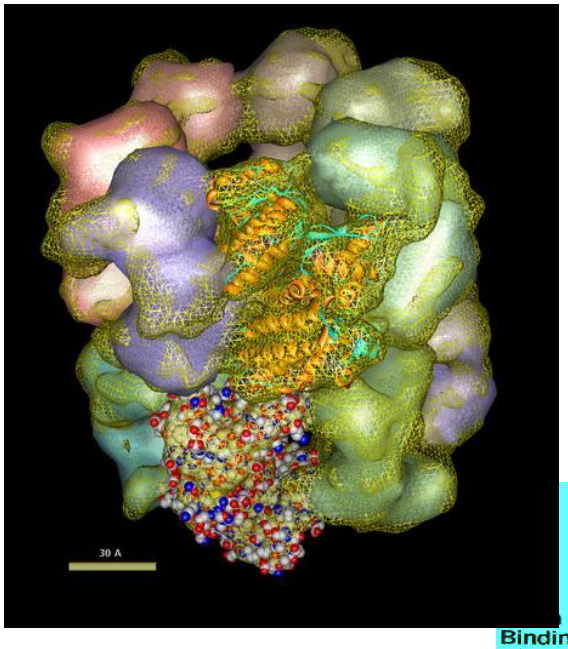
Protein Folding

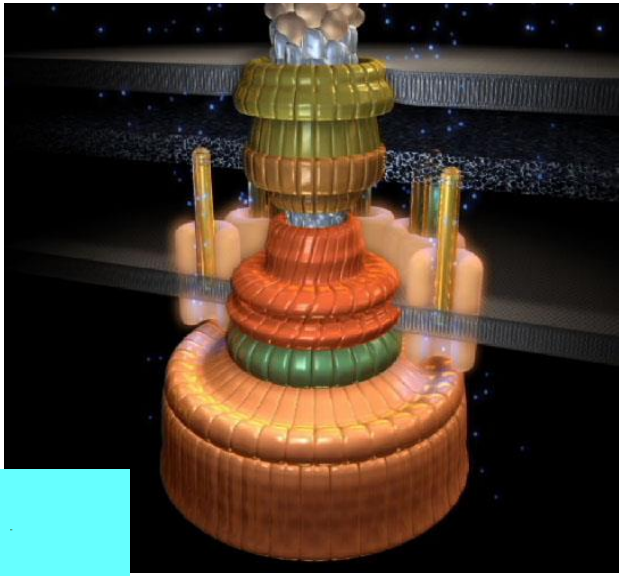
Proteins:

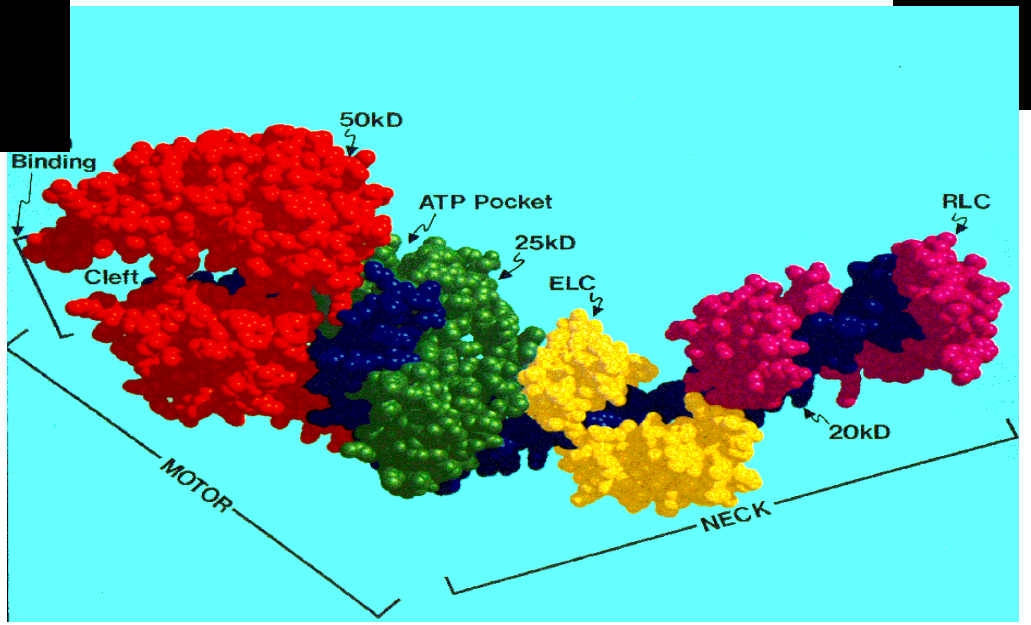
- Proteins are biopolymers that form most of the cellular machinery
- The function of a protein depends on its 'fold' – its 3D structure



Chaperone

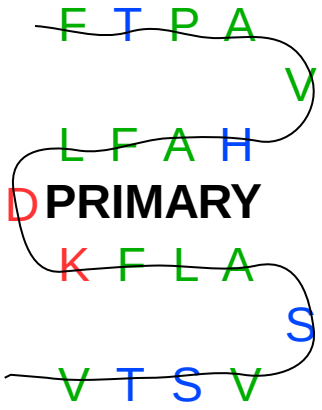
Motor



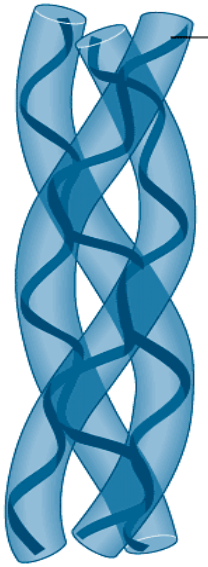


Walker

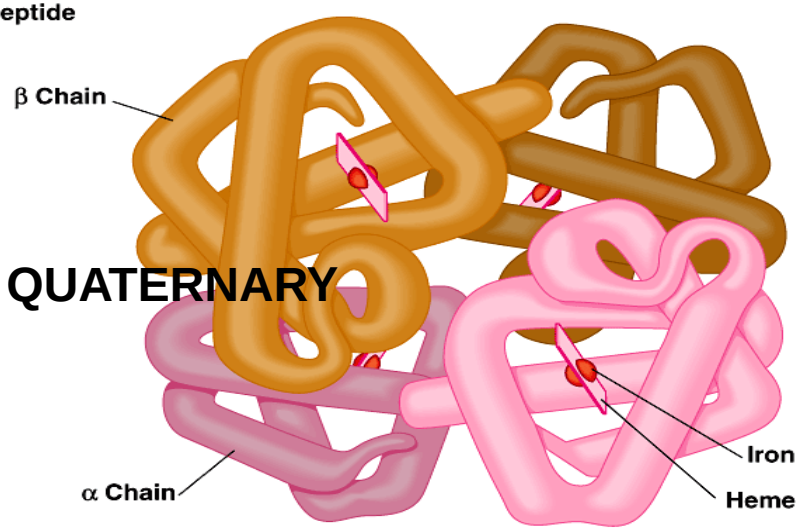
Levels of Folding:



PRIMARY

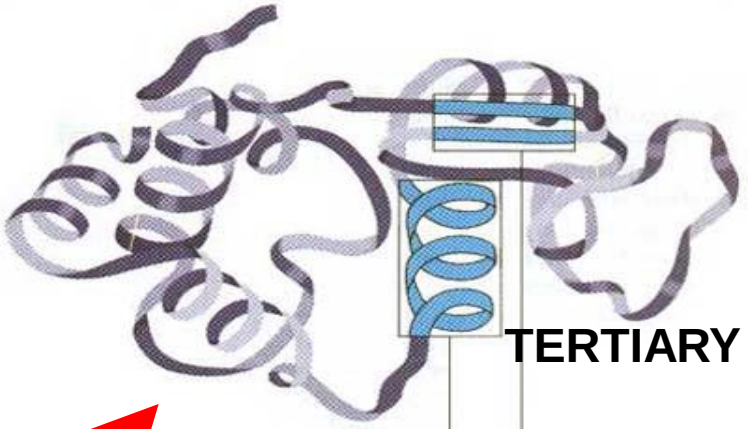


(a) Collagen

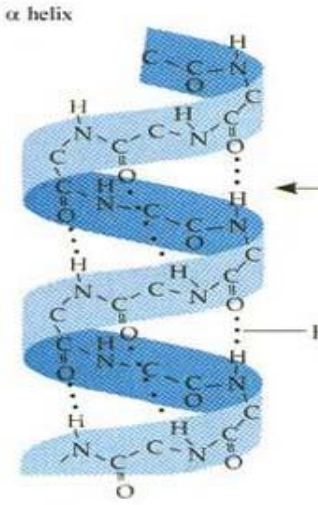


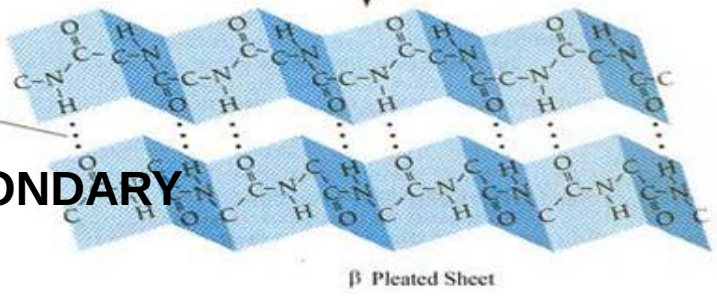
QUATERNARY

(b) Hemoglobin



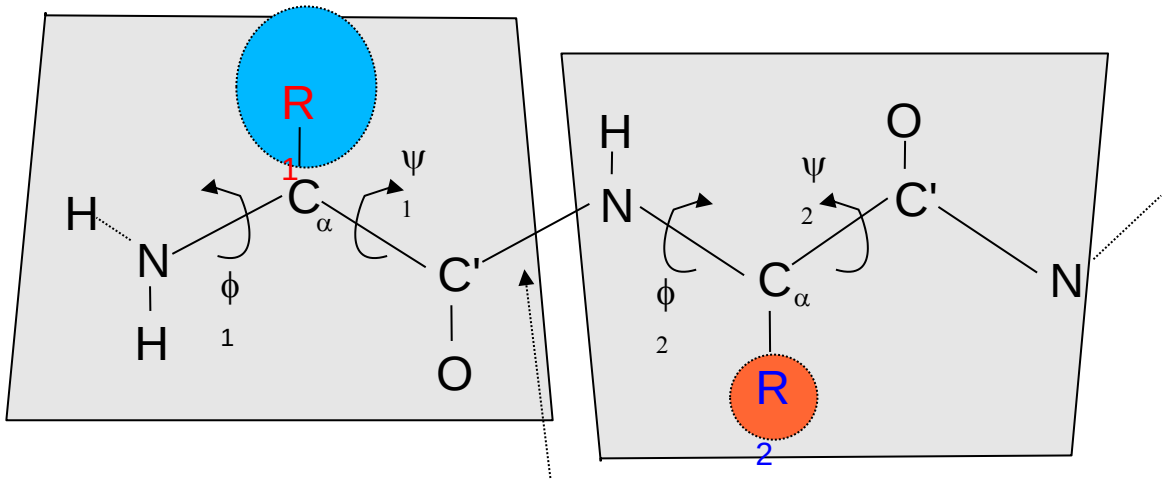
TERTIARY





SECONDARY

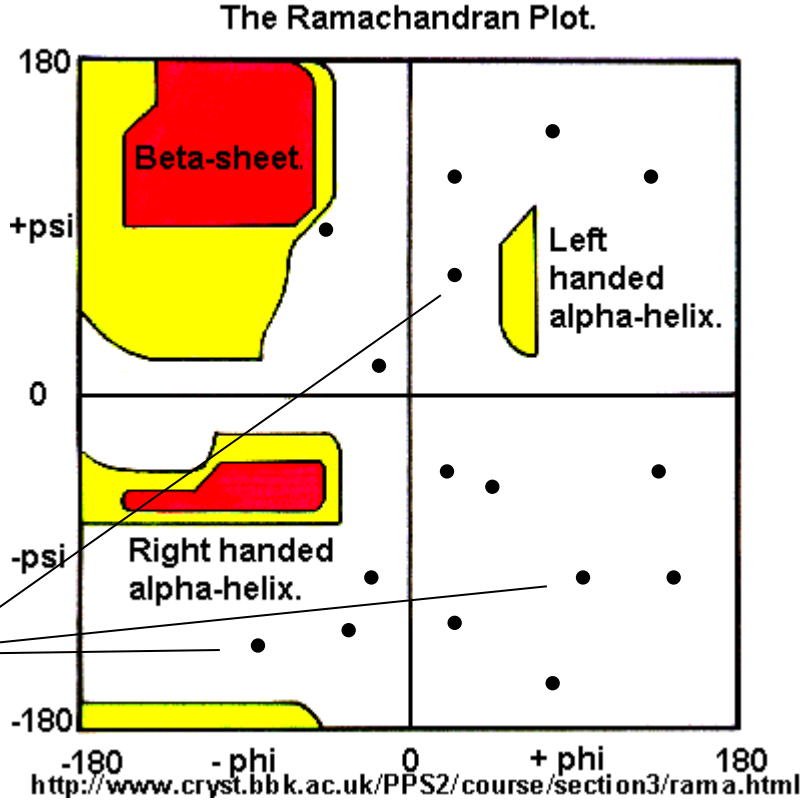
The Backbone



Peptide Bond

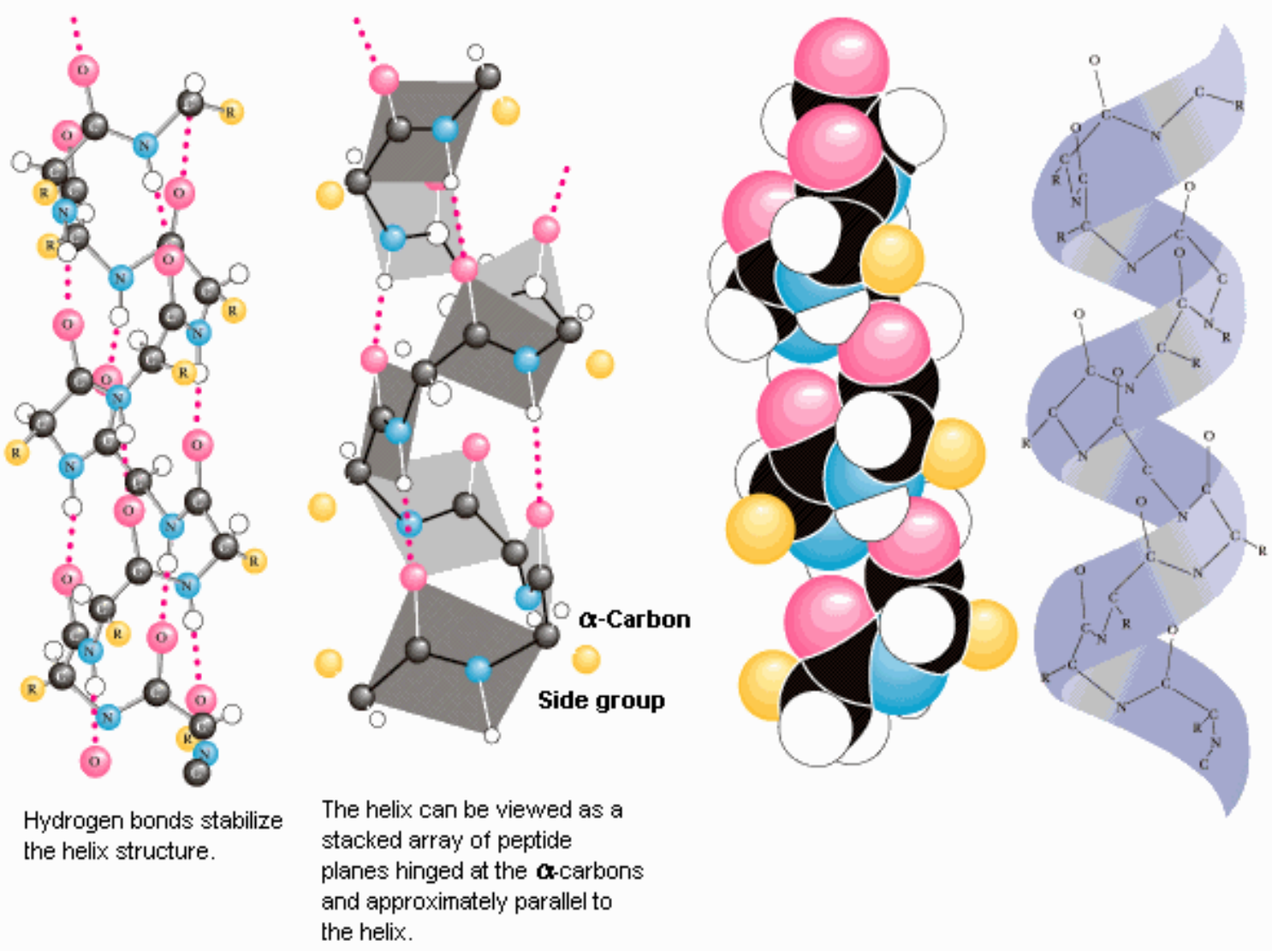
- Amino acids linked together by peptide bonds

Steric constraints lead
only to a subset of possible angles
--> Ramachandran plot



Glycine residues
can adopt many
angles

α Helices

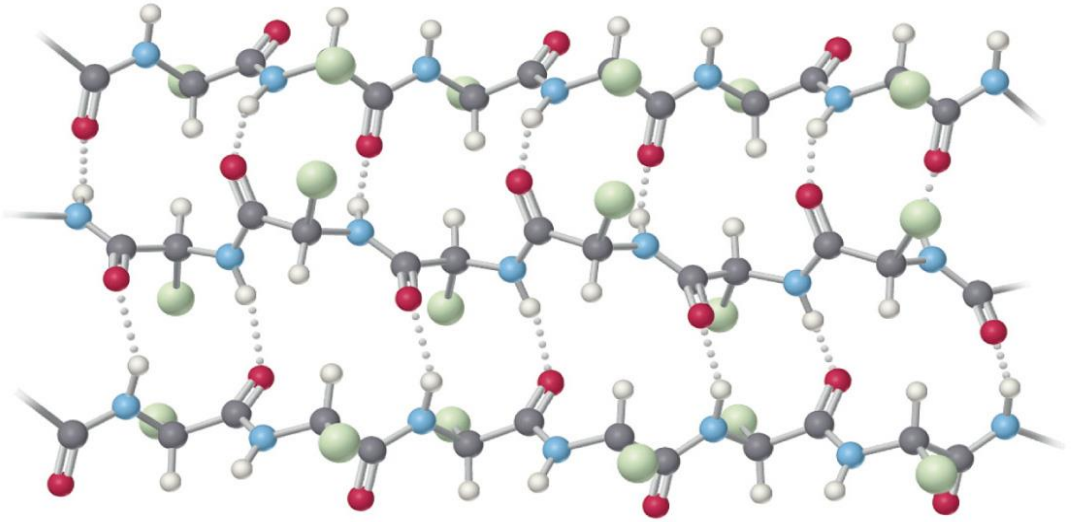


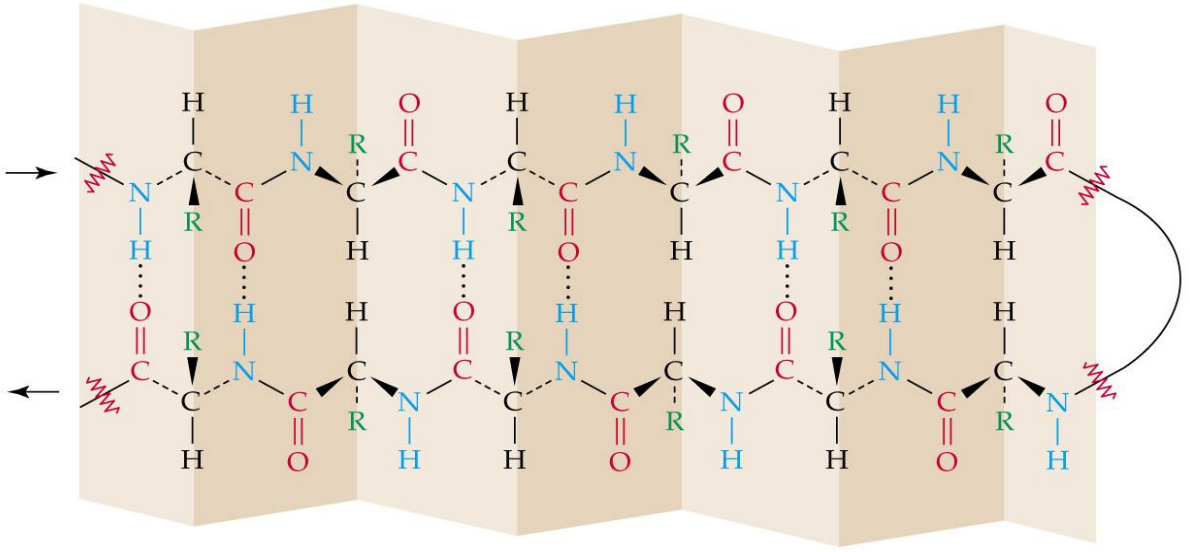
Hydrogen bonds stabilize the helix structure.

The helix can be viewed as a stacked array of peptide planes hinged at the α-carbons and approximately parallel to the helix.

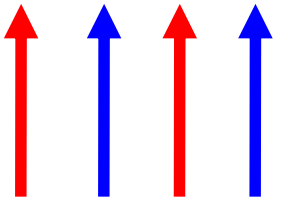
3.6 residues/turn

β Sheets

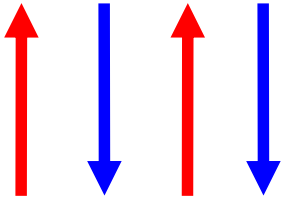




parallel sheet



anti-parallel sheet



other topologies possible
but much more rare

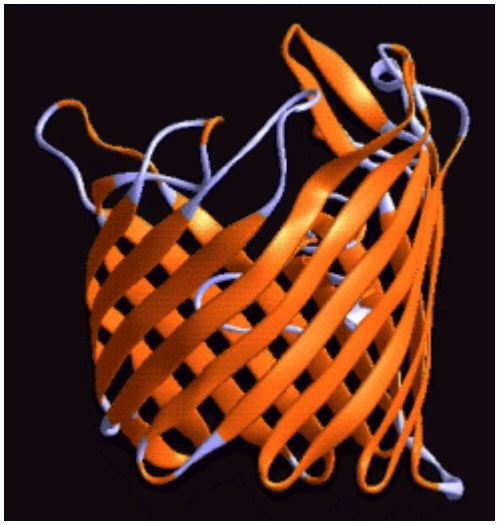
Classes of Folds:

- There are three broad classes of folds: α, β and α+β
- as of today, 103000 known structures --> 1100 folds (SCOP 1.75)

alpha class

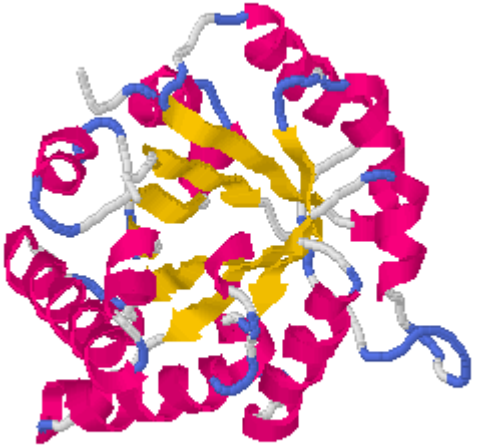
myoglobin – stores oxygen in muscle tissue

beta class



streptavadin – used a lot in biotech, binds biotin

alpha+beta class



TIM barrel – 10% of enzymes adopt this fold, a great template for function

Databases:

SWISSPROT:

contains sequence data of proteins – 100,000s of sequences

Protein Data Bank (PDB):

contains 3D structural data for proteins – 100,000 structures, x-ray & NMR

SCOP:

classifies all known structures into fold classes ~ 1100 folds

Protein Folding:

amino acid sequence $\underset{\text{DESIGN}}{\overset{\text{FOLDING}}{\rightleftharpoons}}$ structure

- naturally occurring sequences seem to have a unique 3D structure

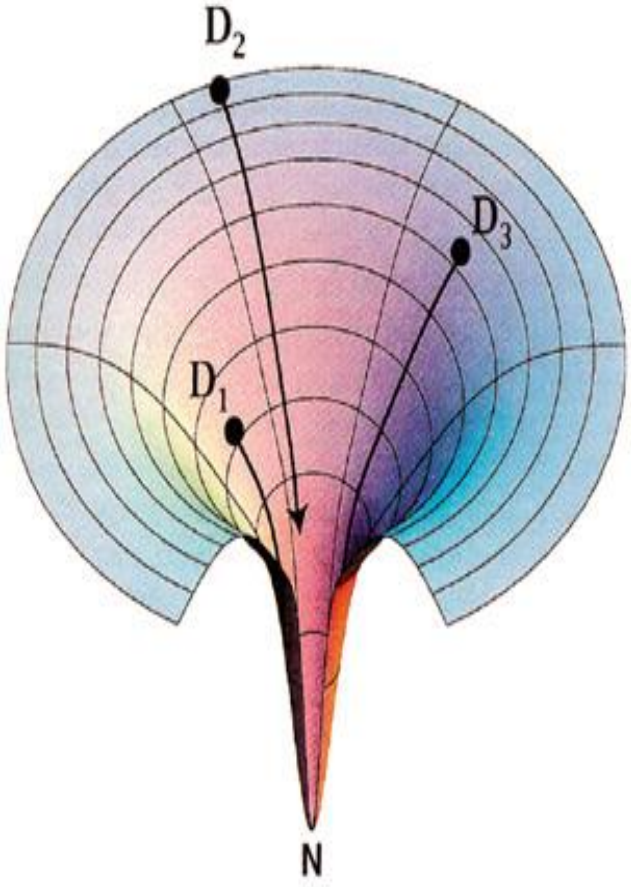
Levinthal paradox: if the polymer doesn't search all of conformation space, how on earth does it find its ground state, and in a reasonable time?

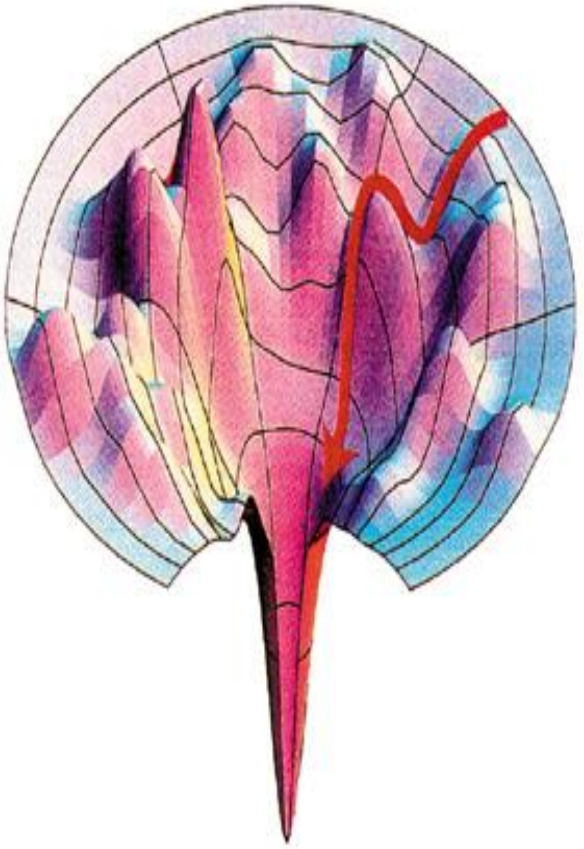
if 2 conformation/residue & dt ~ 10^{-12} -> t=10^{25} years for a protein of L = 150!!!

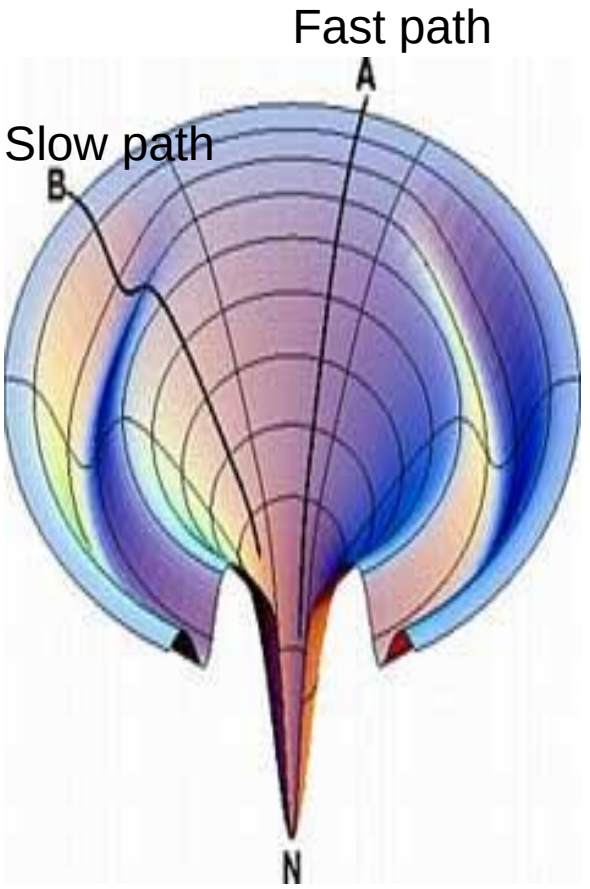
Reality: t = .1 to 1000 s

How do we resolve the paradox?

Paradox Resolved: Funnels







- there are multiple folding pathways on the energy landscape – slow & fast
- If a protein gets stuck (misfolded) there are chaperones to help finish the fold

Factors Influencing folding:

Hydrogen bonding:

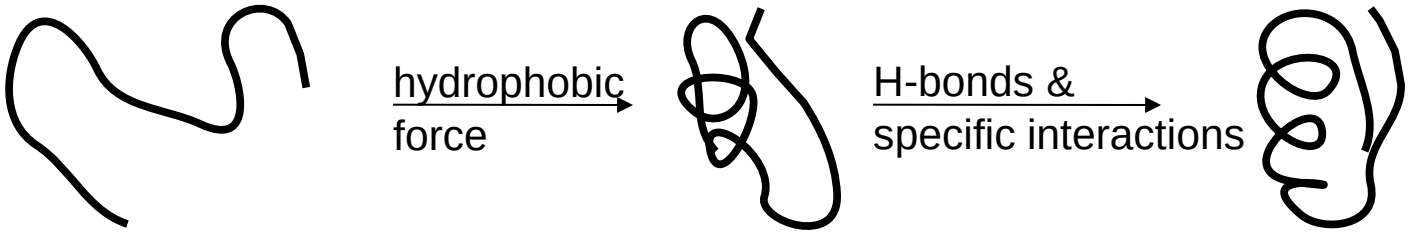
doesn't drive folding since unfolded structure can form H-bonds with H20
drives 2ndary structure formation after compaction

Hydrophobicity:

main driving force
significant energy gain from burying hydrophobic side-chains
leads to much smaller space to search

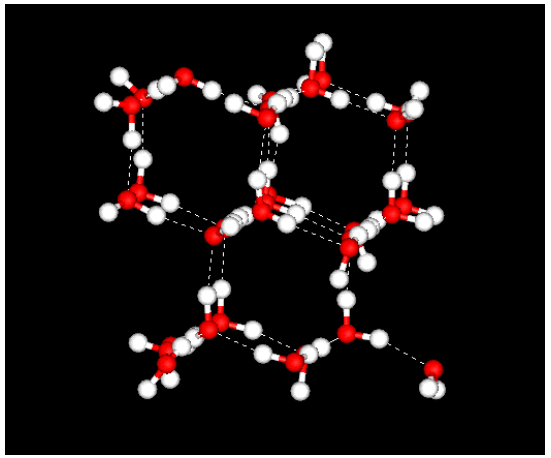
Other interactions:

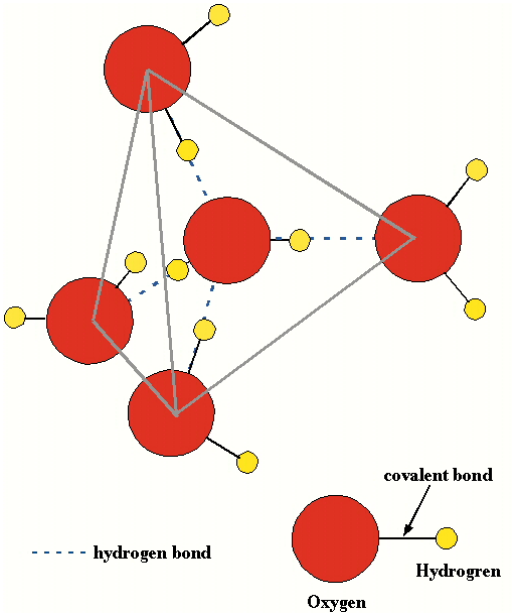
give specificity and ultimately favour final unique state
disulfide bridges = formed between contacting Cystine residues
salt-bridges = formed between contacting -ve and +ve charged residues
secondary structure preferences = from entropy



More on Hydrophobicity:

- Hydrophobicity is an entropic force – water loses entropy due to the presence of non-polar solvent



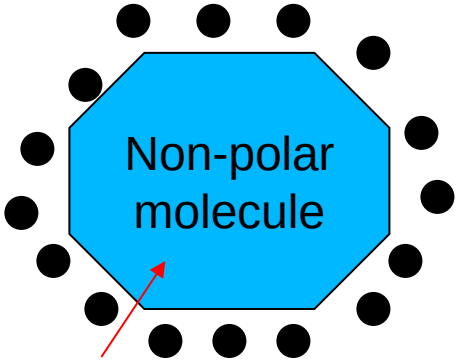


H20 molecules form a tetrahedral structure, and there are 6 hydrogen-bonding Orientations/H20

When a non-polar molecule occupies a vertex → reduces to only 3 orientations

$dS = k \ln 3 - k \ln 6 = - k \ln 2$ → $dG = + kT \ln 2$ costs energy to dissolve

Hydrophobicity and Packing:



Area = A

A non-polar object with area A will disrupt
The local H20 environment

For 1 nm^2 of area ~ 10 H20 molecules are affected

So hydrophobic cost per unit area

$$\gamma = 10\ k\ T\ \ln 2/nm^2 = 7\ k\ T\ /\ nm^2$$

Hydrophobic energy cost = G = γ A

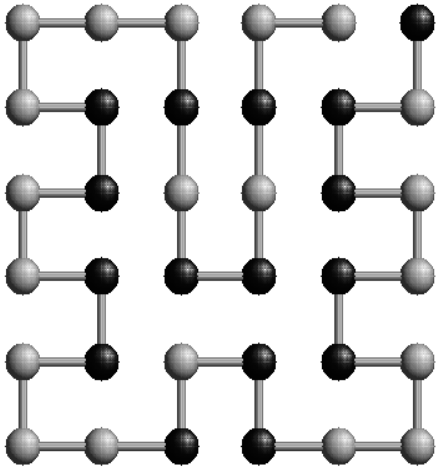
For an O2 molecule in H20, A = 0.2 nm2 so G ~ 1 kT. So O2 easily dissolves in H20

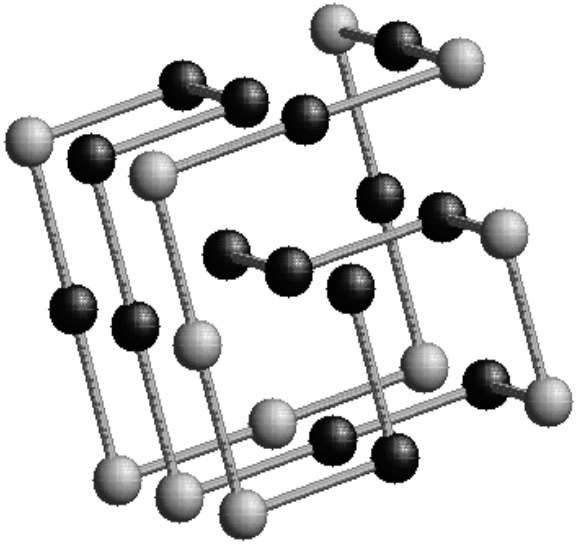
For an octane molecule, G ~ 15 kT, so octane will aggregate so as to minimize the combined exposed area

Simple Models of Folding: Getting at the big picture

- folding proteins in 3D with full atomic detail is HARD!!! essentially unsolved
  --> study tractable models that contain the essential elements

SIMPLE STRUCTURE MODEL = LATTICE MODELS:





- enumerate all compact structures that completely fill a 2D or 3D grid
- can also study non-compact structures by making larger grid

Simple Energy functions:

H-P Models:

- amino acids come in only two types, H = hydrophobic, P = polar
- interactions: H-H, H-P & P-P with $E_{PP} > E_{HP} > E_{HH}$

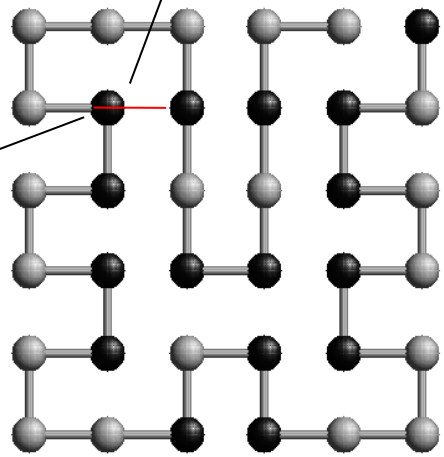
Energy = $\Sigma E_{ij} \Delta(r_i - r_j)$

- could use full blown 20 x 20 E_{ij} matrix = Miyazawa-Jernigan matrix

Solvation Models:

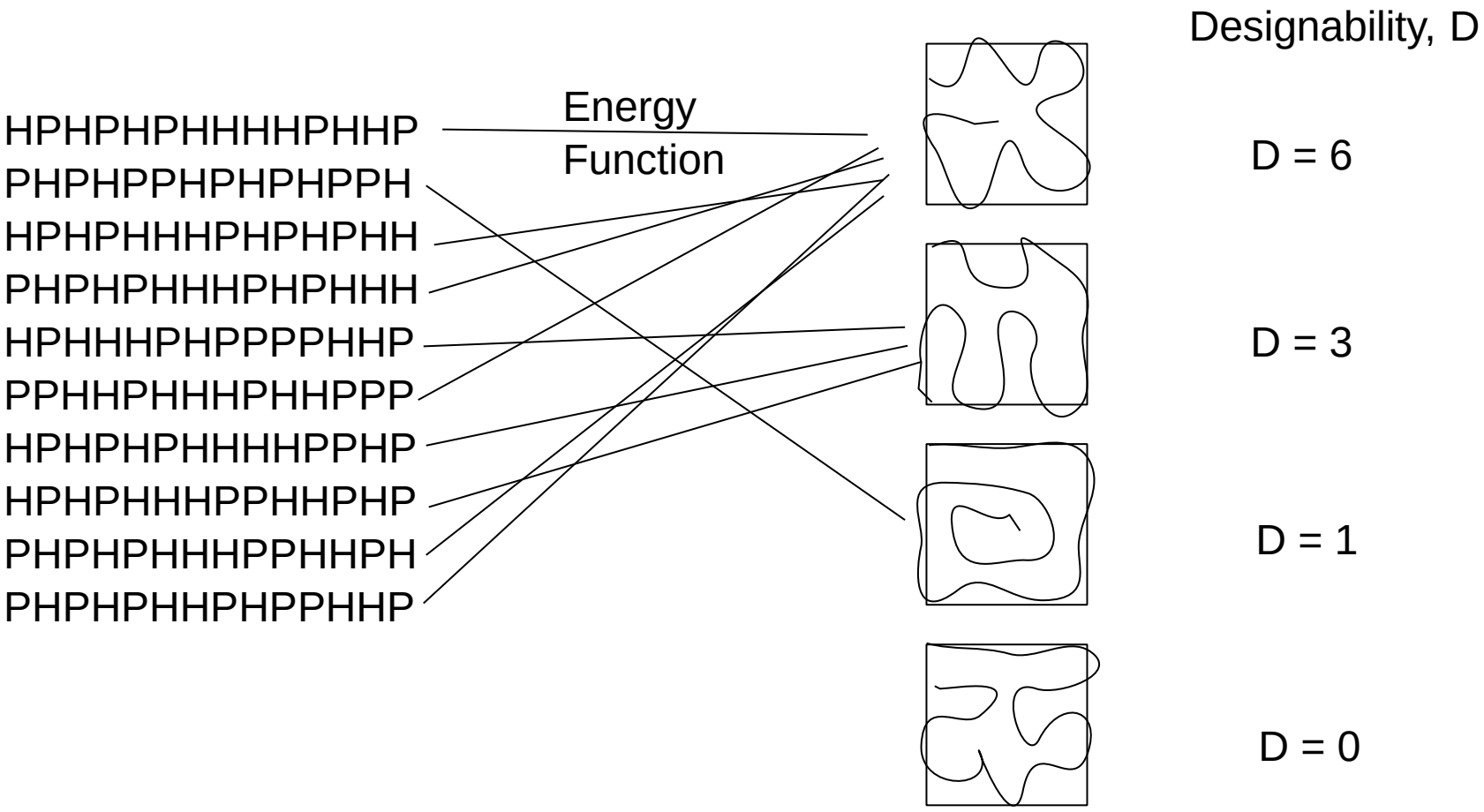
- energy is gained for burying hyrdophobic residues
- if residue is buried, surface exposure, s = 1
- if residue is exposed, surface exposure, s = 0
- hydrophobicity scale: H: h = -1, P: h= 1
- Energy = $\Sigma h_i s_i$

Ground state structure has the lowest energy for given sequence



Model Results: Designability Principle

- Fold random HP sequences, and determine the ground state for each
- Designability = # of sequences which fold into a given structure

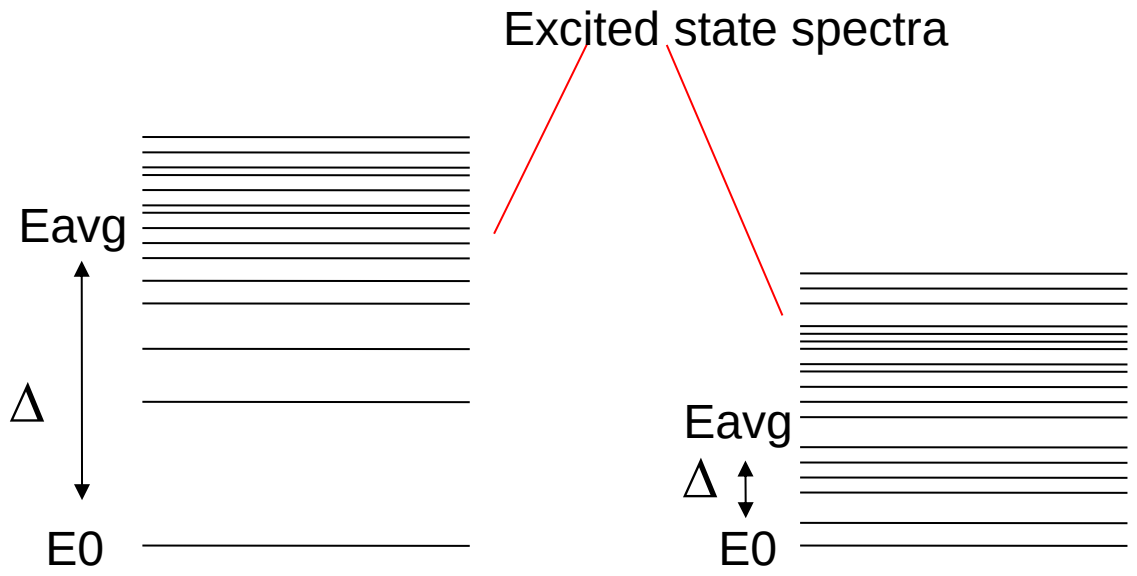


Designability Principle: there are only a few highly designable structure, most structures have very few sequences that fold into them

Thermodynamic Stability

- high designabilty implies mutational stability, does it imply thermodynamic stability?

YES



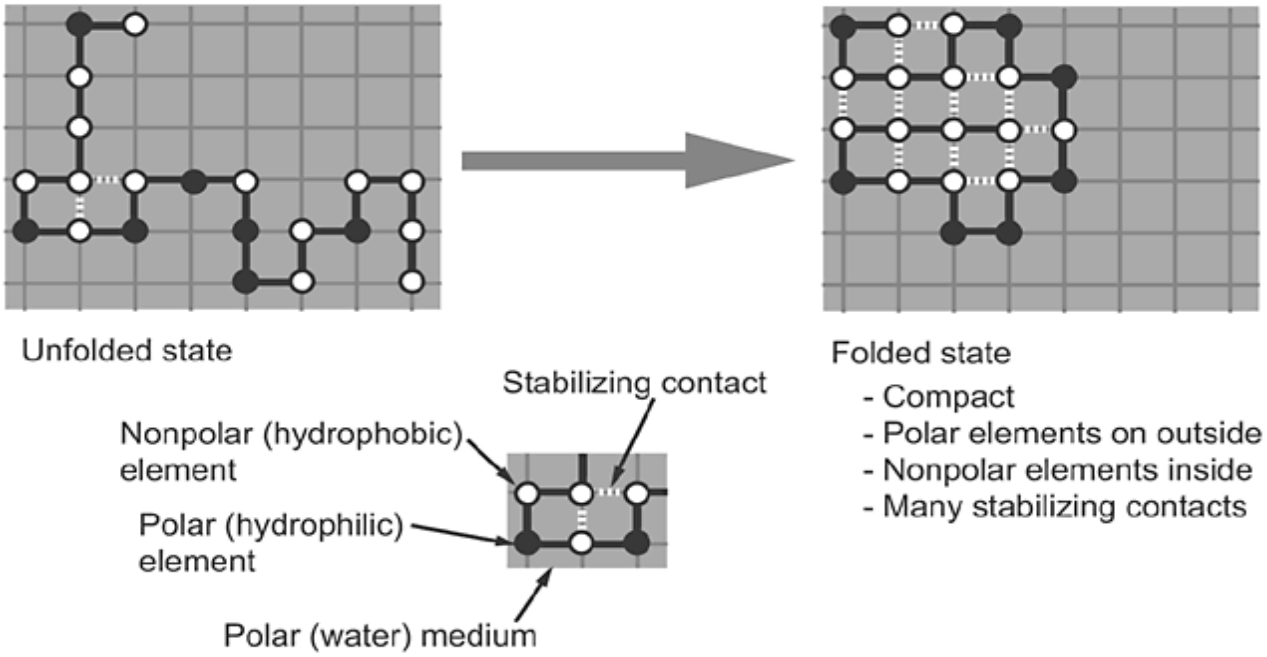
HIGH DESIGN STRUCTURE

LOW DESIGN STRUCTURE

- Highly designable structures are characterized by a large energy gap, Δ

Fast Folding

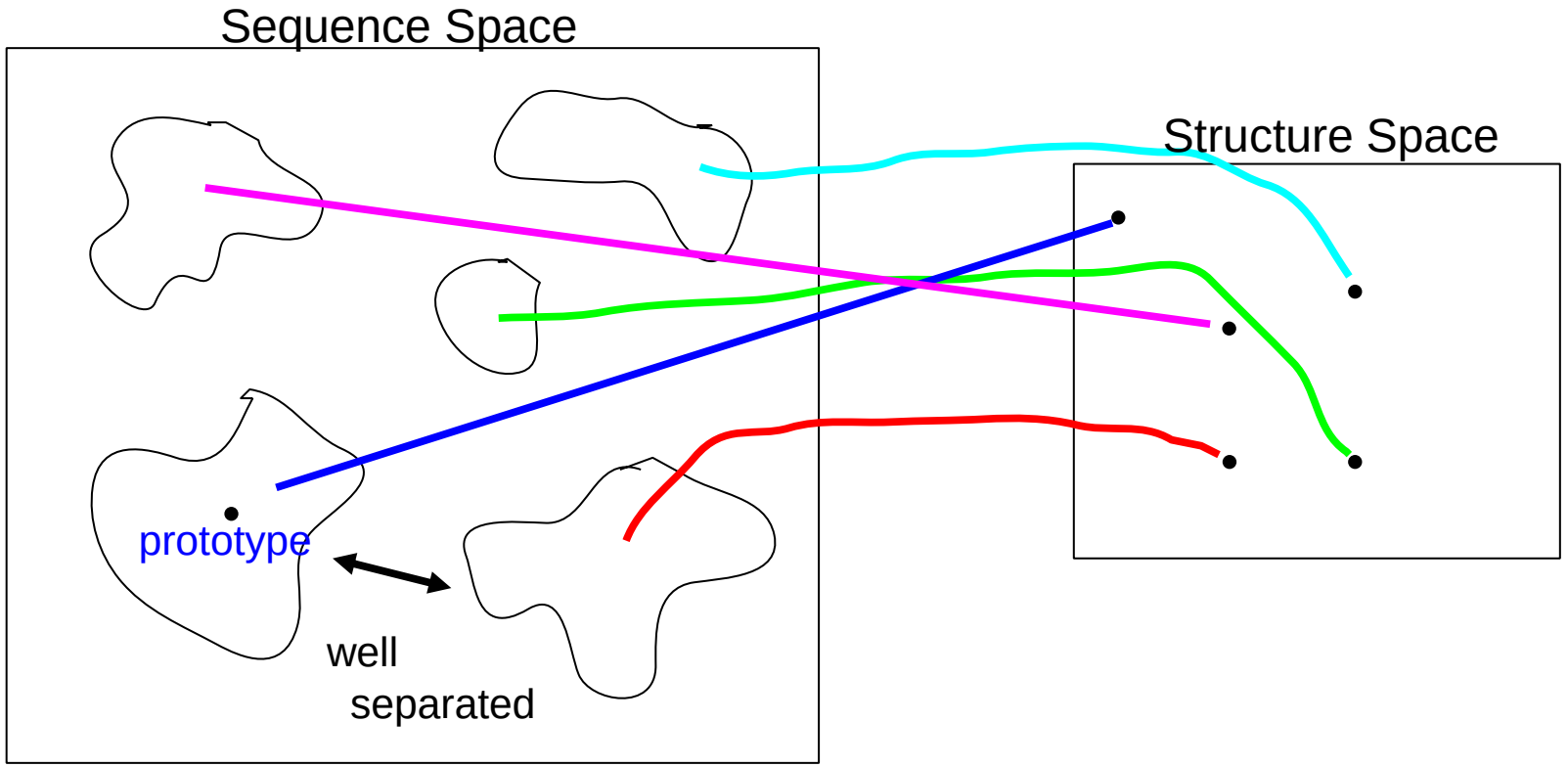
- High designability structures are fast folders, since there are few low lying energy structures to compete with – no kinetic traps
- Low designability structures are slow – have many competing low energy alternatives which act as kinetic traps



- Determine kinetics using Metropolis Monte-carlo

  t ~ # of monte-carlo steps needed to first achieve near native state (90%)

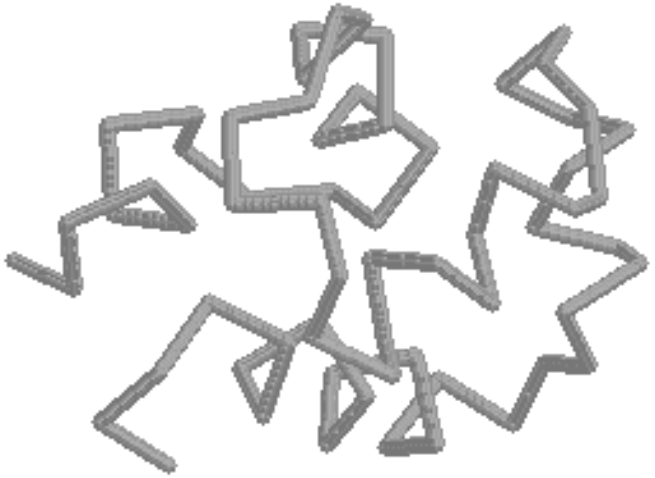
Neutral Networks in Protein Folding:



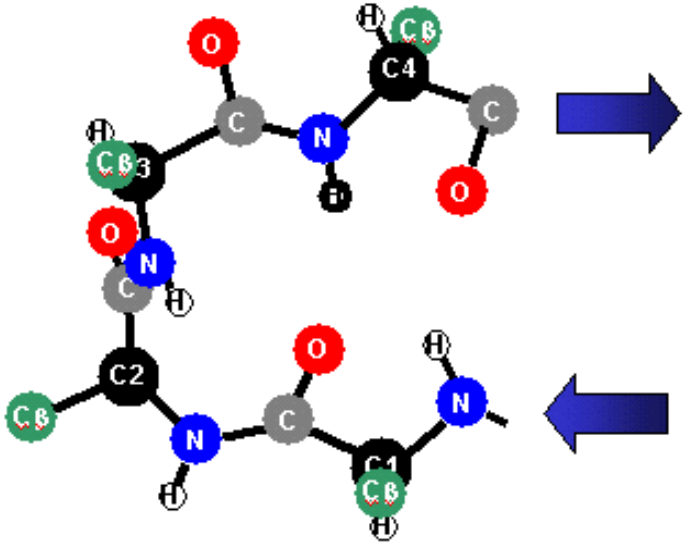
- Just like RNA, designable proteins have well connected neutral networks
- Unlike RNA, these neutral networks are well separated, so they are not space covering
- Prototype sequence tends to have best thermodynamic properties (cluster center)

Protein Folding in the Real World:

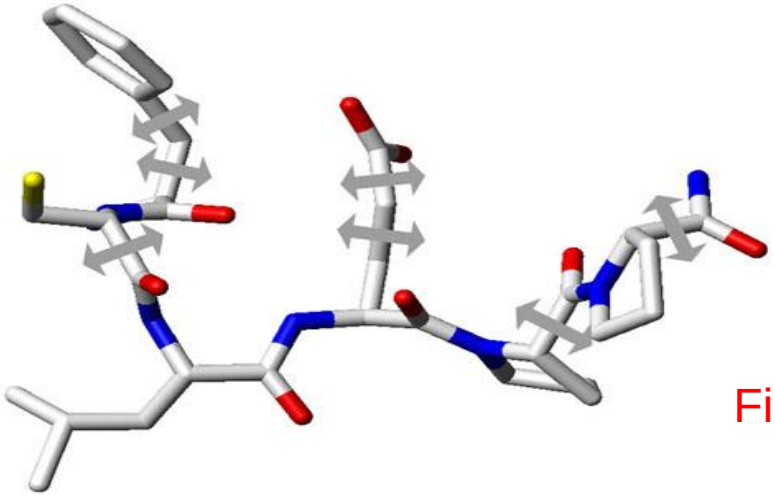
OFF-LATTICE MODELS:



Coarse: just C_α and C_β



Medium: all backbone and C_b



Fine: all atoms and use side chain rotamers

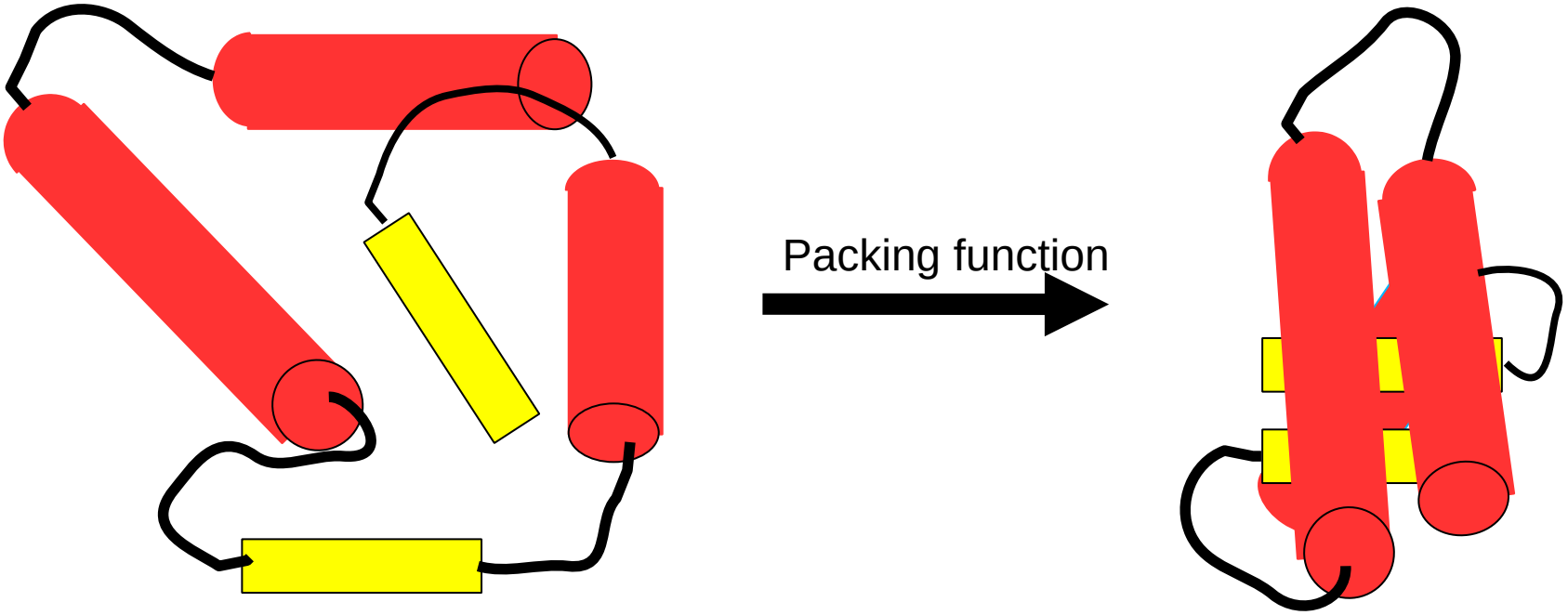
Structure Construction:

Enumerate structures:

- enumerate all structures that are possible using a finite # of (ϕ,ψ) angles
- e.g. 4 pairs, L = 20 --> 4^{20} = 1 x 10^{12} structures!!!

Packing of secondary elements:

- pack together in 3D a fixed set of secondary structural elements
- can go to much larger structures
- must sample the space



Protein Design:

1) Improve natural folds:

give natural proteins new function, stability, kinetics

2) The search for novel folds: for L = 100 --> 100^{20} sequences !!!

There may be sequences that fold into structures not seen in nature

Inverse folding problem: given a structure find a compatible sequence for which the structure is the ground state fold

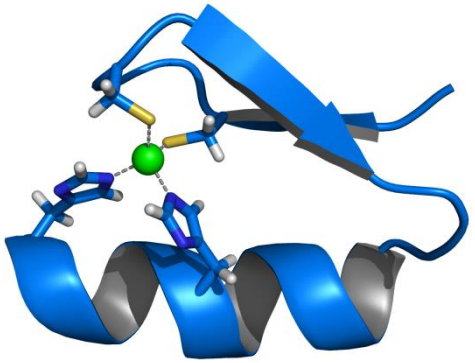
E(s) →

VLMQEGGFVLMS
MQEFTDGVMAA
AAVKRGTWWSR
EFVKLILAAIRST

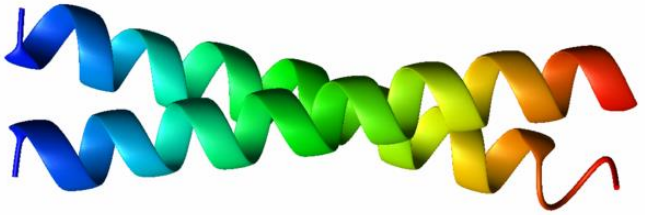
Can we design any structure we want? NO, designability principle.

Successful Designs

Redesigned Zinc Finger (Steve Mayo Lab)

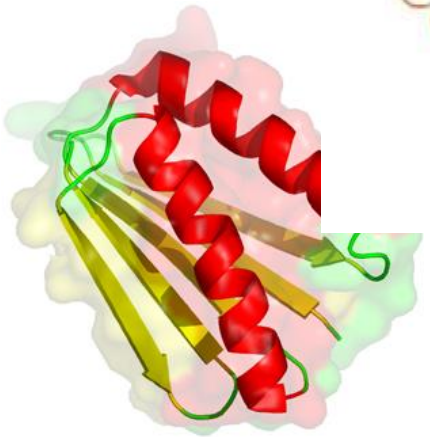


Design of right-handed coiled coil (Harbury & Kim)

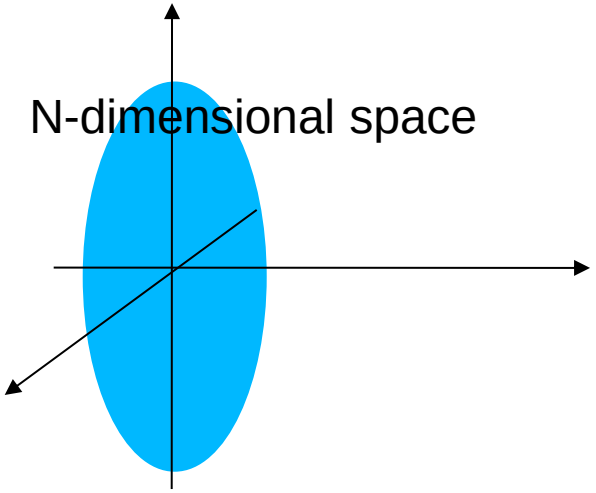


Binary patterning of helical bundle (Michael Hecht Lab)

Design of novel fold (David Baker Lab)



Principal Component Analysis:



data:

$x^i = (x_1, x_2, x_3, ..., x_N)$

with

i = 1, to some large M

Given a distribution of data find
directions along which data has greatest spread

Usefulness: given a huge dimensional dataset
can reduce it to a few important degrees of freedom




(e.g. can decompose large image data sets into
a few simple facial movements)

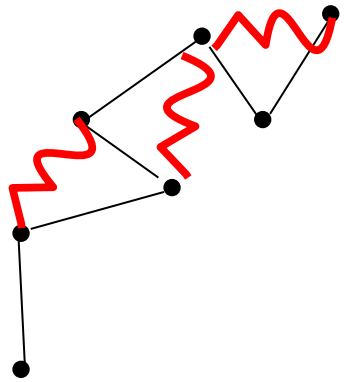
METHOD:

$$\text{covariance matrix} = C_{ij} = 1/(N-1)\ \Sigma_m\ (x^m_i - \langle x_i \rangle)(x^m_j - \langle x_j \rangle)$$

- eigenvalues, eigenvectors of C_{ij} give the directions of largest variation in data

(for proteins = the dominant eigenvectors correspond to the most flexible motions)

Normal Mode Analysis:



Assume motions of molecule are harmonic:

$$V = V0 + dV/dx|_{x0}(x-x0) + \tfrac{1}{2}\, d^2V/dx^2|_{x0}(x-x0)^2$$

$dV/dx|x0 = 0$ and $K_{ij} = d^2V/dx_i dx_j$

Or, place springs between atoms that are closer than R_c

$$V = \Sigma_{ij}\, \tfrac{1}{2}\, K_{ij}\, (x_i - x_j)^2$$

Equations of motion: $M\, d^2\mathbf{x}/dt^2 = -d\, V/\, d\mathbf{x}$

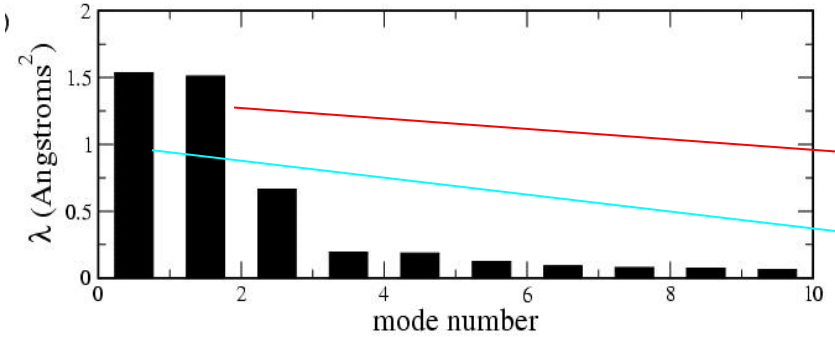
assume $\mathbf{x} = \Sigma\, \mathbf{a}_i \exp(-\omega_i\, t)$ ---> $M\, \omega^2\, \mathbf{x} = K\, \mathbf{x}$

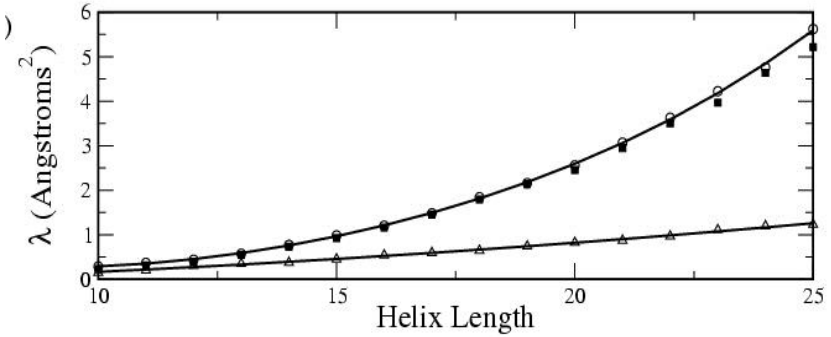
computing eigenvalues of K --> normal (dynamical) modes of the molecule

low-frequency modes = 'soft modes' = global motions of molecule

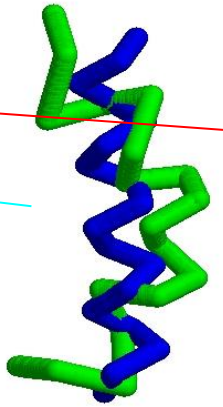
high-frequency modes = local motion of atoms in molecule

PCA Application to Helices:

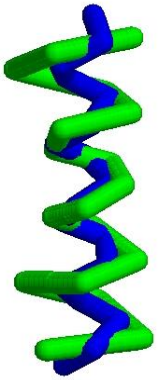




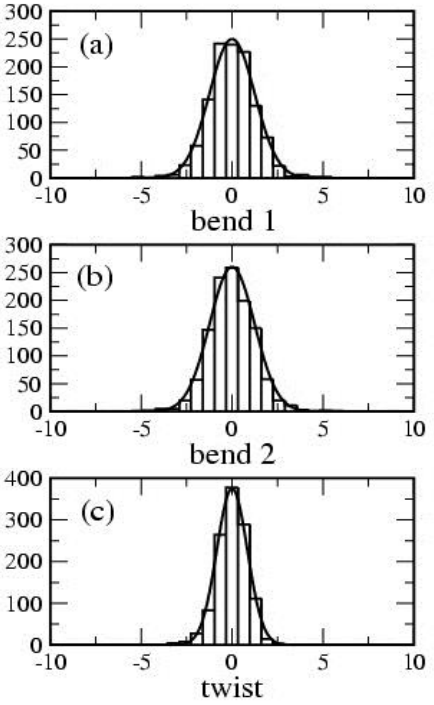
BEND



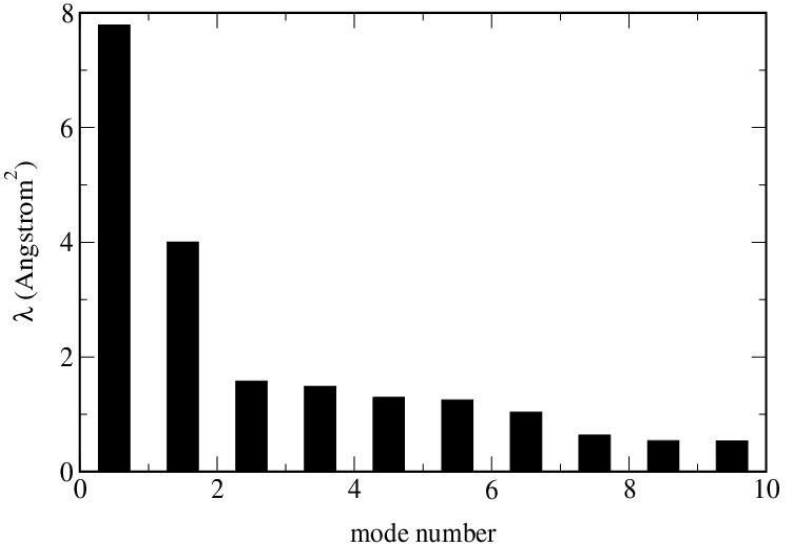
TWIST

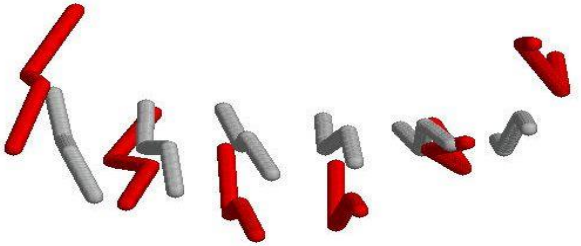


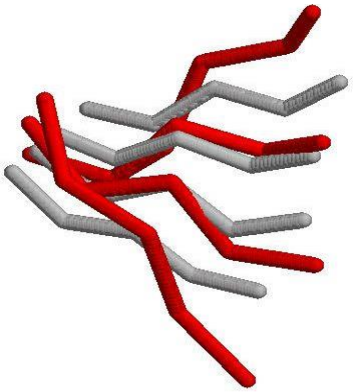
can fit normal modes to those obtained from PCA to extract spring constants

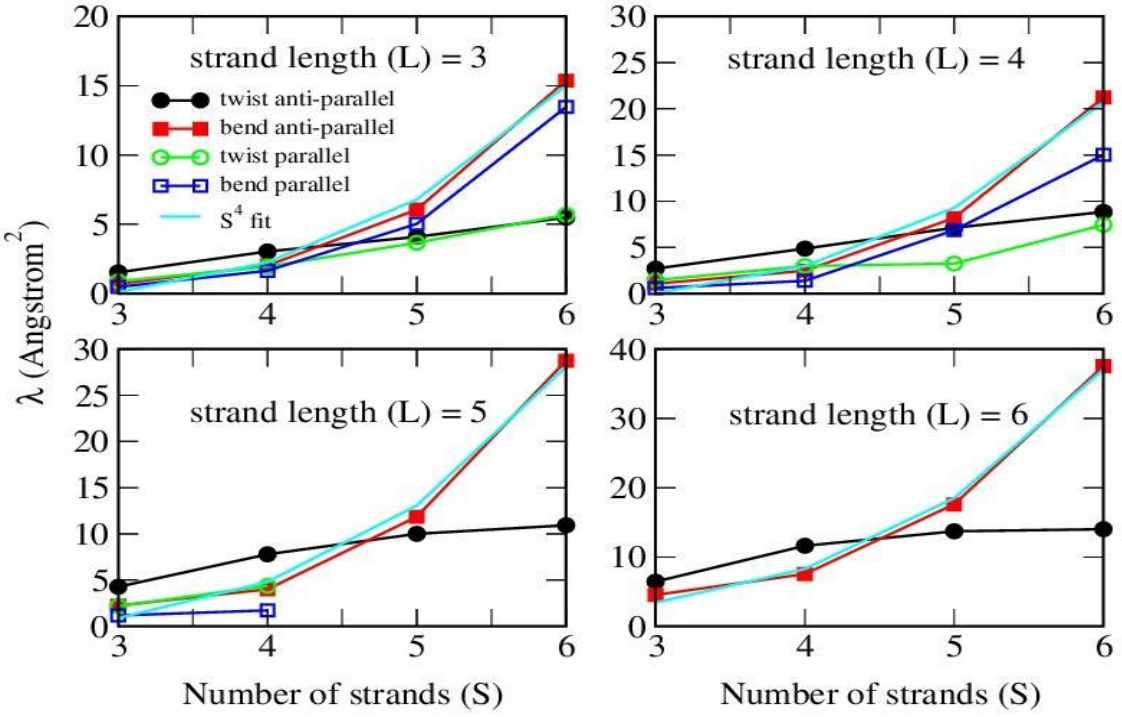


Application to Sheets:





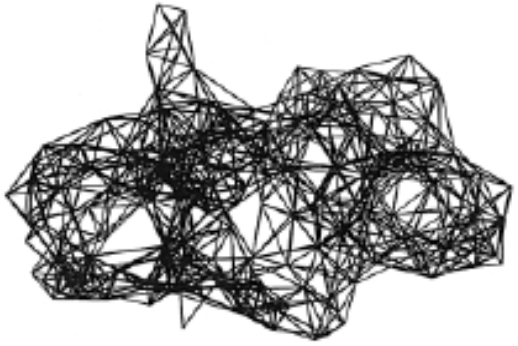




Application 3D structures:

open

open
spring
model



closed
spring
model

- Do normal modes of protein structures correspond to real conformational changes?
  Sometimes.
- Compute springs from complicated potential (requires relaxation), or use simple springs
- Slow modes often overlap well with the conformational change between 'closed' and 'open' configurations.
- Normal modes from 'open' conformation are often in better agreement with real motion

(F. Tama & Y-H Sanejouand, Protein Eng. 14, 1, 2001)