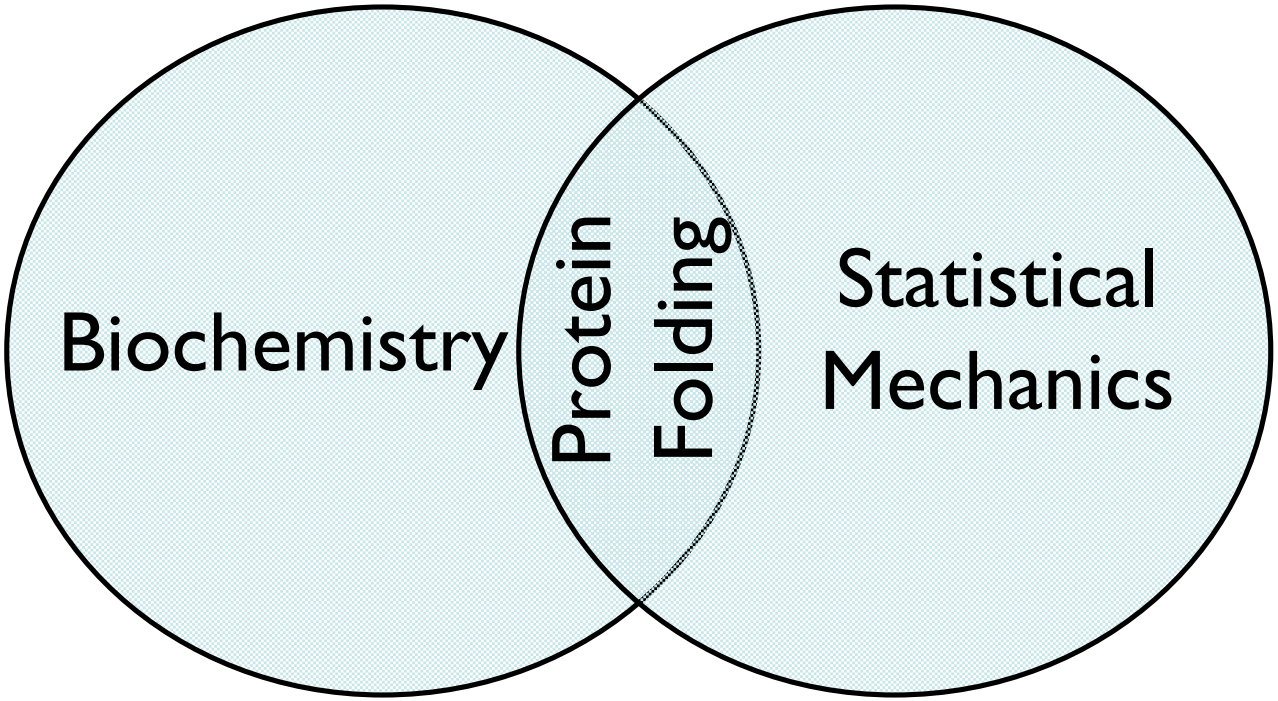
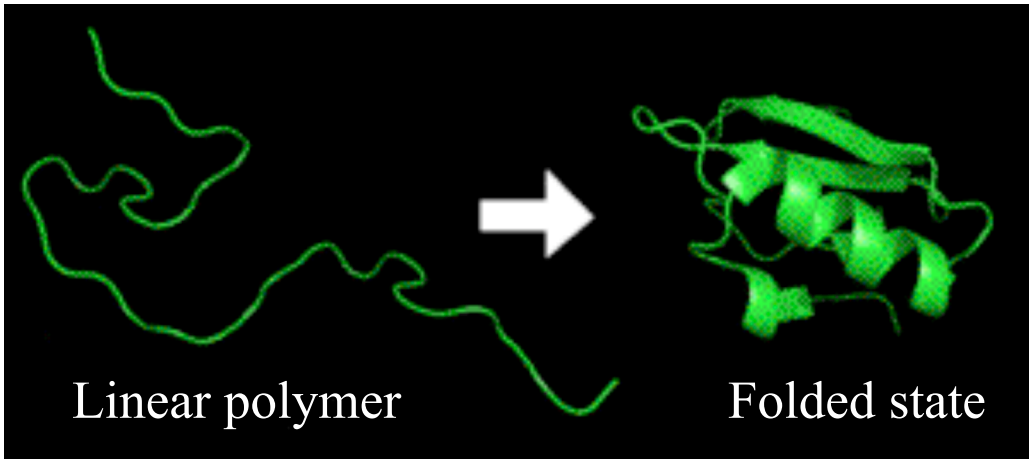
Bioinformatics: Practical Application of Simulation and Data Mining

Protein Folding I

Prof. Corey O'Hern
Department of Mechanical Engineering
Department of Physics
Yale University

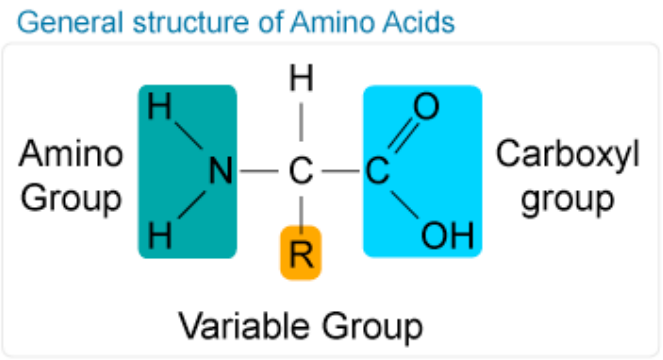

Biochemistry
Protein Folding
Statistical Mechanics

What are proteins?



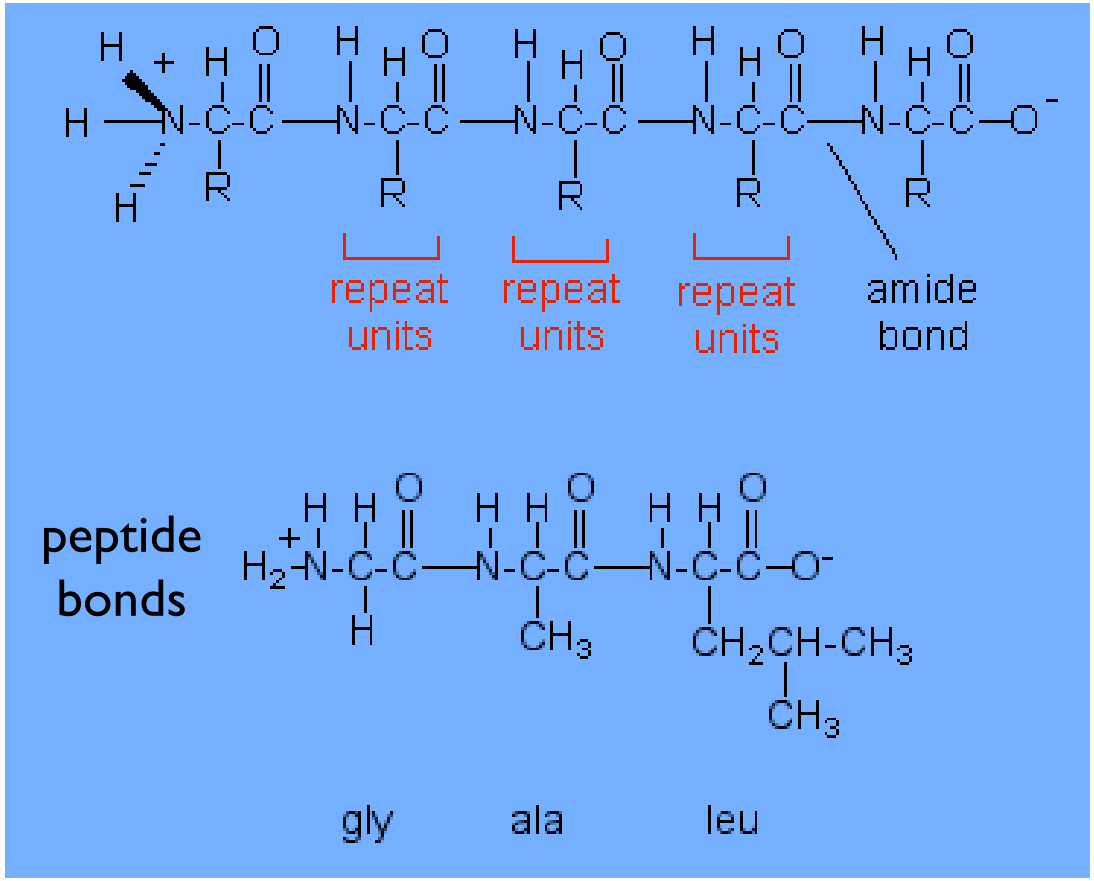
- Proteins are important; e.g. for catalyzing and regulating biochemical reactions, transporting molecules, …
- Linear polymer chain composed of tens (peptides) to thousands (proteins) of monomers
- Monomers are 20 naturally occurring amino acids
- Different proteins have different amino acid sequences
- *Structureless*, extended unfolded state
- Compact, 'unique' native folded state (with secondary and tertiary structure) required for biological function
- Sequence determines protein structure (or lack thereof)
- Proteins unfold or denature with increasing temperature or chemical denaturants

Amino Acids I



N-terminal C_α C-terminal

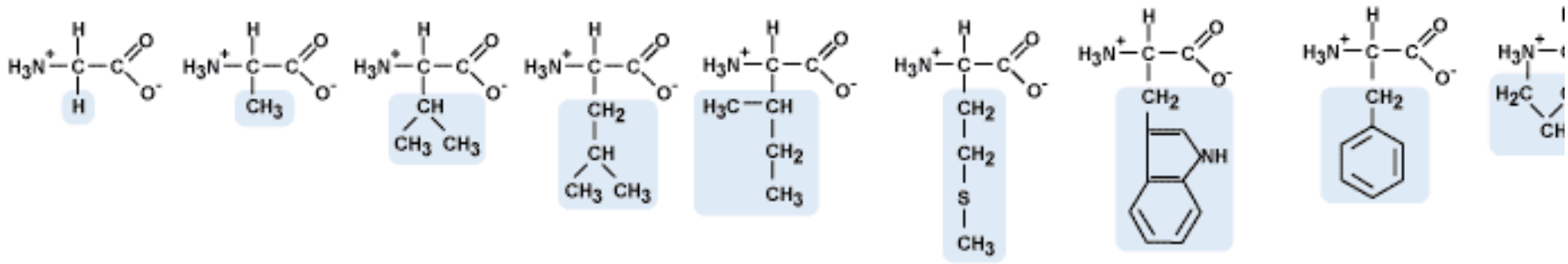
R
variable
side chain



- Side chains differentiate amino acid repeat units
- Peptide bonds link residues into polypeptides

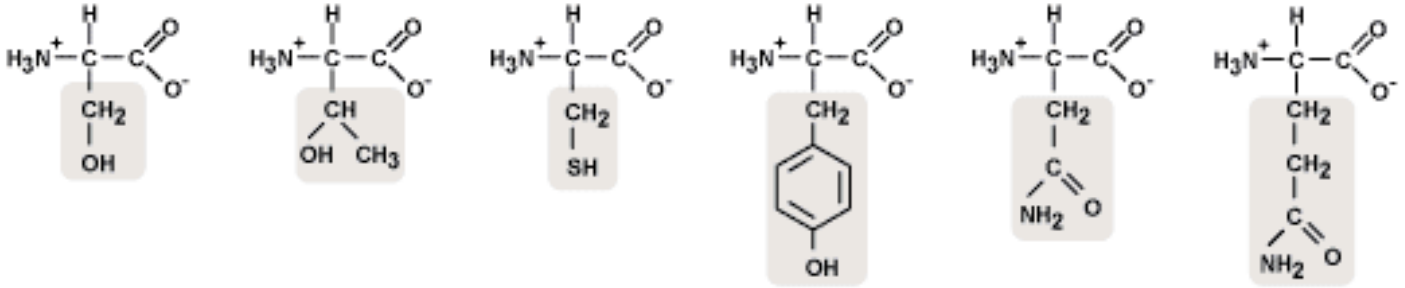
Amino Acids II

NONPOLAR



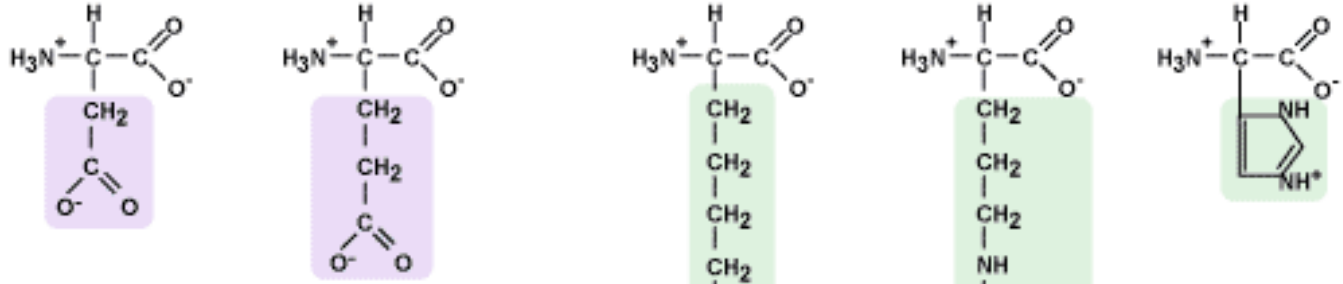
Glycine (Gly) Alanine (Ala) Valine (Val) Leucine (Leu) Isoleucine (Ile) Methionine (Met) Tryptophan (Trp) Phenylalanine (Phe) Proli

POLAR



Serine (Ser) Threonine (Thr) Cysteine (Cys) Tyrosine (Tyr) Asparagine (Asn) Glutamine (Gln)

Electrically Charged



Acidic

Aspartic Acid (Asp) Glutamic Acid (Glu)

Lysine (Lys)

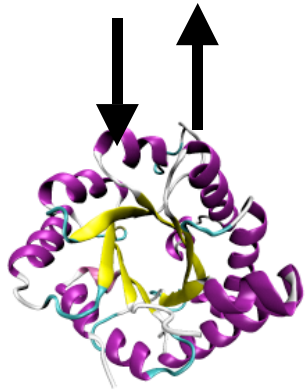
Basic

Arginine (Arg) Histidine (His)

The Protein Folding Problem:

What is ‘unique’ folded 3D structure of a protein based on its amino acid sequence? Sequence → Structure

Lys-Asn-Val-Arg-Ser-Lys-Val-Gly-Ser-Thr-Glu-Asn-Ile-Lys- His-Gln-Pro- Gly-Gly-Gly-…



Driving Forces

- Folding: hydrophobicity, hydrogen bonding, van der Waals interactions, …
- Unfolding: increase in conformational entropy, electric charge…

inside H (hydrophobic)

outside P (polar)

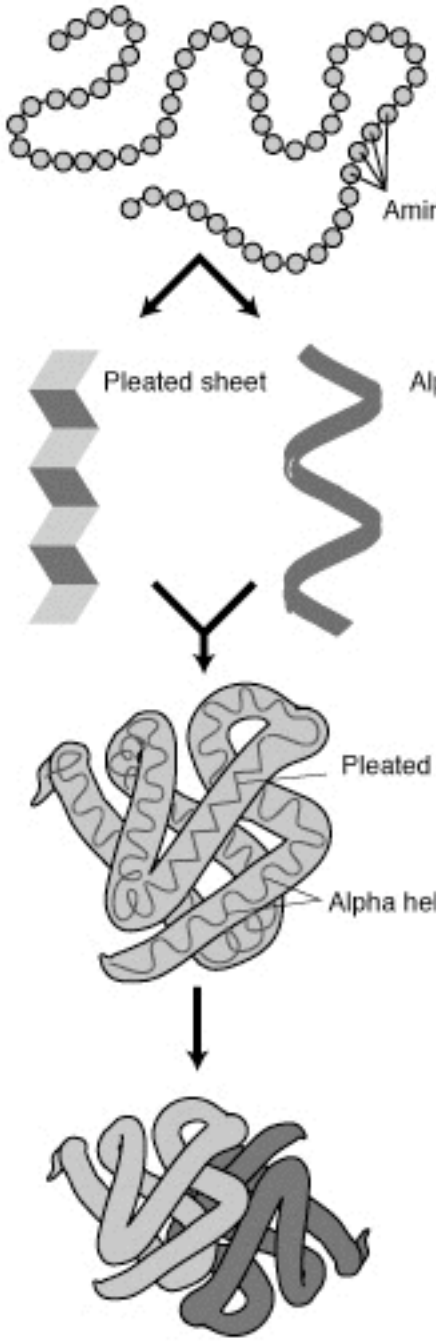
Hydrophobicity index

At pH 2[A]		At pH 7[B]	
Very Hydrophobic			
Leu	100	Phe	100
Ile	100	Ile	99
Phe	92	Trp	97
Trp	84	Leu	97
Val	79	Val	76
Met	74	Met	74
Hydrophobic			
Cys	52	Tyr	63
Tyr	49	Cys	49
Ala	47	Ala	41
Neutral			
Thr	13	Thr	13
Glu	8	His	8
Gly	0	Gly	0
Ser	-7	Ser	-5
Gln	-18	Gln	-10
Asp	-18		
Hydrophilic			
Arg	-26	Arg	-14
Lys	-37	Lys	-23
Asn	-41	Asn	-28
His	-42	Glu	-31
Pro	-46	Pro	-46 (used pH 2)
		Asp	-55

[A] pH 2 values: Normalized from Sereda et al., J. Chrom. 676: 139-153 (1994).
[B] pH 7 values: Monera et al., J. Pept. Sci. 1: 319-329 (1995).

7

Higher-order Structure



Primary protein structure
is sequence of a chain of amino acids

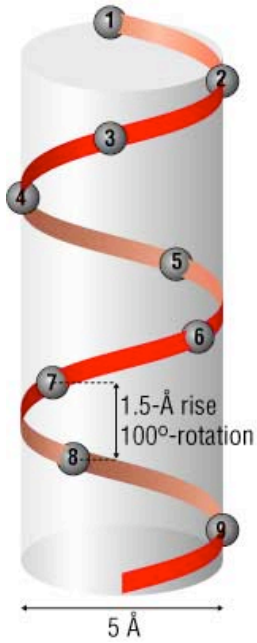
Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds

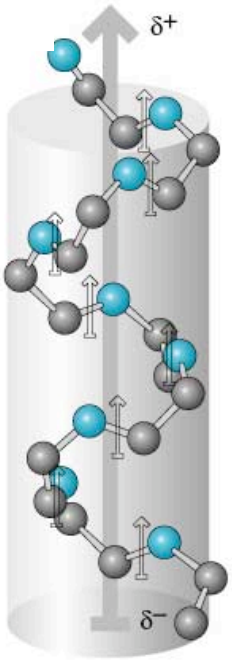
Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.

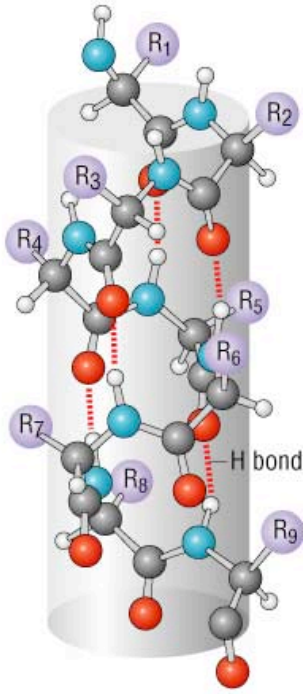
Quaternary protein structure
is a protein consisting of more than one amino acid chain.

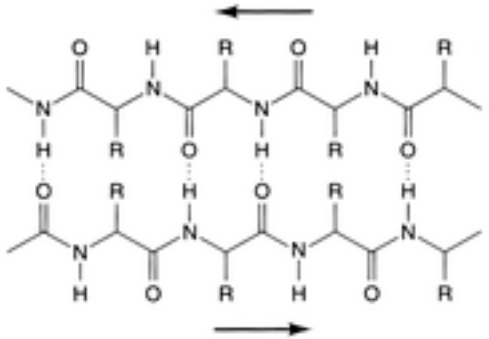
Secondary Structure: Loops, α-helices, β-strands/sheets

α-helix

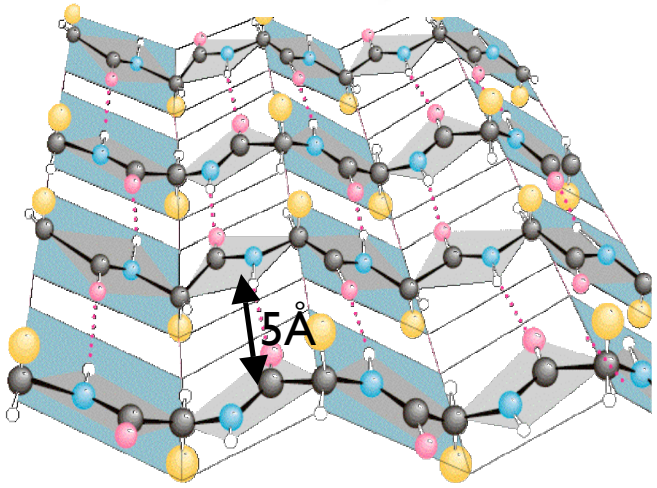








β-strand

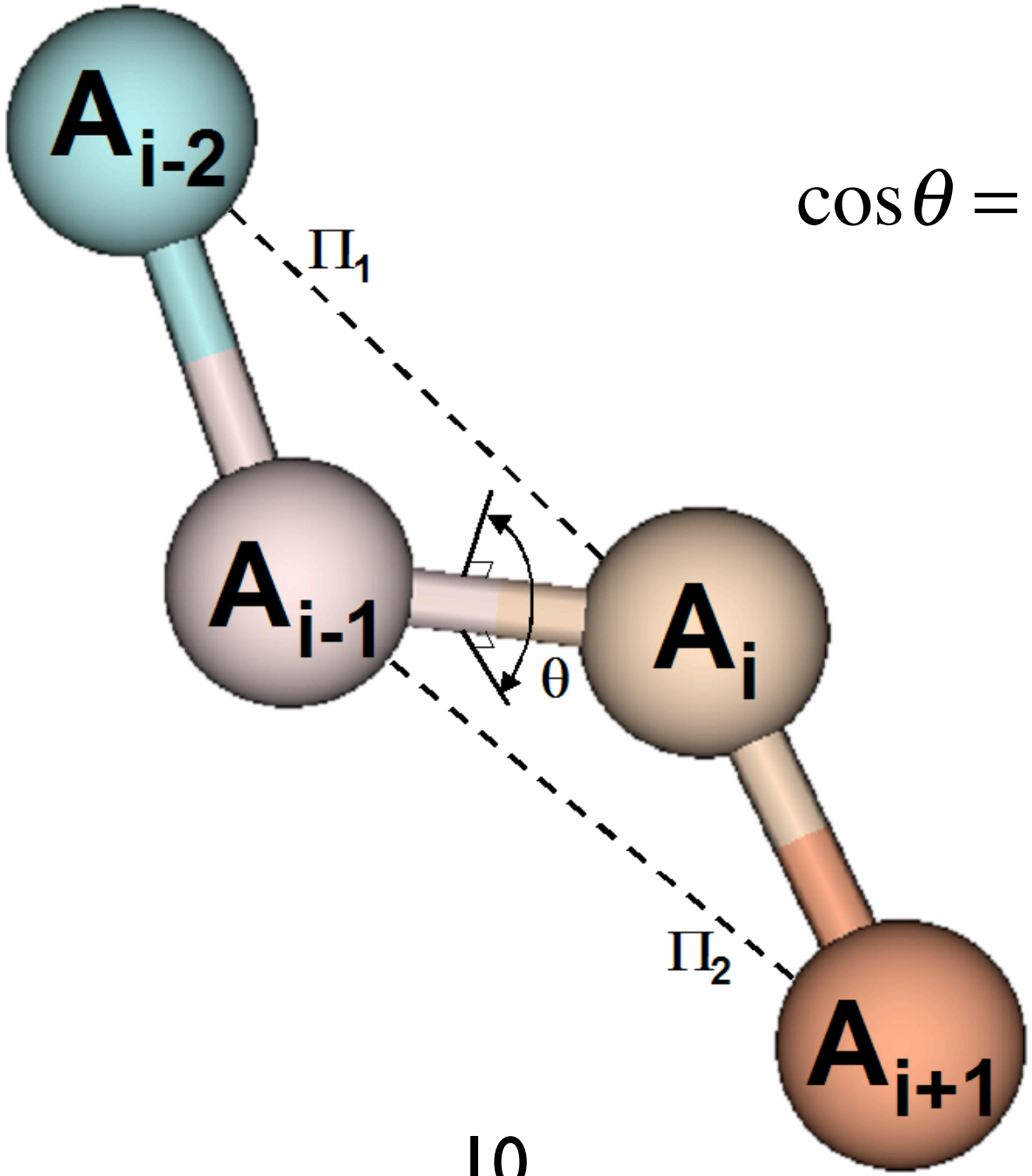


β-sheet

- Right-handed; three turns
- Vertical hydrogen bonds between NH_2 (teal/white) backbone group and C=O (grey/red) backbone group four residues earlier in sequence
- Side chains (R) on outside; point upwards toward NH_2
- Each amino acid corresponds to 100°, 1.5Å, 3.6 amino acids per turn
- $(\phi,\psi)=(-60°,-45°)$
- α-helix propensities: Met, Ala, Leu, Glu

- 5-10 residues; peptide backbones fully extended
- NH (blue/white) of one strand hydrogen-bonded to C=O (black/red) of another strand
- C_α ,side chains (yellow) on adjacent strands aligned; side chains along single strand alternate up and down
- $(\phi,\psi)=(-135°,-135°)$
- β-strand propensities: Val, Thr, Tyr, Trp, Phe, Ile

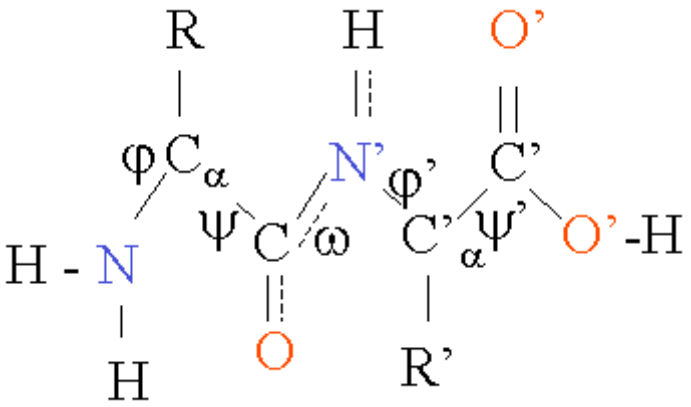
Backbonde Dihedral Angles



$$\cos\theta = \hat{\pi}_1 \bullet \hat{\pi}_2$$

Ramachandran Plot: Determining Steric Clashes

Backbone
dihedral angles

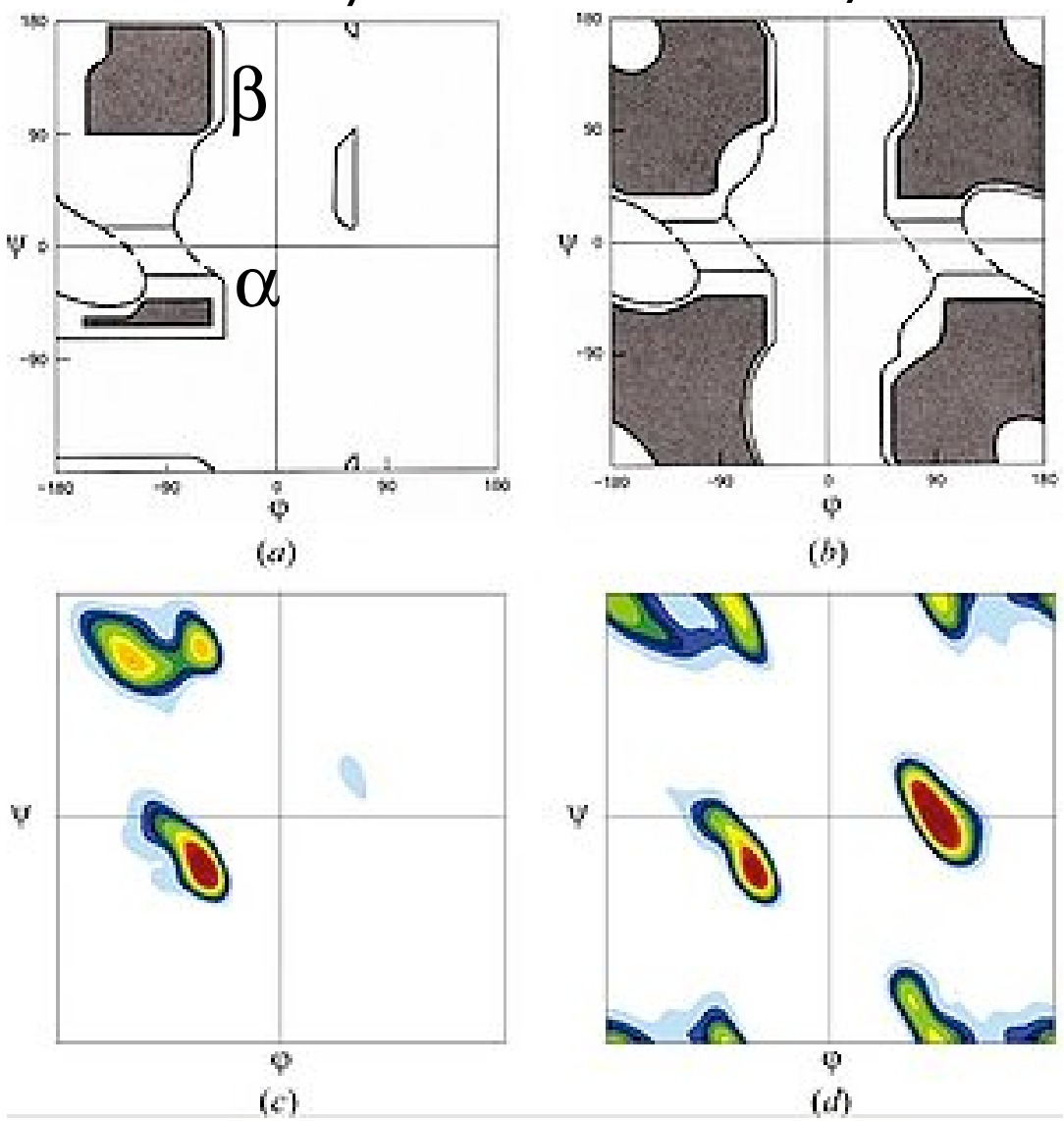


4 atoms define dihedral angle:

$CC_\alpha NC$	ϕ
$C_\alpha N\ CC_\alpha$	ω=0,180°
$N\ CC_\alpha N$	ψ

Non-Gly

Gly



theory

PDB

vdW radii

< vdW radii

backbone flexibility

How can structures from PDB exist outside Ramachadran bounds?

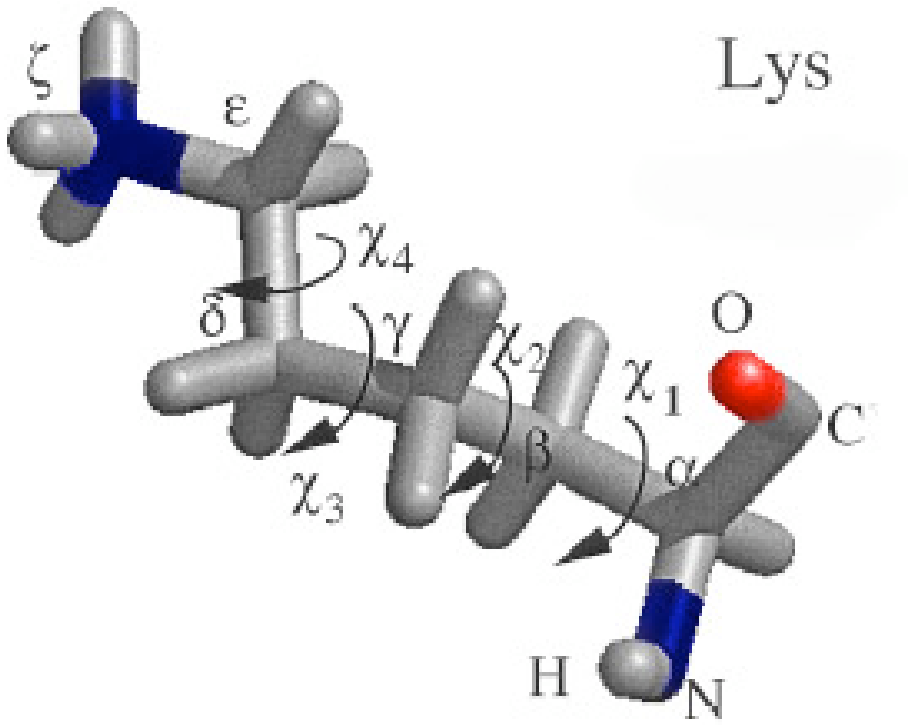
- Studies were performed on alanine dipeptide
- Fixed bond angle (θ=110°) [105°,115°]

J. Mol. Biol. (1963) 7, 95-99

Side-Chain Dihedral Angles

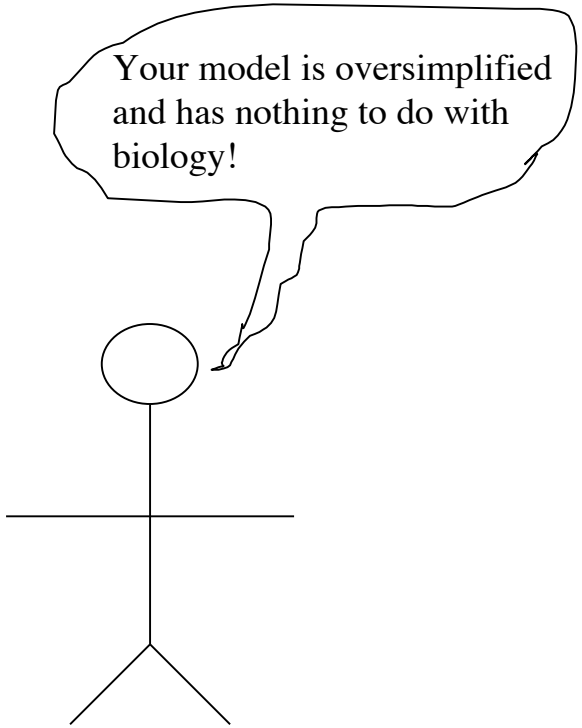
χ_4: Lys, Arg

χ_5: Arg

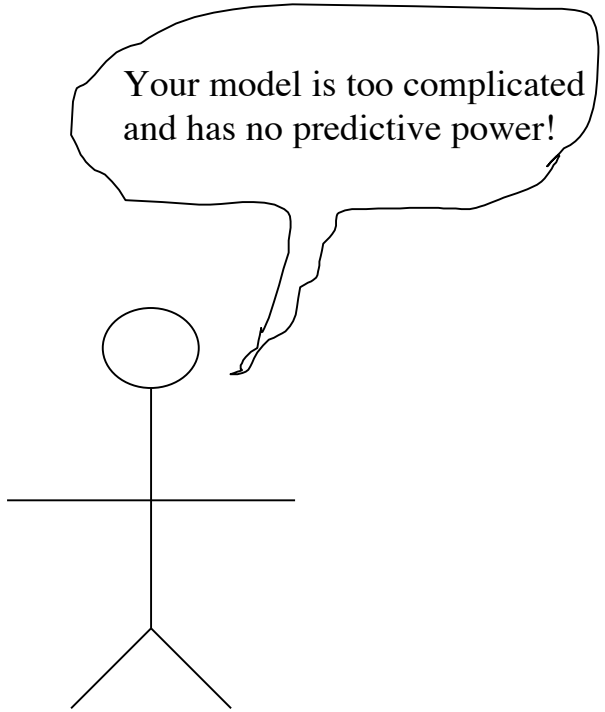


Side chain: C_α-CH_2-CH_2-CH_2-CH_2-NH_3

Use $NC_\alpha C_\beta C_\gamma C_\delta C_\varepsilon N_\zeta$ to define χ_1, χ_2, χ_3, χ_4



Molecular biologist



Biological Physicist

Possible Strategies for Understanding Protein Folding

- For all possible conformations, compute free energy from atomic interactions within protein and protein-solvent interactions; find conformation with lowest free energy…e.g using all-atom molecular dynamics simulations

  Not possible?, limited time resolution

- Use coarse-grained models with effective interactions between residues and residues and solvent

  General, but qualitative

Why do proteins fold (correctly & rapidly)??

Levinthal's paradox:

For a protein with N amino acids, number of backbone conformations/minima

$$N_c \sim \mu^{2N}$$

μ = # allowed dihedral angles

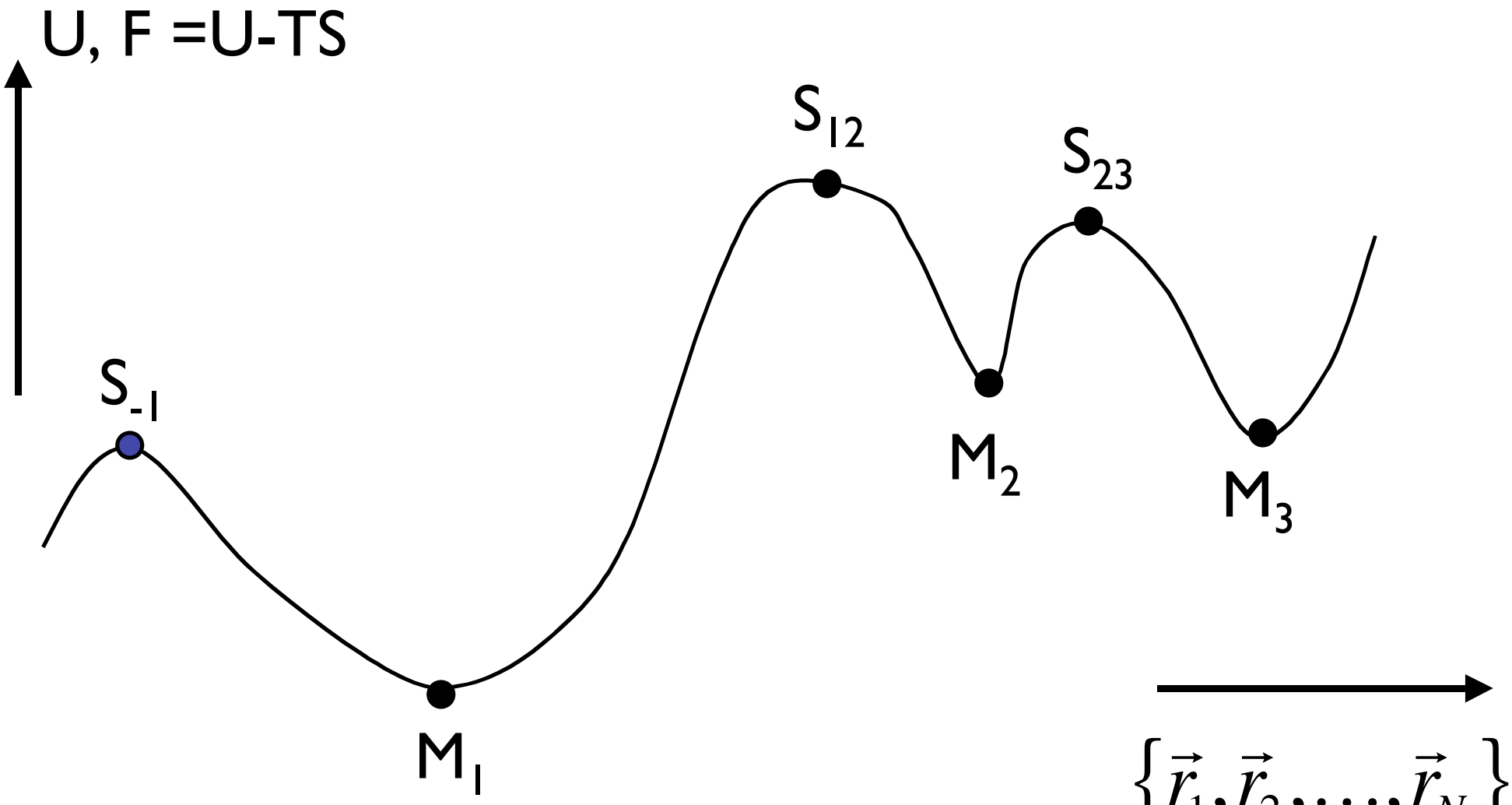
How does a protein find the global optimum w/o global search? Proteins fold much faster.

$N_c \sim 3^{200} \sim 10^{95}$

$\tau_{fold} \sim N_c \, \tau_{sample} \sim 10^{83}$ s vs $\tau_{fold} \sim 10^{-6}$-10^{-3} s

$\tau_{universe} \sim 10^{17}$ s

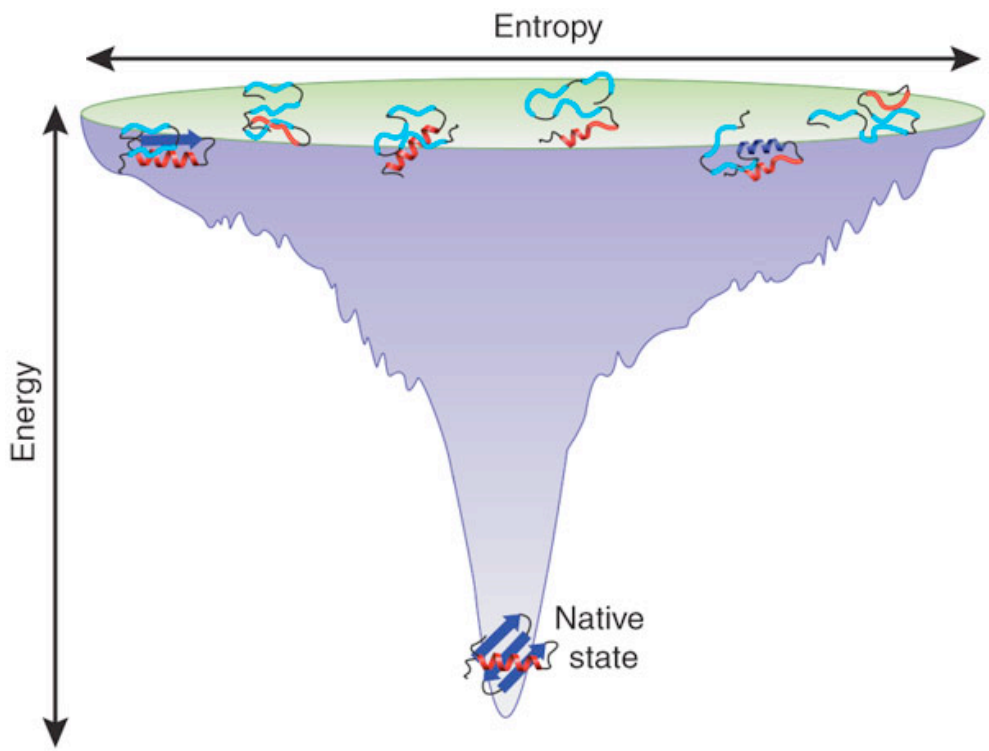
Energy Landscape

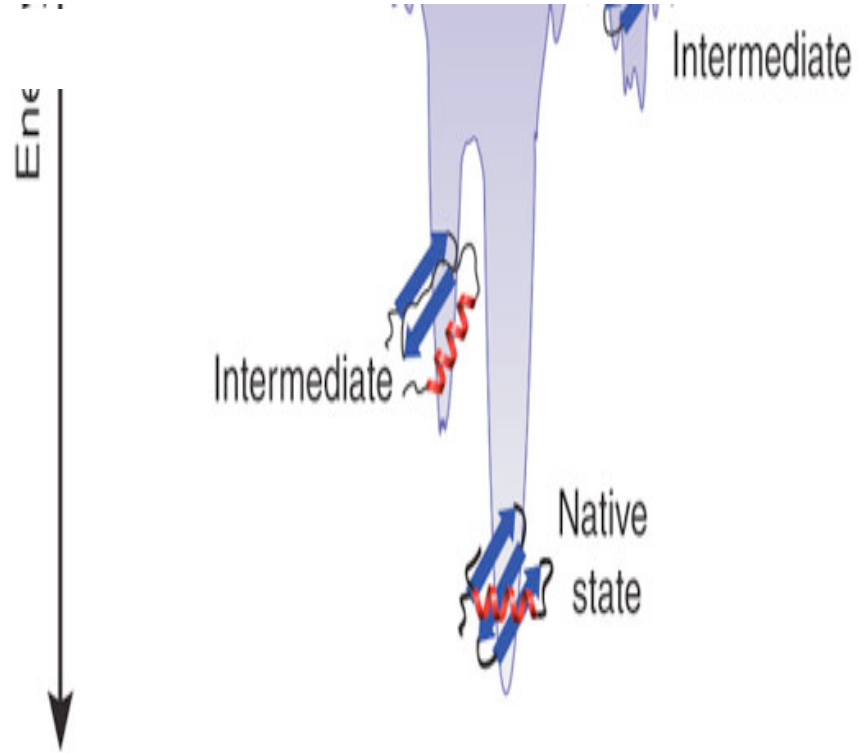


all atomic coordinates; dihedral angles

$$\vec{\nabla} U = 0 \quad \begin{cases} \nabla^2 U > 0 & \text{minimum} \\ \nabla^2 U = 0 & \text{saddle point} \\ \nabla^2 U < 0 & \text{maximum} \end{cases}$$

Roughness of Energy Landscape



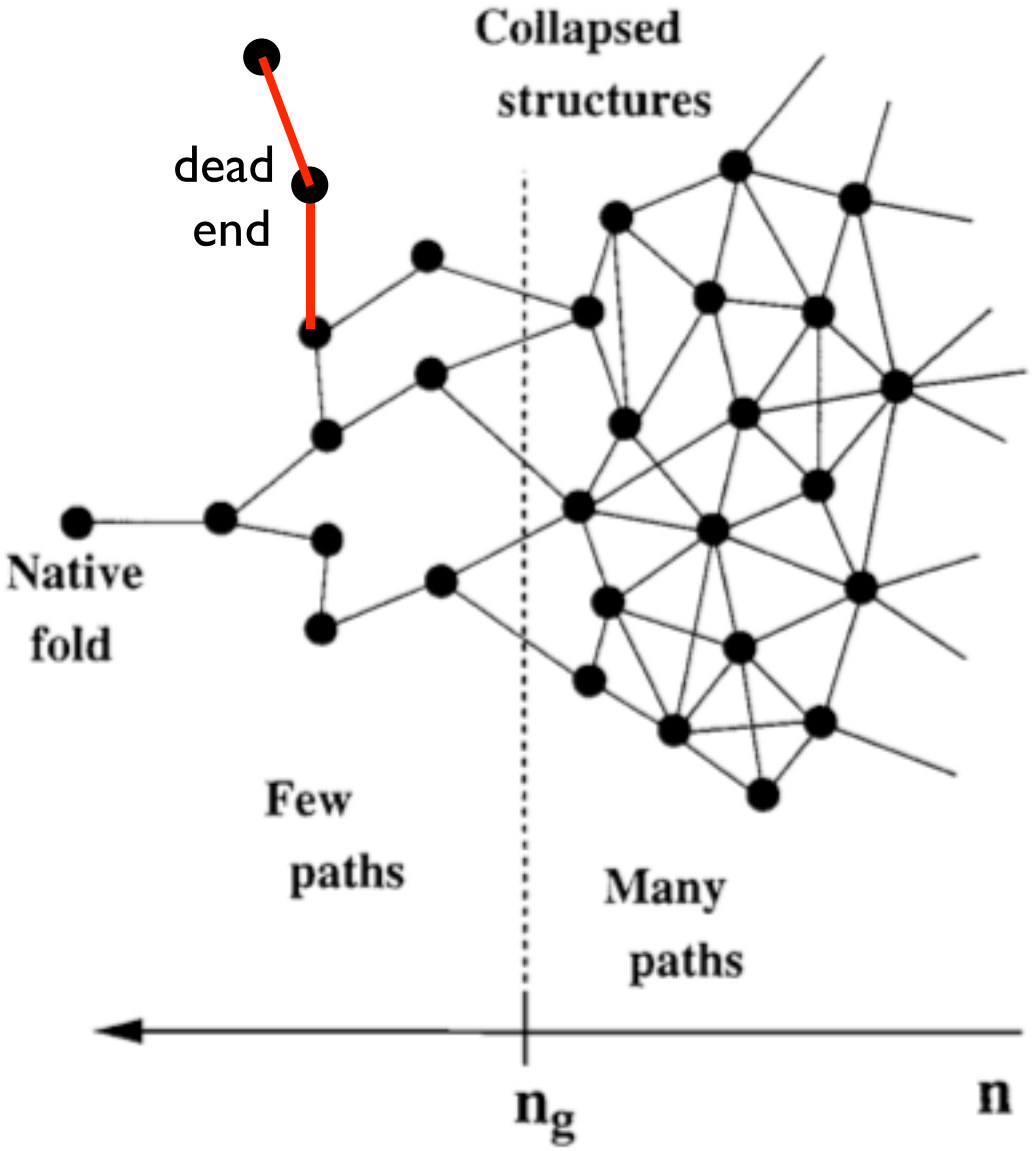


smooth, funneled

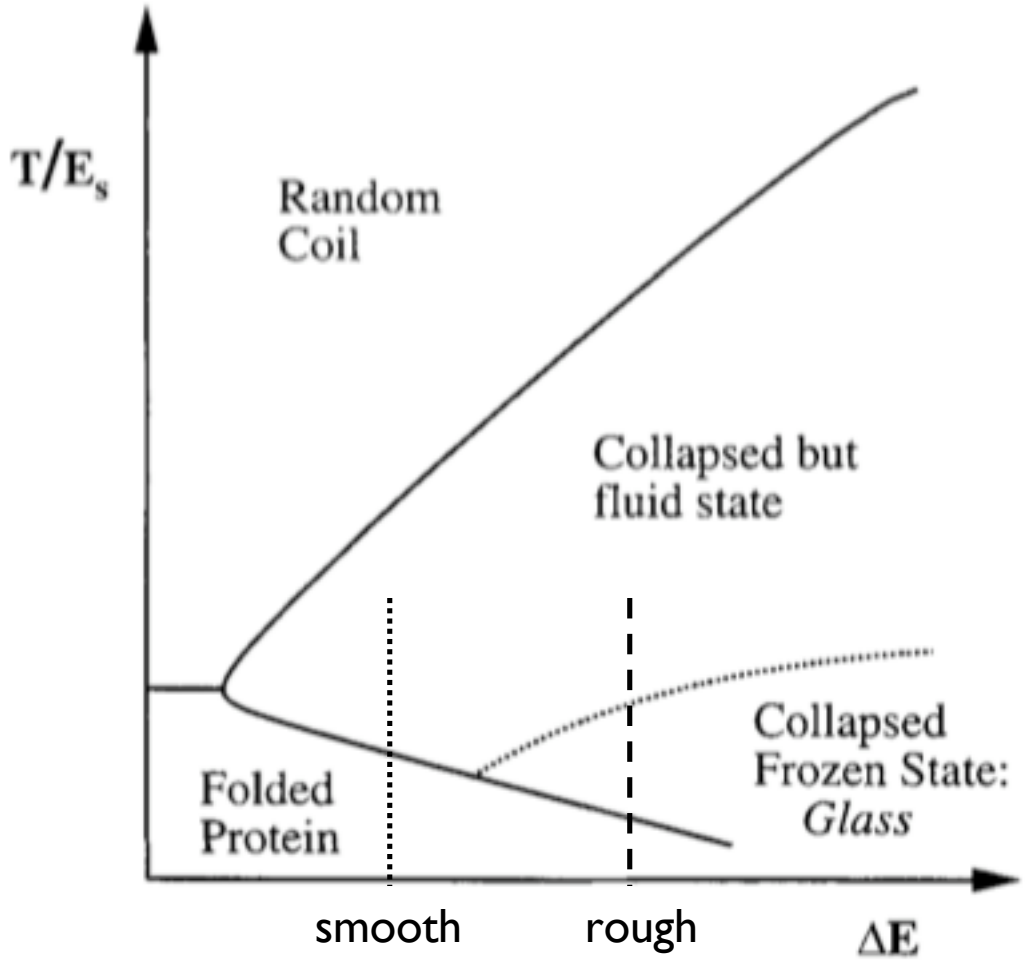
rough

(Wolynes et. al. 1997)

Folding Pathways



Folding Phase Diagram



Open Questions

- What differentiates the native state from other low-lying energy minima?

- How many low-lying energy minima are there? Can we calculate landscape roughness from sequence?

- What determines whether protein will fold to the native state or become trapped in another minimum?

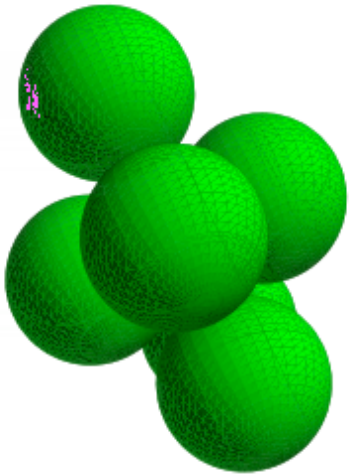
- What are the pathways in the energy landscape that a given protein follows to its native state?

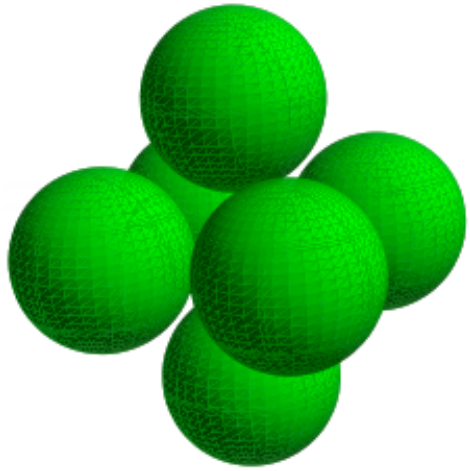
NP Hard Problem!

Digression---Number of Energy Minima for Sticky Spheres

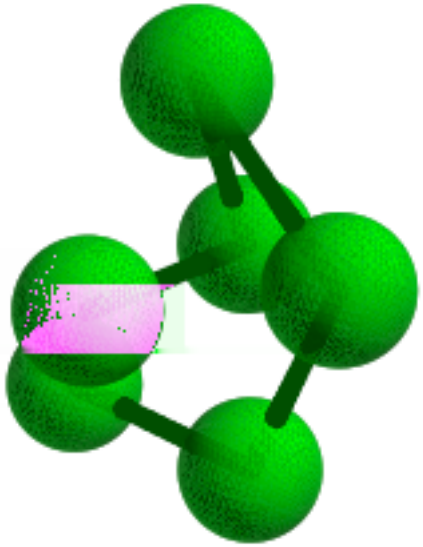
N_m	N_s	N_p
4	1	1
5	1	6
6	2	50
7	5	486
8	13	5500
9	52	49029
10	-	-

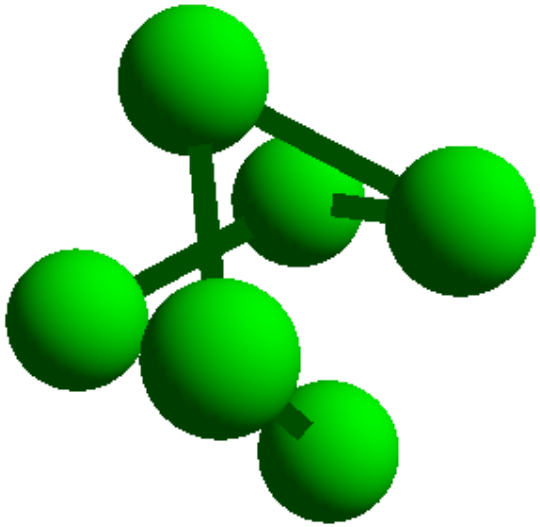
sphere packings





polymer packings





$N_s \sim \exp(aN_m)$;

$N_p \sim \exp(bN_m)$ with $b>a$