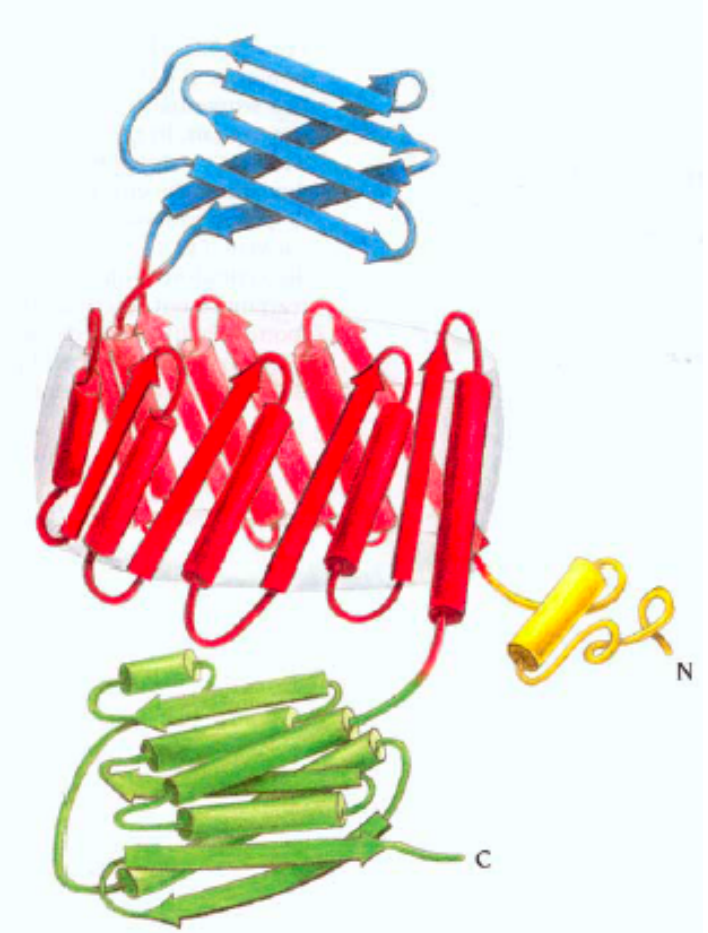
4. Protein Domains

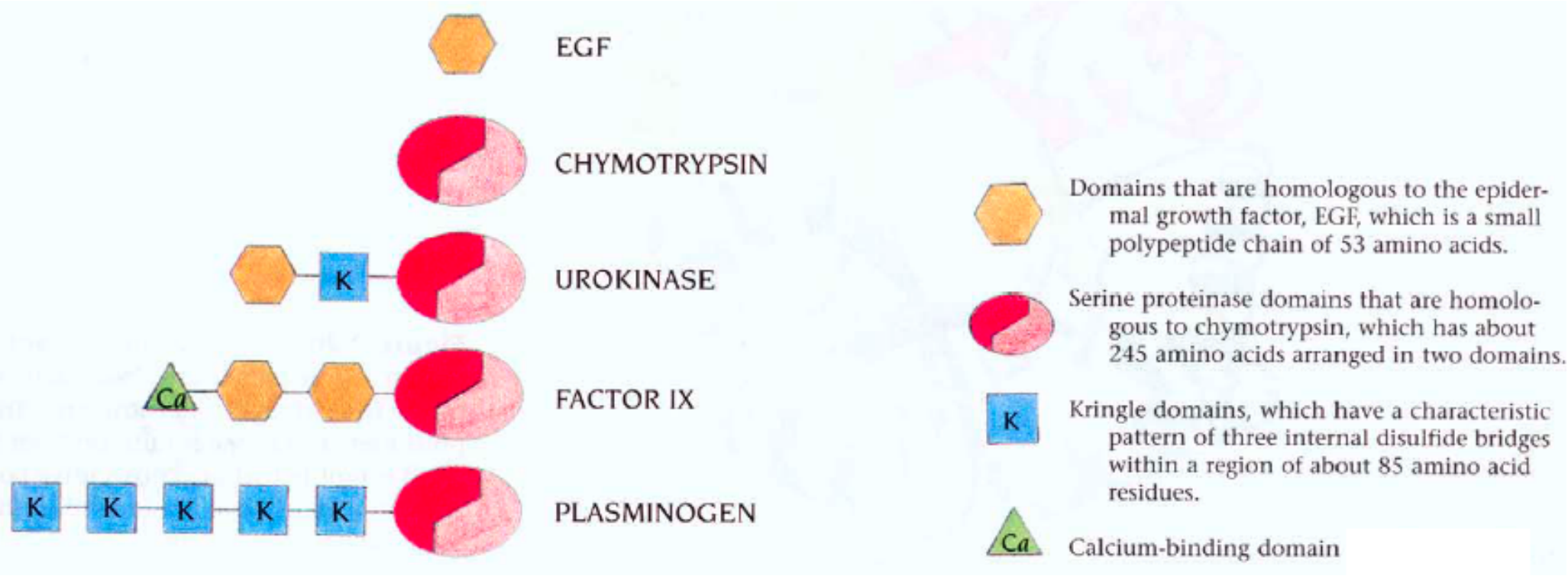
Domain: part of polypeptide chain able to fold independently in a stable structure.
Sometimes are functional units, proteins with different domains usually have combined functions.



Definitions: what are protein domains?

- The polypeptide chain can often fold into one or more distinct regions of structure, called domains.
- Domains are considered as the basic units of folding, function and evolution and often have similar chain topologies (Holm & Sander, 1994).
- Protein domains are often considered as independent or, at the least, semi-independent units, able to fold and in some cases retain function if separated from the parent chain.
- The independent, modular nature of many domains means that they can often be found in proteins with the same domain content, but in different orders, or in different proteins in combination with entirely different domain structures

- Protein can be made by 1 single domain or by many, for example:



- Domains are made of elements of Ilary structure, such as alpha helices and beta strands connected by loops, that constitute “motifs”.
- Adjacent motifs form domains
- The number of combination of motifs that make the domains is limited, some combinations are more favored than other
- Often similar domains occur in different proteins with different aa sequence

Why search for protein domains?

- The identification of domains within a protein sequence is an important precursor for a range of methods.
- Protein structural determination method such as X-ray crystallography and NMR have size limitations which limits their use - they are often employed more successfully when solving smaller domain units rather than whole chains.
- Multiple sequence alignment at the domain level can result in the detection of homologous sequences that are more difficult to detect using a complete chain sequence.
- It is also well known that fold recognition methods perform more reliably if a putative multi-domain target is considered in terms of its constituent domains rather than as a whole chain (Jones & Hadley, 2000).
- All of the method mentioned above are used in order to gain an insight into the structure and ultimately function of a given protein chain. Such results are often best achieved at the domain level.

Protein folds

- Nature in the number of protein fold appears to be limited
- Often proteins have tertiary structures similar to other already known and / or are composed of domains that have tertiary structures similar to others already known, although their composition in the protein may be new
- Nearly all of the fold of the tertiary structures discovered to date appear in many different proteins
- The various combinations of a limited number of fold protein generates protein diversity found in living organisms
- The modular nature of protein structure allows insertions and deletions in the sequence
- Within a domain (or a protein family) insertions / deletions occur at loop without affecting secondary structure elements that form the boundary of the domain

Protein folds and function

- The proteins can be grouped according to the type of fold of its domains ⇒ often (but not always) the function can be recognized by structural and sequence similarities

Example: - a protein containing a kinase domain is almost always kinase

- protein containing a alpha / beta hydrolase domain are almost always hydrolase

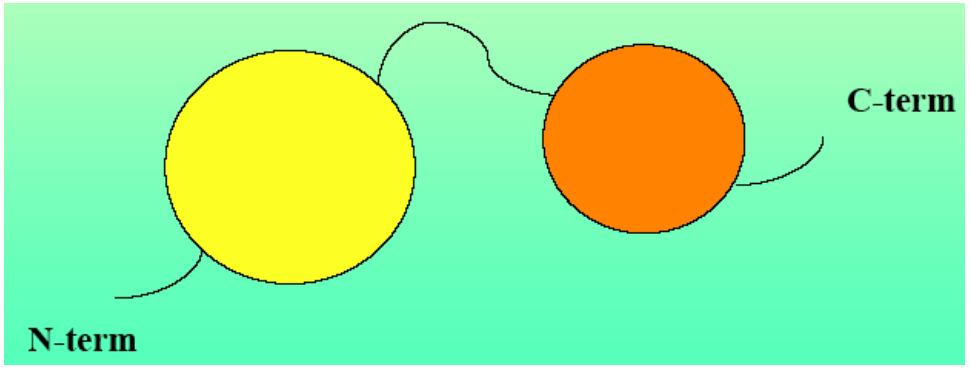
- The coupling overall structure-function in some cases may be weak

Example: - TIM barrel domain: common to many enzymes with very different biochemical functions

Finding protein domains

- The delineation of protein domains within a polypeptide chain can be achieved in several ways:
  - Use of structural data to locate and assign domains:
    - Dali Domain Dictionary (DDD; Dietmann & Holm, 2001),
    - CATH (http://www.cathdb.info/ Orengo et al., 1997)
    - SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/ Murzin et al., 1995)
  - Identification of domains at the sequence level:
    - relies on the detection of global-local sequence alignments between a given target sequence and domain sequences found in databases:
    - Ex. Pfam (Bateman et al., 2000). Difficulties in elucidating the domain content of a given sequence at the sequence homology level arise when searching the target sequence against sequence databases results in a lack of significant matches. In such situations, an ab initio approach to domain assignment from sequence is required. Methods employed by the DomPred server (http://bioinf.cs.ucl.ac.uk/dompred/)

Criteria for the identification of a domain:



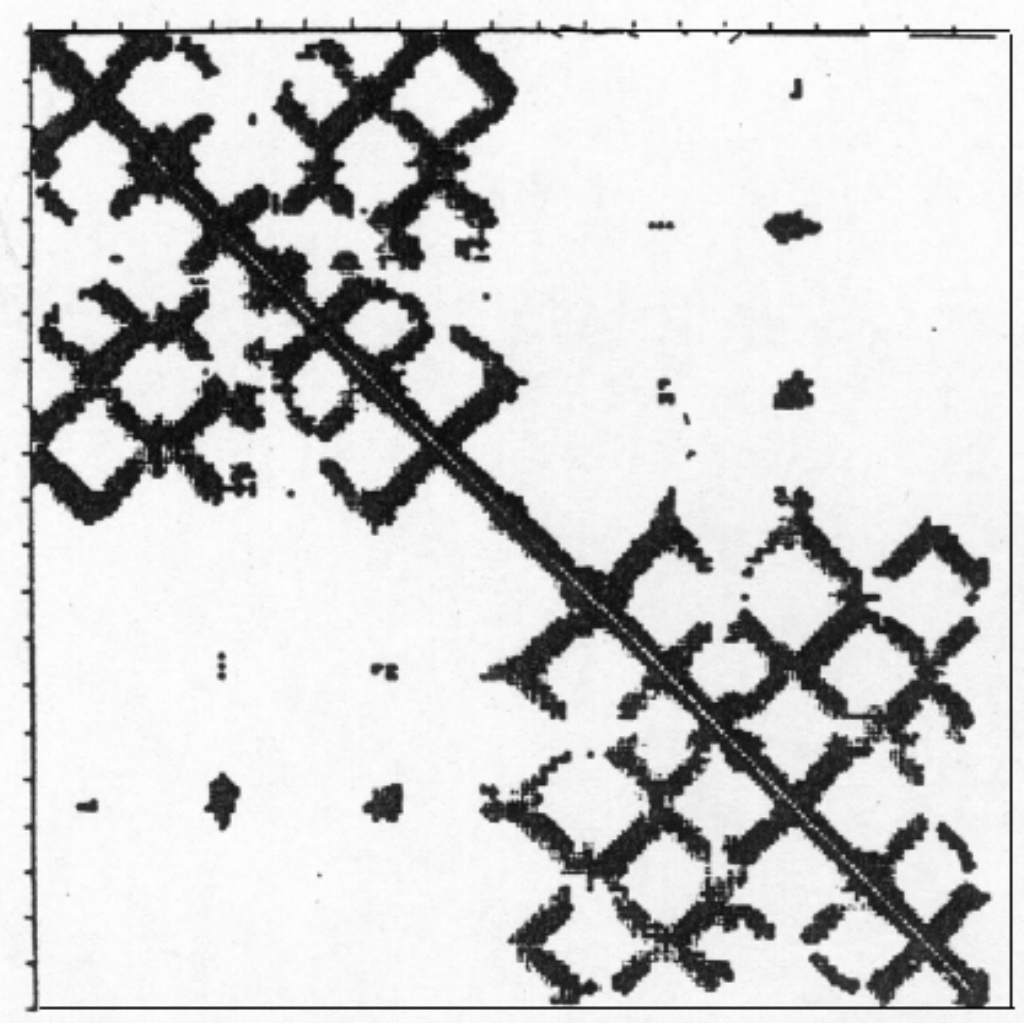
- There are 2 main criteria to define a domain:
  - 1. Diagonal plot
  - 2. Interface parameter (area accessible to the solvent)

1. Diagonal plot

- Procedure:
  - Construction of a squared matrix *(n x n)* where n = n. of aa present in the protein.
  - Measurement of the distance between the C_α for every couple of aa considered
  - Establishment of a threshold under which a point is inserted in the *n x n* matrix (for example 9 Å)
  - The result is a matrix symmetric along the diagonal, where different groups of contacts identify the different domains

Example: Immunoglobulin

- Two domains, one constant and one variable. Zones outside the domains and outside the diagonal represent the points of contacts between the two domains

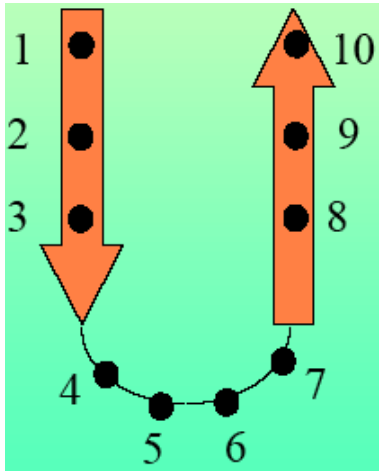


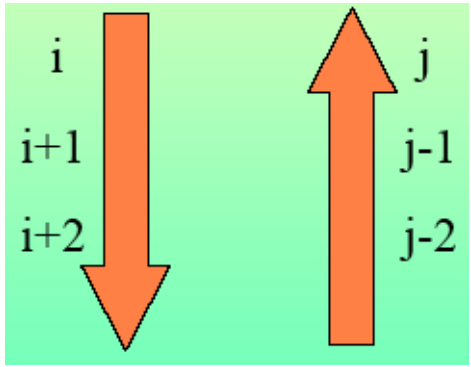
Diagonal plot and Ilary structure

- The diagonal plot can give information on Ilary structure elements.
  - Alpha helices: short range interactions = black dots around the diagonal

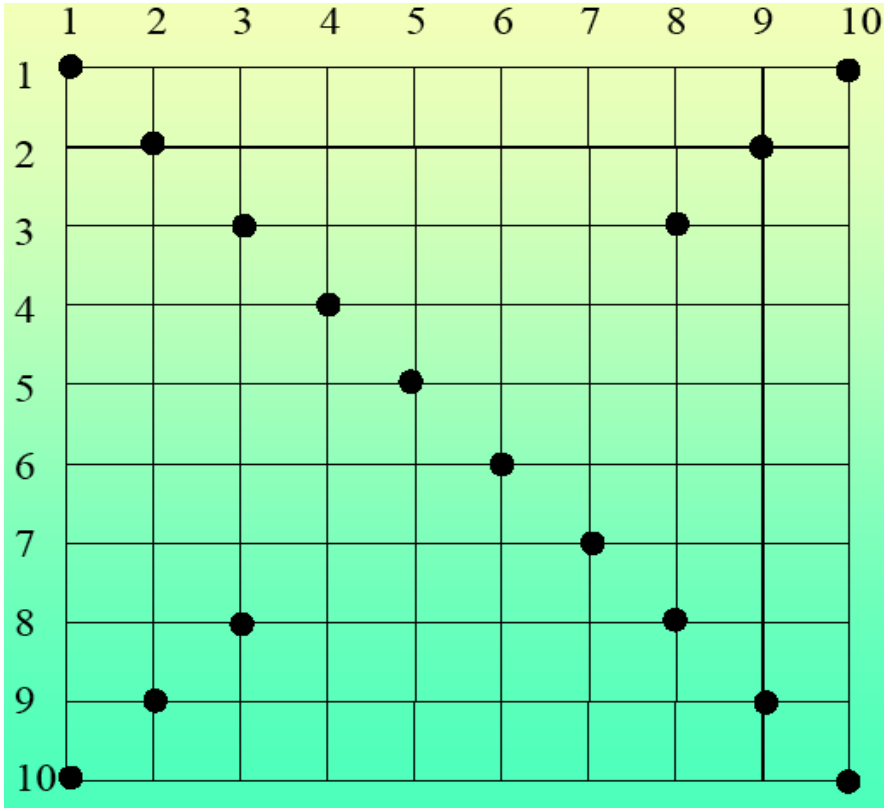
Anti-parallel beta-strands:

- Trend:



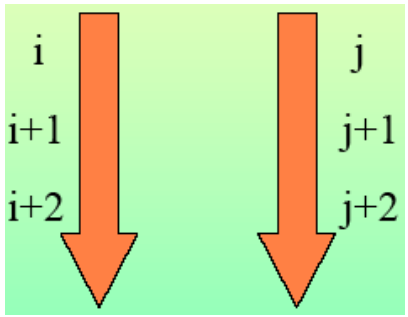


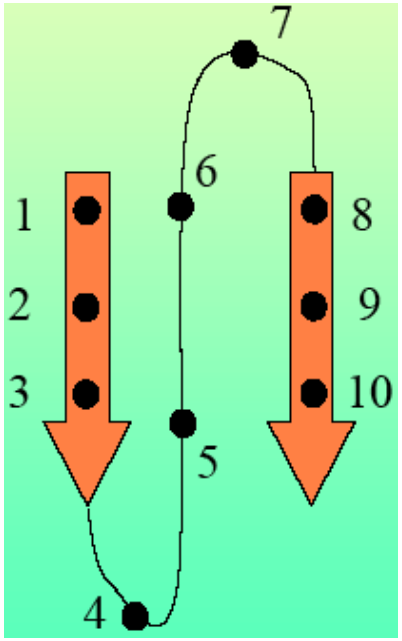
- Dots perpendicular to the diagonal



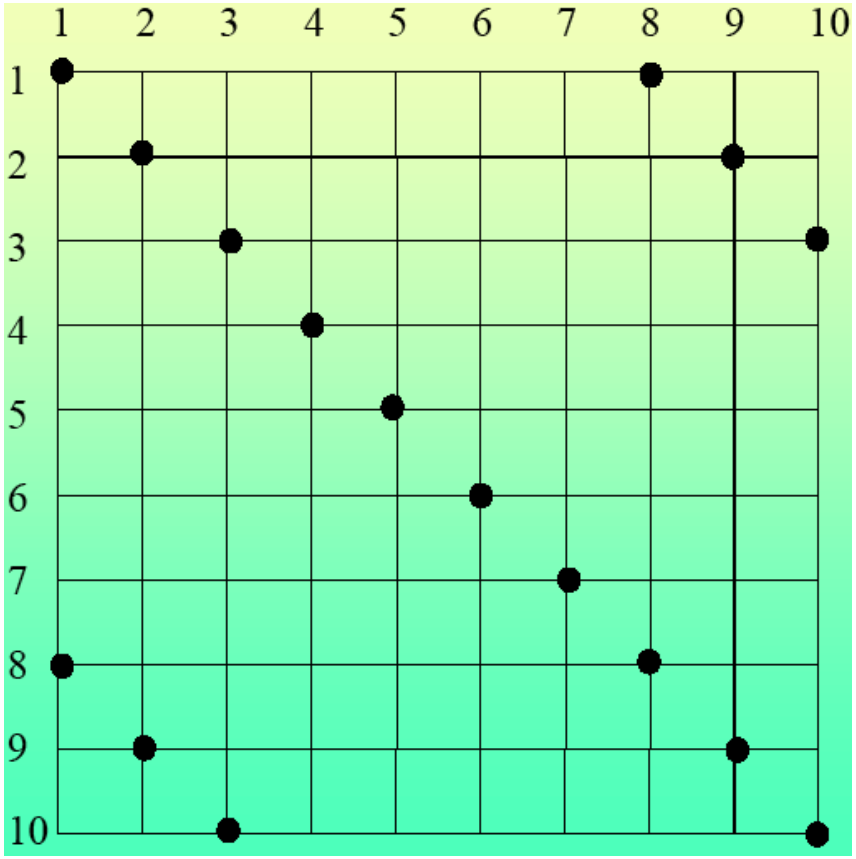
Parallel beta-strands:

- Trend:



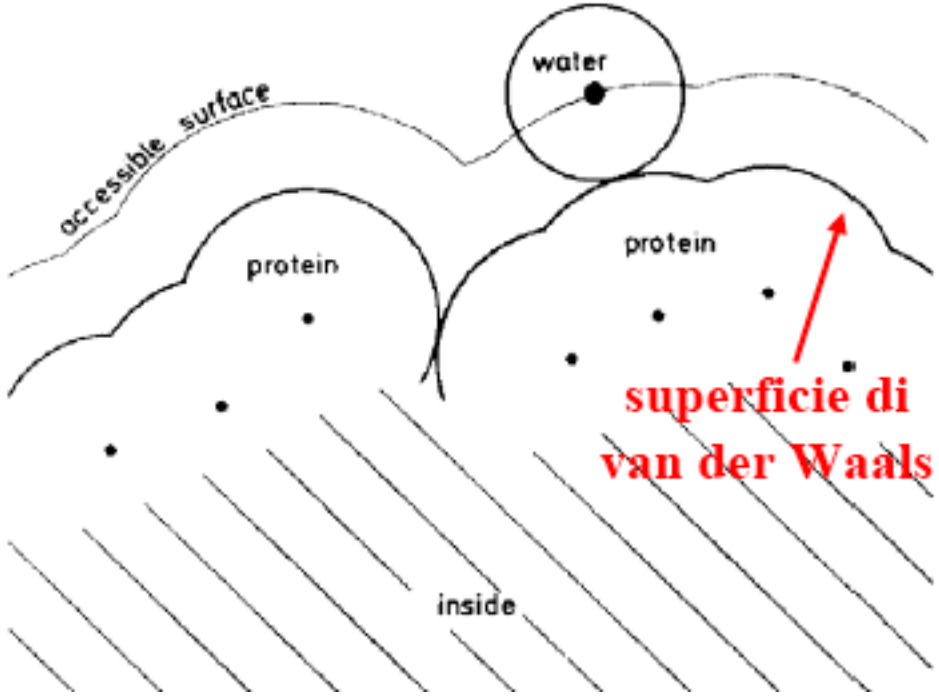


- Dots parallel to the diagonal



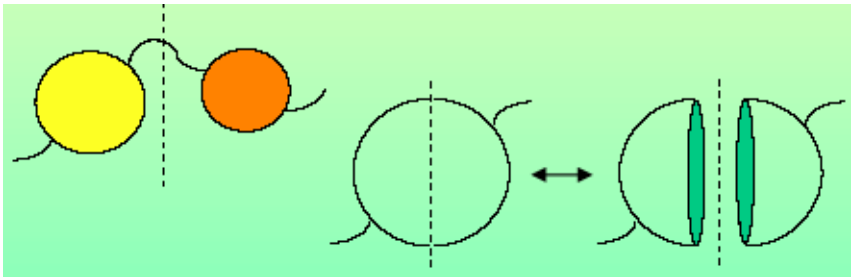
2. Interface parameter

- Procedure
  - This method relies on measurements of the accessible solvent area (ASA) obtained by calculating the area of contact of a spherical probe of radius = 1.4 Å (approx. water molecule).
  - The ASA does not correspond to the van der Waals area of the protein.



- Procedure cont:
  - Suppose the protein is made of *n* aa
  - First we calculate the ASA of the entire protein (A_p)
  - Then we re-calculate the ASA for the protein to which a theoretical cut is applied to one aa at the time (say between aa1 and aa2 first, then between aa2 and aa3 etc etc).
  - At each stage we calculate the ASA of both fragments, A_{pi} and $A_{P(n-i)}$
  - The *interface parameter B* is defined as:

$$B = (A_{pi} + A_{P(n-i)}) - A_P$$



This represents the ASA of the protein that remains exposed as a consequence of the cut

  - The plot of B as a function of n, it is possible to visualize the number of domains and the aa that constitute each domain.

Examples:

IgGL: catena λFAB;

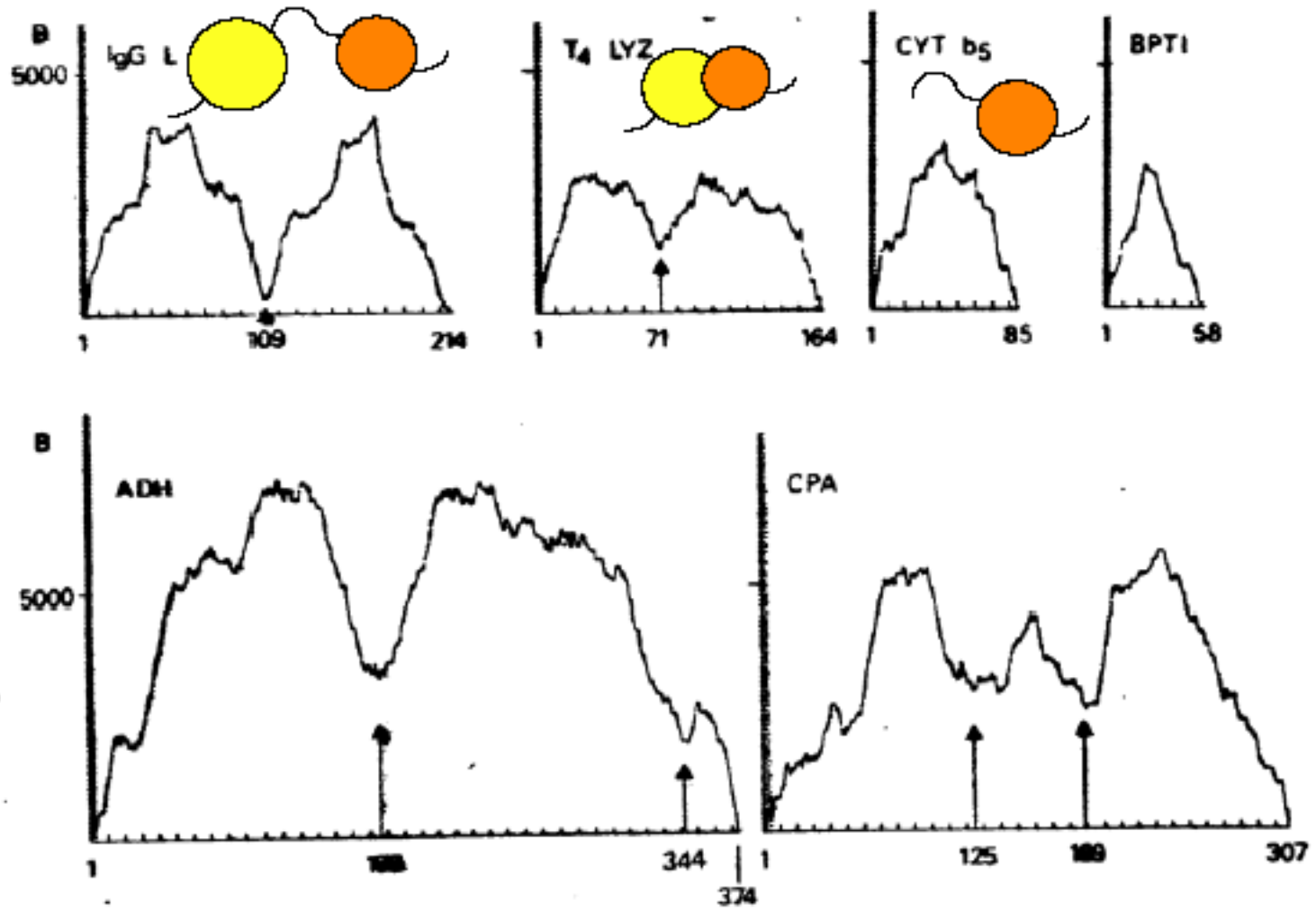
T4 LYZ: lisozima;

CYT *b5: citocromo b5;*

BPTI: inibitore della tripsina bovina;

ADH: alcool deidrogenasi;

CPA: carbossipeptidasiA.



Classification of protein domains

- Michael Levitt and Cyrus Chothia have established a systematic method of classification of protein domains, depending on the llary structure and motifs involved in their formation:
  - Alpha domains
  - Beta domains
  - Alpha/Beta domains
  - Other

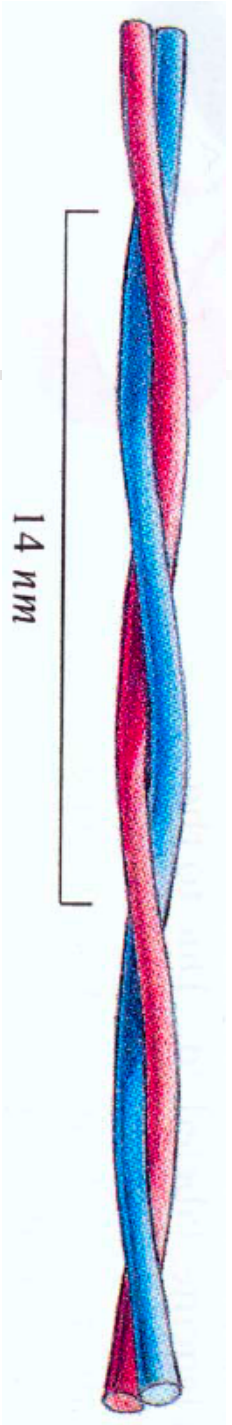
Alpha domains

We will discuss three examples:

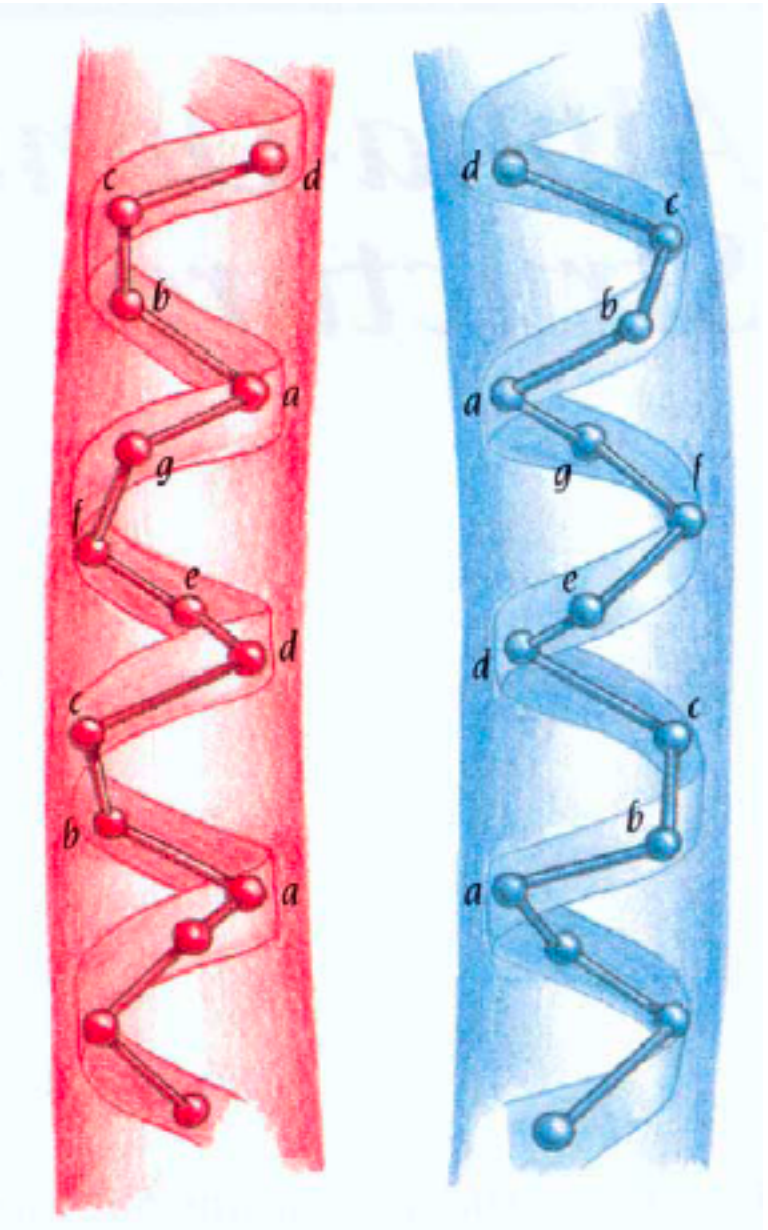
1. The coiled-coil
2. The helical bundle
3. The globin fold

Coiled-coils

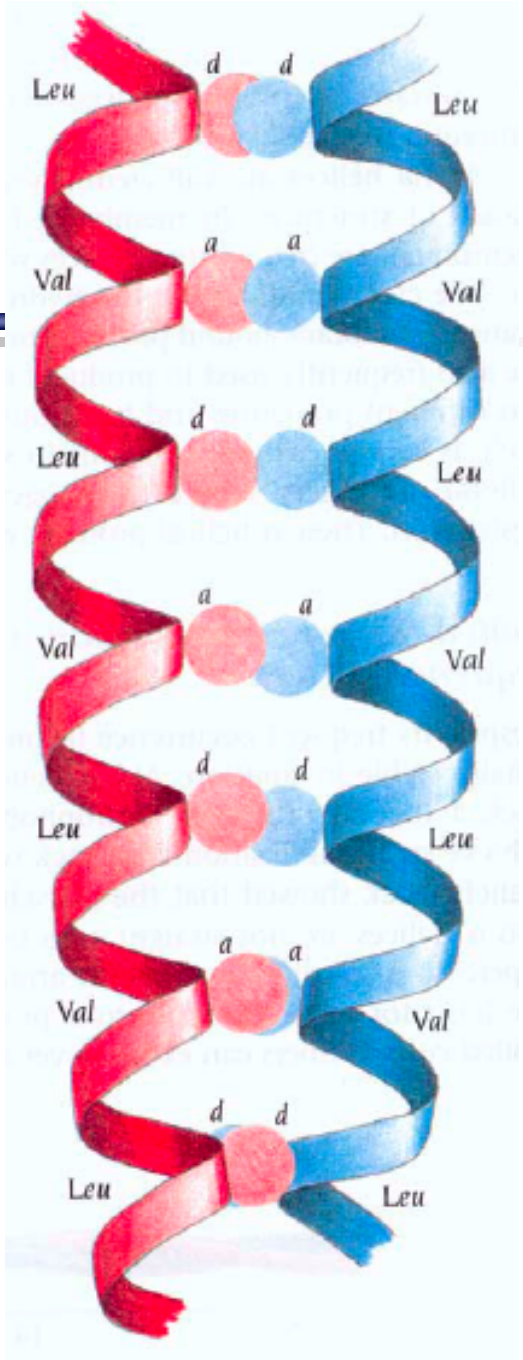
- The most compact and stable structure for alpha helices is *not* to remain isolated, but to maximize interactions between the aa side chains of a pair of helices folded on one another to form a *coiled-coil*
- The coiled-coil are the basis for some fibrous proteins, such as alpha-keratin and myosin, and proteins that bind DNA or RNA



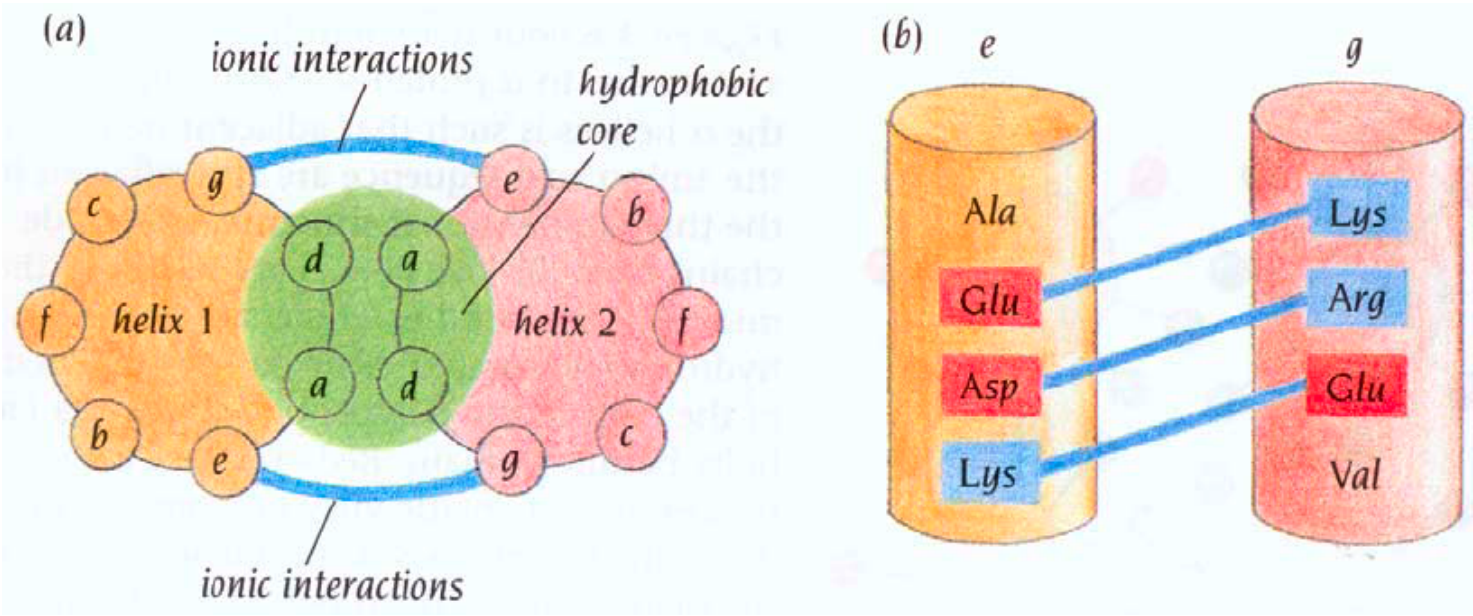
- In the left-handed coiled-coil, two right-handed alpha helices are slightly distorted with 3.5 aa per turn, instead of the standard 3.6.
- The 3.5 aa / turn lead to an identical orientation (and therefore interactions) of the aa side-chains every 7 aa, that is every 2 turns of the coiled-coil
- This asset is called *heptad repeat*, and the aa are labelled by the letters of the alphabet from *a* to *g*.



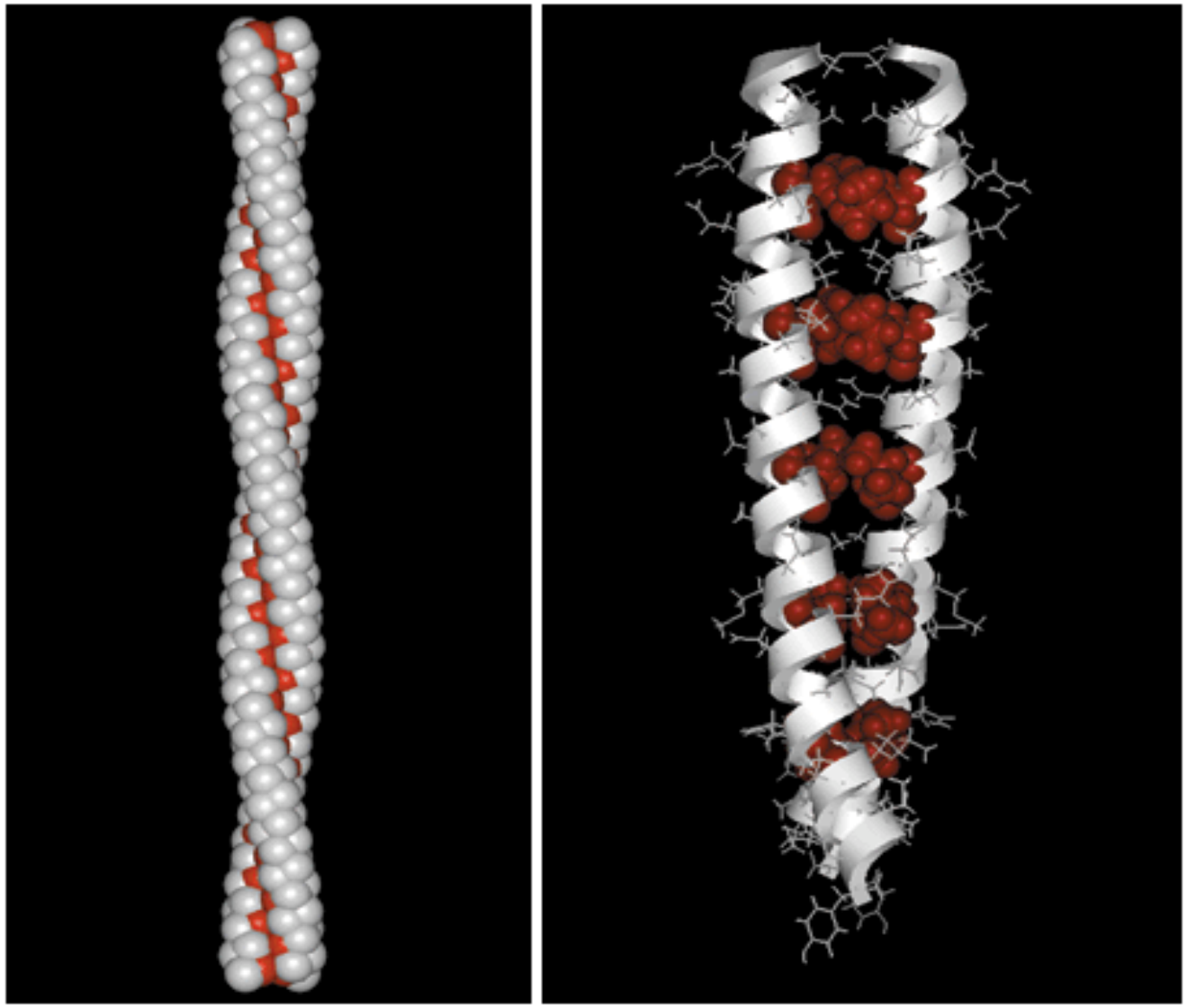
- It follows that if *a* is hydrophobic, then also *d* must be hydrophobic as it is facing the same part of the helix.
- When 2 helices form a coiled-coil, the residue *a* come to pack against each other every 2 turns of the helix; the same happens for *d*
- Therefore residues *a* and *d* form the hydrophobic core of the protein.
- One example of this pattern is the Leucine-zipper, a fold that was first observed in the DNA-binding protein GCN4, shown in the figure.



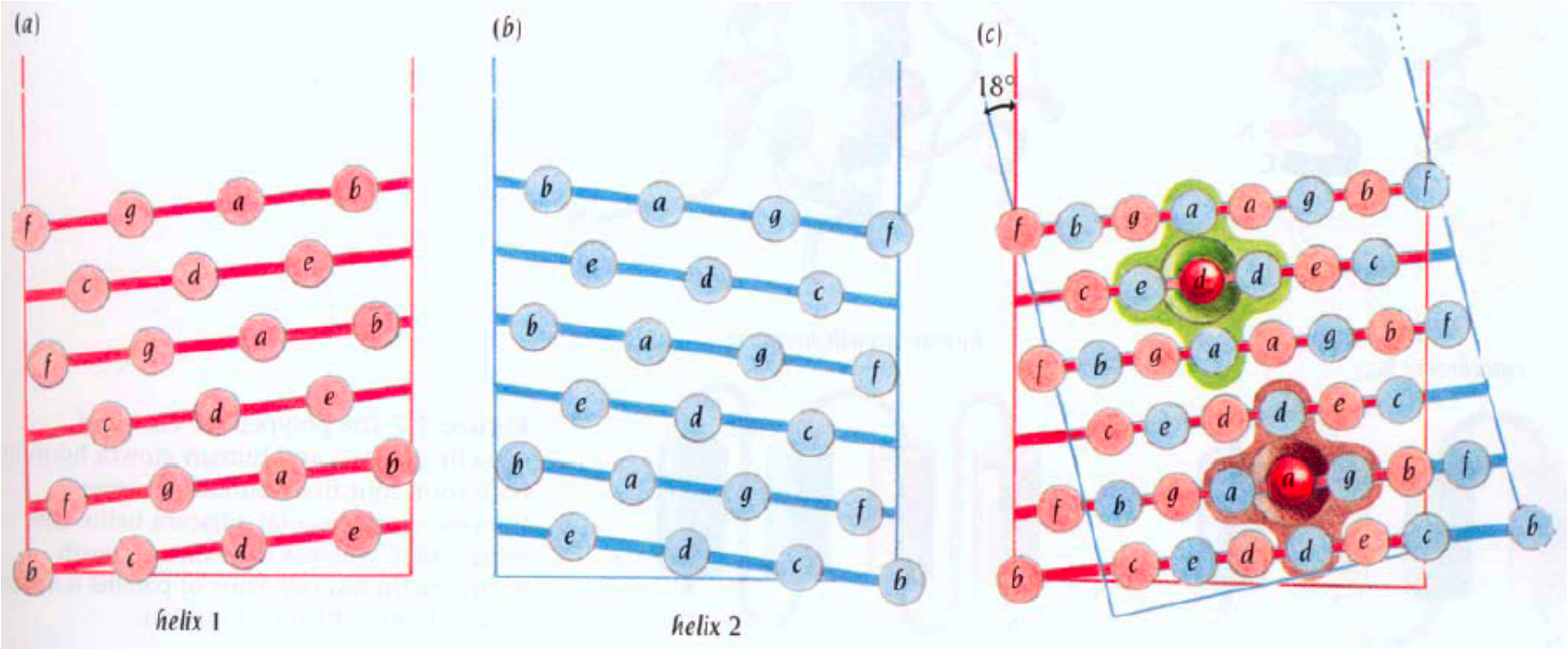
- The aa *e* and *g* will be adjacent to the hydrophobic core made by *a* and *d*. They (*e* and *g*) are usually charged (Glu, Arg, Lys) and they form charge-charge interactions, keeping the alignment and orientation of the coiled-coil:



Leucine Zipper (coiled coil)

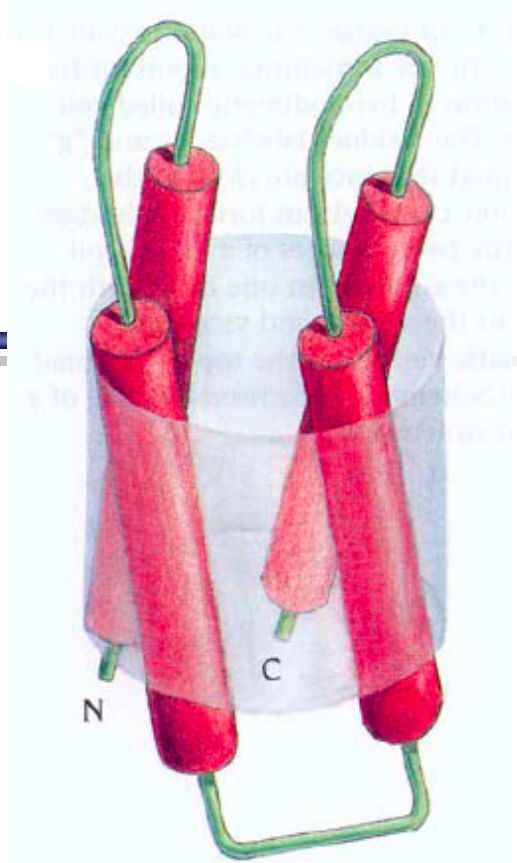


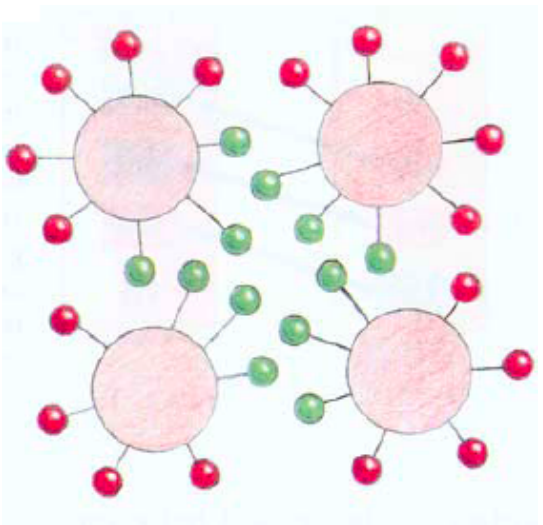
- The two helices of the coiled-coil are packed in such a way that the side-chains of one helix fit into the gaps between the side-chains of the other one (*ridges-into-grooves* model)
- The axis of one helix is tilted by 18° in relation to the other one
- In this way each aa of one helix is surrounded by 4 aa of the opposite helix.



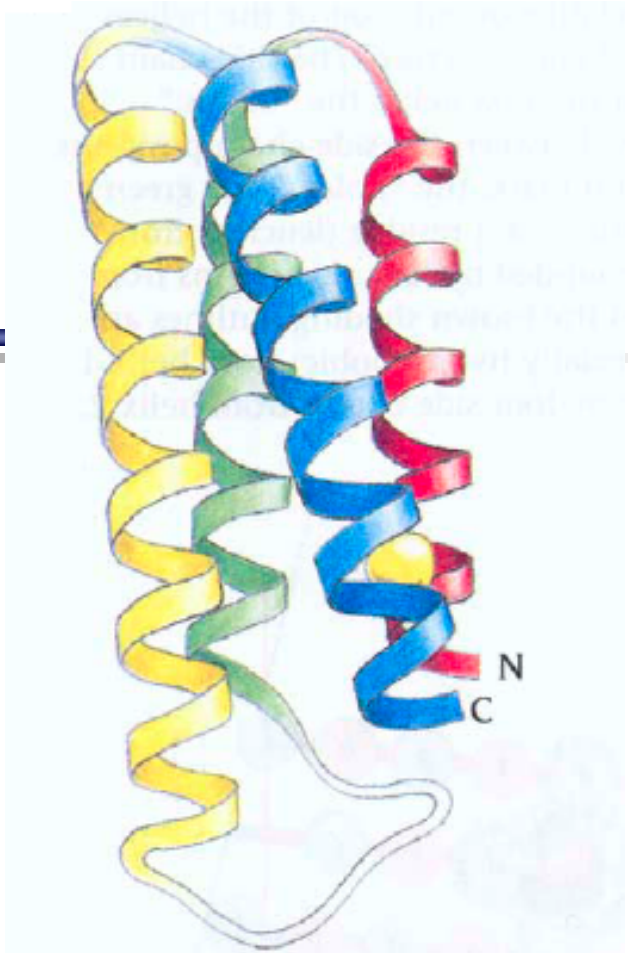
Alpha-helical bundle

- Composed of 4 helices with axes tilted by 20°
- Hydrophobic side-chains are oriented inside, the hydrophilic outside
- The hydrophobic core is tightly packed so that water molecules can't be present

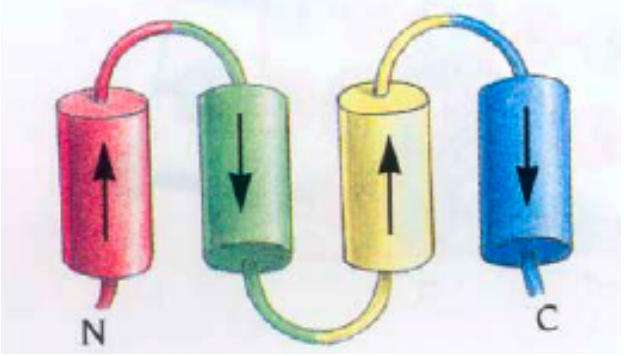




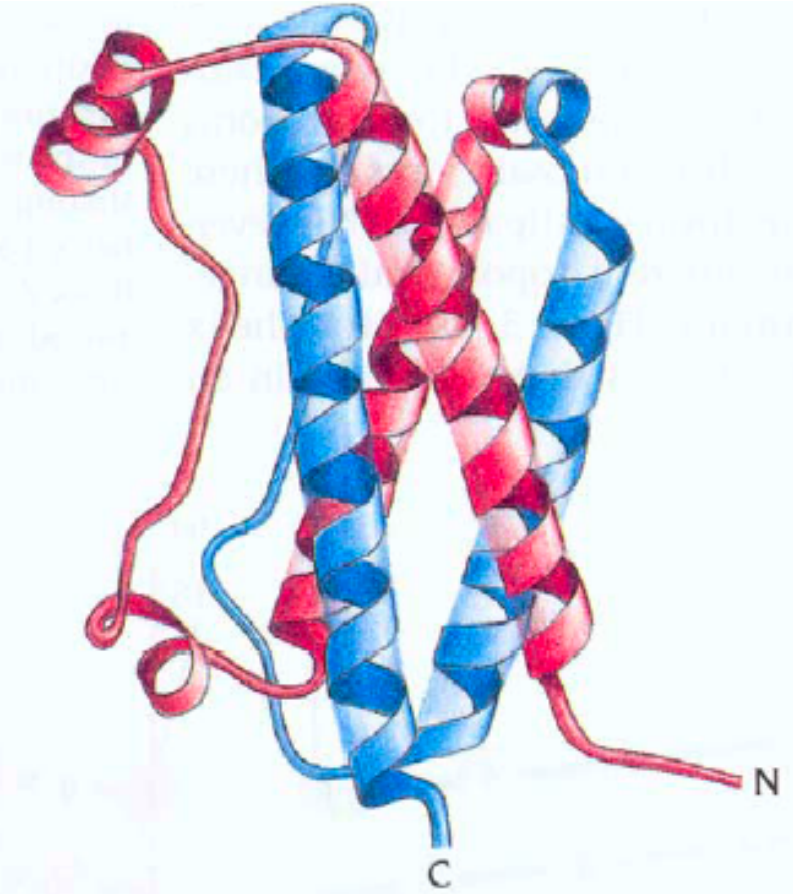
- The alpha-helical bundle is very common in many proteins, for example
  - Mioemerytrin, cyt *c*′, cyt b_{562}, ferritin, tobacco mosaic capsid protein
- Nearly always the helices are anti-parallel



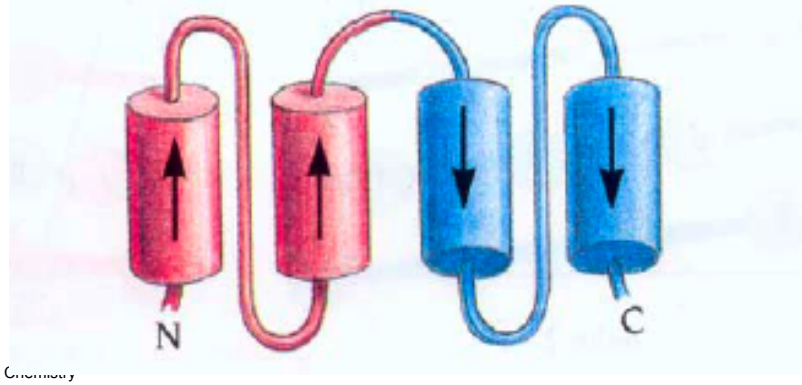
cytochrome b_{562}



- One example of the rare parallel arrangement is the human growth hormone
- Antipallel helices are more stable because of the compensation of the helices macrodipoles

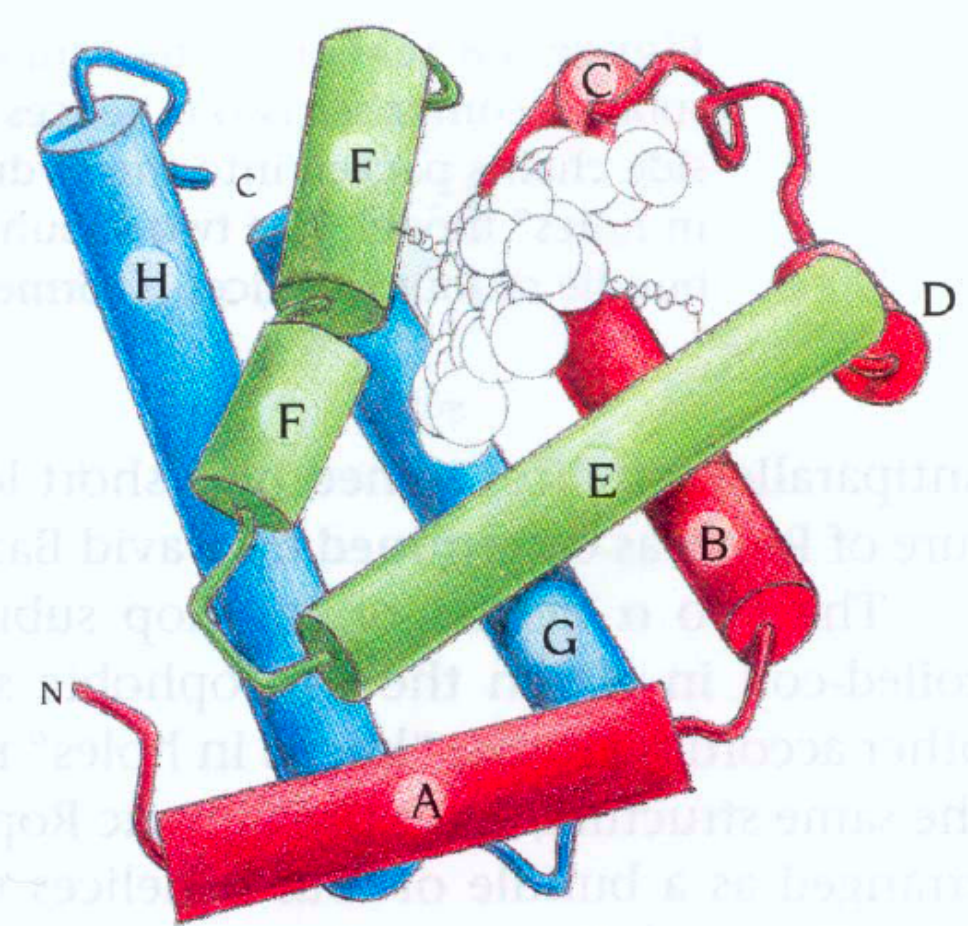


human growth hormone

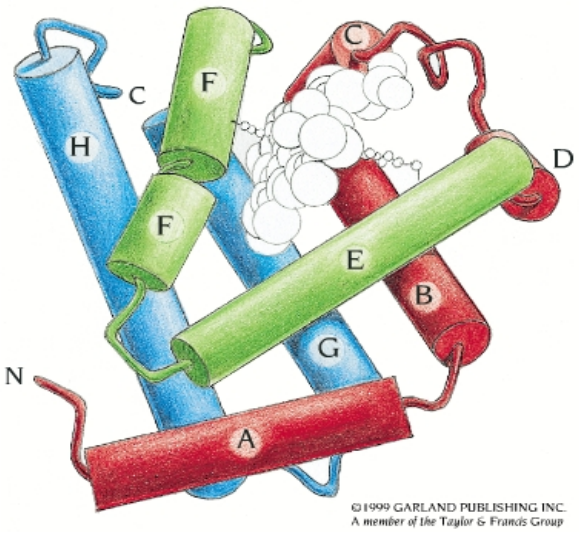


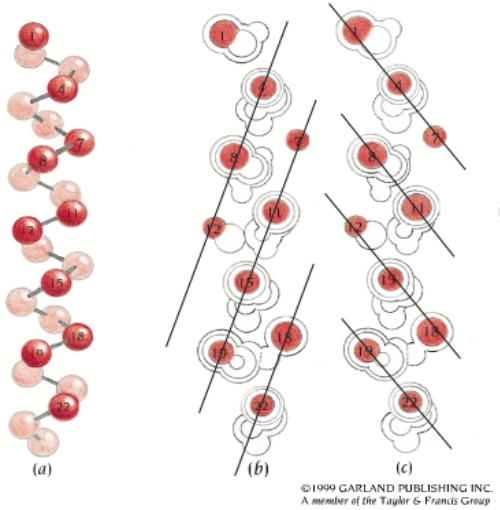
Globin fold

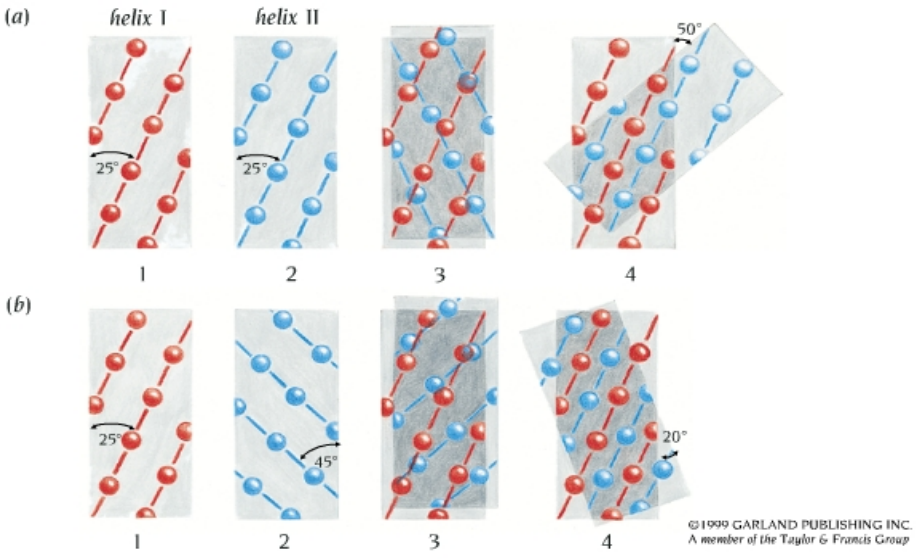
- Constituted by 8 alpha-helices labelled from A to H
- Connected by short loops
- Normally define an hydrophobic pocket
- Helices length varies from 7 aa to 28 aa
- The helices have diverse orientations
- Examples:
  - Myoglobin, hemoglobin, ficocianins



Globin fold

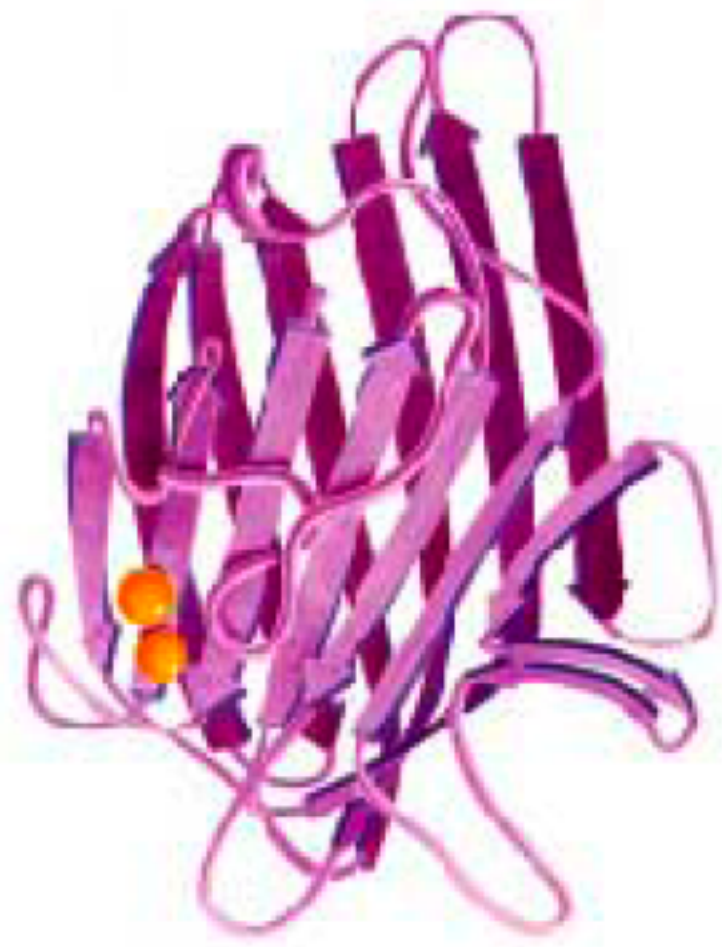






The helices are not parallel but they form a 20° (a) or a 50° angle (b) to adapt ridges-into-grooves

Beta domains



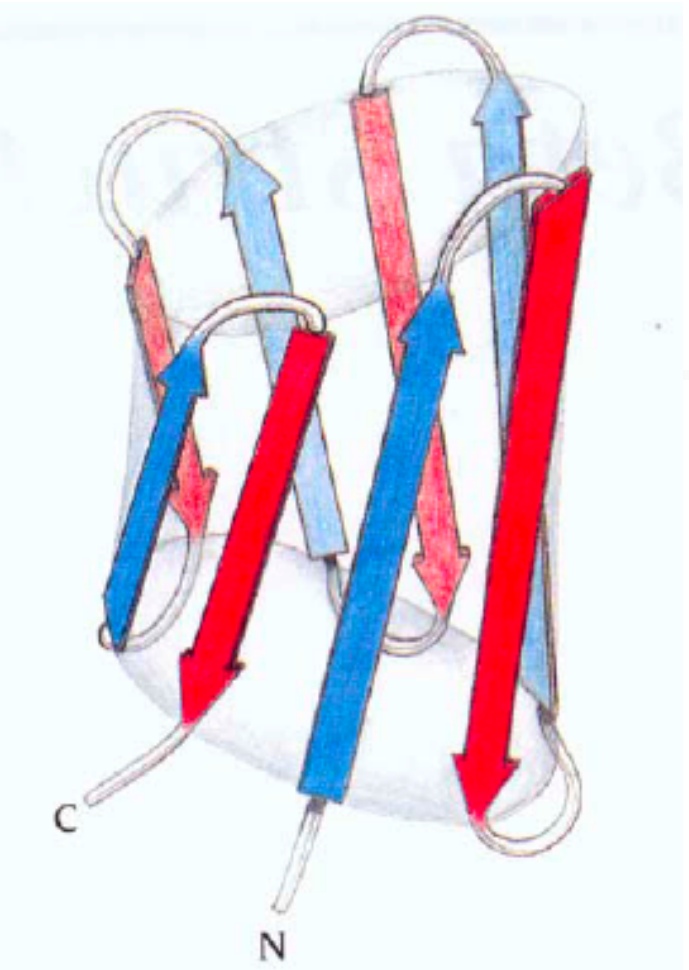
Beta-domains

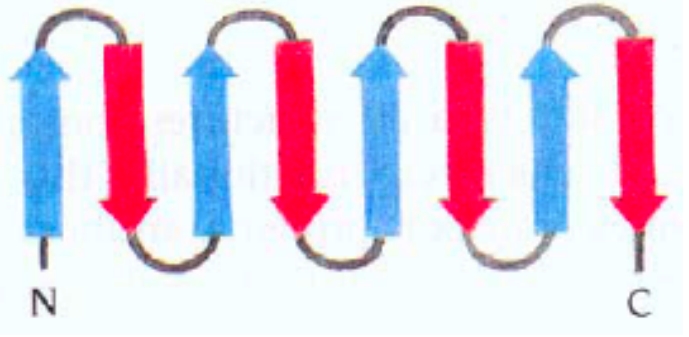
- There are many combinations of beta-strands to form anti-parallel beta-sheets connected by loops, but the most commonly found domains are 3:
  - Up and down
  - Greek key
  - Jelly roll
- Often beta-strands form 2 beta-sheets packed one against the other with an hydrophobic core in between.
- This creates amphipatic sheets, with one face hydrophobic and another hydrophilic
- Another behaviour is seen for the beta helix, that we will see separately

- Two antiparallel sheets can pack against each other to form:
  - Beta-barrel
    - This is a cylindric structure where all strands form hydrogen bonds with one another. The last strands forms hydrogen bonds with the first, closing the barrell.
  - Beta-sandwich
    - Here 2 sheets pack against each other in a way that no hydrogen bonds are formed between the two sheets, but the interactions occur between the side chains of the residues.

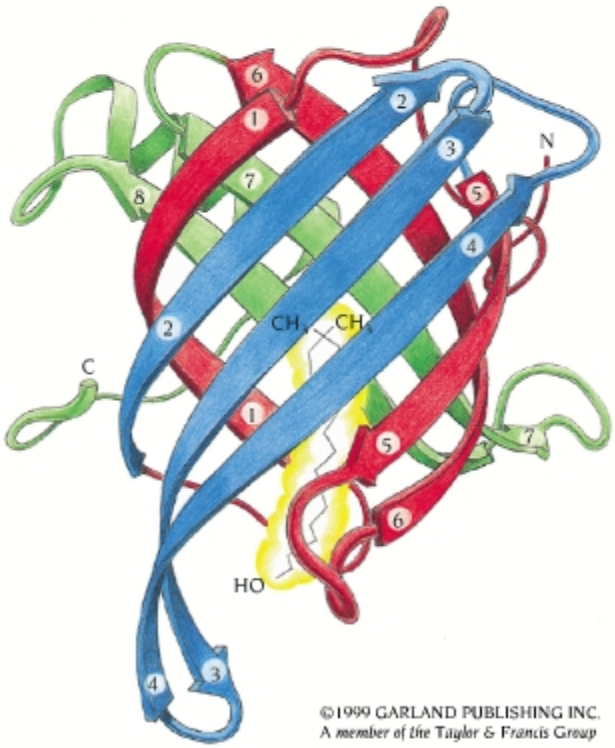
Up and down: beta barrels

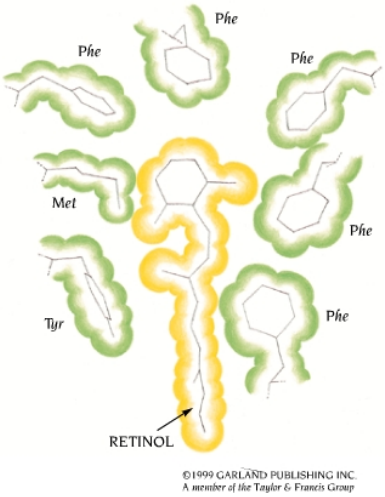
- In the up and down beta domains the beta-strands are connected by short loops or beta-hairpins.
- When the last beta-strands forms hydrogen bonds with the first one we have the beta-barrell
- Usually there are 8 strands in a beta-barrell, such as in the retinol binding protein and the fatty acid binding protein
- Usually the external part of the barrell is formed by hydrophilic residues and the internal part by hydrophobic ones





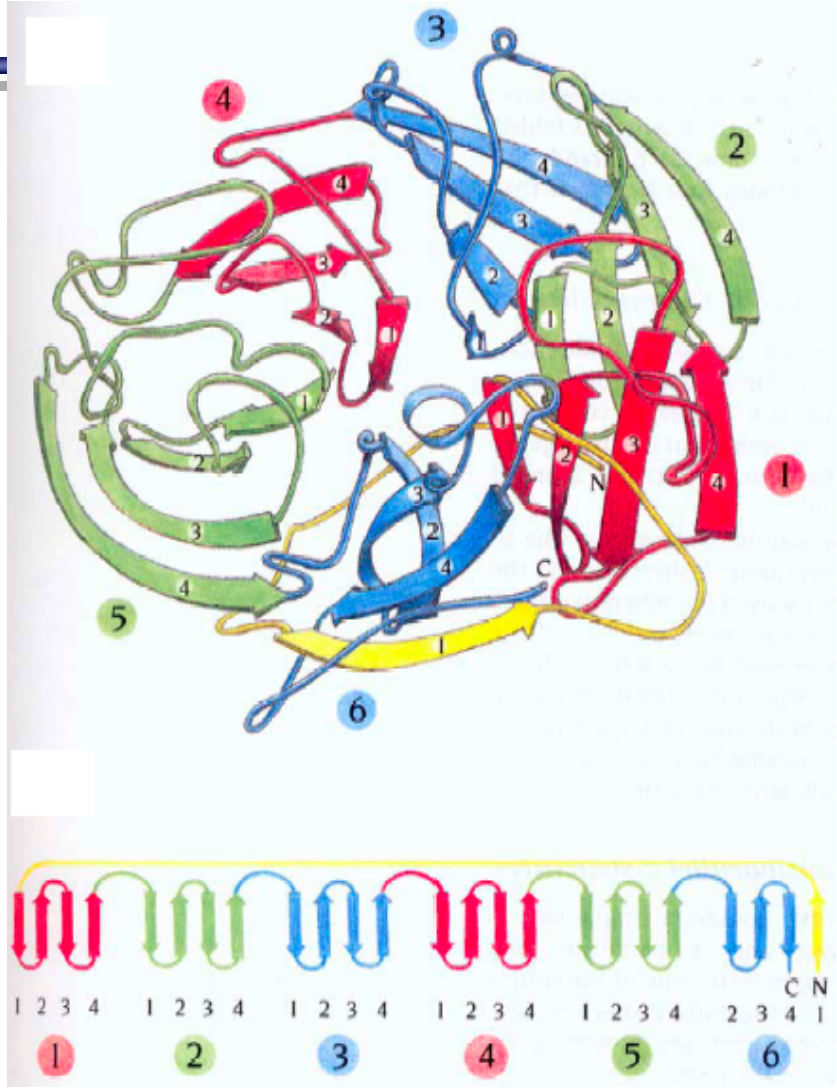
- The retinol binding protein is the carrier protein involved in the transport of retinol (vitamin A alcohol) from the liver storage site to peripheral tissue. Vitamin A is a fat-soluble vitamin necessary for growth, reproduction, differentiation of epithelial tissues, and vision.





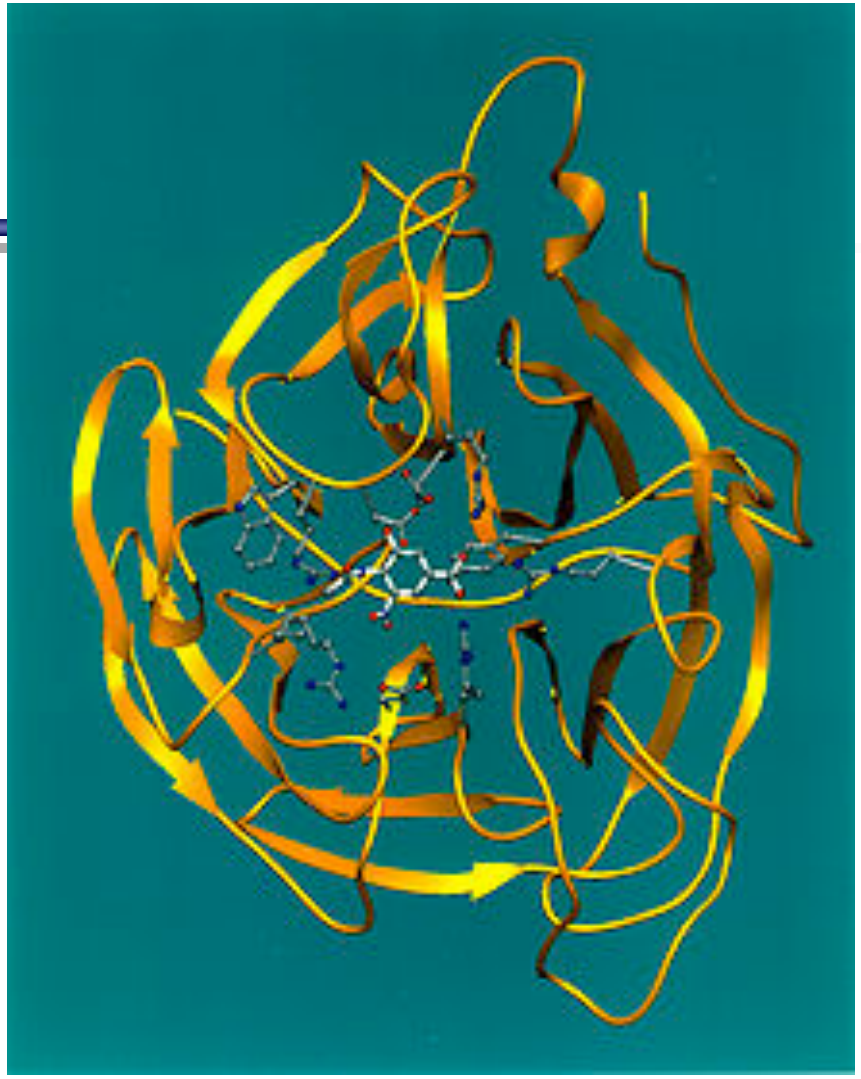
Up and down beta-propeller

- Here 4 beta-strands do not form a barrel but they form 6 sheets that are organised as a propeller
- The loops of every beta-sheet that connect the beta-strands 2 and 3 are all on the same side of the propeller and they form the active site of the protein



Beta-propeller

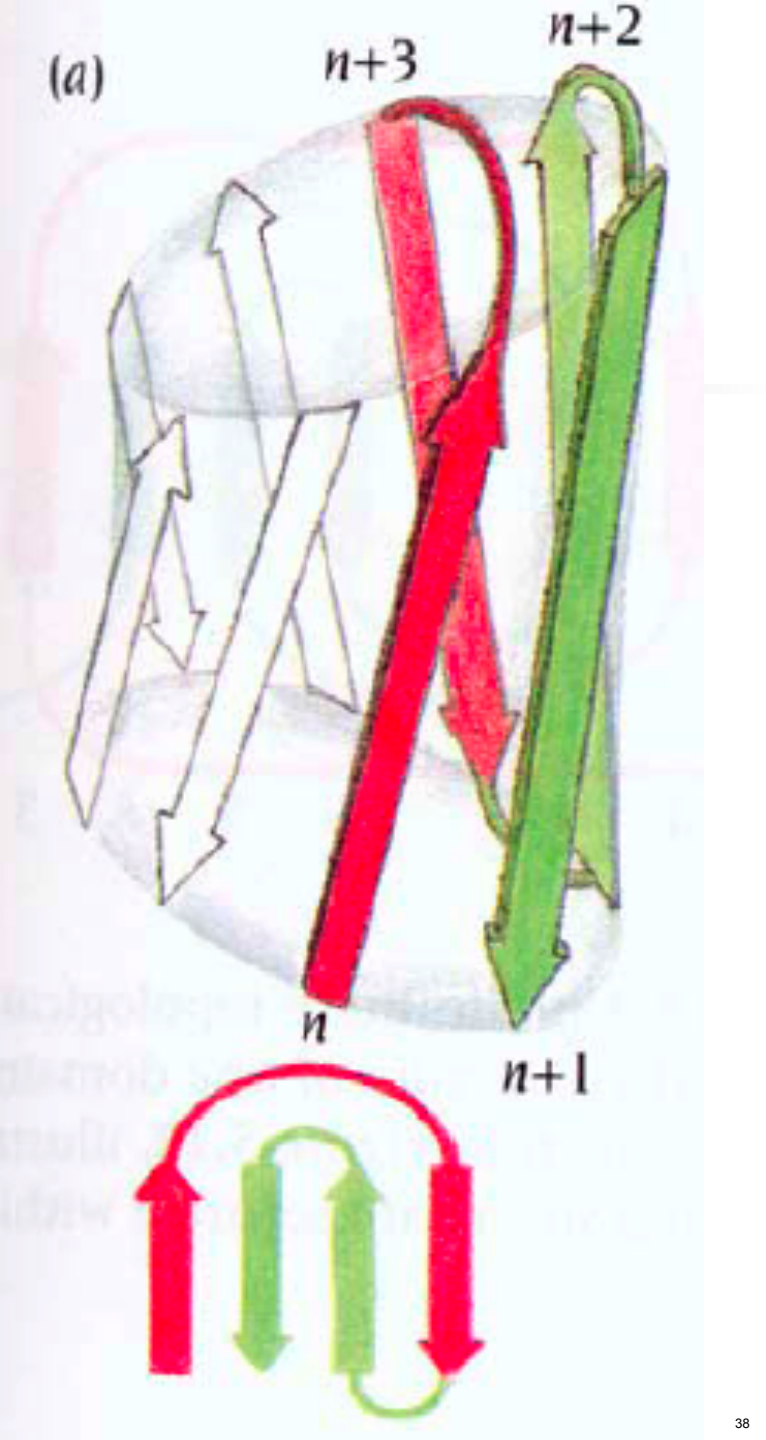
- One example of beta-propeller is the **neuraminidase** of the influenza virus. **Neuraminidase** enzymes are glycoside hydrolase enzymes which cleave the glycosidic linkages of neuraminic acids.
- The most commonly known neuraminidase is the viral neuraminidase, a drug target for the prevention of influenza infection.
- The viral neuraminidases are frequently used as an antigenic determinants found on the surface of the Influenza virus. Some variants of the influenza neuraminidase confer more virulence to the virus than others.
- Neuraminidases, also called sialidases, catalyze the hydrolysis of terminal sialic acid residues from the newly formed virions and from the host cell receptors.Sialidase activities include assistance in the mobility of virus particles through the respiratory tract mucus and in the elution of virion progeny from the infected cell.



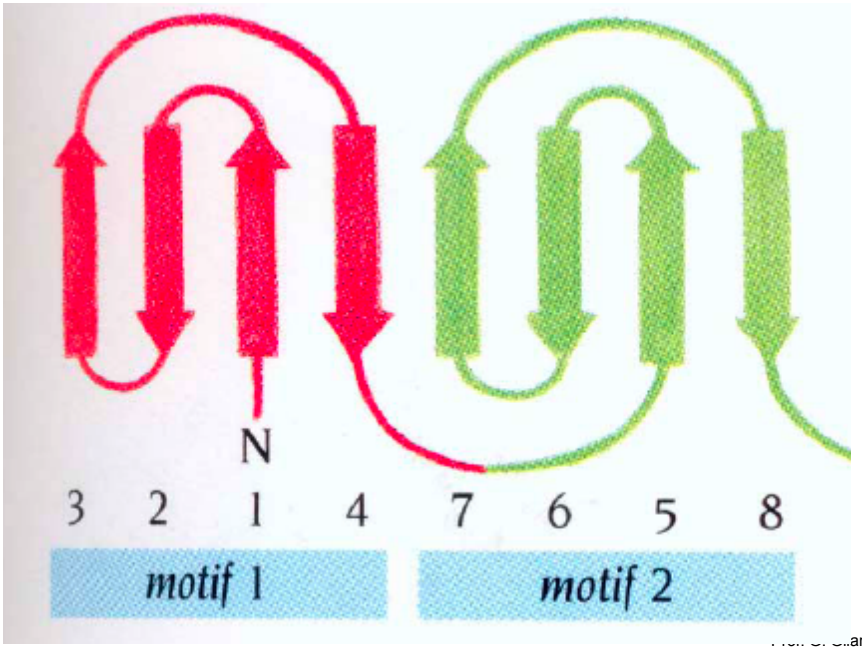
Crystallographic structure of influenza A N9 neuraminidase and its complex with the inhibitor 2-deoxy 2,3-dehydro-N-acetyl neuraminic acid

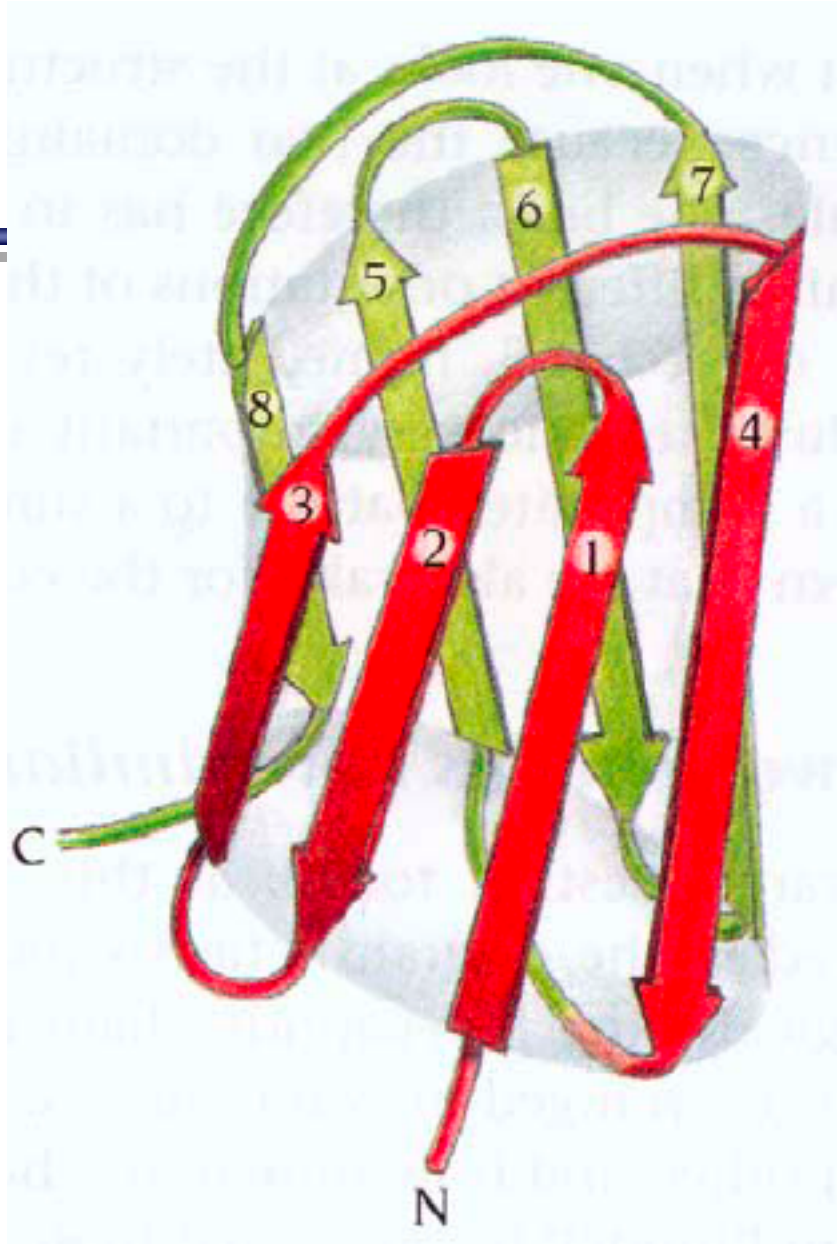
Greek key

- This is a particular type of beta-barrel, where the Greek key connect beta-strands on opposite sides of the barrel



- One example of the Greek key is the fold of the Cu-Zn superoxide dismutase



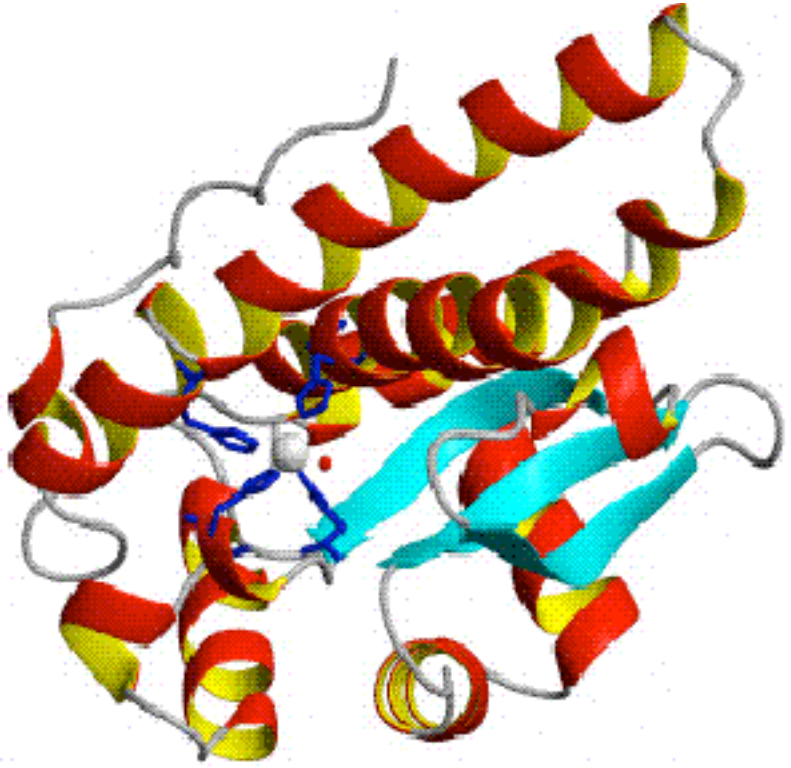


- **Superoxide dismutases** (**SOD**) are a class of enzymes that catalyze the dismutation of superoxide into oxygen and hydrogen peroxide. As such, they are an important antioxidant defense in nearly all cells exposed to oxygen.
- The SOD-catalysed dismutation of superoxide may be written with the following half-reactions :

$$M^{(n+1)+} - SOD + O_2^- \rightarrow M^{n+} - SOD + O_2$$

$$M^{n+} - SOD + O_2^- + 2H^+ \rightarrow M^{(n+1)+} - SOD + H_2O_2.$$

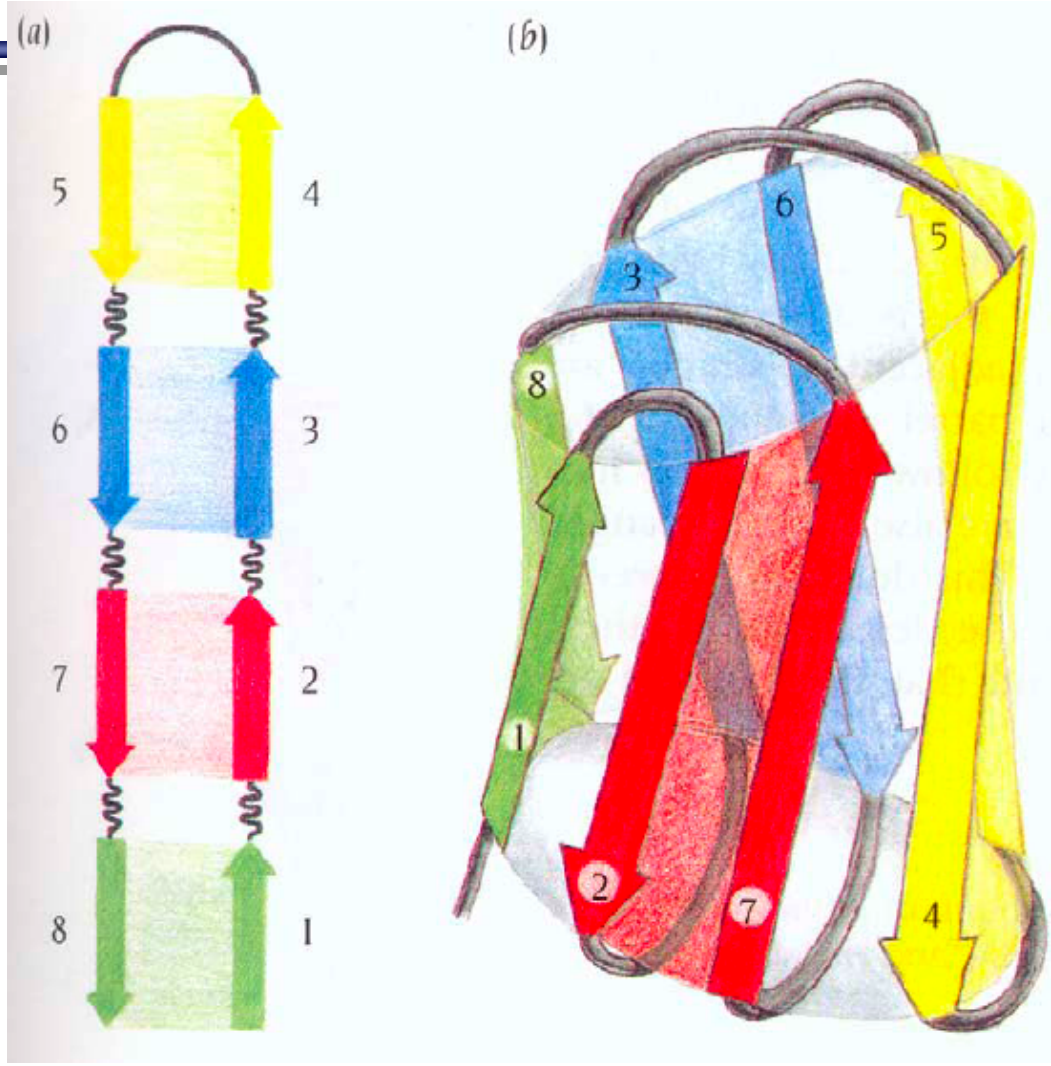
where M = Cu (n=1) ; Mn (n=2) ; Fe (n=2) ; Ni (n=2).



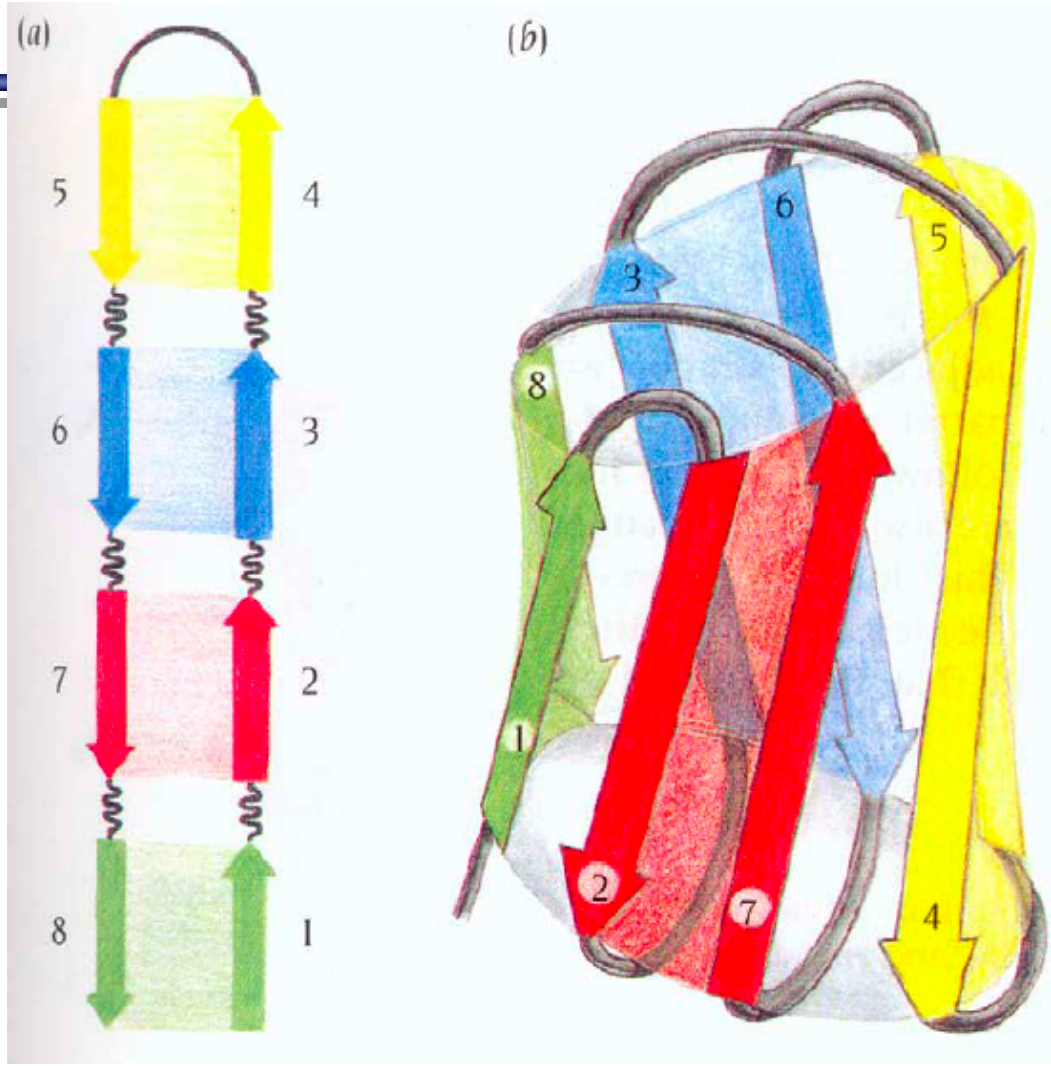
Structure of the monomeric unit of human superoxide dismutase 2.

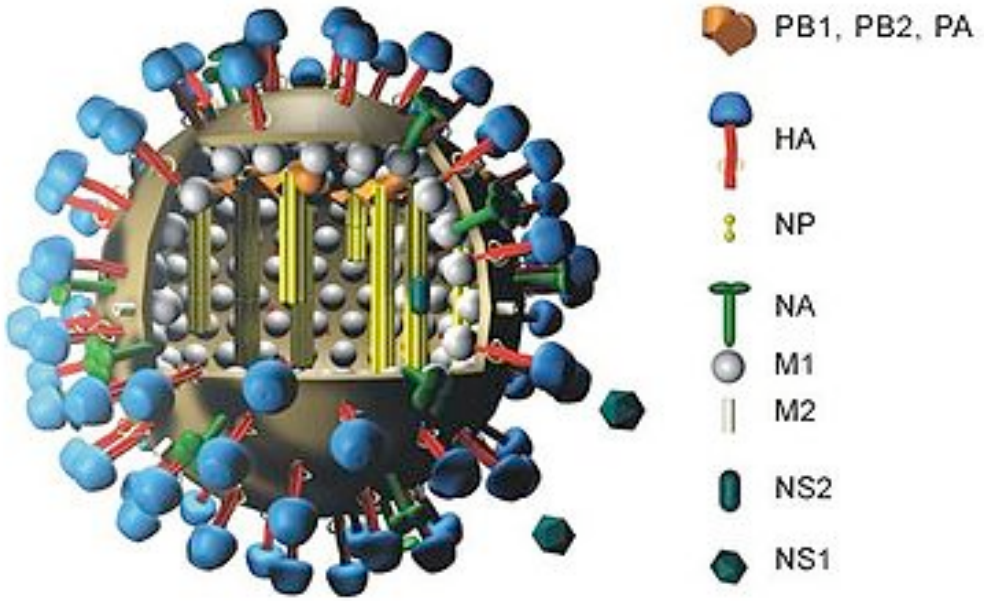
Jelly roll

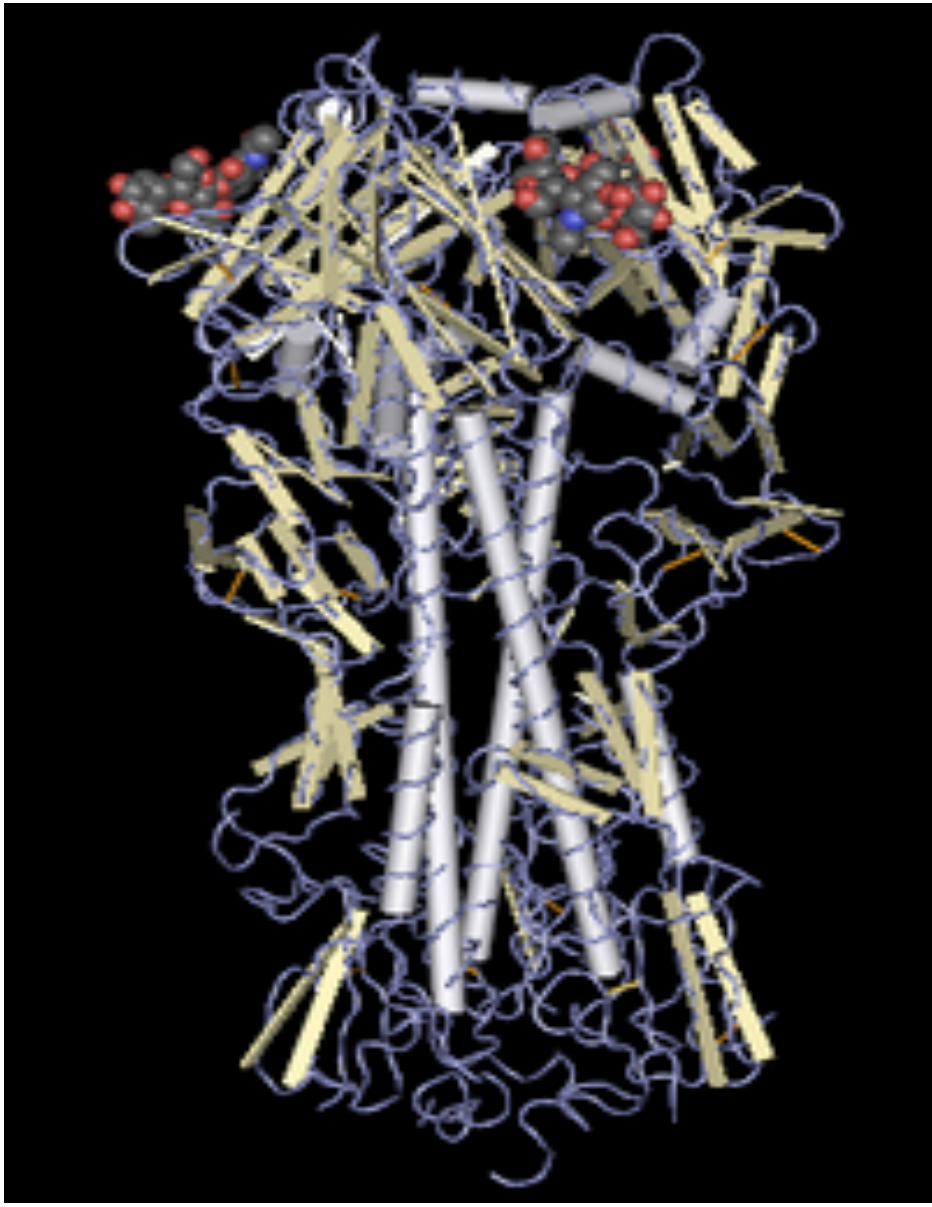
- Imagine a string of 8 antiparallel beta-strands, numbered from 1 to 8, connected two-by-two by hydrogen bonds, involving pairs 1-8, 2-7, 3-6, 4-5
- Imagine now to use this string to wrap a cylinder so that the beta-strands are along the body of the cylinder and the loops go across the bases: this is a jelly roll



- The pairs 1-8, 2-7, 3-6, 4-5 are located in such as way that strand 1 is adjacent to strand 2, 7 to 4, 5 to 6 and 3 to 8. This geometry allows to form hydrogen bonds between these strands.



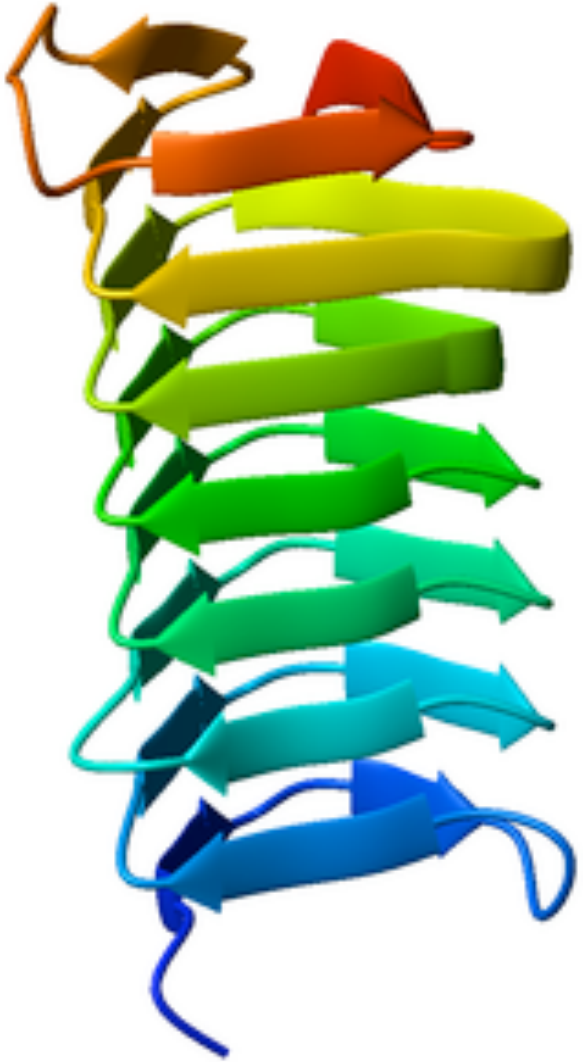




- **Example of Jelly roll is the protein of the capsid of spherical virus, hemmaglutinin**
- Influenza **hemagglutinin** (HA) is found on the surface of the influenza viruses (see HA in fig).
- It is an antigenic glycoprotein. It is responsible for binding the virus to the cell that is being infected.
- The name "hemagglutinin" comes from the protein's ability to cause red blood cells (erythrocytes) to clump together ("agglutinate") *in vitro*.

Beta-helix

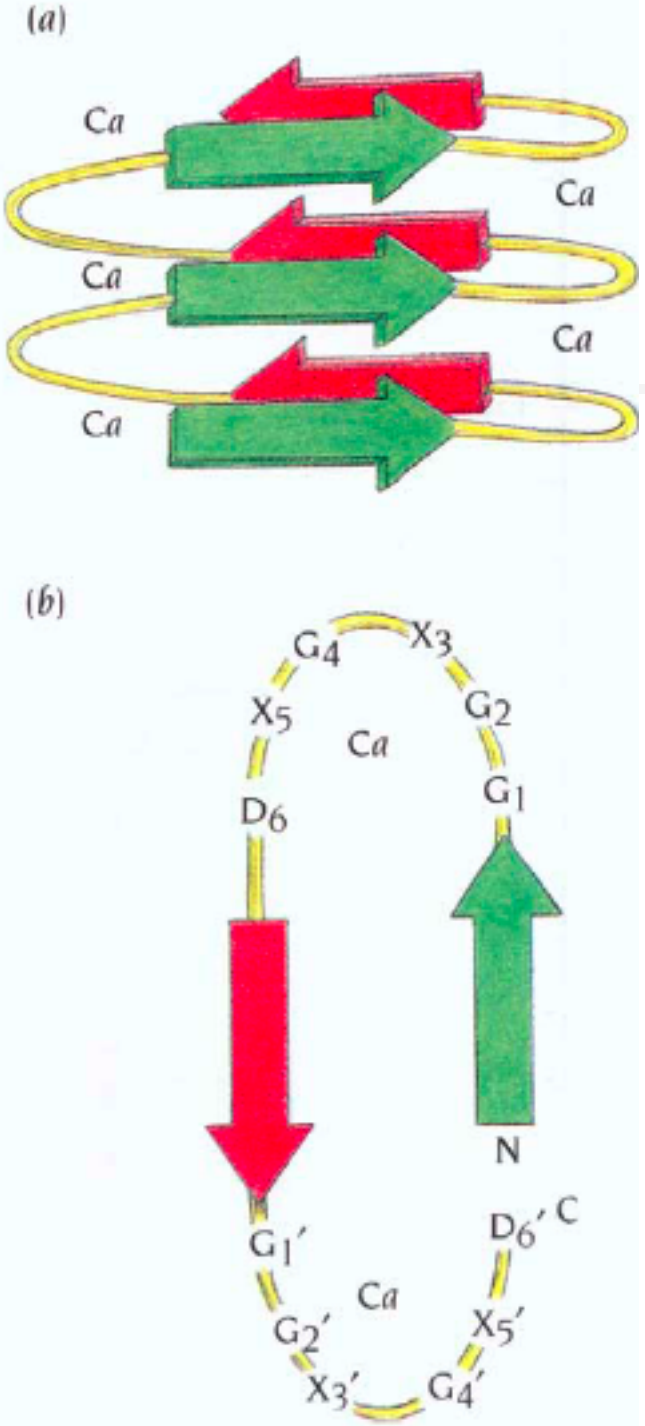
- This was for the first time seen in 1993 in the bacterial enzyme **pectate liase**. After that it has been seen in other extracellular bacterial proteins and in the “tailspike protein” of the batceriophage P22.
- In this structures the polypeptide chain is folded to form a large helix made of beta-strands linked by ample loops
- There are 2 types of beta-helices:
  - Beta-helices with 2 beta-sheets
  - Beta-helices with 3 beta-sheets



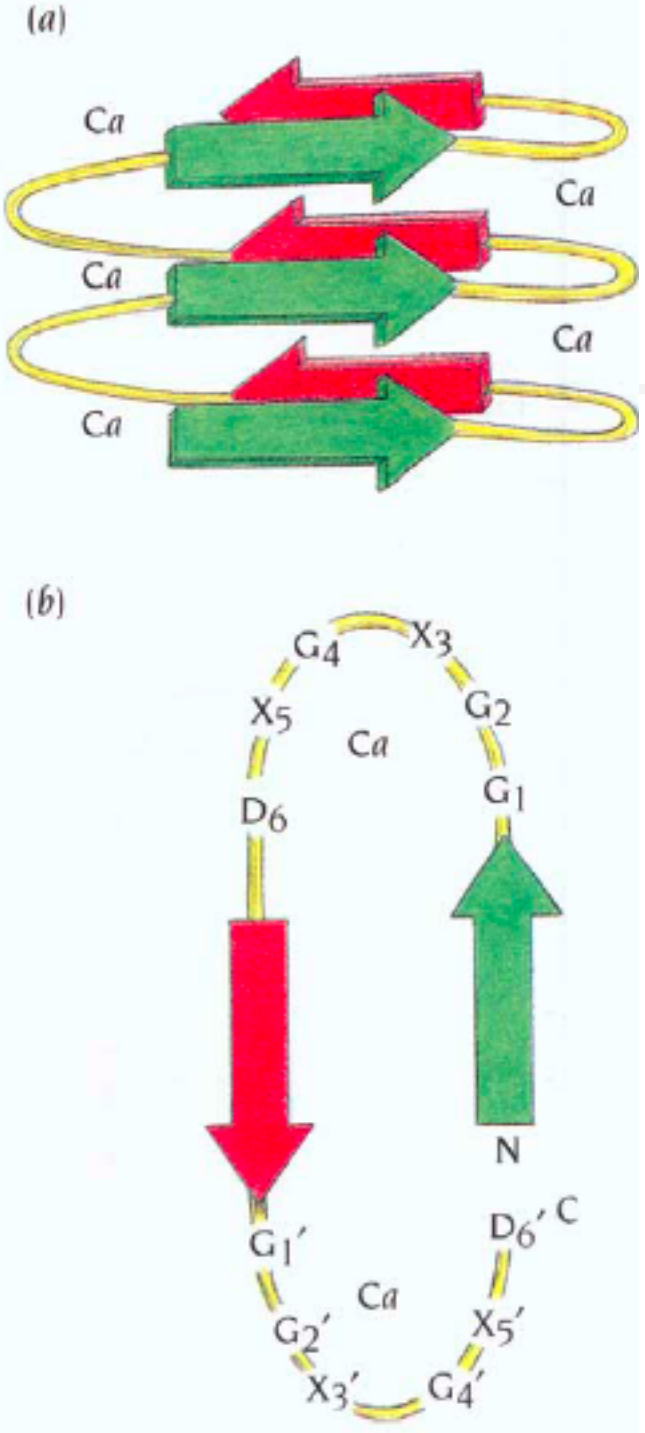
Monomeric, left-handed β-helix antifreeze protein from the spruce budworm *Choristoneura fumiferana*

Beta helix with 2 beta sheets

- This is the most simple beta-helix observed in bacterial extracellular proteinases.
- Each turn of the helix is made by 2 beta strands and 2 loops
- This structural unit is repeated 3 times to form an right-handed helix made of 2 parallel beta-sheets each made of 3 strands with an internal hydrophobic core
- The basic structural unit of the beta-helix with 2 beta sheets is made of 18 aa, 3 in each beta strand and 5 in the loop.
- The consensus sequence of this basic unit is $(GLY\text{-}GLY\text{-}X\text{-}GLY\text{-}X\text{-}ASP\text{-}X\text{-}U\text{-}X)_2$ where X= any aa, U=hrdrophobic aa, often Leu

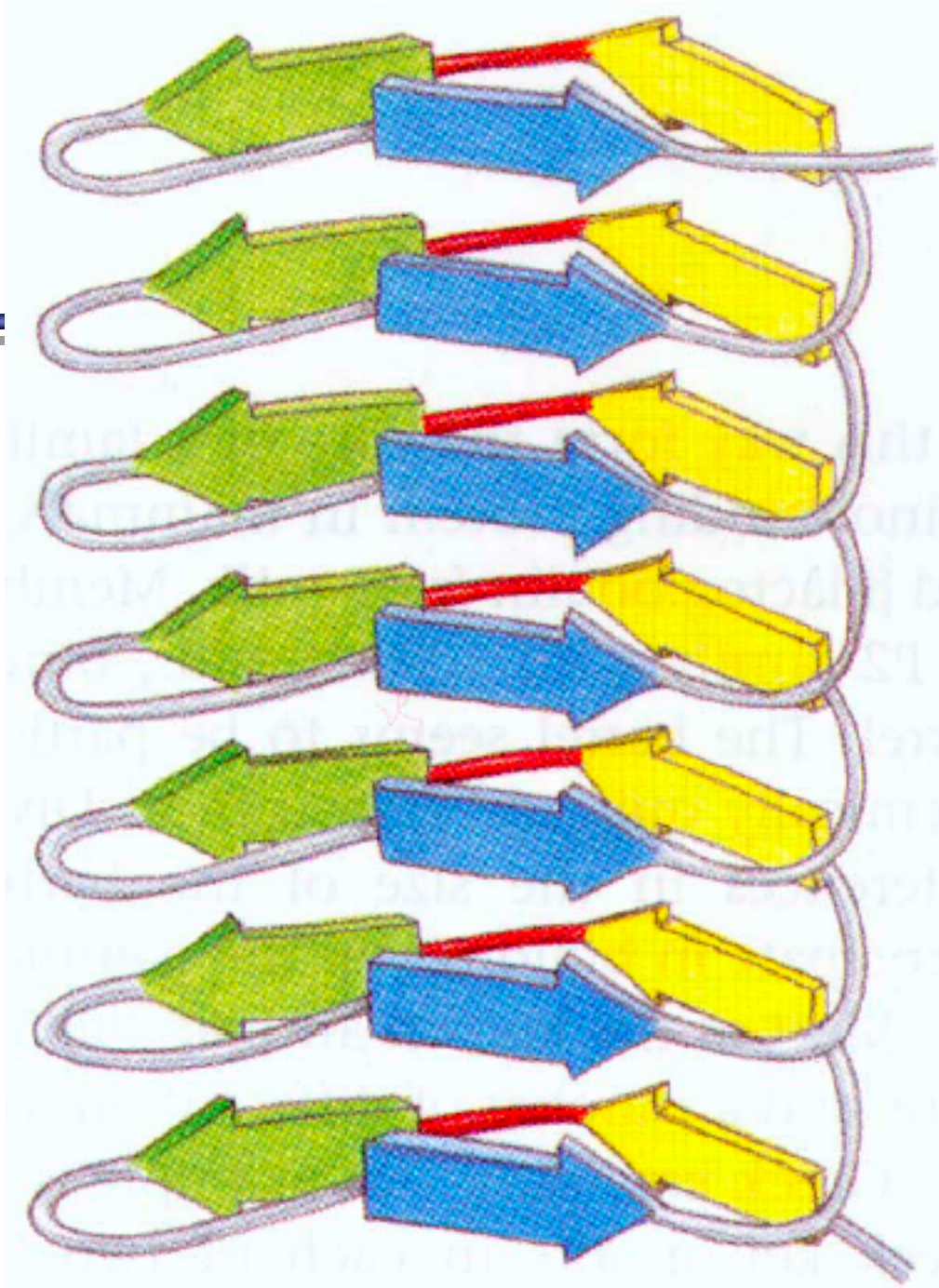


- In the consensus sequence, (GLY-GLY-X-GLY-X-ASP-X-U-X)$_2$, the first 6 aa form a loop and the other 3 a short beta strand, with the hydrophobic chain of U oriented towards the protein core
- The loops are stabilised by Ca^{2+} that binds to the Asp
- This consensus sequence can be used to find this kind of folds in sequence database

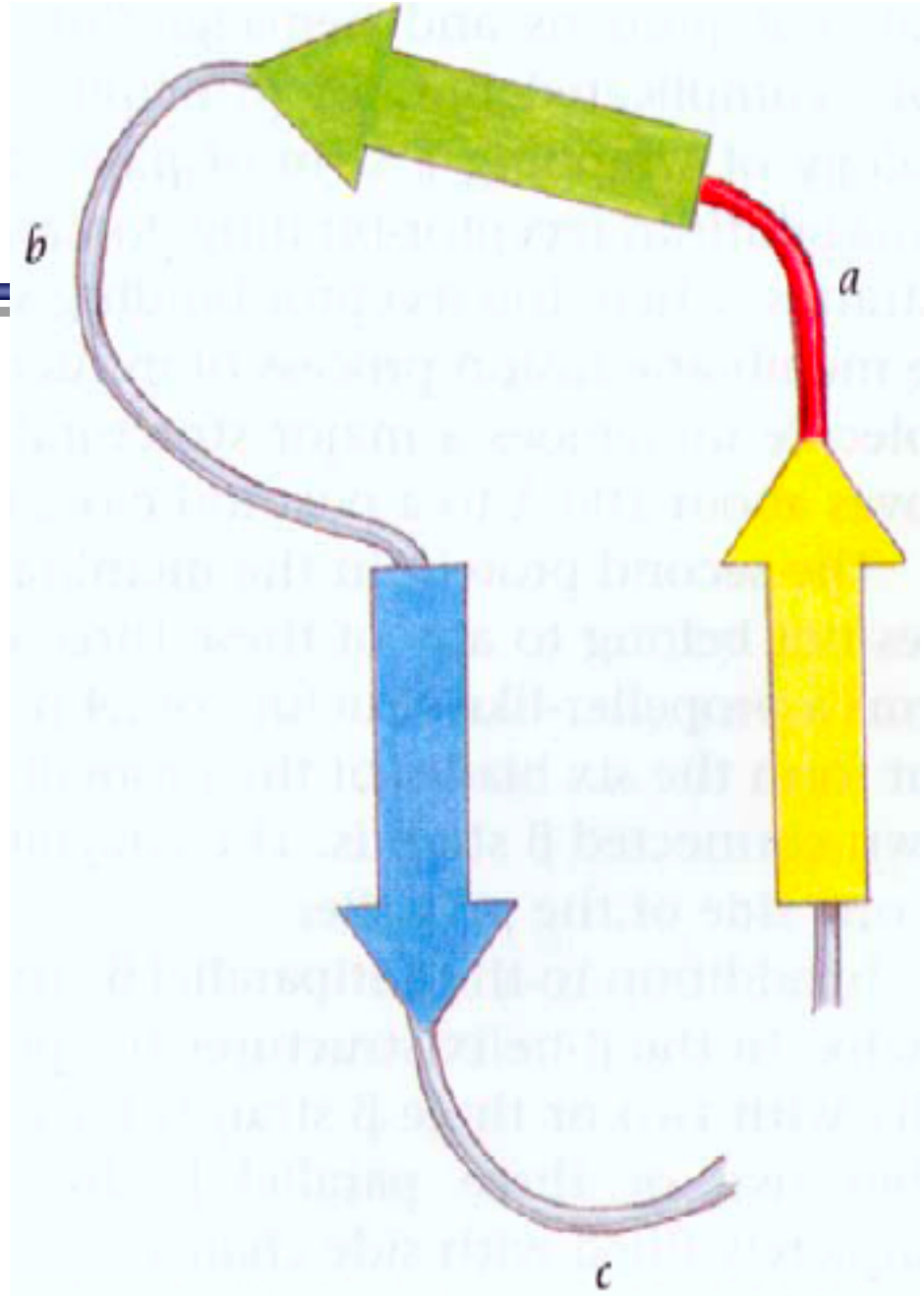


Beta helix with 3 beta sheets

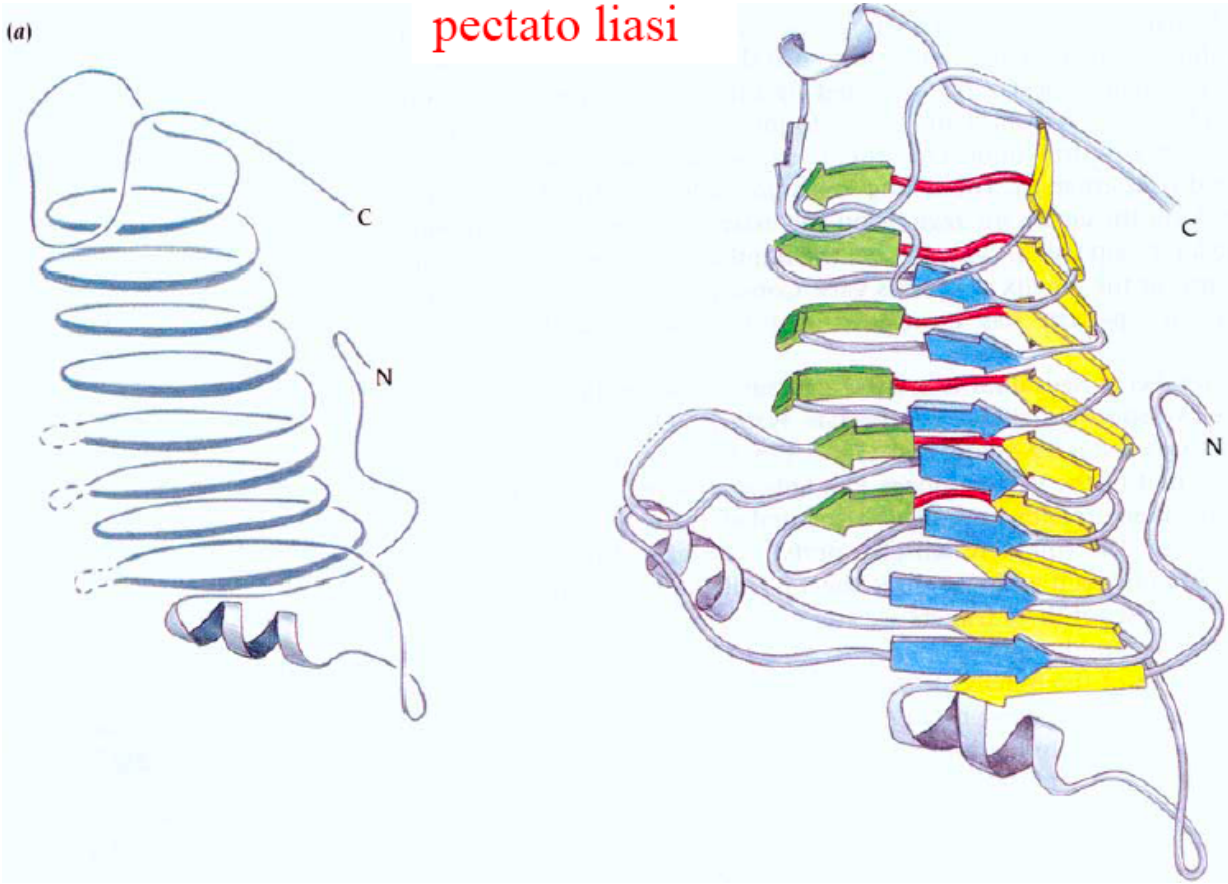
- This beta helix is more complex and is typical of the pectate liase and tailspike protein of phage P22
- Each turn is made of 3 beta strands, each made of 3 to 5 aa, connected by 3 loops
- The beta helix is therefore made of 3 parallel beta sheets that form the sides of a prism
- The section of the beta helix is not triangular, because of the orientation of the loops.



- Two of the sheets are located one next to the other, like in the 2 beta strands helix, and the third is nearly perpendicular to the other two
- One loop (a) is short and normally made of 2 aa and leading to a 120• tilt between the strands that links.
- The other 2 loops (b,c) are longer and are variable in dimensions and conformation. The longs loops are projected to the external part and can form the active site of enzymes
- The variability of the loops does not allow to formulate a consensus sequence for this type of domain.



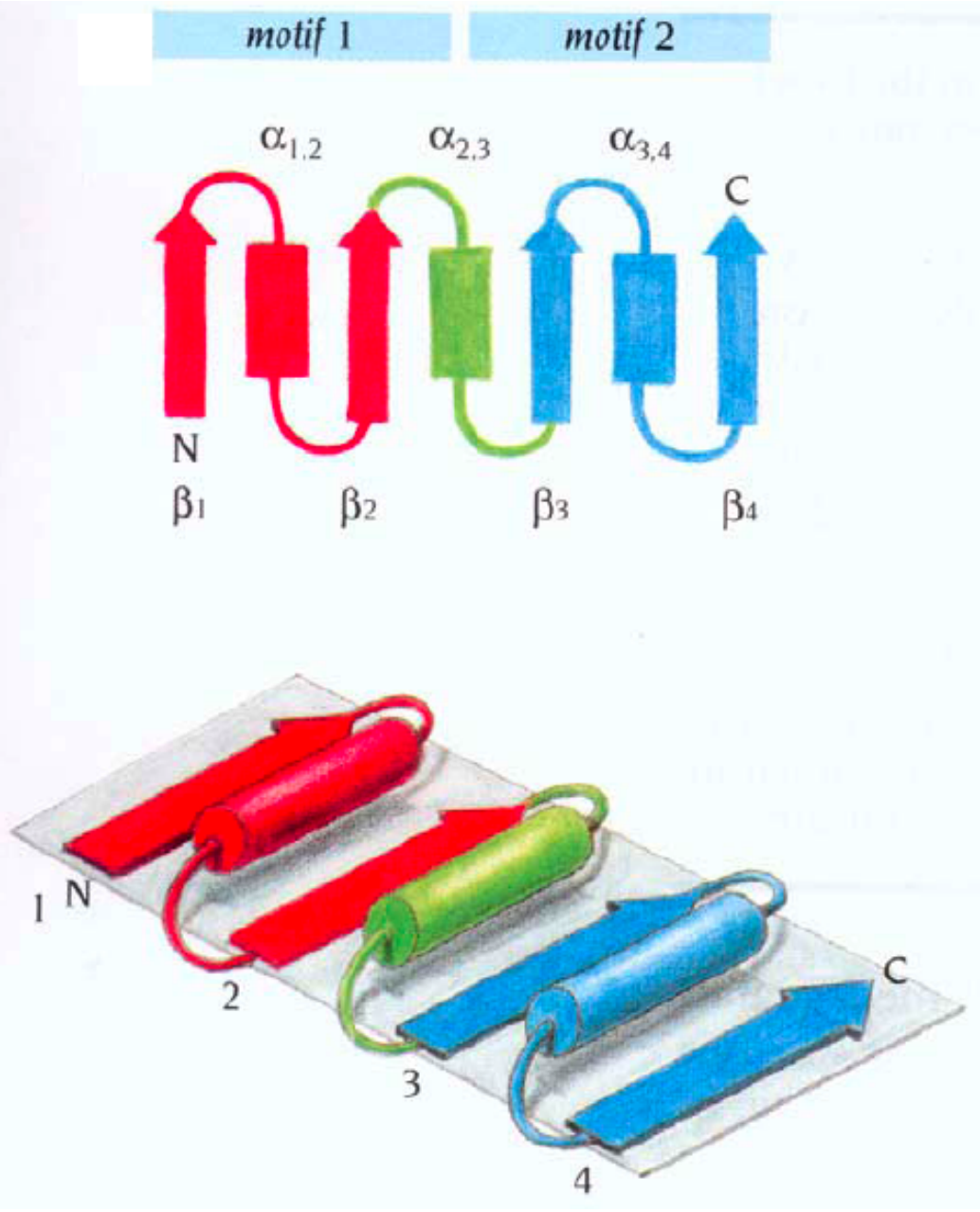
- The number of turns in the beta helices with 3 sheets is higher than in the beta helices with 2 sheets
- For example the pectate liase has 7 turns, with a length of 34Å and a diameter of 17-27Å, the helix of the tailspike of phage P22 has 13 turns



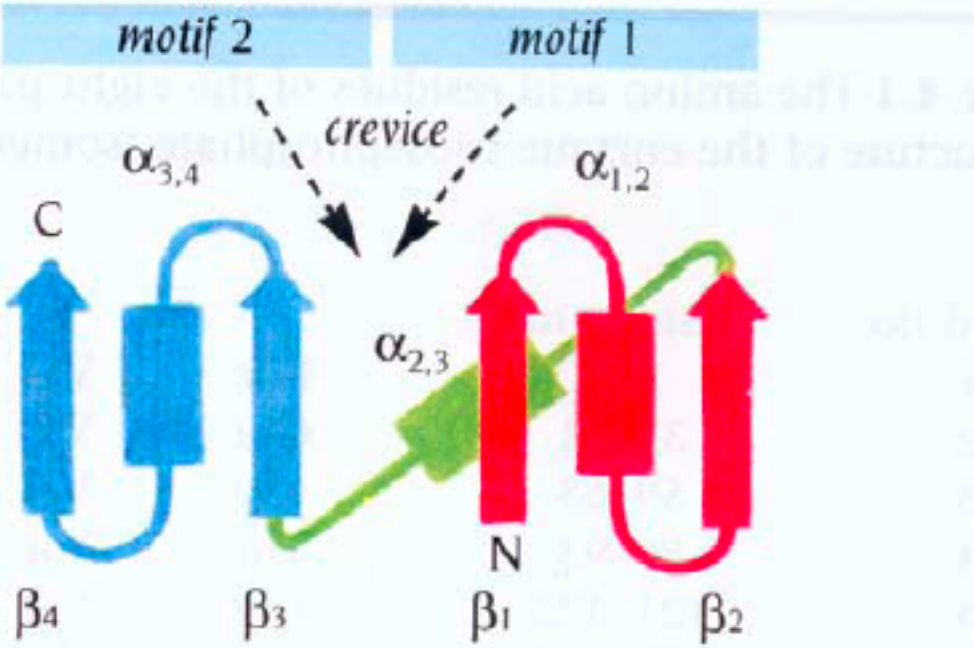
Alpha-Beta domains

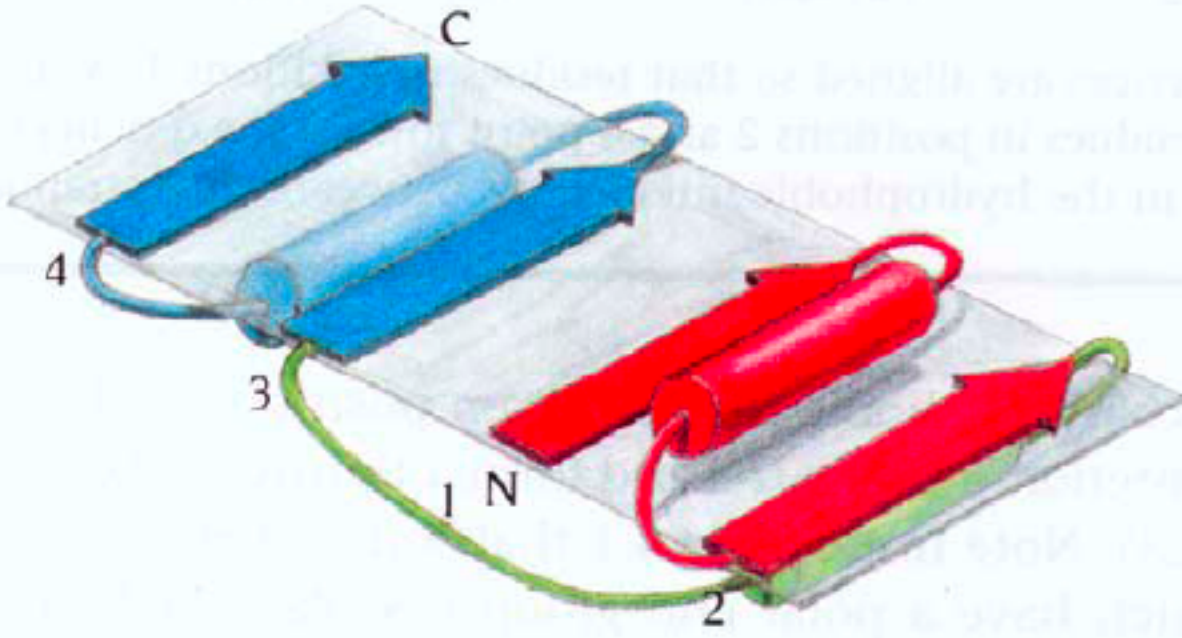
- The cross-over connection is the basic unit that constitute 3 common alpha-beta domains often found in proteins:
  - Alpha/beta barrel
  - Leu rich motifs (horse shoe folds)
  - Alpha/beta open sheet
  - Two cross-over connections can be linked in 2 different modes, involving an alpha helix, to form a parallel beta-sheet made of 4 beta-strands:

- 1 – The alpha helices are both on the same side of the parallel beta-sheet
  - The last beta strand of the first cross over connection is adjacent to the first beta strand of the second cross over connection
  - The strands are in the order 1-2-3-4
  - This is the typical layout of the alpha/beta barrel and of the horse shoe fold



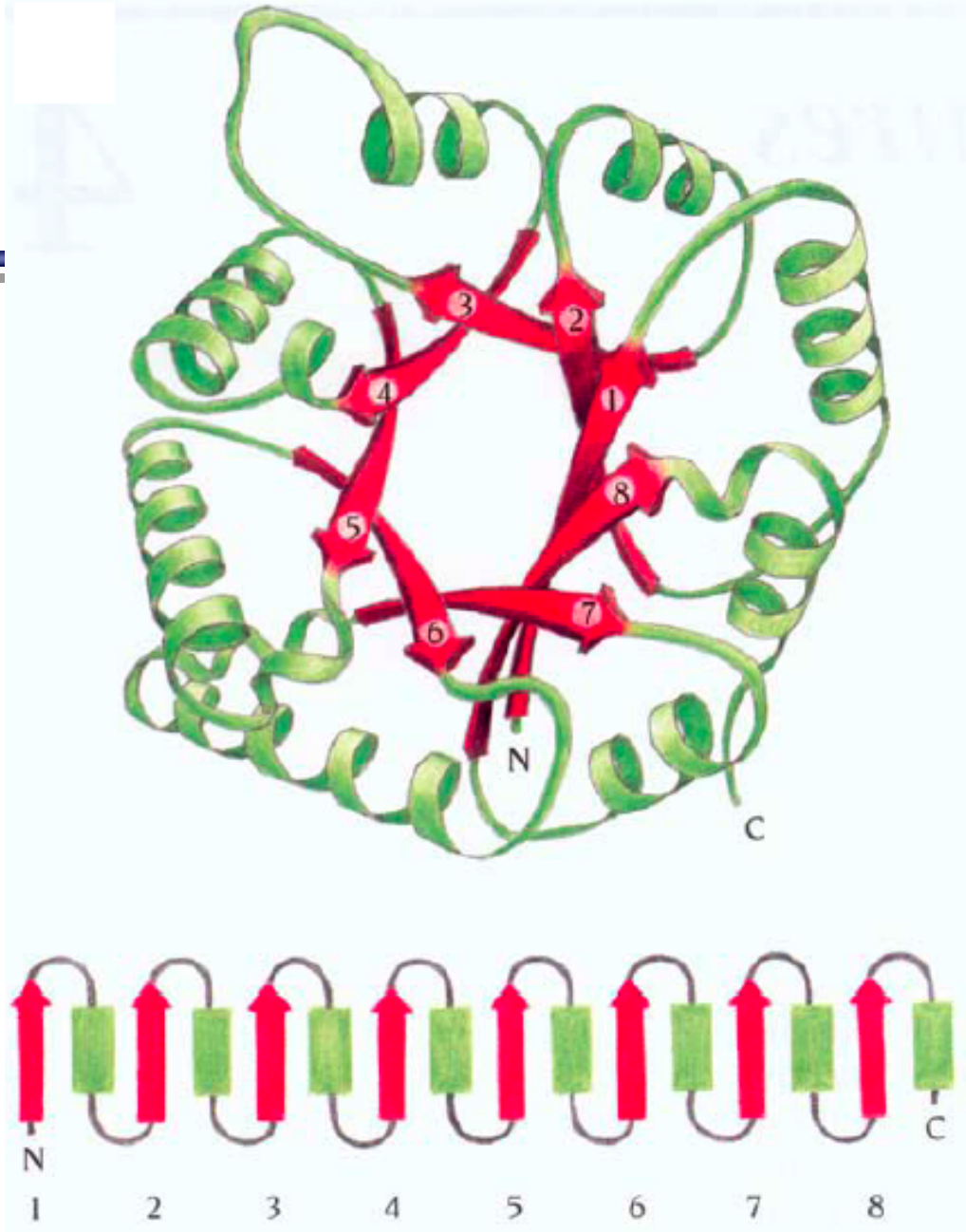
- 2 – The alpha helices are on opposite sides of the parallel beta-sheet
  - The last beta strands of each cross over connection are adjacent
  - The strands are in the order 4-3-1-2
  - This is the typical layout of the alpha/beta open sheet



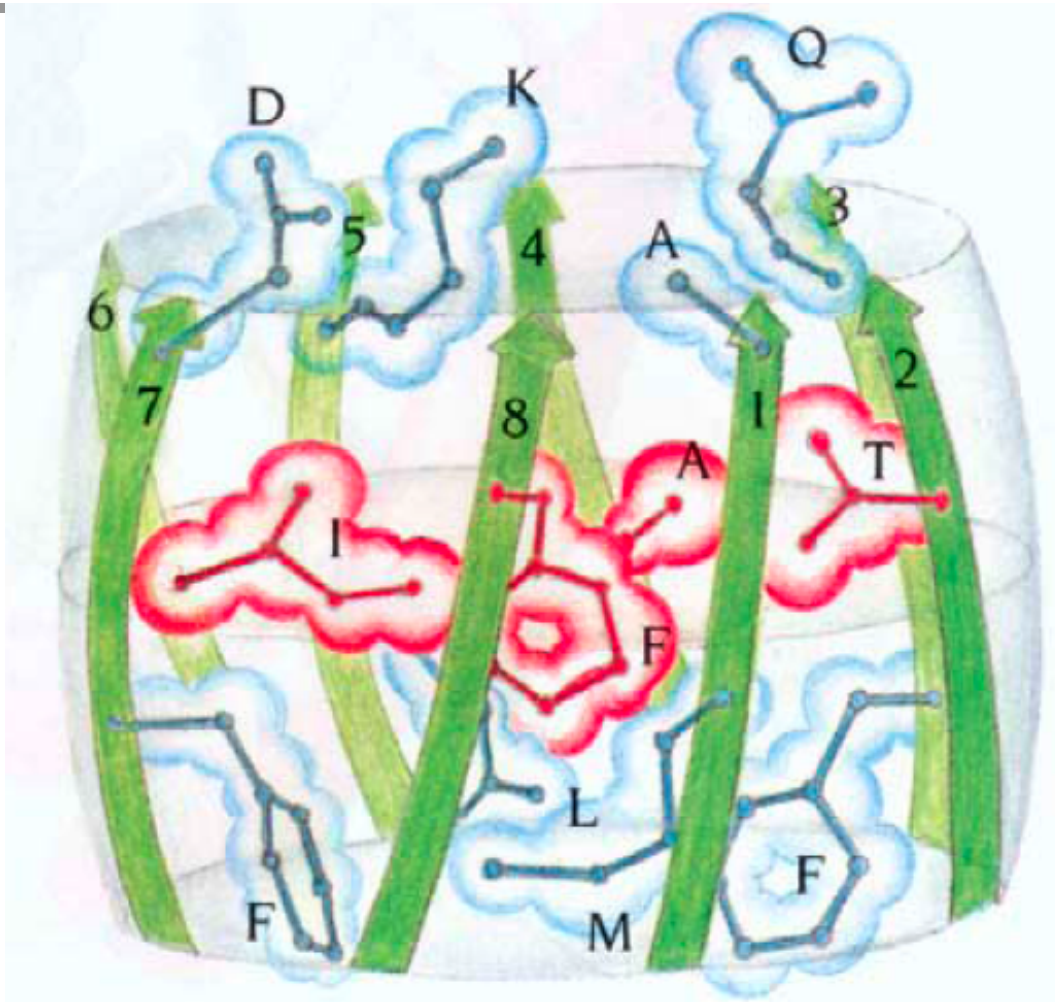


Alpha-beta barrel

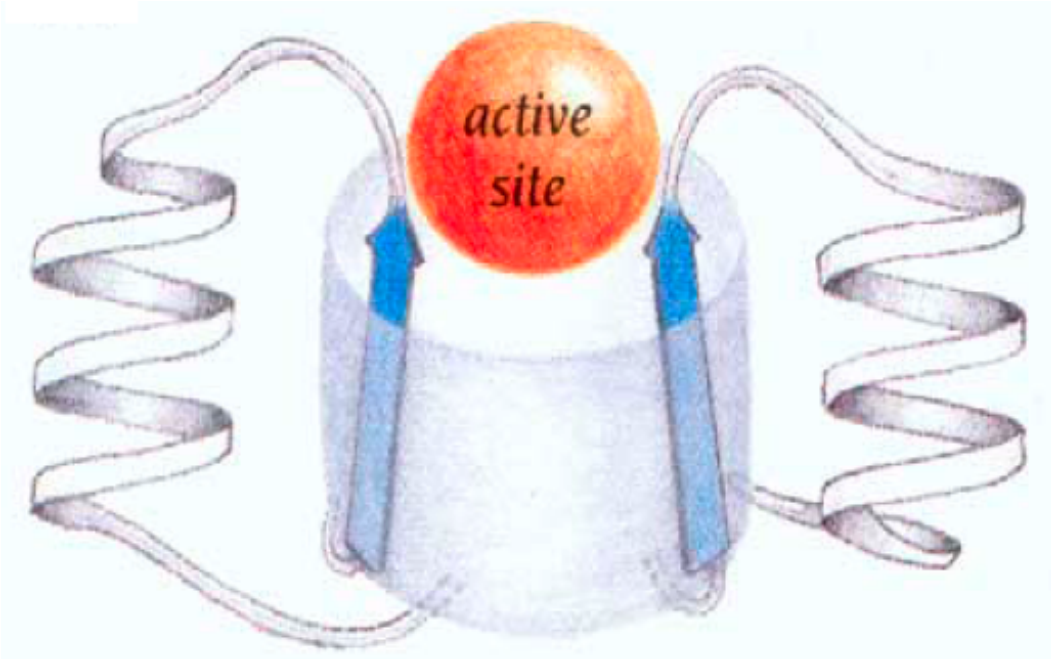
- This is made of 8 parallel beta strands positioned such that the the 8th strands is adjacent and makes H-bonds with the 1st.
- Normally the cross over connections between the parallel beta strands are alpha helices.
- There also are alpha/beta barrels made of 10 parallel beta strands
- The alpha/beta barrel is a domain that normally involves at least 200 aa. It is typical of many enzymes
- It is also called TIM barrel from the structure of the enzyme triose-phosphate isomerase



- In the alpha/beta barrel the hydrophobic side chains of the alpha helices are packed against the hydrophobic side chains of the beta strands.
- Usually these interactions involve residues of Val, Ile, Leu, that form the 40% of the aa present on the barrel
- The other side chains of the aa (1,3,5,7…) of the beta strands are oriented towards the inside of the barrel and form a hydrophobic core tightly packed
- The edges of the barrel is made of polar aa Arg, Lys, Asp capable of interacting with the solvent

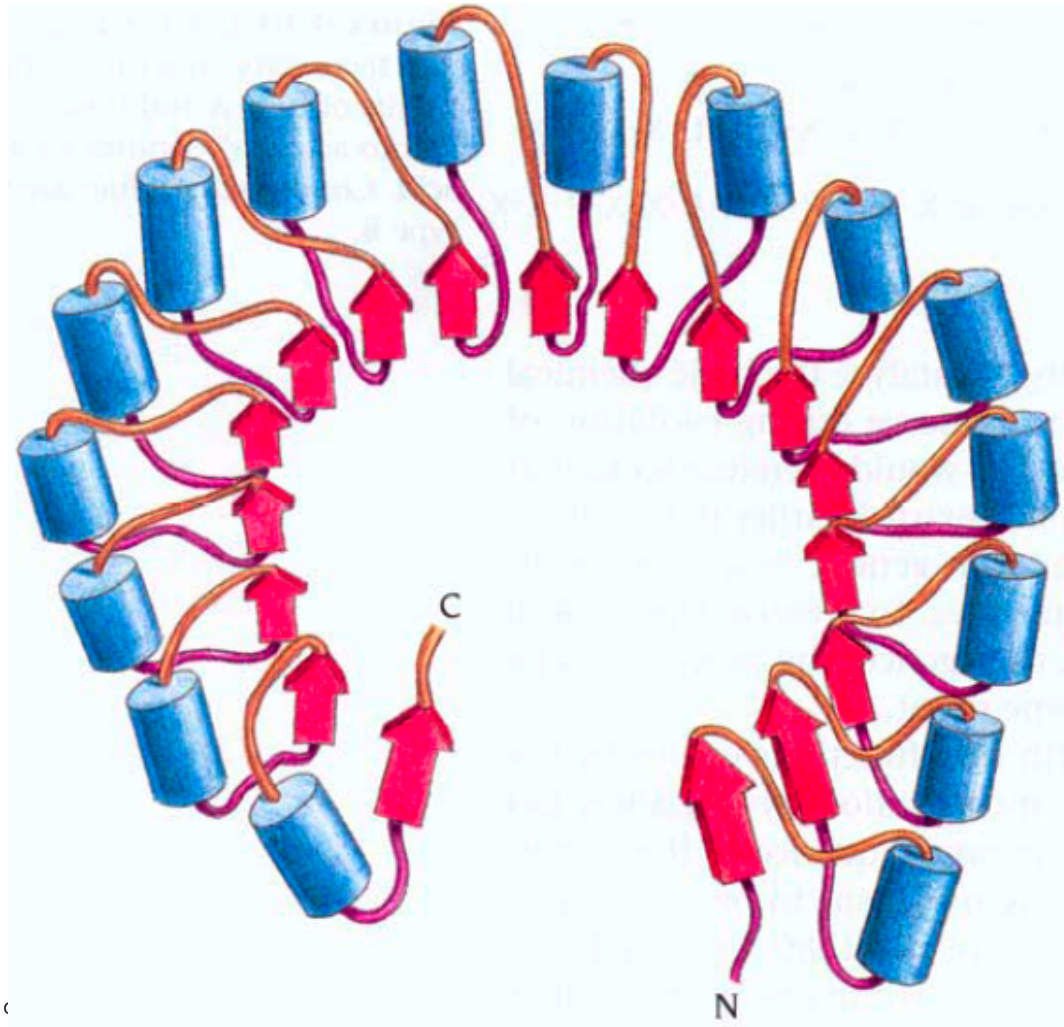


- All alpha/beta barrel known belong to enzymes with activities such as sugar isomerase, phosphate transfer, degradation of polysaccharides
- In all structures the active site is located in the same position, in a pocket made by the 8 loops that connect the C-ter ends of the beta strands with the N-ter ends of the helices
- The aa that are involved in substrate binding and catalysis belong to these loops



Horse shoe fold

- The beta-strands form a parallel sheet and all the alpha helices are on the same side of the sheet
- The beta-sheet forms an open arched structure, with the alpha helices on the outside of the arch and the beta sheet inside.
- The interface between the beta sheet and the alpha helices form an hydrophobic core.
- The other side of the beta sheet is hydrophilic, exposed to the solvent, a characteristic typical of this domain

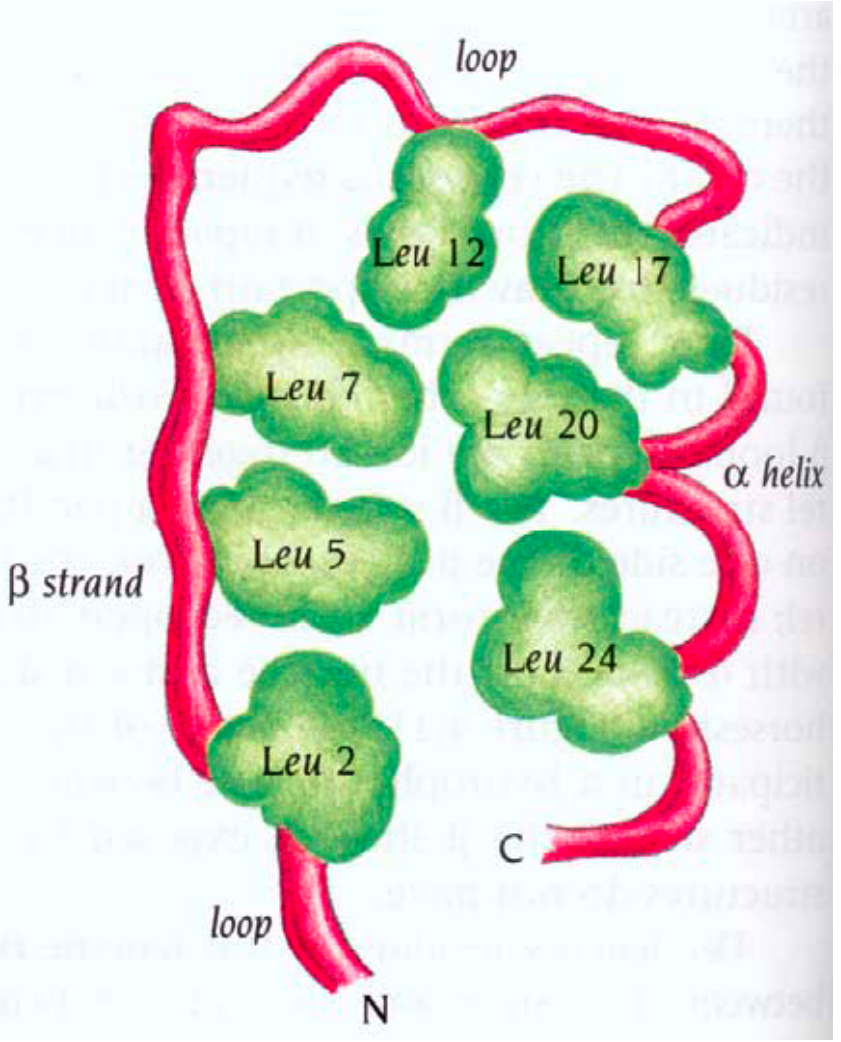


- The horse shoe domains show homologous sequences rich of Leu.
- These are repetitions of 20-30 aa, identified in more that 60 different proteins with these domains, among which cellular receptors, adhesion molecules, DNA repair proteins
- The homologous sequences contain a typical pattern of Leu that is crucial for the beta-loop-alpha structure
- The connections between the beta-loop-alpha are very similar to those present in the alpha/beta barrels.

	2	5	7	12	20	24
(*tipo A*) NH_2	-X-**L**-E-X-**L**-X-**L**-X-X-C-X-**L**-T-X-X-X-C-X-X-**L**-X-X-a-**L**-X-X-X-X-					
(*tipo B*) NH_2	-X-**L**-R-E-**L**-X-**L**-X-X-N-X-**L**-G-D-X-G-a-X-X-**L**-X-X-X-**L**-X-X-P-X-X-					

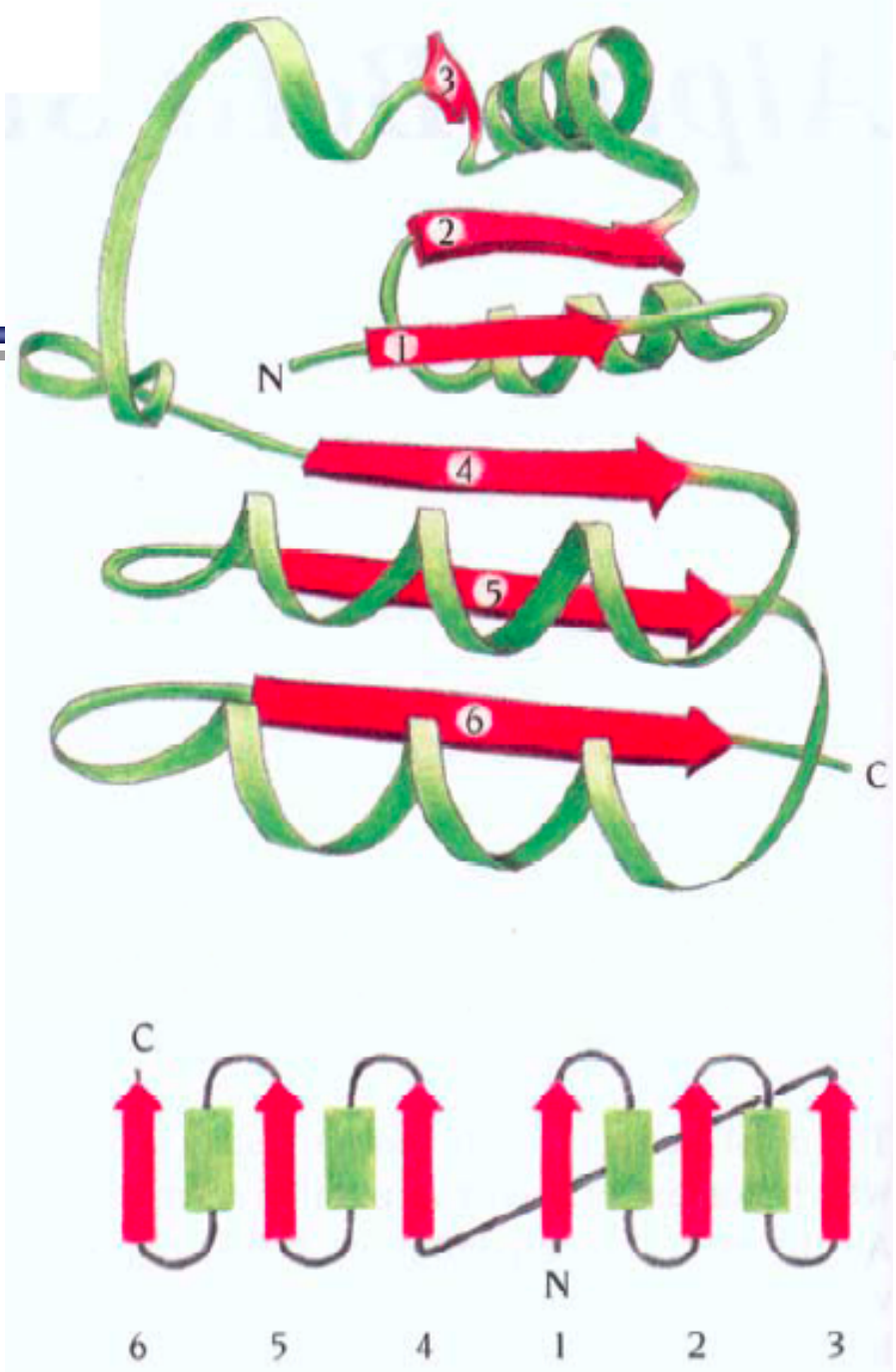
(a = aminoacido alifatico)

- The Leu belonging to the two motifs form the hydrophobic core between the beta sheet and the alpha helix
- The Leu 2, 5 and 7 of the beta sheet pack against the Leu 20 and 24 of the alpha helix to form the central part of the hydrophobic zone
- Leu 12 of the loop and Leu 17 of the alpha helix also contribute to the hydrophobic core
- The analysis of over 500 pairs of Leu rich sequences from 68 different proteins have shown that aa 17, 20 and 24 can be different from Leu but must always be hydrophobic. Leu 2, 5, 7 and 12 are not varied.

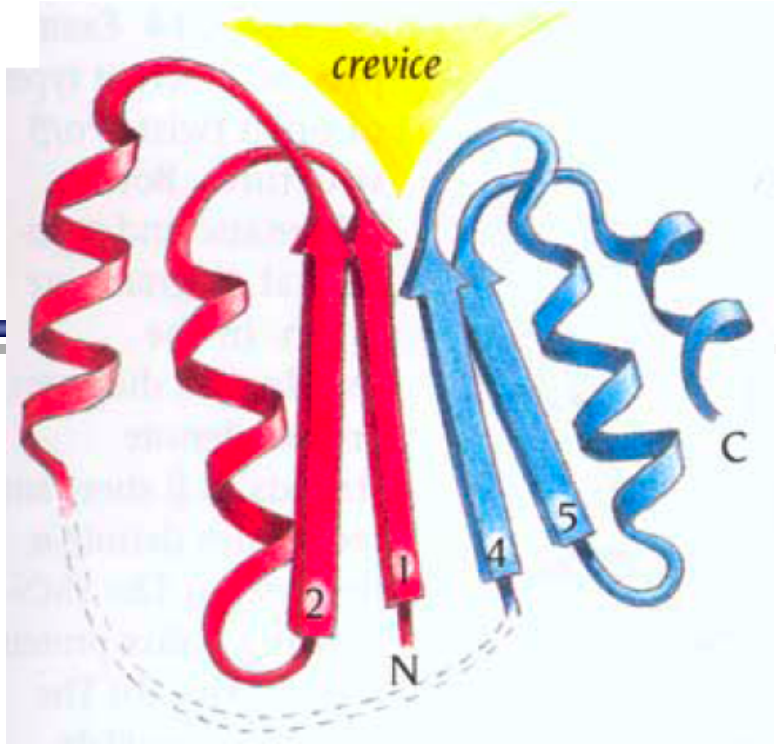


Alpha/beta open sheet

- In the alpha/beta open sheet the alpha helices are located on both sides of a parallel beta sheet
- Each beta strands offers the side chains of hydrophobic aa that back against those of the alpha helices, one for each side of the beta sheet
- There are always two beta strands (see 1 and 4 in fig) in the internal part of the beta sheet that have connections with the adjacent beta strands on opposite sides. One loop is located above, a second below the plane of the beta strand

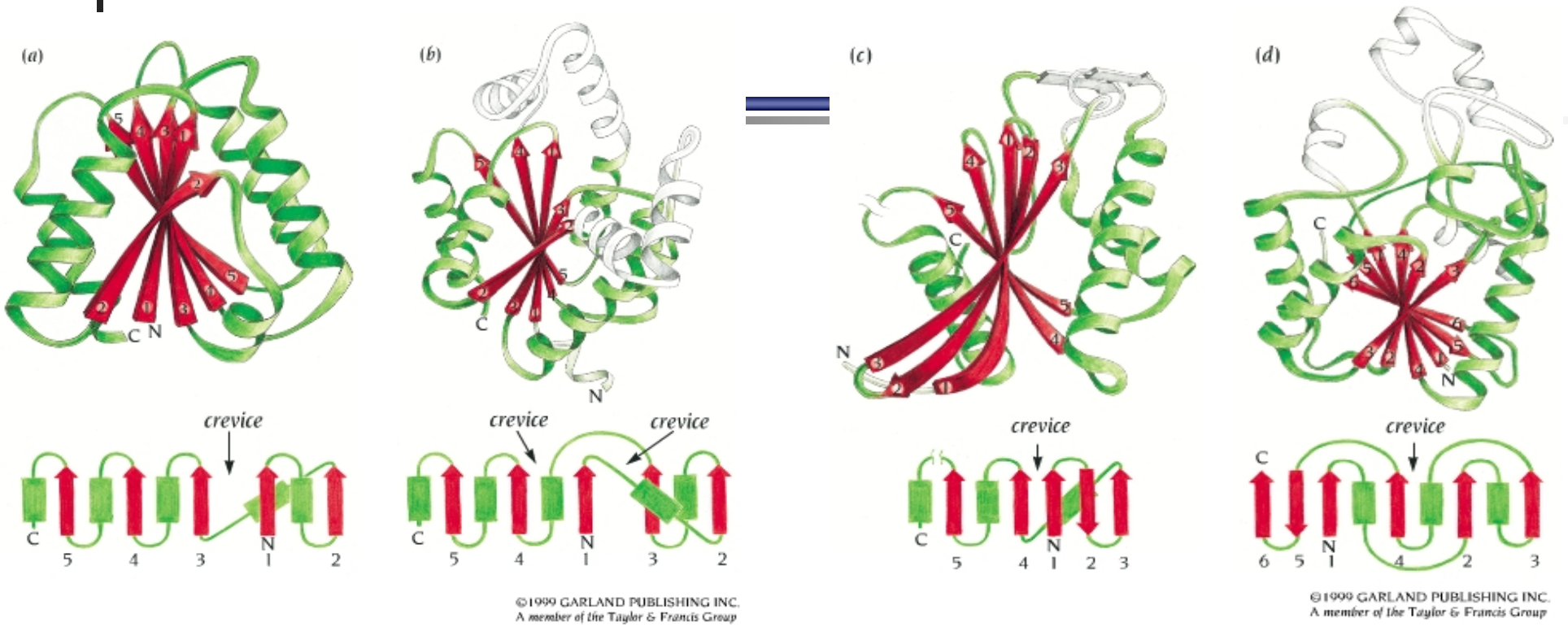


- The point where there is the inversion of the beta sheet that is covered by the alpha helices is called “switch point”
- The switch point is at the C-ter of the beta sheet and it defines a pocket where are located the binding sites of this class of proteins
- There are no geometric restrictions on the number of the beta strands involved: the number varies from 4 to 10



Alpha/beta domains

Sheets



a) Flavodoxin, a redox protein containing FMN

b) Adenilate kinase catalysing the reaction AMP + ATP ⇔ 2ADP

c) ATP binding domain of esokinase catalysing glucose phosphorylation

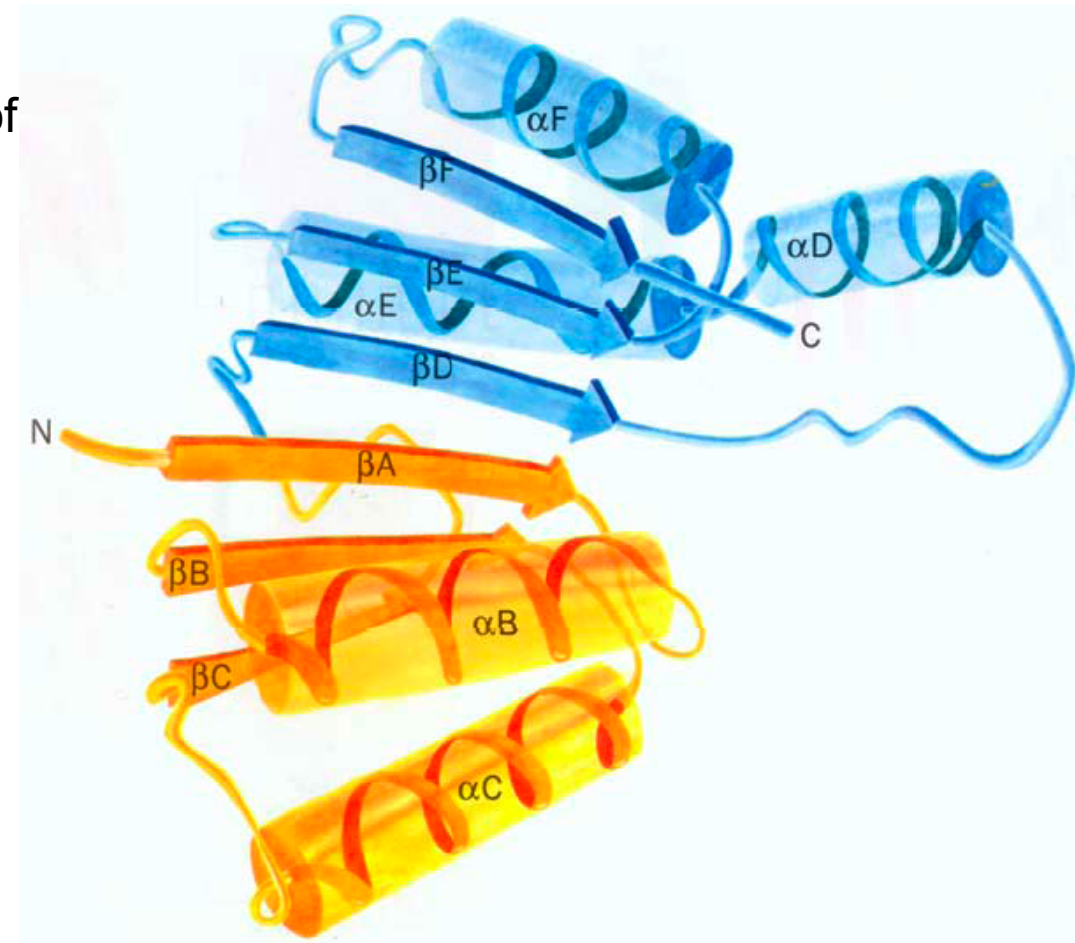
d)Phosphoglycerate mutase catalysing phosphate transfer from C3 to C2

Dinucleotide binding: the Rossmann fold

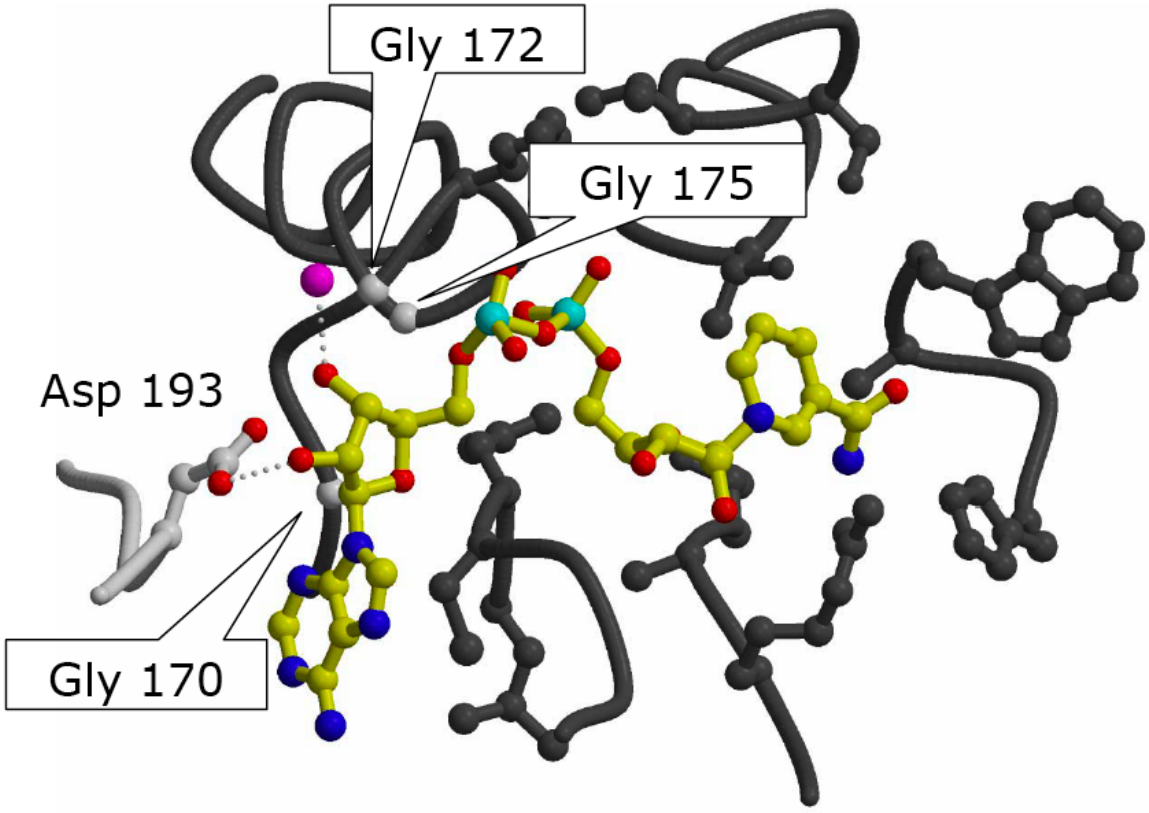
- Many enzymes, in particular the dehydrogenases, have a domain capable of binding the dinucleotide (NAD, FAD)
- This domain is made of 2 units beta-alpha-beta-alpha-beta
- This domain was first identified in 1970 in lactate dehydrogenase by M. Rossmann, and for this reason is now called Rossman fold
- The Rossmann fold is made by the consensus sequence:

GXGXXG(17X)D

- This sequence is involved in nucleotide binding

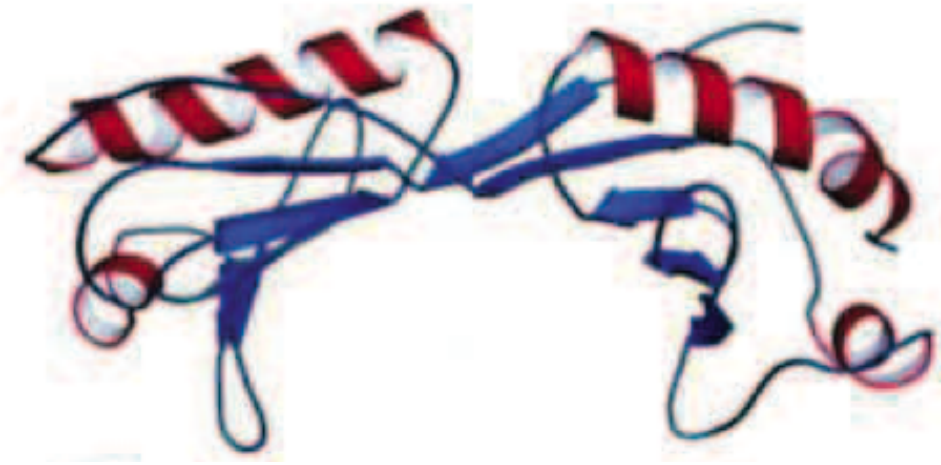


- The residues GXGXXG are in the loop between the first beta strand and the following alpha helix
- This loop allow the interaction with the pyrophosphate group
- At the end of the second parallel beta strand the COO^- of the side chain of Asp or Glu interact with the diol of the ribose



Alpha+beta domains

- Domains that contain both β sheets that α-helices, but they are separated
- There are no particular principles regarding their structural organization
- Groups of interacting α-helices
- Antiparallel or mixed β-sheets



Example: saddle α + β protein that binds the TATA box (start site of gene transcription in eukaryotes)