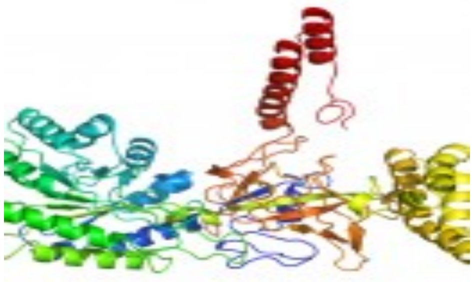
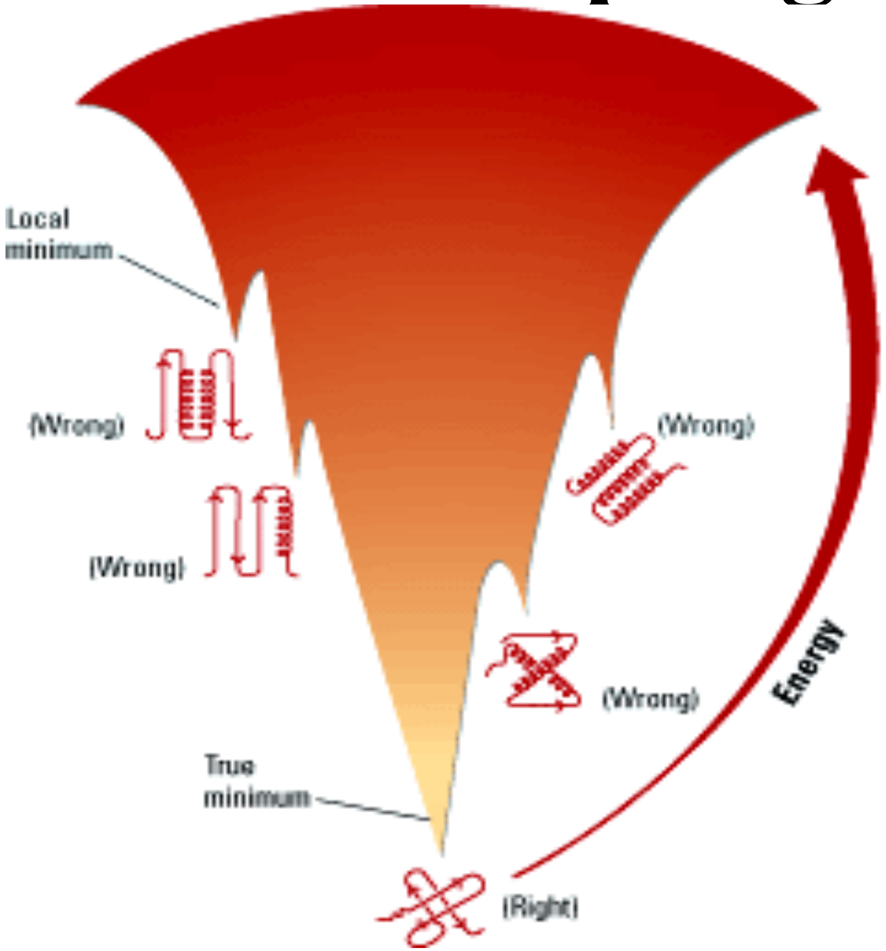


Template Free Protein Structure Modeling

Jianlin Cheng, PhD

Professor
Department of EECS
Informatics Institute
University of Missouri, Columbia
2018

Protein Energy Landscape & Free Sampling

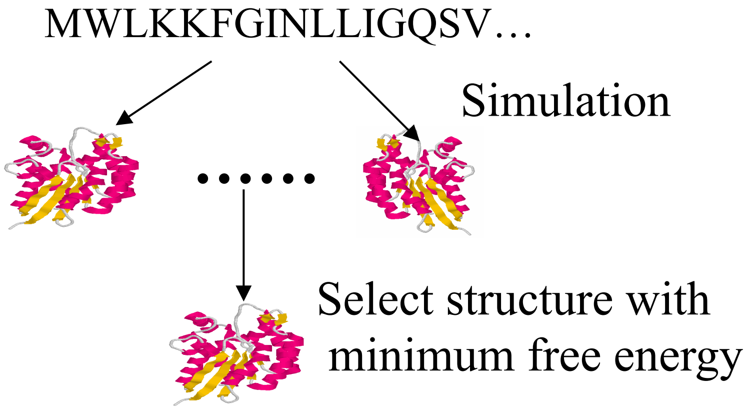


http://pubs.acs.org/subscribe/archive/mdd/v03/i09/html/willis.html

Two Approaches for 3D Structure Prediction

- **Ab Initio Structure Prediction**

  Physical force field – protein folding
  Contact map - reconstruction



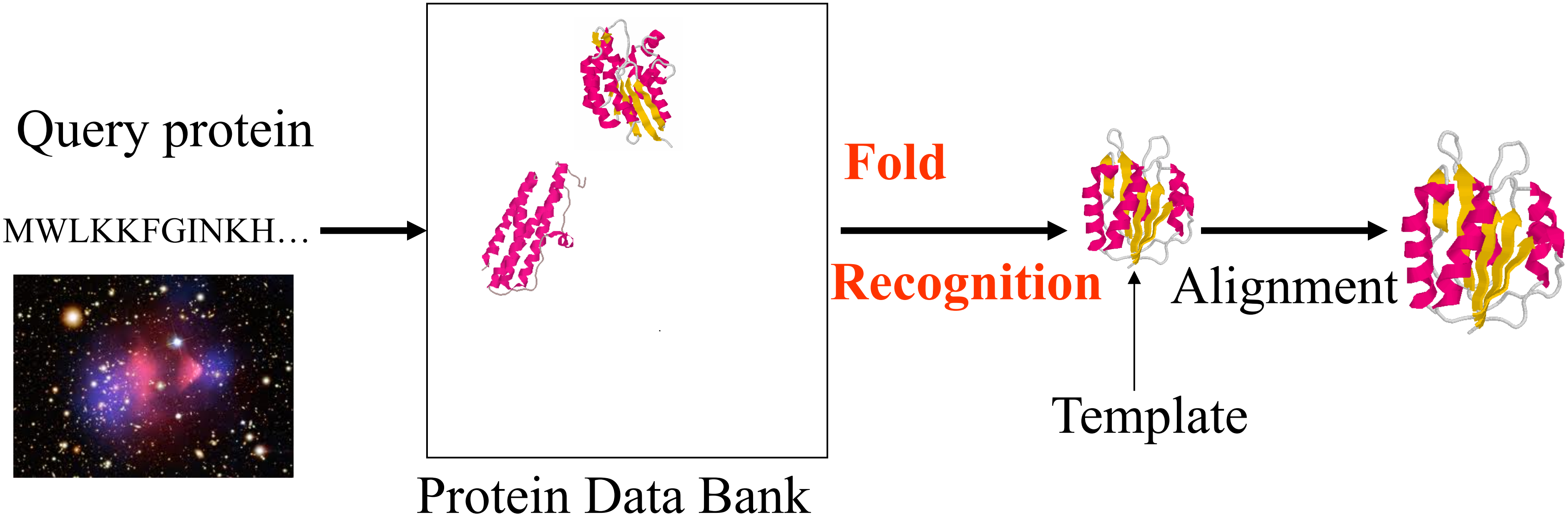
MWLKKFGINLLIGQSV…

Simulation

……

Select structure with minimum free energy

- **Template-Based Structure Prediction**



Query protein

MWLKKFGINKH…

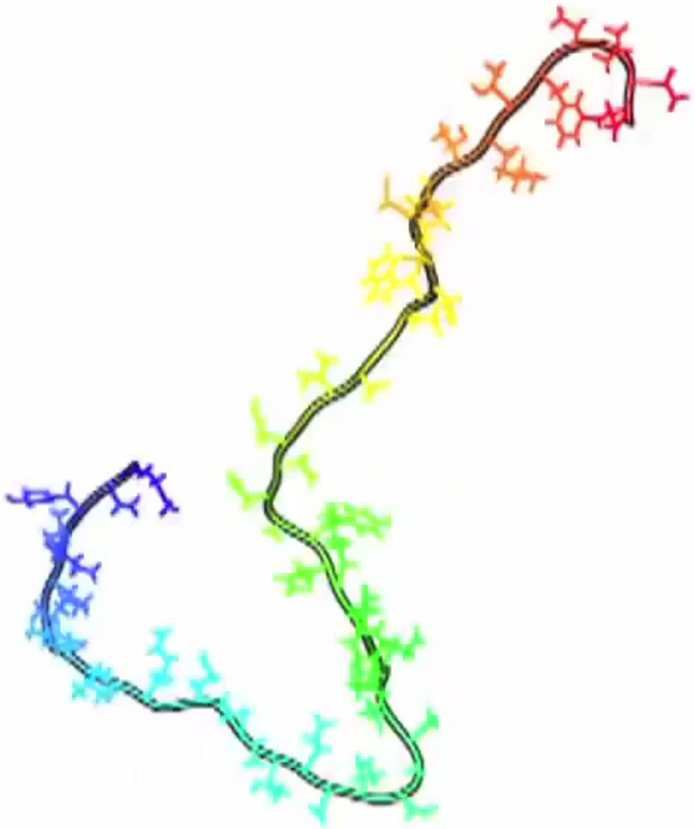
Protein Data Bank

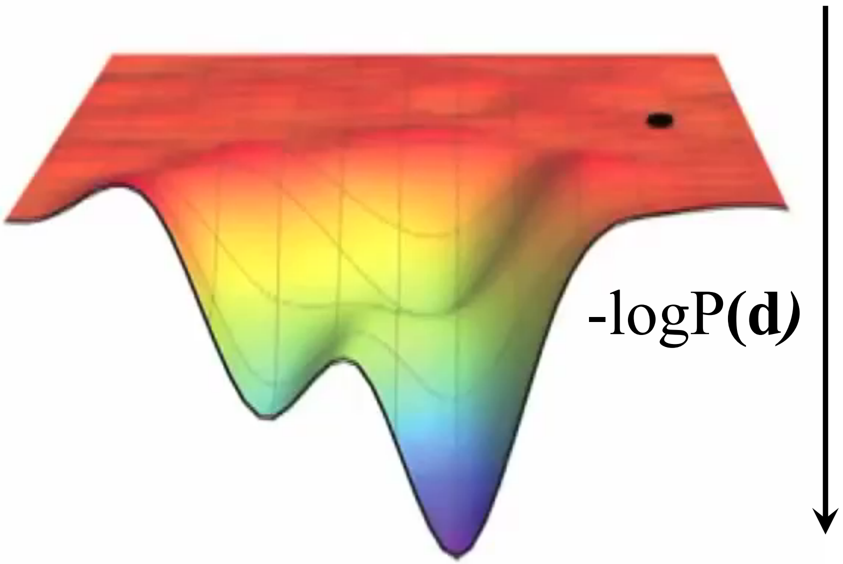
Fold Recognition

Template

Alignment

Demo of Our Protein Structure Prediction Software (FUSION)





Funnel-shaped landscape

Energy Functions

- T. Lazaridis, M. Karplus. Effective energy functions for protein structure prediction. Current Opinion in Structural Biology. 2000
- A. Liwo, C. Czaplewski, S. Oldiej, H.A. Scheraga. Computational techniques for efficient conformational sampling of proteins. 2008
- K. Simons et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. JMB. 1997. (Rosetta – a case study) -- reading assignment due Feb. 26

Protein Energy Function

- The native state of a protein is the state of lowest free energy under physiological conditions
- This state corresponds to the lowest basin of the effective energy surface.
- The term ‘effective energy’ refers to the free energy of the system (protein plus solvent)

Two Kinds of Energy Functions

- Physical effective energy function (PEEF): fundamental analysis of forces between particles
- Statistical effective energy function: data derived from known protein structures (e.g., statistics concerning pair contacts and surface area burial)

Statistical Effective Energy Function (SEEF)

- Less sensitive to small displacements
- Because of their statistical nature, they can, in principle, include all known and unrecognized, physical effects.
- Works better for protein structure prediction

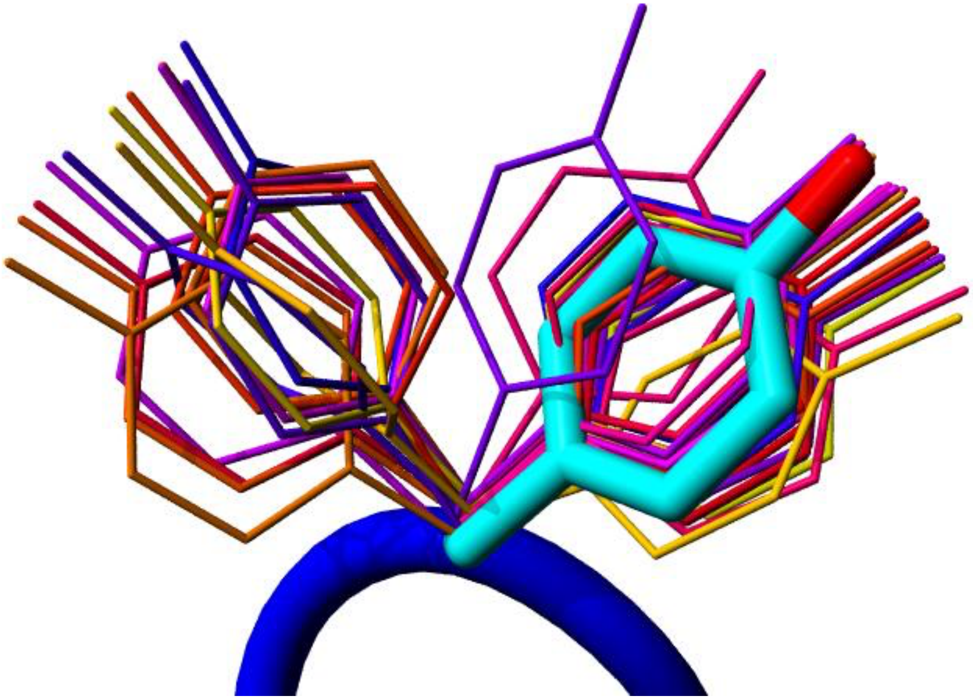
SEEF

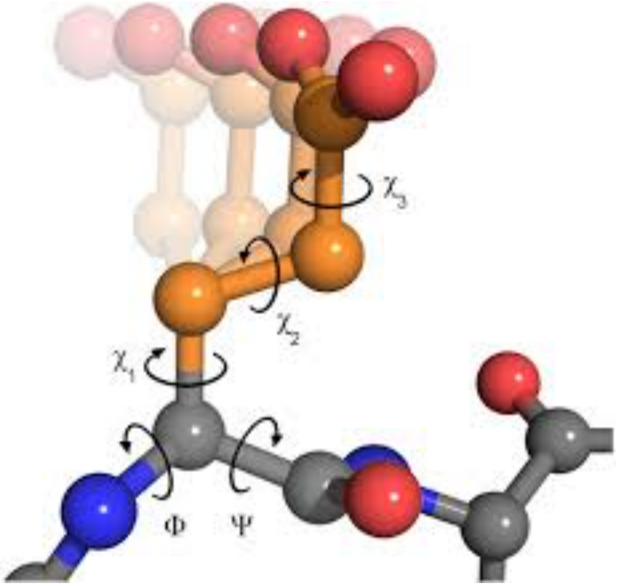
- Employ a reduced representation of the protein: a single interaction center at Ca or Cb for each residue.
- Basic idea: log (P_{ab} / P_a * P_b). P_{ab}: is the observed probability that residues a and b are in contact. P_a is frequency of a and P_b is the frequency of b
- Energy = -log (P_{ab} / P_a * P_b)
- More info: use secondary structure, solvent accessibility, distance as conditions.

Energy Terms

- Pairwise contact potentials
- Hydrogen bonds
- Torsion angle
- Burial energy (solvation energy)
- Sidechain orientation coupling, rotamer energy

Rotamer Energy





http://dunbrack.fccc.edu/scwrl4/

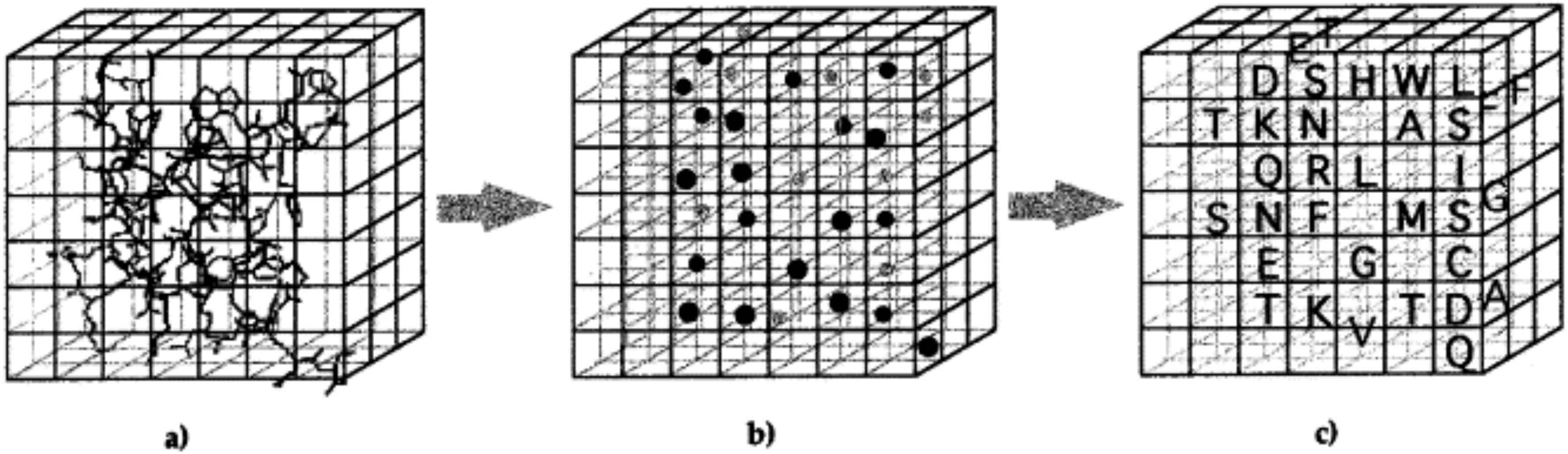
Physical / Statistical Effective Energy Function (PEEF)

- CHARMM implementation (http://www.charmm.org)
- AMBER implementation (http://ambermd.org)
- Dfire energy: http://sparks-lab.org/tools-dfire.html (program)
- RW energy: http://zhanglab.ccmb.med.umich.edu/RW/ (program available)

Benchmark

- Can a function select a native structure from a large pool of decoys?
- Can a function be used effectively in conformation sampling to generate a high proportion of near-native conformations?

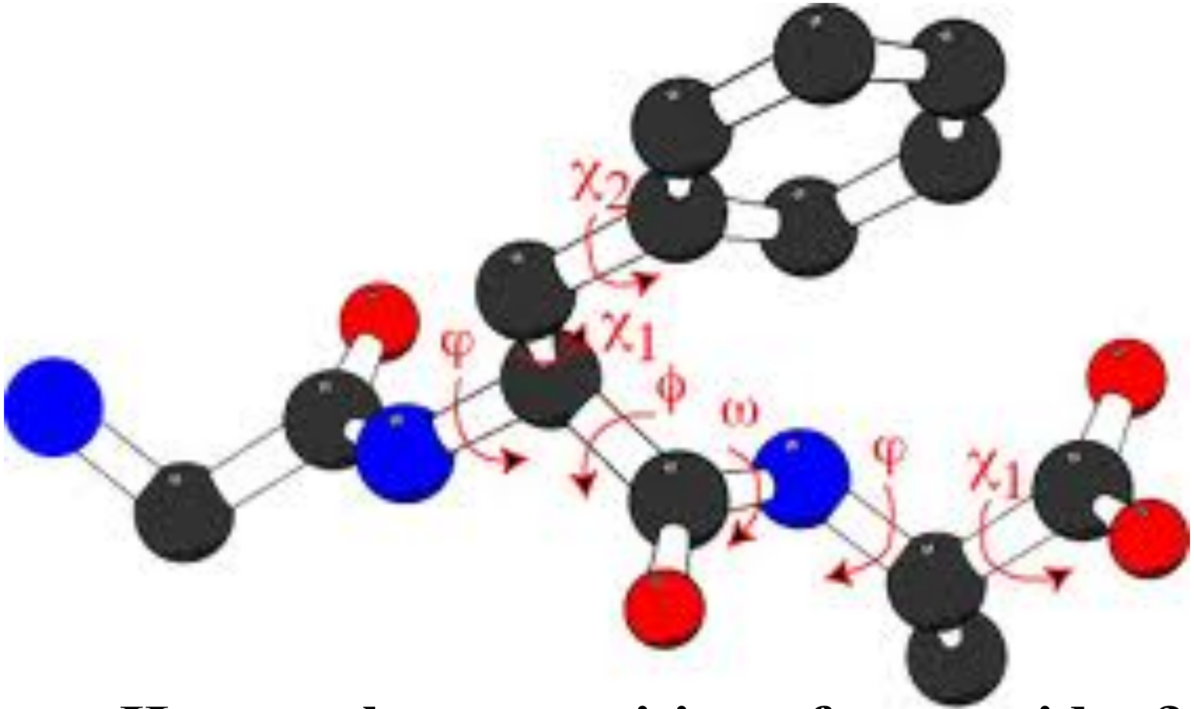
Representation for Conformation Sampling



How to change position of one residue?

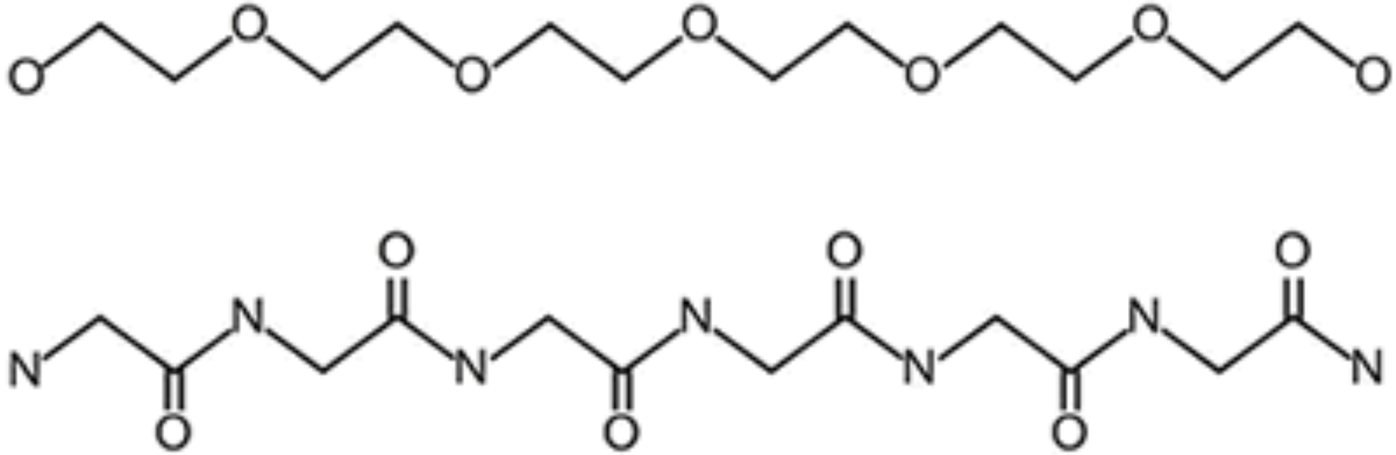
ITASSER: http://zhanglab.ccmb.med.umich.edu/I-TASSER/

Torsion Angles



How to change position of one residue?

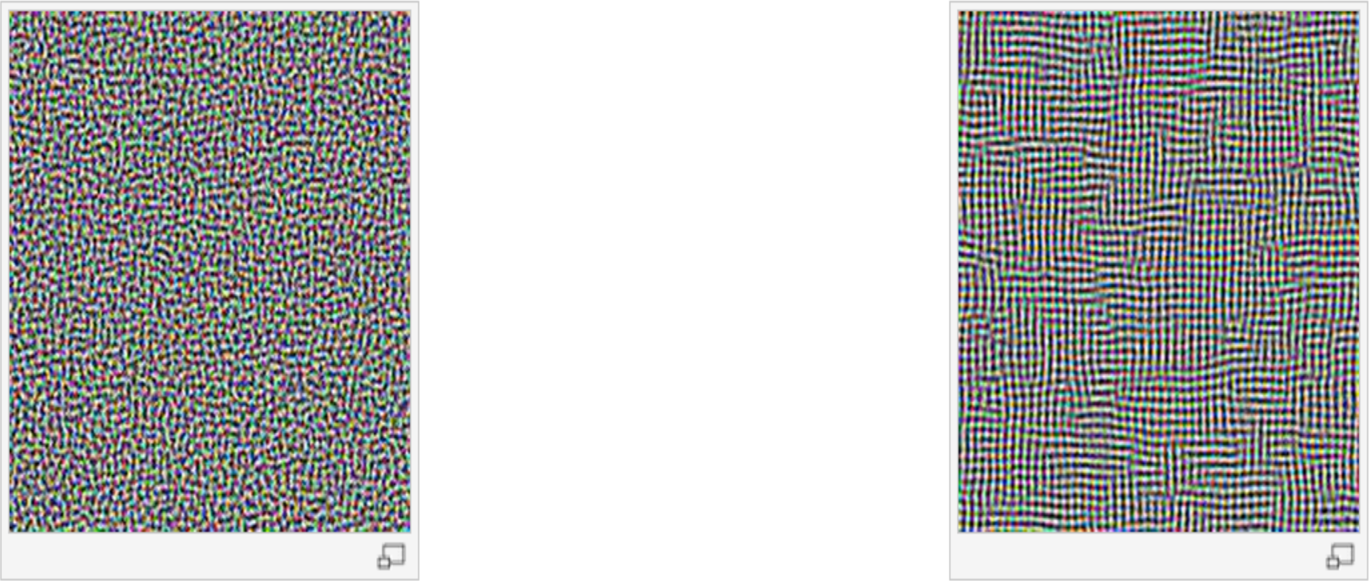
Vector Space



Simulated Annealing

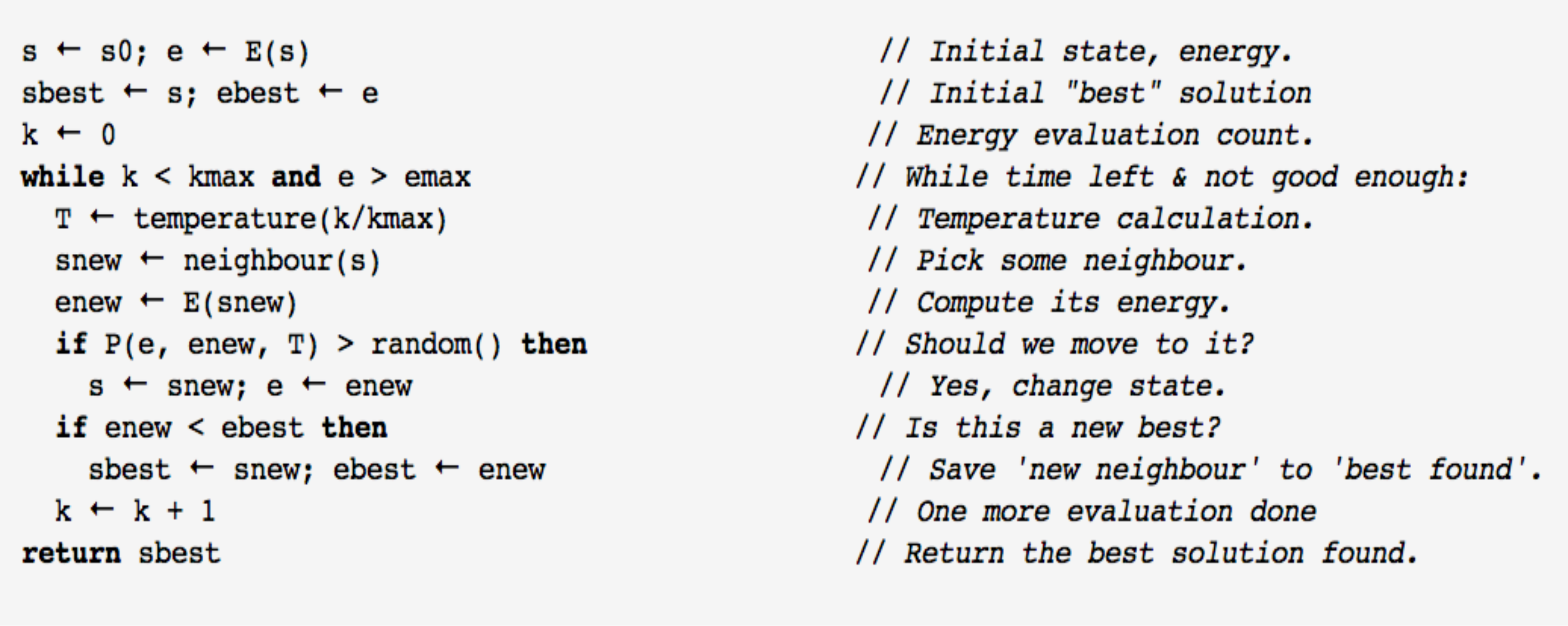
- Accept a move based on a probability related to temperature, e.g., P ~ e^ (-ΔE / T)
- Temperature (T) controls the degree of exploration. Higher temperature, more exploration? Why?
- Temperature decreases as the sampling process progresses (from iteration to iteration): cooling schedule

An Example



Example illustrating the effect of cooling schedule on the performance of simulated annealing. The problem is to rearrange the pixels of an image so as to minimize a certain potential energy function, which causes similar colours to attract at short range and repel at a slightly larger distance. The elementary moves swap two adjacent pixels. These images were obtained with a fast cooling schedule (left) and a slow cooling schedule (right), producing results similar to amorphous and crystalline solids, respectively.

Pseudo Code



```
s ← s0; e ← E(s)                                // Initial state, energy.
sbest ← s; ebest ← e                            // Initial "best" solution
k ← 0                                          // Energy evaluation count.
while k < kmax and e > emax                    // While time left & not good enough:
  T ← temperature(k/kmax)                      // Temperature calculation.
  snew ← neighbour(s)                          // Pick some neighbour.
  enew ← E(snew)                               // Compute its energy.
  if P(e, enew, T) > random() then             // Should we move to it?
    s ← snew; e ← enew                          // Yes, change state.
  if enew < ebest then                         // Is this a new best?
    sbest ← snew; ebest ← enew                  // Save 'new neighbour' to 'best found'.
  k ← k + 1                                    // One more evaluation done
return sbest                                   // Return the best solution found.
```

A TFM Example: Rosetta

- K. Simons, C. Kooperberg, E. Huang, D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. JMB, 1997.

Rosetta: https://www.rosettacommons.org

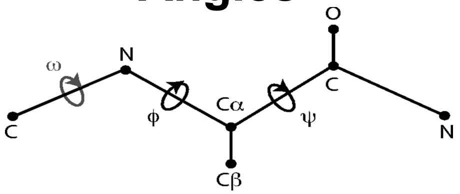
Reading assignment due: Feb. 22

Basic Idea

- Short sequence segments are restricted to the local structures adopted by the most closely related sequences in the PDB
- Use the observed local conformations of similar local sequences to reduce sampling space

Fragment Assembly (e.g. Rosetta)

Angles



Coordinates

↔ (X, Y, Z)

SDDEVYQYIVSQ**V**KQYGIEPAELLSRK**Y**GDK
AKYHLSQ

Randomly
pick 9 residues

Reduce
search
space!

9-Residue Fragment DB

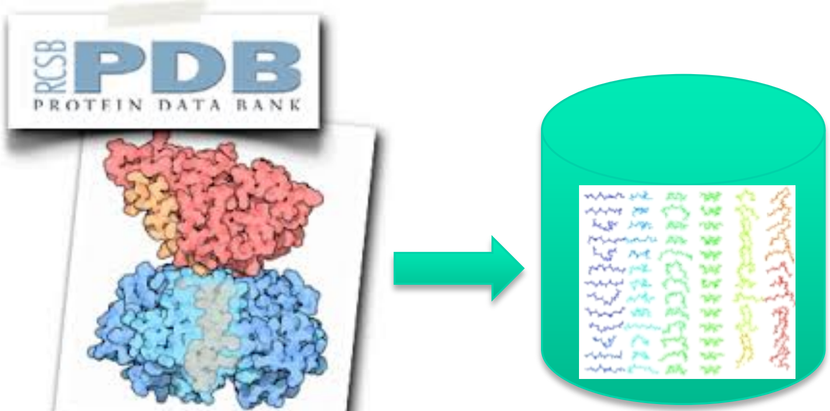
Fragment	Angles
SDDEQYQRK	(130,-120, …)
….	
….	

Find a similar fragment
Replace angles

Two ways of obtaining fragments

- Database-based approach: https://www.rosettacommons.org
- Model-based approach: http://sysbio.rnet.missouri.edu/FRAGSION/

Shortcomings of Fragment Assembly Approach Based on Database Search



~80,000 proteins

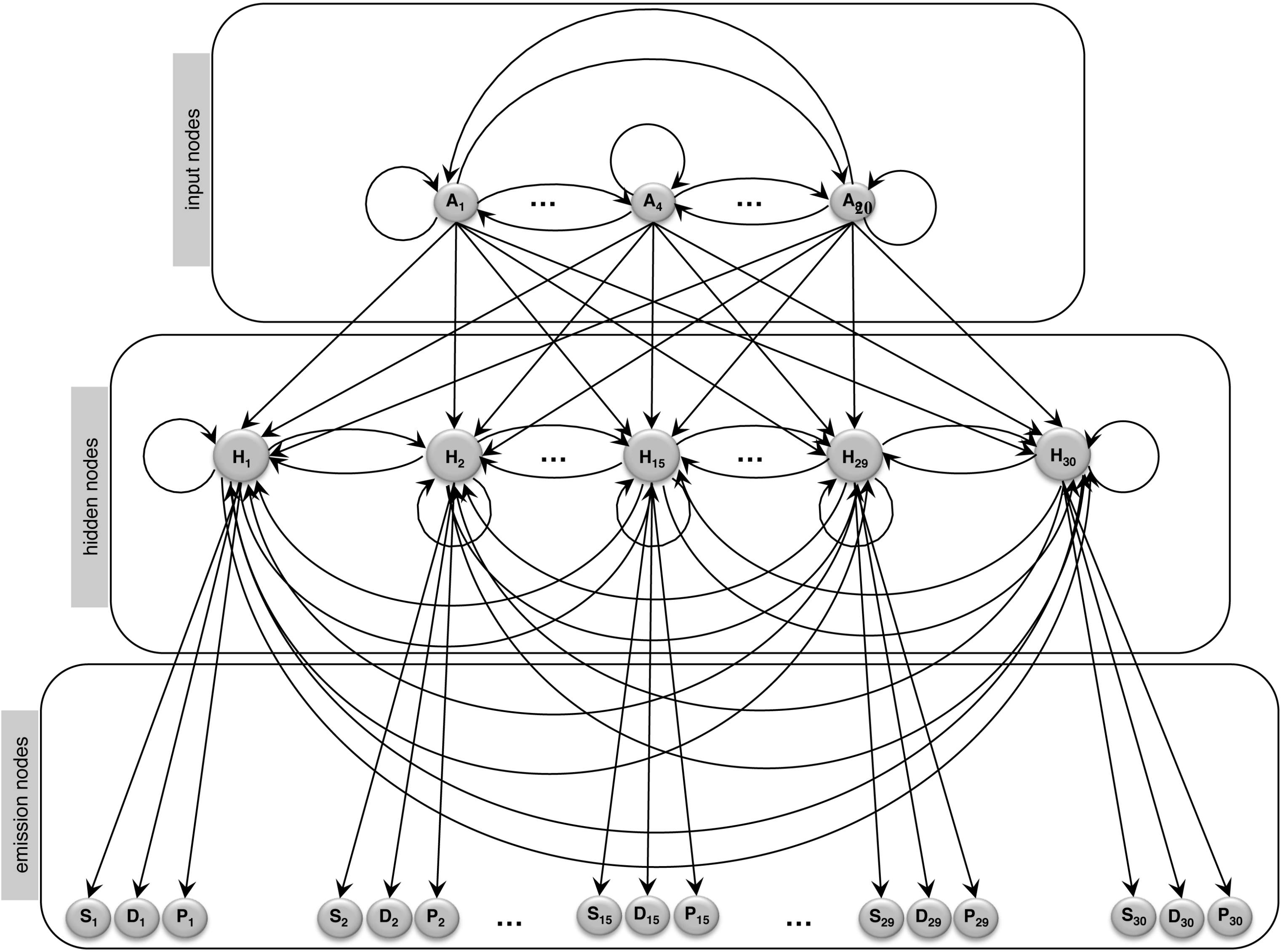
Fragment Structure Database

- Incomplete coverage
- Computationally expensive
- Restricted to small proteins

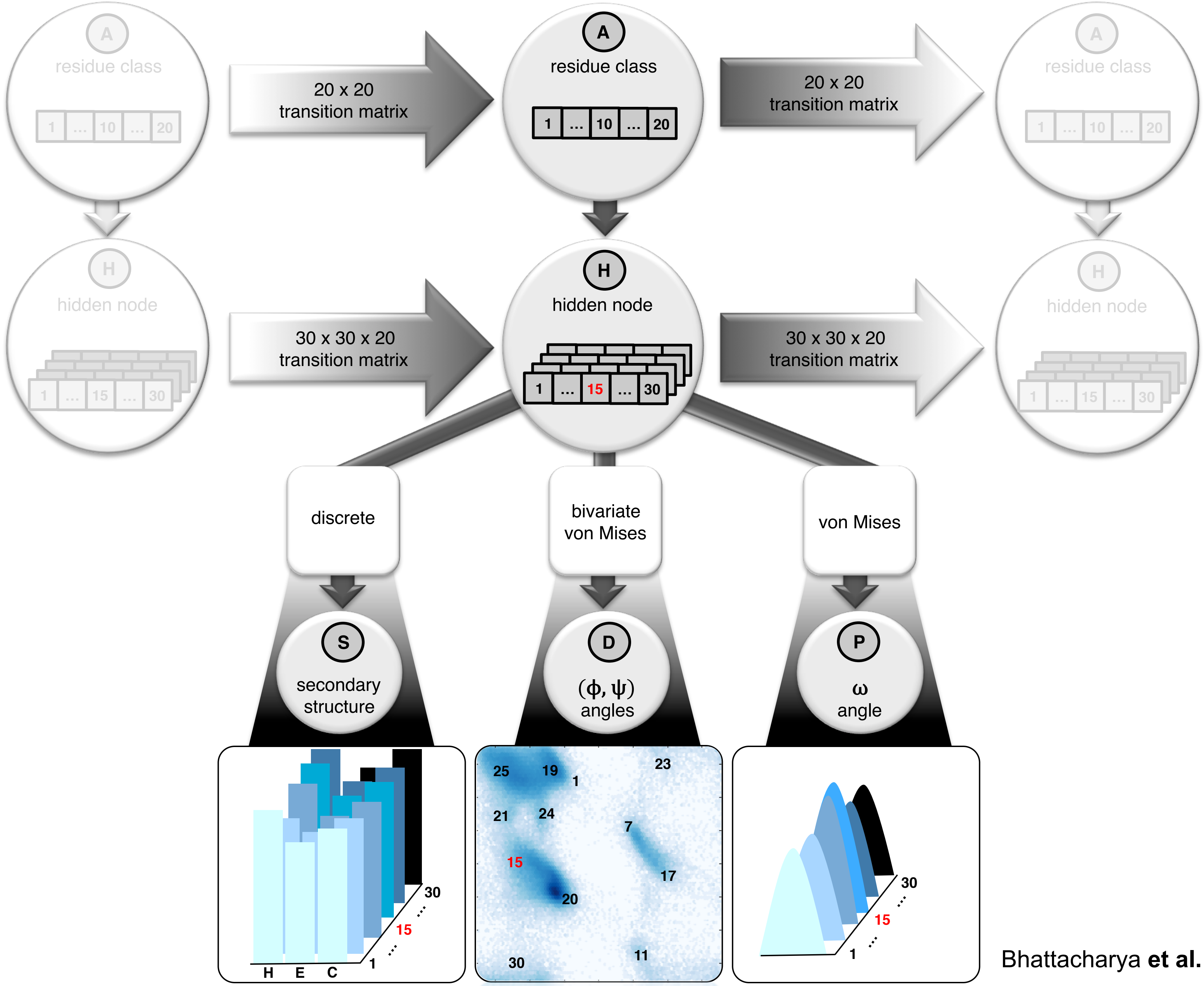
IOHMM (Input-Output Hidden Markov Model)

to model protein conformational space

Bhattacharya & Cheng, Bioinformatics, 2016
Bhattacharya & Cheng, Scientific Reports, 2015



$$P(\mathbf{S},\mathbf{D},\mathbf{P}) = \sum_{\mathbf{H}} P(\mathrm{H}_1 \mid \mathrm{A}_1)\, P(\mathrm{S}_1 \mid \mathrm{H}_1)\, P(\mathrm{D}_1 \mid \mathrm{H}_1)\, P(\mathrm{P}_1 \mid \mathrm{H}_1) \prod_{\mathrm{i}=2}^{n} P(\mathrm{H}_i \mid \mathrm{H}_{i-1}, \mathrm{A}_i)\, P(\mathrm{S}_i \mid \mathrm{H}_i)\, P(\mathrm{D}_i \mid \mathrm{H}_i)\, P(\mathrm{P}_i \mid \mathrm{H}_i)$$



A residue class

20 x 20 transition matrix

A residue class 1 ... 10 ... 20

20 x 20 transition matrix

A residue class

H hidden node

30 x 30 x 20 transition matrix

H hidden node 1 ... 15 ... 30

30 x 30 x 20 transition matrix

H hidden node

discrete

bivariate von Mises

von Mises

S secondary structure

D (ϕ, ψ) angles

P ω angle

Bhattacharya **et al.** 2015

Parameter Learning

using EM algorithm

- **1,740 experimentally solved proteins**
- **270,350 observations**
- **Training using stochastic EM algorithm**

Bhattacharya **et al.** 2015; Van **et al.** 2005; Paluszewski **et al.** 2010

Selecting optimal model

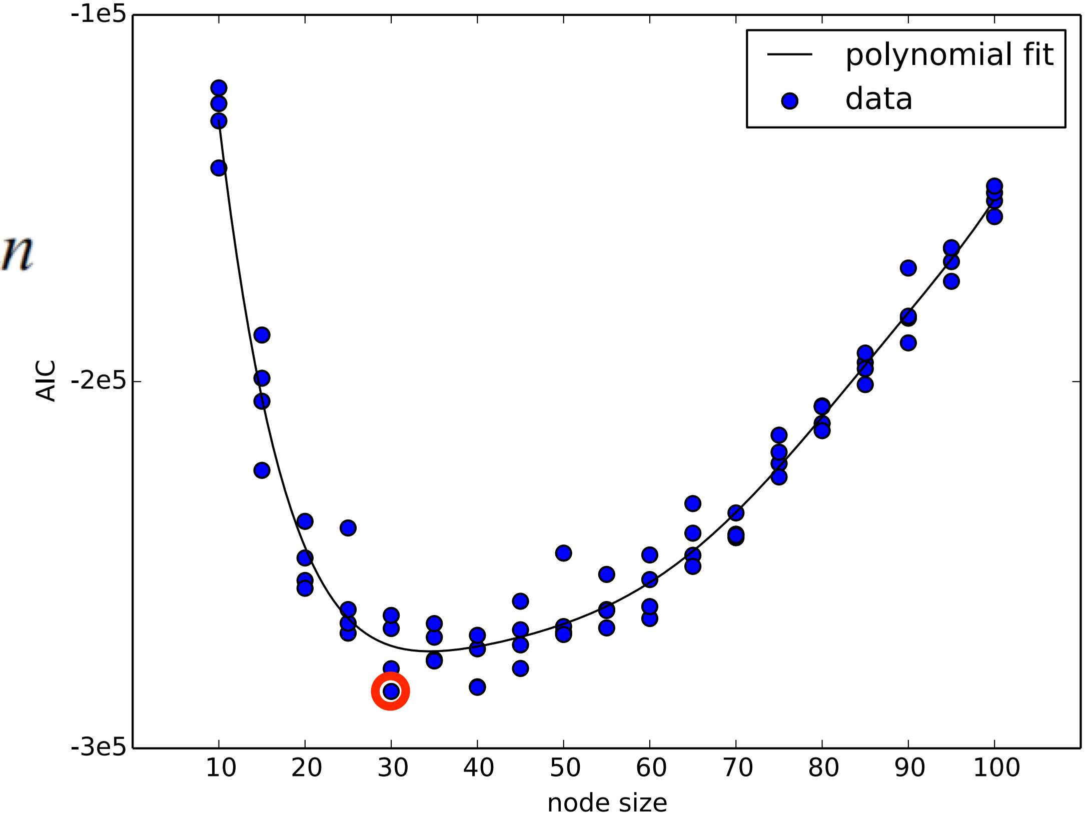
using information theory

$$AIC(n) = -2\log L\,(d|m) + 2n$$

$\Big\{$ $L(d|m)$: likelihood

d : data

n : parameters



30 hidden nodes
7,812 parameters

Bhattacharya **et al.** 2015; Burnham **et al.** 2002

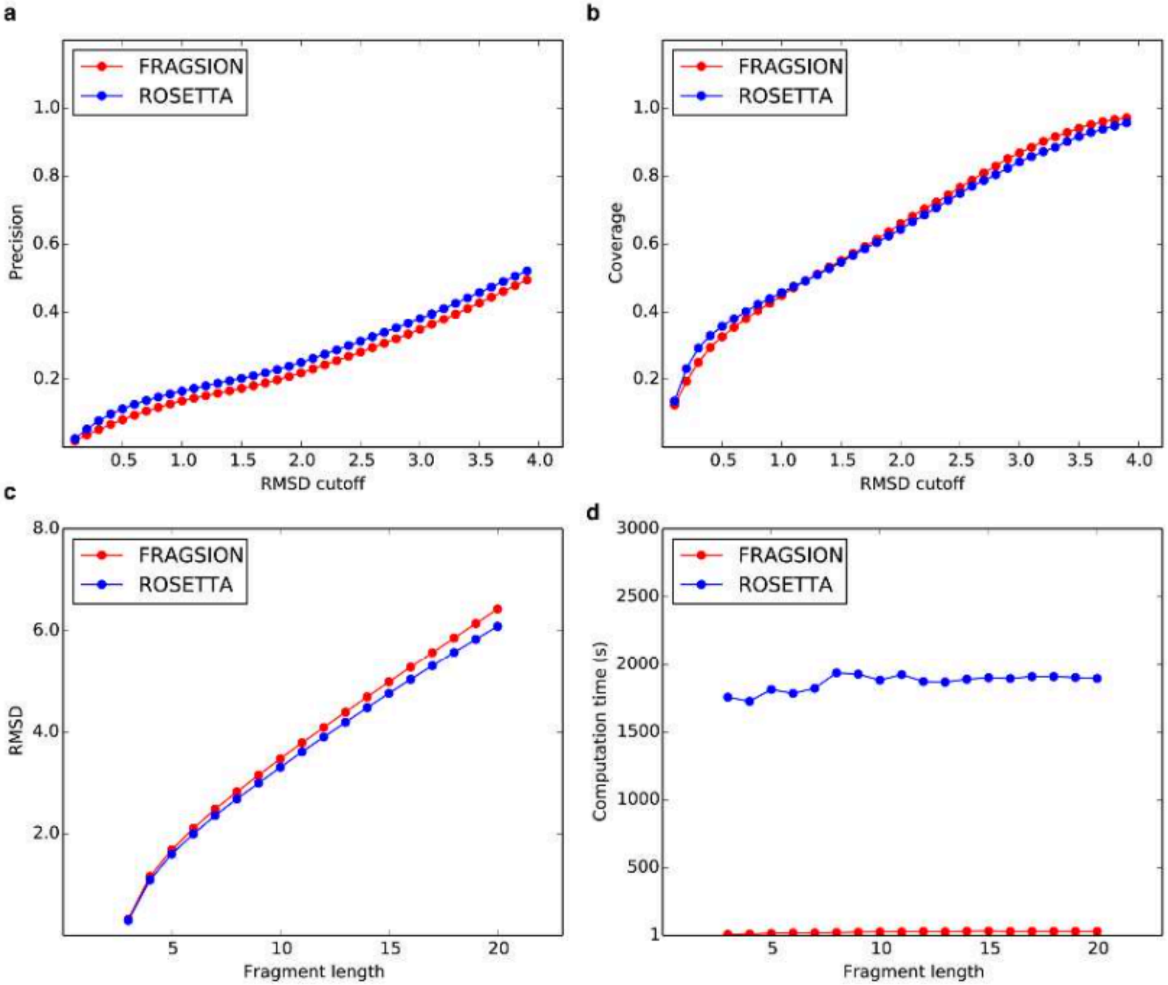


Fig. 1. Comparison between FRAGSION and ROSETTA. Precision (a), coverage (b) at various RMSD cutoffs and RMSD (c), computation time (d) at different fragment lengths averaged over the dataset generated by FRAGSION (red) and ROSETTA (blue).

Function of IOHMM Model of Protein Conformation

- Sample the conformation of a (sub) sequence of any size
- Software: Fragsion: http://sysbio.rnet.missouri.edu/FRAGSION/

Protein Folding Video

- https://www.youtube.com/watch?v=HBONCqN9U4k

Scoring Functions of Selecting Local Conformations

- Knowledge-based potential functions
- Bayesian scoring function

$$P(structure \mid sequence) = P(structure) \times \frac{P(sequence \mid structure)}{P(sequence)}$$

One native assumption is P(structure) = 1 / # of structures.

P(a structure)

- 0 for configurations with overlaps between atoms
- Proportional to exp(-radius of gyration^2) for all other configurations.
- Independent of secondary structure elements

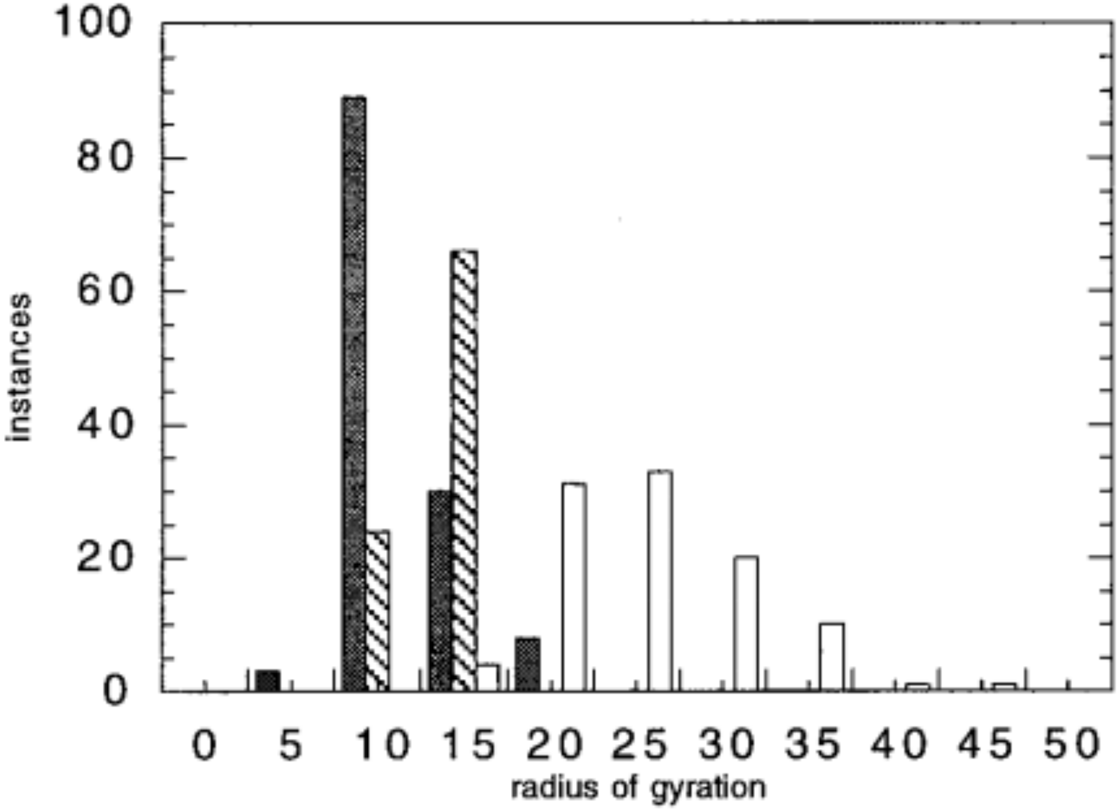
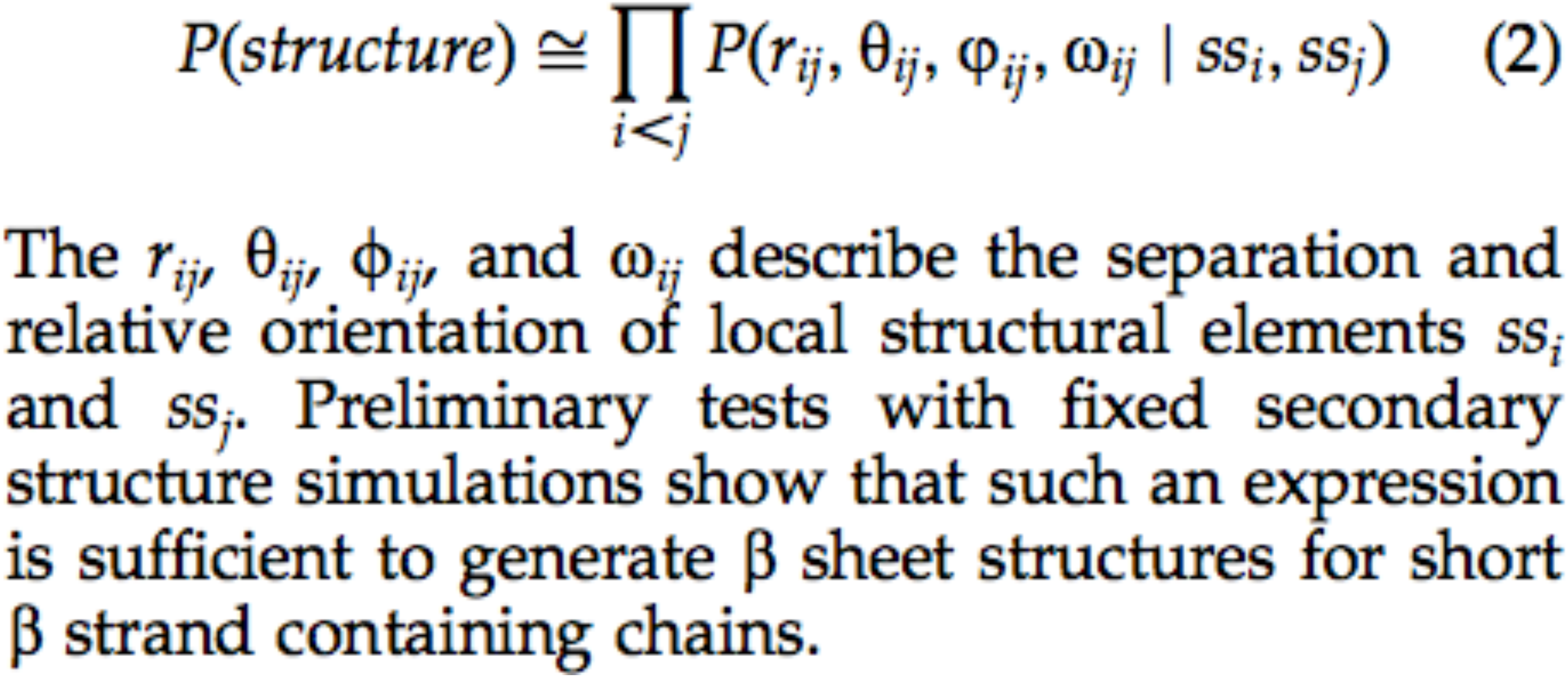


Figure 1. Comparison of the radii of gyrations of simulated and native structures. 100 structures were generated for chains of 100 residues by splicing together protein fragments as described in Methods using either no scoring function (open bars), or the square of the radius of the gyration as the scoring function (hatched bars). Histograms were computed using 5 Å bins. The distribution of radii of gyrations for the small (50 to 150 residue) proteins in the pdbselect 25 set is shown for comparison (filled bars).

Considering Beta-Sheet Pairing



$$P(structure) \cong \prod_{i<j} P(r_{ij}, \theta_{ij}, \varphi_{ij}, \omega_{ij} \mid ss_i, ss_j) \qquad (2)$$

The r_{ij}, θ_{ij}, ϕ_{ij}, and ω_{ij} describe the separation and relative orientation of local structural elements ss_i and ss_j. Preliminary tests with fixed secondary structure simulations show that such an expression is sufficient to generate β sheet structures for short β strand containing chains.

Scoring – P(Sequence | Structure)

$$P(aa_1, aa_2, \ldots, aa_n \mid \mathit{structure}) \cong \prod_i P(aa_i \mid E_i)$$

$$\times \prod_{i<j} \frac{P(aa_i, aa_j \mid r_{ij}, E_i, E_j)}{P(aa_i \mid r_{ij}, E_i, E_j) P(aa_j \mid r_{ij}, E_i, E_j)} \tag{8}$$

E_i can represent a variety of features of the local structural environment around residue i.

Implementation

- Second term: for pairs separated for more than 10 residues along the chain
- Buried environment: >16 other Cb atoms within 10 Angstrom of the Cb atom of the residue; otherwise, exposed

Negative Log of Interaction Probability Function

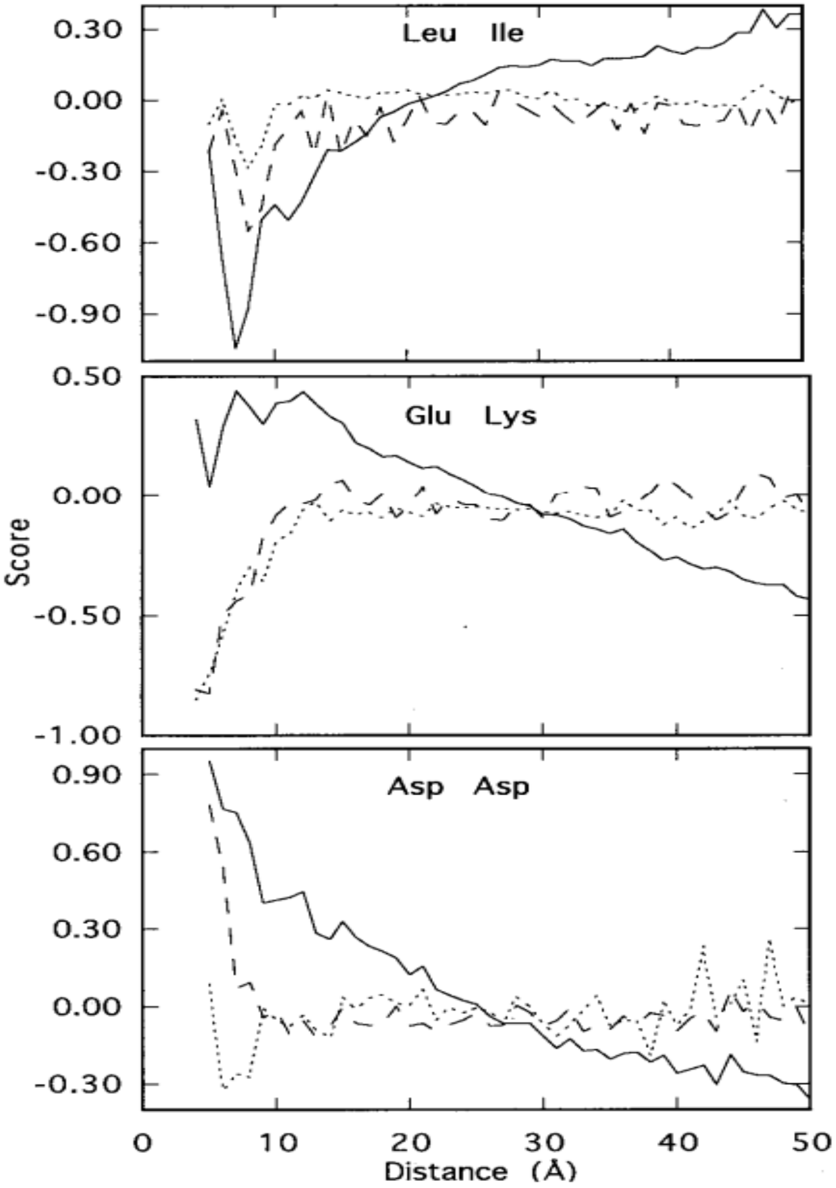


Figure 4. Comparison of the negative logarithms of equation (5) and the residue pair specific second term in equation (8) for sequence separations greater than ten. Residues with greater than 16 neighbors were considered buried. Continuous lines, equation (5); dotted lines, equation (8) both residues buried; broken line, equation (8) both residues exposed.

Structure Generation

- Initialization:

$$P(structure \mid sequence) \cong e^{-radius\ of\ gyration^2} \times \prod_{i<j} \frac{P(r_{ij} \mid aa_i, aa_j)}{P(r_{ij})} \quad (6)$$

Splicing together fragments of proteins of known structure with similar local sequences and evaluating them initially using equation.

Simulated Annealing

- Low scoring conformations with distributions of residues similar to those of known proteins are resampled by simulated annealing in conjunction with a simple move set that involves replacing the torsion angles of a segment of the chain with the torsion angles of a different protein fragment with a related amino acid sequence.
- The simulated conformation is evaluated by (8)

Methods

- Structures are represented using a simplified model consisting of heavy atoms of the main-chain and the C_b atom of the side chain.
- All bond lengths and angles are held constant according to the ideal geometry of alanine (Engh & Huber 91); the only remaining variables are the backbone torsional angles.

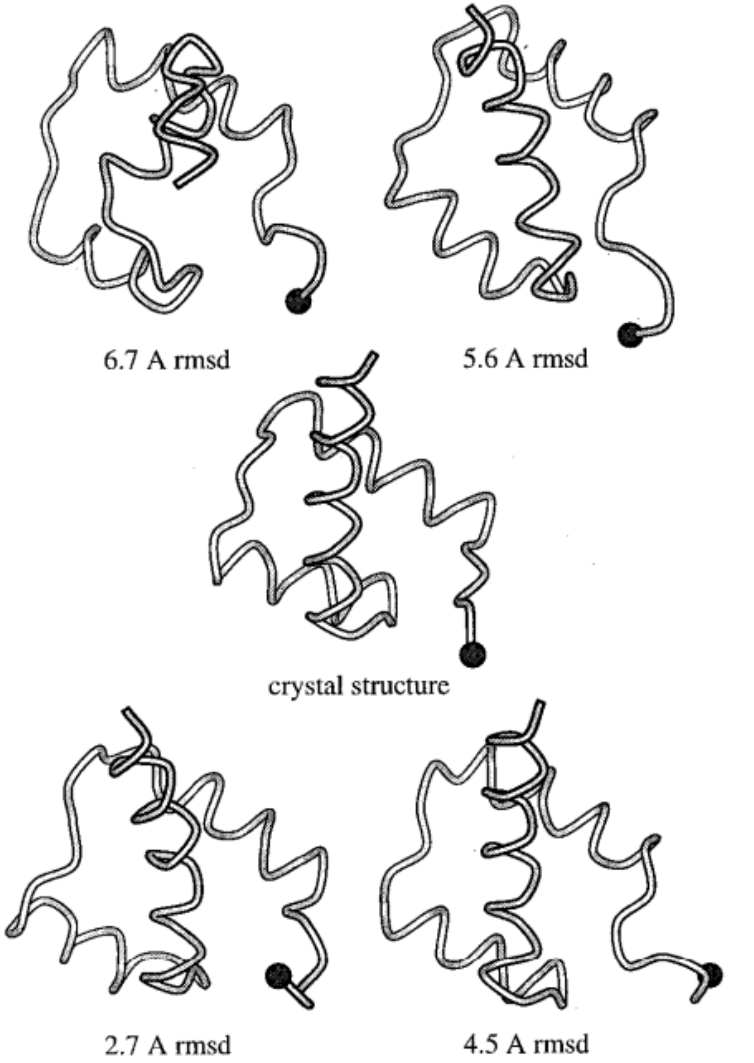
Fragment Databases

- Nimers / trimers (sequences) and their conformations extracted from known structures in the database
- Identify sequence neighbors: simple amino acid frequency matching score.

Simulation

- The starting configuration in all simulations was the fully extended chain.
- A move consists of substituting the torsional angles of a randomly chosen neighbor at a randomly chosen position for those of the current configuration.
- Moves which bring two atoms within 2.5 Angstrom are immediately rejected; other moves are evaluated according to the Metropolis criterion using the scoring equation.
- Simulated annealing was carried out by reducing the temperature from 2500 to 10 linearly over the course of 10,000 cycles (attempted moves).

Simulated Structure Examples



6.7 A rmsd

5.6 A rmsd

crystal structure

2.7 A rmsd

4.5 A rmsd

Figure 5. Simulated homeodomain structures with different rms deviations from the native structure. The N termini are displayed as black spheres.

Table 1. Folding simulation results

	<7 Å	<6 Å	<5 Å	<4 Å	Lowest	
	rmsd	rmsd	rmsd	rmsd	rmsd	*Q*
A. *Unconstrained simulations*						
Homeodomain						
dist_env filter + msa (100)	65	47	31	17	2.75	−1.7
dist_env filter − msa	63	45	31	16	2.75	−1.8
No filter	63	48	38	8	2.75	−1.5
Random sequence	31	11	1	0	4.89	−0.2
Random fragments	16	4	1	0	4.73	−0.6
Random all	6	2	0	0	5.82	0
Calbindin						
dist_env filter + msa (64)	31	17	2	0	4.70	−1.7
dist_env filter − msa	24	14	1	0	4.70	−1.9
No filter	17	3	2	0	4.86	−1.4
Random sequence	3	0	0	0	6.18	−0.2
Random fragments	6	1	0	0	5.71	−0.4
Random all	0	0	0	0	7.63	0
Protein A						
dist_env filter	96	95	93	41	3.29	−2.3
No filter	86	85	77	41	3.16	−2.0
Random sequence	33	25	8	1	3.52	−0.2
Random fragments	48	32	9	1	3.97	−0.6
Random all	32	14	1	0	4.58	0
Cro repressor						
dist_env filter + msa (4)	39	18	8	0	4.20	−1.7
dist_env filter − msa	35	20	10	0	4.20	−1.9
No filter	24	11	4	0	4.26	−1.5
Random sequence	7	1	0	0	5.95	−0.3
Random fragments	5	0	0	0	6.14	−0.7
Random all	0	0	0	0	7.26	0
Protein G						
dist_env filter + msa (5)	3	0	0	0	6.33	−1.5
dist_env filter − msa	2	0	0	0	6.33	−1.5
No filter	1	0	0	0	6.89	−1.2
Random sequence	0	0	0	0	8.43	−0.4
Random fragments	0	0	0	0	7.80	−0.6
Random all	0	0	0	0	8.35	0

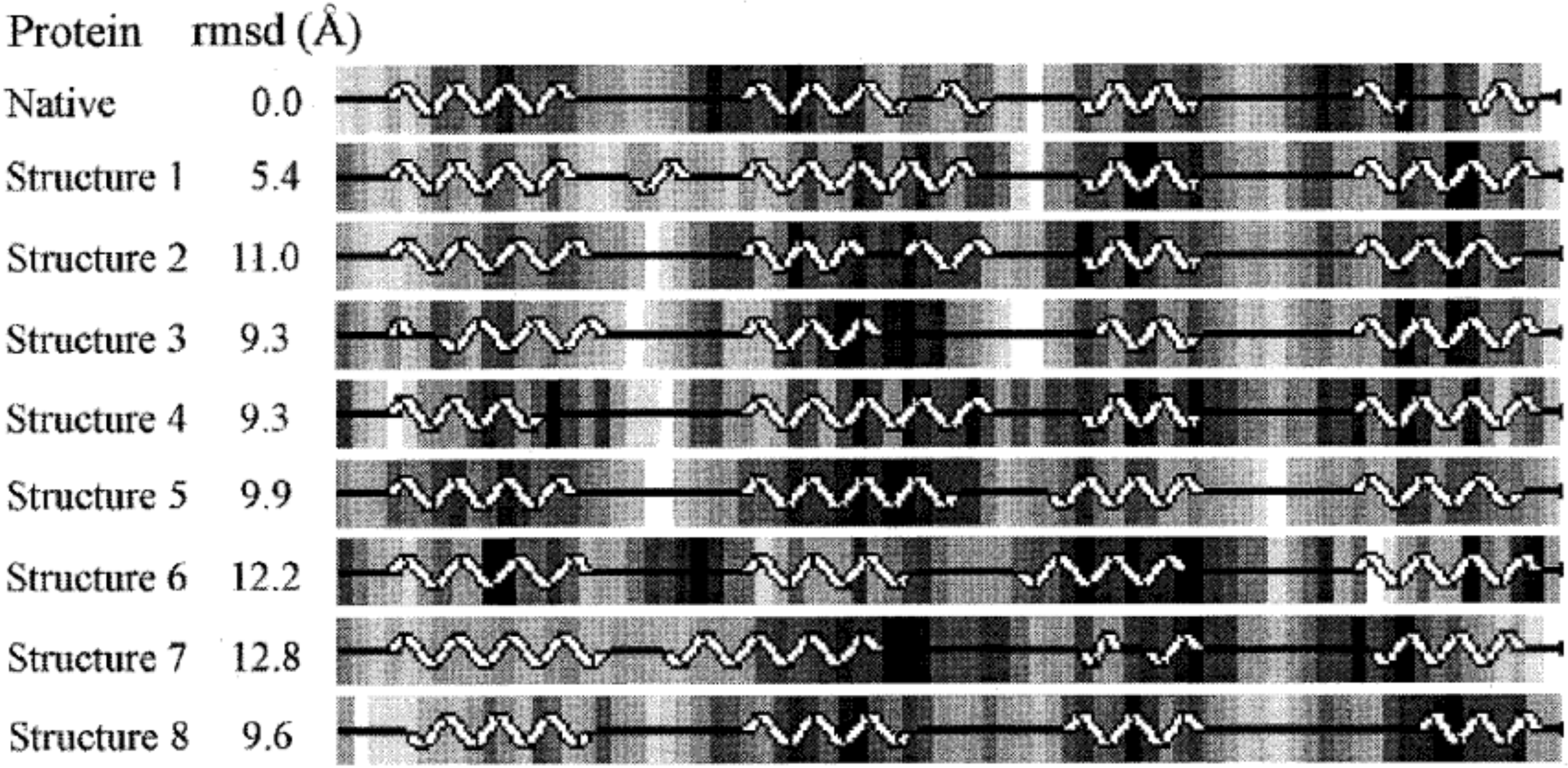


Figure 6. Solvent accessibility and secondary structure of a number of simulated non-native calbindin structures as depicted by PROCHECK (Laskowski *et al.*, 1993). The structures were randomly drawn from the simulated structure set prior to filtering. The rmsd to the native structure is shown in the second column; the rmsd between all pairs of structures is greater than 5 Å. White, solvent accessible; black, buried.

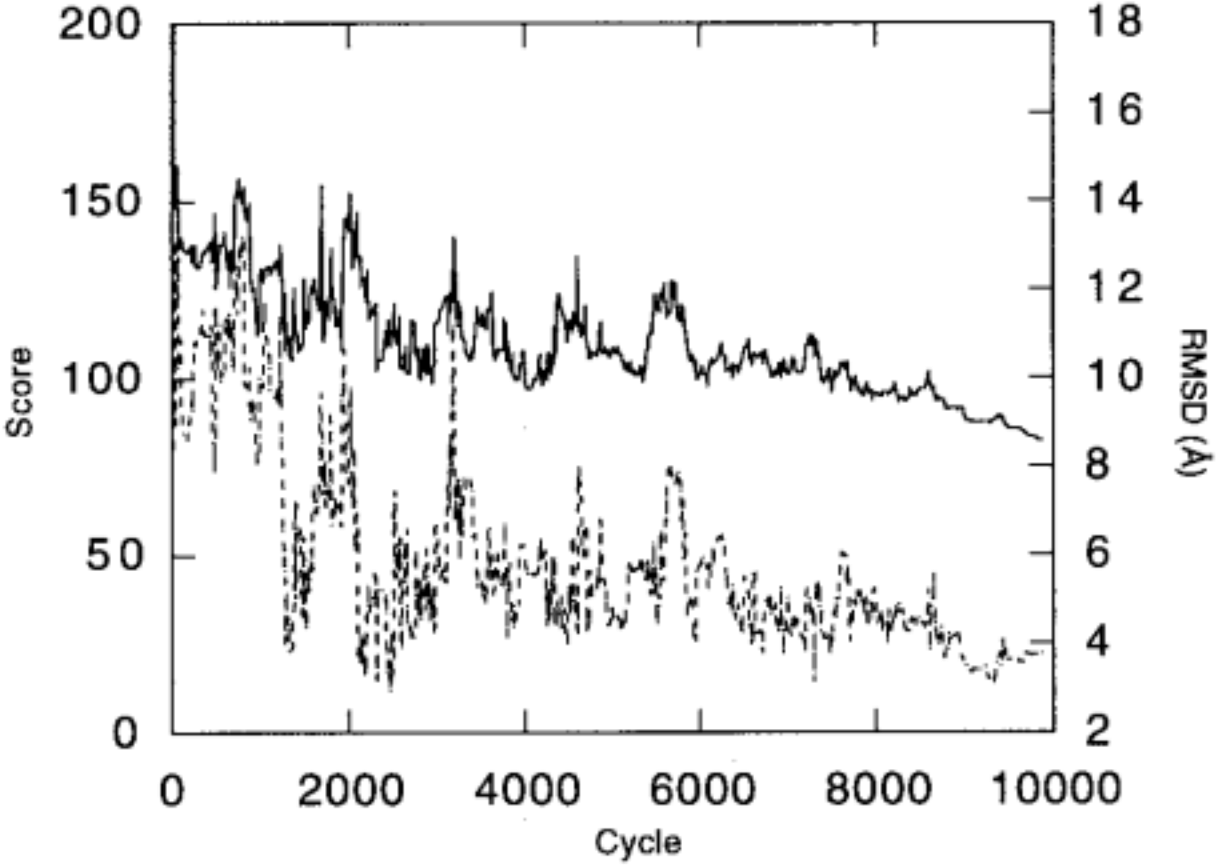


Figure 7. Progression of a homeodomain folding simulation. Continuous line, score; broken line, rmsd from the native structure. A cycle is an attempted replacement of the current torsion angles of a segment of the structure with the torsion angles of a fragment from the protein database with similar local sequence.

Table 2. Origins of fragments contributing to final simulated structures

	Structure I	Structure II
Residue	(2.7 Å rmsd, 2.1 Å dme)	(3.0 Å rmsd 2.1 Å dme)
1	Methyltransferase (1hmy)	Endonuclease III (1abk)
2	Creatinase (1chm)	Endonuclease III (1abk)
3	Cytochrome *c* (1ccr)	Endonuclease III (1abk)
4	Cytochrome *c* (1ccr)	Recoverin (1rec)
5	Cytochrome *c* (1ccr)	Recoverin (1rec)
6	Barley seed protein (1bw4)	Recoverin (1rec)
7	Hydrolase inhibitor (1hle)	3-isopropyl malate DH (1hex)
8	Ribose binding protein (2dri)	3-isopropyl malate DH (1hex)
9	HIN recombinase (1hcr)	Proteinase inhibitor (1cew)
10	HIN recombinase (1hcr)	Proteinase inhibitor (1cew)
11	HIN recombinase (1hcr)	Proteinase inhibitor (1cew)
12	Aspartate aminotransferase (1ars)	Histidine binding protein (1hsl)
13	Apolipoprotein-E3 (1lpe)	Cutinase (1cus)
14	Apolipoprotein-E3 (1lpe)	Leghemoglobin (1gdm)
15	Apolipoprotein-E3 (1lpe)	Leghemoglobin (1gdm)
16	Glutathione transferase (1gst)	Leghemoglobin (1gdm)
17	Glutathione transferase (1gst)	Uteroglobin (1utg)
18	Acyl transferase (3cla)	Uteroglobin (1utg)
19	Interleukin-10 (1ilk)	Uteroglobin (1utg)
20	Thermolysin (8tln)	Alpha-parvalbumin (1rtp)
21	Immunoglobin FC (1fc2)	Adenovirus fiber protein (1knb)
22	Immunoglobin FC (1fc2)	Adenovirus fiber protein (1knb)
23	Immunoglobin FC (1fc2)	Adenovirus fiber protein (1knb)
24	Dihydrofolate reductase (3dfr)	Alpha-parvalbumin (1rtp)
25	Dihydrofolate reductase (3dfr)	Phosphotransferase (1npk)

The proteins from which the final torsion angles of two simulated homeodomain structures originate are indicated for residues 1 to 25 of both structures.

Table 3. *Z*-scores for native-like conformations with different scoring functions

	1FC2A	1HDD	2CRO	4ICB	Average
Surface	−0.52	−0.23	−0.38	−0.48	−0.40
HF	−0.46	−0.68	−0.04	−0.69	−0.47
Contact(HL)	−0.41	−0.19	0.08	−0.38	−0.23
Contact(MJ)	−0.30	−0.13	0.08	−0.59	−0.24
Shell	−0.41	−0.48	−0.55	−1.05	−0.63
Shelltop	−0.39	−0.37	−0.42	−1.02	−0.55
Histogram	0.00	−0.04	−0.70	−0.48	−0.31
VdW(HL4)	−0.36	−0.69	−0.39	−1.31	−0.69
Shellm	−0.43	−0.54	−0.66	−0.59	−0.56
Shelltopm	−0.38	−0.56	−0.64	−0.89	−0.62
Eq(8)	−0.32	−0.69	−1.12	−0.87	−0.75
Eq(8) + msa	−0.32	−0.79	−1.08	−1.29	−0.87

The cutoff below which conformations were taken to be native-like was 4 Å rmsd for protein A and the homeodomain, and 5 Å rmsd for calbindin and cro repressor. The *Z*-scores (the number of standard deviations separating the scores of the native-like conformations from the ensemble average) were calculated over ensembles of 500 conformations for each protein generated using the "no filter" condition of Table 1.

Project 2

- Develop a simple prototype of fragment assembly template-free modeling system

Project Plan

- Representative protein structure database: http://bioinfo.mni.th-mh.de/pdbselect/
- Fragment generation: Rosetta (database approach) or FRAGSIOIN (model-based approach)
- Energy function: Rosetta 3, Dfire energy function (executable available), Yang Zhang's RW potential (executable available), or something else
- Sampling approach
- Testing: 3 CASP11 TFM targets
- Present your plan Monday (March 7)

Technical Issues and Resources

- Conversion between torsional angles and Cartesian coordinates: https://www.rosettacommons.org/content/conversion-dihedral-angle-representation-cartesian-representation;
- Convert coordinates to torsion angles http://www.math.fsu.edu/~quine/MB_10/6_torsion.pdf
- RW potential: http://zhanglab.ccmb.med.umich.edu/RW/
- Dfire energy: https://www.rosettacommons.org/manuals/archive/rosetta3.4_user_guide/df/d11/classcore_1_1scoring_1_1methods_1_1dfire_1_1_d_f_i_r_e___energy.html
- Rosetta: https://www.rosettacommons.org/
- FRAGSION: http://sysbio.rnet.missouri.edu/FRAGSION/

UniCon3D – Open Source Software

- Use HMM to sample angles
- Use sequential fragment assembly to build 3D structures
- Small code base
- Easy to use
- Reference: Bhattacharya, Cao and Cheng. UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. Bioinformatics, 2016.
- Tool: https://github.com/multicom-toolbox/UniCon3D

Rosetta & MULTICOM Resource

- How to Use Rosetta: http://www.cs.huji.ac.il/~fora/81855/exercises/81855_ex3_2012.pdf
- Rosetta Online Document: https://www.rosettacommons.org/manuals/archive/rosetta3.4_user_guide/index.html
- Fragsion: http://sysbio.rnet.missouri.edu/FRAGSION/

Project Schedule

- Wednesday (Feb. 28): project discussion
- Wednesday (March 7): plan presentation
- Monday (March 19): presentation and discussion of results