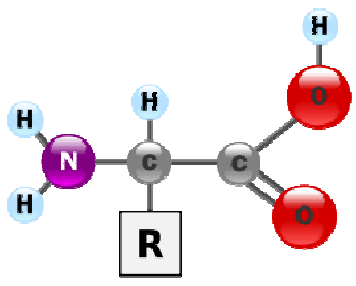
Basics on Protein Structure

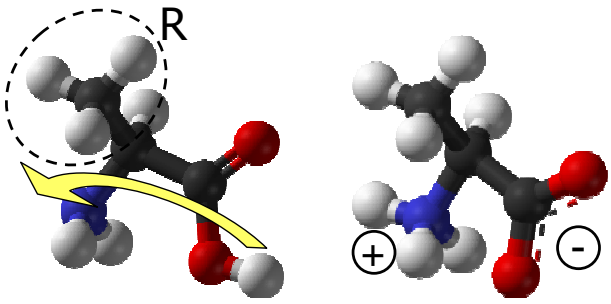
Building Blocks: Amino Acids

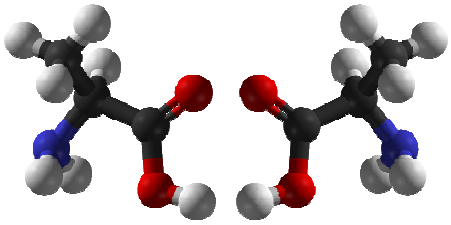
- Proteins/polypeptides are composed of **amino acids**, linked in chains
- Each amino acid consists of a central carbon (known as C^{α}) attached to four units:
  - H - hydrogen
  - NH_2 – amino group
  - COOH – carboxyl group
  - a distinguishing side-chain (known as *R* group)



More on Amino Acids...

- Actually, in usual conditions (PH≈7), amino acids show a different charge distribution (in figures, Alanine)





L-isomer: "CORN" rule

D-isomer

- Because of C^{α}, two different mirror images for the molecule are possible: **only the L-isomer is present in proteins** as we know them in Nature

Amino Acids: How Many

- There are 20 amino acids: humans are able to synthesize about a dozen of them
- 8 amino acids must be directly provided by our food: they are named "essential amino acids"
- In most proteins (not all!), the 20 amino acids occur with similar frequencies

The Hydrophobic Effect

- The hydrophobic effect is the ***tendency to form intermolecular aggregates in an aqueous medium***.

- At the macroscopic level, this effect shows up as an apparent repulsion between a given substance and water (e.g. oil and water cannot be mixed)
- The hydrophobic effect plays a crucial role in determining protein structure and function

Amino Acids Table

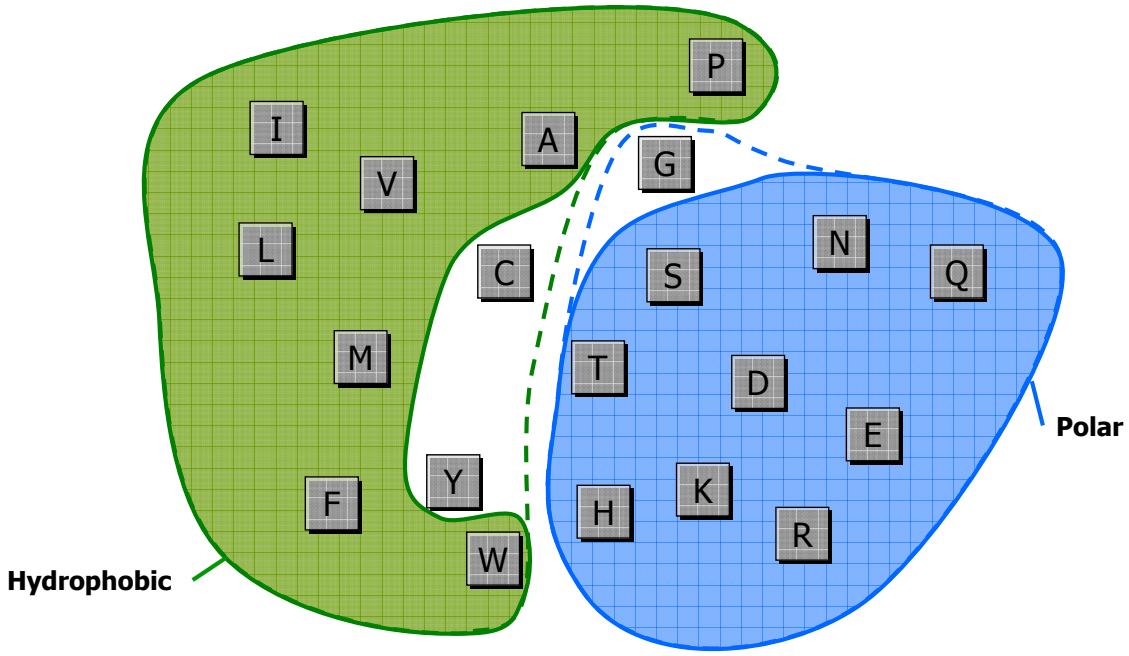
- Here amino acids are listed according to their Kyte-Doolittle index, which reflects hydrophobic/hydrophilic properties

Isoleucine I, Ile 4.5	Valine V, Val 4.2	Leucine L, Leu 3.8	**Phenil-alanine** F, Phe 2.8	Cysteine C, Cys 2.5
Methionine M, Met 1.9	Alanine A, Ala 1.8	Glycine G, Gly -0.4	Threonine T, Thr -0.7	Serine S, Ser -0.8
Tryptophan W, Trp -0.9	Tyrosine Y, Tyr -1.3	Proline P, Pro -1.6	Histidine H, His -3.2	Glutamic Acid E, Glu -3.5
Glutamine Q, Gln -3.5	Aspartic Acid D, Asp -3.5	Asparagine N, Asn -3.5	Lysine K, Lys -3.9	Arginine R, Arg -4.5

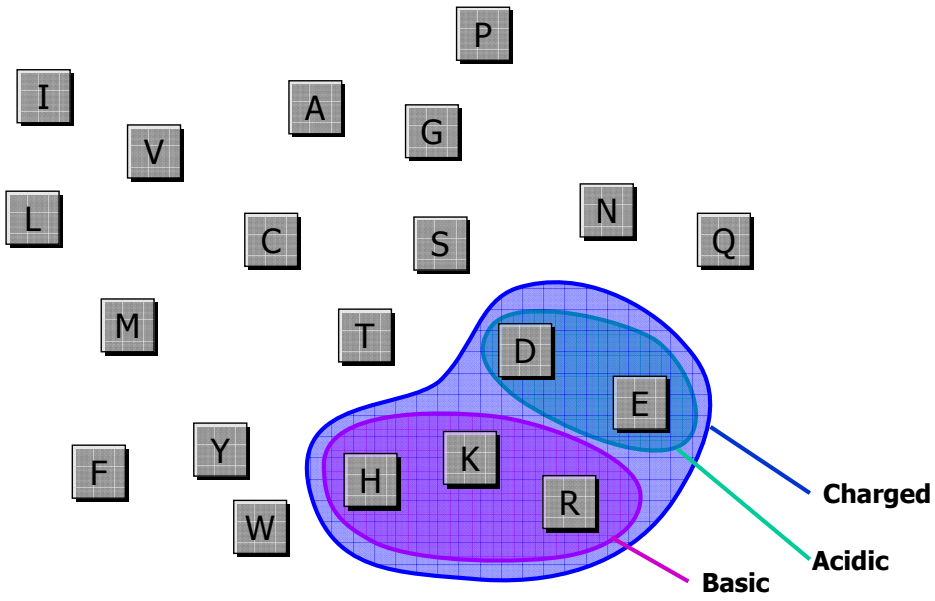
Classifying Amino Acids

- Amino acids can me classified according to different properties, due to the specific side chain:
  - Polarity (H -nonpolar, or hydrophobic
    vs. P - polar, or hydrophilic)
  - Aliphatic/aromatic R
  - Acidic/basic R
  - Size
  - Etc.
- Many different classifications are present in literature; in the following slides, classification roughly refers to what assumed in the *VMD* software
- **See, e.g. http://www.mcb.ucdavis.edu/courses/bis102/AAProp.html**

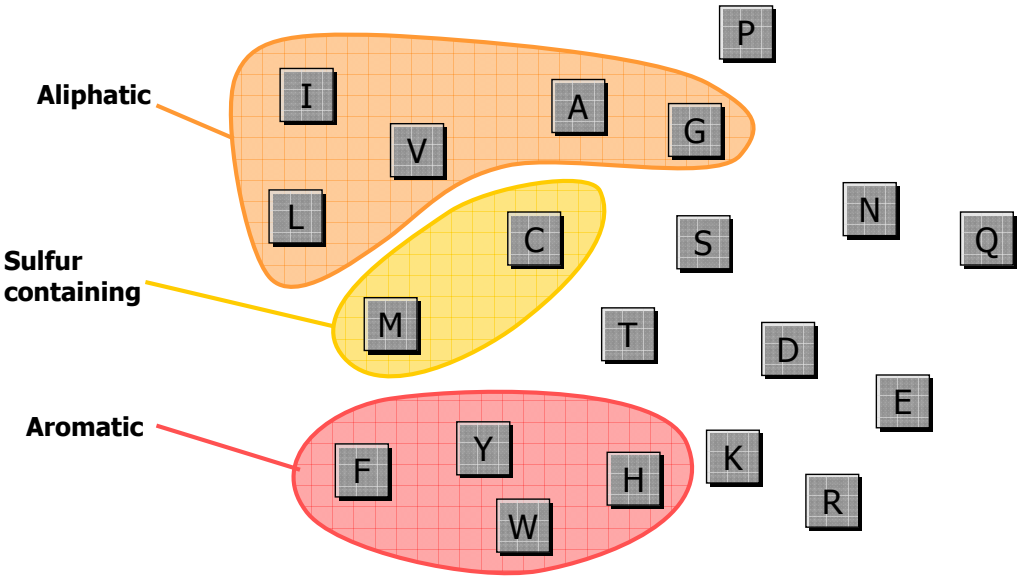
A.A. Classification (I)



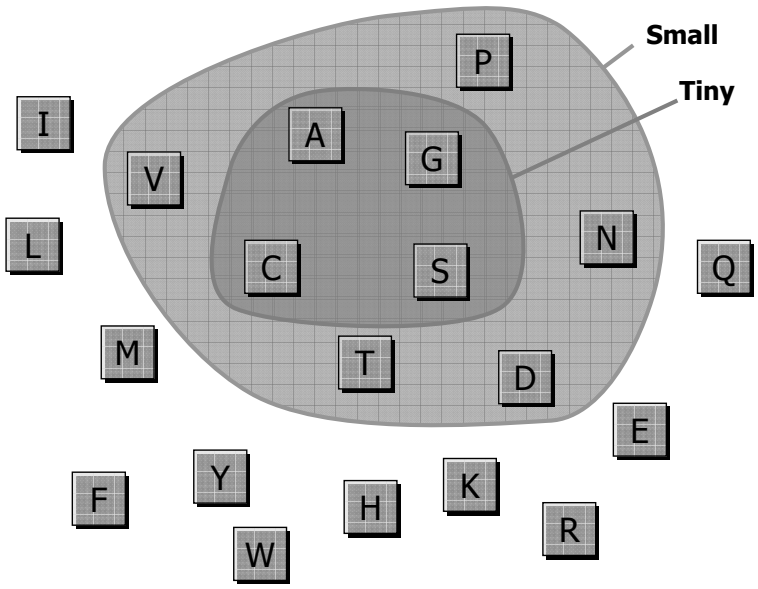
A.A. Classification (II)



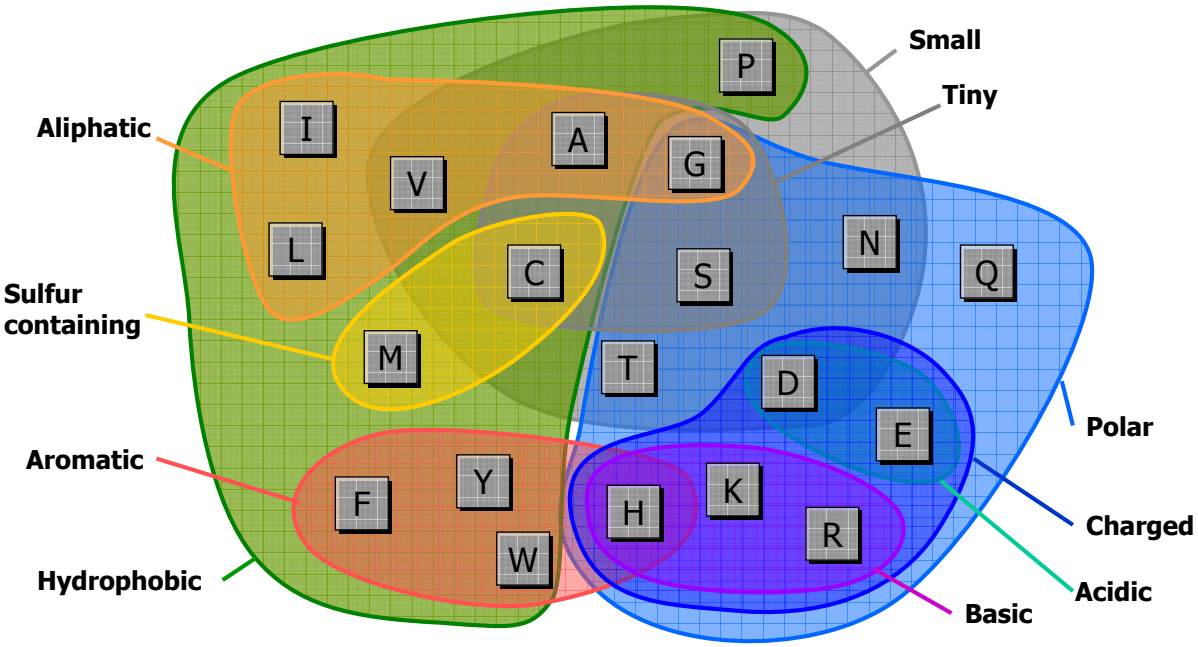
A.A. Classification (III)



A.A. Classification (IV)

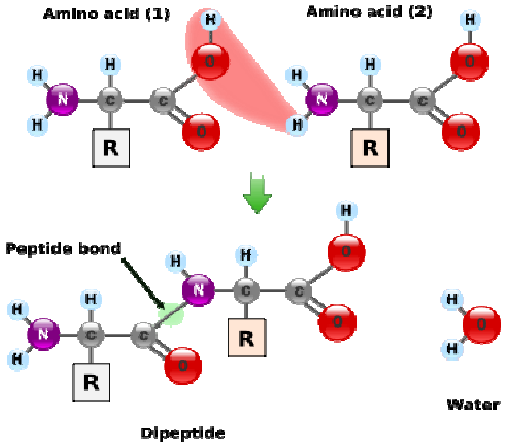


A.A. Classification Overview



Linking Amino Acids (I)

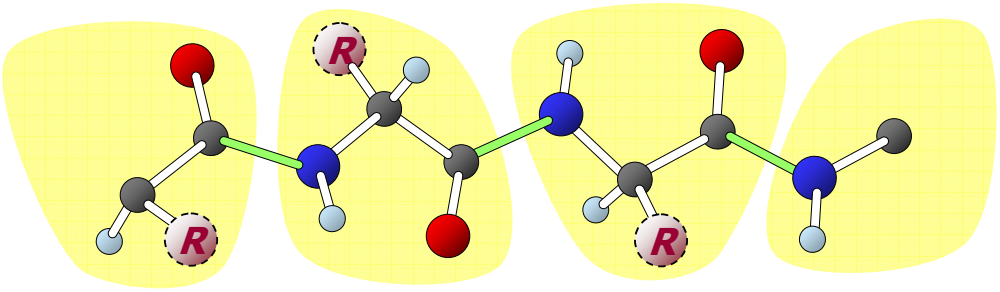
- Joining amino acids, we obtain a *polypeptide*
- The carboxyl C of one amino acid joins the amino nitrogen of another amino acid, forming the (rigid) **peptide bond** C-N and releasing one water molecule



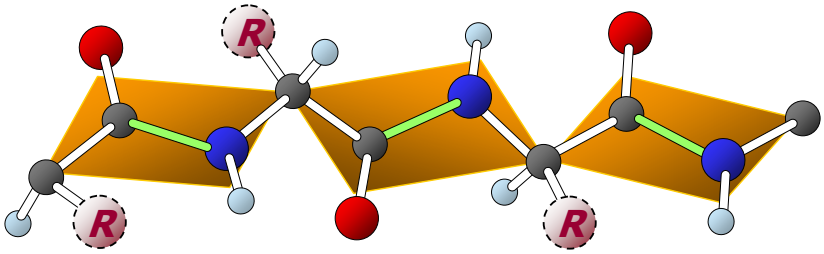
Linking Amino Acids (II)

- A peptide chain thus has a N-terminus and a C-terminus (aka as Amino/Carboxy termini); the standard orientation is from N- to C-
- The peptide chain can be viewed as a **backbone** (or main chain), connecting Alpha Carbons, and a number of **side chains**, one for each R group.
- Sometimes, other covalent bonds may occur between Sulfur atoms in different residues.
- Plenty of additional non-covalent bonds contributes to the stabilization of the whole molecule.

Pointing Out Chain Units



Units in the chain can be associated to the composing amino acids (***residues***)



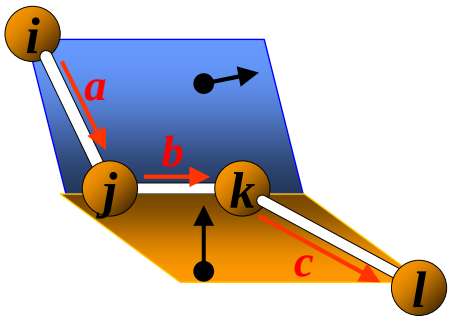
Otherwise: rigid planar structures around the peptide bond (***rigid peptide units***)

Primary Structure

- The simplest description of a protein is given by the sequence (from N- to C- terminus) of the composing residues: such a list is known as **Primary Structure**
- The primary structure can be related (often, but not always, in a straightforward way) to the DNA sequence in the corresponding gene
- Example: primary structure of *bovine plasma retinol-binding protein* (1HBQ)

ERDCRVSSFRVKENFDKARFAGTWYAMAKKDPEGLFLQDNIVAEFSVDENGQM
SATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWI
IDTDYETFAVQYSCRLLNLDGTCADSYSFVFARDPSGFSPEVQKIVRQRQEEL
CLARQYRLIPHNGYCNGKSERNIL

Dihedral Angles



$$\cos(\tau) = \cos(n_{ab} \bullet n_{bc})$$

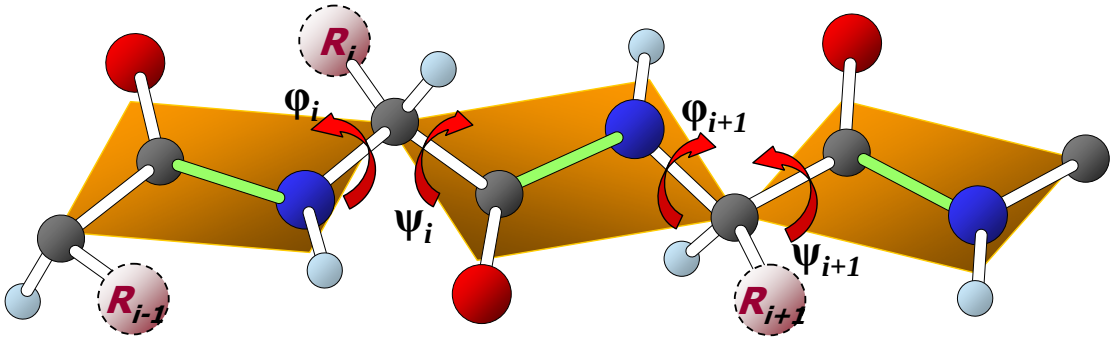
$$\mathrm{sgn}(\tau) = \mathrm{sgn}(a \wedge b \bullet c)$$

- To specify 3D arrangements of atoms, *dihedral angles* are often used
- Let's consider a sequence of four bonded atoms i, j, k, l in space: ijk define a plane, jkl another.
- The dihedral angle τ is the angle between the vectors normal to the two planes.
- If τ=0, all the atoms are co-planar, and i al l are "on the same side" (*cis*)
- If τ=π, all the atoms are co-planar, and i al l are "on opposite sides" (*trans*)

Relative Position of Peptide Units

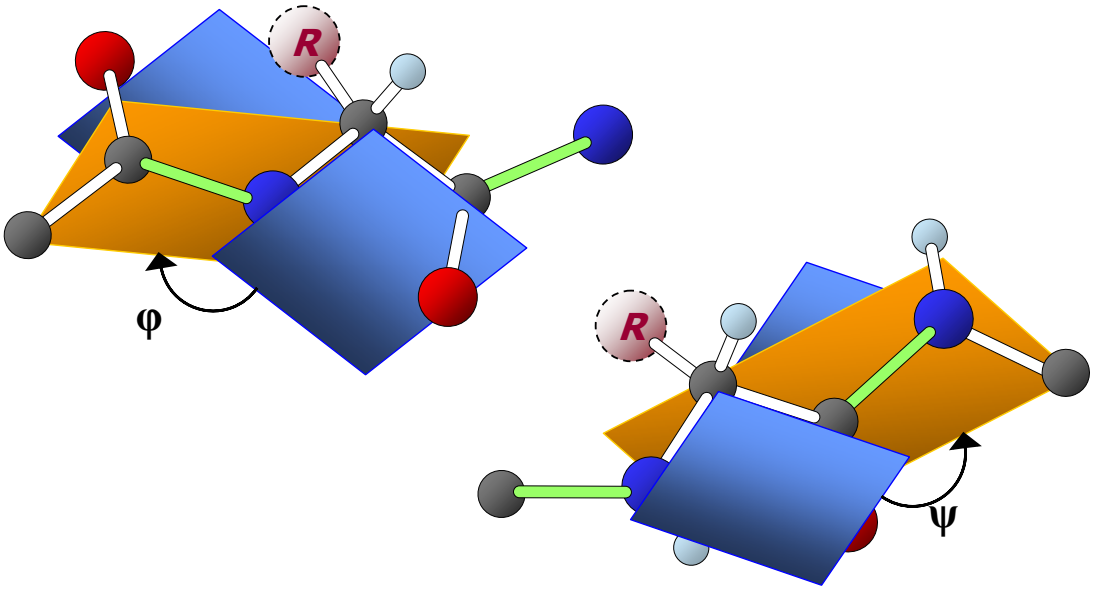
Each peptide unit can rotate around the two covalent bonds at their C^{α} ends.

- The dihedral angle around the N-C^{α} bond is named **phi** (φ) (C'_{i-1}-N_i-C^{α}_i–C'_i)
- The dihedral angle around the C^{α}-C′ bond is named **psi** (ψ) (N_i-C^{α}_i–C'_i-N_{i+1})
- The dihedral angle ω around the peptide bond is constant, usually $\approx\pi$ (trans)



The 3D displacement of the backbone is fully determined by the values φ and ψ at each C^{α}.

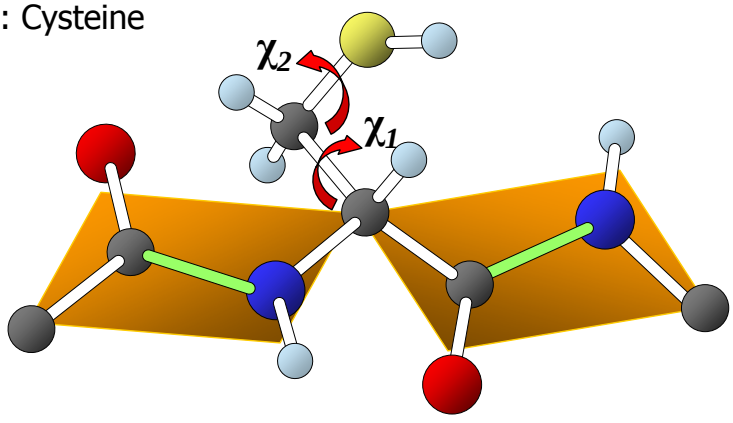
Phi and Psi Angles



Relative Positions in Side Chains

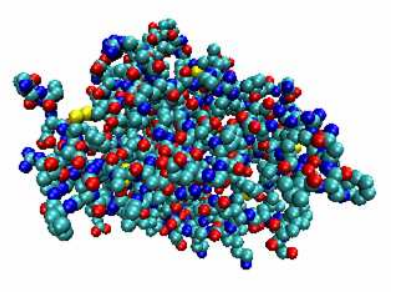
- The exact position of carbons in side-chains can be described by a sequence of angles (starting from C^{α}), known as χ_1, χ_2, χ_3, etc.
- *Rotameric structures* of a protein are those with the same $\{\varphi, \psi\}$ but different side-chain conformations (i.e. different χs)

In the example: Cysteine



Native & Denaturated States

- Proteins chains in ordinary conditions fold into low-energy, aggregated conformations known as ***native states***.
- Typical biological functions of proteins are carried out in native states.
- An "unfolded" protein, yet keeping its primary structure, is said to be in a ***denaturated state***.
- Denaturation of a protein occurs by different kinds of external stress and/or heat

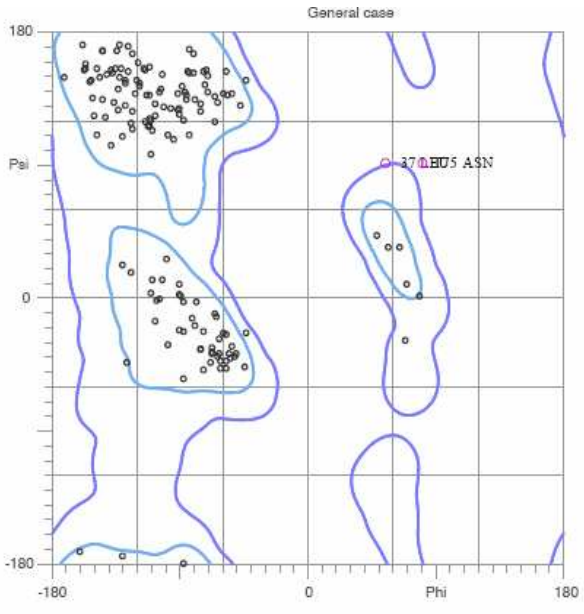


In fig, a representation of the native state of 1HBQ

Ramachandran Plot

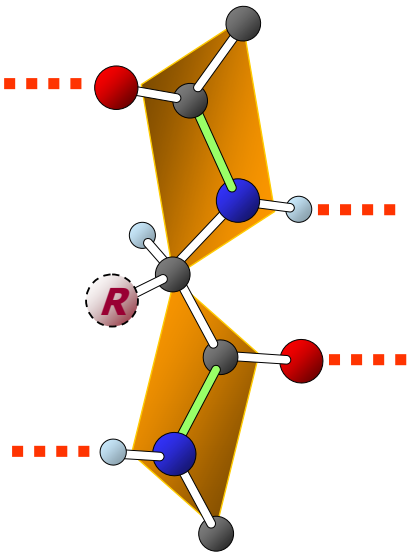
- In native state,

  (φ, ψ) pairs do not take any possible values: their distribution can be pointed out in the *Ramachandran Plot*
- Because of steric factors, some areas in the plot cannot be populated
- In fig: (φ, ψ) pairs for 1HBQ



Chains Hydrophobicity

- The internal part of a protein almost contains hydrophobic side chains
- (Globular) proteins usually fold packing hydrophobic side chains in the interior of the molecule, yielding a **hydrophobic core** and leaving a hydrophilic surface
- Conversely, the main chain is highly hydrophilic, with NH as H-bond donor and C'=O as H-bond acceptor



Protein Structure is Hierarchical

Primary (sequence: what we have seen so far)

Secondary (*local* folding)

Tertiary (long-range, *global* folding)

Quaternary (multimeric assemblies)

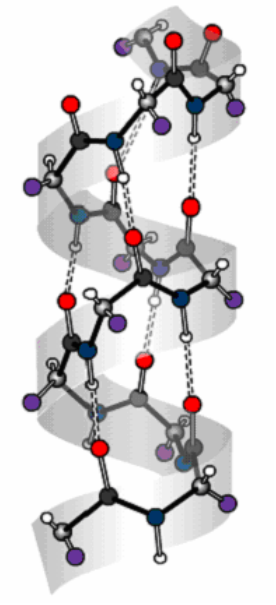
Supramolecular (large-scale assemblies)

Secondary Structure

- Residues in close positions (not necessarily adjacent) can be arranged in a few typical conformations:
  - **Alpha helices**
  - **Beta sheets**
  - (Loops)
- Such regular elements are stabilized by H bonds between N and C′=O groups in the main chain, with close values for (φ, ψ) on consecutive residues
- Secondary elements build up a stable, low flexible framework for the molecule

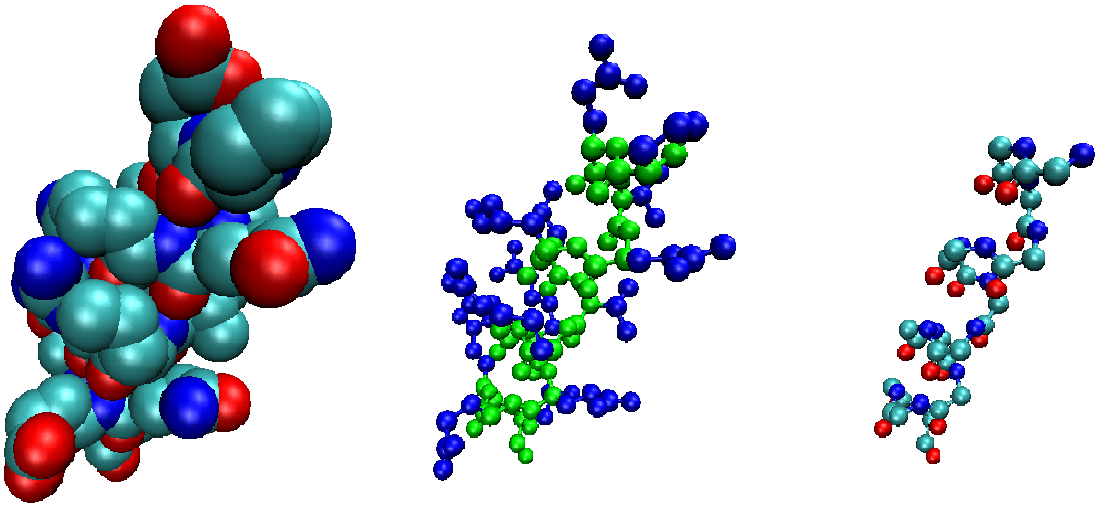
Alpha Helices

- In an α helix, the C′=O group of residue *n* is joined by an H bond to the NH of residue *n*+4
- All C′=O and NH are joined, except the NH of the first residue and the C′=O of the last residue
- Features:
  - 3.6 residues per turn
  - (φ, ψ) ≈ (-60°, -50°)
  - Length (globular proteins): from 4/5 up to >40 residues; on average, 10
  - Rise per residue: 1.5 Å



Example: α-Helix in 1HBQ

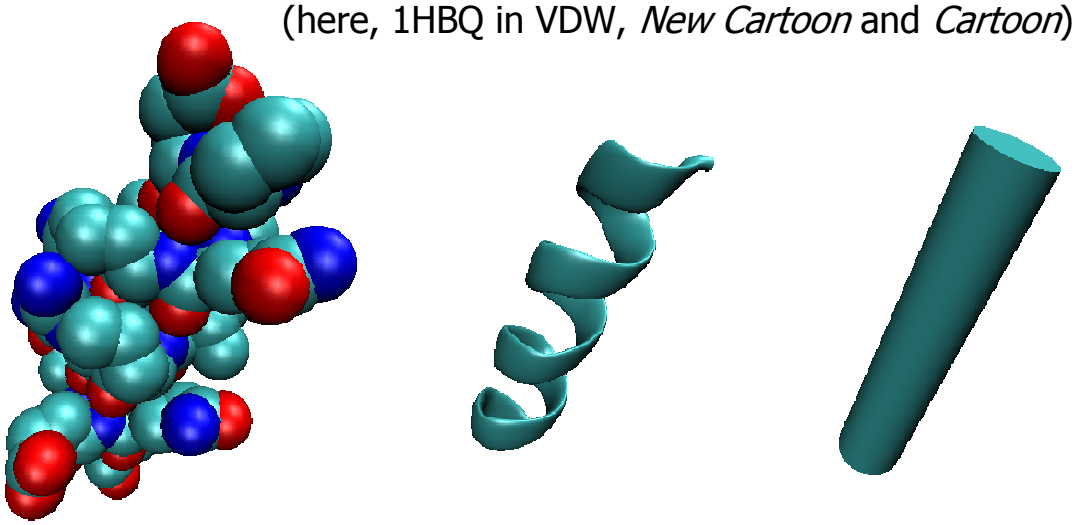
- Shown: VDW, all atoms, only backbone atoms
- Note: all side chains are arranged out of the helix atoms



Representation of α-Helices

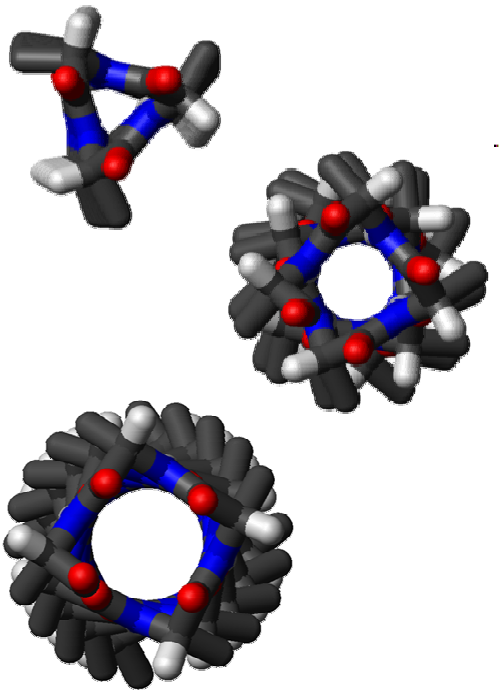
- Different representations are aimed at showing different structural details

(here, 1HBQ in VDW, *New Cartoon* and *Cartoon*)



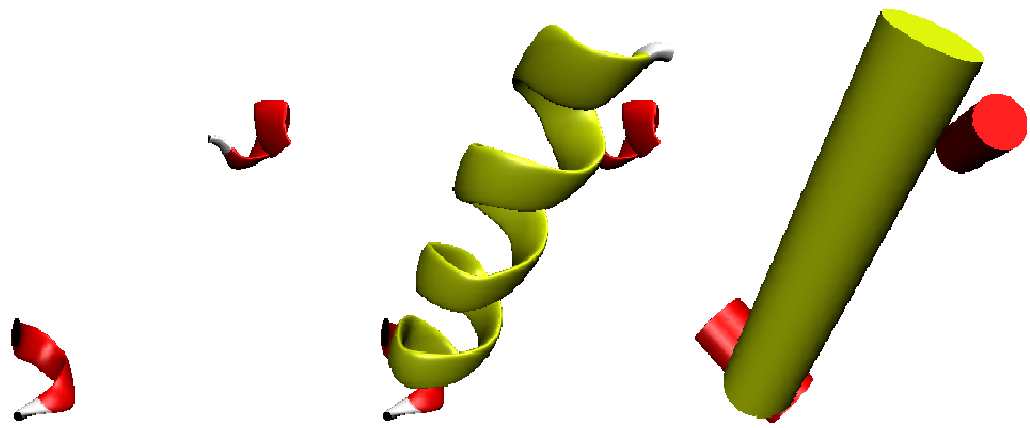
Other (rare) Helices

- Other two types of helices can be (rarely) found, usually as very short chunks
- In a ***3_{10} helix***, C′=O of residue *n* is joined to the NH of residue *n*+3
  - (φ, ψ) ≈ (-49°, -26°)
  - 3 residues per turn
- In a ***π helix***, C′=O of residue *n* is joined to the NH of residue *n*+5
  - (φ, ψ) ≈ (-55°, -70°)
  - 4.1 residues per turn
- Pictures: internal space of 3_{10}, α- and π helices



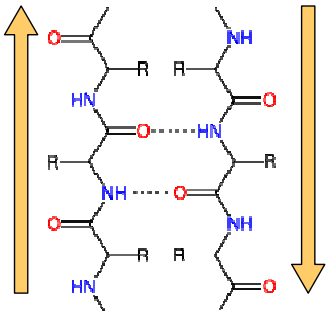
Example: 3_{10}-Helices in 1HBQ

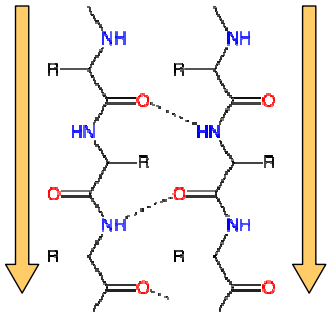
- Shown: 3_{10} helices, 3_{10} helices + α helix



Beta Sheets

- ***Beta sheets*** are made of (2+) stretched portions of the main chain (***beta strands***), placed side by side by H bonds
- Beta sheets usually show a pleated, twisted shape

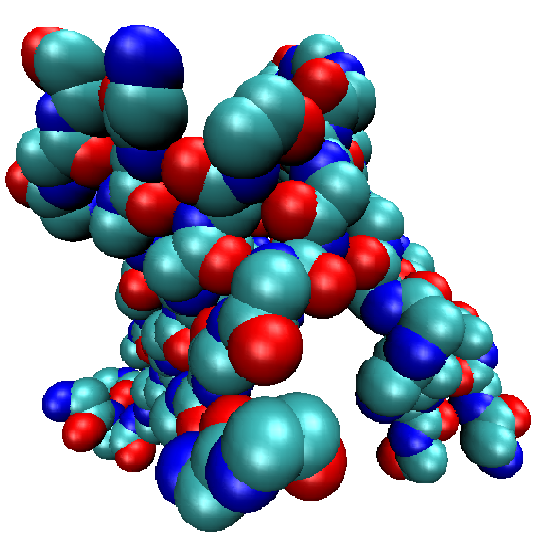


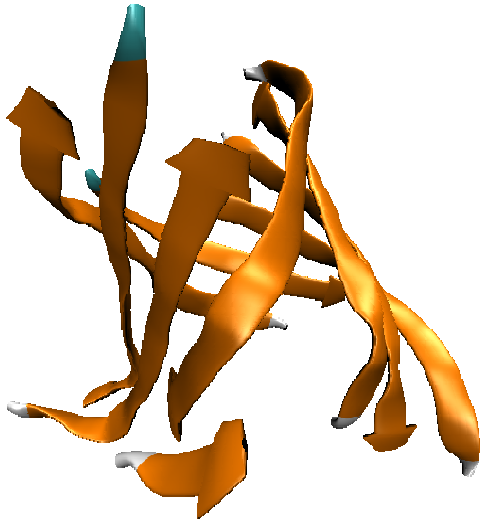


- Two possible geometries: antiparallel $(\varphi, \psi) \approx (-139°, +135°)$
  parallel $(\varphi, \psi) \approx (-119°, +113°)$

Example: β-Sheet in 1HBQ

- Shown: VDW (only backbone atoms), *New Cartoon*





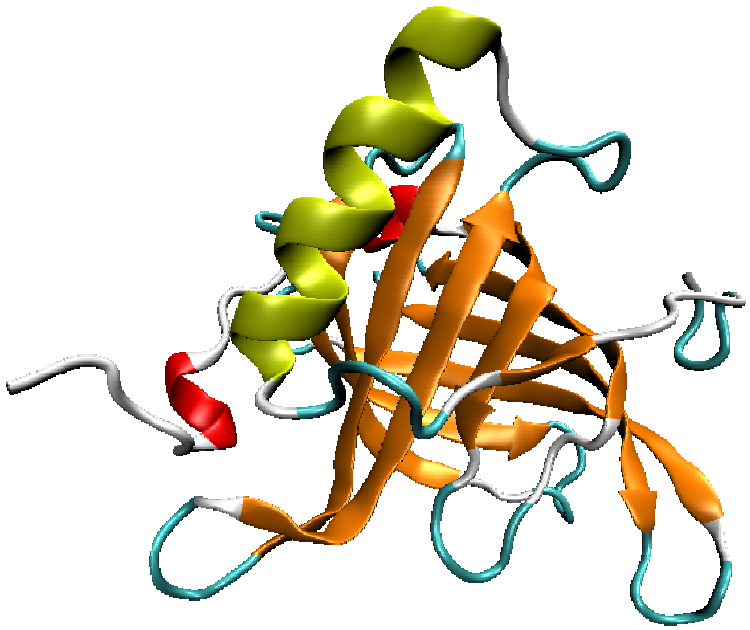
Turns

- Apart helices and β sheets,
  a close approach (< 7 Å) of two C^{α} is known as ***turn***

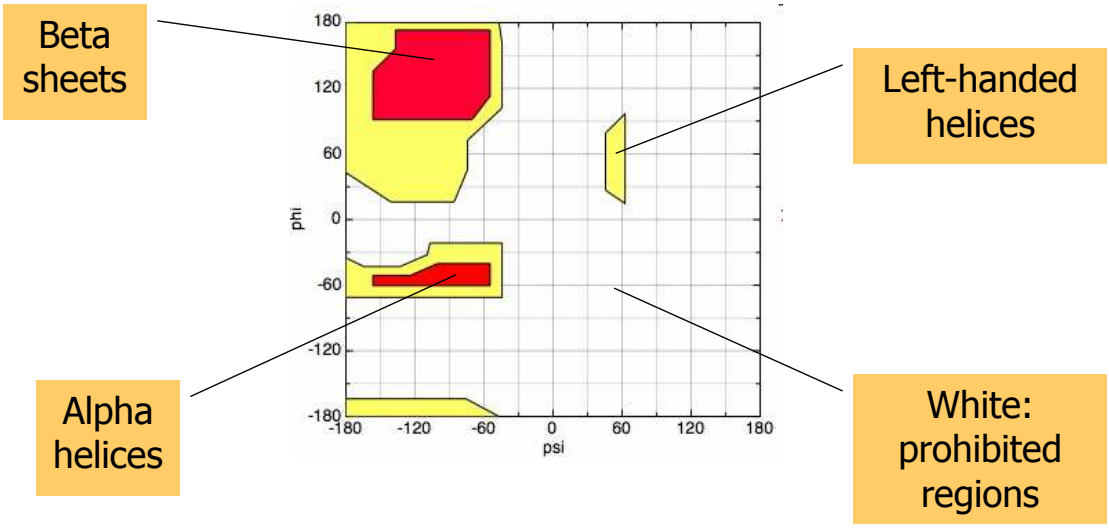
- A turn may be either related to H bonding or not
- Turns are classified according to the number of residues separating the involved C^{α} :
  - β-turn ($C^{\alpha}_{i} \rightarrow C^{\alpha}_{i+3}$) – the most common
  - γ-turn ($C^{\alpha}_{i} \rightarrow C^{\alpha}_{i+2}$) α-turn ($C^{\alpha}_{i} \rightarrow C^{\alpha}_{i+4}$) π-turn ($C^{\alpha}_{i} \rightarrow C^{\alpha}_{i+5}$)

- In case the backbone direction reverses because of a turn, such turn is commonly called ***hairpin***
- Glycine and Proline are usually common in turn regions

Full Example on 1HBQ



2-ry Structs on Ramachandran



DPSS Code & DSSP Algorithm

- DPSS: Dictionary of Protein Secondary Structure
  - **T** (generic turn), **G** (3_{10} helix), **H** (α helix), **I** (π helix)
  - **E** (β sheet), B (single-pair β sheet), **S** (bend), ' ' (space: none of the previous ones; aka **C** - coils / **L** - loops)
- DSSP: Define Secondary Structure of Proteins (**given the atomic coordinates**, each residue is annotated with DPSS codes)

  Steps:

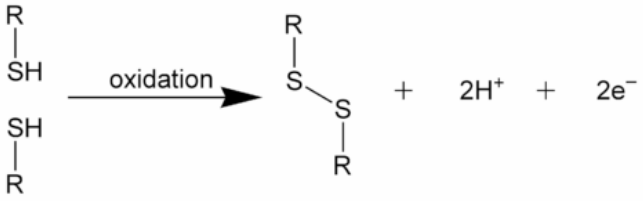
  1. Find H bonds (just *by electrostatic definition*)
  2. Compare hydrogen bonding pattern at each residue with known patterns

Tertiary Structure

- The term "tertiary structure" refers to ***the actual 3D displacement of atoms in a protein***, as found in its native state
- Tertiary structure often (especially in globular proteins) shows a core with packed hydrophobic residues: this contributes to the molecular stability
- 3D architecture of a protein can be *topologically* described by specific arrangements of secondary elements (and super-secondary ones, see ahead), that account for about 90% of the atom content
- Tertiary structure is often made stable also by disulfide bonds between different cysteine residues

Disulfide Bonds

- A disulfide bridge is a covalent bond originated by the oxidation of two -SH groups belonging to different cysteine residues



- A disulfide bond is characterized by its χ_{ss} dihedral angle $C^{\beta}-S^{\gamma}-S^{\gamma}-C^{\beta}$: $\chi_{ss} \approx \pm 90°$.
- Methionine (the other sulfur-containing amino acid) cannot form disulfide bonds
- Hair and feathers are mainly made of *keratins*; such proteins are linked together by disulfide bonds.

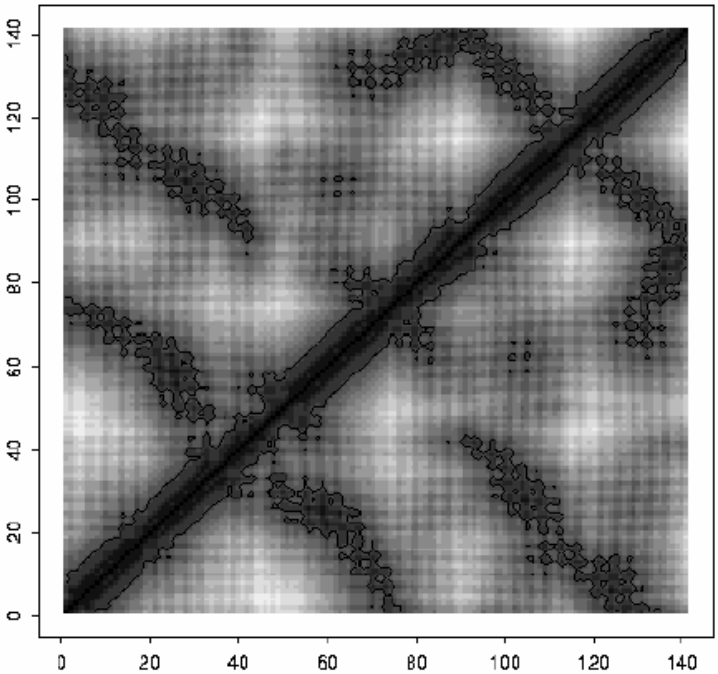
Distance & Contact Maps

- The distance r_{ij} between each possible pairs of atoms (a_i, a_j) can be reported onto a symmetrical *distance matrix* $\{r_{ij}\}$
- $\{r_{ij}\}$ is usually restricted to backbone Cαs only; in this case, *i* and *j* refer to the residue sequence number.
- Given $\{r_{ij}\}$ and a threshold distance *t*, the *contact map* $C_m(t)$ with elements $\{c_{ij}\}$ can be plotted: it's a symmetrical matrix with binary elements, defined as

  $c_{ij} = 1$ if $r_{ij} < t$, 0 otherwise

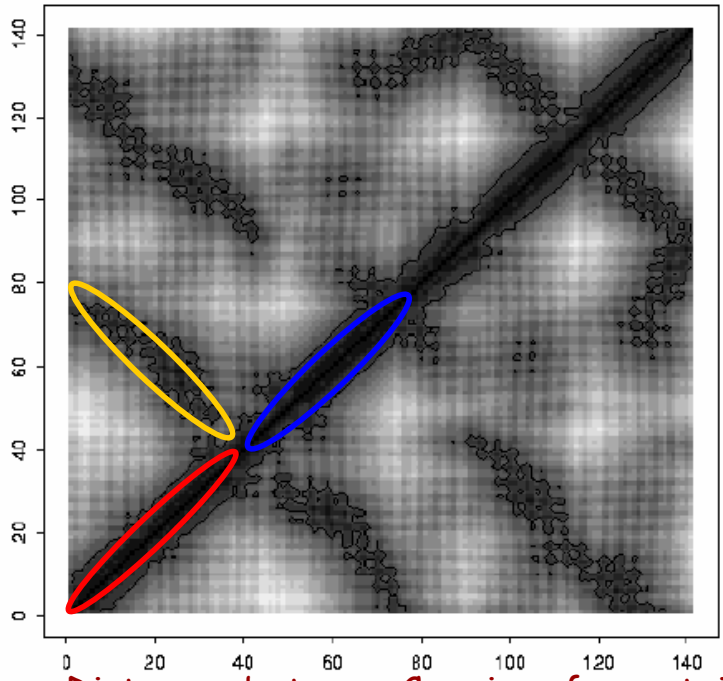
- Contact maps are a synthetic tool to visualize overall tertiary arrangements in proteins.
- Secondary elements determine specific patterns on contact maps

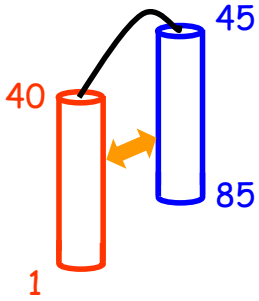
Intra-Molecular Distance Matrix



Distances between C_α pairs of a protein with 142 residues.
Darker squares represent shorter distances.

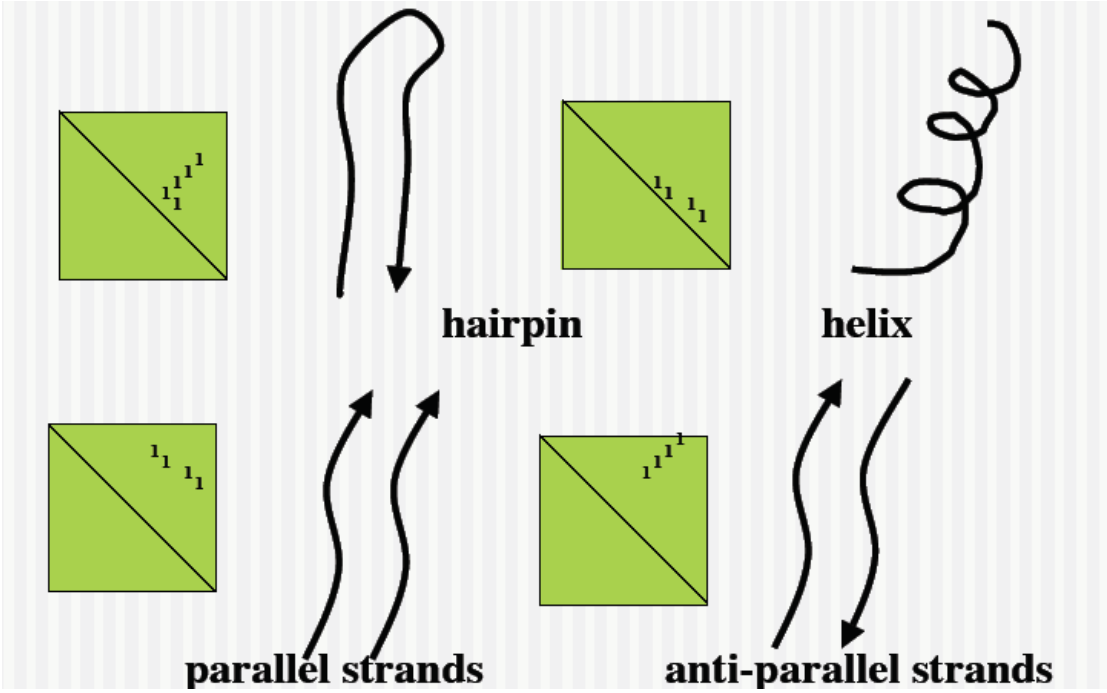
Intra-Molecular Distance Matrix





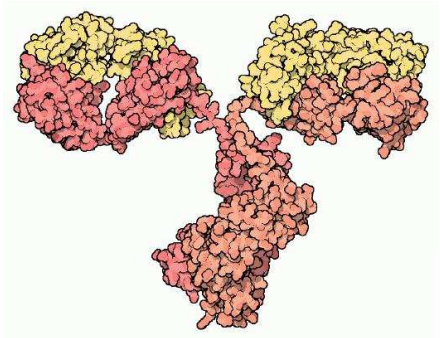
Distances between C_α pairs of a protein with 142 residues.
Darker squares represent shorter distances.

Intra-Molecular Distance Matrix

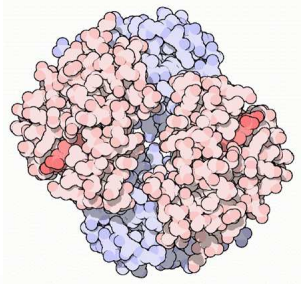


2ary-and-half Structures: Motifs

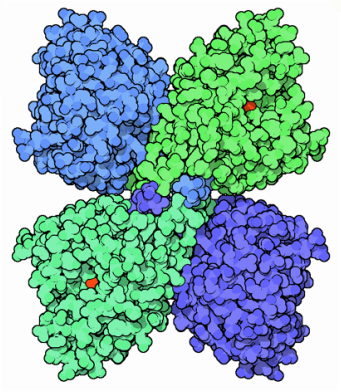
4-ary Structure: Examples



Antibodies



Hemoglobin



Tyrosine Hydroxylase

The Folding Process

Protein Stability

- Protein stability can be evaluated with $\Delta G \approx 5 \div 15$ kcal/mol (1 cal=4.2 J)
- Gibbs free energy: G=U+pV-TS, i.e. G=H-TS
  - U: internal energy; p: pressure; V: volume
  - T: temperature; S: entropy; H: enthalpy
- $\Delta H - T\Delta S \leq 0$ (second law of thermodynamics)
- With constant T and p, $\Delta G \leq 0$

Structure Determination

- X-ray crystallography
- NMR spectrometry