Protein Folding and Design

Carl Kingsford
02-251

Proteins

mRNA
$\Sigma = \{A,C,G,U\}$

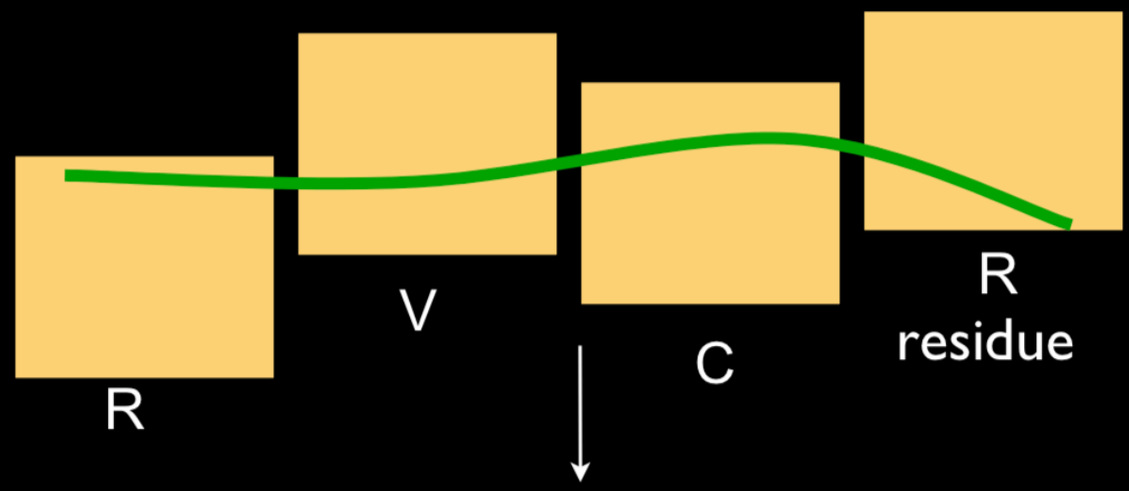
↓

protein
$|\Sigma| = 20$ amino acids

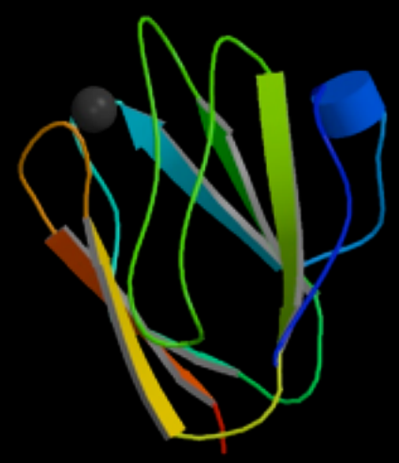
AGG GUC UGU CGA

↓ ↓ ↓ ↓

R V C R

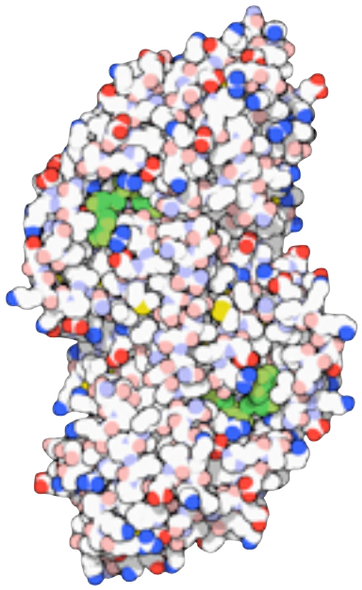


Amino acids with flexible side chains strung together on a backbone

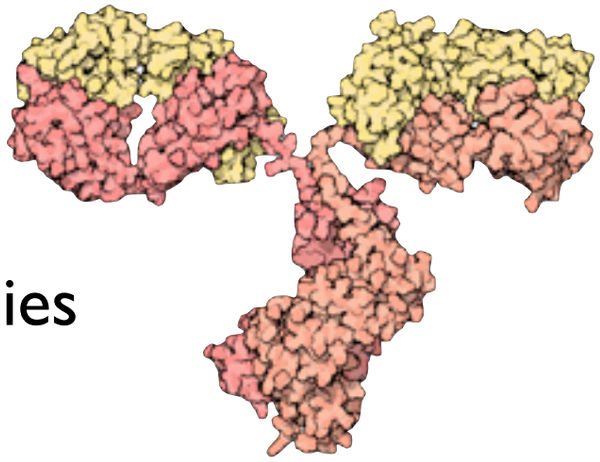


Function depends on 3D shape

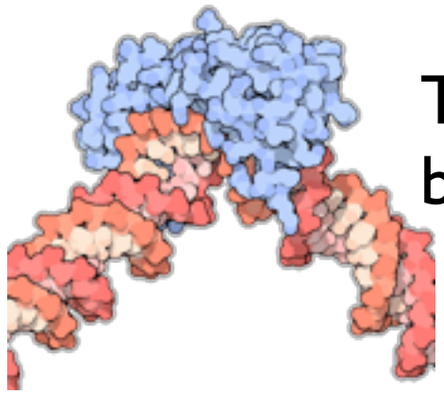
Examples of Proteins



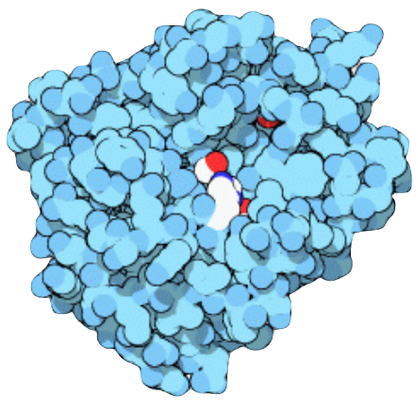
Alcohol dehydrogenase



Antibodies



TATA DNA binding protein

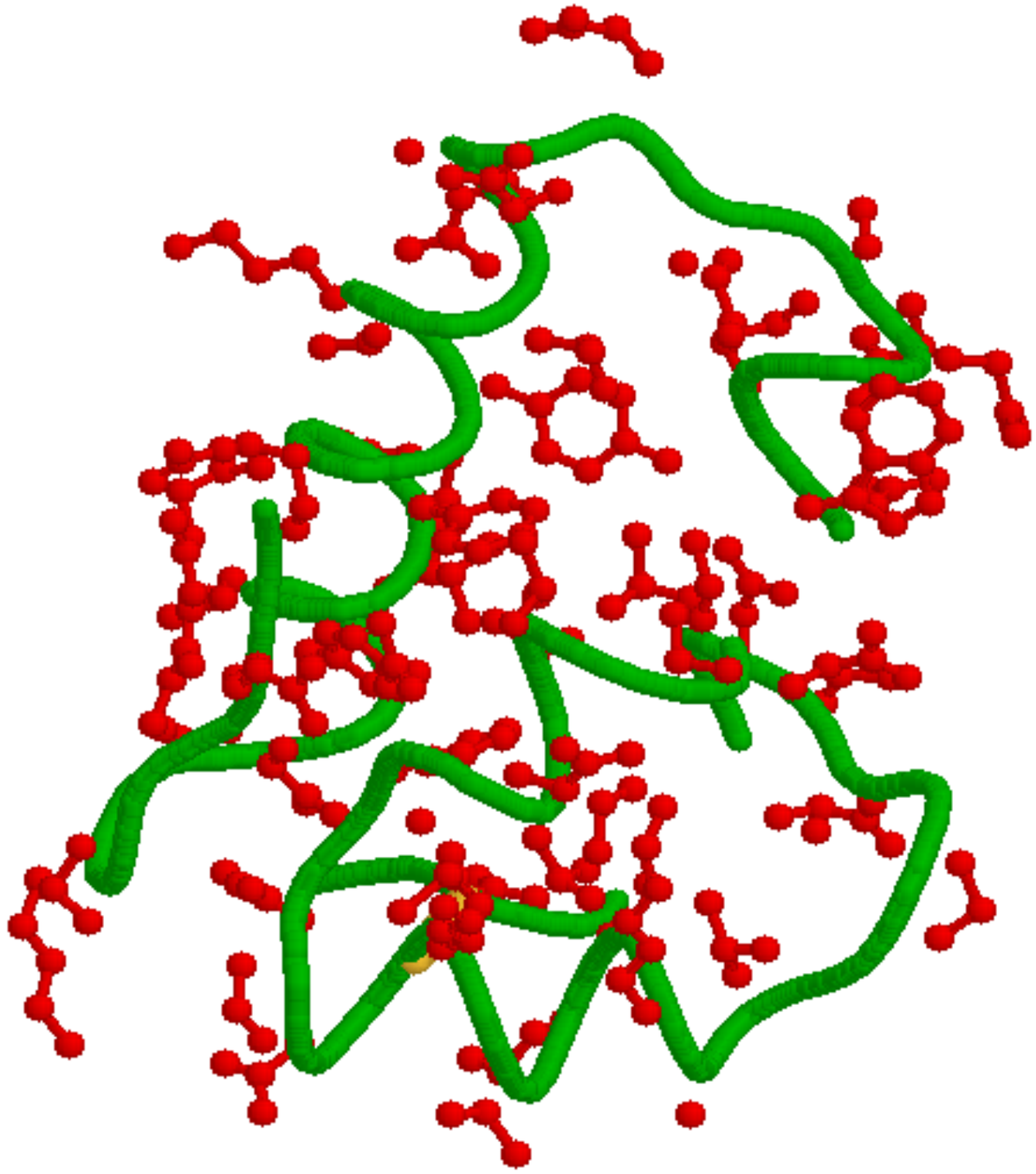


Trypsin: breaks down other proteins

Collagen: forms tendons, bones, etc.

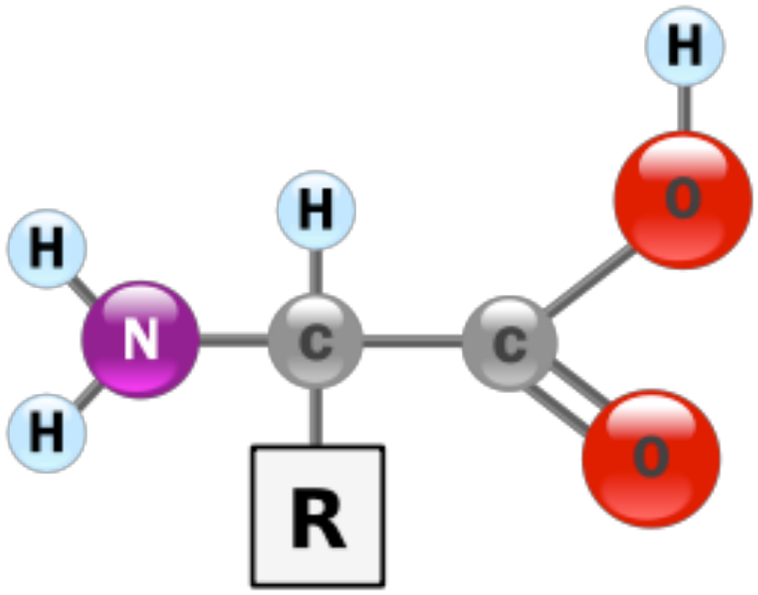
Examples of "Molecules of the Month" from the Protein Data Bank
http://www.rcsb.org/pdb/

Protein Structure



Backbone

Side-chains



Twenty standard Amino Acids

Nonpolar, aliphatic R groups

Glycine · Alanine · Valine

Leucine · Methionine · Isoleucine

Aromatic R groups

Phenylalanine · Tyrosine · Tryptophan

Positively charged R groups

Lysine · Arginine · Histidine

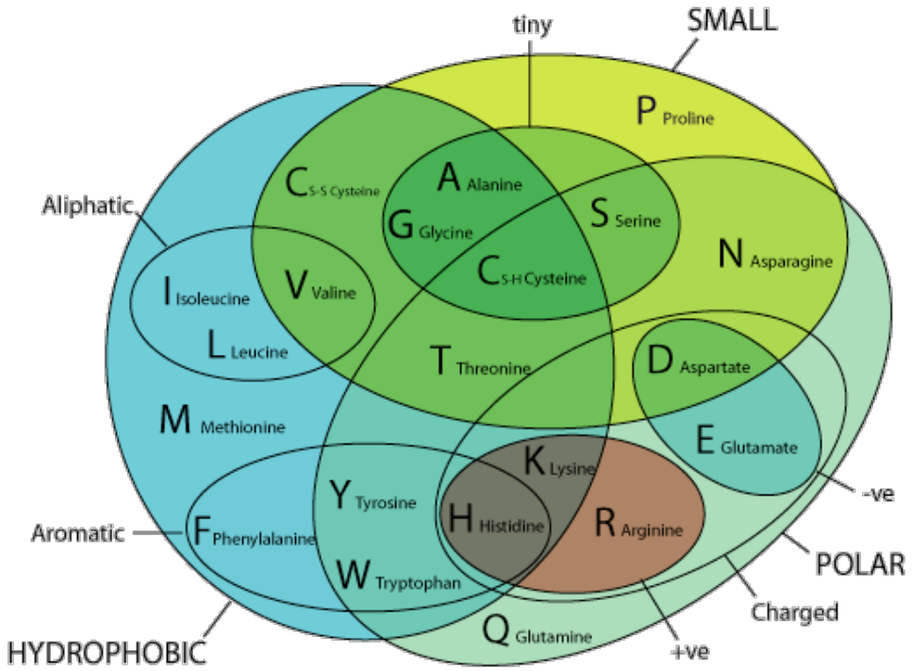
Polar, uncharged R groups

Serine · Threonine · Cysteine

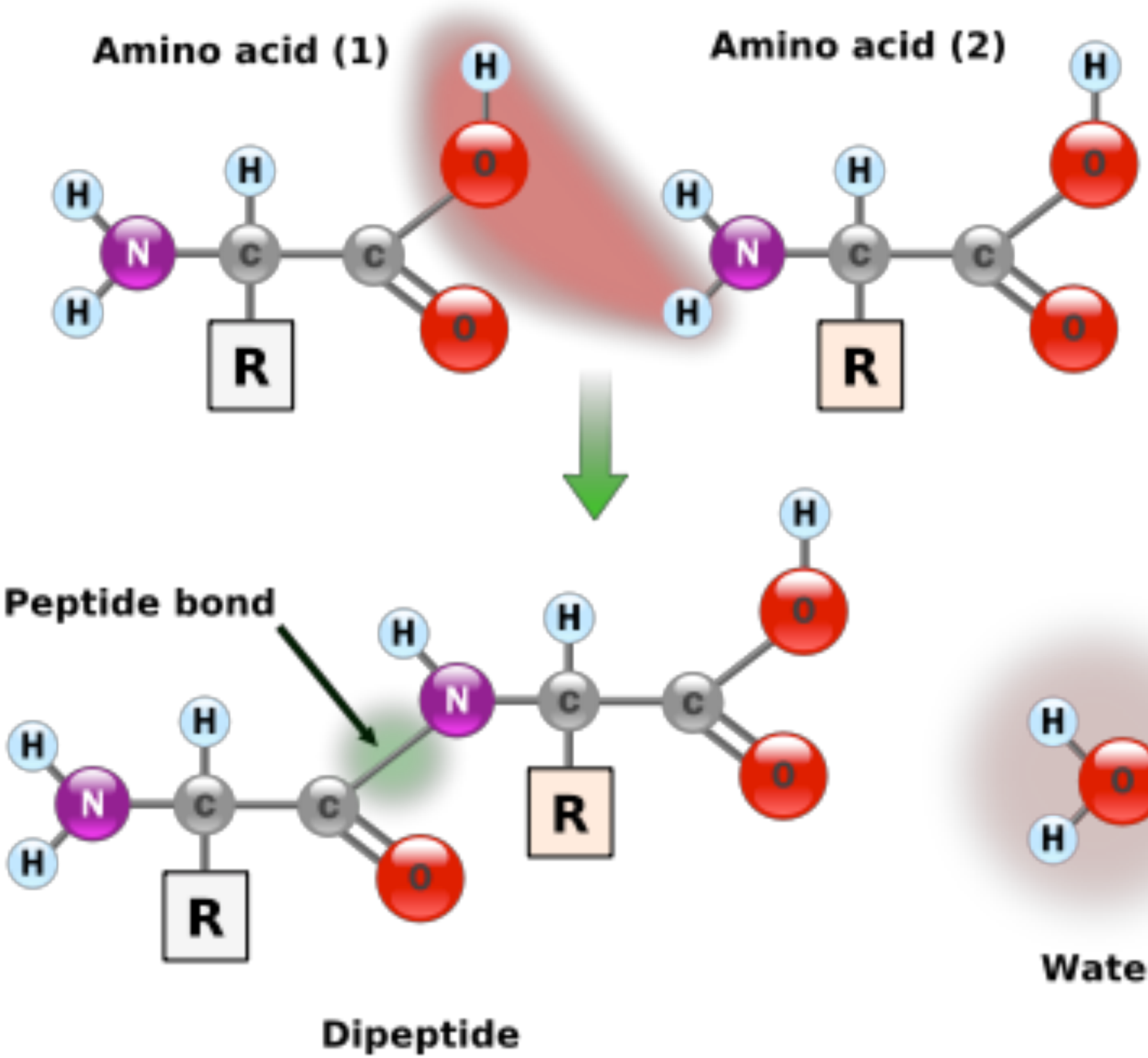
Proline · Asparagine · Glutamine

Negatively charged R groups

Aspartate · Glutamate



http://www.jalview.org/help/html/misc/properties.gif

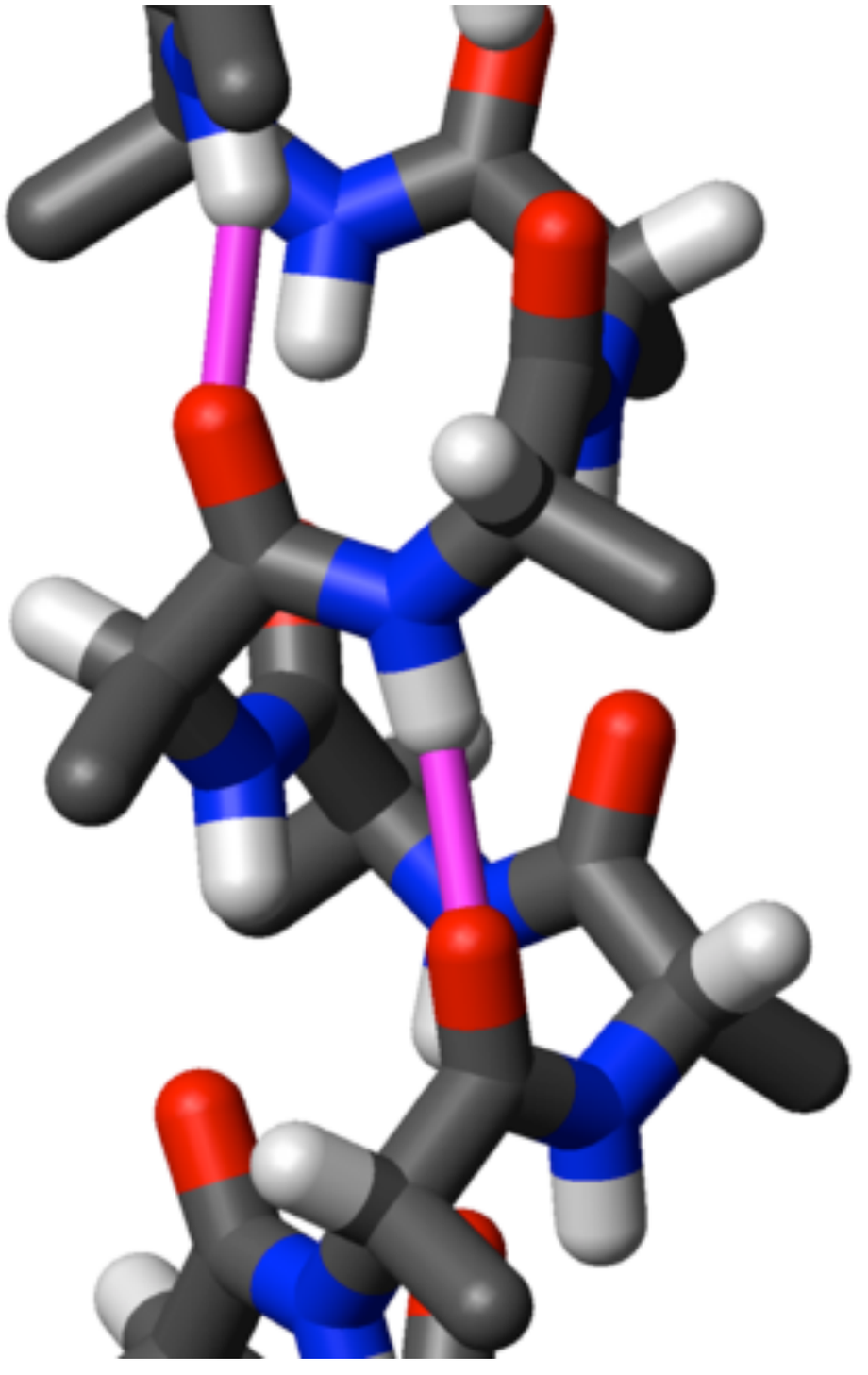

Amino acid (1)
Amino acid (2)
Peptide bond
Dipeptide
Water
H
O
N
C
R

Alpha helix

Beta sheet

1tim

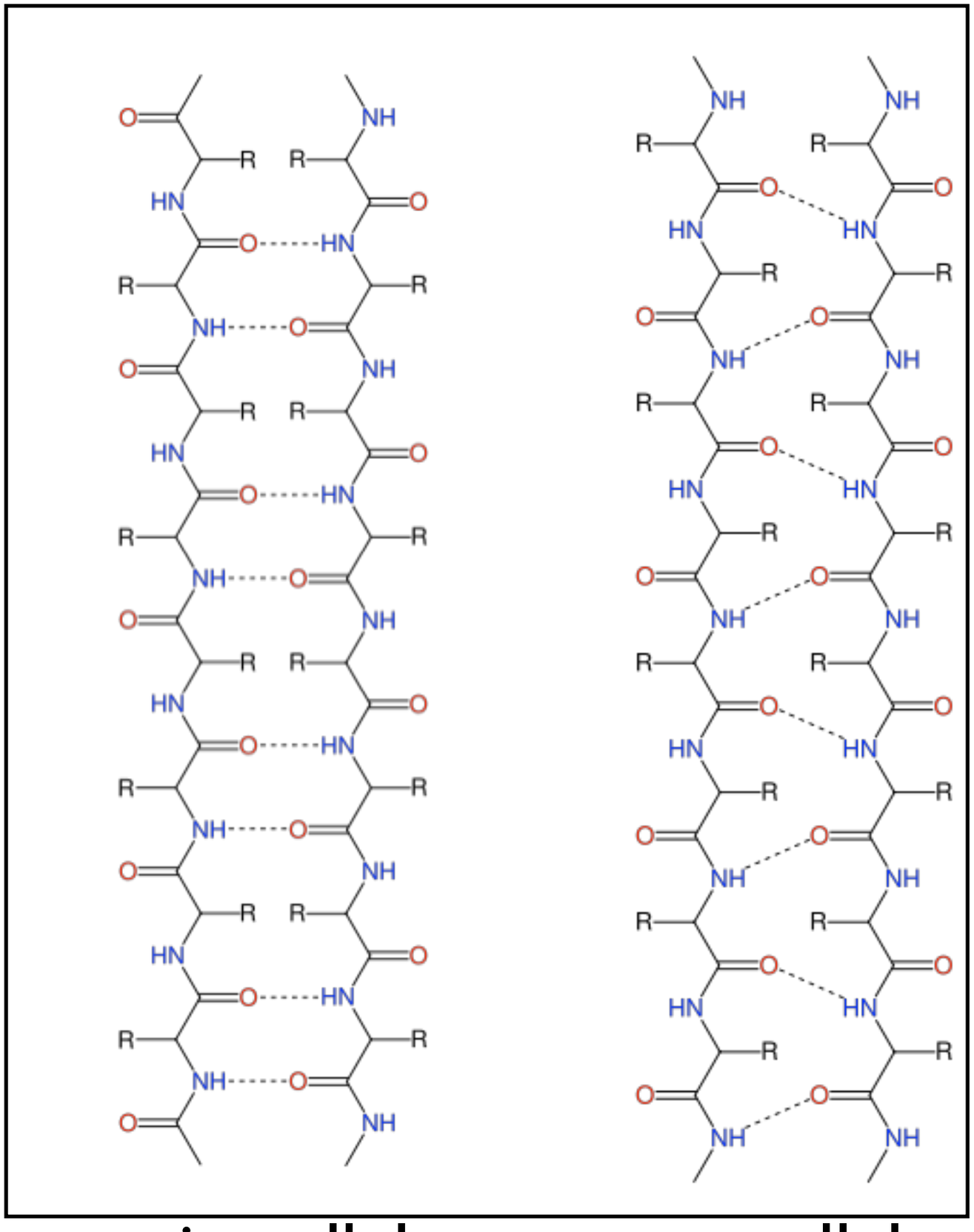
Alpha Helix



C'=O of residue *n* bonds to
NH of residue n + 4

Suggested from theoretical consideration
by Linus Pauling in 1951.

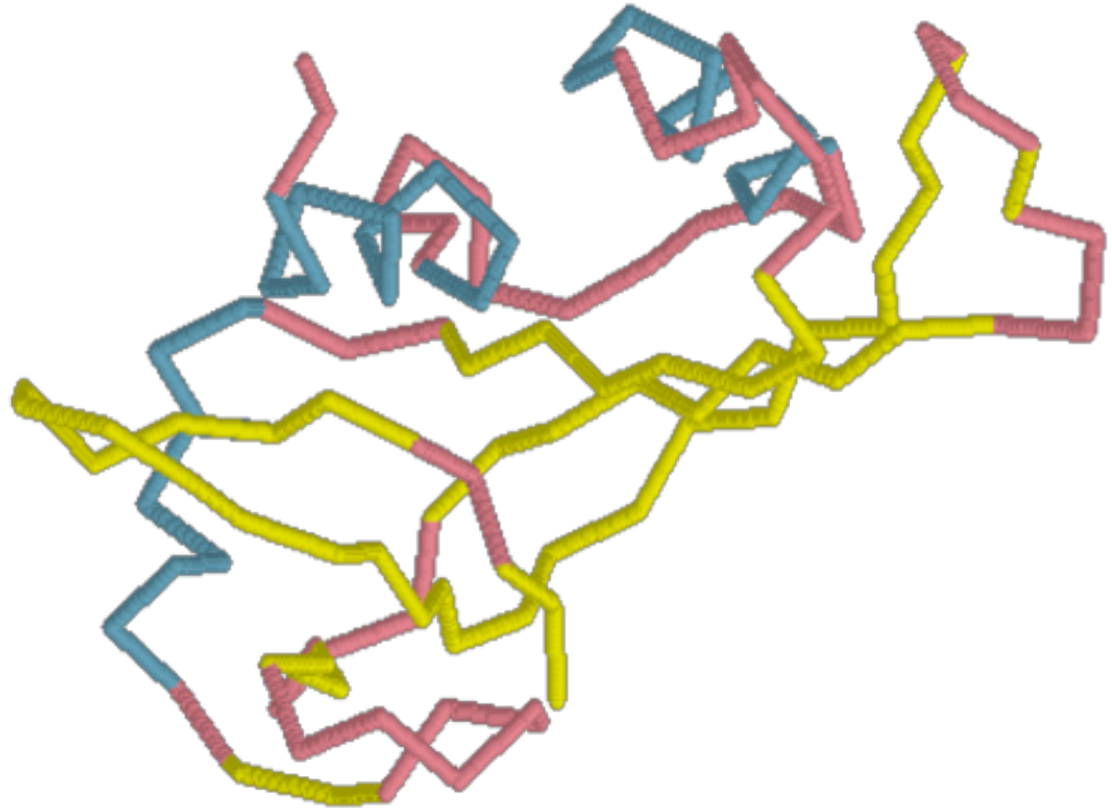
Beta Sheets



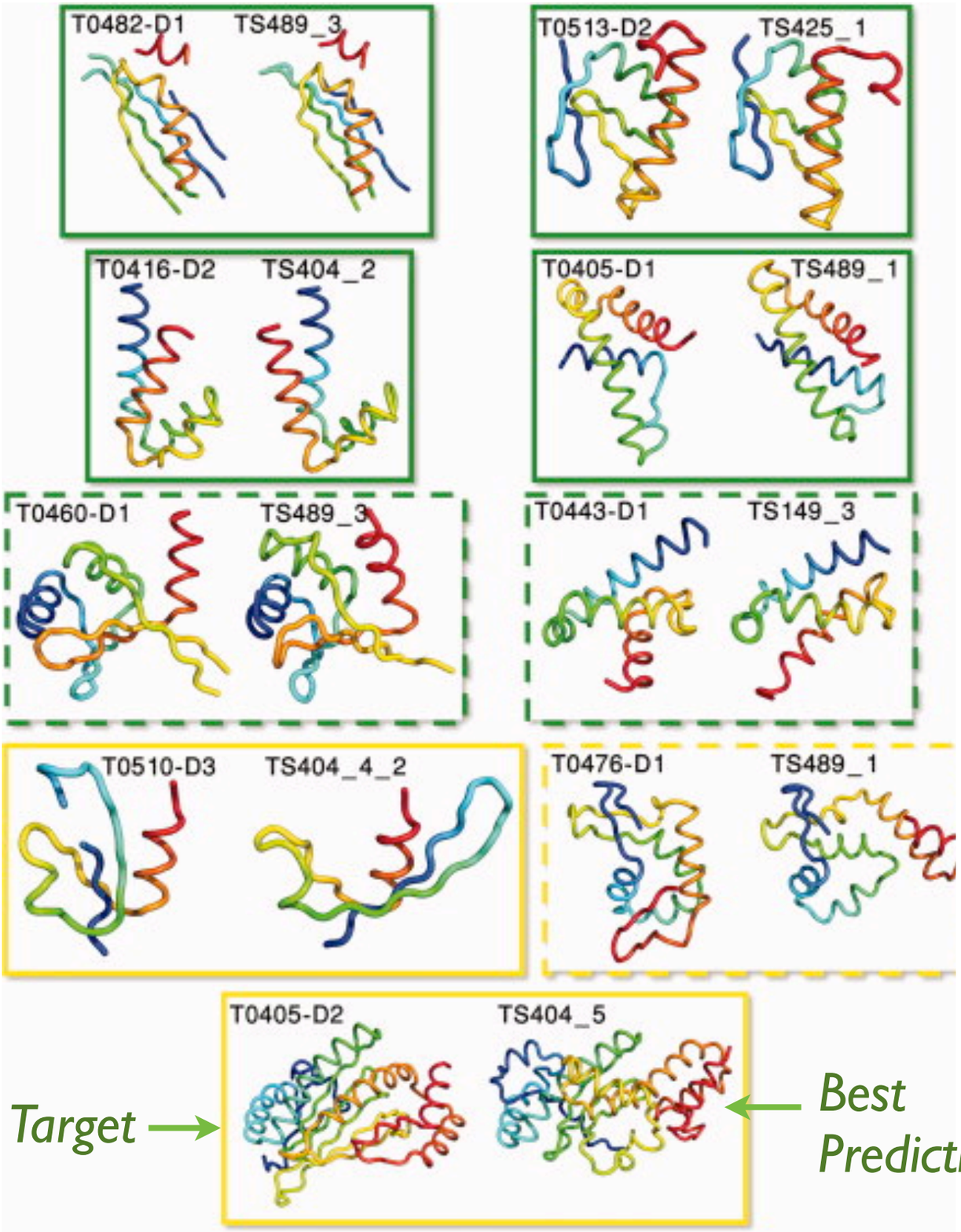
Structure Prediction

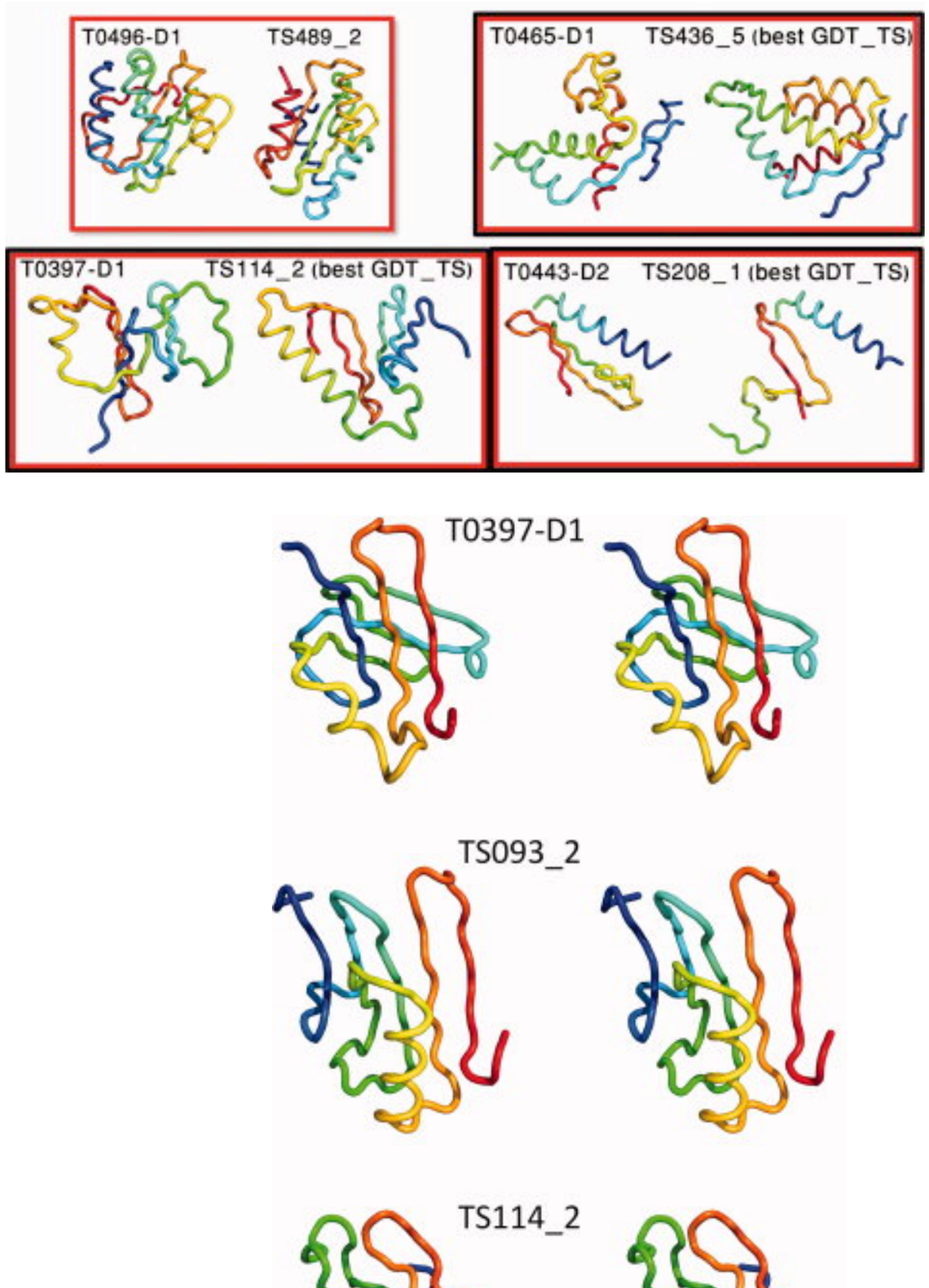
Given: KETAAAKFERQHMDSSTSAASSSN…

Determine:



CASP8

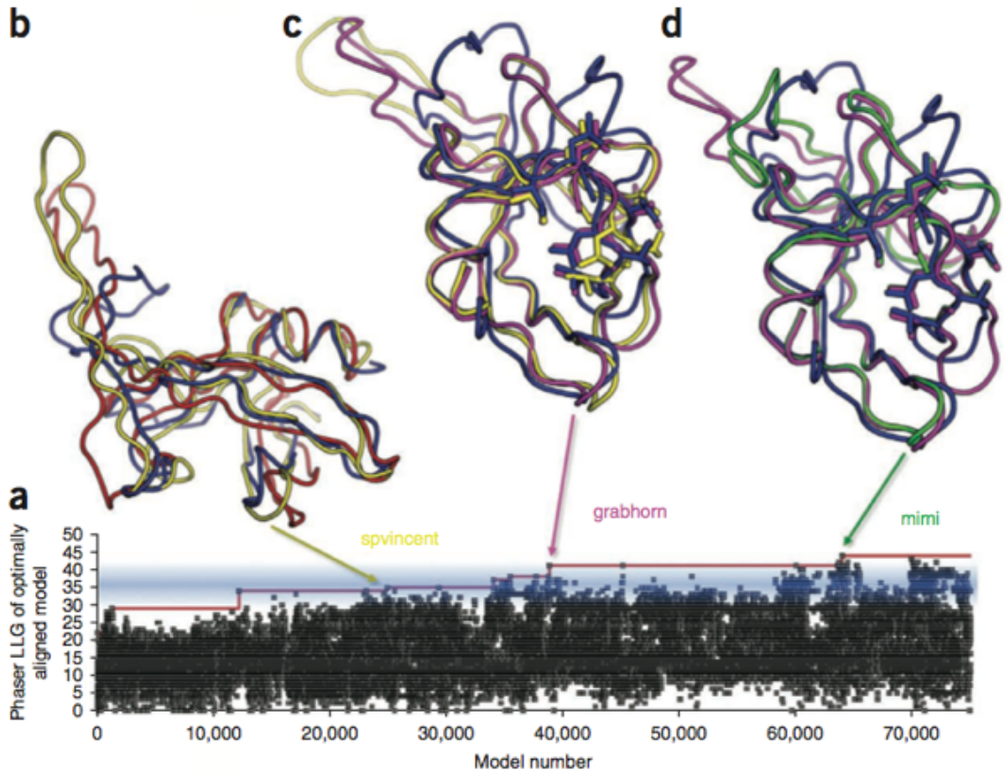




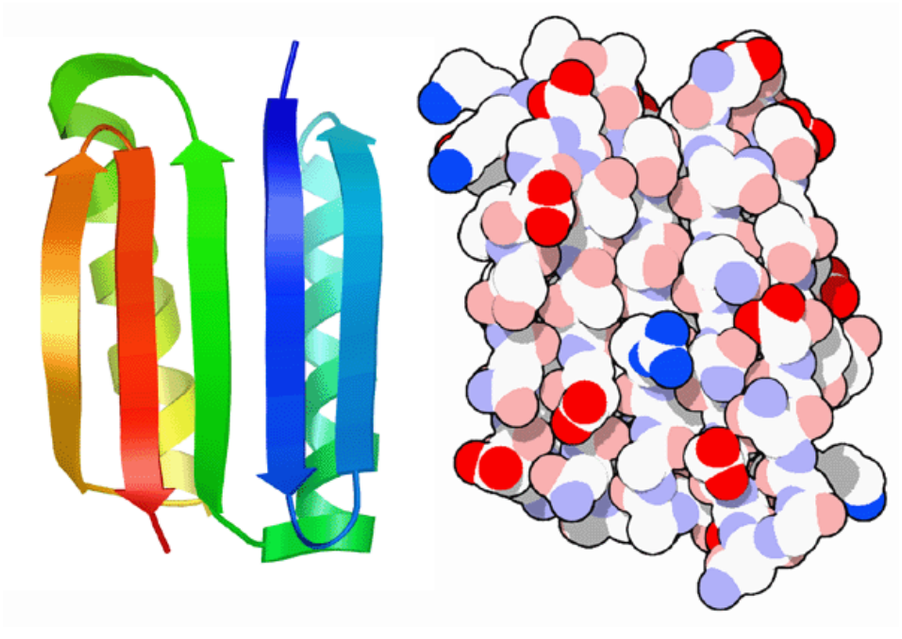
Ben-David et al, 2009

Critical Assessment of protein Structure Prediction

Structure Prediction & Design Successes

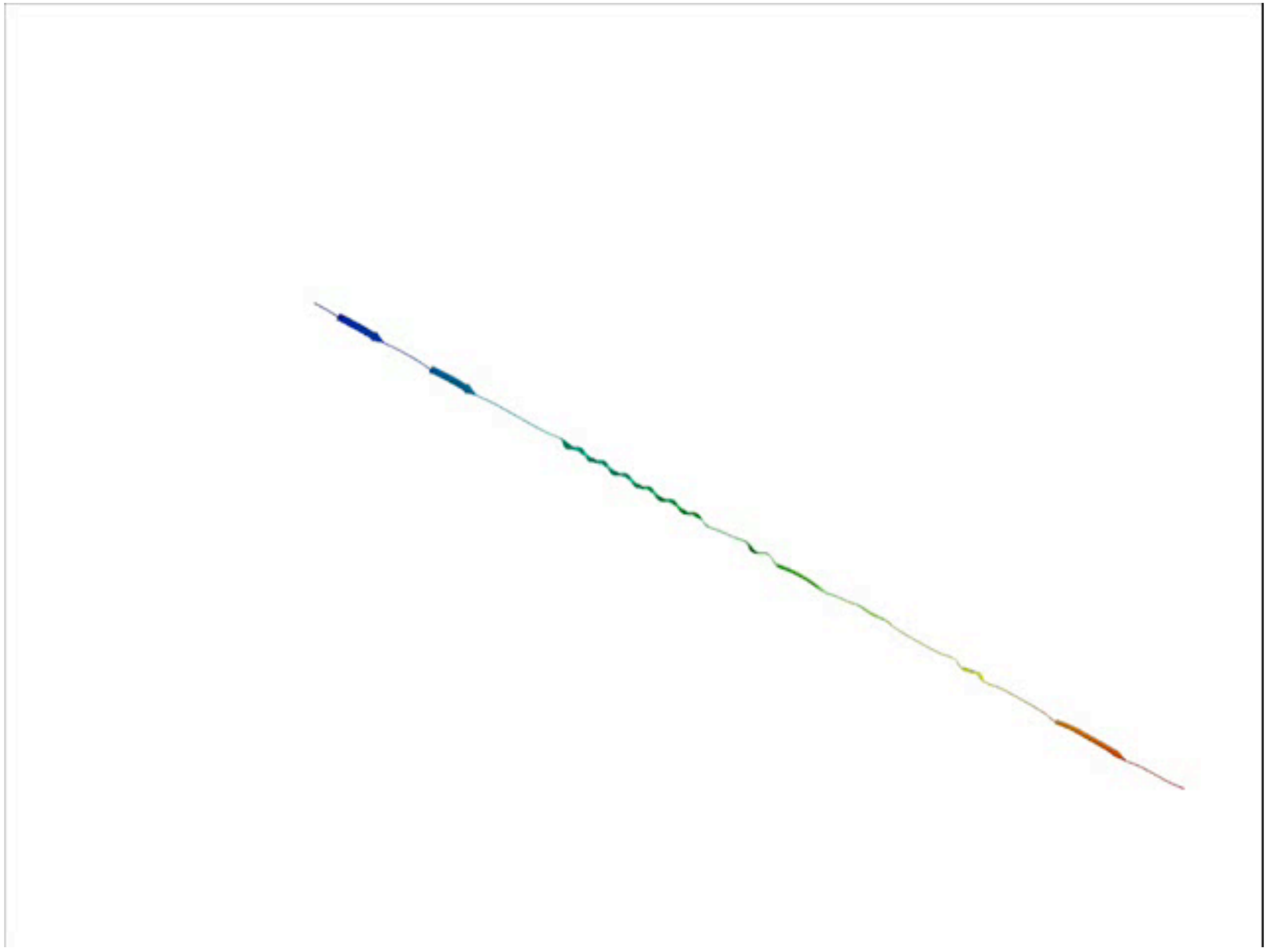


FoldIt players determination the structure of the retroviral protease of Mason-Pfizer monkey virus (causes AIDS-like disease in monkeys). [Khatib et al, 2011]



Top7: start with unnatural, novel fold at left, designed a sequence of amino acids that will fold into it. (Khulman et al, *Science*, 2003)

Folding Ubiquitin with Rosetta@Home



http://boinc.bakerlab.org

Rosetta@Home Algorithm (High-level)

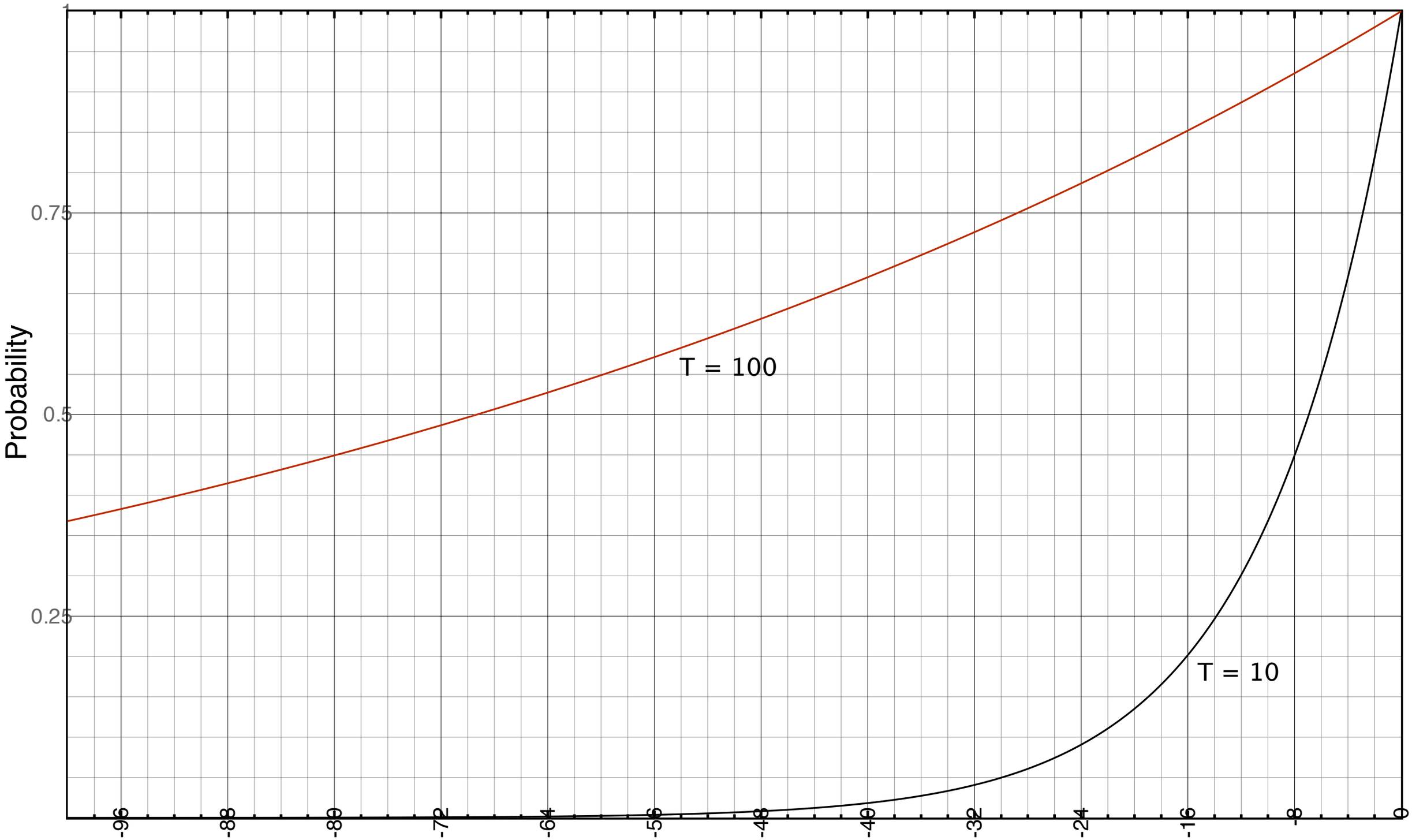
```
S = linear, unfolded chain
While some part of chain hasn't been moved a lot:
    Move part of S to get structure S'
    If energy(S') < energy(Best):
        Best = S'
    If energy(S') < energy(S):
        S = S'
    Else with probability related to energy(S) - energy(S'):
        S = S'
```

exp((energy(S) - energy(S'))/T)

Stage 1: uses big moves and a simple energy function

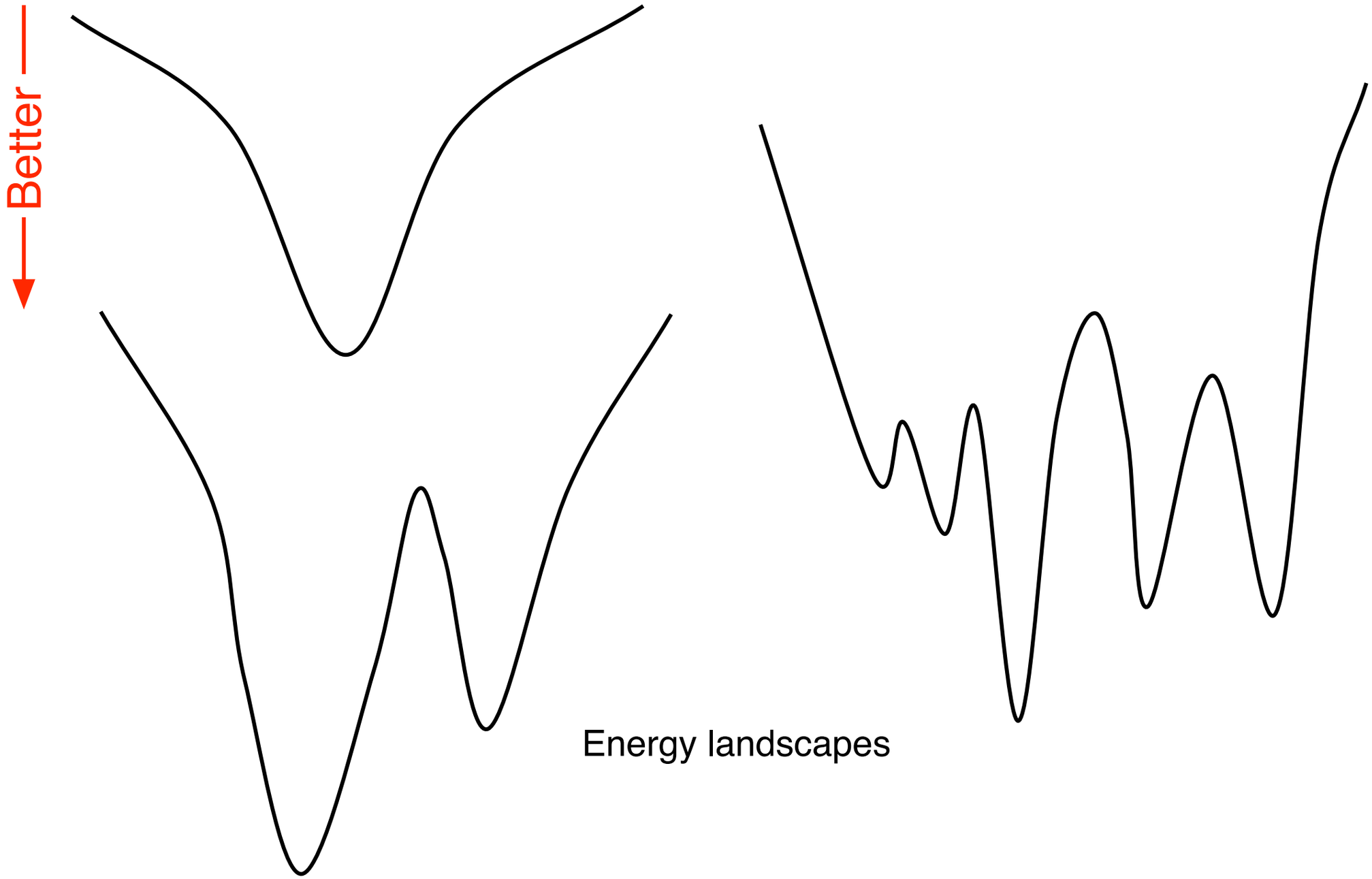
Stage 2: uses small moves and a complex energy function

exp(Δenergy)/T)



When T is large, more likely to accept a "bad" move.

Avoiding Local Minima

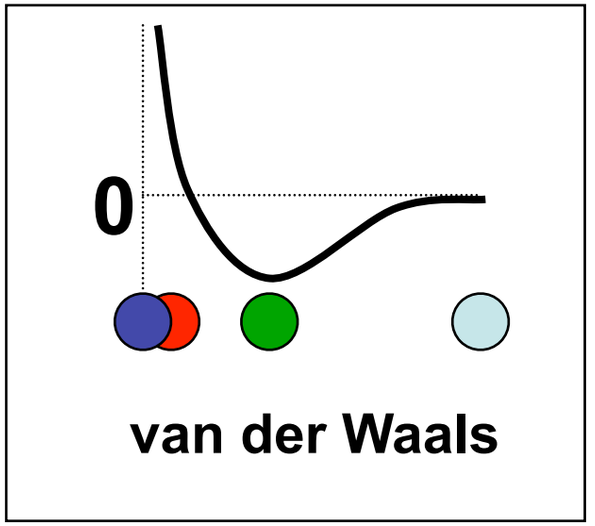


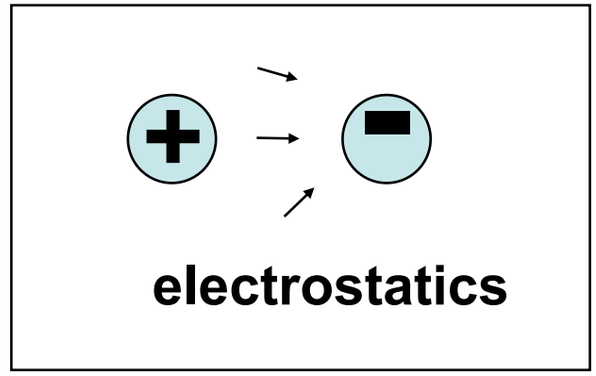
At low values of T, you will walk down towards a local minima.
At high values of T, you may jump out of a valley.

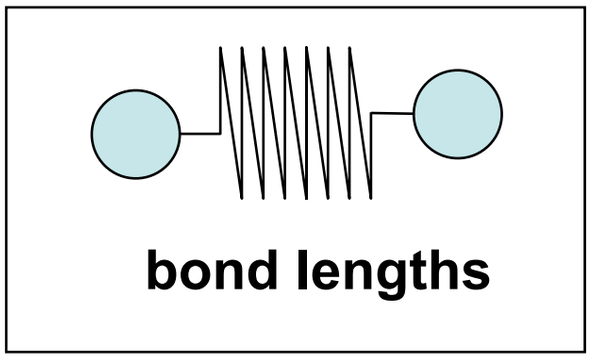
Simulated annealing idea: start with a high value of T and decrease over time (cooling schedule).

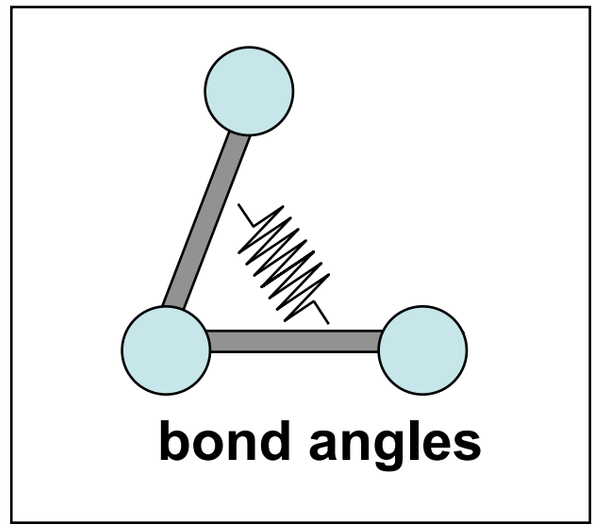
Determining the Energy

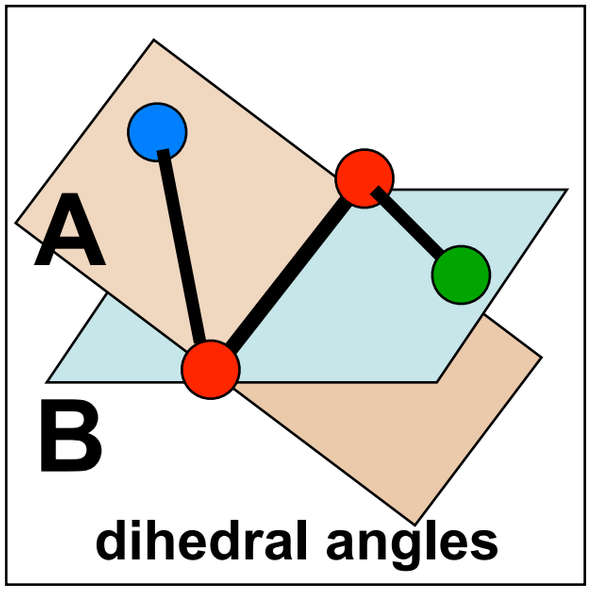
- Energy of a protein conformation is the sum of several energy terms.
- "Force Fields" such as CHARMM and AMBER give explicit approximations to each of these terms.

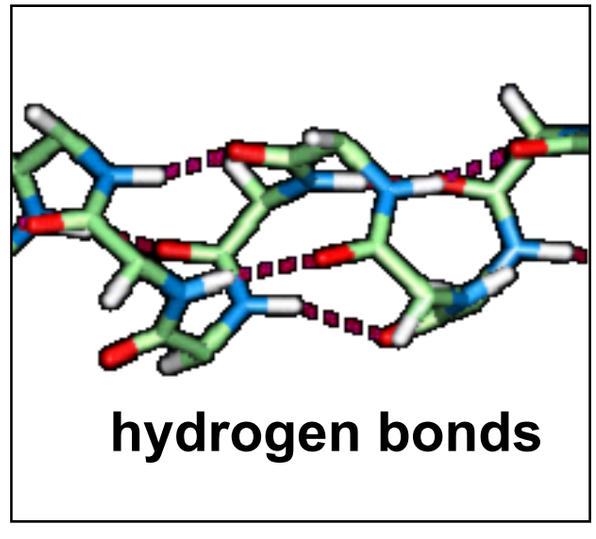












Energy Function (AMBER) Details

calculate the potential energy of a protein structure

$$V(r^N) = \sum_{\text{bonds}} \frac{1}{2} k_b (l - l_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_a (\theta - \theta_0)^2$$

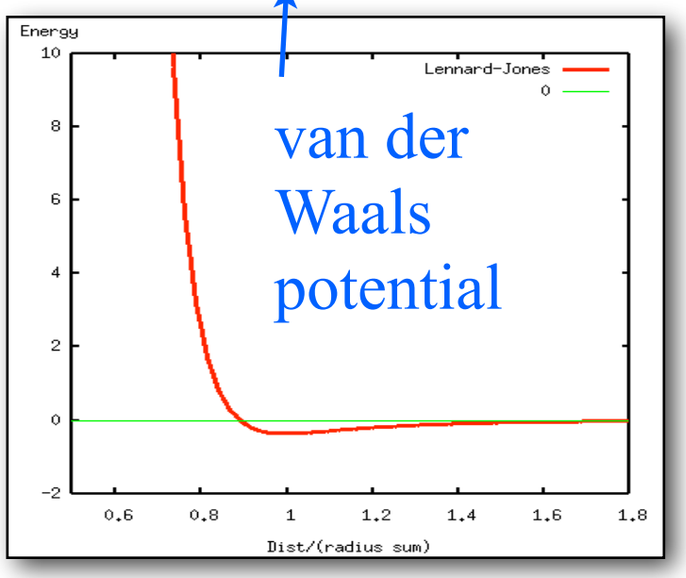
← Hook's law, spring of ideal length l_0 or θ_0 and tension k_b, k_a

$$+ \sum_{\text{torsions}} \frac{1}{2} V_n [1 + \cos(n\omega - \gamma)]$$

← function dependent on how much a bond is twisted

$$+ \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \left\{ \epsilon_{i,j} \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

Sum over all pairs of atoms



van der Waals potential

electrostatic between particles of charge q_i and q_j: derived from Coulumb's law

equations typeset courtesy Wikipedia

Protein Structure Summary 1

Protein structure vital in understanding protein function.

Prediction of protein structure is a very hard computational problem

- Some notable successes over the last ≈15 years
- Based on carefully constructed energy functions
- Main algorithmic tool: simulated annealing-like randomized algorithms that efficiently explore the space of conformations

2013 Nobel Prize in Chemistry

© Nobel Media AB. Photo: A. Mahmoud

Martin Karplus

Prize share: 1/3

© Nobel Media AB. Photo: A. Mahmoud

Michael Levitt

Prize share: 1/3

© Nobel Media AB. Photo: A. Mahmoud

Arieh Warshel

Prize share: 1/3

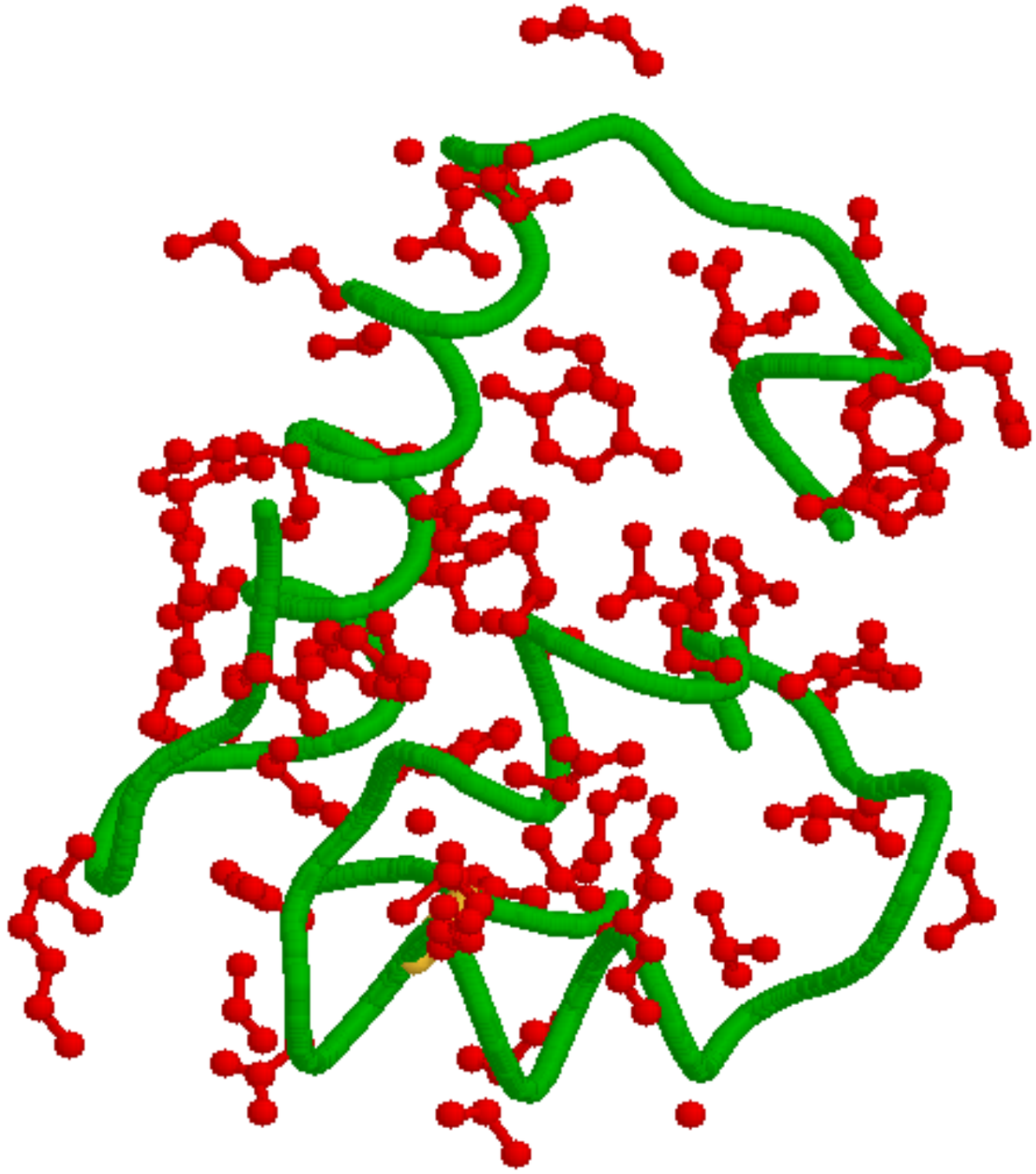
“Chemists used to create models of molecules using plastic balls and sticks. Today, the modelling is carried out in computers. In the 1970s, **Martin Karplus**, **Michael Levitt** and **Arieh Warshel** laid the foundation for the powerful programs that are used to understand and predict chemical processes. Computer models mirroring real life have become crucial for most advances made in chemistry today.”

https://www.nobelprize.org/prizes/chemistry/2013/summary/

Side-Chain Positioning

A key step in structure prediction & protein design

Protein Structure



Backbone

Side-chains

Side-chain Positioning

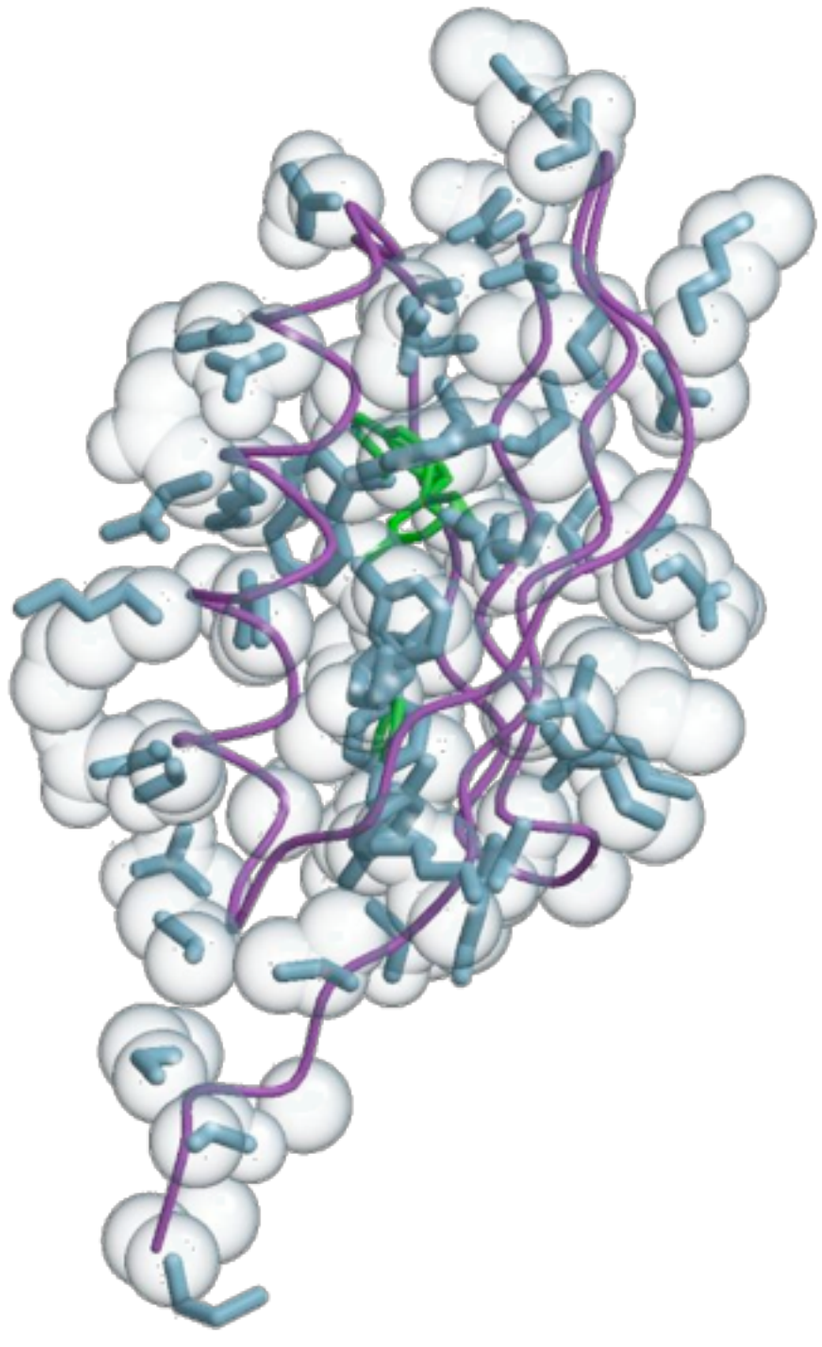
Given:

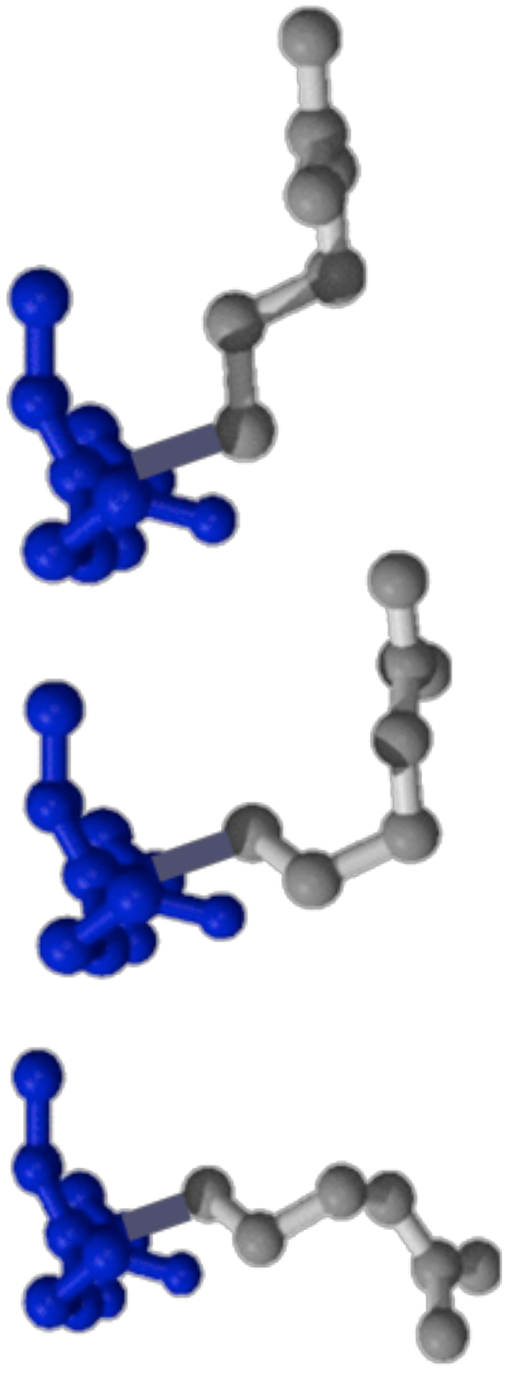
- amino acid sequence
- position of backbone in space

Find best 3D positions for side chains

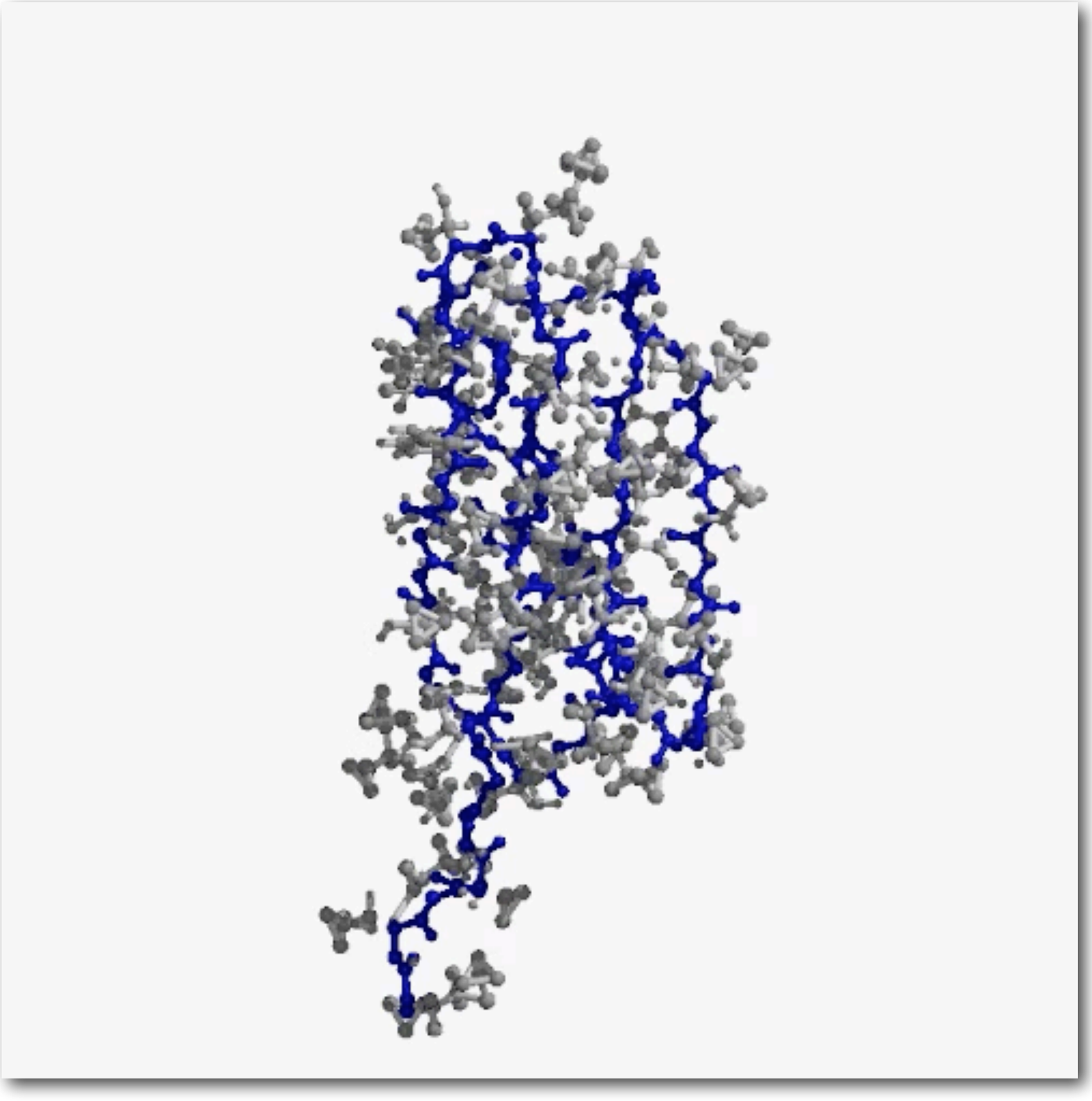
"Best" = lowest-energy

Discrete formulation reasonable using *rotamers*





3 rotamers of Arg



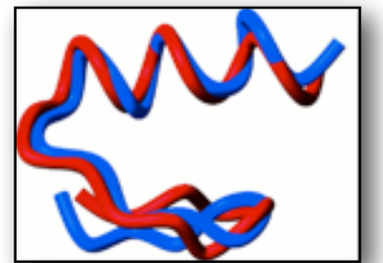
Applications

Homology modeling

- Rapid, low-cost structure determination

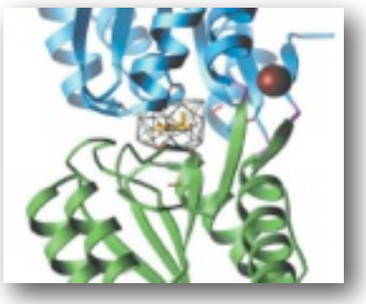
Protein design

- Find sequence that folds into a given shape
- e.g. redesign of zinc finger that folds without zinc, (Dahiyat+97)



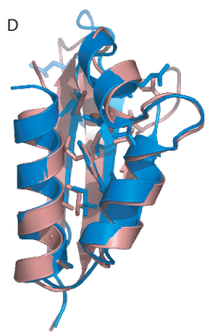
Ligand binding

- e.g. novel binding pockets (Looger+03)



Subroutine in flexible backbone prediction

- e.g. (Bradley+,2005)



Rosetta

At a **very** high level, the most success protein structure prediction software does the following:

Repeat:

- Generate many candidate backbones
- Optimize positions of side-chains
- Select promising structures to refine

Leaver-Fay et al. Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol.* **487**: 545–574 (2011) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4083816/

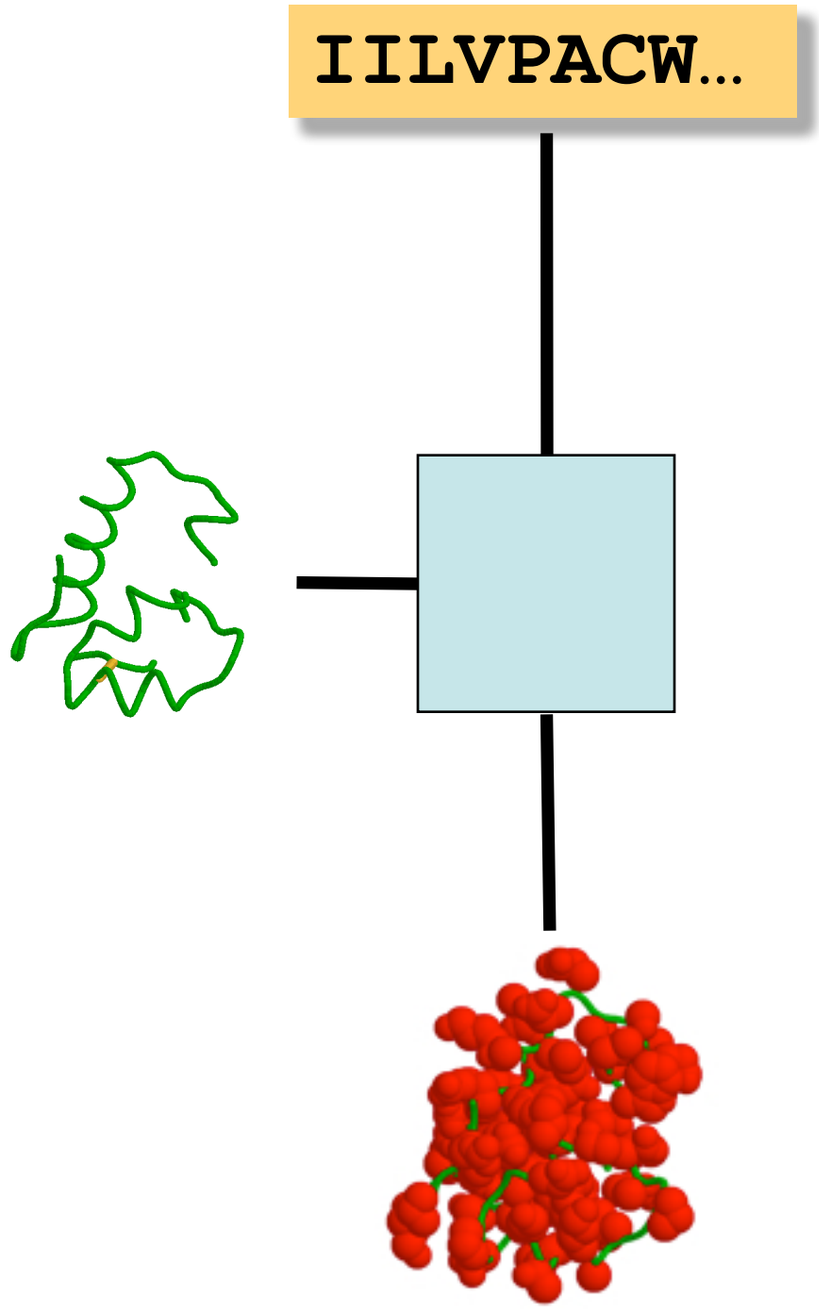
Side-chain Positioning Problem

Given:

- fixed backbone
- amino acid sequence

Find the 3D positions for the side-chains that minimize the energy of the structure

Assume lowest energy is best



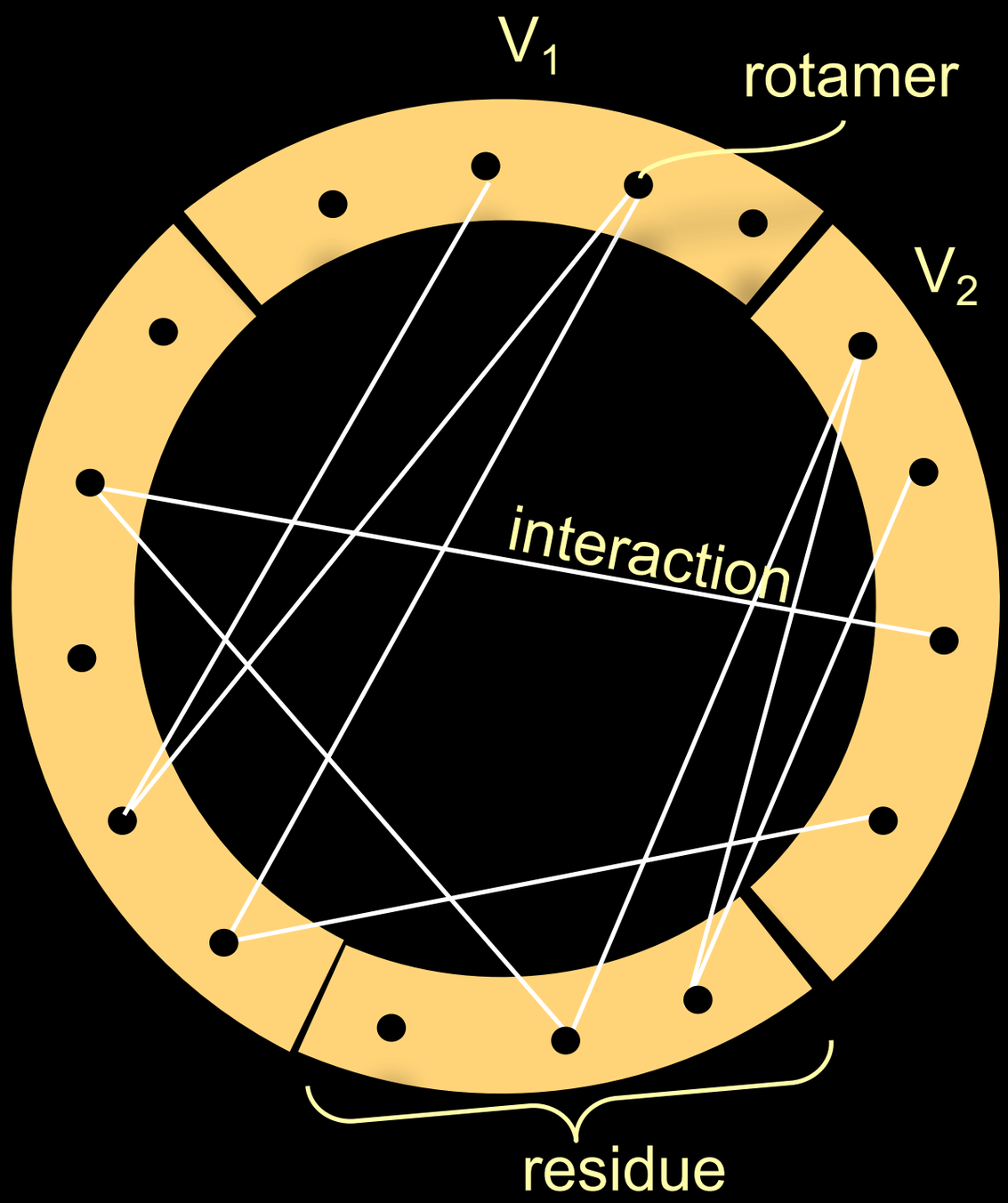
Graph Problem

Graph with part V_i for each side chain:

- node for each rotamer
- edge {u,v} represents the interaction between u and v

Weights:

- E(u) = self-energy
- E(u,v) = interaction energy

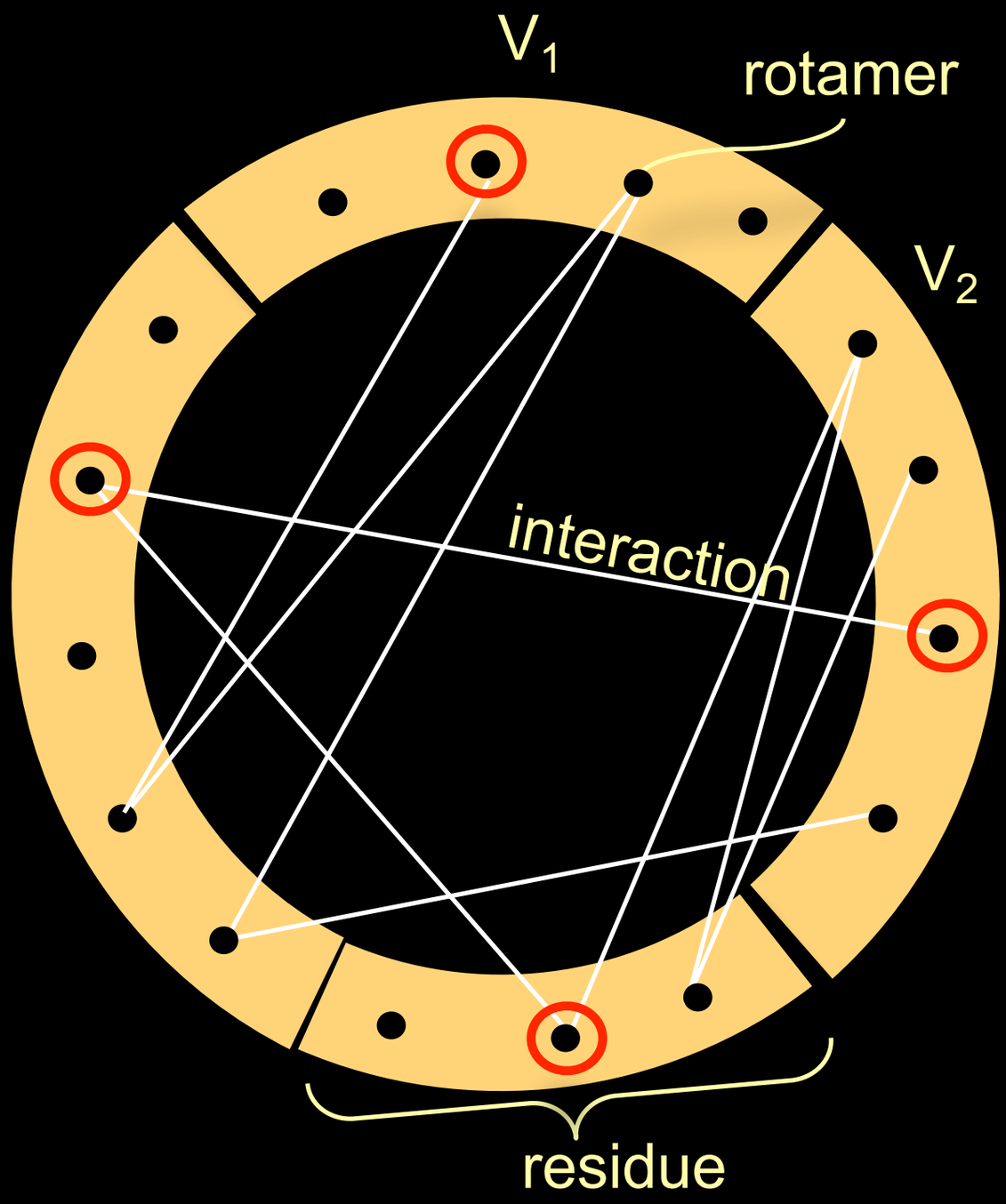


Graph Problem

Solution is one node from each part

Cost of solution is cost of induced subgraph

Goal: pick one node from each position to minimize the cost of the induced subgraph



Hardness

NP-hard to approximate the minimum energy within a factor of cn where $c > 0$ and n = # of rotamers (CKS04)

$\Rightarrow$ Little hope for a fast algorithm that guarantees good solutions

Proposed Solutions

Local search
- Monte Carlo (Xiang+01)
- Simulated annealing (Lee+91, Kuhlman+00)
- Many others

Graph heuristics
- Scwrl (Bower+97, Canutescu+03)
- Dead-end elimination (Desmet+92,...)
- & others (Samudrala+98, Bahadur+04)

Mathematical programming
- Semidefinite **(Chazelle, K, Singh, 04)**
- Linear/integer (Althaus+00; Eriksson+01; **KCS, 05**)

⇒ Flexible, practical framework to find optimal solutions.

Integer Programming

- General optimization framework:
  - Describe system by set of variables

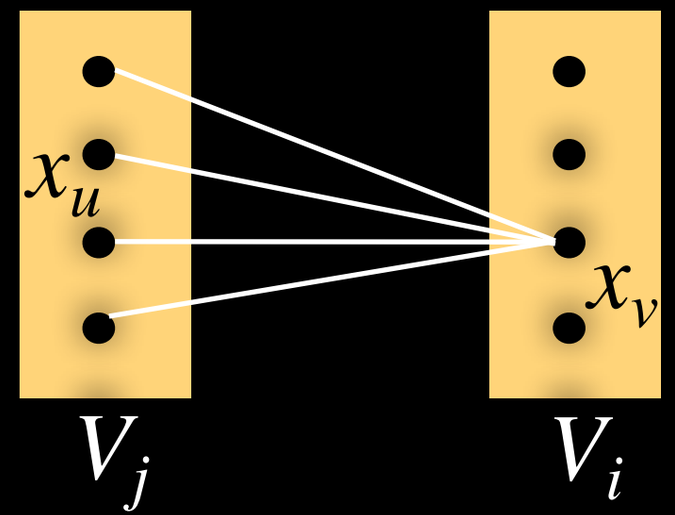
IP :=
- Minimize a linear function.
- Subject to linear constraints (= or ≥).
- While requiring the variables to be {0, 1}.

- Computationally hard, but many advanced solver packages:
  - CPLEX, COIN-OR, ABACUS, FortMP, LINGO, …

Integer Programming Formulation

Binary variables x_u for each node

Binary variables x_{uv} for each edge



Minimize $\sum_u E_u x_u + \sum_{u,v} E_{uv} x_{uv}$

subject to:

1. $\sum_{u \in V_j} x_u = 1$ for every residue j

2. $\sum_{u \in V_j} x_{uv} = x_v$ for every residue j, node v

Why Integer Programming?

Optimal solutions

- Eliminate any effect of local search
- Help to improve energy functions
- Assess quality of heuristic methods

Very good IP solvers available

Ensemble of near-optimal solutions

- Several design candidates
- Confidence in solution

Linear Programming Relaxation

$$x_u, x_{uv} \in \{0, 1\}$$

Integer Program

Enforcing binary constraints is hard.

Guarantees finding an optimal choice of rotamers.

$$0 \leq x_u, x_{uv} \leq 1$$

Linear Program

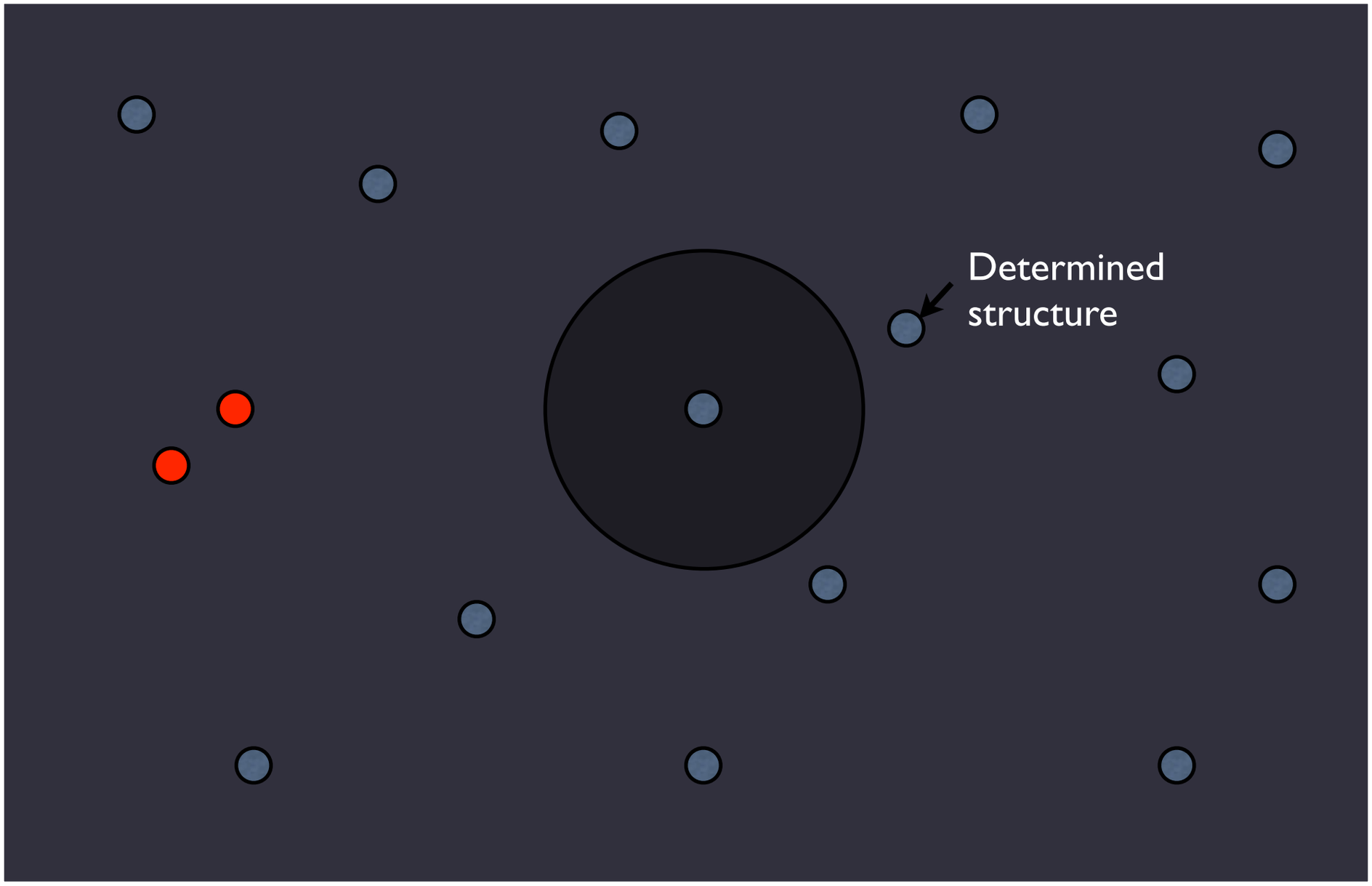
Computationally easier.

May return fractional solution.

If integral, done.

If not, either round or add new constraints

Structural Genomics



Space of all protein structures

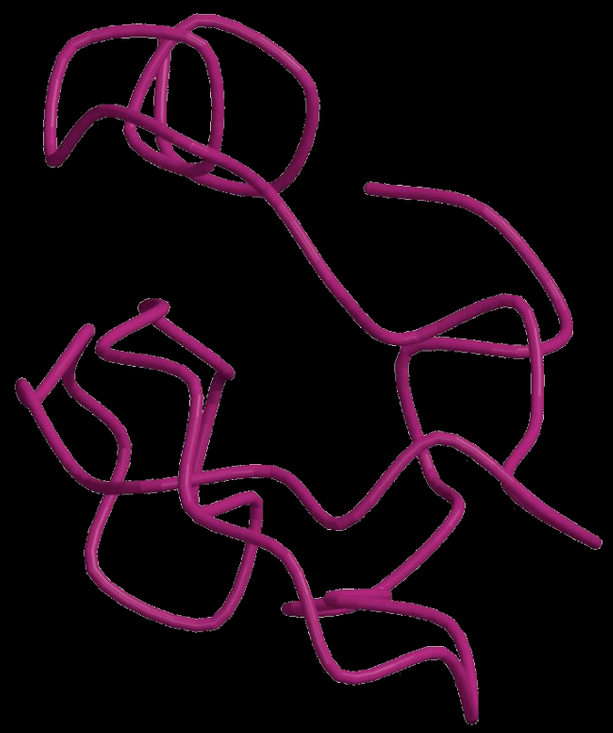
Homology Modeling

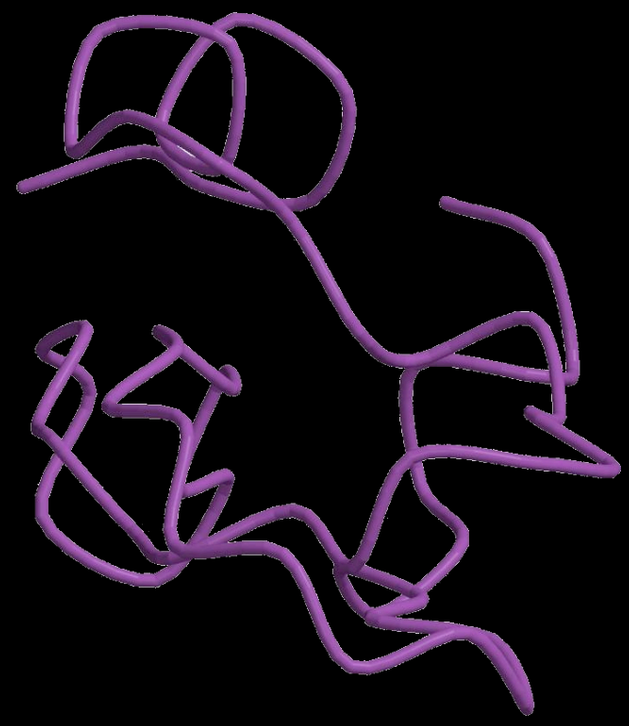
Similar sequences ⇒ similar backbones

Use known backbone of similar protein to predict new structure

```
1dtk   XAKYCKLPLRIGPCKRKIPSFYYKWKAKQCLPFDYSGCGGNANRFKTIEECRRTCVG-
5pti   RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA
```

1dtk: toxin in venom of Dendroaspis polylepis





5pti: bovine pancreatic trypsin inhibitor

Homology Modeling

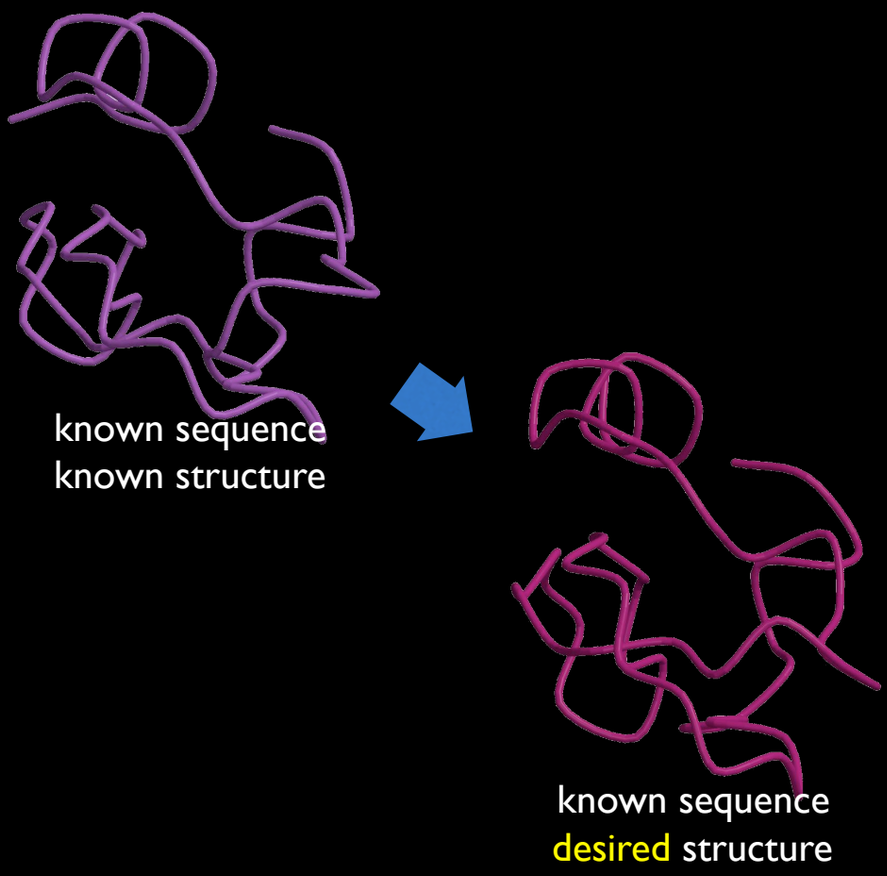
33 homology modeling problems

- 49 to 220 variable residues
- 723 to 4154 nodes
- 29 to 87% sequence similarity

< 12 minutes of computation

The LP relaxation was integral in 31 problems

Can solve the remaining 2 with additional branch & bound



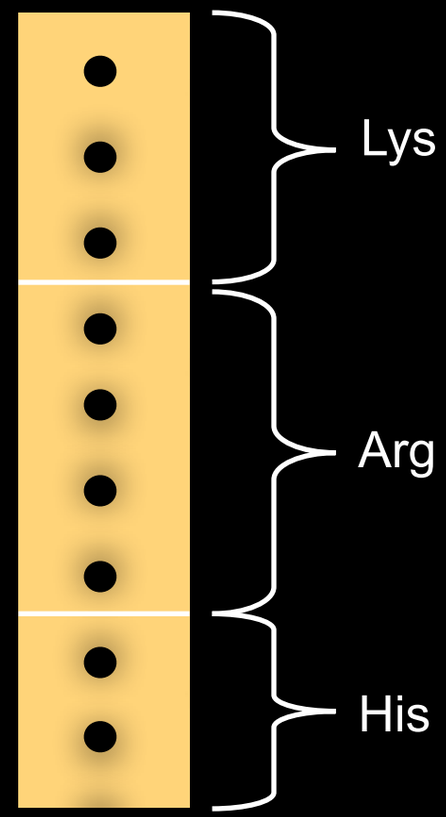
Design Problems

Want to design a sequence that will fold into a given backbone

- Output is an amino acid sequence

Assumption: a sequence that fits well onto this backbone will fold into it

Put rotamers for several amino acids into each graph part



Redesign Tests

- Redesigned 25 protein cores
  - Energy function best suited to solvent inaccessible residues

  ⇒ Fixed surface residues

- Group amino acids into classes:

  AVILMF / HKR / DE / TQNS / WY / P / C / G

- Problem sizes:
  - 11 to 124 residues
  - 552 to 6,655 rotamers

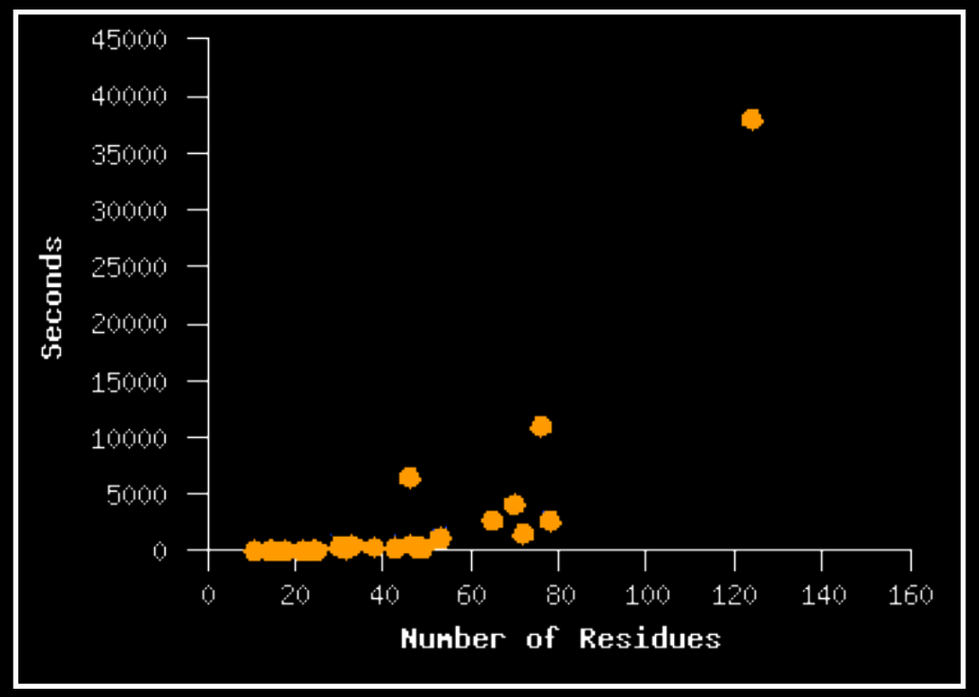
Design Results

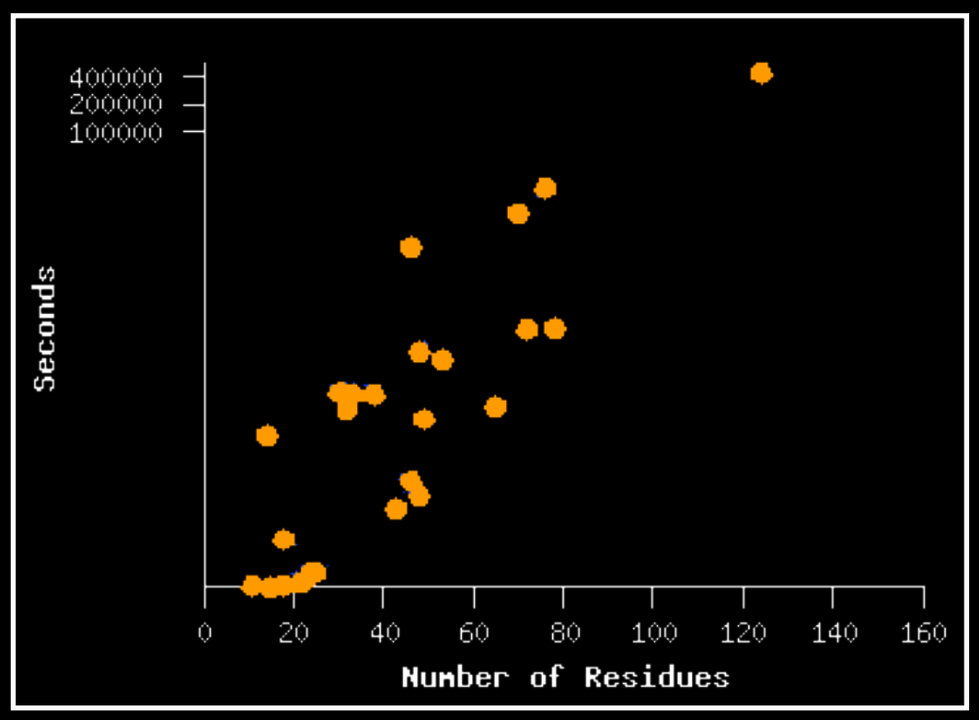
Redesigned 25 protein cores
- 11 to 124 residues
- 552 to 6,655 nodes

LP much slower (20 hours)

Only 6 integral out of 25

After DEE, can solve IP for remaining problems:
- one took 125 hours
- remaining 18 took 13 hours





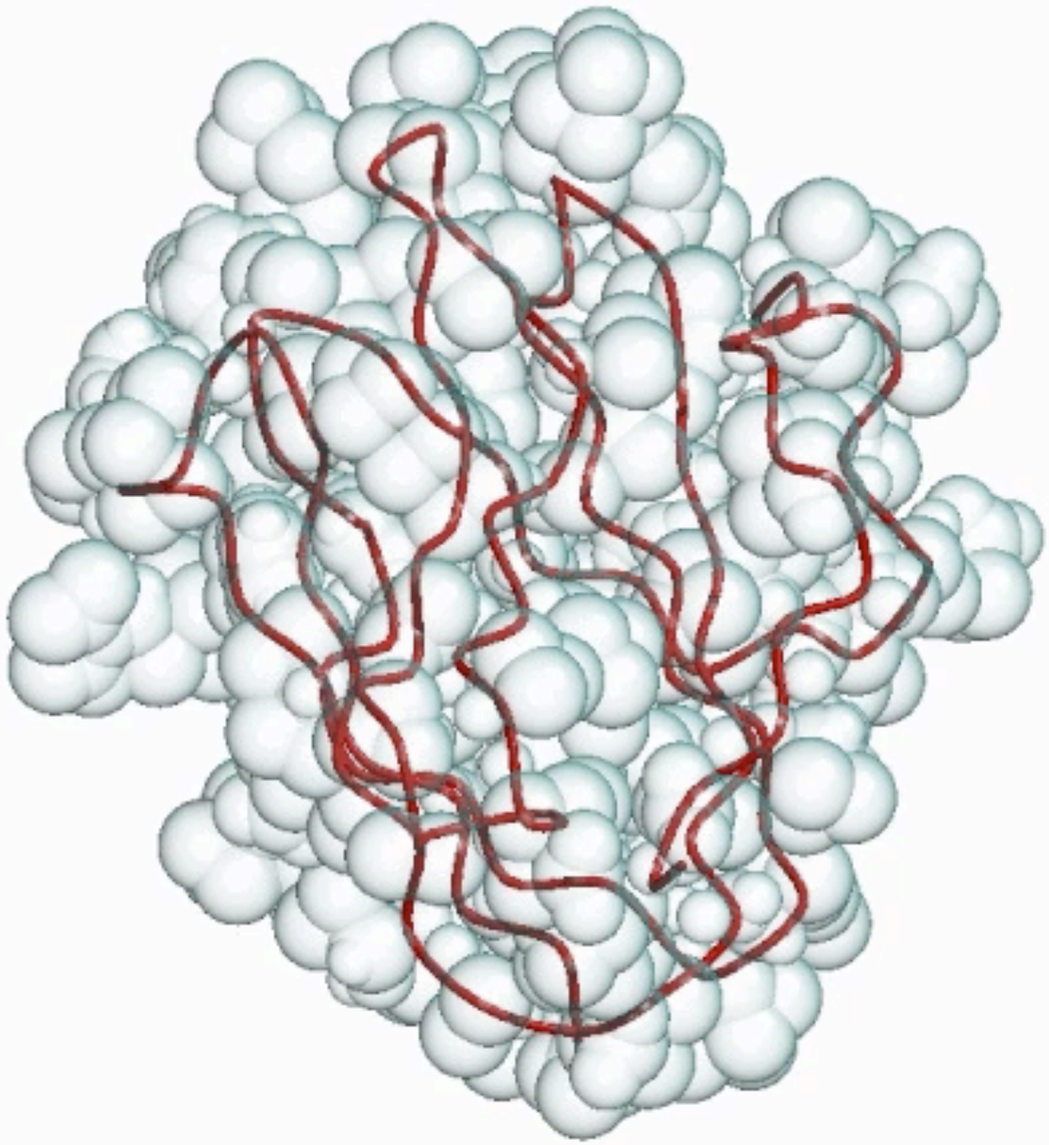
Near-Optimal Solutions

- Near-optimal solutions are useful:
  - Several candidates for protein design
  - Confidence in solution
- Can be found with integer program formulation
- To exclude m previously found solutions, add constraints:

$$\sum_{u \in S_k} x_u \leq p - 1 \quad \text{for } k = 1, \ldots, m$$

where S_k is set of chosen nodes for solution k

Near-Optimal Solutions



1aac - best 597 solutions.

⇦Required only that some residue change

- Can also require, say, core residue change
- Or force several residues to move at once

Thus,

- Side-chain positioning is a biologically useful problem with a nice combinatorial problem behind it
- Linear / integer programming effective method for finding optimal side-chain positions
- Empirical difficulty ≠ theoretical hardness
- Design problems appear to yield harder search problems than homology modeling