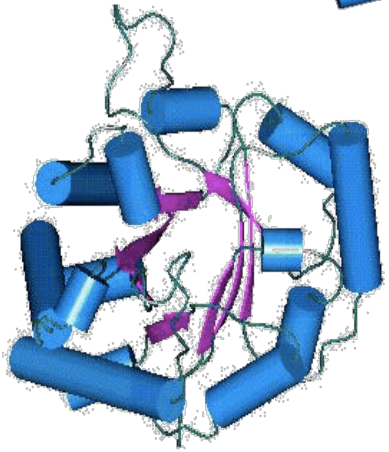
## 2. Protein folds

**Myoglobin**

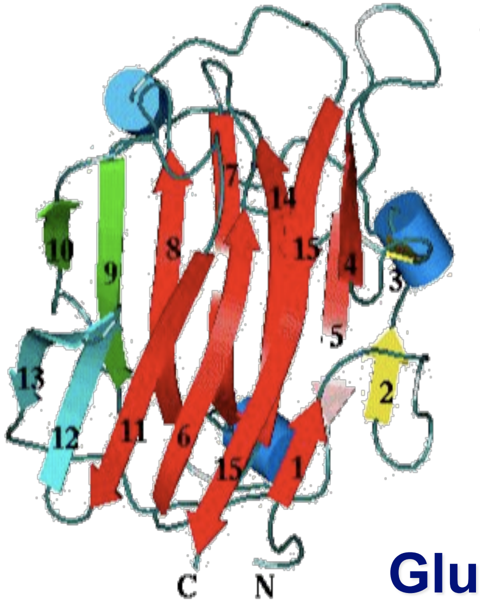


**Carboxypeptidase**

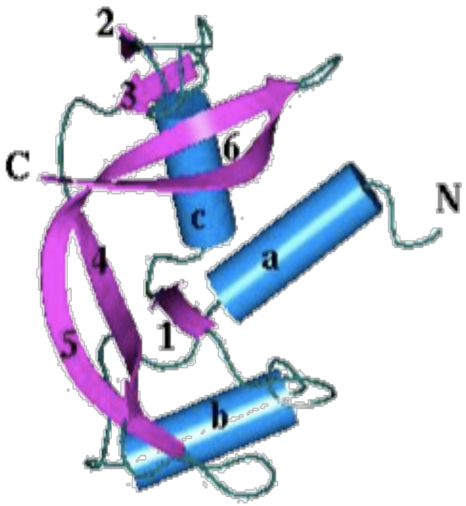




**Triose Phosphate Isomerase**



**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

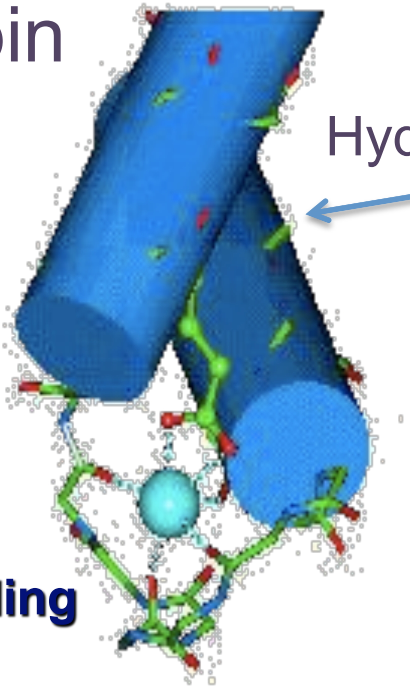
# 3. Soluble and globular proteins

## 3. Supersecondary structure

### 2. α−α

Ca$^{+2}$ EF hand

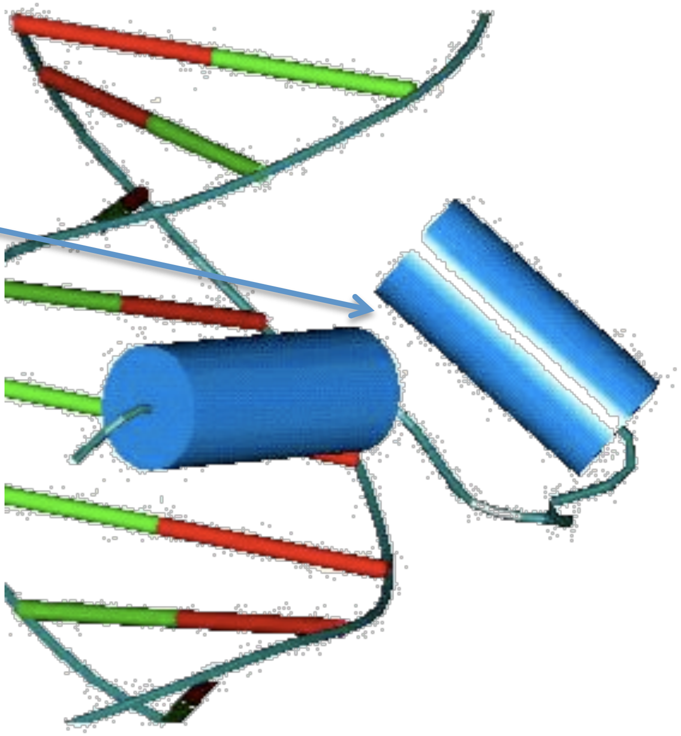
$\alpha$ hairpin



**Ca$^{+2}$ binding**

Hydrophobic core

DNA binding motif
(Cro repressor)

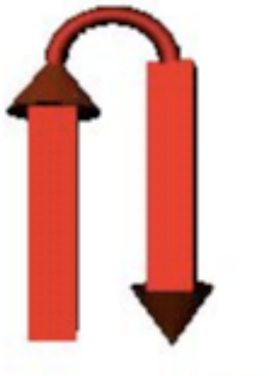


# 3. Soluble and globular proteins

## 3. Supersecondary structure

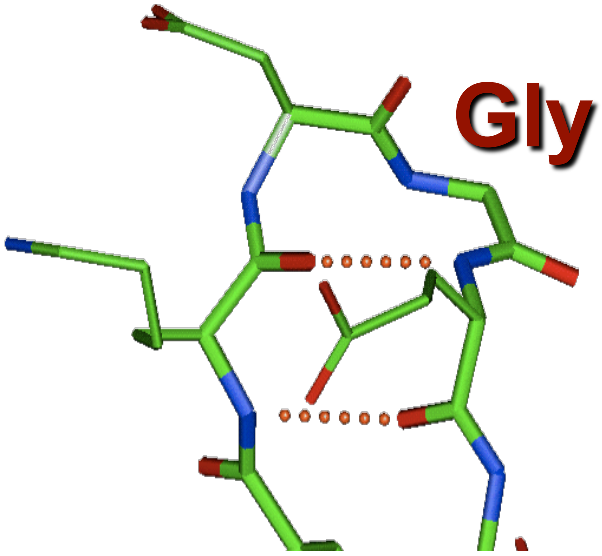
### 3. β−β

β−β hairpin



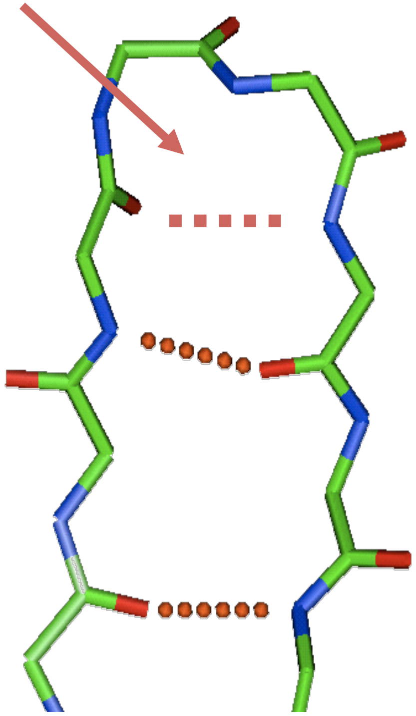
Anti-parallel
β-ladder

Turn β = 2 Aa



Type I’

H.B. lost



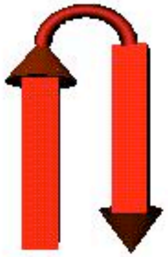
Type II

# 3. Soluble and globular proteins

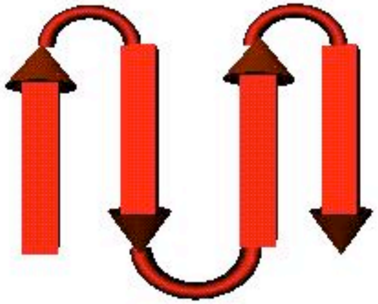
## 3. Supersecondary structure

### 3. β−β

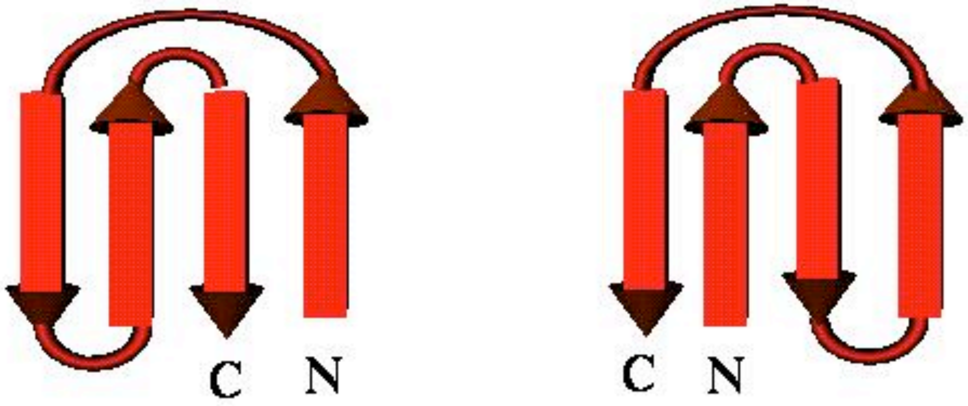
β hairpin



β meander



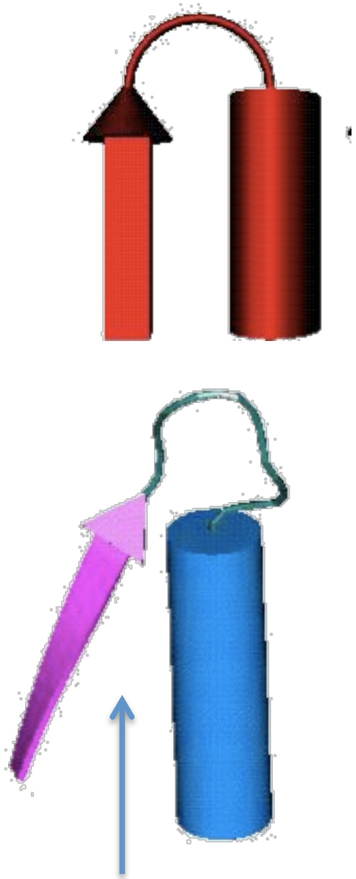
Greek Key



# 3. Soluble and globular proteins

## 3. Supersecondary structure

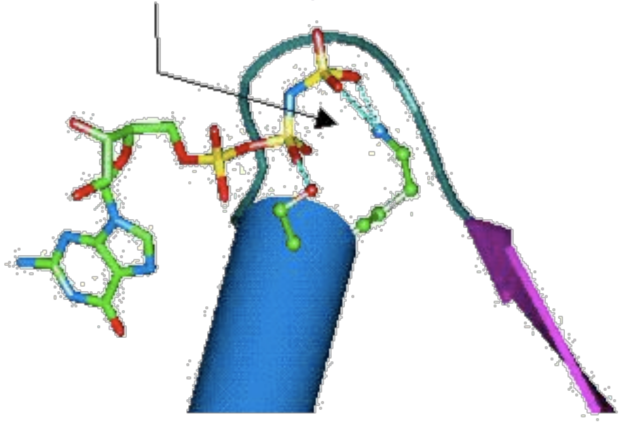
### 4. β−α & α−β



Hydrophobic core

P-loop

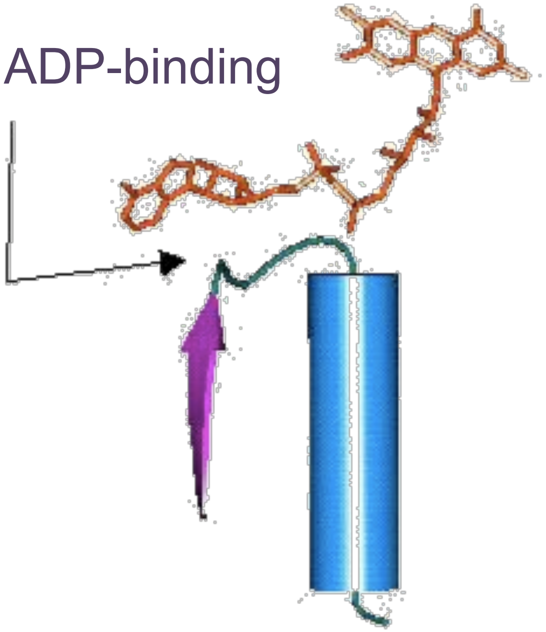
GDP-binding



Ras P21

FAD/NADP binding loop

FAD/NADP-binding

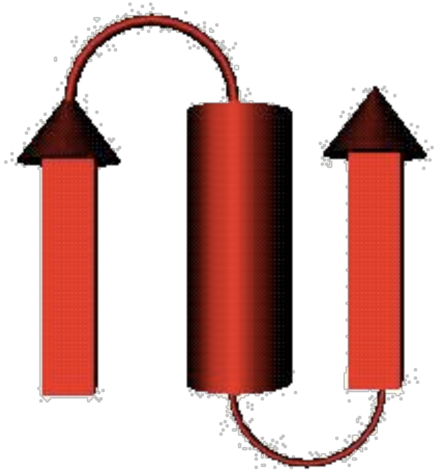


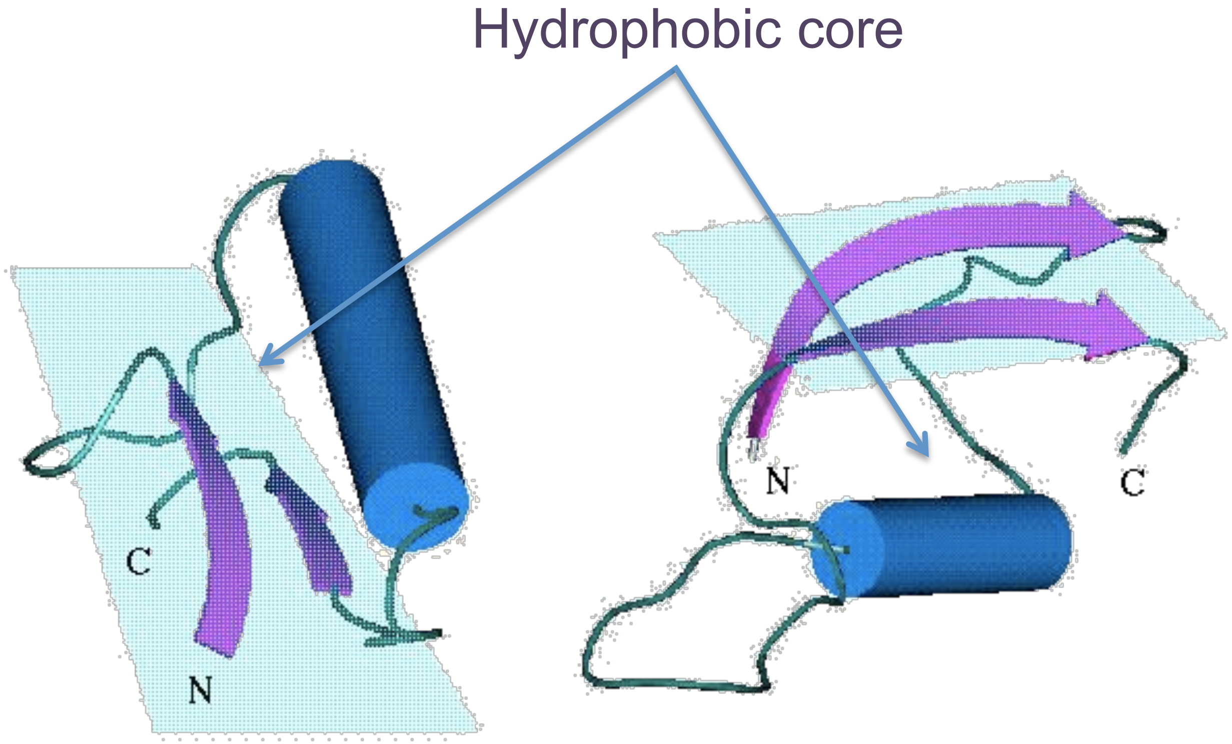
P-Hydroxybenzoate-Hydroxylase

# 3. Soluble and globular proteins

## 3. Supersecondary structure

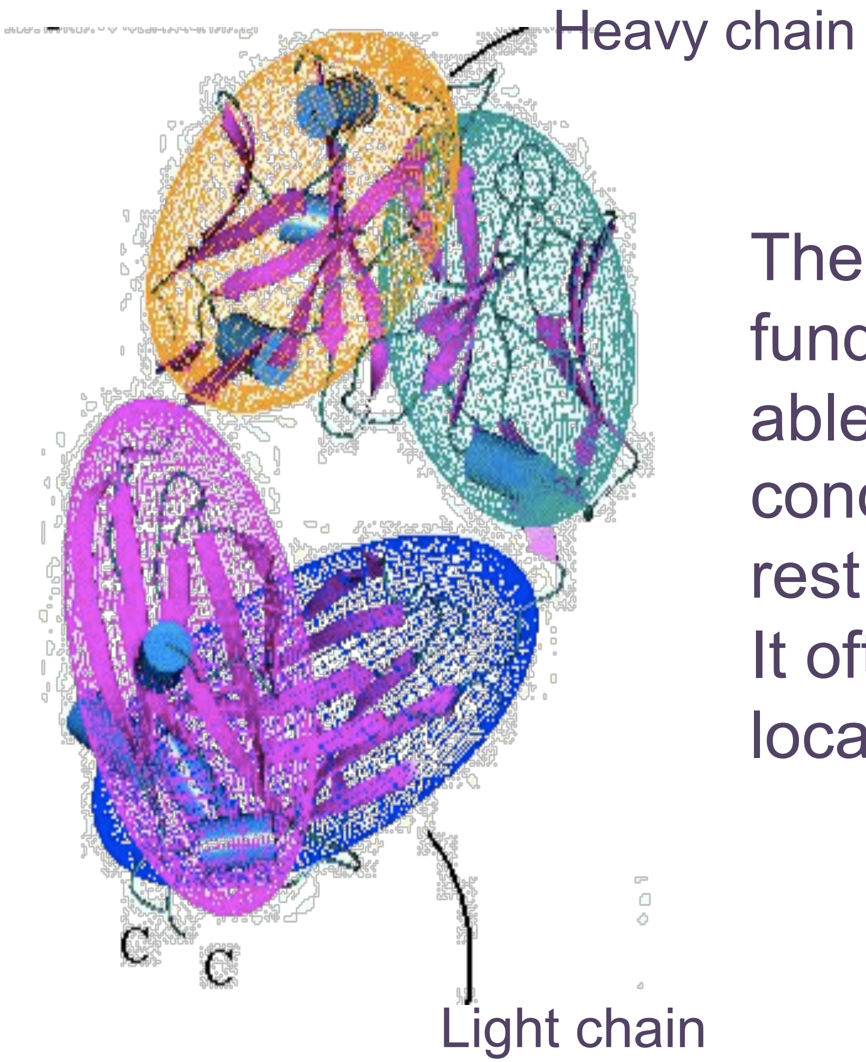
### 5. β−α−β





**Right handed** **Left handed**

# 3. Soluble and globular proteins
## 4. Domains
### 1. Definition

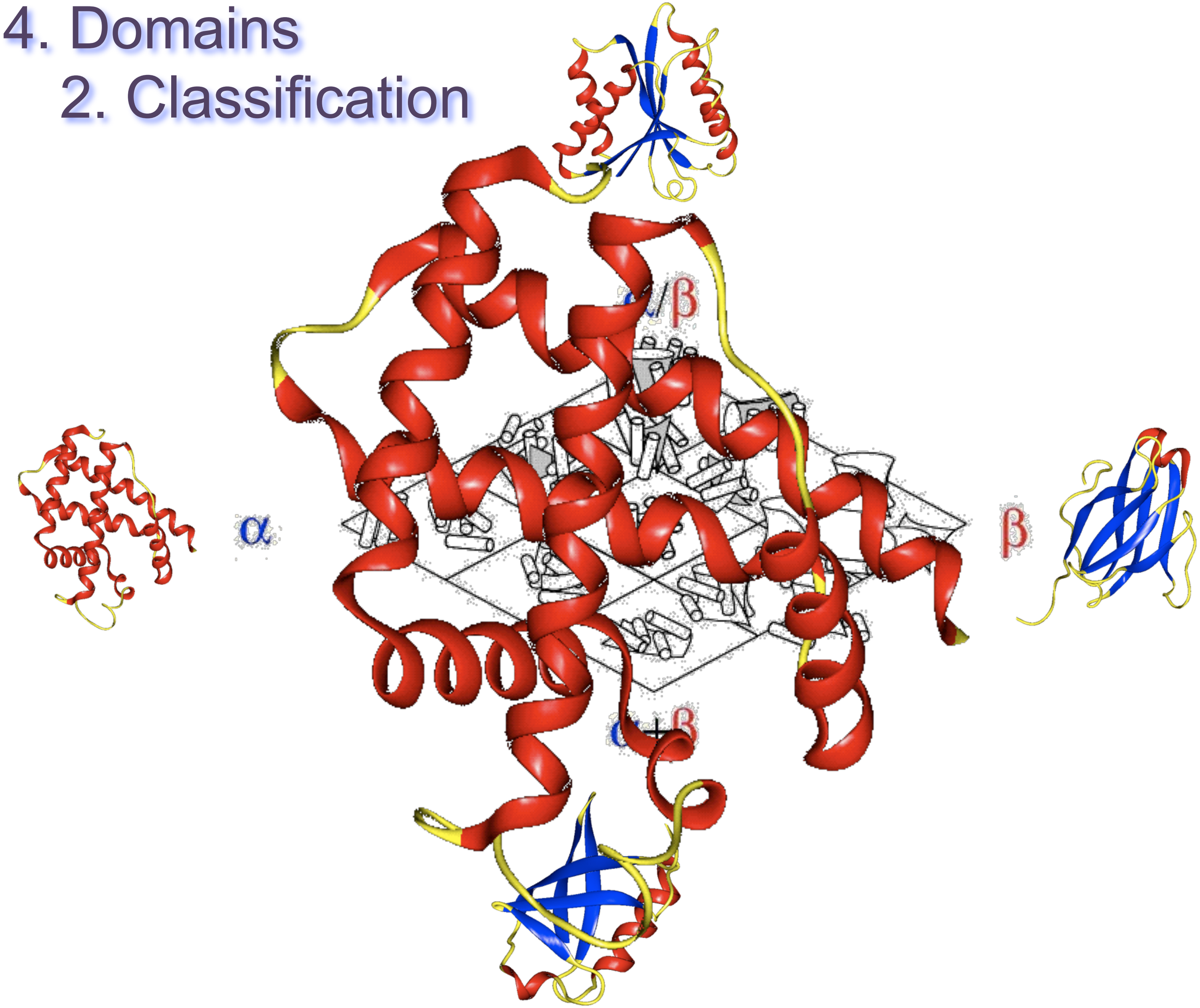


The protein domain is defined as the fundamental unit of 3D structure, able to fold by itself in the right conditions with independence of the rest of the protein.
It often corresponds to functional local and compact units of a protein

# 3. Soluble and globular proteins

## 4. Domains

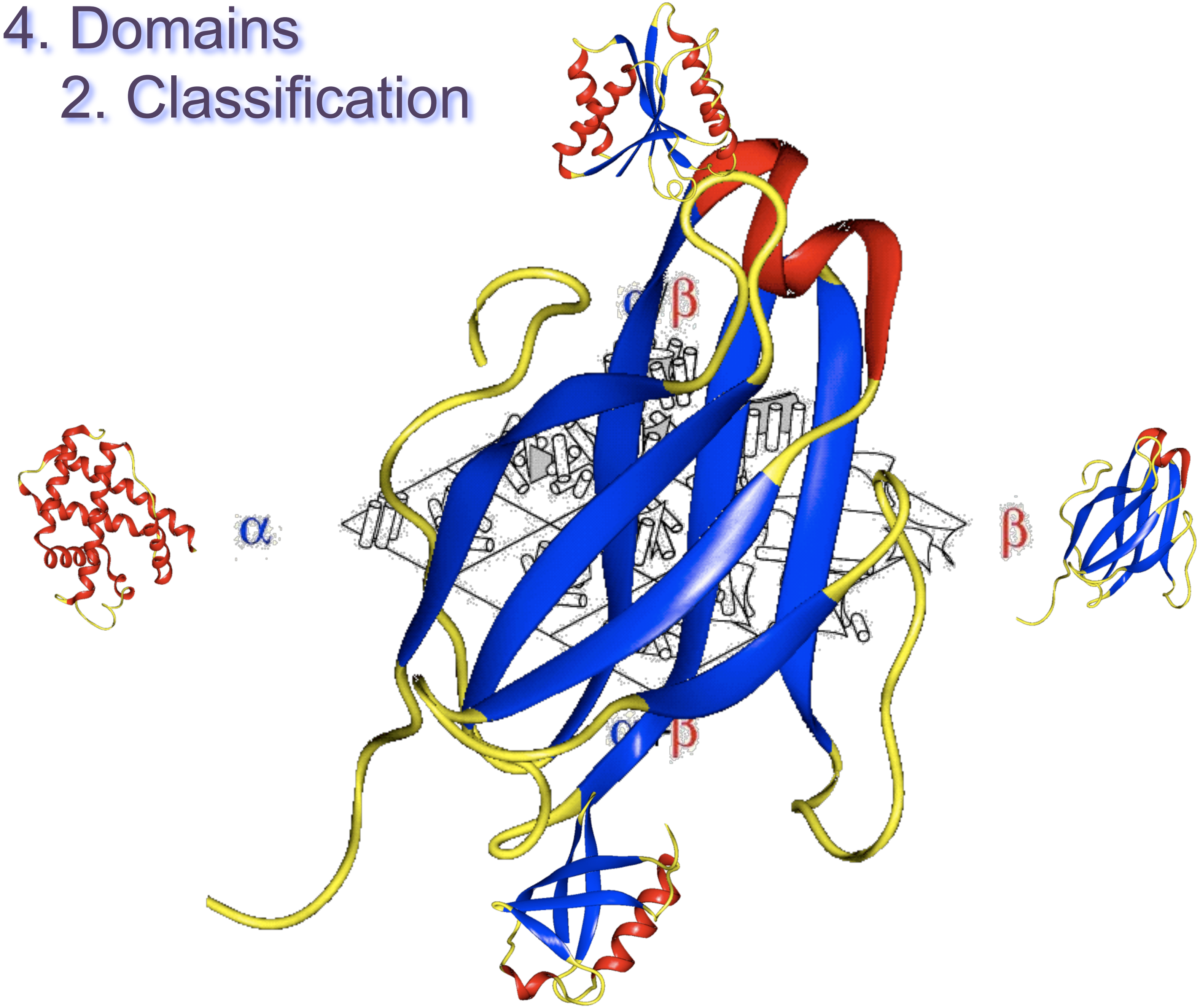
### 2. Classification



# 3. Soluble and globular proteins

## 4. Domains

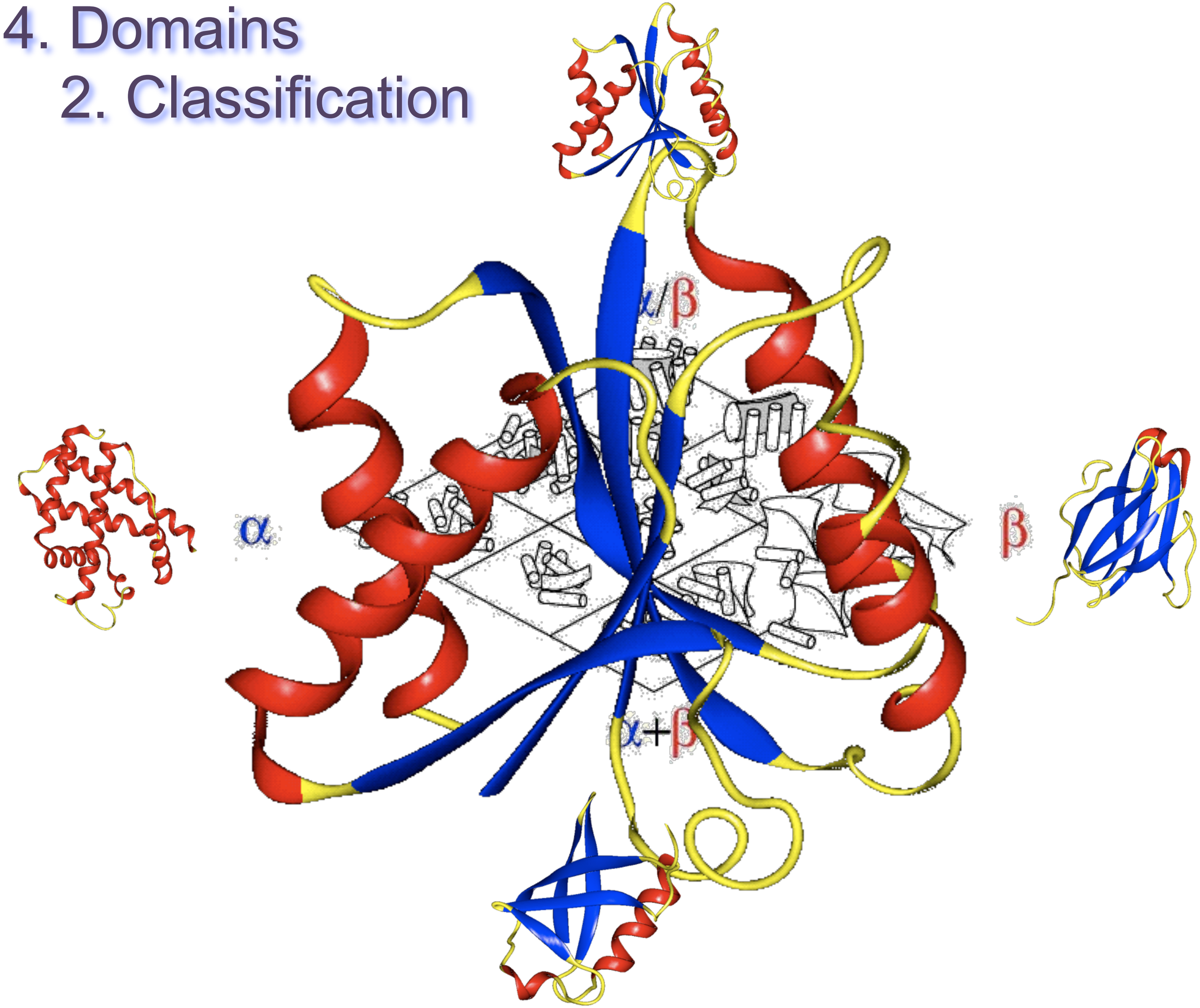
### 2. Classification



# 3. Soluble and globular proteins

## 4. Domains

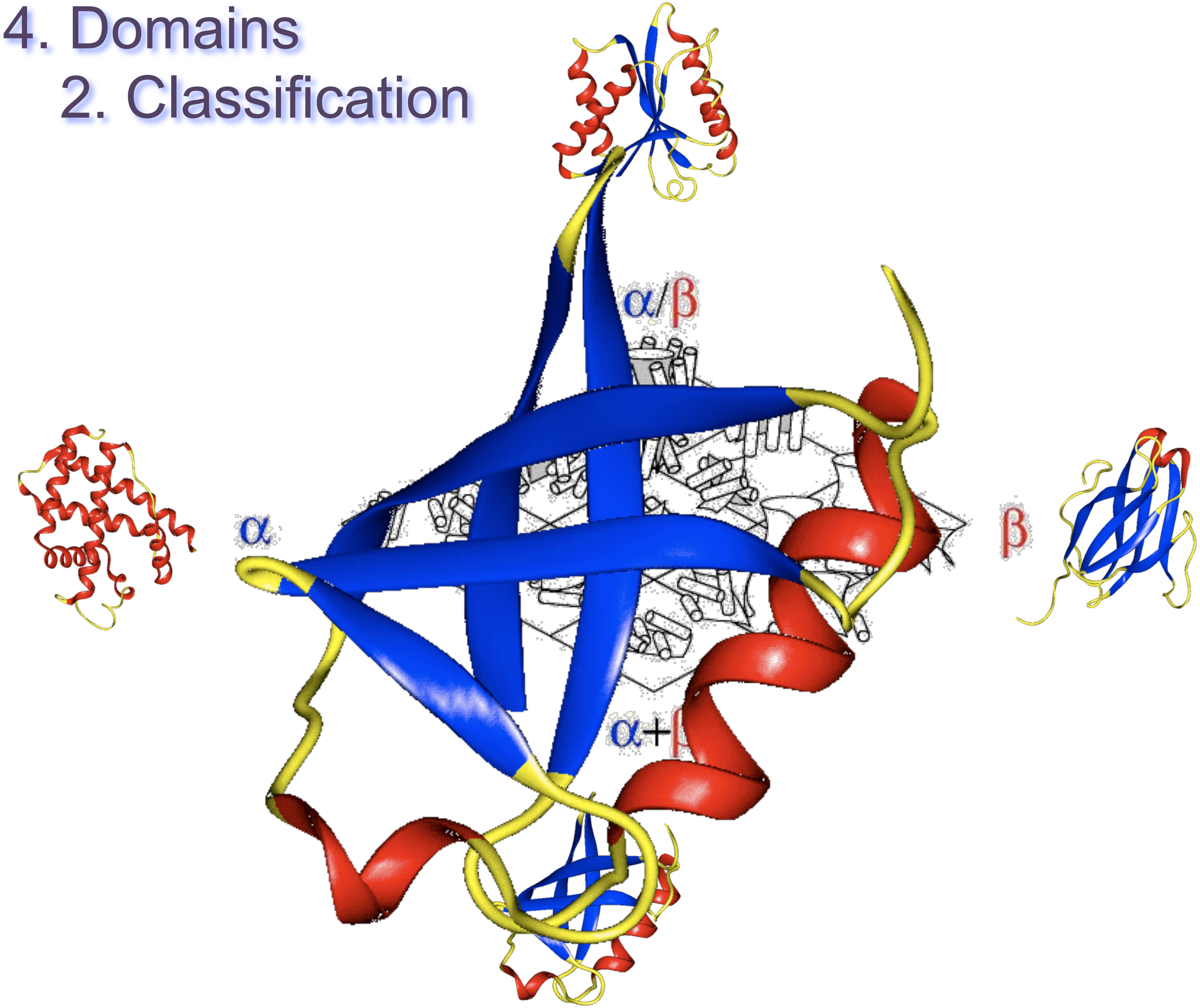
### 2. Classification



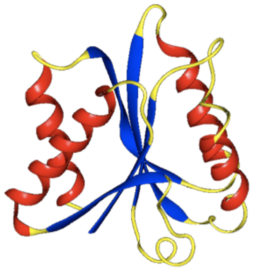
# 3. Soluble and globular proteins

## 4. Domains

### 2. Classification



# 3. Soluble and globular proteins
## 4. Domains
### 2. Classification



**5th class**

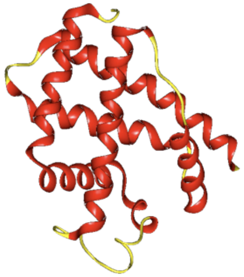
Multi-domain proteins α and β

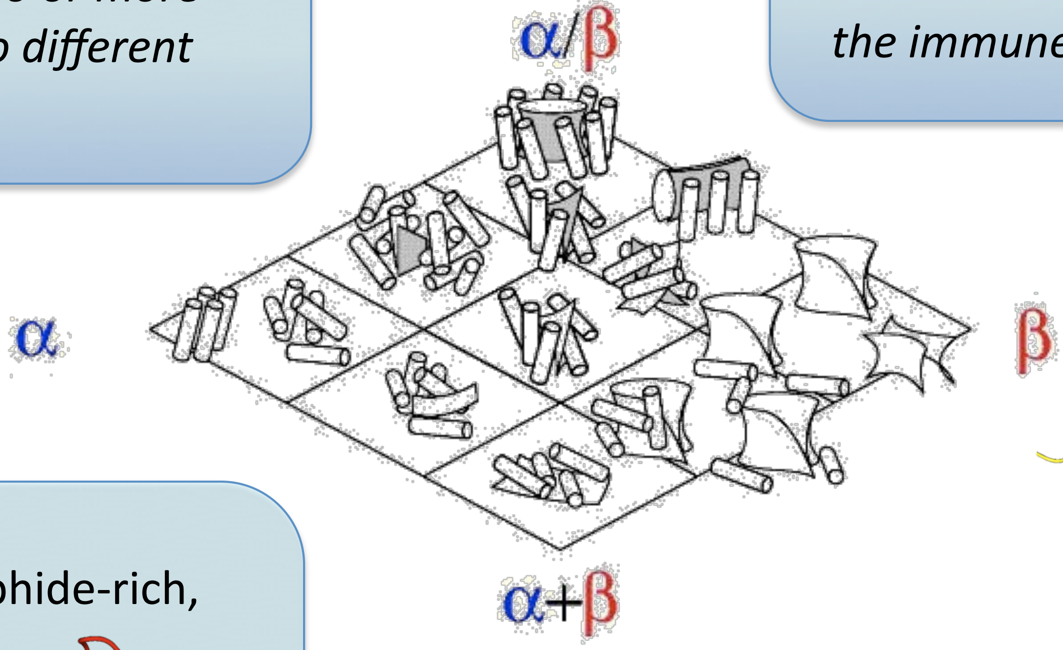
*Folds consisting of two or more domains belonging to different classes*

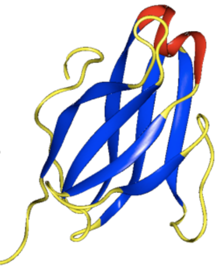
**6th class**

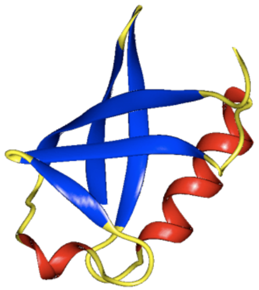
Membrane and cell surface proteins and peptides.

*Does not include proteins in the immune system*



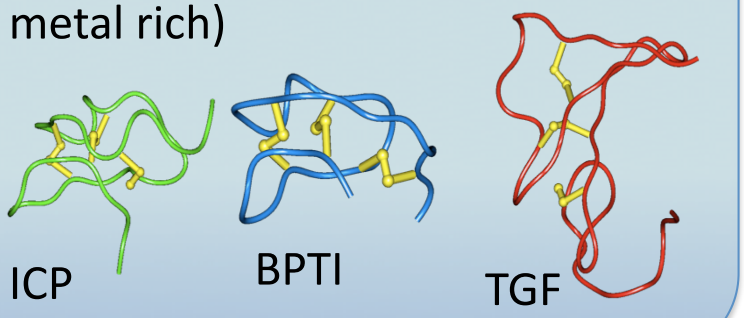






**7th class**

Small proteins (disulphide-rich, metal rich)

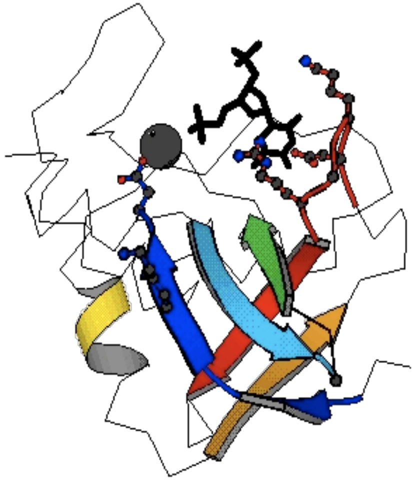


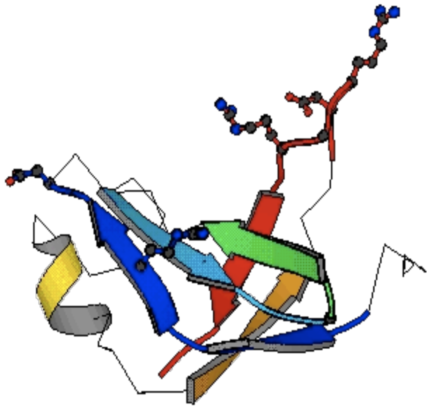
# 3. Soluble and globular proteins

## 4. Domains

### 3. Function-Sequence-Structure

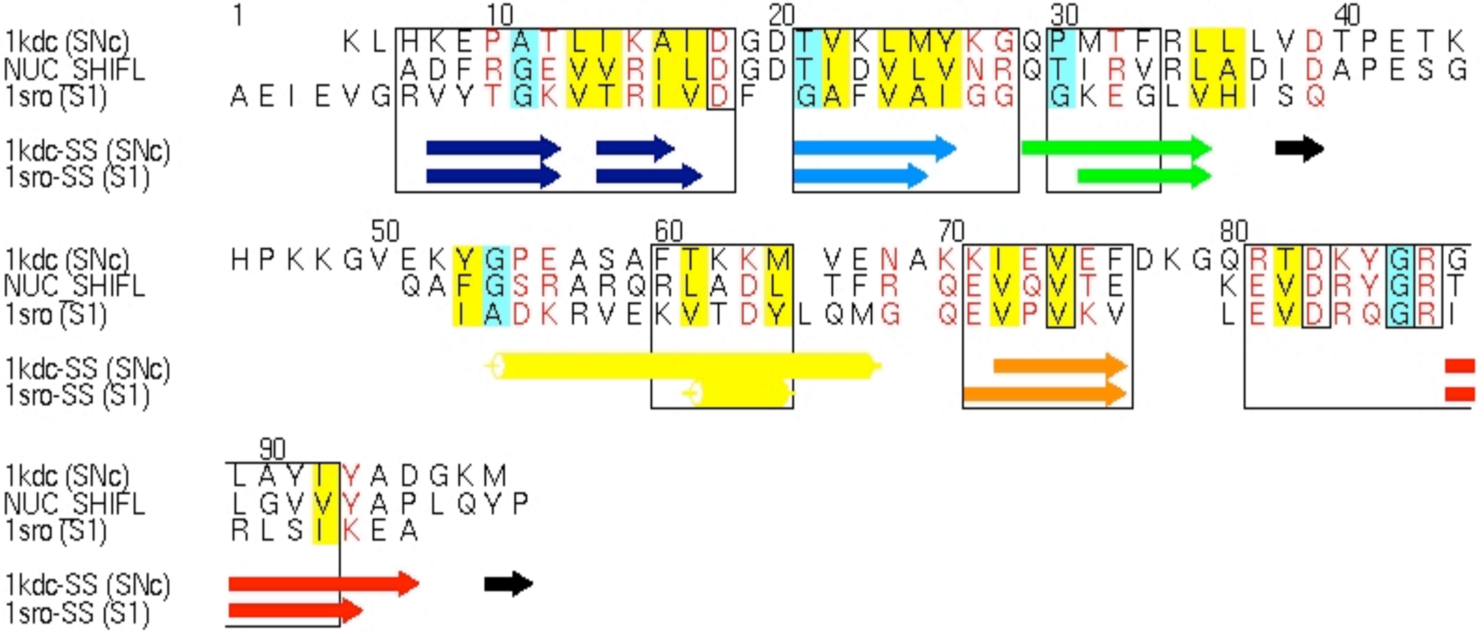
Stapphylococcyl nuclease





S1 RNA binding domain

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

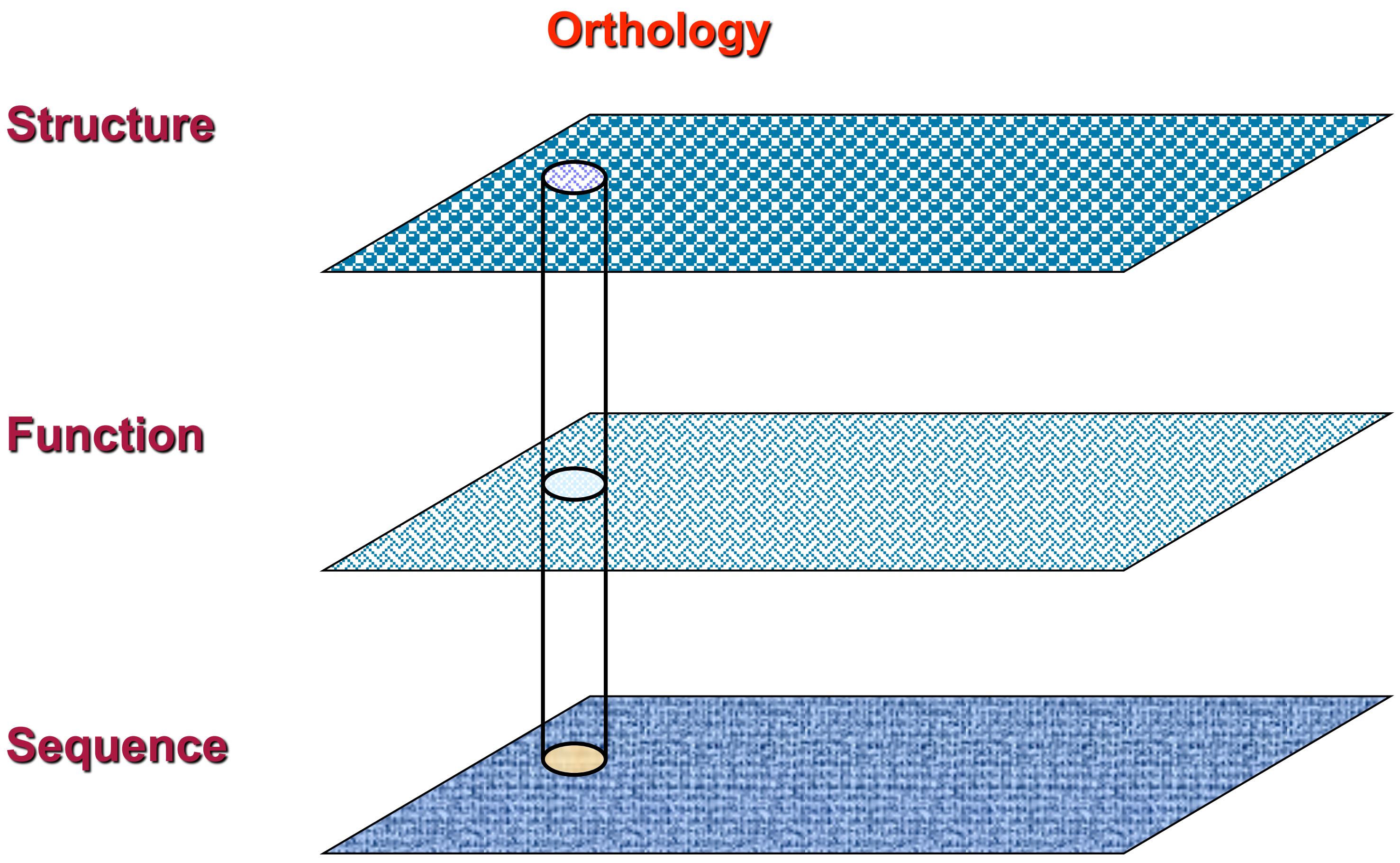


```
              1        10        20        30        40
1kdc (SNc)        KLHKEPATLIKAIDGDTVKLMYKGQPMTFRLLLVDTPETK
NUC_SHIFL           ADFRGEVVRILDGDTIDVLVNRQTIRVRLADIDAPESG
1sro (S1)     AEIEVGRVYTGKVTRIVDF GAFVAIGG GKEGLVHISQ
1kdc-SS (SNc)
1sro-SS (S1)

                   50        60        70        80
1kdc (SNc)    HPKKGVEKYGPEASAFTKKM VENAKKIEVEFDKGQRTDKYGRG
NUC_SHIFL           QAFGSRARQRLADL TFR QEVQVTE   KEVDRYGRT
1sro (S1)             IADKRVEKVTDYLQMG QEVPVKV   LEVDRQGRI
1kdc-SS (SNc)
1sro-SS (S1)

               90
1kdc (SNc)    LAYIYADGKM
NUC_SHIFL     LGVVYAPLQYP
1sro (S1)     RLSIKEA
1kdc-SS (SNc)
1sro-SS (S1)
```

# 3. Soluble and globular proteins

## 4. Domains

### 3. Function-Sequence-Structure



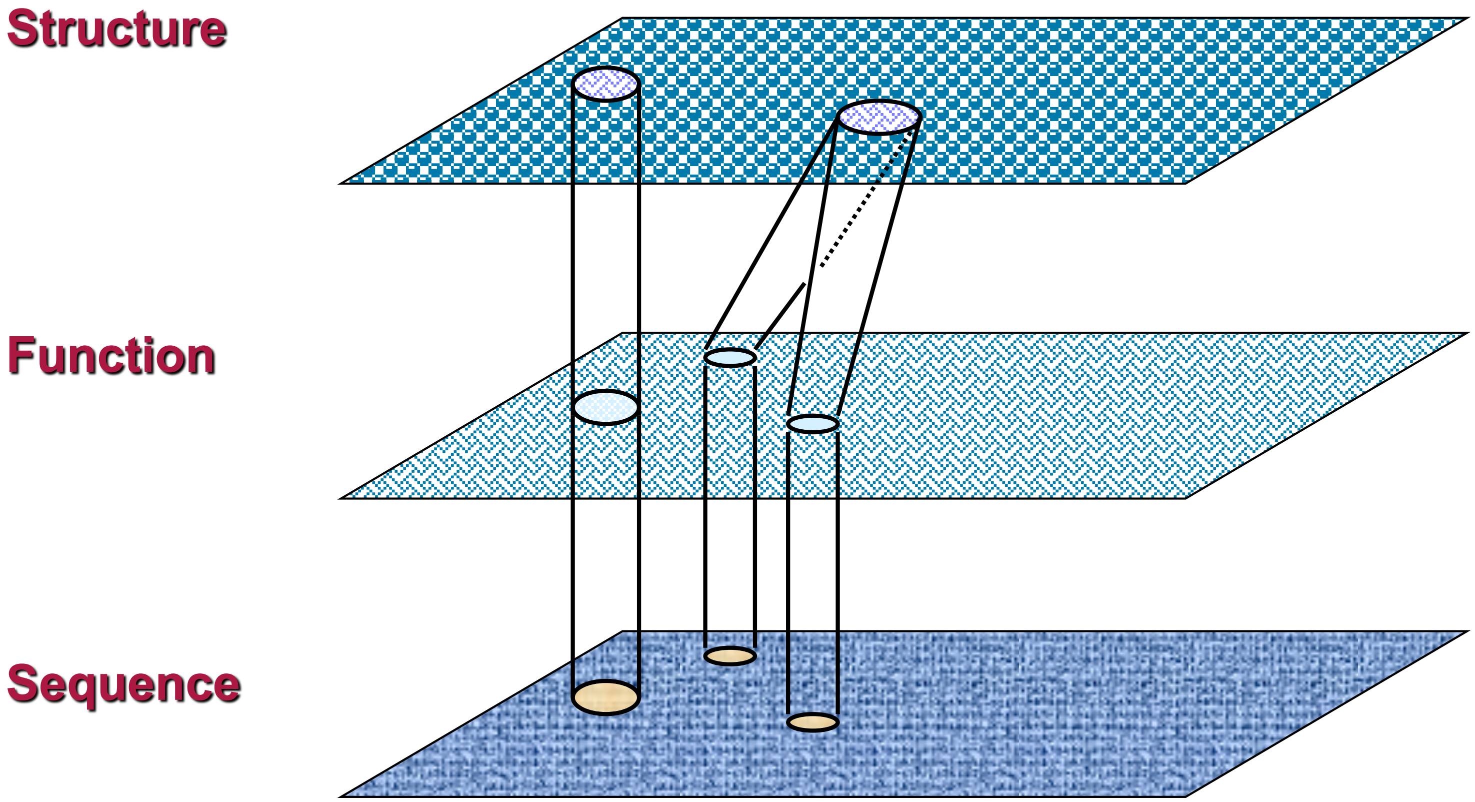
**Homology**: two proteins are homologous if they are the products of genes that evolved from the same ancestor

# 3. Soluble and globular proteins

## 4. Domains

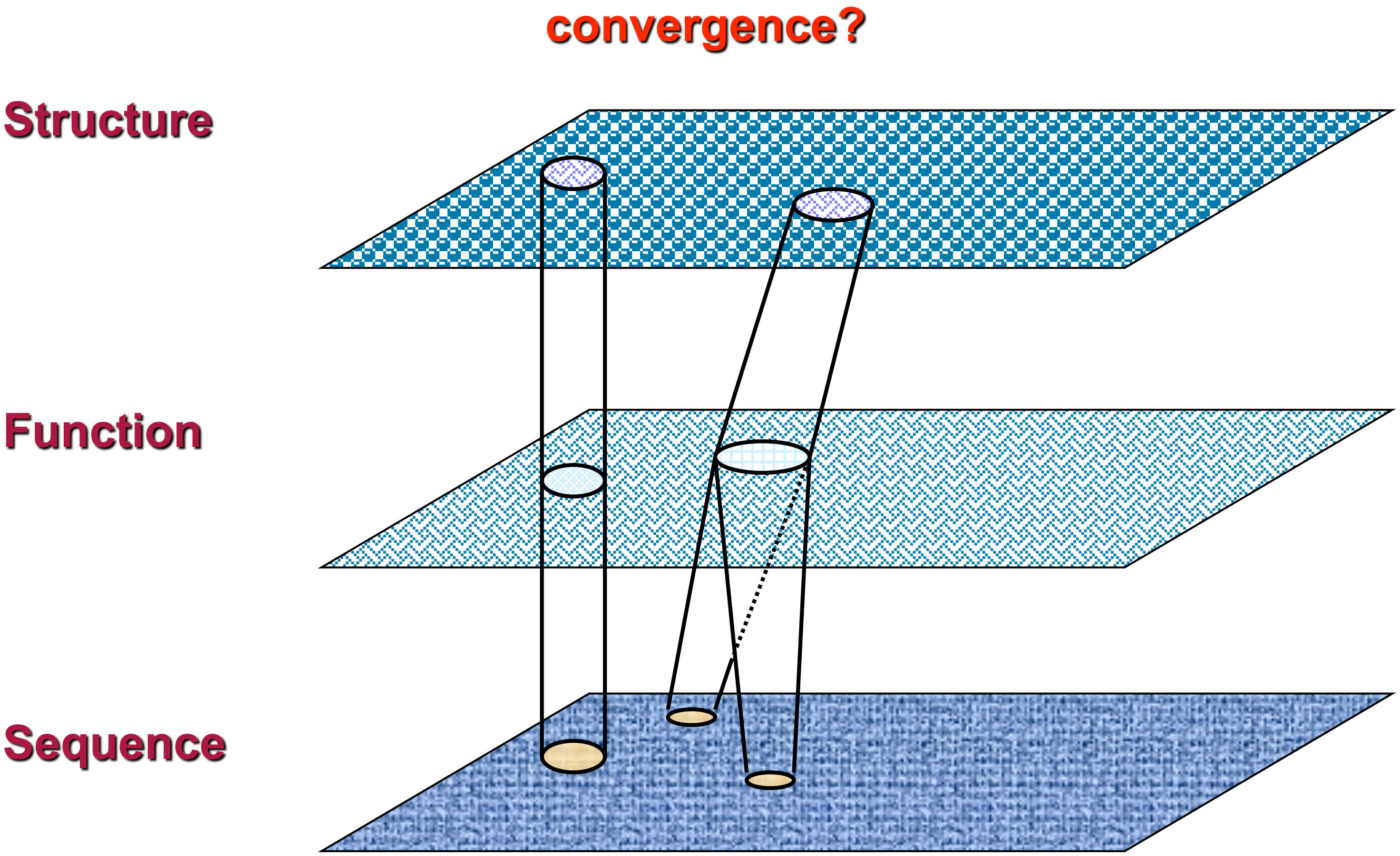
### 3. Function-Sequence-Structure

**Convergence (Analogy)**



**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

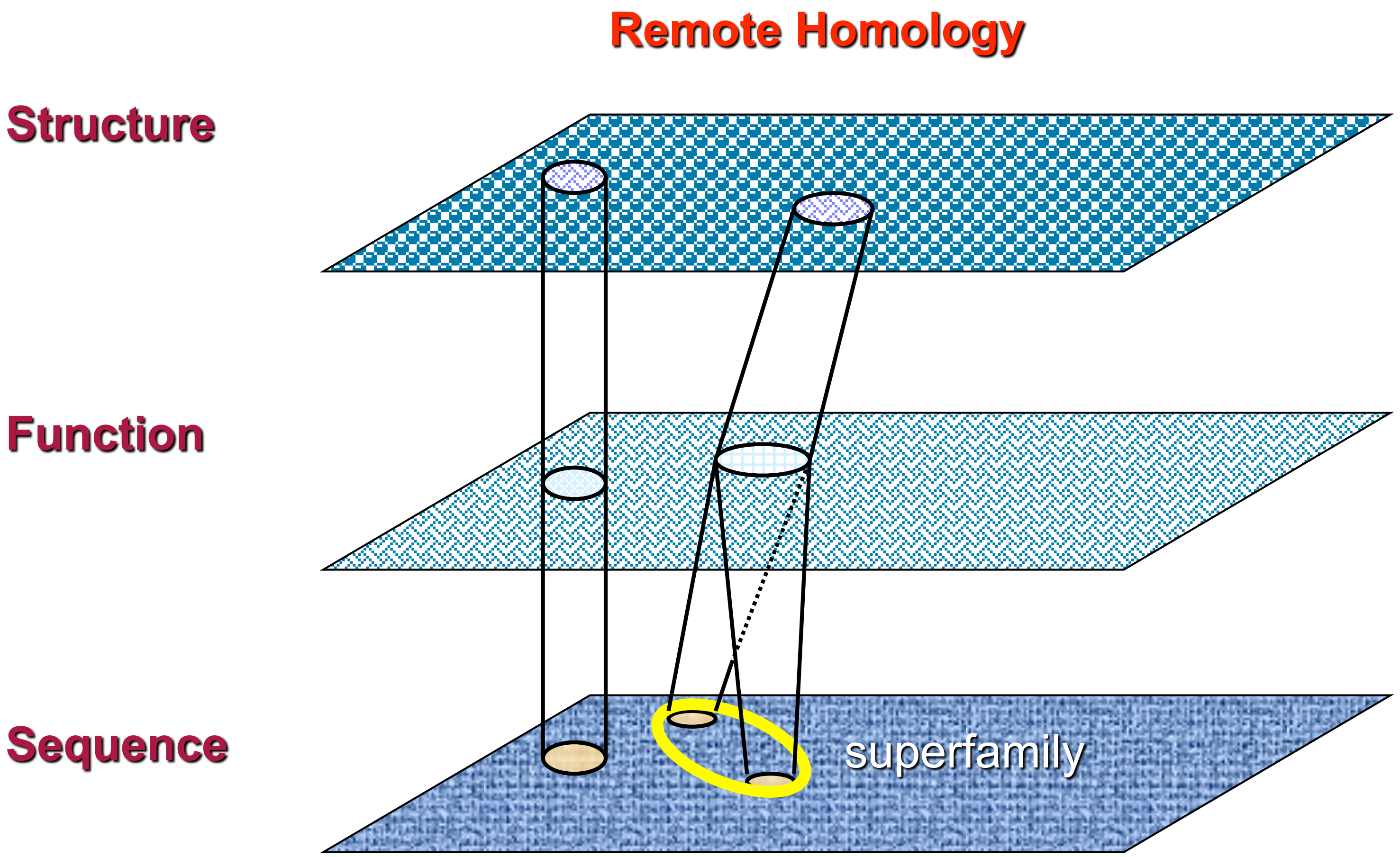


**Homology**: two proteins are homologous if they are the products of genes that evolved from the same ancestor

# 3. Soluble and globular proteins

## 4. Domains

### 3. Function-Sequence-Structure

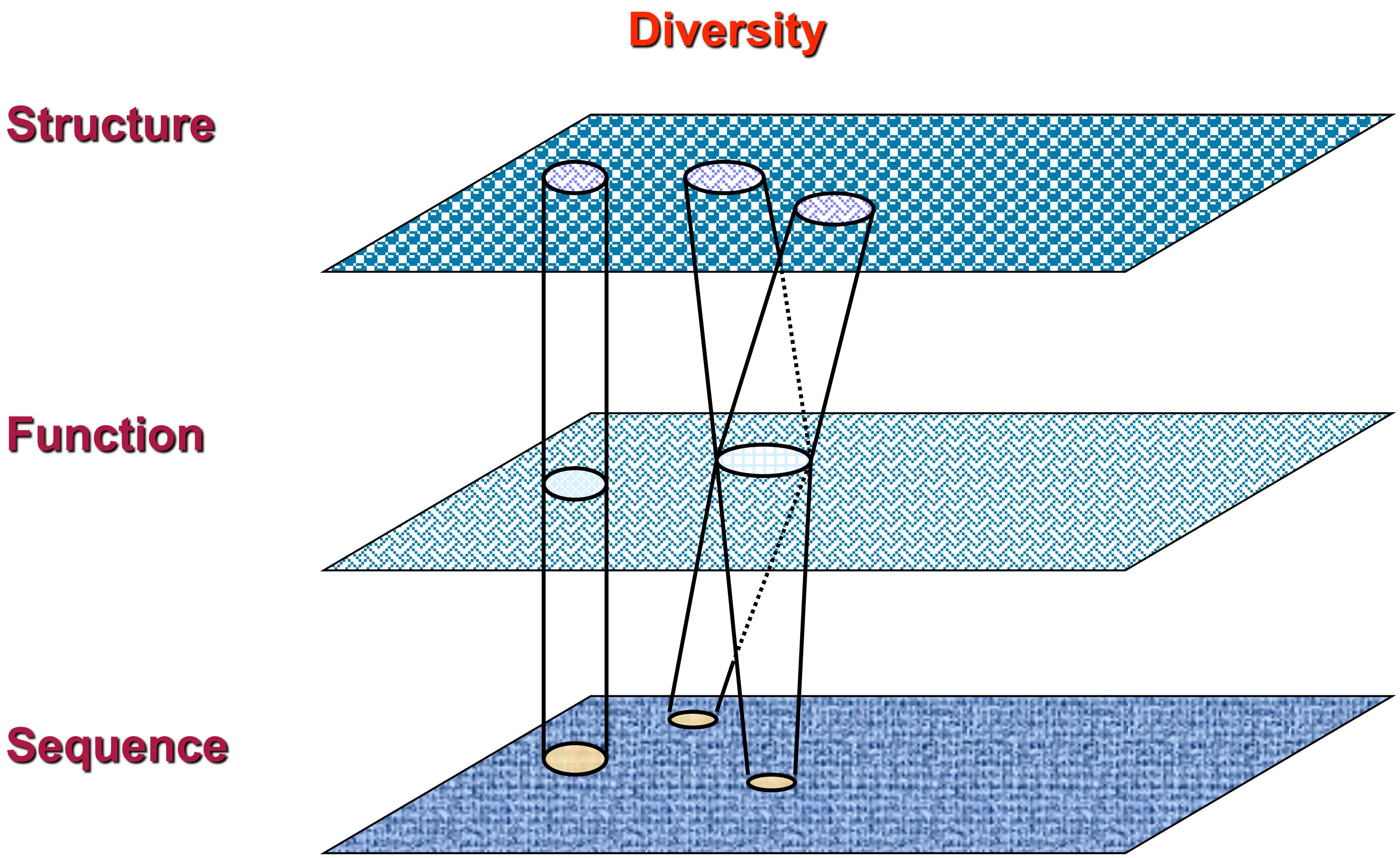


**Homology**: two proteins are homologous if they are the products of genes that evolved from the same ancestor

# 3. Soluble and globular proteins

## 4. Domains

### 3. Function-Sequence-Structure

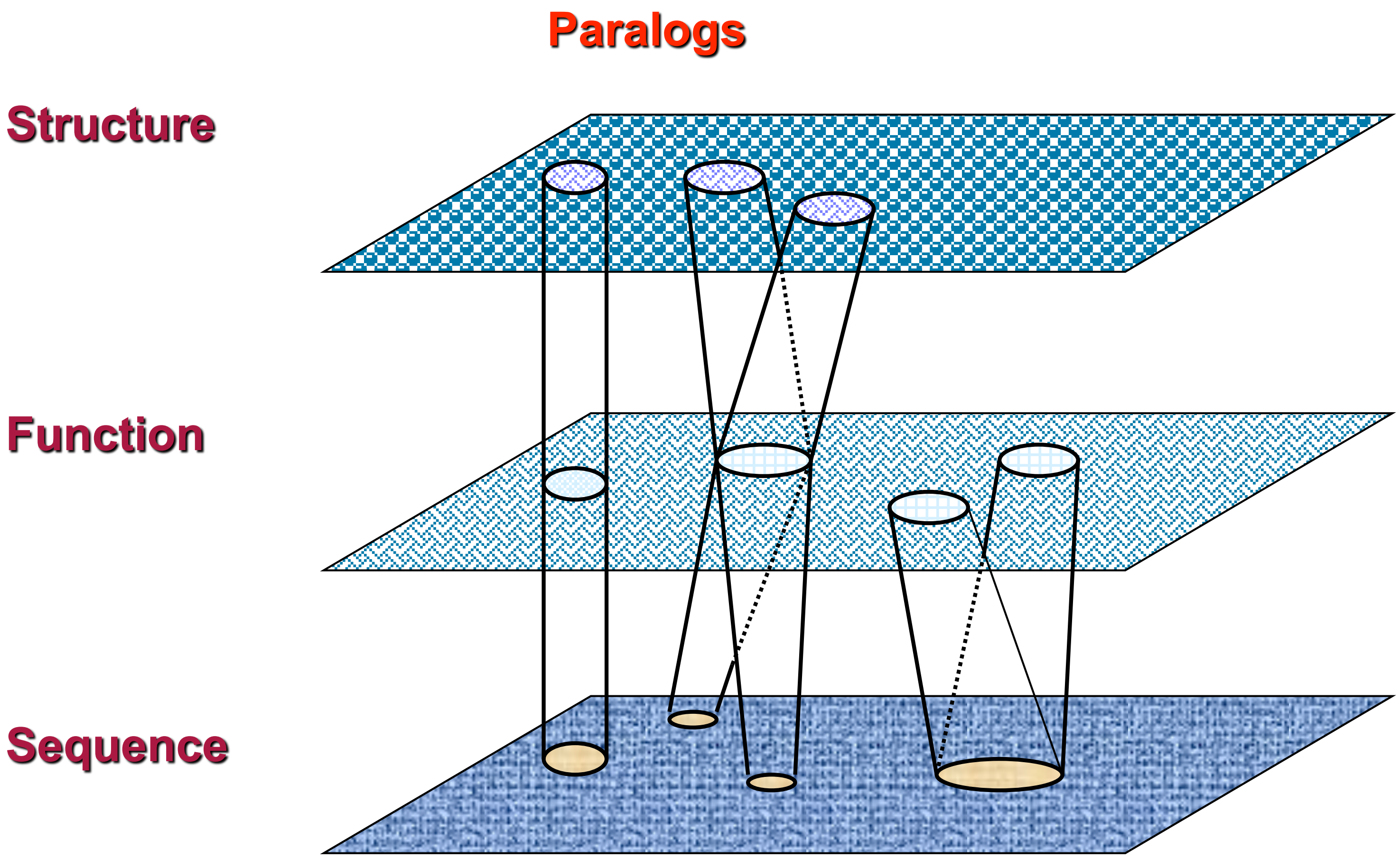


**Homology**: two proteins are homologous if they are the products of genes that evolved from the same ancestor

# 3. Soluble and globular proteins

## 4. Domains

### 3. Function-Sequence-Structure



**Paralogs**

**Structure**

**Function**

**Sequence**

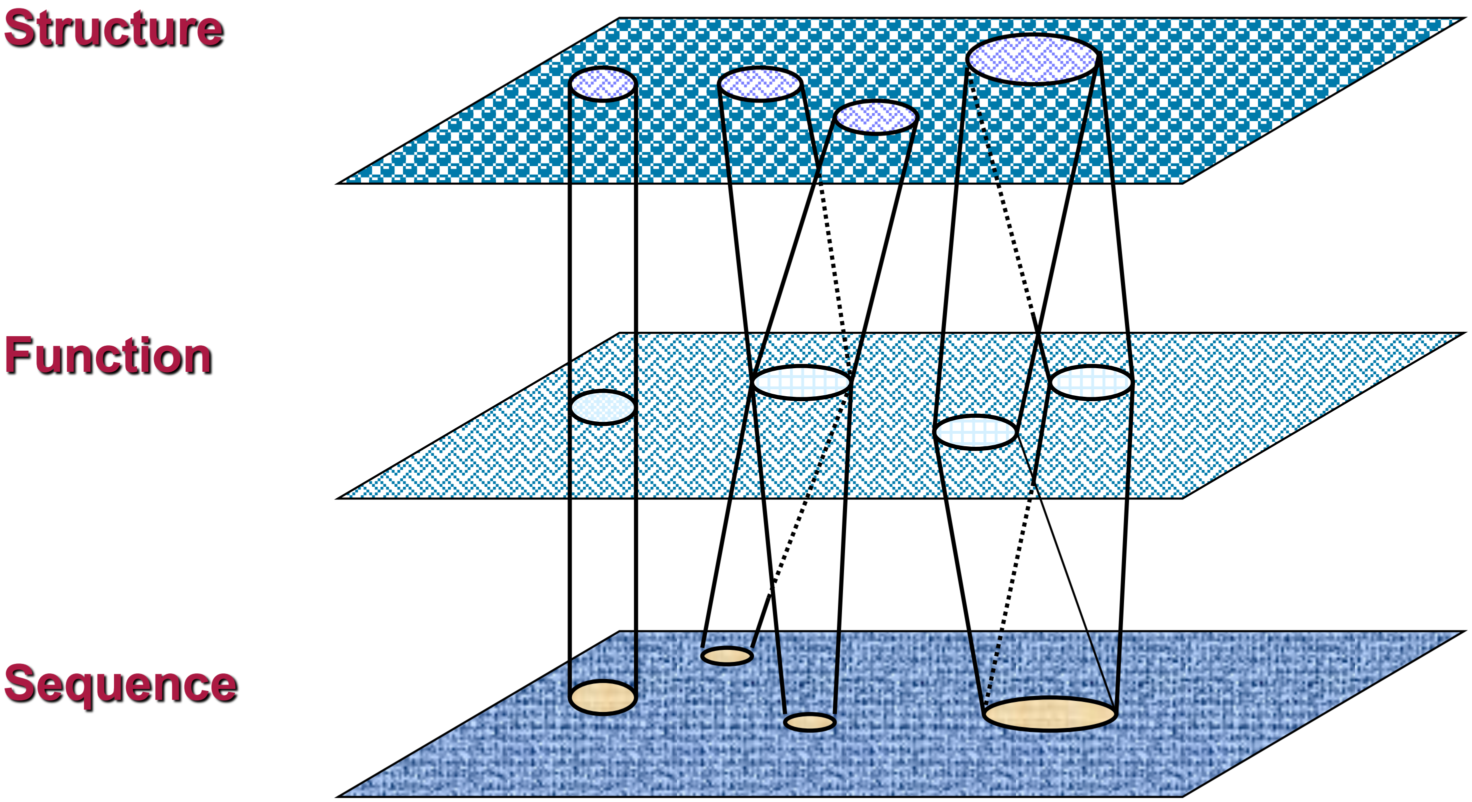
**Homology**: two proteins are homologous if they are the products of genes that evolved from the same ancestor

# 3. Soluble and globular proteins

## 4. Domains

### 3. Function-Sequence-Structure

**Annotation problem**

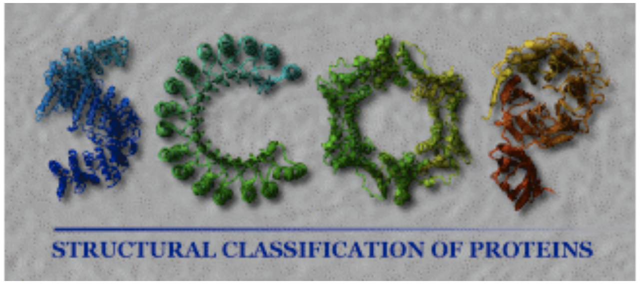


**Homology**: two proteins are homologous if they are the products of genes that evolved from the same ancestor

# 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.

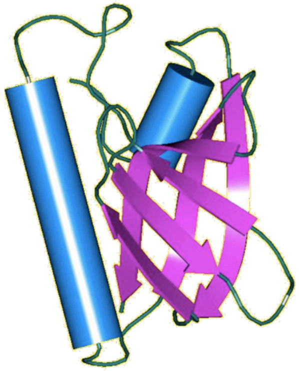
Family: It groups proteins clearly related by homology. In general we assume they are homologs if the alignment has >30% ID, they have the similar structure and similar function

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

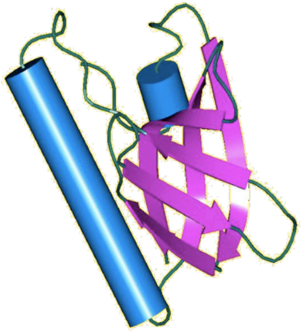
# 3. Soluble and globular proteins
## 4. Domains
### 4. SCOP

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



Enterotoxin

Homology



Cholera toxin

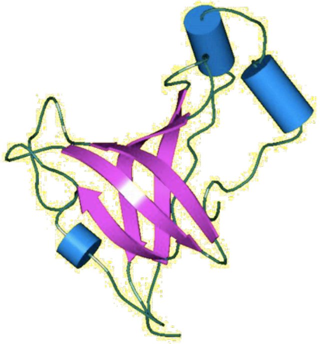
80% Id

Remote Homology

TSS toxin

8.8% Id

Analogs



Aminoacyl tRNA synthetase

4.4% Id

# 3. Soluble and globular proteins

## 4. Domains

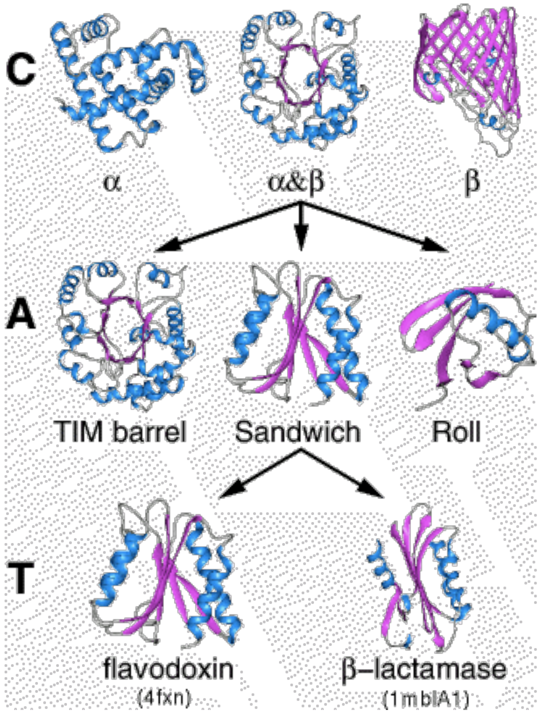
## 5. CATH

- **Class(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.

Class 1
Mainly $\alpha$

Class 3
Mixed
$\alpha$ & $\beta$

Class 2
Mainly $\beta$



- **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%)