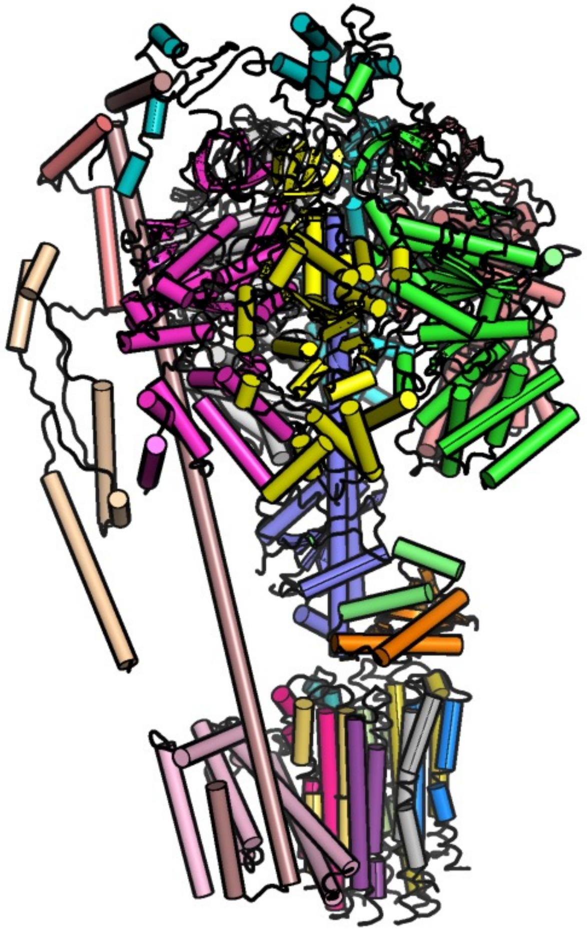


Structural Bioinformatics

Genome 541
Spring 2023

Lecture 1: Protein Structure

Frank DiMaio (dimaio@uw.edu)

HW #0: Getting PyMol and PyRosetta

Today’s class will introduce protein structure and PyMol
Thursday’s class will provide a hands-on demo of PyRosetta

PyMol:

DOWNLOAD URL: https://pymol.org/ep
USERNAME: jun2021
PASSWORD: betabarrel

PyRosetta:

DOWNLOAD URL: https://www.pyrosetta.org/downloads
USERNAME: teaching
PASSWORD: scorefunction

Example ~/.condarc

```
channels:
- https://USERNAME:PASSWORD@conda.rosettacommons.org
- conda-forge
- defaults
```

(pymol + PDB intro demo)

Motivation: Why do we care about macromolecular structure?

Sequence → Structure → Function

- Structure determines function, so understanding structure helps our understanding of function

Structure more conserved than sequence

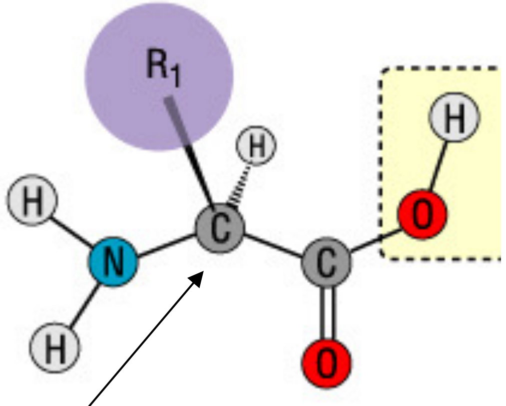
- Structure allows identification of more distant evolutionary relationships

Structure is encoded in sequence

- Understanding the determinants of structure allows design and manipulation of proteins

Proteins are Polymers of Amino Acids

Amino acids

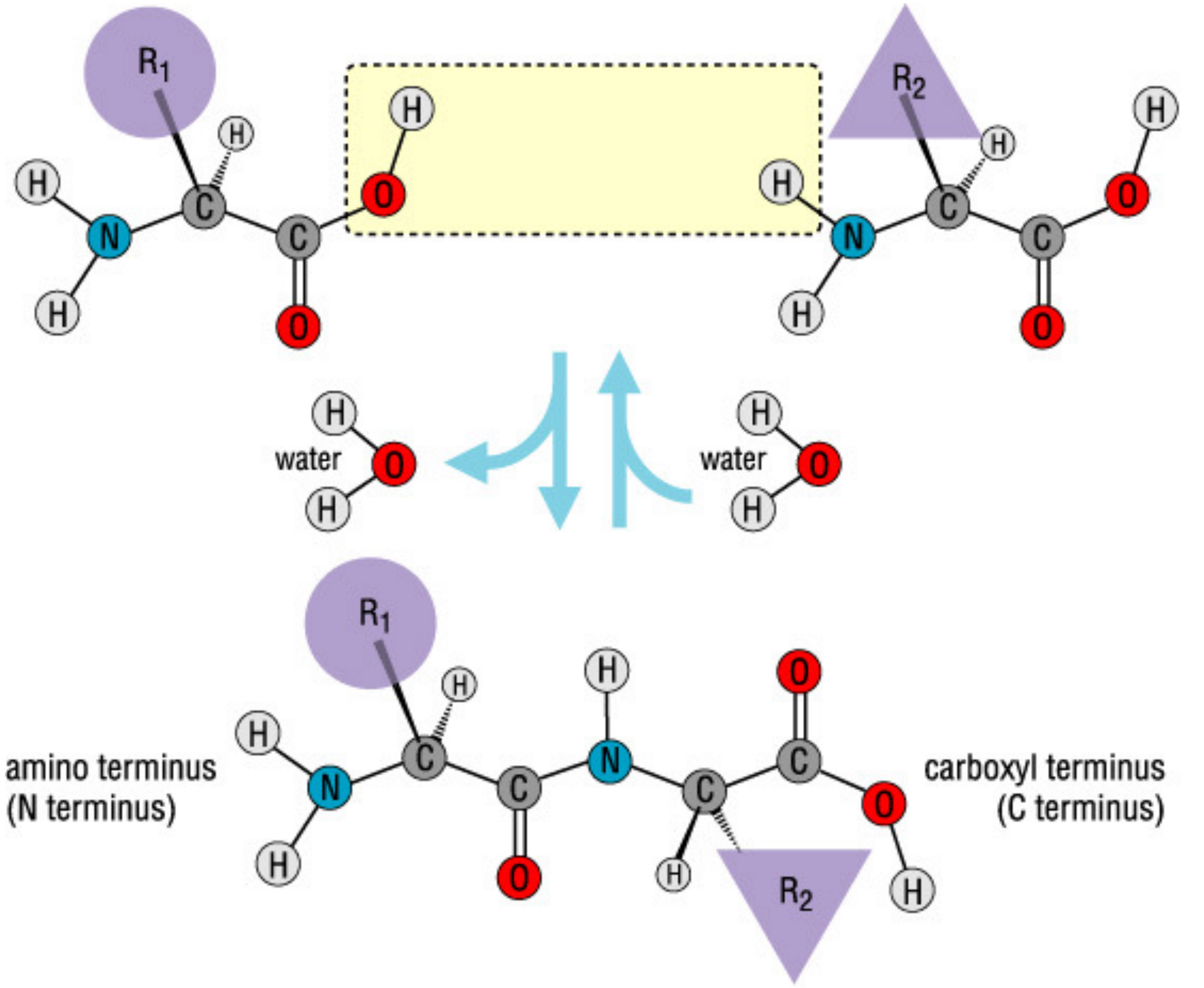


Amino acids have chiral centers

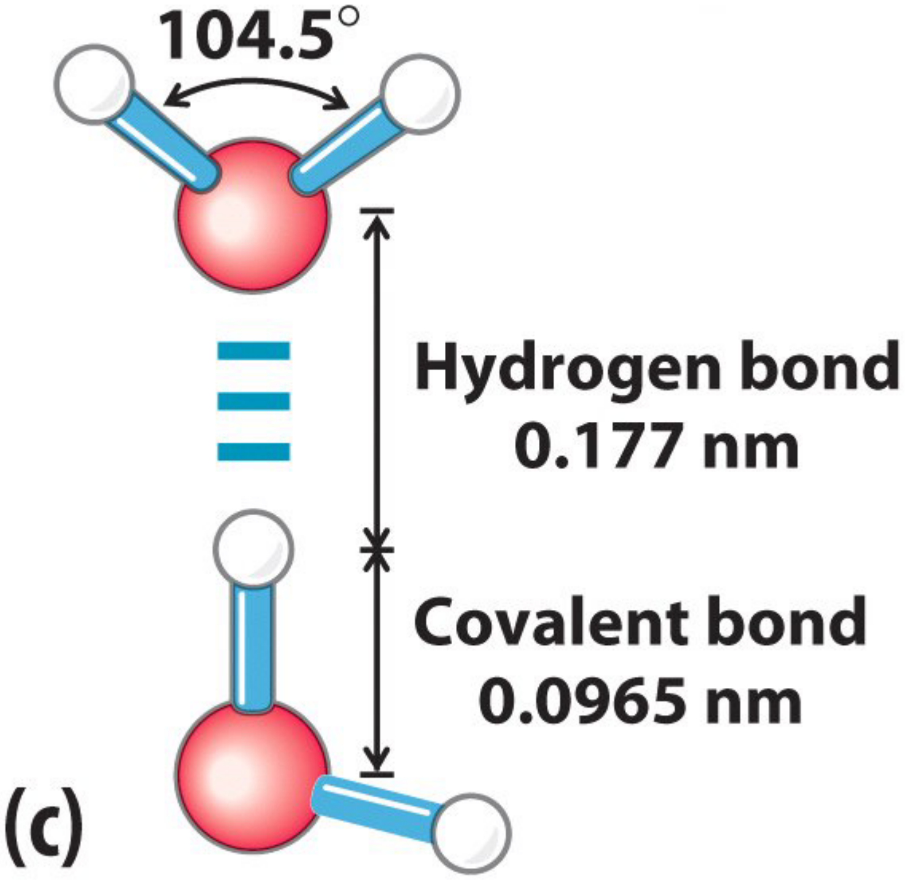
Proteins are Polymers of Amino Acids

Amino acids

polypeptide



Water and hydrogen bonds

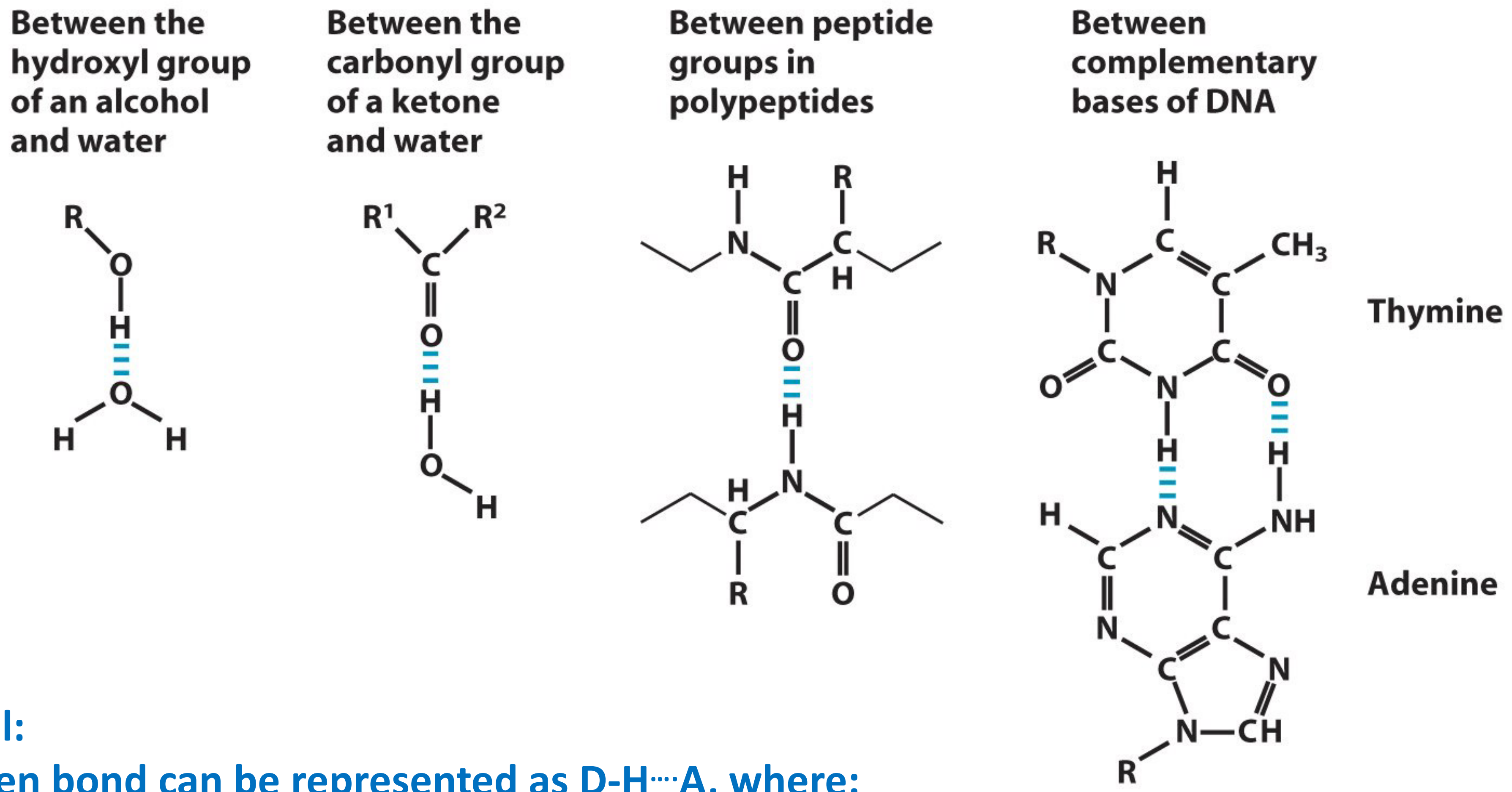


Important:
The O-H distance of ~1.77 Å in an H-bond is *smaller* than the sum of :

- **the H vdW-radius of ~1.2 Å**
- **the O vdW-radius of ~1.4 Å,**

10 Å = 1 nm = 10^{-9} m

Hydrogen bonds in general



Between the hydroxyl group of an alcohol and water

Between the carbonyl group of a ketone and water

Between peptide groups in polypeptides

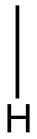
Between complementary bases of DNA

In general:
A hydrogen bond can be represented as D-H····A, where:

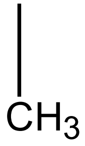
D-H = weakly acidic “donor” group, such as O-H, N-H
A = weakly basic “acceptor” atom such as O, N

Non-polar or Hydrophobic Amino Acids

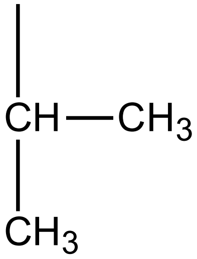
Glycine (Gly, G)



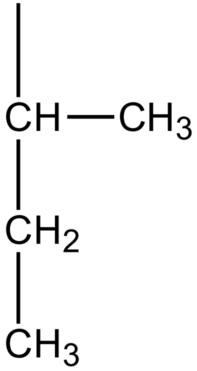
Alanine (Ala, A)



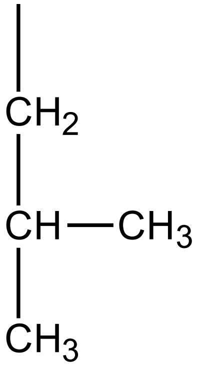
Valine (Val, V)



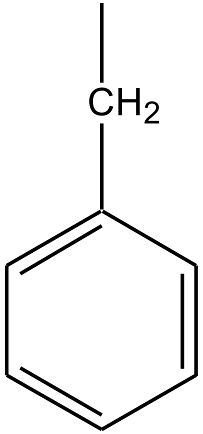
Isoleucine (Ile, I)



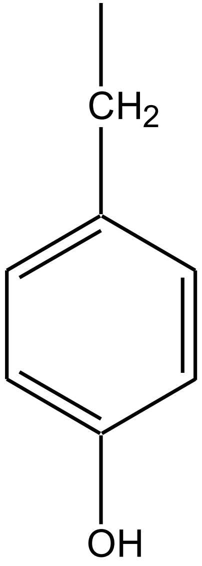
Leucine (Leu, L)



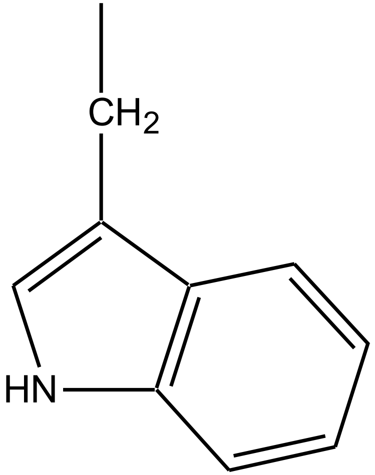
Phenylalanine (Phe, F)



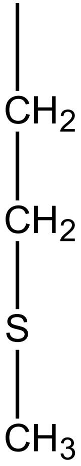
Tyrosine (Tyr, Y)



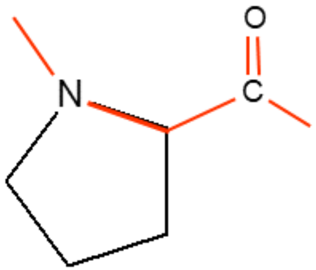
Trptophan (Trp, W)



Methionine (Met, M)



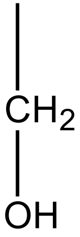
Proline (Pro, P)



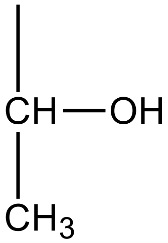
Backbone bonds: red
Side chain bonds: black

Polar or Hydrophilic Amino Acids

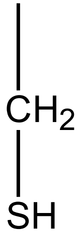
Serine (Ser, S)



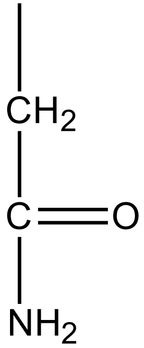
Threonine (Thr, T)



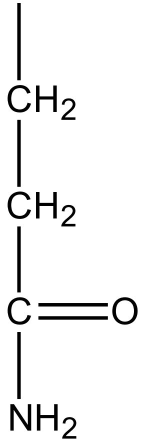
Cysteine (Cys, C)



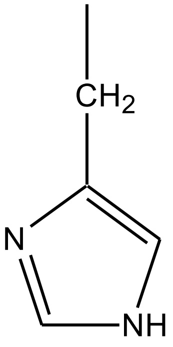
Asparagine (Asn, N)



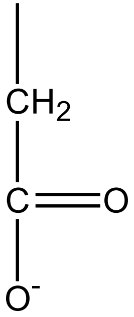
Glutamine (Gln, Q)



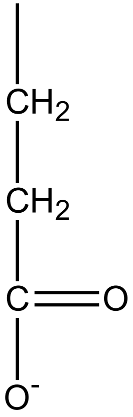
Histidine (His, H)



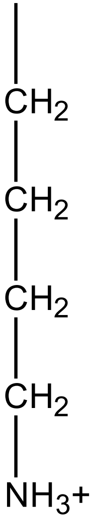
Aspartic Acid (Asp, D)



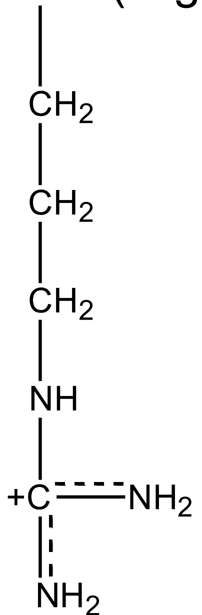
Glutamic Acid (Glu, E)



Lysine (Lys, K)

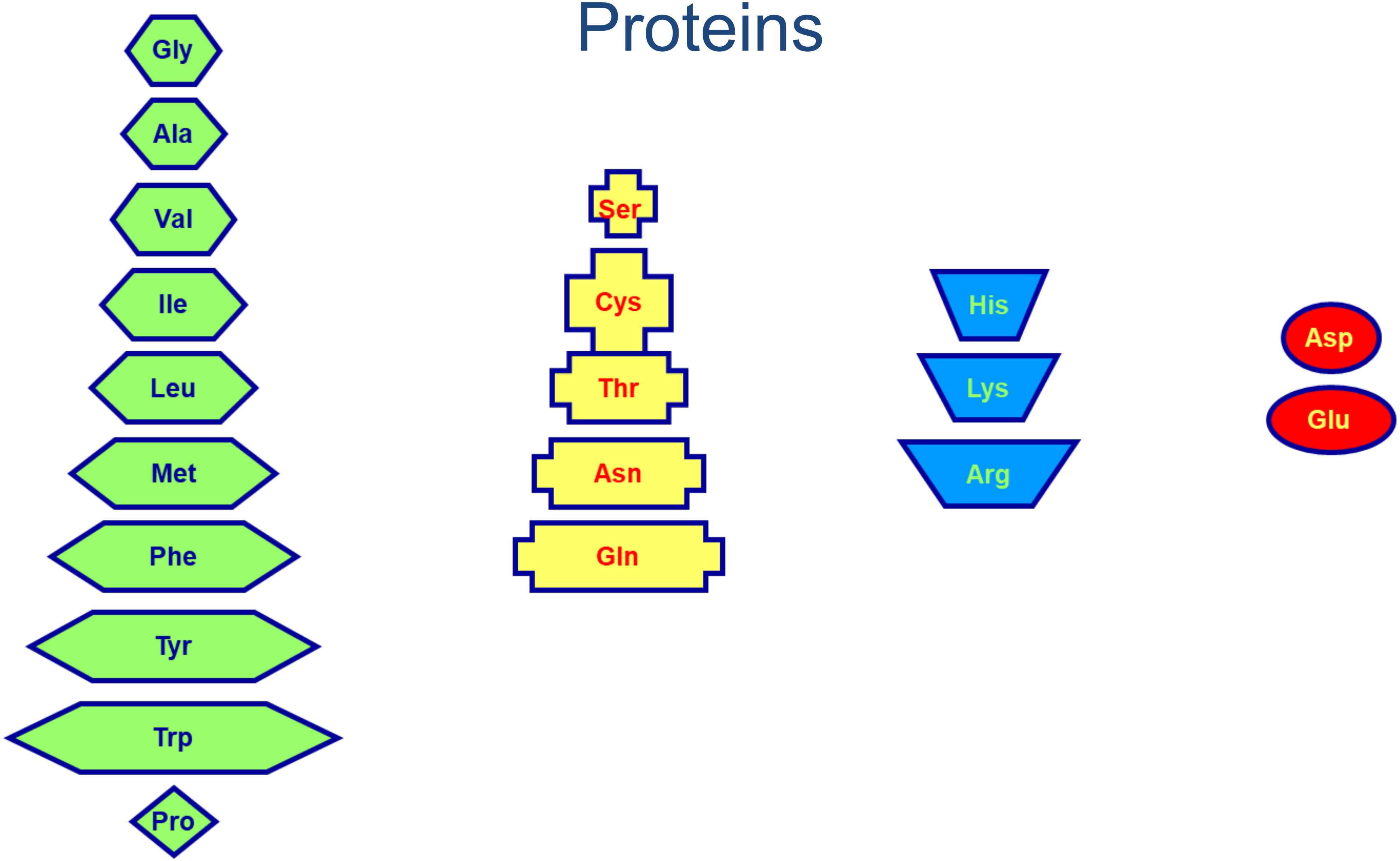


Arginine (Arg, R)



(pymol 1ubq – show how to display sequence, explain atom coloring, select a specific amino acid type)

The Building Blocks of All Proteins



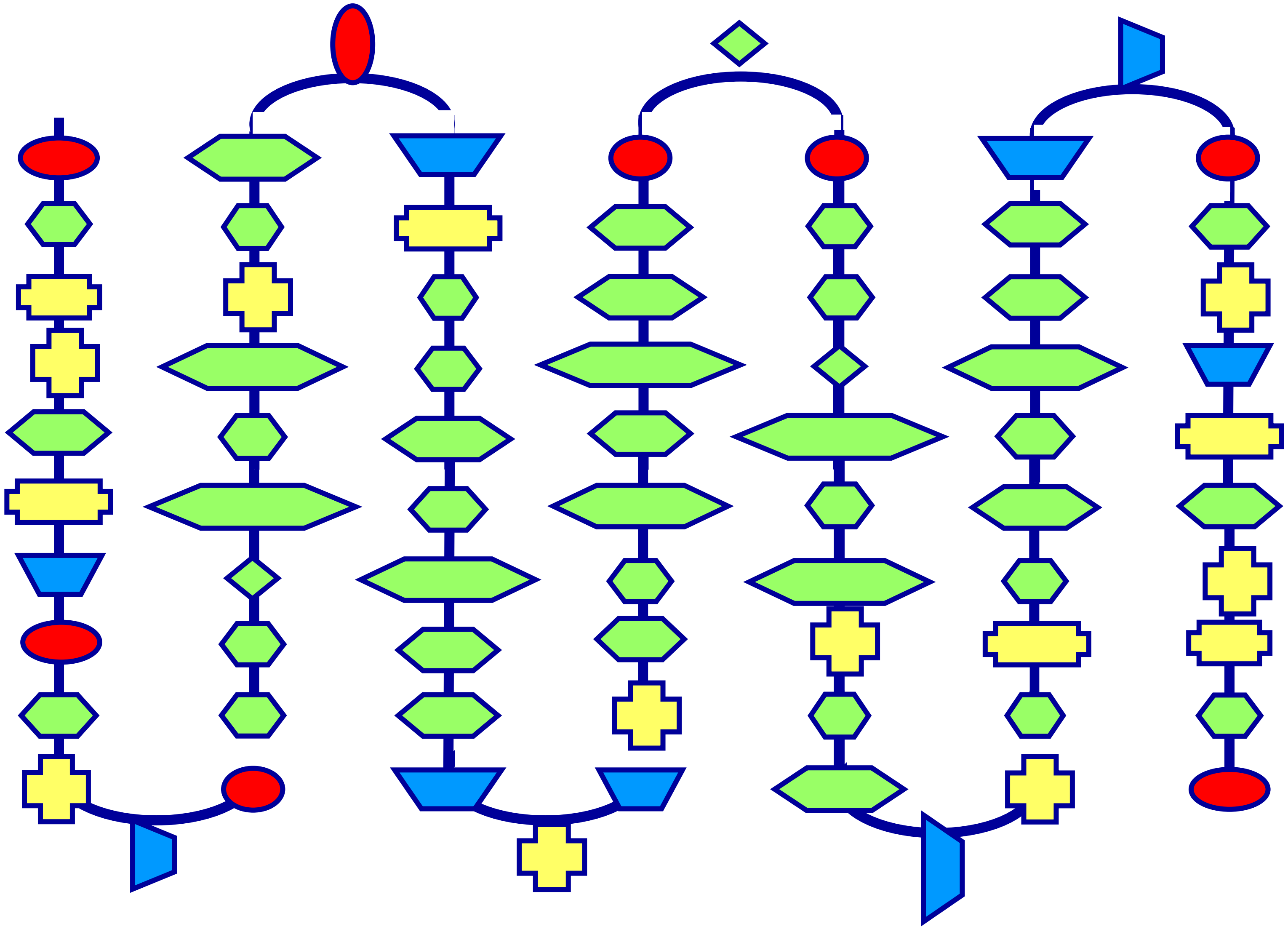
Gly
Ala
Val
Ile
Leu
Met
Phe
Tyr
Trp
Pro

Ser
Cys
Thr
Asn
Gln

His
Lys
Arg

Asp
Glu

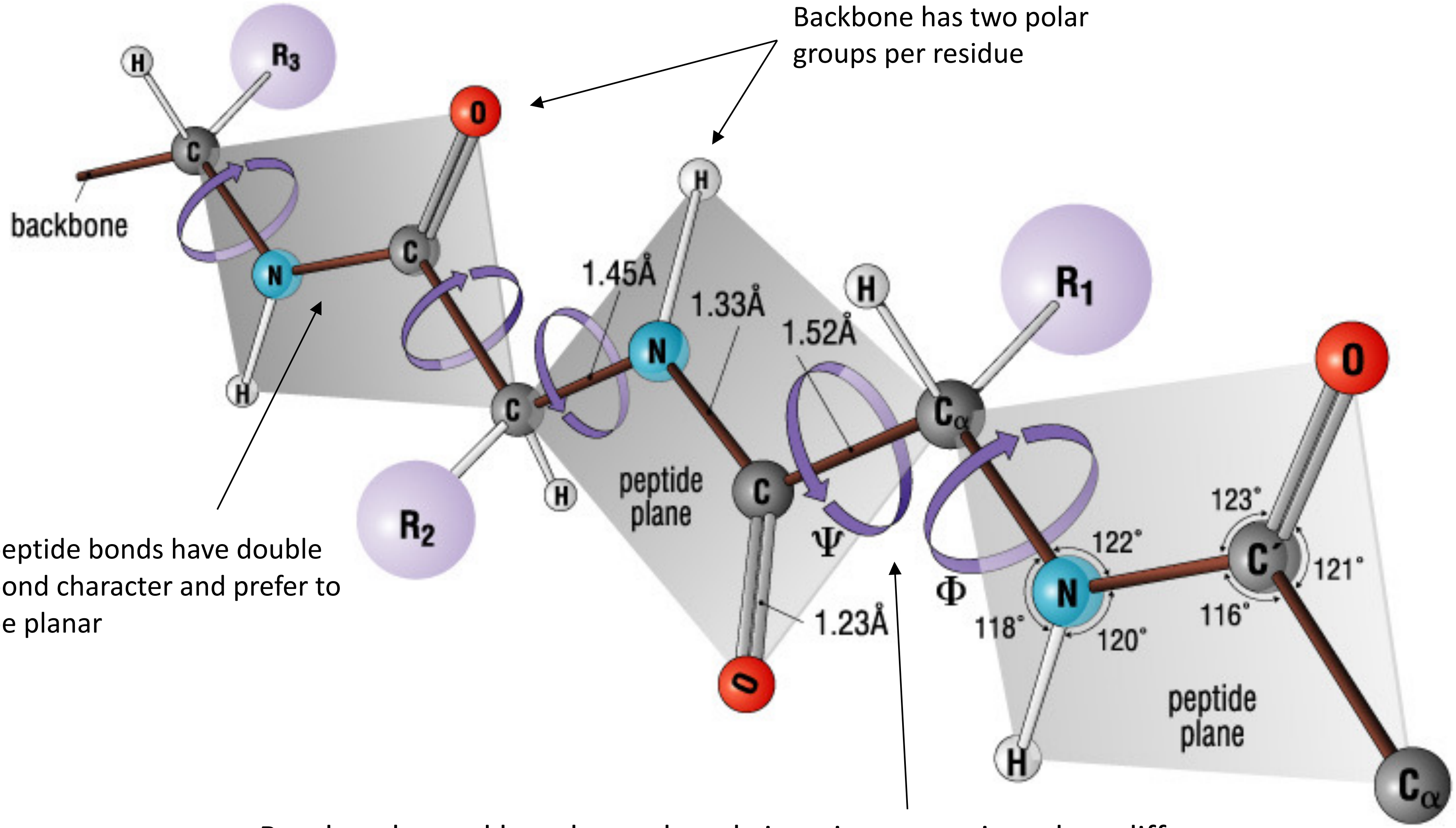
A Polypeptide Chain



Linking amino acids by forming peptide units.

The order of the amino acids is called the "Primary Structure" of a protein

General Features of Polypeptides



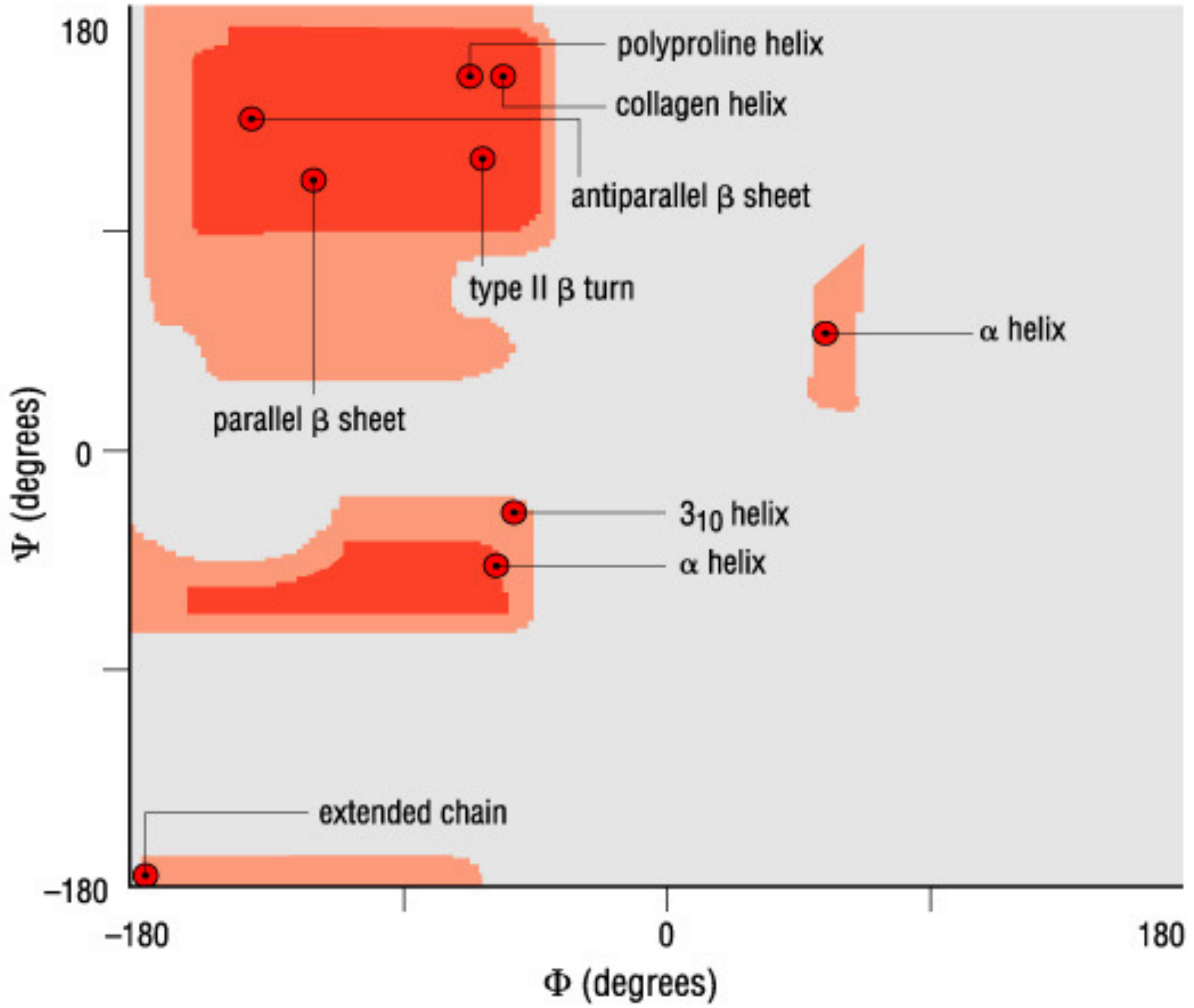
Backbone has two polar groups per residue

peptide bonds have double bond character and prefer to be planar

Bond angles and lengths are largely invariant, proteins adopt different conformations by varying phi and psi

(pymol -> show how to measure distances, angles and torsions)

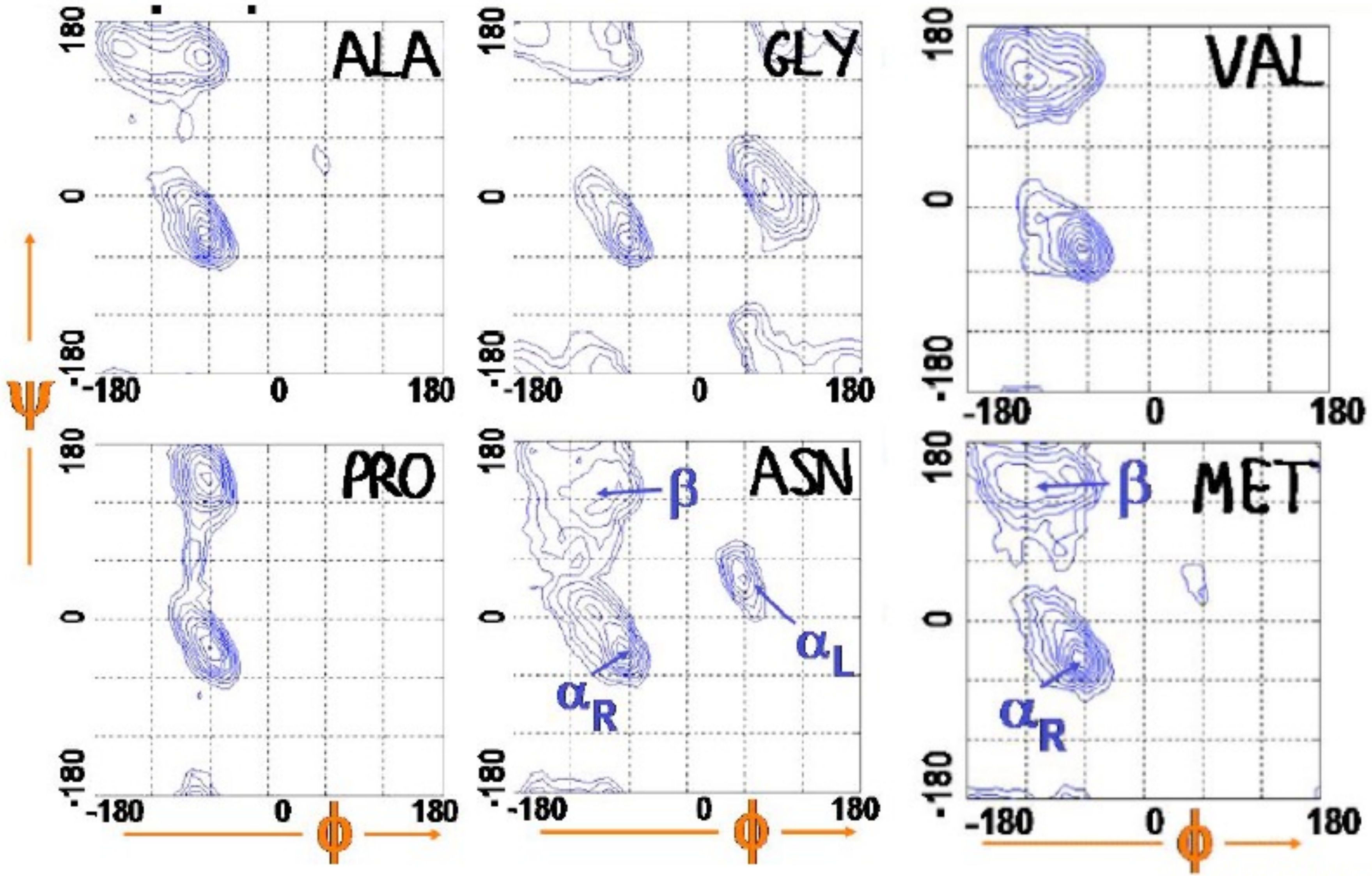
Ramachandran (Φ,Ψ) Plot



(pymol -> exploration of Ramachandran space)

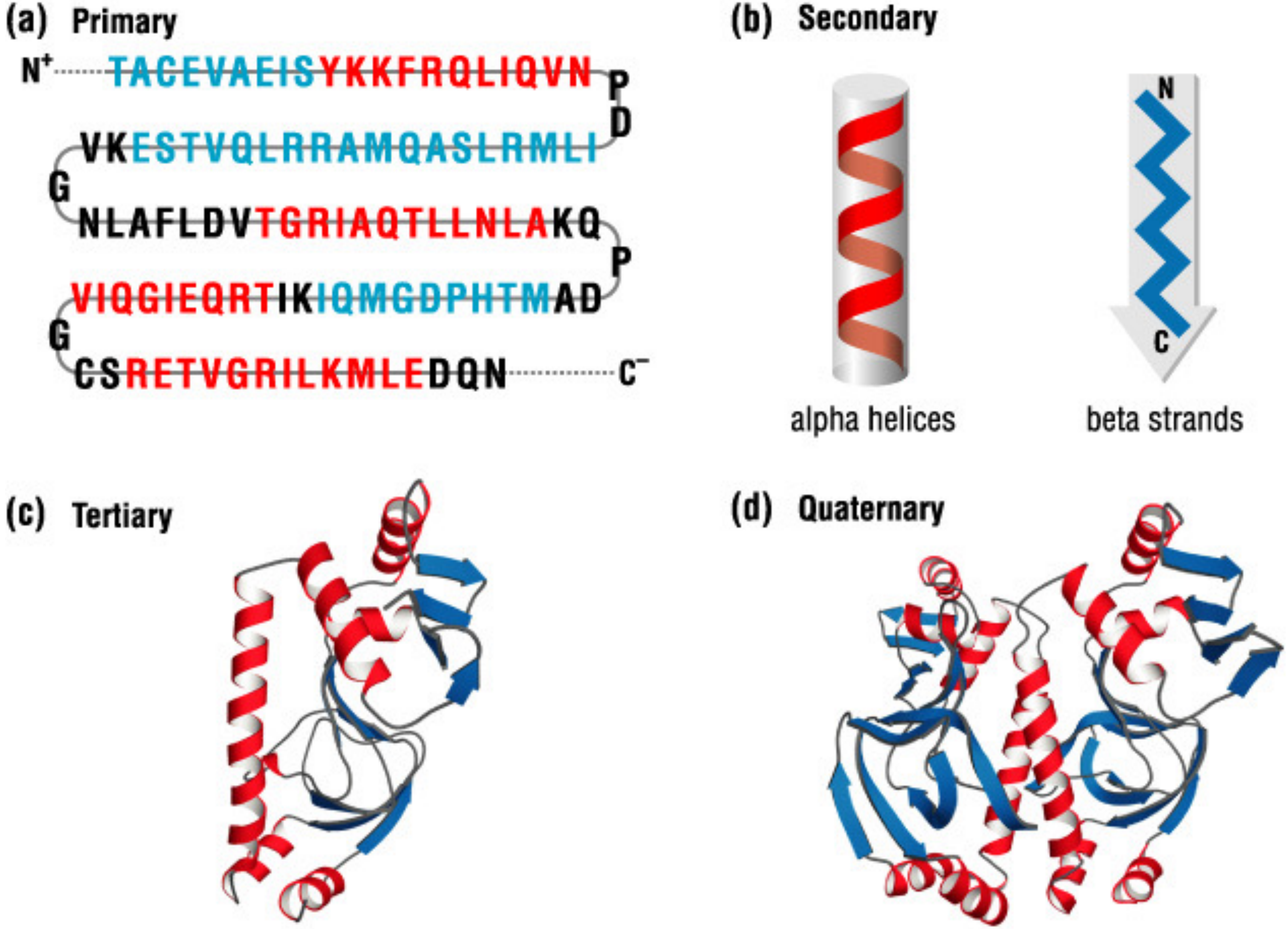
https://pymolwiki.org/index.php/Set_phipsi

Sidechain dependence of Ramachandran angles



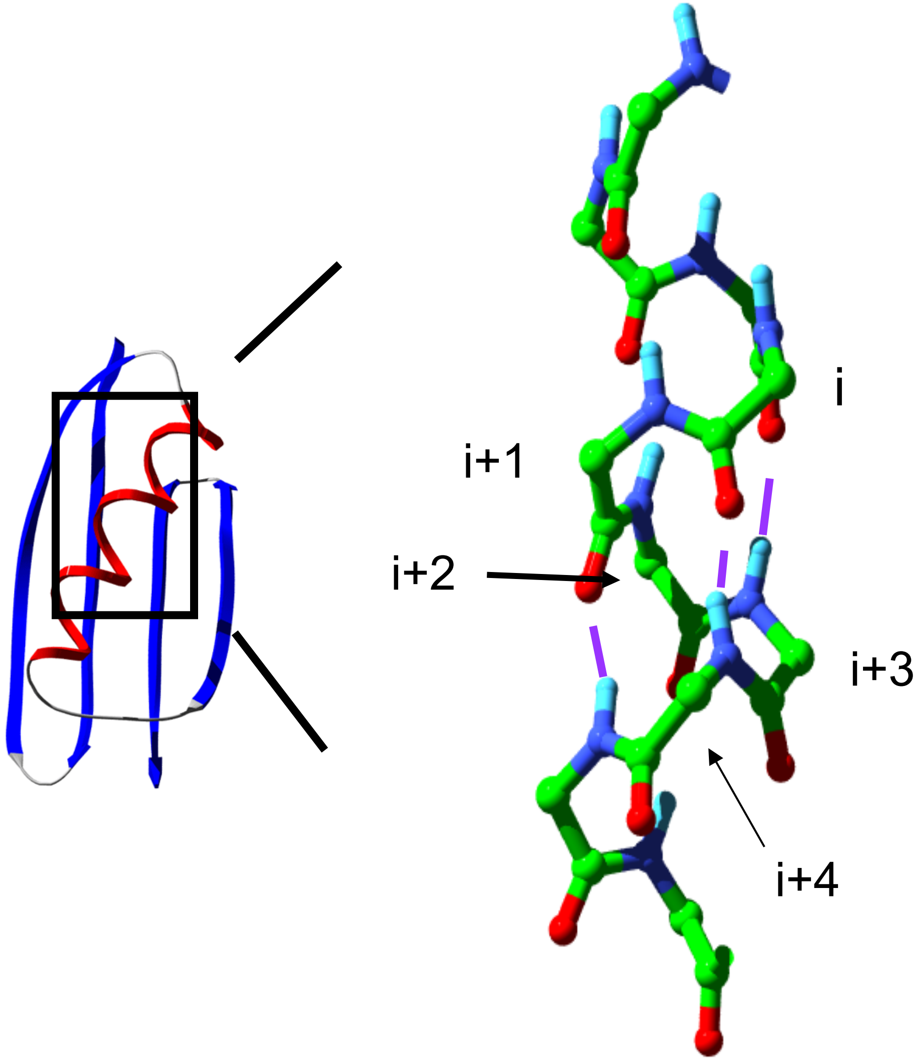
- Torsion preferences vary for different sidechains
- Most look like alanine because of clashes with Cβ

Higher-order Structure



(pymol -> show cartoon representation)

Protein Secondary Structure: The α-helix



Purple: Hydrogen Bonds

Red: Oxygen

Dark Blue: Nitrogen

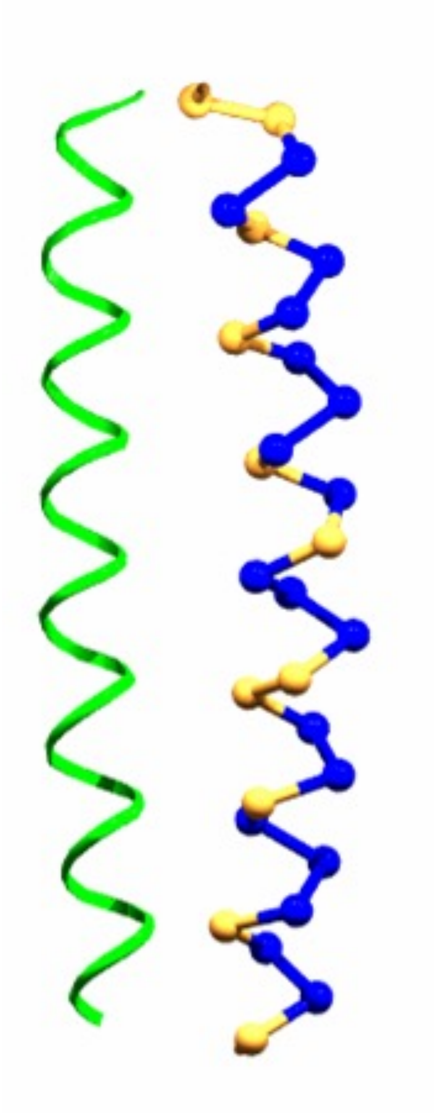
Light Blue: Hydrogen

Green: Carbon

A standard α-helix has hydrogen bonds between *residues i and i+4.*

(pymol show hydrogen bonds in helix)

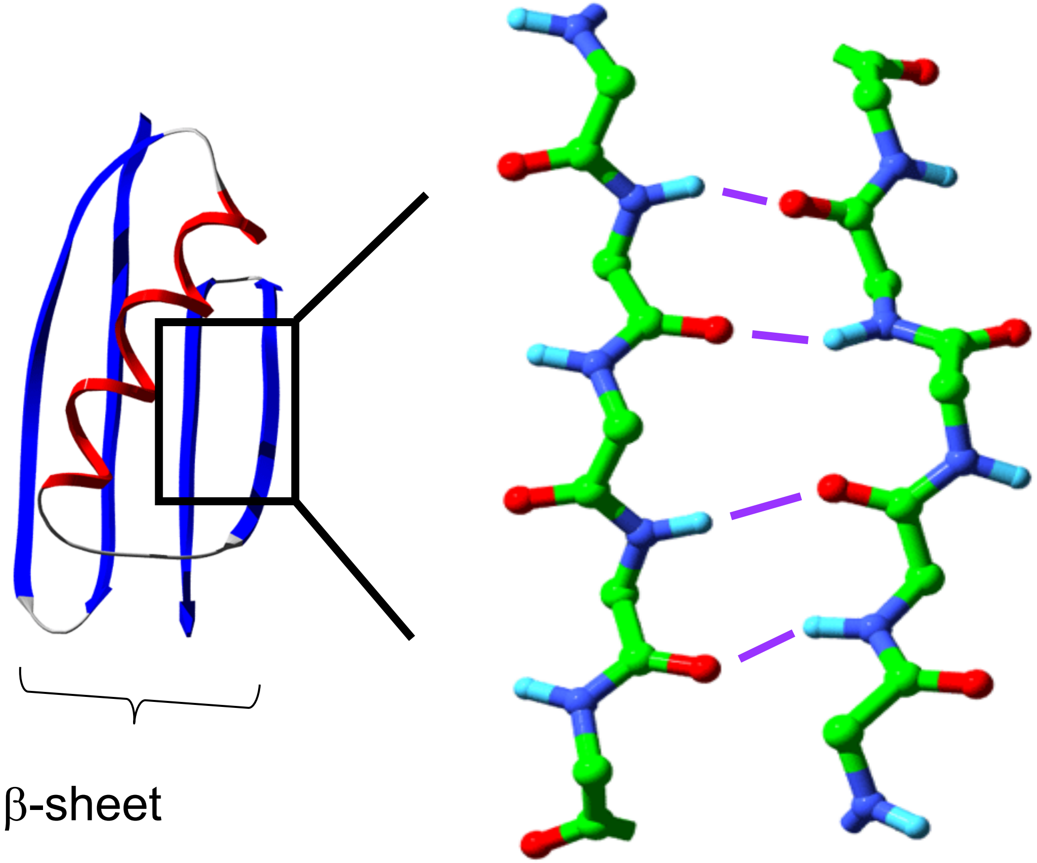
Amphipathic α-Helix



Yellow: hydrophobic amino acids
Blue: hydrophylic amino acids

$$\text{Val} \underbrace{\text{ – Lys – Glu – }}_{3} \text{Leu} \underbrace{\text{ – Leu – Asp – Lys – Val}}_{4} \text{ - Glu}$$

Protein Secondary Structure: The β-strand



Purple: Hydrogen Bonds

Red: Oxygen

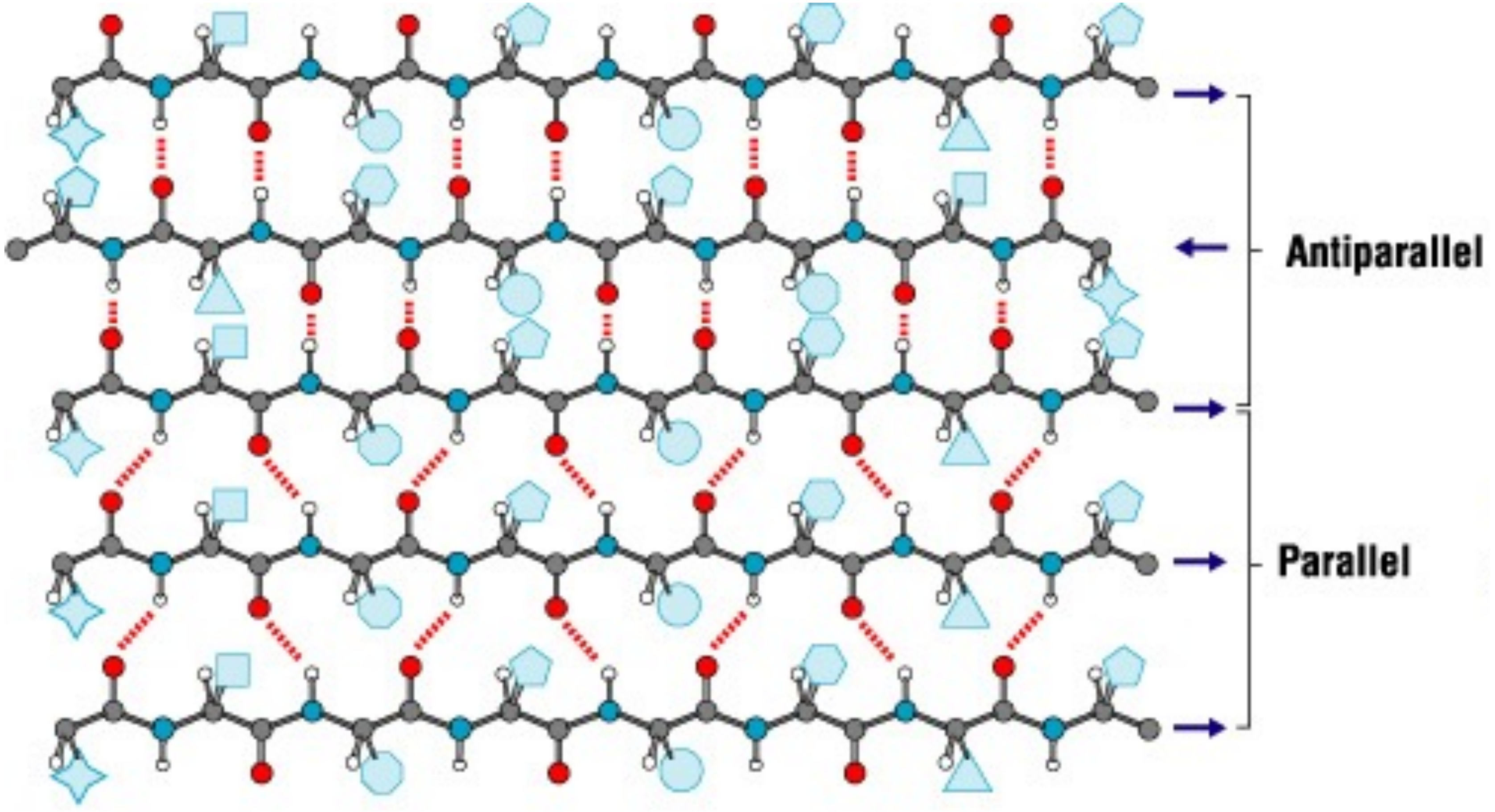
Dark Blue: Nitrogen

Light Blue: Hydrogen

Green: Carbon

β-strands come together to form β-sheets (the interaction can be either parallel or anti-parallel).

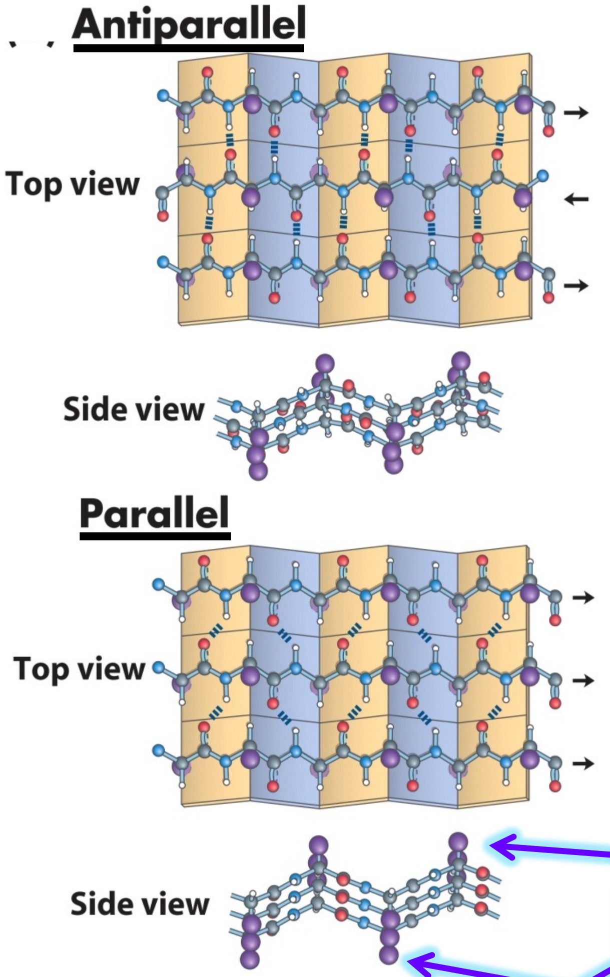
Parallel vs Antiparallel β-strand Interactions

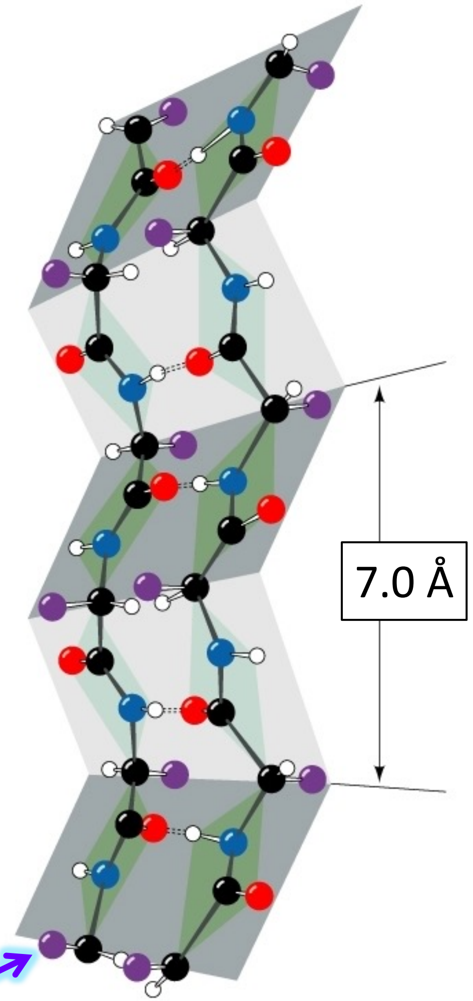


(pymol show beta sheets)

β-sheets form a “pleated sheet”

Note:
There are also many MIXED ß-sheets, with some strands parallel and others anti-parallel

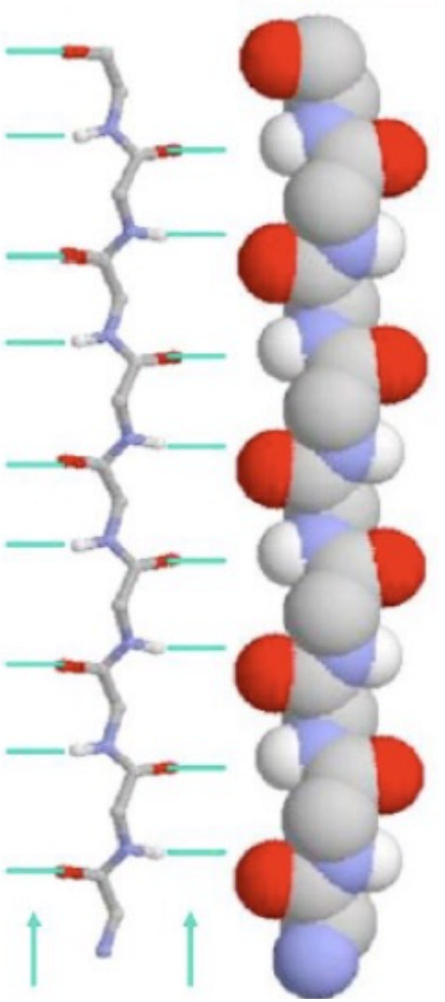




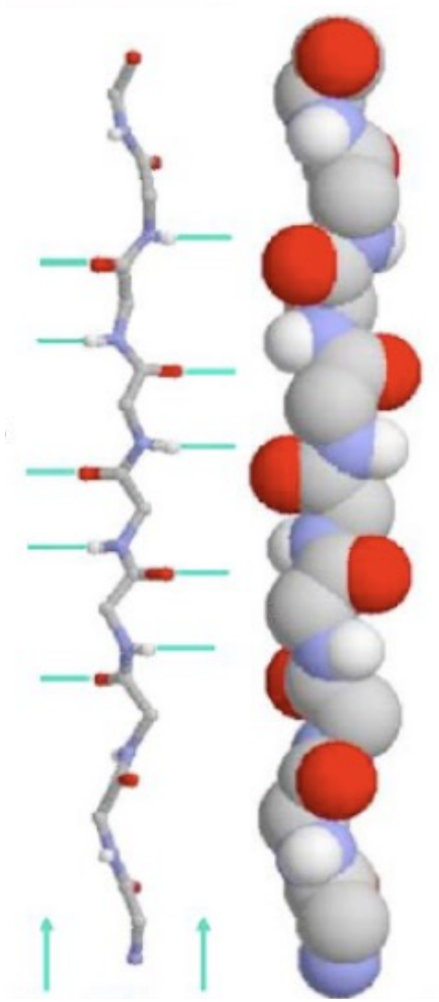
C^{β} of side chain

In both parallel and anti-parallel β-sheets:
The side chains point alternatingly in opposite directions

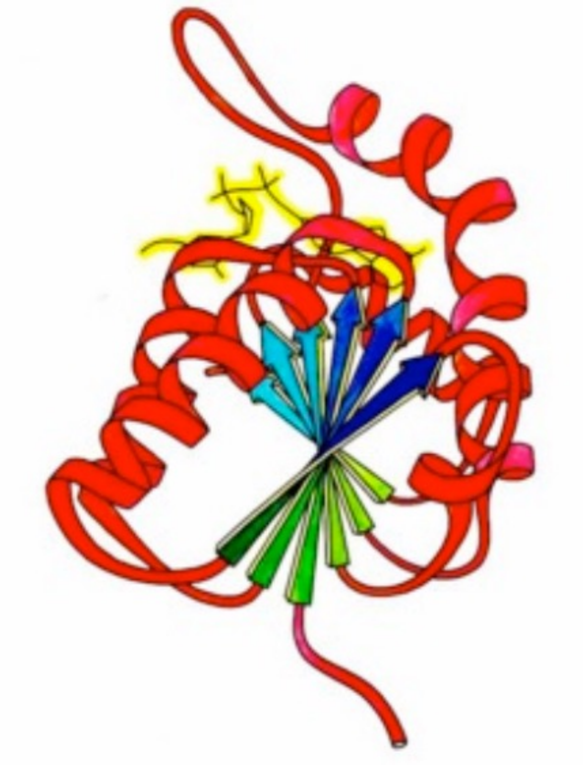
β-strands: why are they twisted?



A fully extended chain is flat

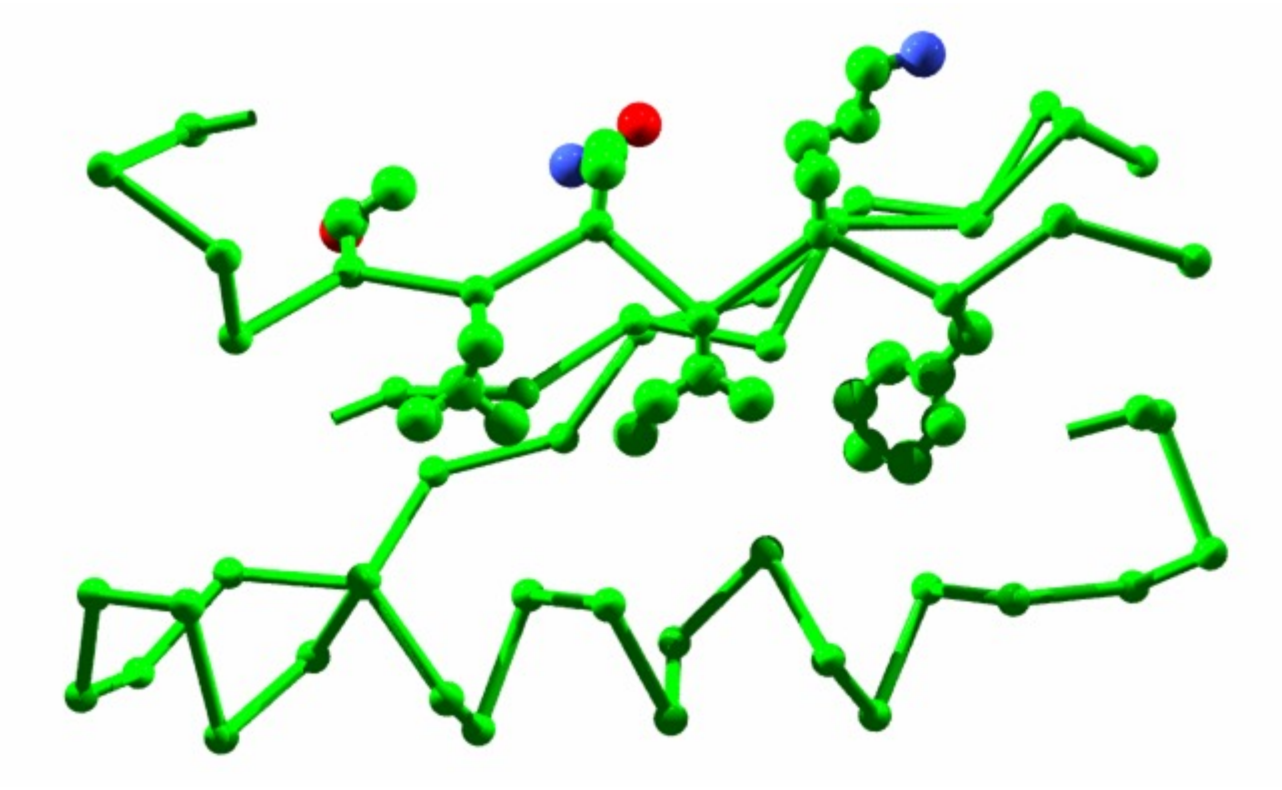


Real beta strands twist and are not flat



Lactate Dehydrogenase domain 1, end view

Hydrophobic / hydrophilic patterning in β-strands



$\underbrace{\text{Thr – Leu – Asn}}_{2}$ – Ile – Lys - Phe

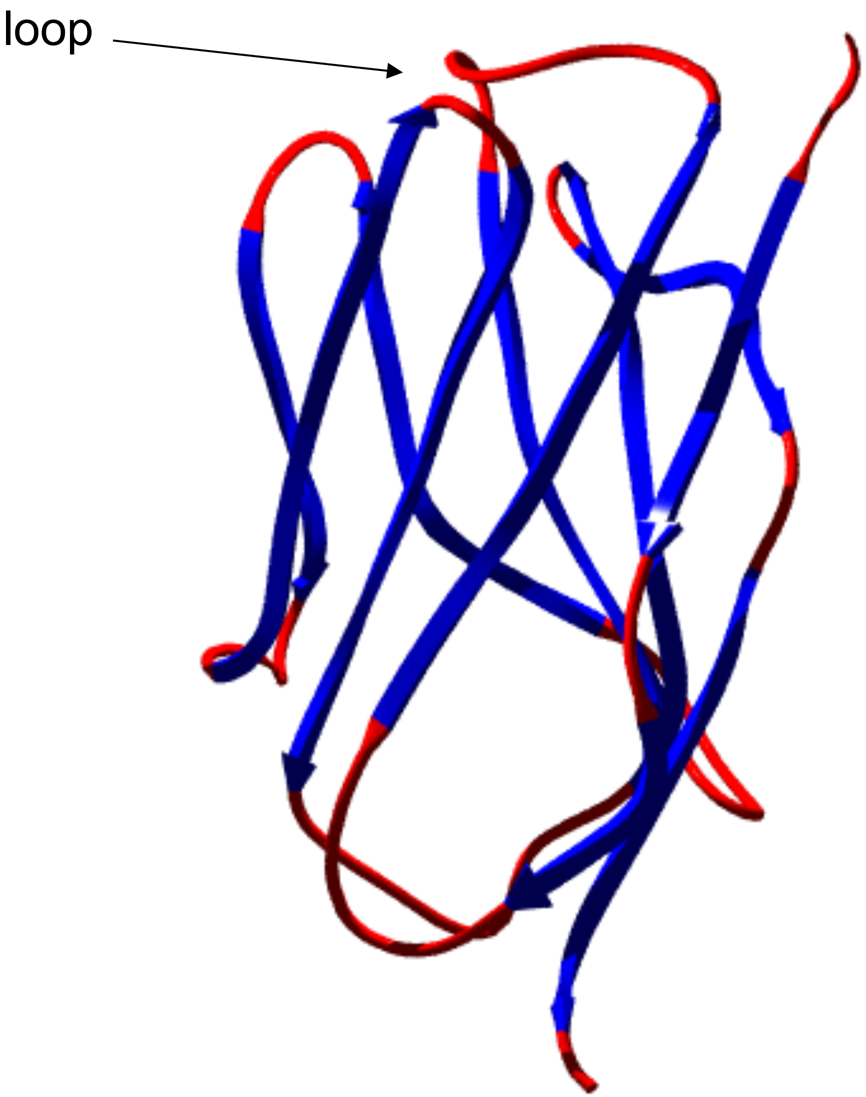
(pymol -> show hydrophobic patterning in beta sheet)

Protein Secondary Structure: Loops and Turns

Example: an antigen binding domain of an antibody

Active site residues and binding residues are often found in loops.

Turns are short loops (2-4 residues), and typically have more regular structure than loops.



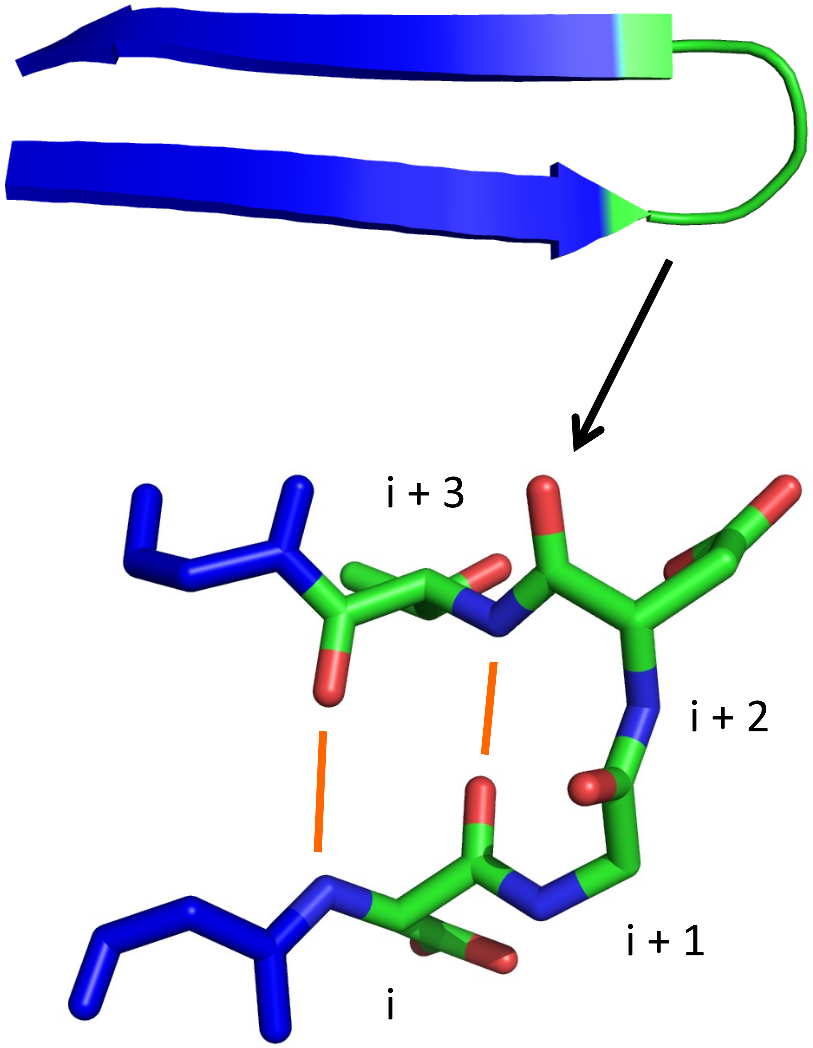
Between secondary and tertiary structure

- **Supersecondary structure**: arrangement of elements of same or different secondary structure into *motifs*; a motif is usually not stable by itself.

- **Domains**: A domain is an independent unit, usually stable by itself; it can comprise the whole protein or a part of the protein.

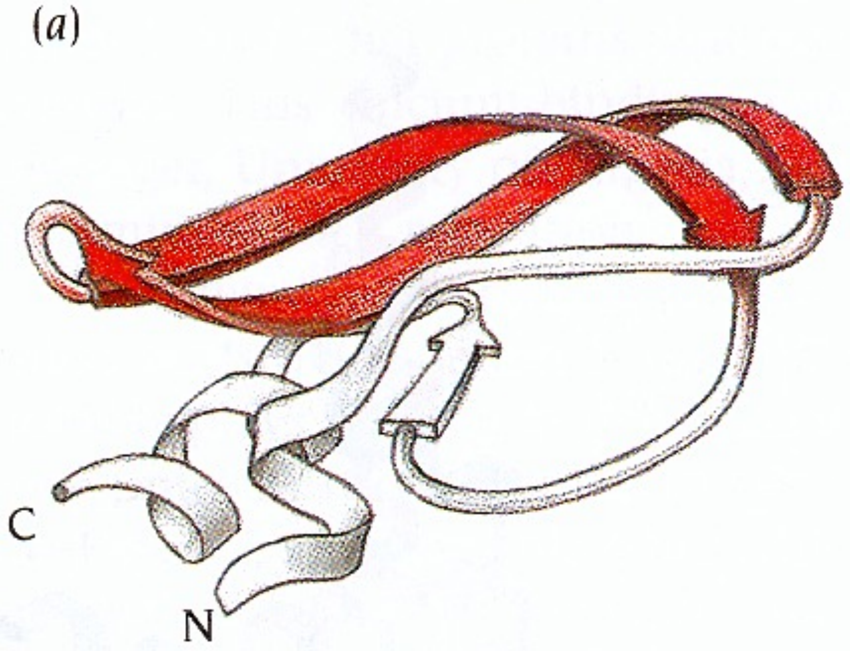
β-hairpin: Most common form of tight turn

type	Φ_{i+1}	Ψ_{i+1}	Φ_{i+2}	Ψ_{i+2}
I	-60	-30	-90	0
I’	60	30	90	0
II	-60	120	80	0
II’	60	-120	-80	0

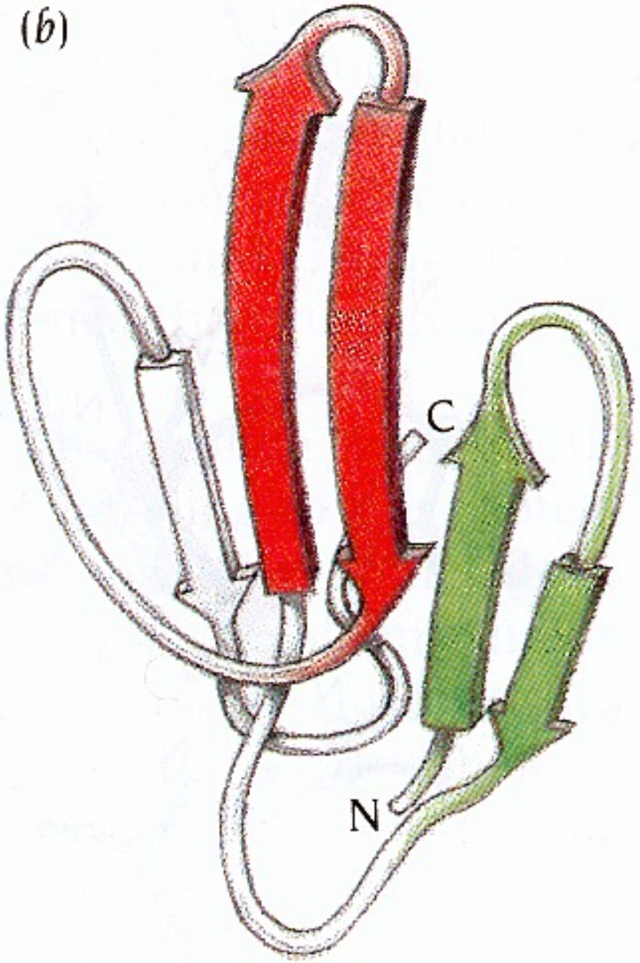


Type II’

β-hairpin: Most common form of tight turn

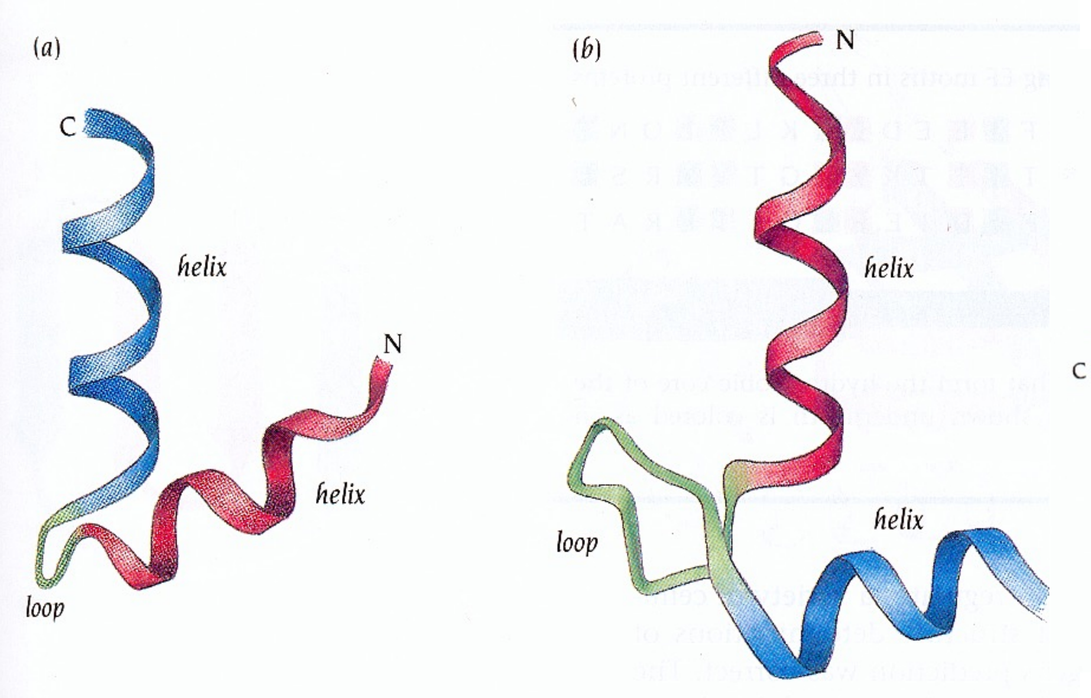


Example of a β-hairpin in bovine pancreatic trypsin inhibitor– BPTI.



Example of a protein with two β-hairpins: erabutoxin from whale.

The helix-turn-helix motif



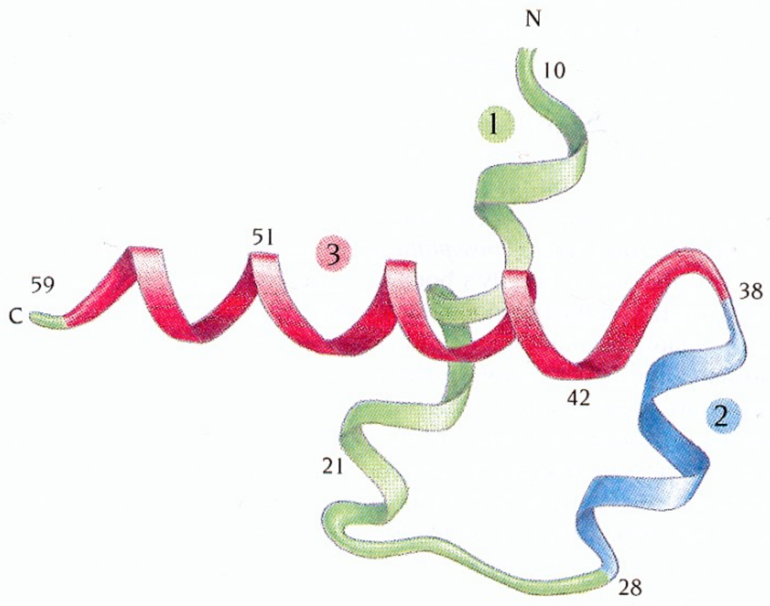


Figure 9.8 Schematic diagram of the three-dimensional structure of the Antennapedia homeodomain. The structure is built up from three α helices connected by short loops. Helices 2 and 3 form a helix-turn-helix motif (blue and red) similar to those in procaryotic DNA-binding proteins. (Adapted from Y.Q. Qian et al., *Cell* 59: 573–580, 1989.)

- This motif is characteristic of proteins binding to the major DNA grove.
- The proteins containing this motif recognize palindromic DNA sequences.
- The second helix is responsible for nucleotide sequence recognition.

The helix-turn-helix motif

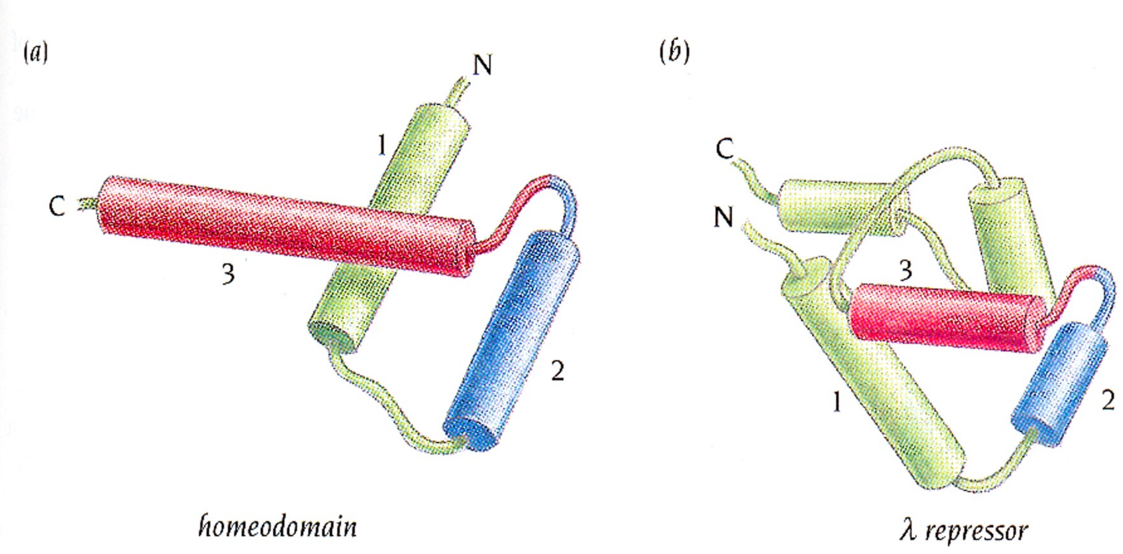


Figure 9.9 Comparison of the helix-turn-helix motifs in homeodomains (a) and λ repressor (b). The recognition helix (red) of the homeodomain is longer than in the procaryotic repressor motif. In addition the first helix of the homeodomain [(green in (a)] is oriented differently.

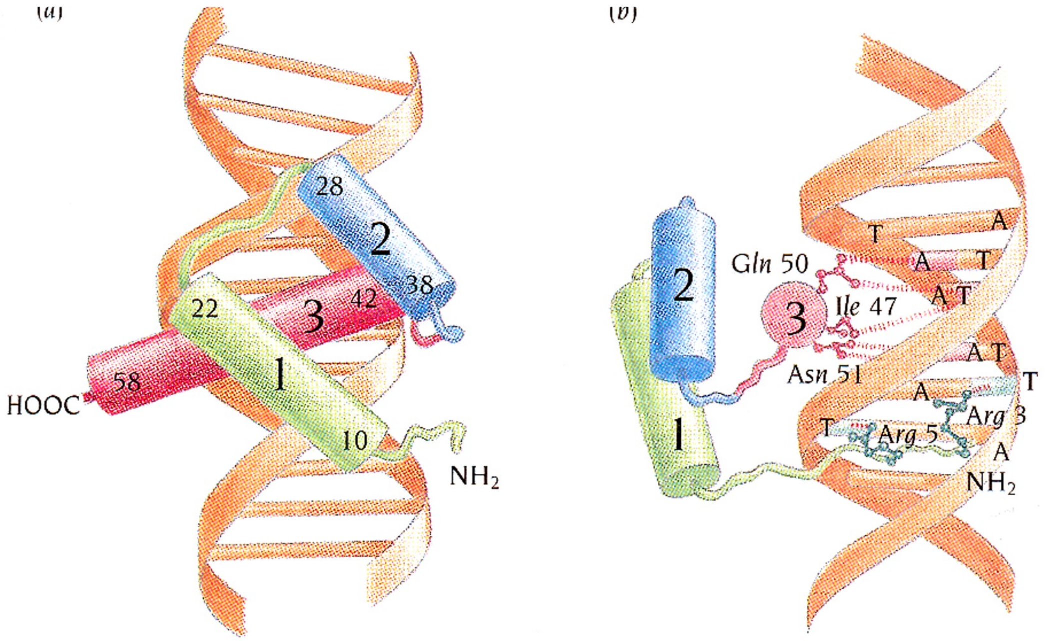
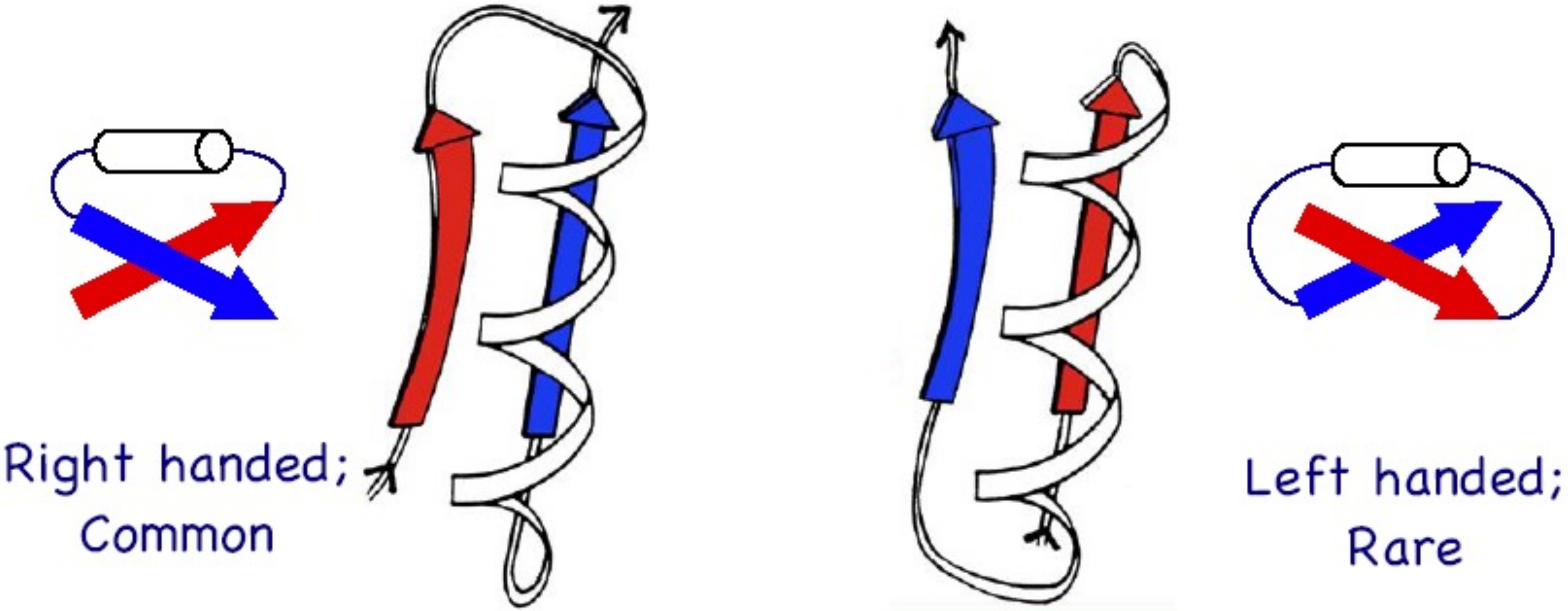


Figure 9.10 Schematic diagrams illustrating the complex between DNA (orange) and one monomer of the homeodomain. The recognition helix (red) binds in the major groove of DNA and provides the sequence-specific interactions with bases in the DNA. The N-terminus (green) binds in the minor groove on the opposite side of the DNA molecule and arginine side chains make nonspecific interactions with the phosphate groups of the DNA. (Adapted from C.R. Kissinger et al., *Cell* 63: 579–590, 1990.)

βαβ motif



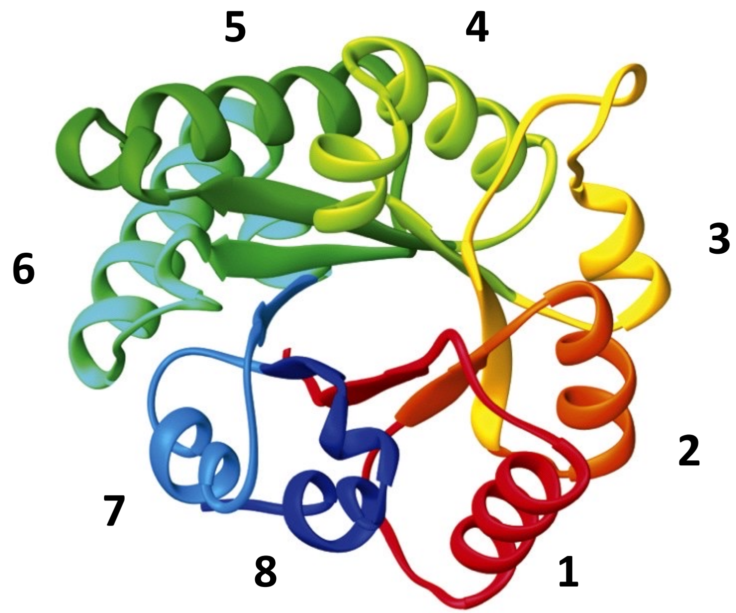
Why?

- Shorter connections in right-handed topology?
- Accessibility to helix termini for hydrogen bonding?
- Trapped ends?

Triose Phosphate Isomerase (TIM)

A domain which occurs in a many proteins.

Note the “β-barrel” in the center surrounded by α-helices



Note the 8-fold repeated β-α motif

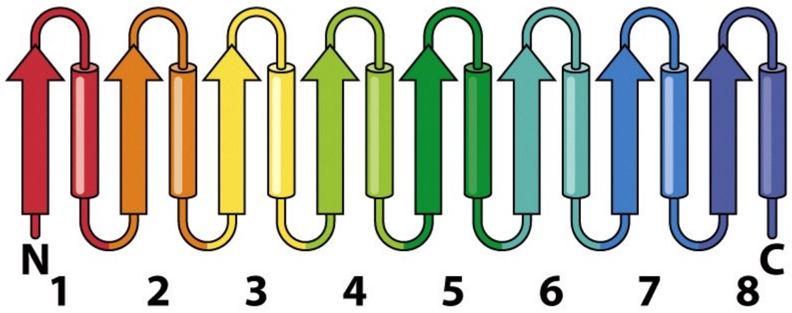


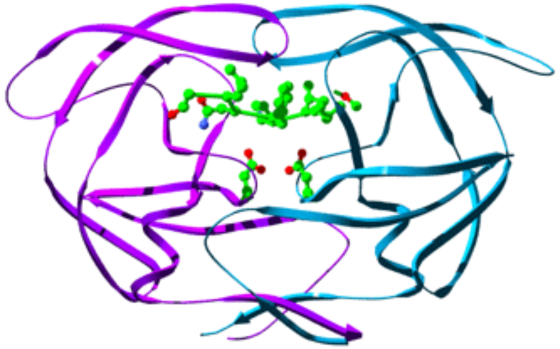
Figure 6-30c

The “TIM barrel” : α/β class topology

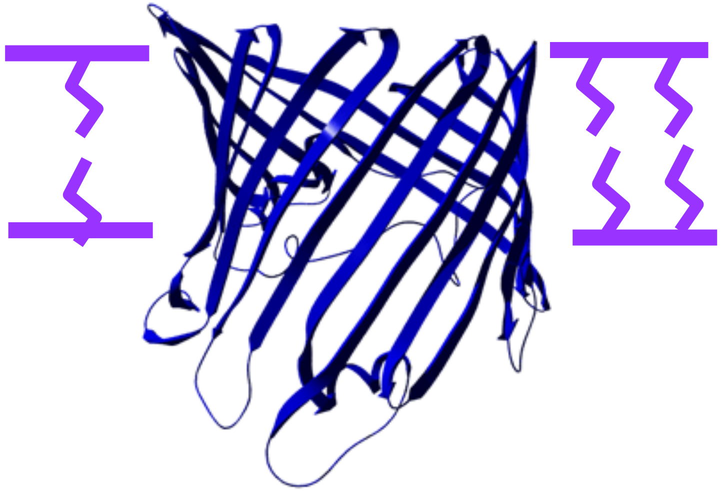
Protein Tertiary Structure

- Most proteins adopt a unique three-dimensional structure that is essential to the biological role they perform. Protein structures can be divided into three groups: **globular proteins**, **fibrous proteins, and integral membrane proteins.**

Examples:



HIV protease
(globular)



Porin
(membrane)

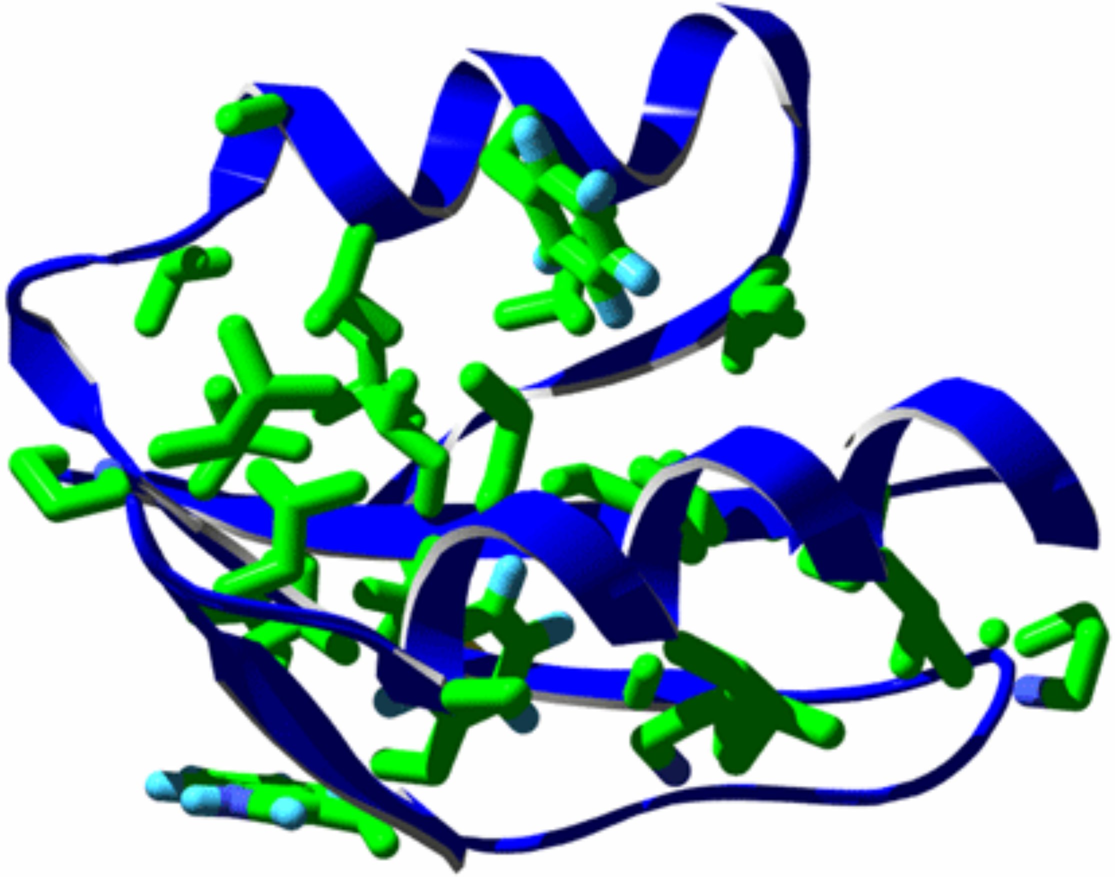
Collagen
(fibrous)

Most globular proteins share these characteristics

1) **Hydrophobics on the inside**

2) Close packing

3) Most polar groups involved in a hydrogen bond



Hydrophobic residues of procarboxypeptidase

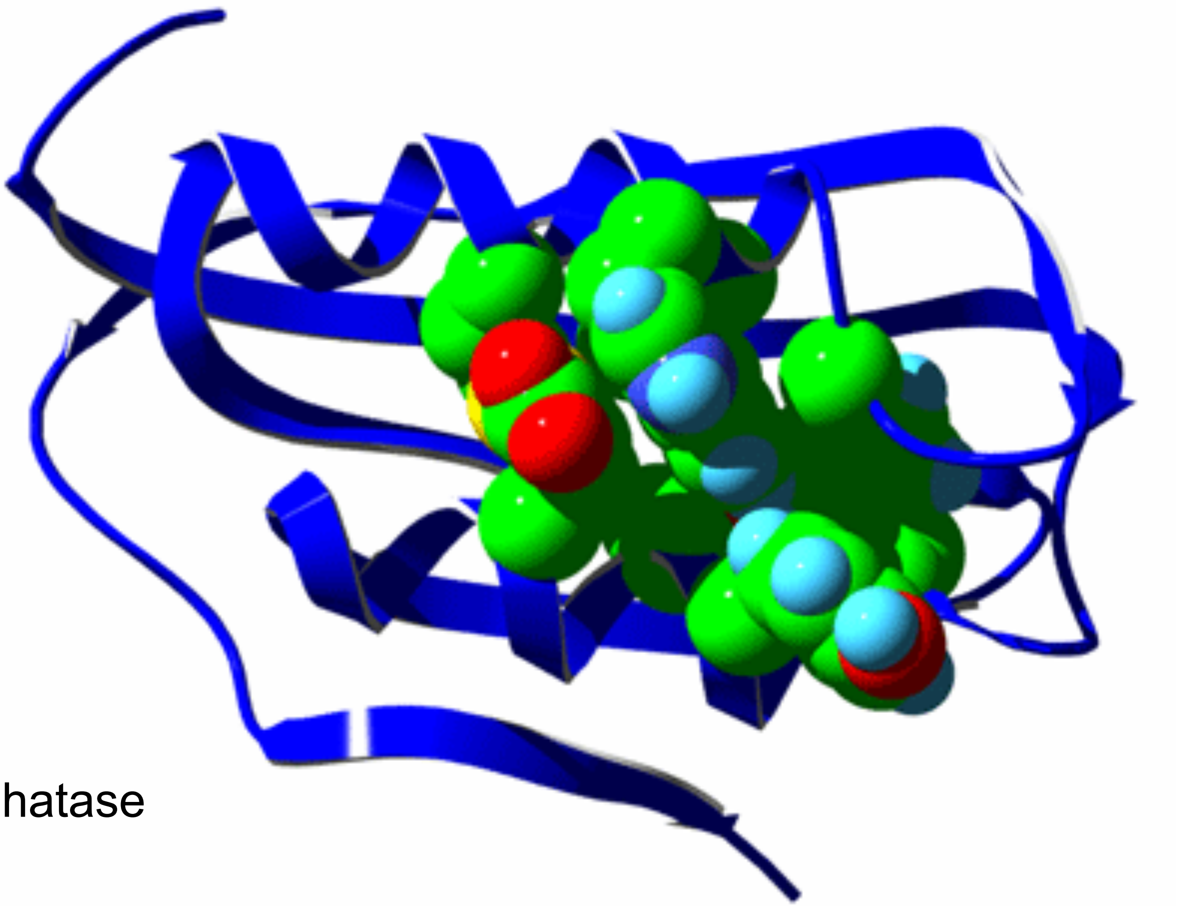
(pymol 1urr – highlight hydrophobics)

Most globular proteins share these characteristics

1) Hydrophobics on the inside

2) **Close packing**

3) Most polar groups involved in a hydrogen bond



acylphosphatase

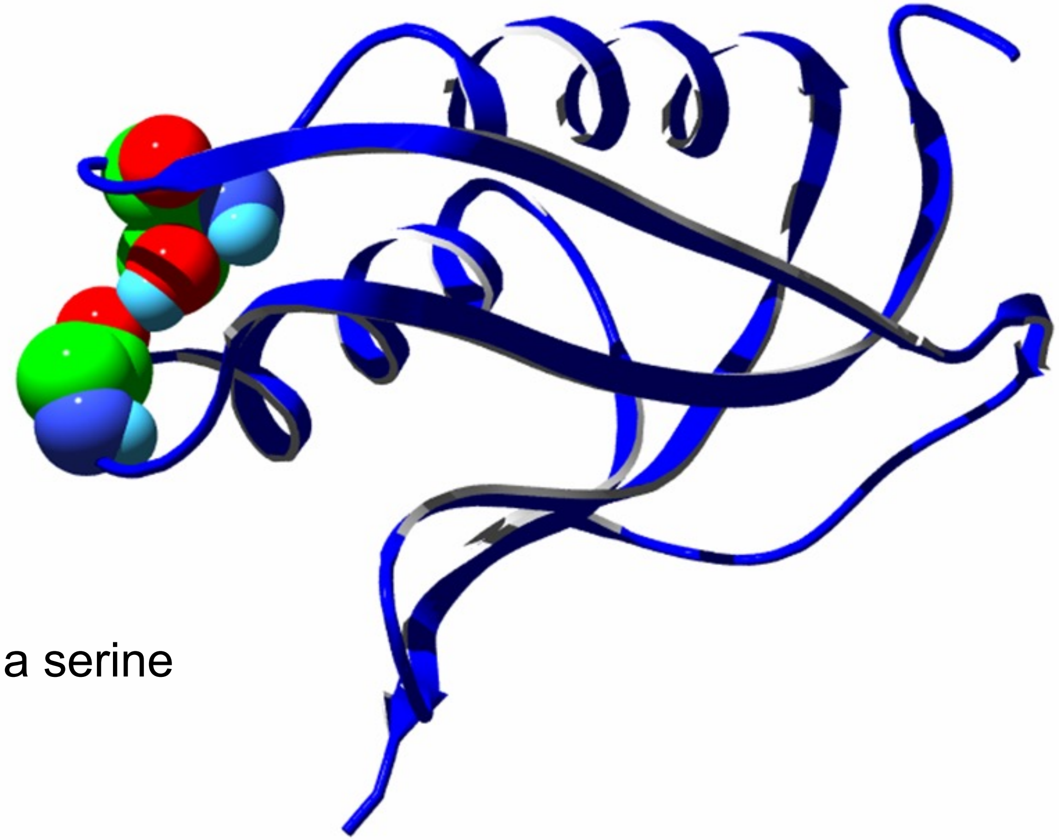
(pymol 1urr – surface & sphere view)

Most globular proteins share these characteristics

1) Hydrophobics on the inside

2) Close packing

3) **Most polar groups involved in a hydrogen bond**



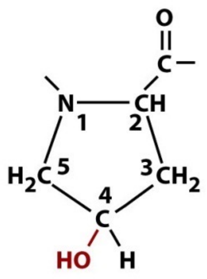
Hydrogen bond between a serine and a backbone carbonyl

Fibrous Proteins

- highly elongated molecules that generally function as structural materials
- their sequences are usually highly repetitive

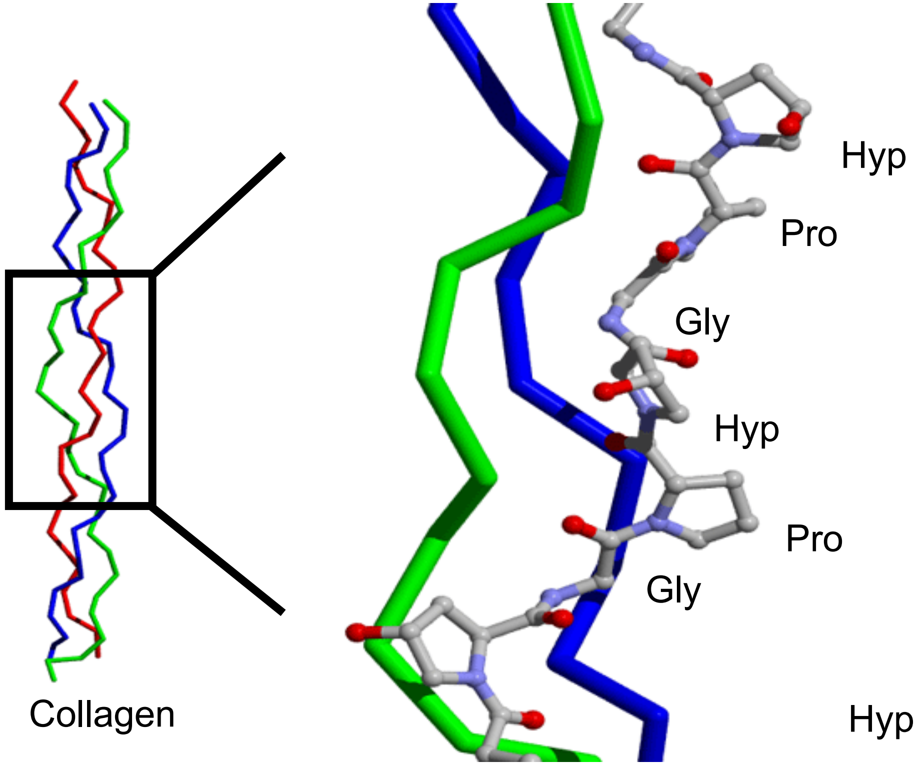
Collagen - a structural component in bone, cartilage, tendon

Sequence: G-X-Y



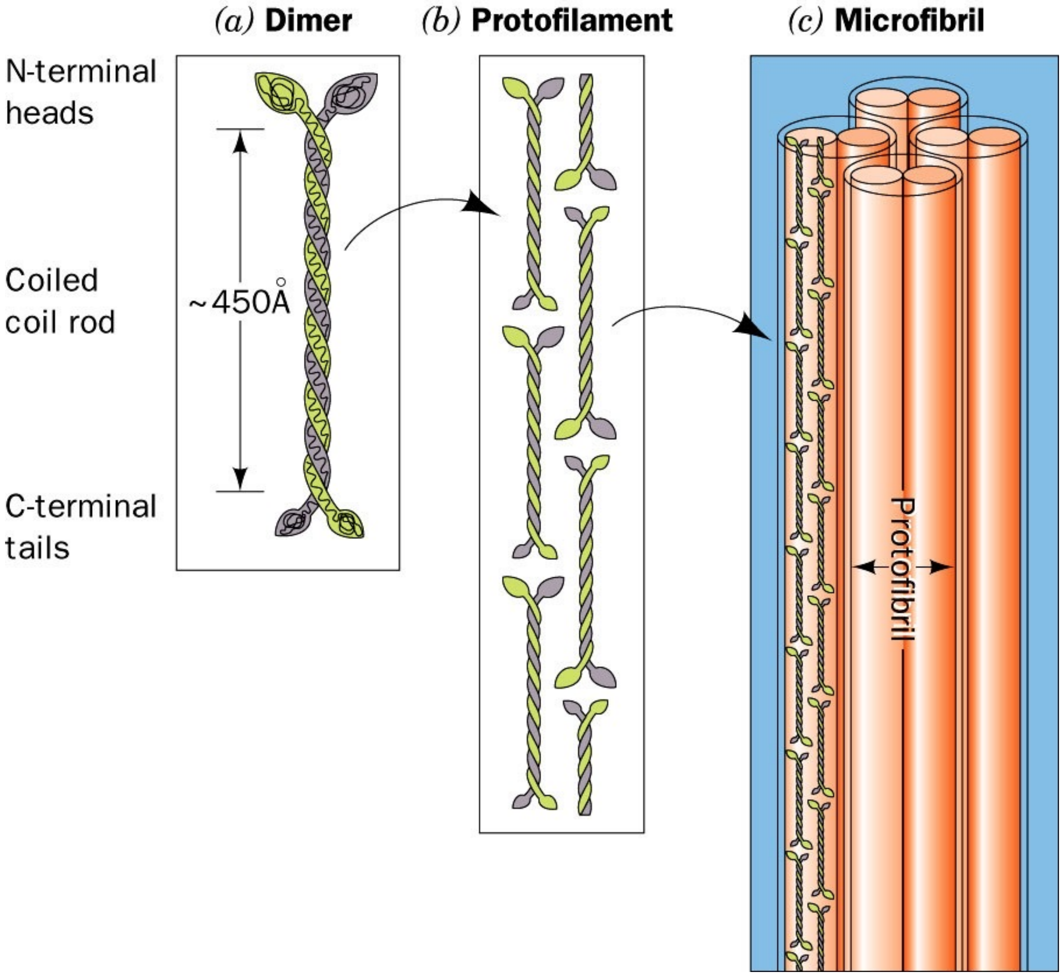
4-Hydroxyprolyl residue (Hyp)

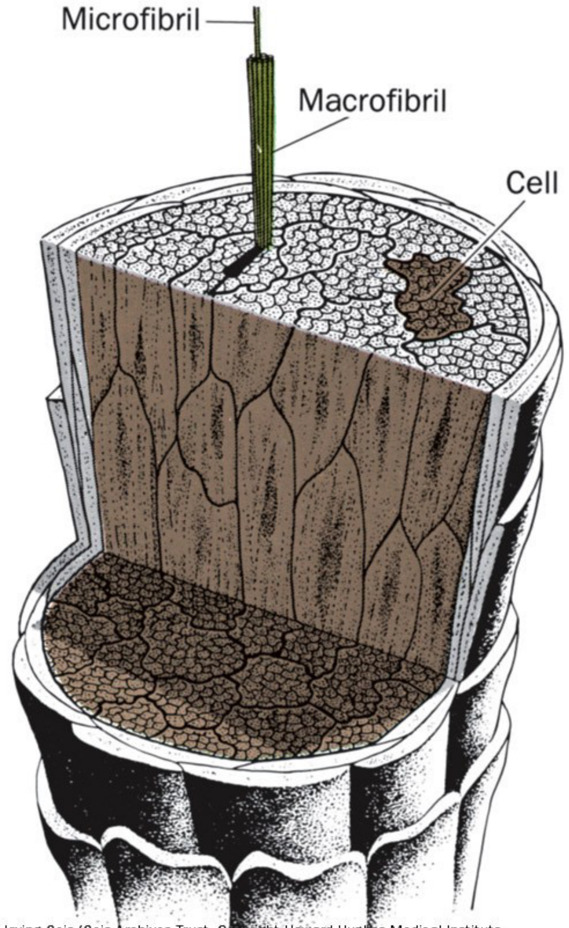
(Vitamin C is required for the enzyme making 4-HydroxyPro)



Hyp = 4-hydroxyproline

α-keratin - the principal protein of mammalian hair, nails, skin



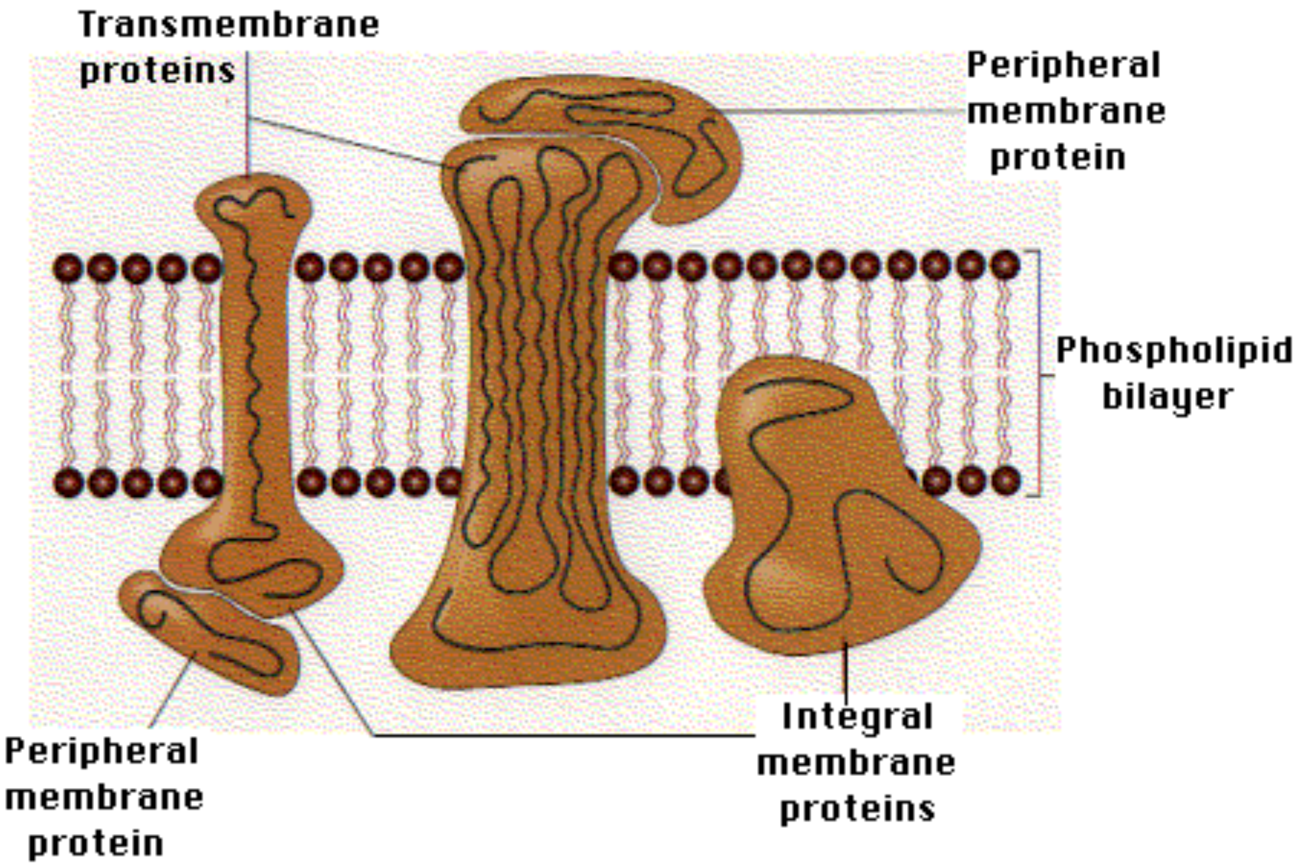


Irving Geis/Geis Archives Trust. Copyright Howard Hughes Medical Institute. Reproduced with perimssion.

The central 310-residue portion of α-keratin has a pseudo-repeat sequence **a**-**b**-**c**-**d**-**e**-**f**-**g** with nonpolar residues at **a** and **d**.

Membrane Proteins

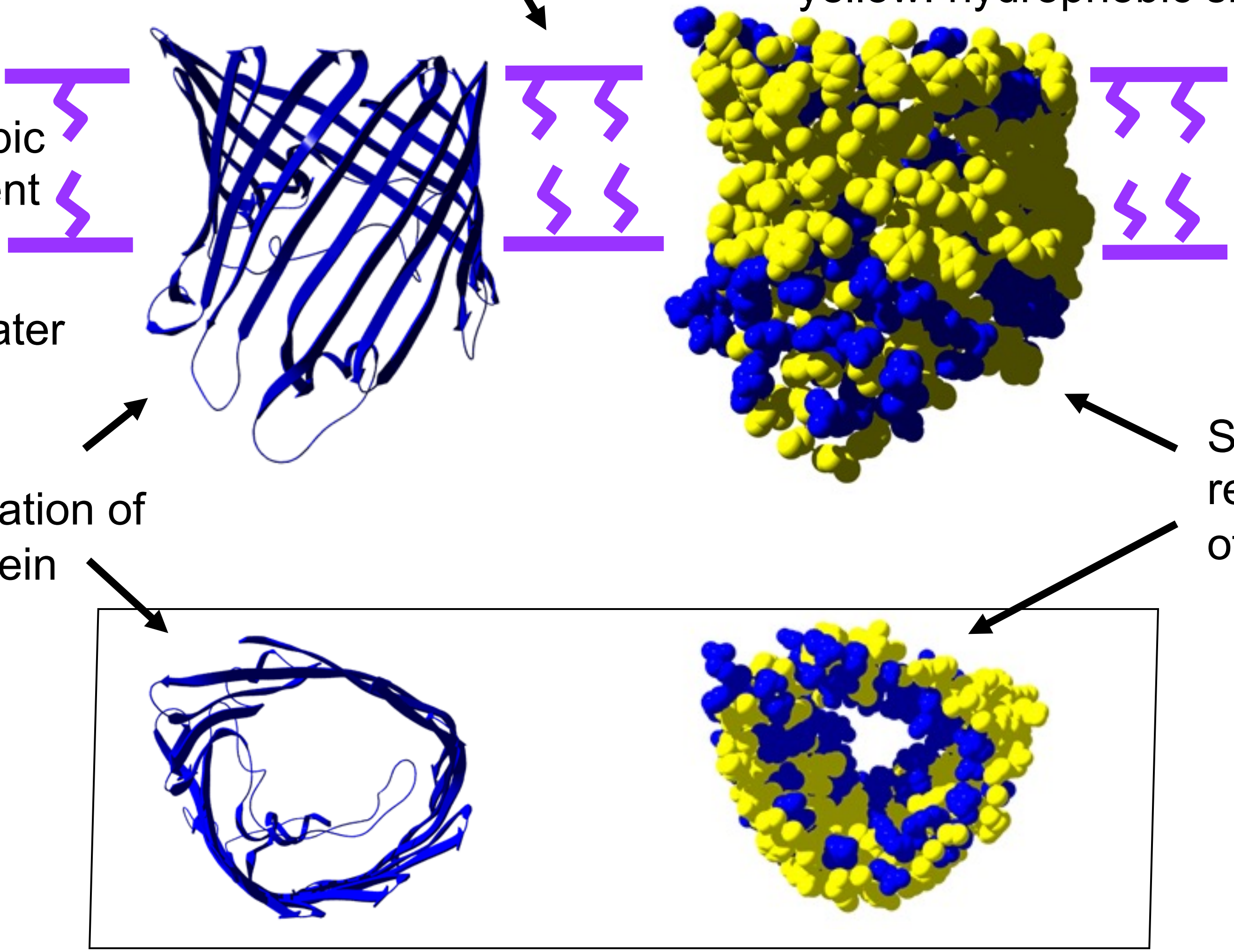
- ~30% of human proteins are membrane proteins
- ~70% of therapeutics are directed towards membrane proteins


Transmembrane proteins
Peripheral membrane protein
Phospholipid bilayer
Integral membrane proteins
Peripheral membrane protein

Membrane proteins are important for:

1) ion and solute transport
2) detection of external signals, e.g. hormones
3) cell-to-cell recognition

Membrane Proteins: hydrophobic residues are found on the exterior



membrane

blue: hydrophilic sidechains
yellow: hydrophobic sidechains

Hydrophobic environment

water

Ribbon representation of porin protein

Space filling representation of porin protein

membrane

(pymol -> show 2POR)

Membrane proteins
are often either all-α or all-β

The protein avoids placing main chain C=O and NH groups in the hydrophobic bilayer)

Bacteriorhodopsin

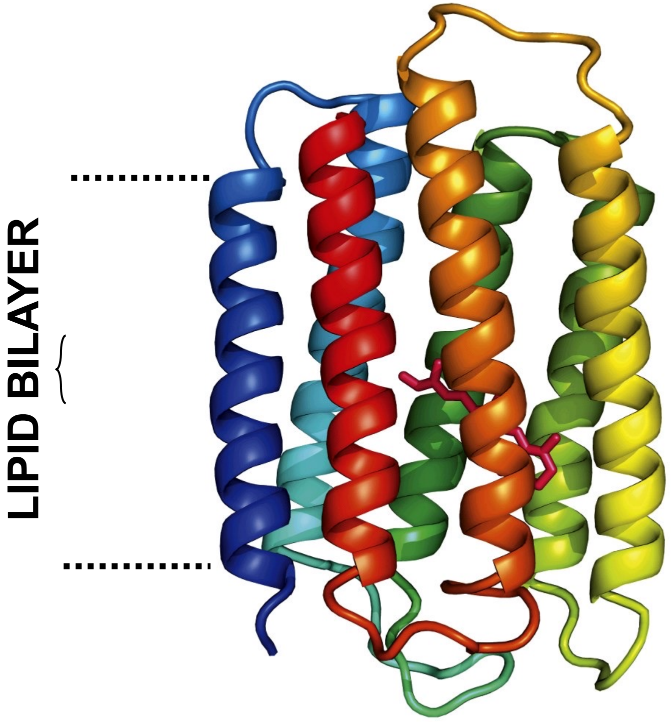


Figure 9-22
© 2013 John Wiley & Sons, Inc. All rights reserved.

α-HELICES crossing the membrane

OmpF Porin

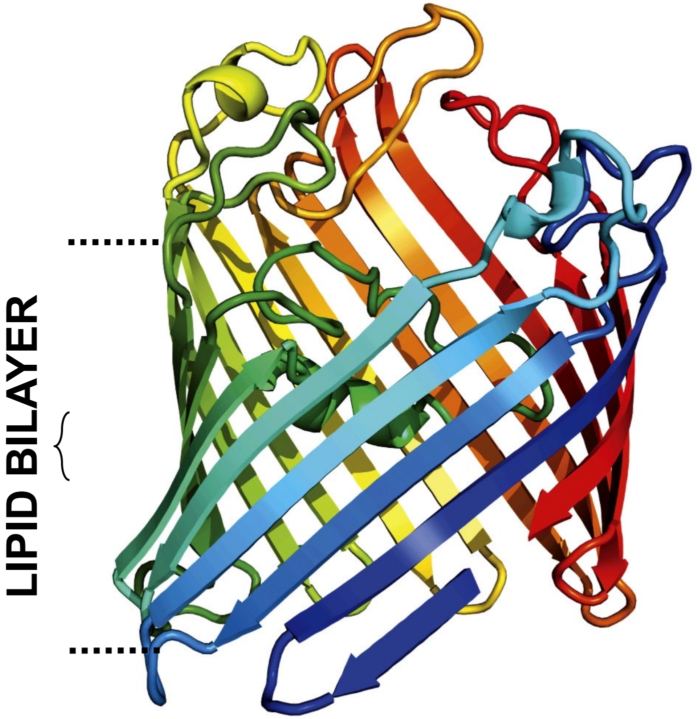
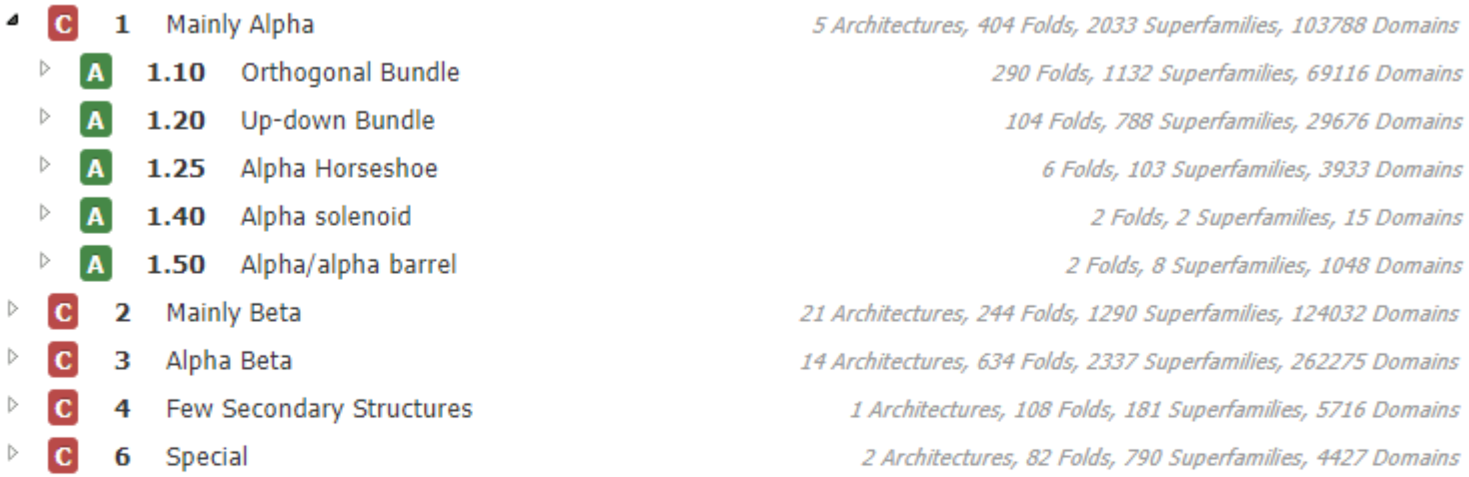


Figure 9-23a
© 2013 John Wiley & Sons, Inc. All rights reserved.

β-BARREL crossing the membrane

CATH

http://www.cathdb.info/browse/tree

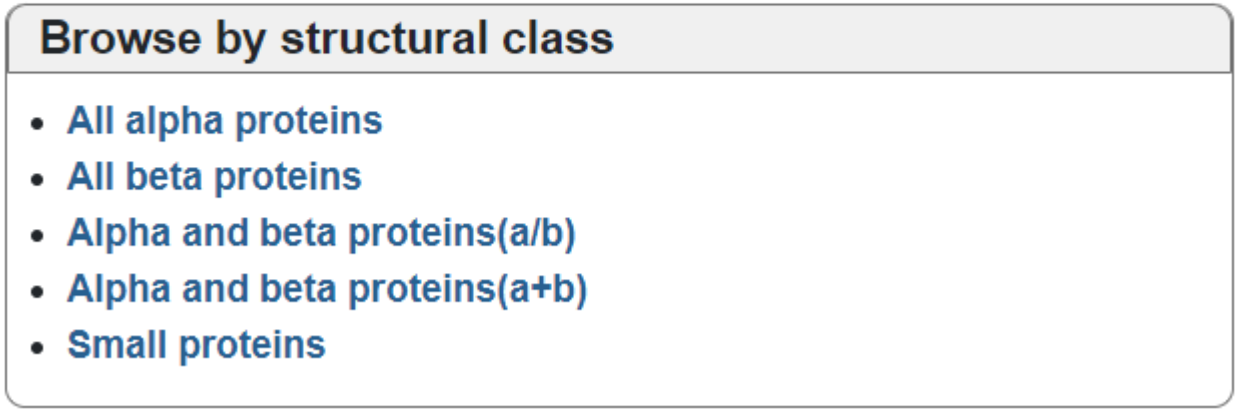


C	1	Mainly Alpha	5 Architectures, 404 Folds, 2033 Superfamilies, 103788 Domains
A	1.10	Orthogonal Bundle	290 Folds, 1132 Superfamilies, 69116 Domains
A	1.20	Up-down Bundle	104 Folds, 788 Superfamilies, 29676 Domains
A	1.25	Alpha Horseshoe	6 Folds, 103 Superfamilies, 3933 Domains
A	1.40	Alpha solenoid	2 Folds, 2 Superfamilies, 15 Domains
A	1.50	Alpha/alpha barrel	2 Folds, 8 Superfamilies, 1048 Domains
C	2	Mainly Beta	21 Architectures, 244 Folds, 1290 Superfamilies, 124032 Domains
C	3	Alpha Beta	14 Architectures, 634 Folds, 2337 Superfamilies, 262275 Domains
C	4	Few Secondary Structures	1 Architectures, 108 Folds, 181 Superfamilies, 5716 Domains
C	6	Special	2 Architectures, 82 Folds, 790 Superfamilies, 4427 Domains

- a combination of manual and automated hierarchical classification
- four major levels:
  - Class (C) – based on secondary structure content
  - Architecture (A) – based on gross orientation of secondary structures
  - Topology (T) – based on connections and numbers of secondary structures
  - Homologous superfamily (H) – based on structure/function evolutionary commonalities
- provides useful geometric information (e.g. architecture)
- partial automation may result in examples near fixed thresholds being assigned inaccurately

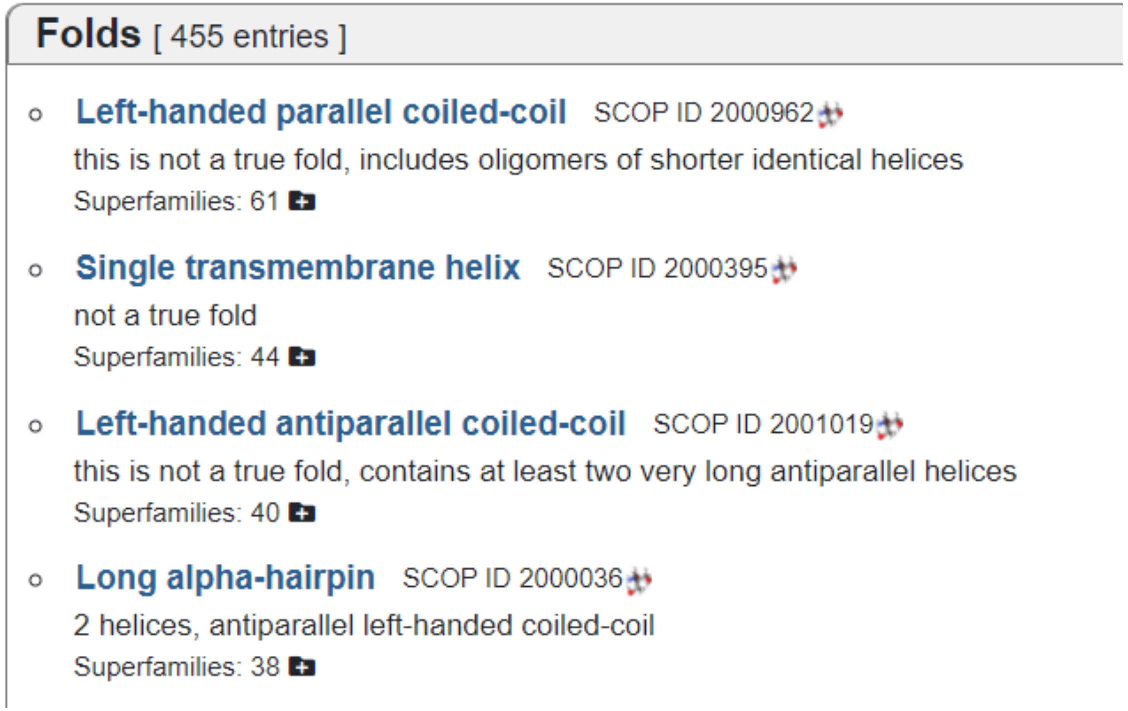
SCOP

https://scop.mrc-lmb.cam.ac.uk/



Browse by structural class

- All alpha proteins
- All beta proteins
- Alpha and beta proteins(a/b)
- Alpha and beta proteins(a+b)
- Small proteins



Folds [455 entries]

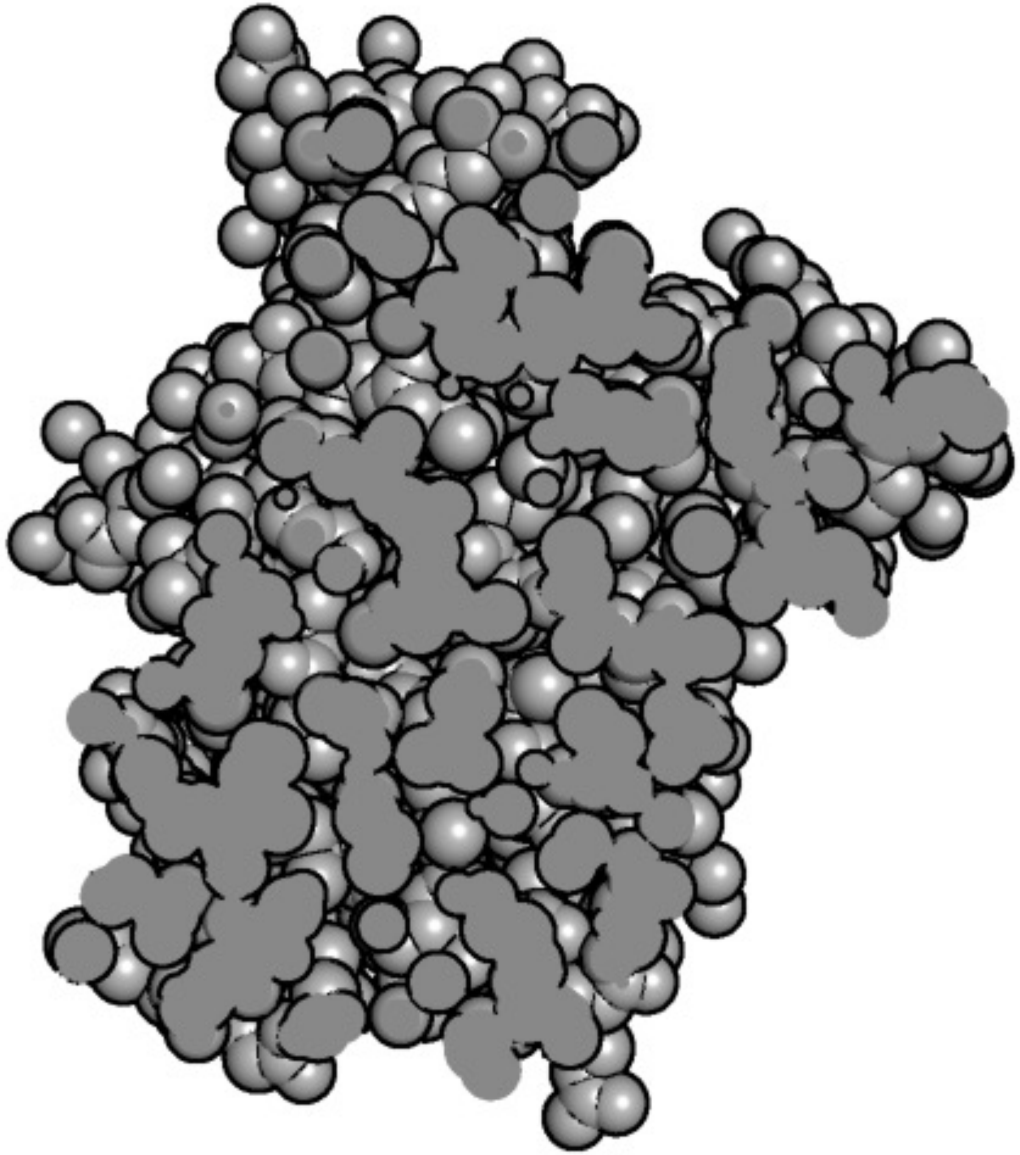
- Left-handed parallel coiled-coil SCOP ID 2000962
  this is not a true fold, includes oligomers of shorter identical helices
  Superfamilies: 61
- Single transmembrane helix SCOP ID 2000395
  not a true fold
  Superfamilies: 44
- Left-handed antiparallel coiled-coil SCOP ID 2001019
  this is not a true fold, contains at least two very long antiparallel helices
  Superfamilies: 40
- Long alpha-hairpin SCOP ID 2000036
  2 helices, antiparallel left-handed coiled-coil
  Superfamilies: 38

- a purely manual hierarchical classification
- Six levels:
  - Class (CL)
  - Fold (CF)
  - Superfamily (SF)
  - Family (FA)
  - Protein (PR)
  - Protein species (SP)
- provides detailed evolutionary information
- manual process influences update frequency and equally exhaustive examination

From Structure to Function

- Proteins are <u>not static</u>
  - Conformational change is critical in performing function
  - Intrinsically disordered proteins transition between ordered and disordered as part of their function
- Proteins are <u>modular</u>
  - Many proteins are comprised of independent folding domains
  - Many proteins function as multi-subunit complexes
- Some proteins require other cofactors/metals to function

Atoms are closely packed in the interior of a protein

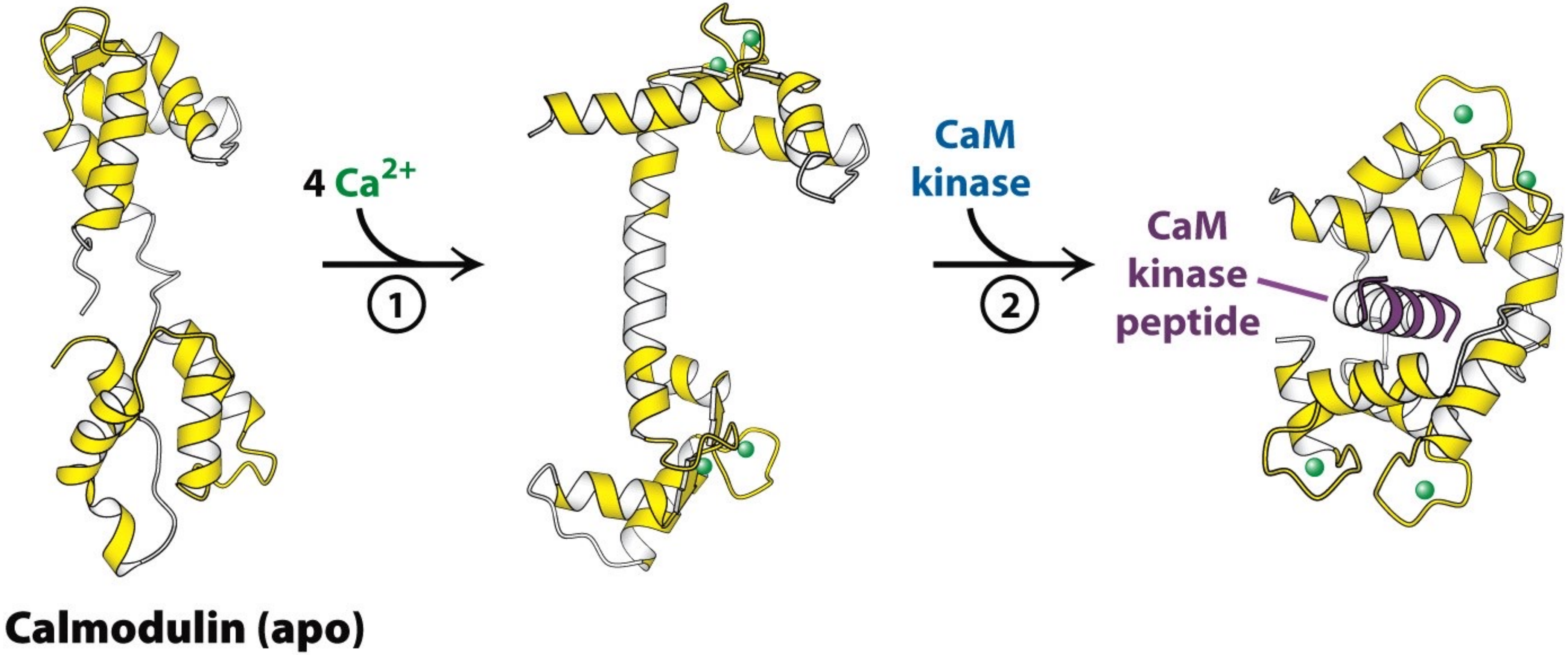


Proteins are usually packed as tightly as organic crystals

However, there are two types of motion which are critical:

1. Thermal motion around equilibrium positions of all protein atoms;
2. Functional motions ("conformational change") in response to
   - encounters with other molecules
   - changes in pH

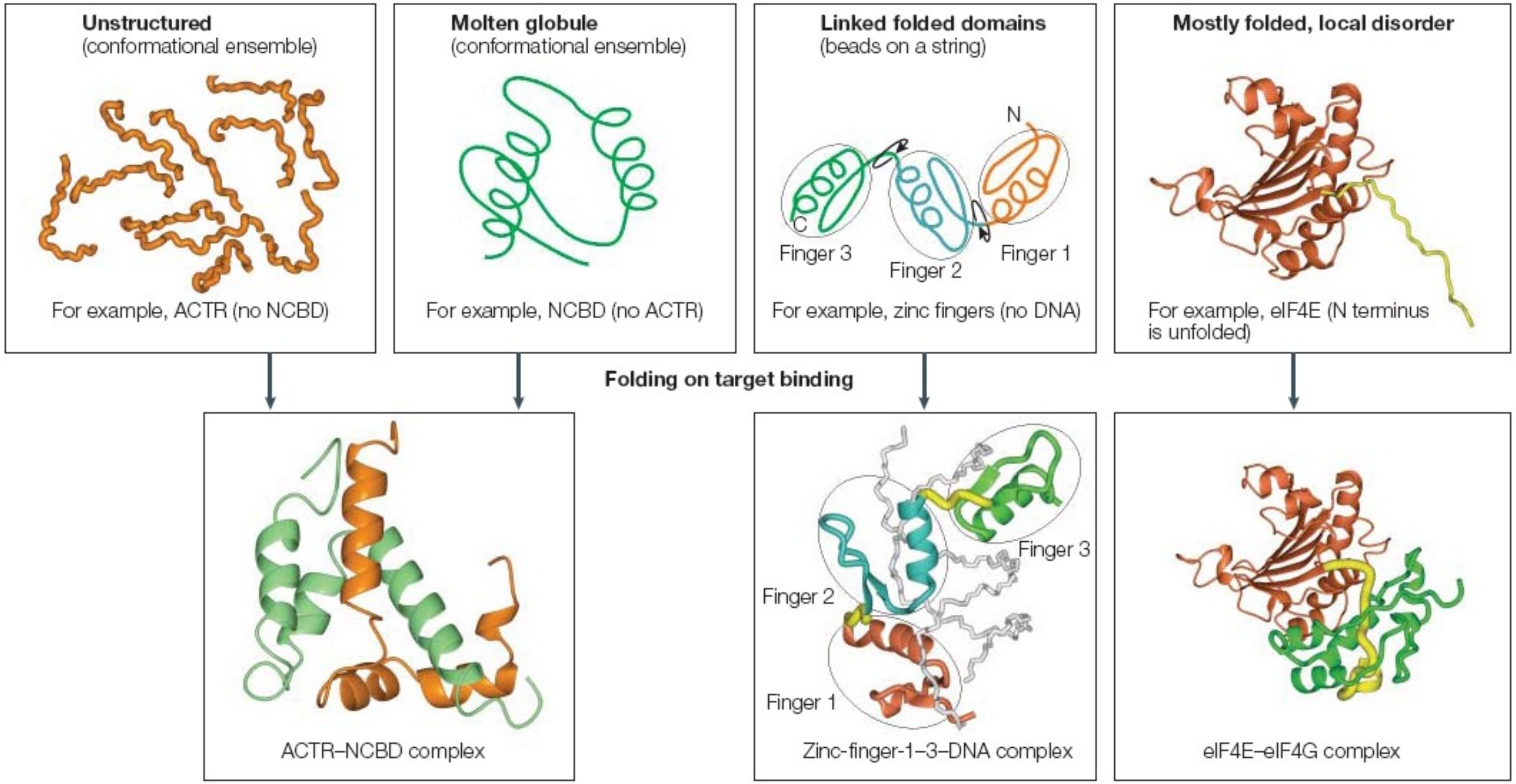
Conformational Change: Calmodulin



Protein structure is important.
Yet, without functional conformational changes of proteins,
life would be pretty miserable.

pdb ids: 1DMO, 3CLN, 1IQ5

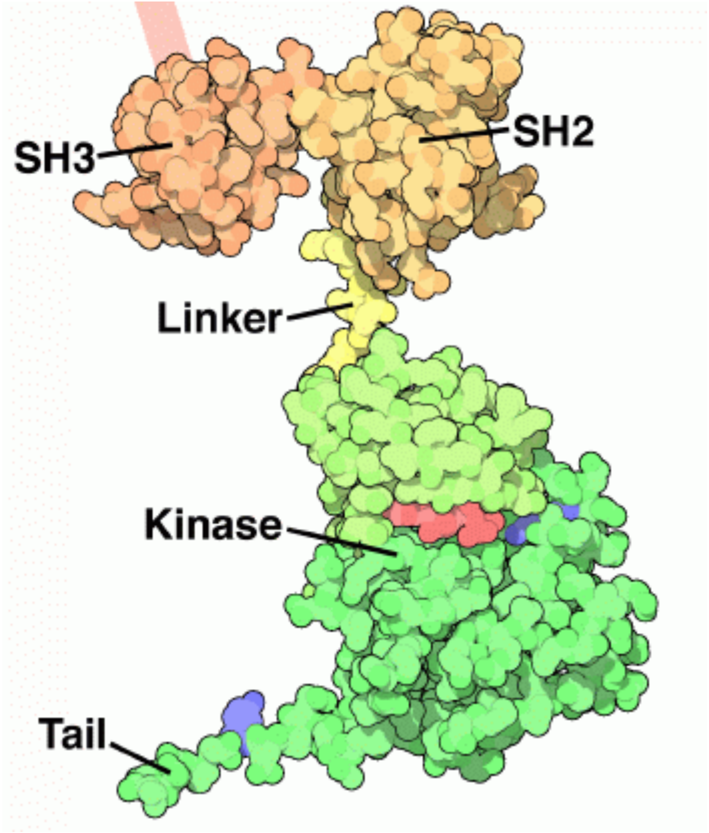
Many Intrinsically Unfolded Proteins Adopt Structure Upon Binding Partner Molecules


Unstructured
(conformational ensemble)
For example, ACTR (no NCBD)
Molten globule
(conformational ensemble)
For example, NCBD (no ACTR)
Linked folded domains
(beads on a string)
N
C
Finger 3
Finger 2
Finger 1
For example, zinc fingers (no DNA)
Mostly folded, local disorder
For example, eIF4E (N terminus
is unfolded)
Folding on target binding
ACTR–NCBD complex
Finger 3
Finger 2
Finger 1
Zinc-finger-1–3–DNA complex
eIF4E–eIF4G complex

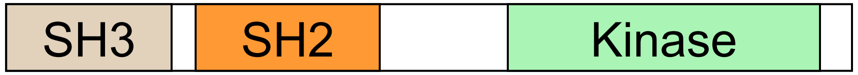
Dyson and Wright (2005) Nat Rev Mol Cell Biol. 6:197-208

Multi-domain proteins

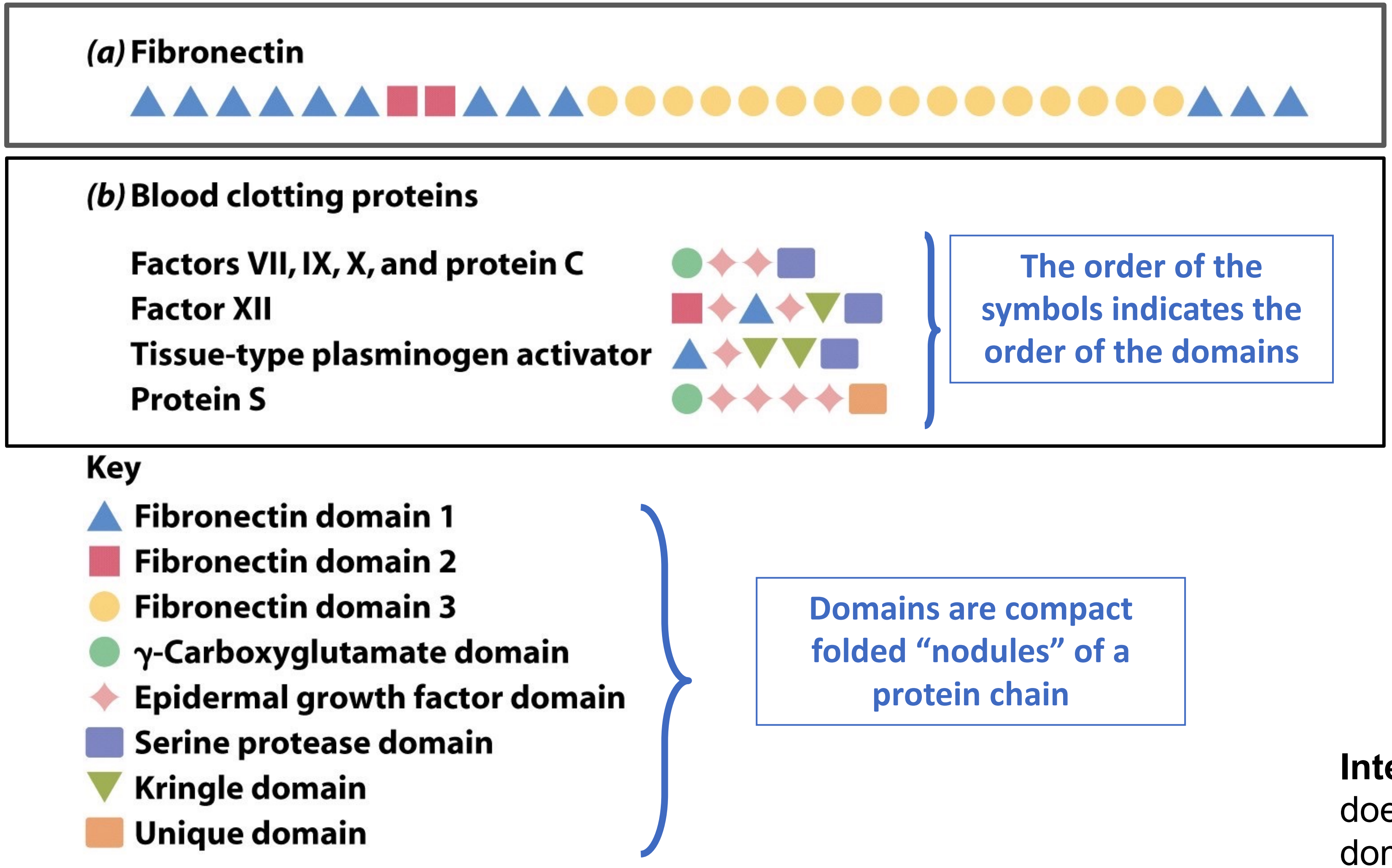
- Many proteins contain ‘independent’ domains connected by linkers. It is common to combine recognition domains with activation domains. By piecing domains together in new ways it is possible to create new functions.



Example: Src tyrosine kinase. The SH3 domain recognizes substrate and the kinase domain phosphorylates the substrate.



Multi-domain proteins are very common


(a) Fibronectin
(b) Blood clotting proteins
Factors VII, IX, X, and protein C
Factor XII
Tissue-type plasminogen activator
Protein S
Key
Fibronectin domain 1
Fibronectin domain 2
Fibronectin domain 3
γ-Carboxyglutamate domain
Epidermal growth factor domain
Serine protease domain
Kringle domain
Unique domain

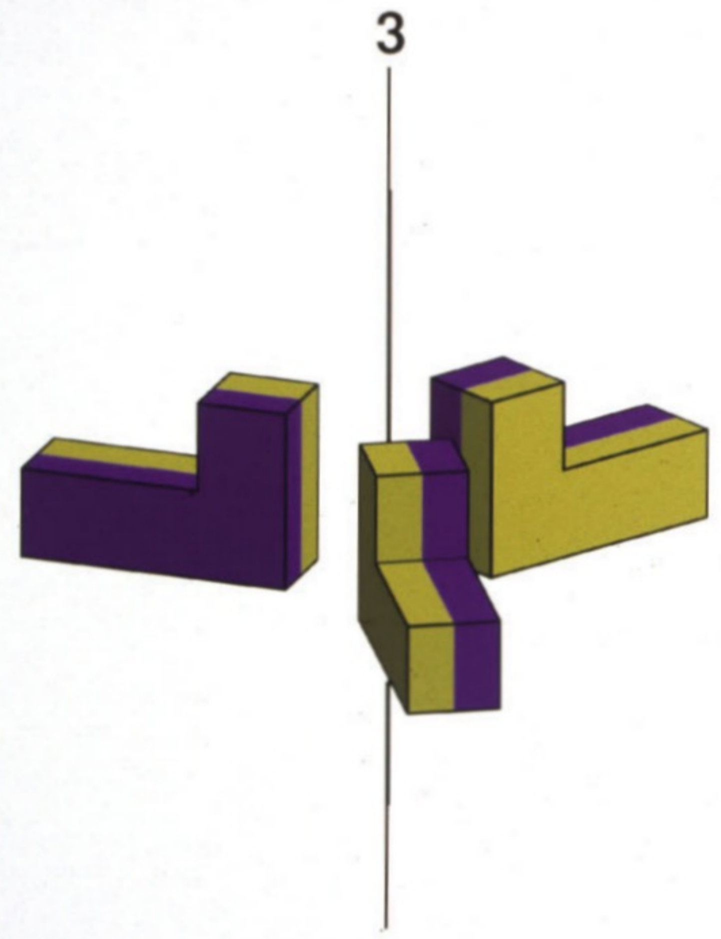
The order of the symbols indicates the order of the domains

Domains are compact folded “nodules” of a protein chain

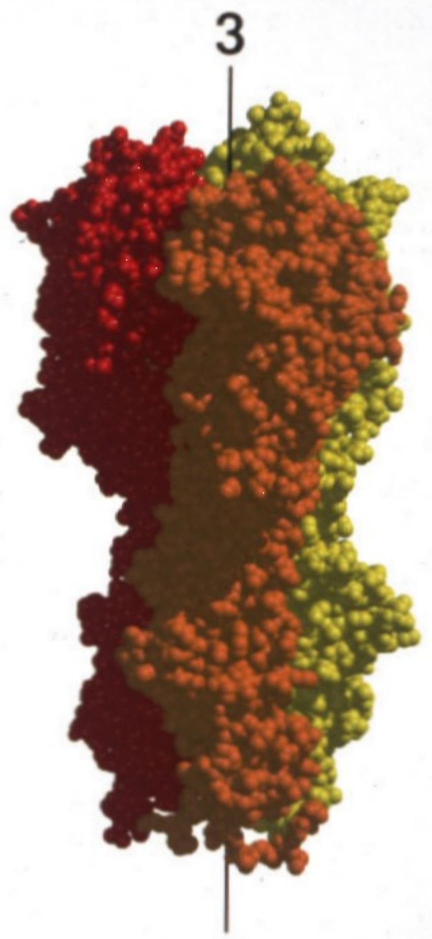
Living organisms often string domains together into one protein chain and then modify each domain for a specific function

Interesting fact: the human genome does not contain more types of protein domains than more primitive organisms, but rather just puts them together in more complicated ways.

A Trimer with cyclic C_3 Point Group Symmetry



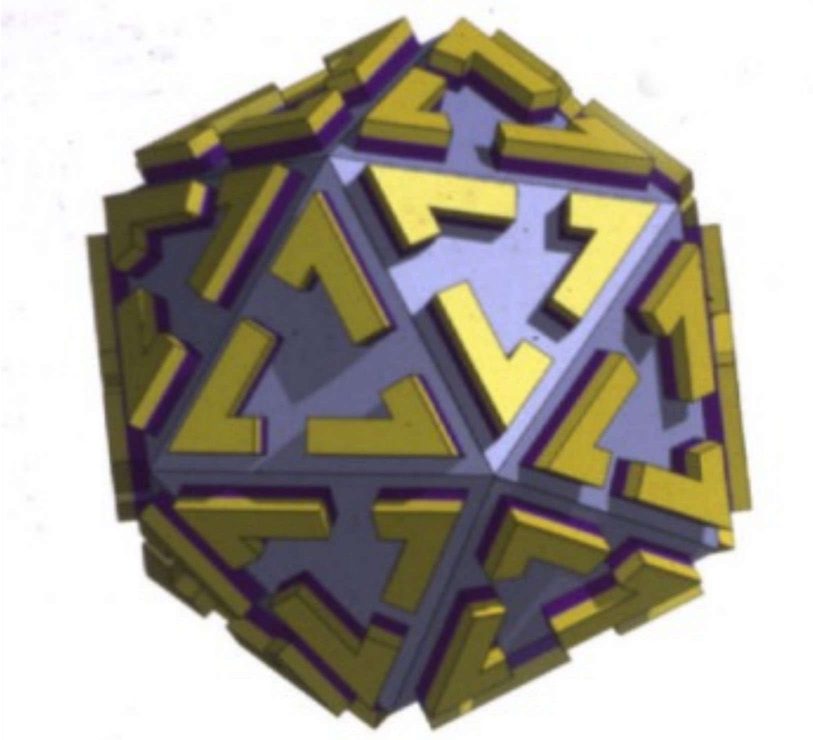
Schematic



Haemagglutinin

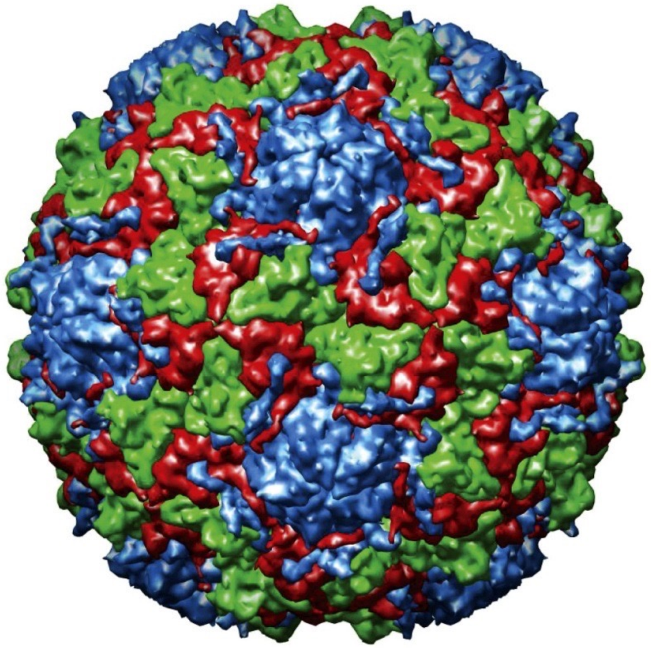
From the surface of the influenza virus

Some viruses have icosahedral symmetry



Icosahedral symmetry generates 60 equivalent objects out of ONE object.

There are 20 triangles per icosahedron, so from the figure above it is quite easy to calculate that there are 60 golden objects with the shape of a "1" per icosahedron

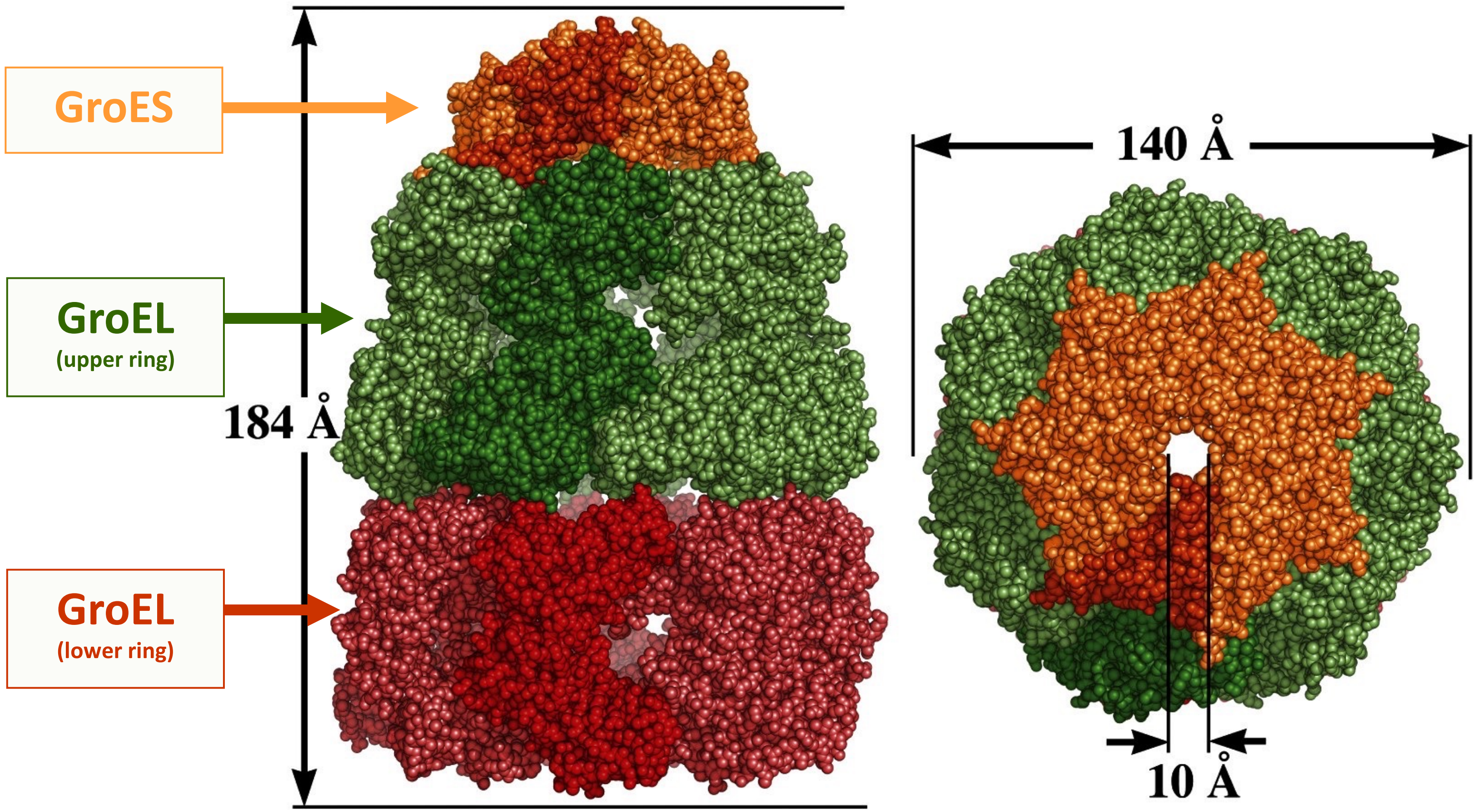


Spherical viruses with icosahedral symmetry have often **N×60** equivalent protein subunits in the capsid surrounding the RNA or DNA in a virus particle (where **N** is an integer).

The virus above has
3 × 60 = 180 proteins in its "capsid".

Inside the capsid above is the viral RNA (Poliovirus looks like the virus above).

The GroEL/GroES chaperone: Outside Architecture

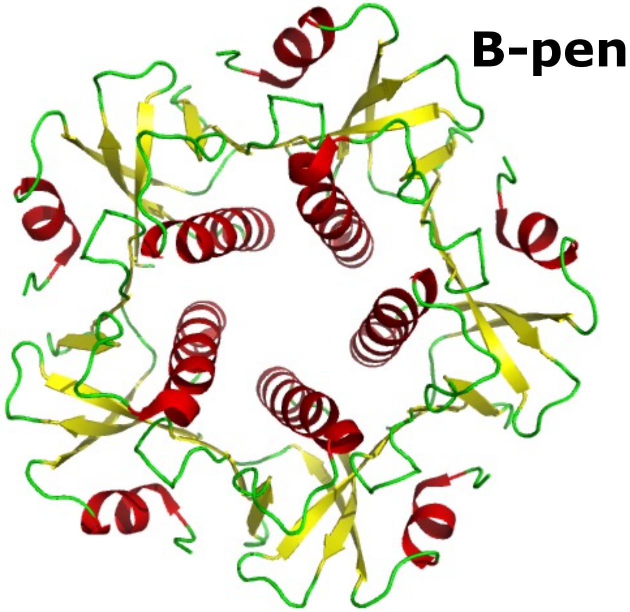


The $(GroES)_7$-$(GroEL)_{14}$-$(ADP)_7$ complex.

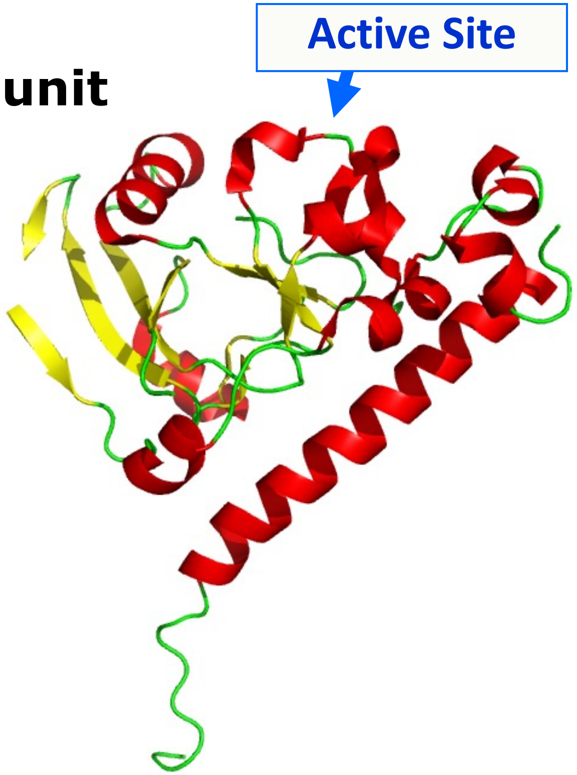
Note different conformations of the two, upper and lower, GroEL rings.
The GroES ring and the two GroEL rings have all 7-fold C7 symmetry.

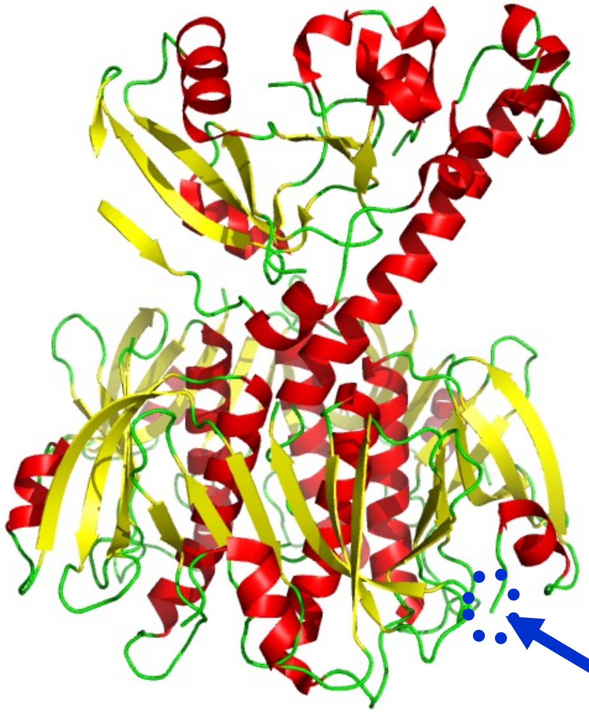
Assembly of the AB_5 holotoxin

Cholera Toxin & Enterotoxin



B-pentamer + **A-subunit**





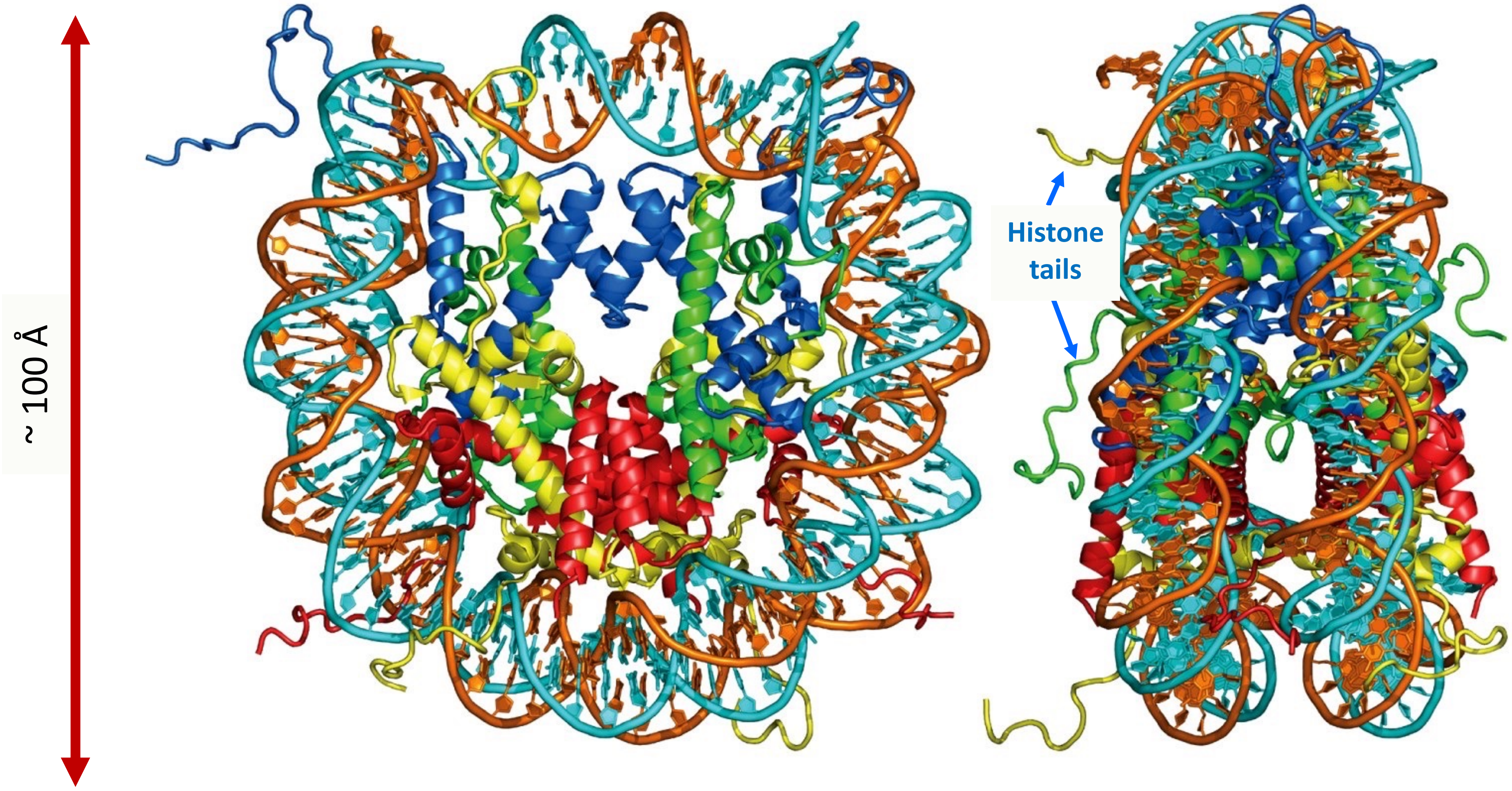
One of five Receptor Binding Sites

AB_5
Holotoxin

Functions:

- The B-subunit binds to human cell surface receptors.
- The A-subunit modifies a key human protein inside the cell.

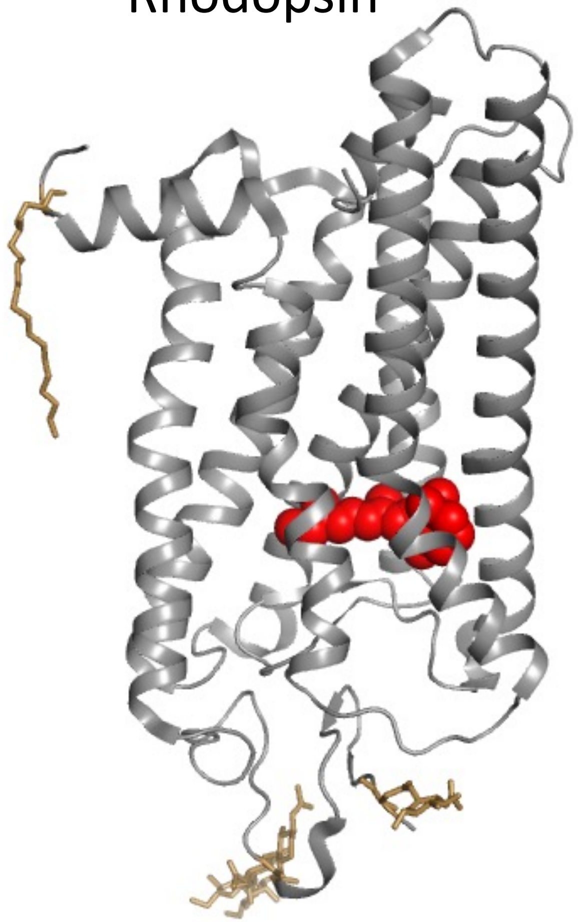
The Nucleosome: a protein + DNA assembly



- Nucleosomes are the building blocks of chromosomes.
- In the centre of the nucleosome there are eight (2x4) proteins called “histones”.
- A double stranded DNA helix (~146 base pairs) wraps around this histone core.
- The histones are shown as ”ribbons” in the centre of the nucleosome

Many proteins feature co-factors

Rhodopsin



The protein of “vision”

A “membrane protein”

Note schematic representations of α-helices

The molecule in red is “retinal”

Brown: “posttranslational modifications”

Myoglobin

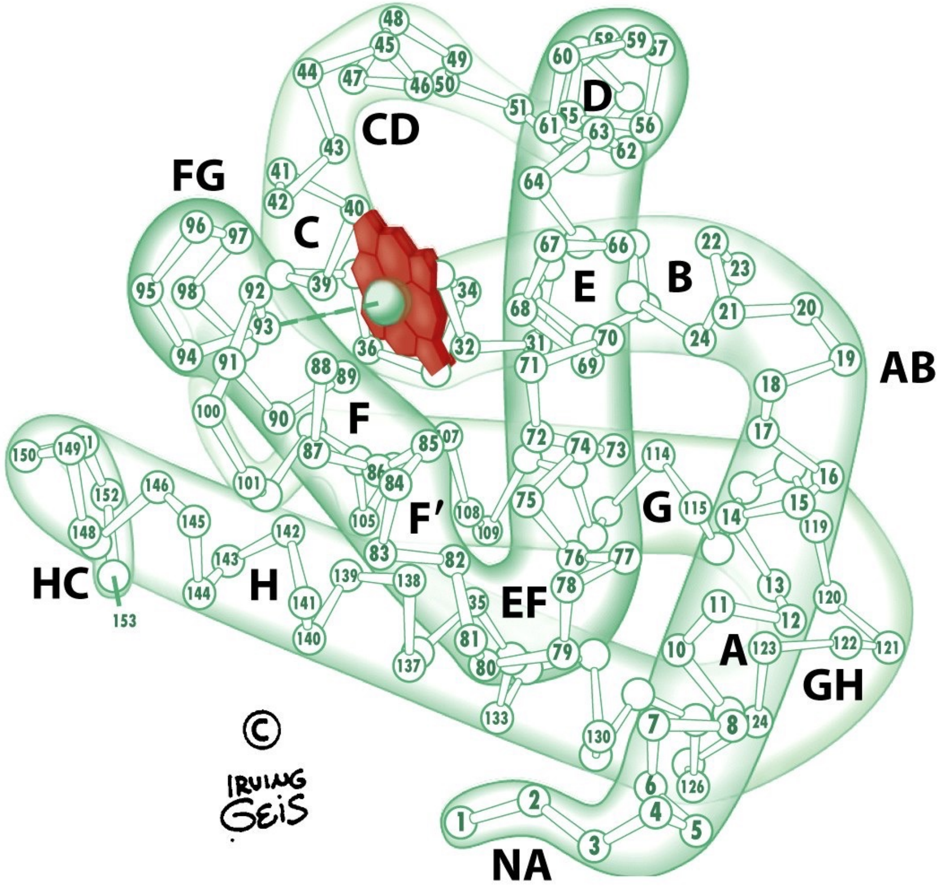


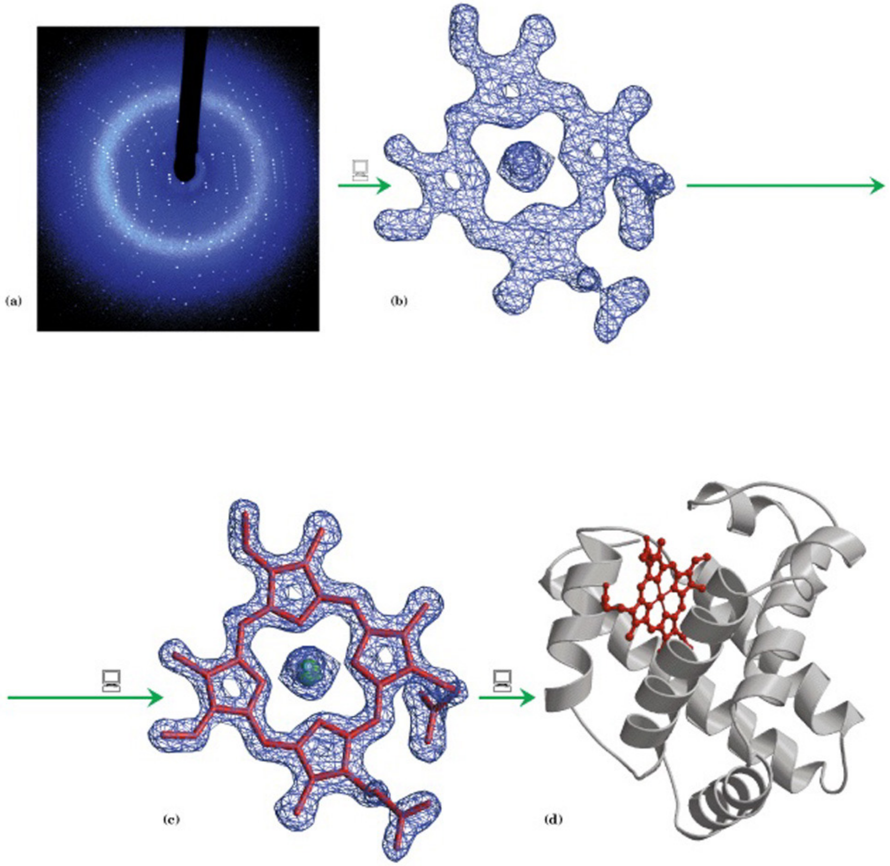
Figure 7-1 Fundamentals of Biochemistry, 2/e

Heme group in red with spherical Fe(II) ion in center.

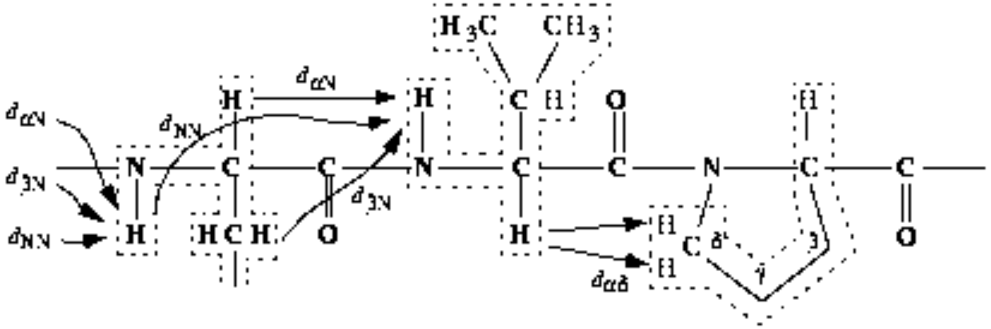
- The eight helices are labeled A to H.
- Helix-connecting loops are AB, BC, etc

X-Ray Crystallography

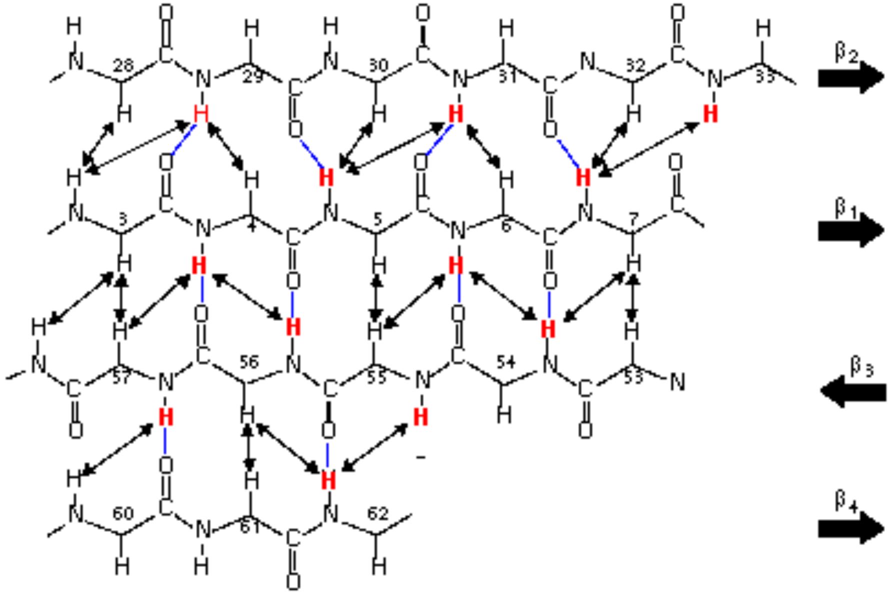
- crystallize and immobilize single, perfect protein
- bombard with X-rays, record scattering diffraction patterns
- determine electron density map from scattering and phase via Fourier transform:
- use electron density and biochemical knowledge of the protein to refine and determine a model



NMR Spectroscopy



determining constraints



using constraints to determine secondary structure

- protein in aqueous solution, motile and tumbles/vibrates with thermal motion
- NMR detects chemical shifts of atomic nuclei with non-zero spin, shifts due to electronic environment nearby
- determine distances between specific pairs of atoms based on shifts, “constraints”
- use constraints and biochemical knowledge of the protein to determine an ensemble of models

Cryo-electron microscopy