Recitation 4-2

EF lectures #12 & 13

Fundamentals of Protein Structure

Announcements

- Pset #3 was due today at noon
- Pset #4 should be released sometime Friday

Intro to Protein Structure

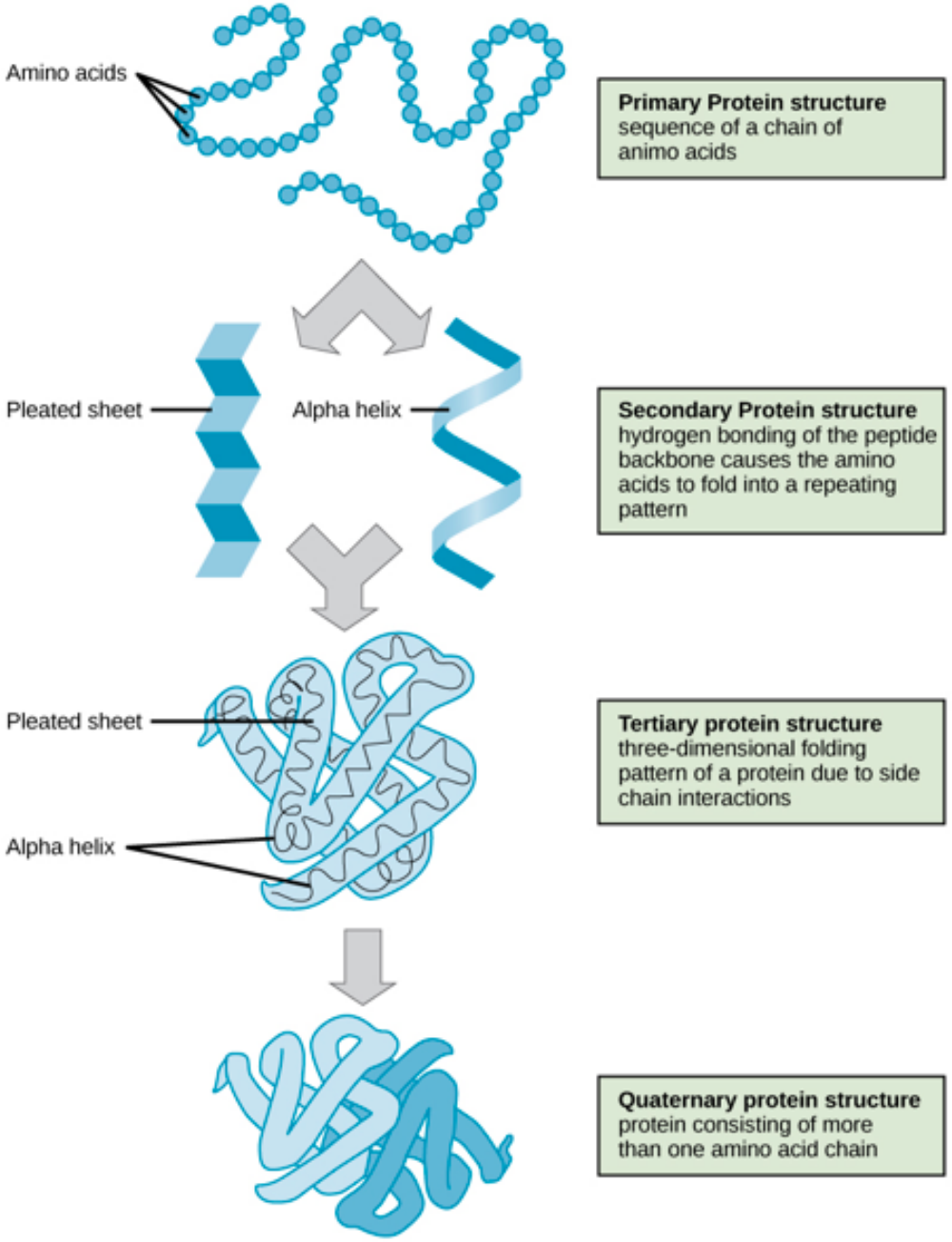


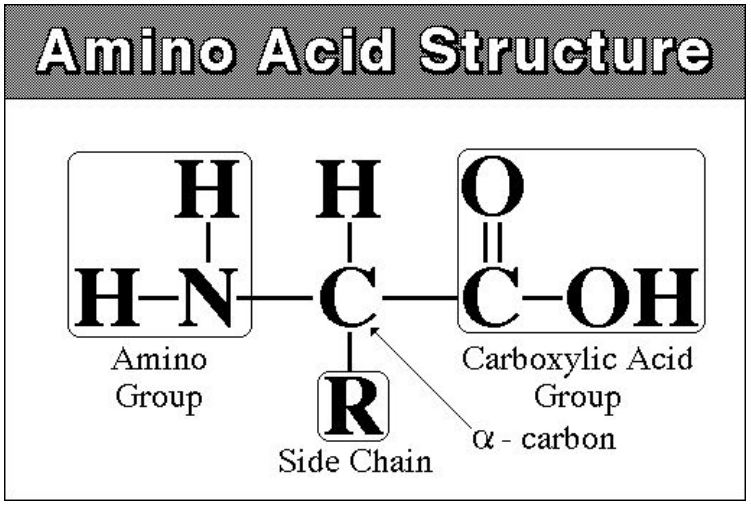
Figure 9: The four levels of protein structure can be observed in these illustrations. (credit: modification of work by National Human Genome Research Institute)

Courtesy of National Human Genome Research Institute. Image is in the public domain.

Four levels of protein structure:

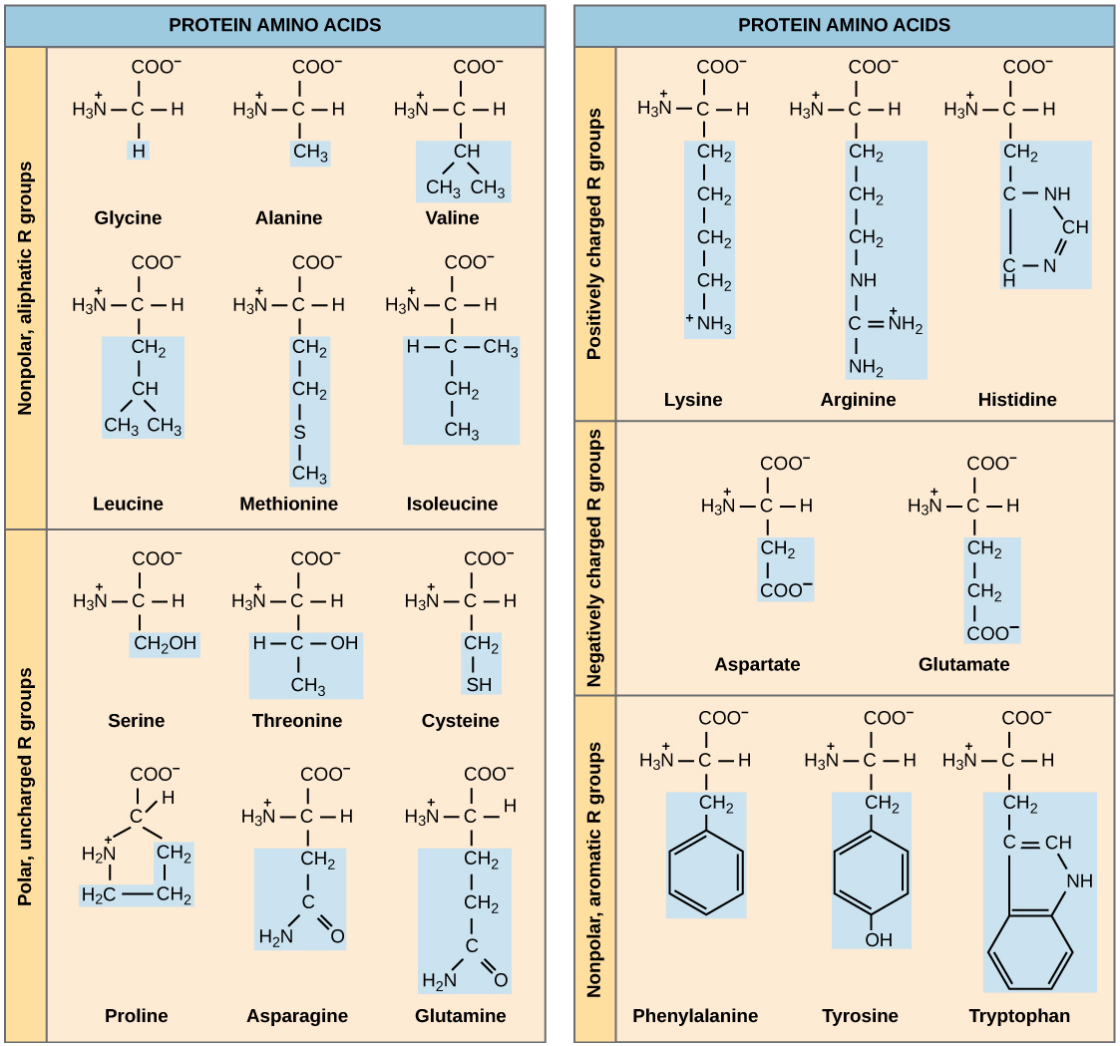
- **Primary:** the amino acid sequence
- **Secondary:** structures formed through interactions of the peptide backbone (α-helix, β-sheet most common)
- **Tertiary:** folding due to side chain interactions
- **Quaternary**: noncovalent association of 2+ folded polypeptides (may or may not be relevant)

Amino acid (=residue) structure



http://legacy.owensboro.kctcs.edu/gcaplan/anat/notes/amino_acid_structure_2.jpg

- 20 naturally occurring side chains; each has its own 3 letter (e.g. Met) and 1 letter (e.g. M) abbreviation
- Chemical properties (polar/ nonpolar/aromatic ring/ charge) and size are important for protein structure and function

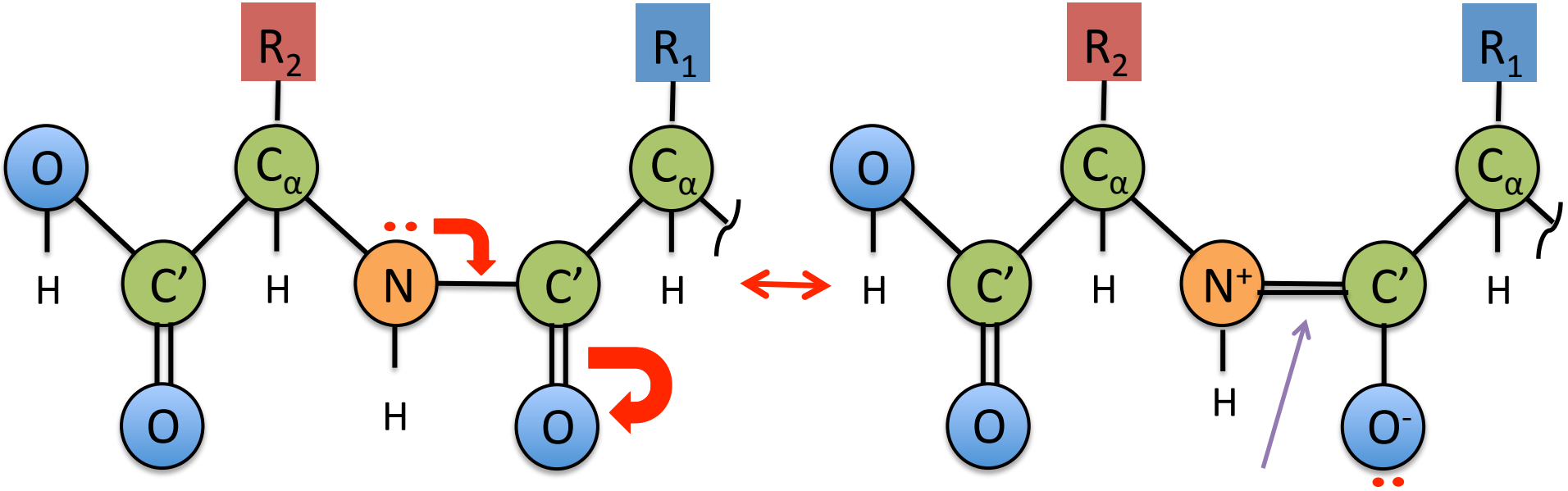


http://cnx.org/content/m44522/latest/Figure_15_01_01.jpg

Biochemistry Review

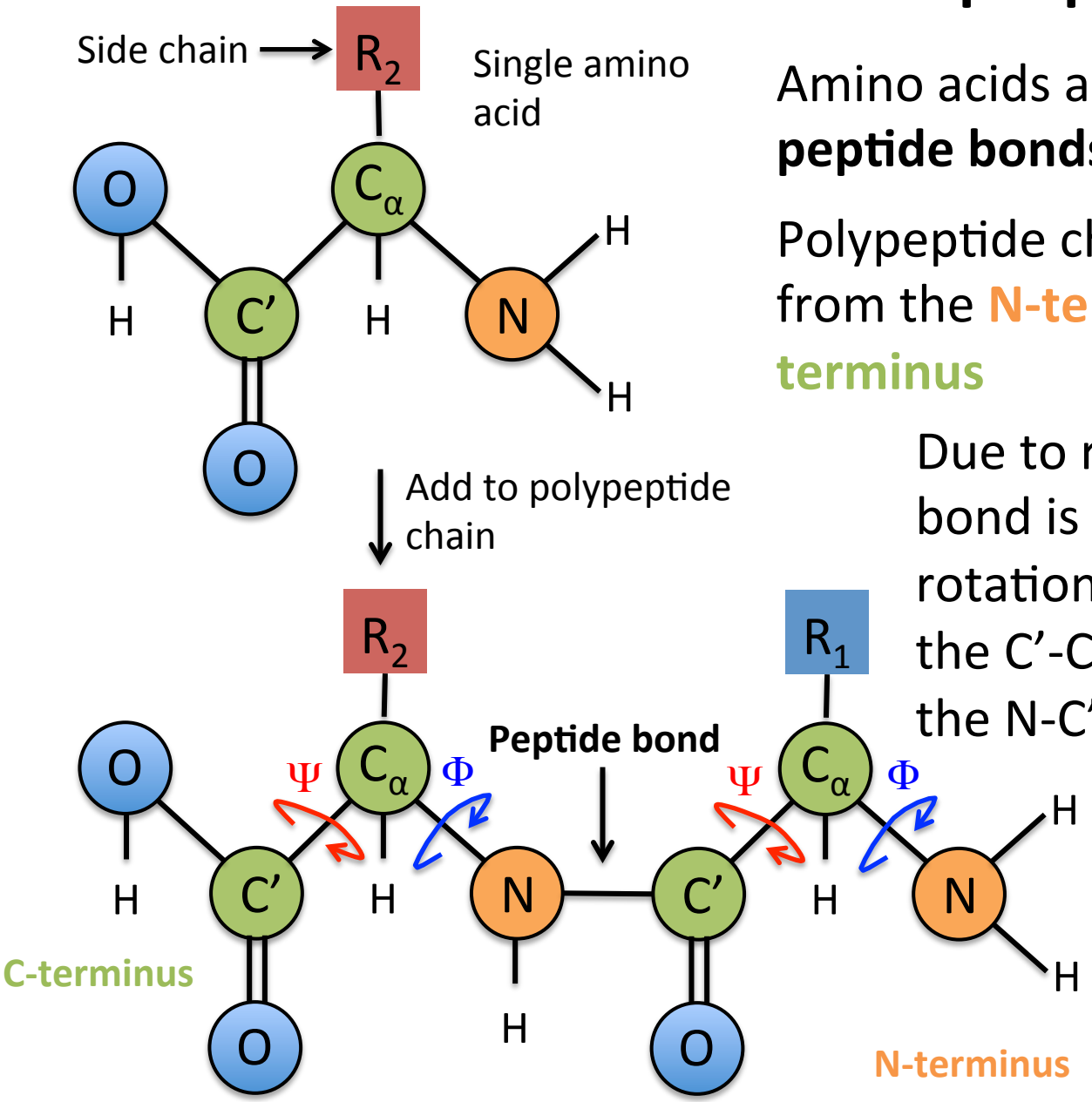
- Covalent bonds share electrons between atoms to fill their outer electron orbitals
  - -Single bond: 2 e^- (1 pair) shared
  - -Double bond: 4 e^- (2 pairs) shared
- Single bonds can rotate around bond axis; double bonds are more restricted and cannot

-Amino acids have resonance structures: electrons are delocalized within the molecule and cannot be represented by one single structure



The resonance structure gives the peptide bond *partial double bond character*, and its rotation is restricted

Amino acids and the peptide bond



Amino acids are connected by **peptide bonds**

Polypeptide chains are extended from the **N-terminus** to the **C-terminus**

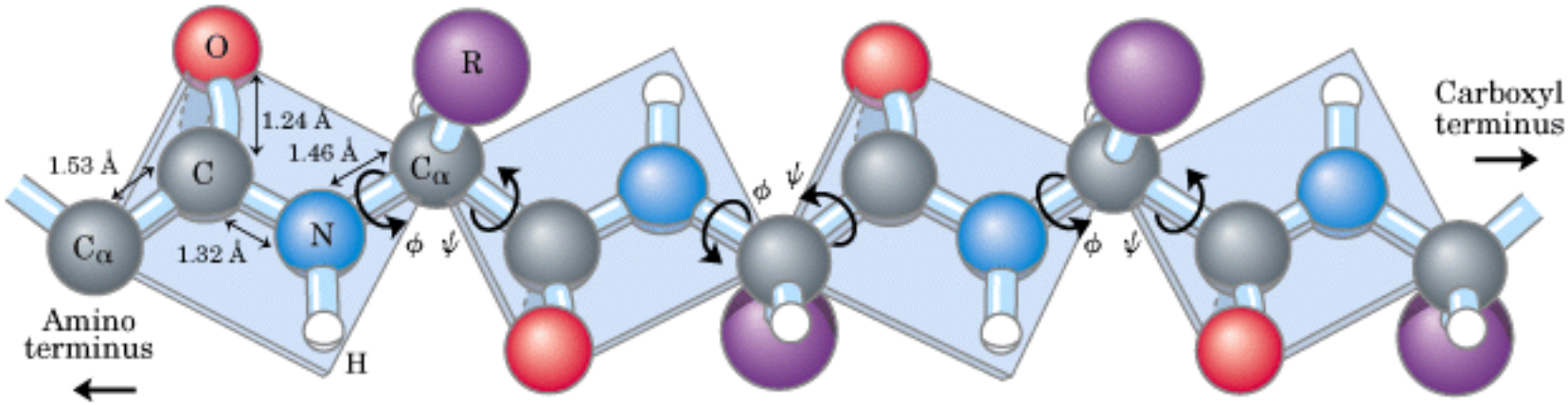
Due to resonance, the peptide bond is rigid, meaning that rotation can only occur along the C'-C_α and C_α-N bonds, not the N-C' bonds

C'-C_α : Ψ (psi)

C_α-N : Φ (phi)

Peptide chain

Can imagine each peptide as a planar "square", with rotation allowed only at the two opposite corners of the square (see below)

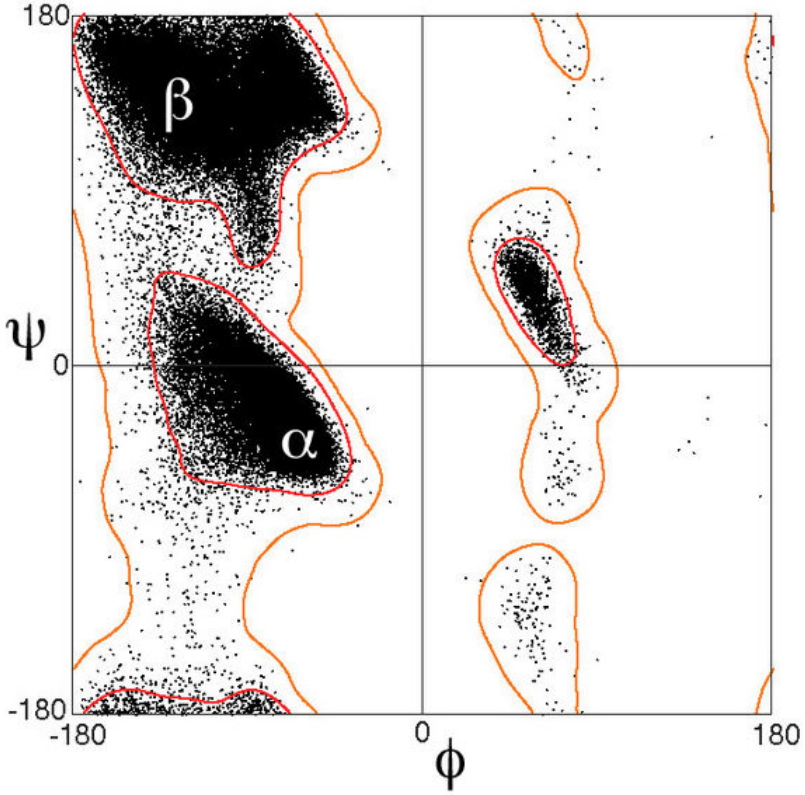


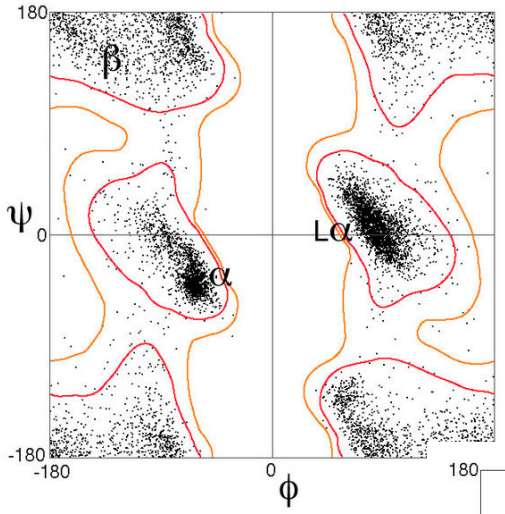
The Ψ and Φ angles are further restricted due to steric hindrance between the side chains and the peptide backbone – actual possible conformations are quite limited

We can explore the set of Ψ and Φ conformations that are favorable using a Ramachandran plot

Ramachandran Plot

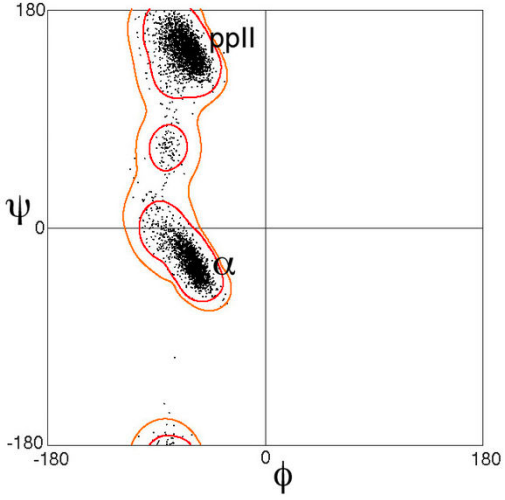
Shows which values of Ψ and Φ are possible for one type (or more generally, all) amino acids in a protein – most combinations of Ψ and Φ are forbidden due to steric hindrance





(**left**) Plot for glycine – gly side group is just -H, so has less restricted conformations due to less steric hindrance

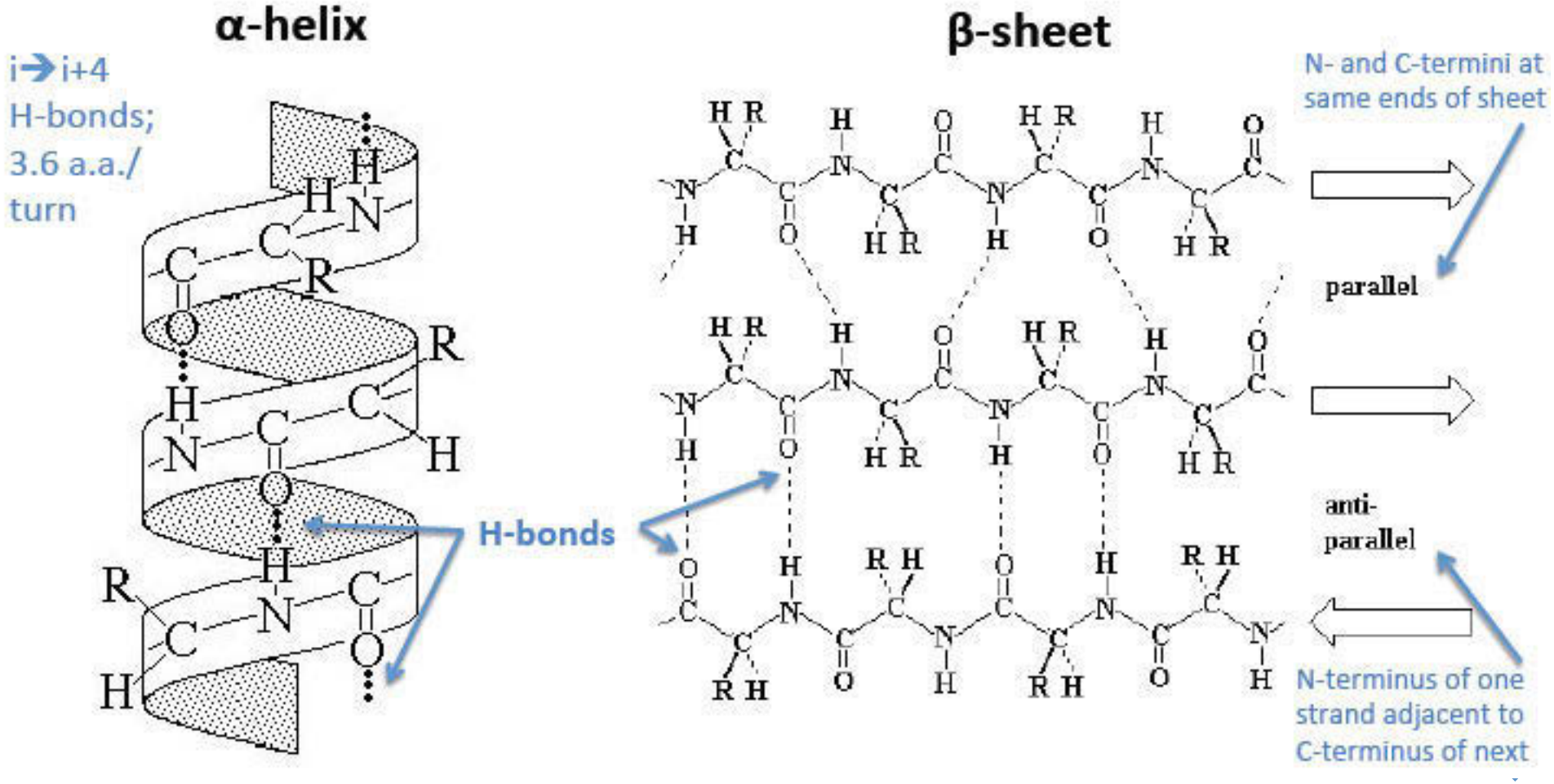
(**right**) Proline has an unusual structure (amine N bound to 2, not 1 C) that restricts its possible conformations



General case plot: uses data from nearly 100,000 residues from 500 structures (excluding Gly, Pro, and a.a. before Pro)

Secondary structure: α-helices and β-sheets

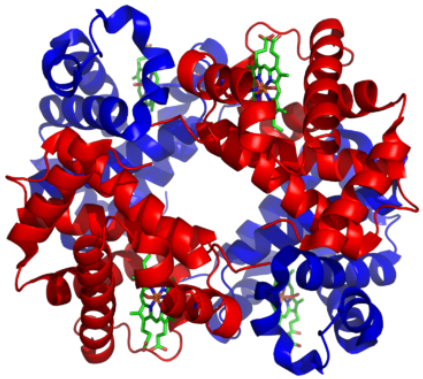
Form due to favorable interactions (hydrogen bonding) between molecules of the peptide backbone (not the side chains)



Secondary structure: α-helices and β-sheets

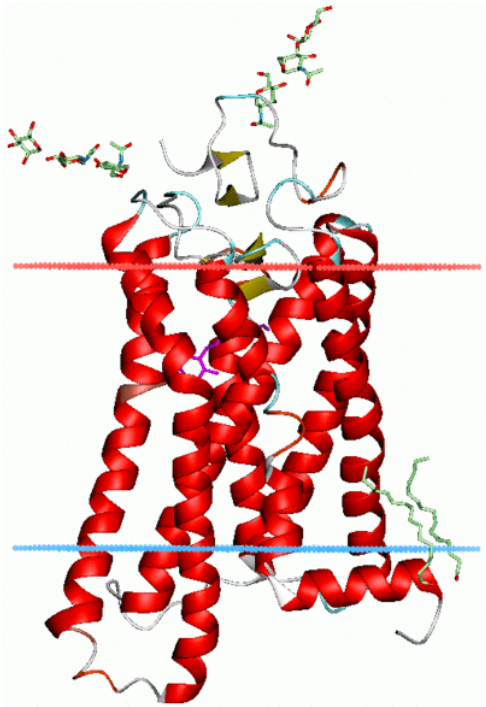
Form due to favorable interactions (hydrogen bonding) between molecules of the peptide backbone (not the side chains)

α-helix



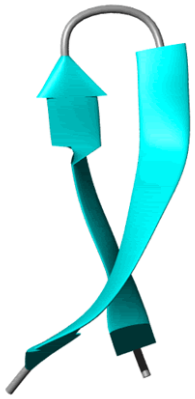
Hemoglobin

http://en.wikipedia.org/wiki/Alpha-helix



Rhodopsin

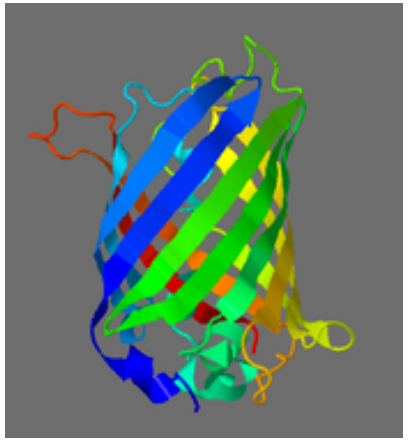
β-sheet



Courtesy of Isabella Daidone. Image in the public domain.

Courtesy of Jason Koval & Kevin Cartwright. Images in the public domain.

http://en.wikipedia.org/wiki/Beta-sheet



GFP: β-barrel

http://en.wikipedia.org/wiki/Beta_barrel

Courtesy of Christopher King. License: CC-BY.

Secondary structure: α-helices and β-sheets

Form due to favorable interactions (hydrogen bonding) between molecules of the peptide backbone (not the side chains)

α-helix

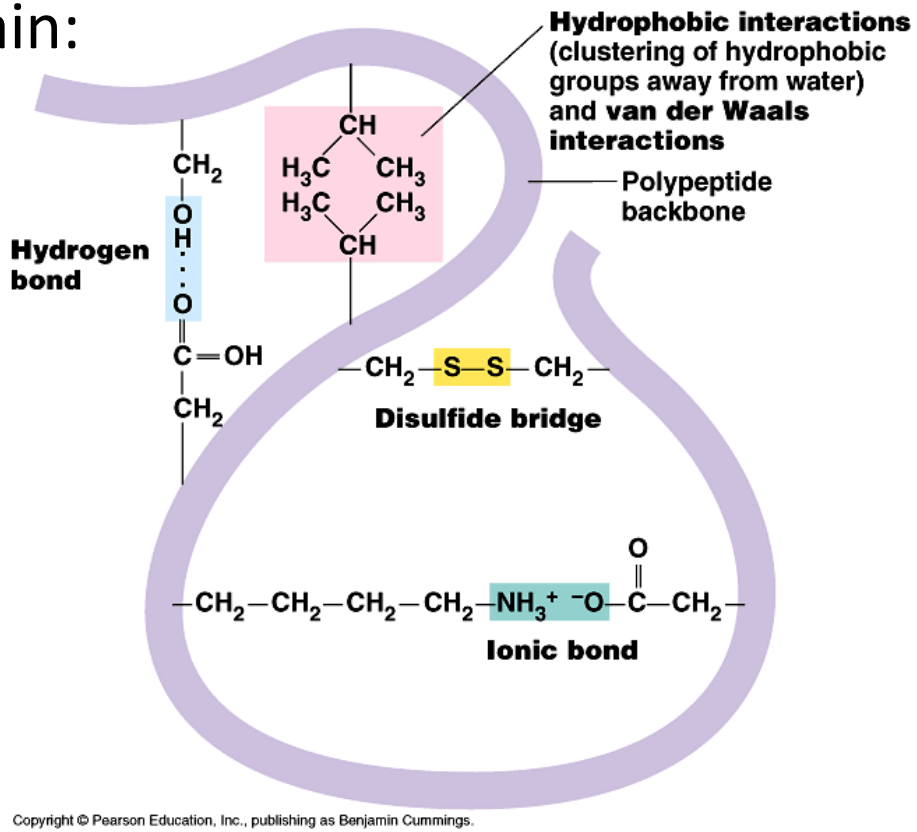
-If one side of helix faces interior of protein and one faces aqueous exterior, the helix will likely be amphipathic (=possesses both hydrophilic and hydrophobic properties)

Image of helical wheel representation of an amino acid sequence removed due to copyright restrictions.

Tertiary structure: side-chain interactions

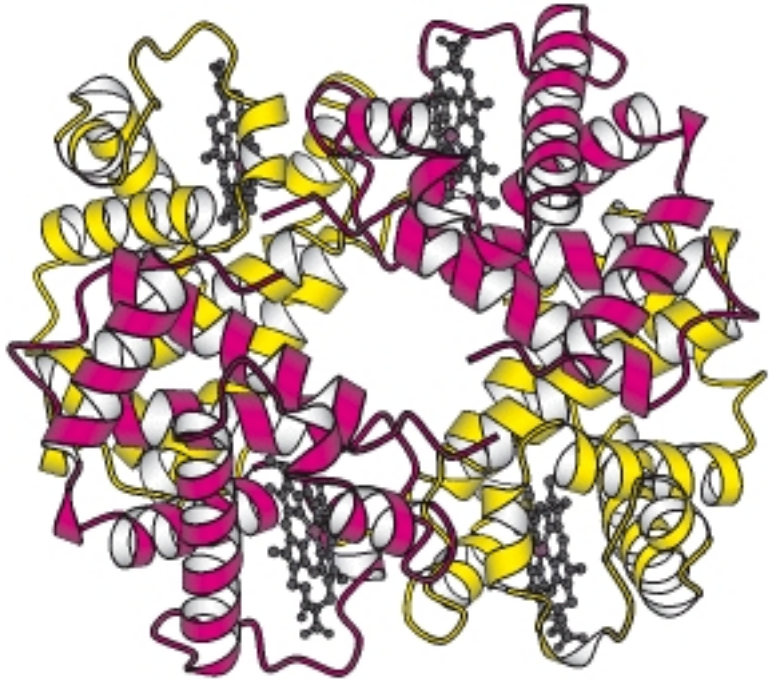
- Helices, sheets and other secondary structure elements are combined to produce the complete structure, largely through side chain interactions between amino acids that are far apart along the peptide chain:

- **Disulfide bridge (bond)**: strong covalent bonds that form between 2 Cysteine residues (S-S bond)
- **Hydrophobic interactions**: hydrophobic side chains tend to be packed away inside the protein, hydrophilic side chains on the outside so H_2O can form H-bonds with them
- **Interactions between charged residues** (ionic bonds)
- **Hydrogen bonding** between side chains



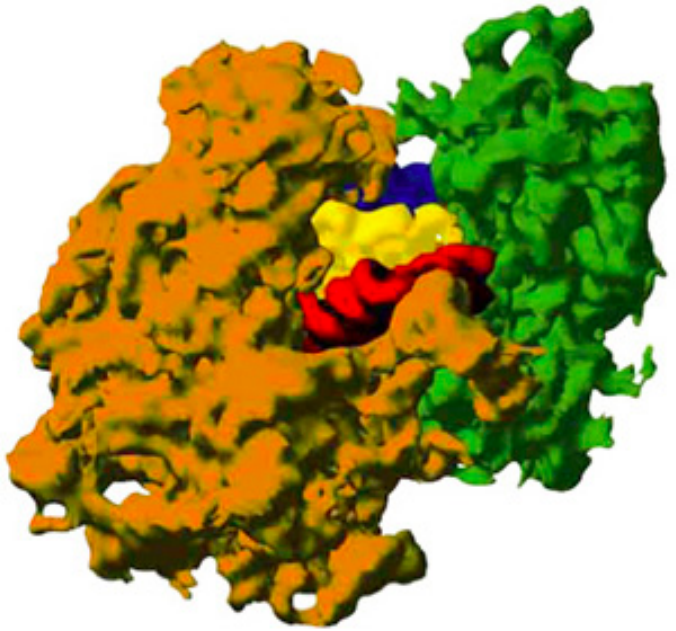
Quaternary Structure: assembling multiple peptide subunits

- Many proteins contain more than one polypeptide chain (subunits) which interact to form the functional protein
  - Maintained by interchain interactions



The $\alpha_2\beta_2$ Tetramer of Human Hemoglobin

From Biochemistry. 5th edition.

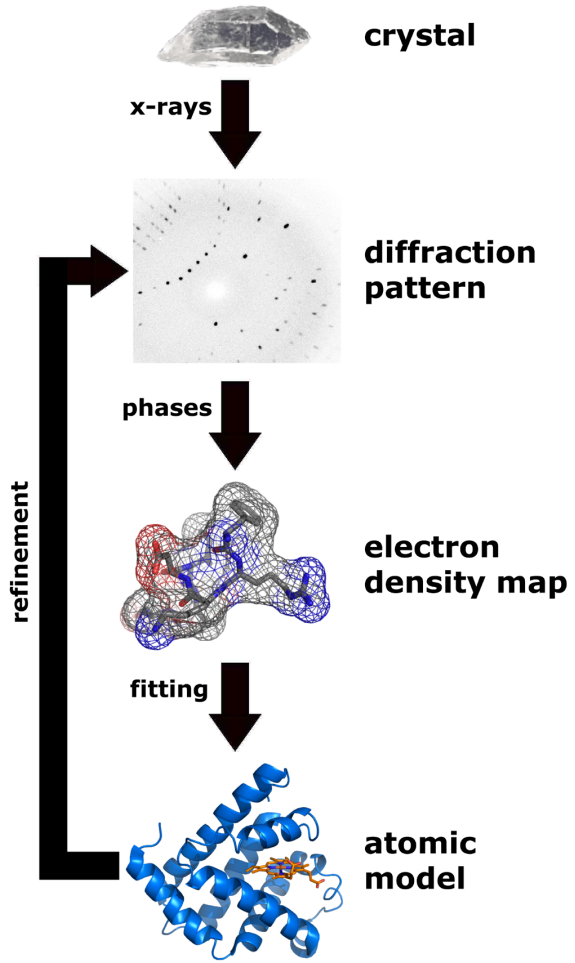


The ribosome consists of a large (orange) and small (green) protein subunit (also contains RNA)

http://publications.nigms.nih.gov/thenewgenetics/chapter1.html

Protein structure computationally inferred through 2 methods

- 1. X-ray crystallography (most common)
  - Crystalline atoms cause a beam of incident X-rays to diffract into many specific direction
  - Crystallizing a protein is difficult!

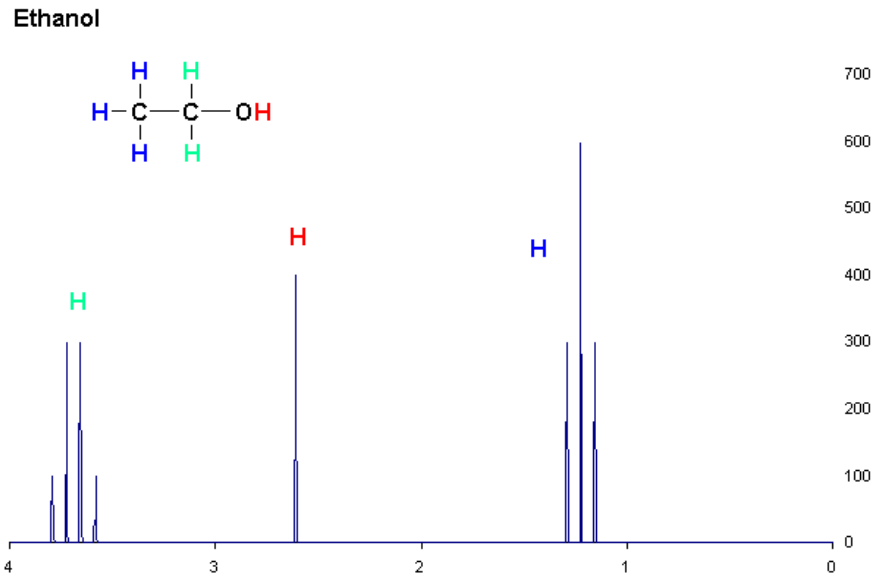


Courtesy of Thomas Splettstößer. Used with Permission.

http://en.wikipedia.org/wiki/X-ray_crystallography

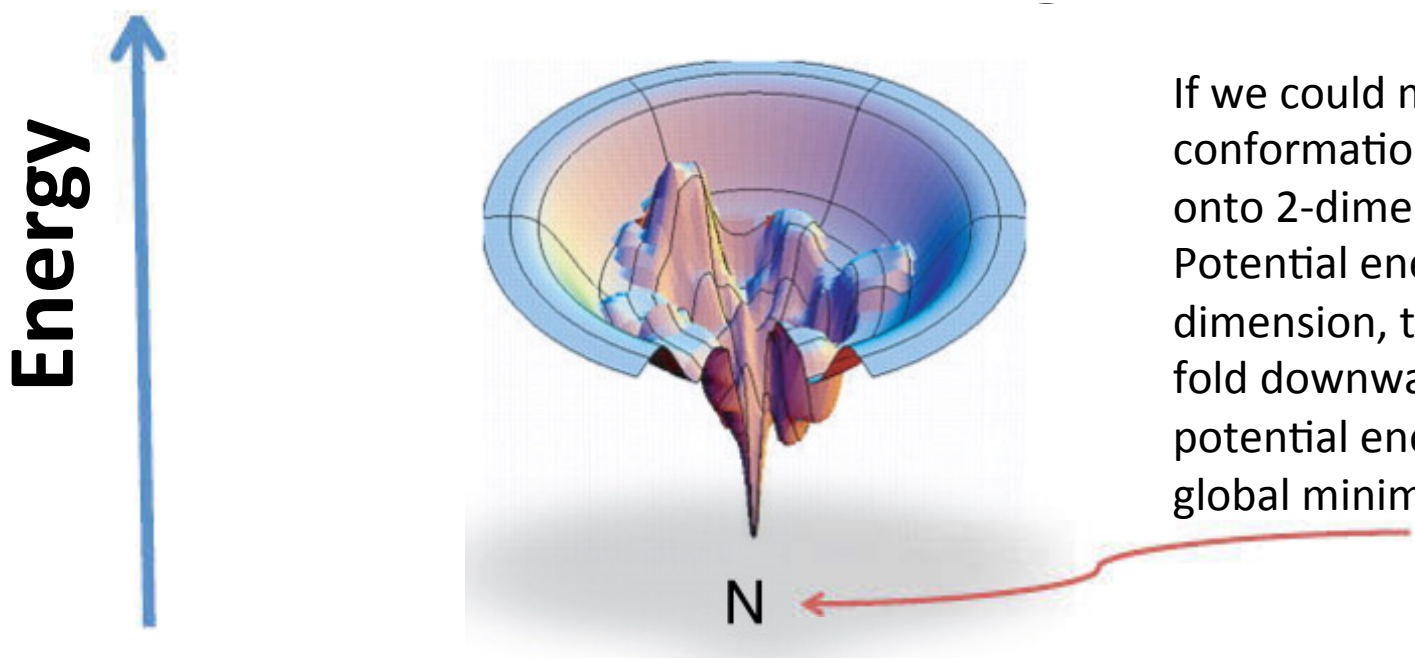
Protein structure computationally inferred through 2 methods

- 2. NMR (nuclear magnetic resonance) spectroscopy: exploits the magnetic properties of atomic nuclei
  - Intramolecular magnetic field around an atom in a molecule changes the resonance frequency, giving access to details of the electronic structure of a molecule
  - Usually limited to small (<35 kDa) proteins (more common for small, organic molecules)
  - Used for intrinsically disordered proteins or others that can't be crystallized



http://en.wikipedia.org/wiki/NMR_spectroscopy

Proteins usually assume structures that minimize potential energy



If we could map all possible conformations of a protein onto 2-dimensions with Potential energy as the 3rd (z-) dimension, the protein would fold downward along the potential energy surface to the global minimum

For a particular structure, how do we compute its potential energy? 2 main approaches:

1. Physical explanation for forces (CHARMM)
2. Statistical comparison of structure's components to those observed in other proteins (Rosetta)

1. CHARMM (Physicist's Approach)

- Chemistry at HARvard Macromolecular Mechanics (developed by Martin Karplus's group – 2013 Nobel Prize in Chemistry)
- Describe physical force fields, which may be approximate but represent identifiable sources
- $U_{CHARMM} = U_{bonded} + U_{non-bonded}$

$$U_{bond} = \sum_{bonds} K_b(b - b^0)^2,$$

$$U_{angle} = \sum_{angles} K_\theta(\theta - \theta^0)^2,$$

Bond stretching and angle bending modeled as harmonic oscillators

$$U_{UB} = \sum_{Urey-Bradley} K_{UB}(b^{1-3} - b^{1-3,0})^2,$$

$$U_{dihedral} = \sum_{dihedrals} K_\varphi((1 + \cos(n\varphi - \delta)),$$

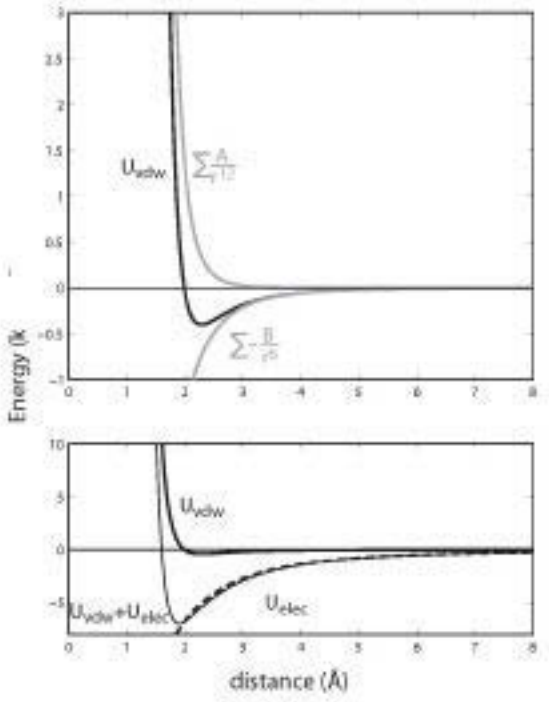
$$U_{improper} = \sum_{impropers} K_\omega(\omega - \omega^0)^2, \text{ and}$$

$$U_{CMAP} = \sum_{residues} u_{CMAP}(\Phi, \Psi)$$

bonded terms describe interactions between chemically bonded neighbors

1. CHARMM (Physicist's Approach)

- Describe physical force fields, which may be approximate but represent identifiable sources
- $U_{CHARMM} = U_{bonded} + U_{non-bonded}$



Lennard-Jones Potential – approximates interactions between a pair of neutral atoms or molecules

$$U_{vdw} \equiv U_{LJ} = \sum_{nonb.pairs} \varepsilon_{ij} \left[\left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{6} \right],$$

$$U_{elec} = \sum_{nonb.pairs} \frac{q_i q_j}{\epsilon r_{ij}}.$$

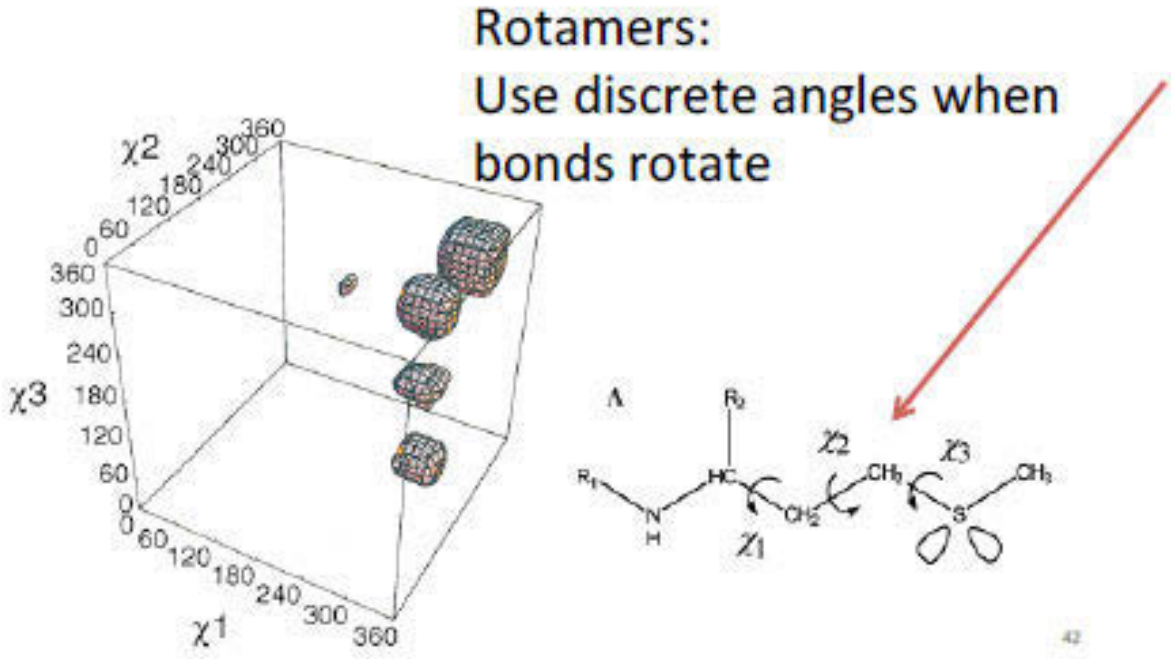
potential energy due to electrostatic interactions

r^{-12} term: describes repulsion at short ranges due to overlapping electron orbitals

r^{-6} term: describes attraction at long ranges (van der Waals force, or dispersion force)

2. Rosetta (Statistician's Approach)

- Assigns energy values for many of the same forces as CHARMM, but does so by comparing the protein's conformation with those of other observed structures instead of from first principles (e.g. modeling forces as harmonic oscillators, etc.)



Example: angles of methionine side chain are observed to only take on a small number of conformations (rotamers)

-Analogous to Ramachandran plot, but for side chain angles instead of phi/psi angles of protein backbone

2. Rosetta (Statistician's Approach)

- Assigns energy values for many of the same forces as CHARMM, but does so by comparing the protein's conformation with those of other observed structures instead of from first principles (e.g. modeling forces as harmonic oscillators, etc.)

$$Energy = w_1 * term_1 + w_2 * term_2 + \ldots$$

Rosetta energy terms (Pset 3 Q3)

Rosetta Full-atom Scoring Functions		
Van der Waals net attractive energy	FA	fa_atr
Van der Waals net repulsive energy	FA	fa_rep
Hydrogen bonds, short and long-range, (backbone)	FA/CEN	hbond_sr_bb, hbond_lr_bb
Hydrogen bonds, short and long-range, (side-chain)	FA	hbond_sc, hbond_bb_sc
Solvation (Lazaridis-Karplus)	FA	fa_sol
Dunbrack rotamer probability	FA	fa_dun
Statistical residue-residue pair potential	FA	fa_pair
Intra-residue repulsive Van der Waals	FA	fa_intra_rep
Electrostatic potential	FA	hack_elec
Disulfide statistical energies (S-S distance, etc.)	FA	dslf_ss_dst, dslf_cs_ang, dslf_ss_dih, dslf_ca_dih
Amino acid reference energy (chemical potential)	FA/CEN	ref
Statistical backbone torsion potential	FA/CEN	rama
Van der Waals "bumps"	CEN	vdw
Statistical environment potential	CEN	env
Statistical residue-residue pair potential (centroid)	CEN	pair
Cb		cbeta

Methods for Refining Structures

- Once we have a starting structures, how can we refine it? Lower potential energy by:
  - Changing protein backbone phi/psi angles
  - Side-chain "packing" of amino acid R groups
- 3 approaches to refine structures:
  - 1. Energy Minimization
  - 2. Molecular Dynamics
  - 3. Simulated Annealing

1. Energy Minimization

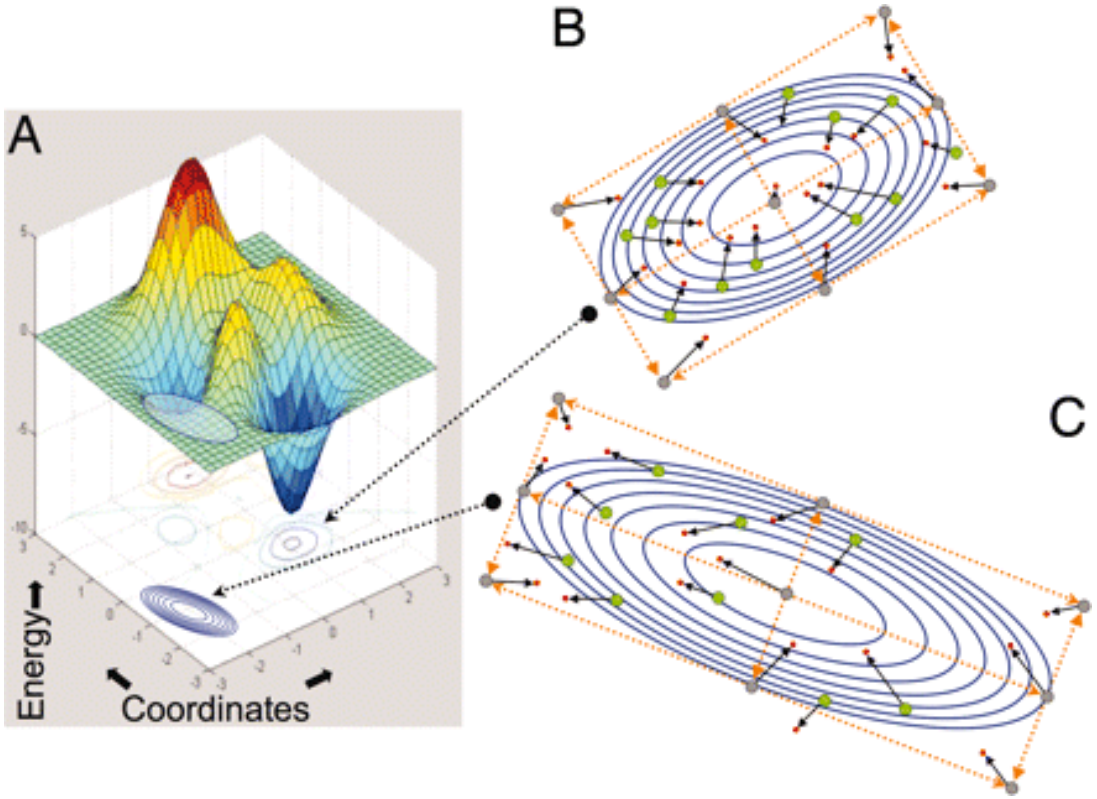
- Deterministic process of going downward on potential energy surface

- Perform gradient (multidimensional analog of derivative) descent along potential energy surface
- Downward on potential energy surface is going in the direction of the force field since

$$F(\vec{x}) = -\nabla U(\vec{x})$$

for atomic positions $\vec{x}$

- Local search since you only go immediately downward to local minimum, which may or may not be global minimum



Courtesy of the National Academy of Sciences. Used with permission.
Source: Summa, Christopher M., and Michael Levitt. "Near-native Structure Refinement using in Vacuo Energy Minimization." *Proceedings of the National Academy of Sciences* 104, no. 9 (2007): 3177-82.

http://www.pnas.org/content/104/9/3177

2. Molecular Dynamics

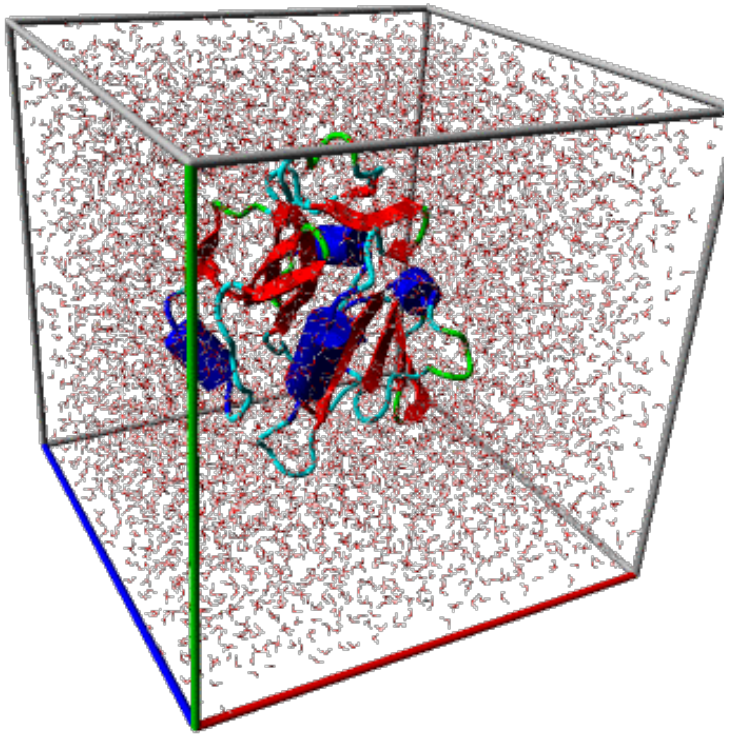
- Calculate force field between all atoms in protein and surrounding environment (solvent, lipid bilayer, etc.)
- From force field, can calculate velocity, and use this to update positions over very small timescale

  (t_i-t_{i-1} ~ femtoseconds):

$$(t_i) = x(t_{i-1}) + v(t_{i-1}) \times (t_i - t_{i-1})$$

$$v(t_i) = v(t_{i-1}) - \frac{\nabla U(t_{i-1})}{m} \times (t_i - t_{i-1})$$

- Then recalculate force field based on new atomic positions, repeat…
- Very computationally expensive since pairwise interactions between thousands of atoms at each timestep x billions or more timesteps



Courtesy of Elmar Krieger, Yasara. Used with permission.

http://www.yasara.org/dhfr.gif

3. Simulated Annealing

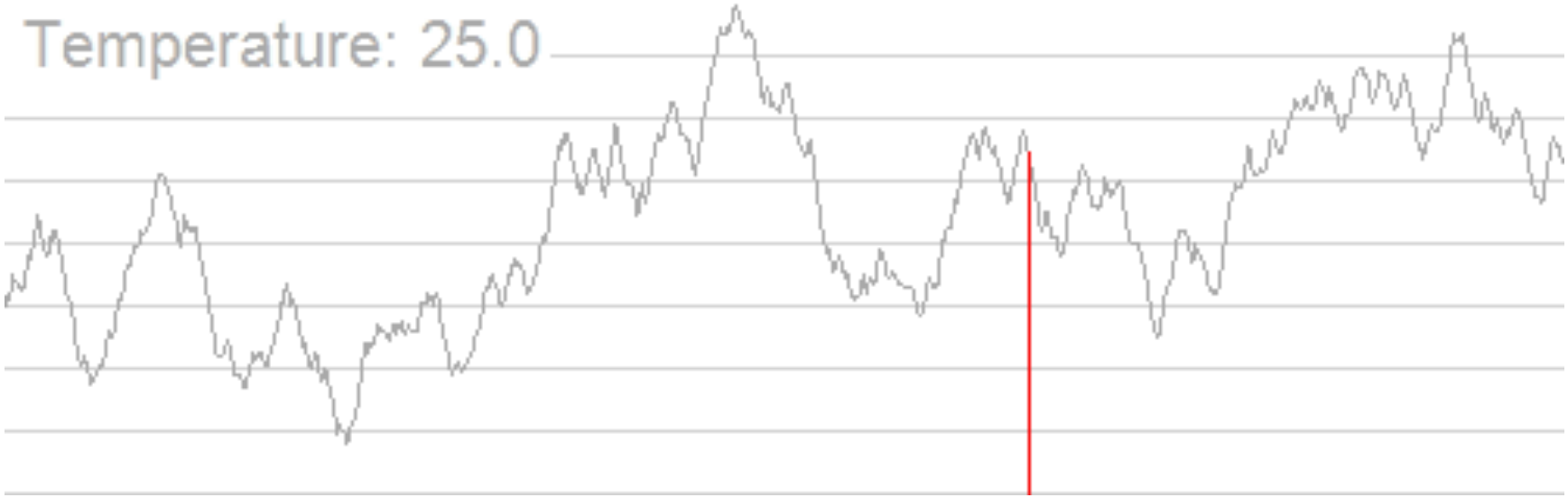
- Stochastic search of protein conformations
- Metropolis-Hastings algorithm is an implementation of Simulating Annealing that generates sample states of a thermodynamic system
  - Monte-Carlo method (as was Gibbs Sampler)
- Idea: Sample a new conformation by perturbing protein's current structure
  - If lower energy, accept new conformation
  - If higher energy, accept it with probability proportional to difference in energy between two structures
    - Unlike Energy Minimization, it is possible to move upward on the potential energy surface, which may allow us to escape a local minimum and find the global minimum

3. Simulated Annealing

- Name and idea come from metallurgy
  - High temperature – more molecular movement
    - Metropolis-Hastings algorithm for protein folding: more sampling of protein conformations / movement along the potential energy surface
    - High probability of jumping from local minimum to another region of the potential energy surface
  - Lower temperature – less molecular movement
    - Low probability of escaping current local minimum – used for for small refinements of protein structure
  - Annealing schedule specifies how temperature changes from high to low over time

3. Simulated Annealing

-Example searching for maximum instead of minimum:



Courtesy of Kingpin13. Image in the public domain.

http://en.wikipedia.org/wiki/File:Hill_Climbing_with_Simulated_Annealing.gif#file

3. Simulated Annealing

- Metropolis-Hastings acceptance criterion:
  - If test conformation has lower energy, always accept it
  - If test conformation has higher energy, accept it with probability:

$$\frac{P(S_{\text{test}})}{P(S_{\text{current}})} = \frac{\left(e^{-E_{\text{test}}/kT}\right)/Z(T)}{\left(e^{-E_{\text{current}}/kT}\right)/Z(T)} = e^{-(E_{\text{test}}-E_{\text{current}})/kT}$$

Probability of each conformation follows Boltzmann distribution

Z(T) is a normalization constant that cancels

k is Boltzmann constant; can refer to *kT* as "temperature"

  - At higher temperature, exponent is closer to 0 so acceptance probability is closer to 1 (more likely to accept higher energy conformational changes)

Structure discovery using homology

- In determining a structure, if regions of your query protein are homologous (evolutionarily related) to other proteins of known structure, they likely adopt the homologous structure/fold
- Align your query protein to those in the PDB:
  - High (50%+) sequence similarity
  - Medium (20%-50%) sequence similarity
  - Low (<20%) sequence similarity

Structure discovery using homology

- In determining a structure, if regions of your query protein are homologous (evolutionarily related) to other proteins of known structure, they likely adopt the homologous structure/fold
- Align your query protein to those in the PDB:
  - High (50%+) sequence similarity
    - Can be confident the structures are very similar – generally only need to refine regions where alignment is poor.

Structure discovery using homology

- In determining a structure, if regions of your query protein are homologous (evolutionarily related) to other proteins of known structure, they likely adopt the homologous structure/fold
- Align your query protein to those in the PDB:
  - Medium (20%-50%) sequence similarity
    - Try several alignments, and refine structure resulting from each. Choose the final based on lowest energy.

Structure discovery using homology

- In determining a structure, if regions of your query protein are homologous (evolutionarily related) to other proteins of known structure, they likely adopt the homologous structure/fold
- Align your query protein to those in the PDB:
  - Low (<20%) sequence similarity
    - Lots of possible starting structures from different alignments. Need to do more aggressive refinement since final structure may be significantly different than starting structure; choose final structure based on lowest energy.

Structure discovery without homology

- Monte Carlo search of backbone phi/psi angles
  - Choose small region of 3-9 a.a.s and set angles to those of similar peptide in PDB
  - Calculate energy of structure with new angles of the 3-9 a.a.s
  - Accept according to Metropolis criterion
- Repeat this 36,000 times to get 1 final structure

Repeat 20,000 times to get 20,000 structures

- Cluster the many structures into small number of representative groups; more sophisticated refinement of groups

7.91J / 20.490J / 20.390J / 7.36J / 6.802 / 6.874 / HST.506 Foundations of Computational and Systems Biology
Spring 2014