A Probabilistic Graphical Model for *Ab Initio* Folding

Jinbo Xu

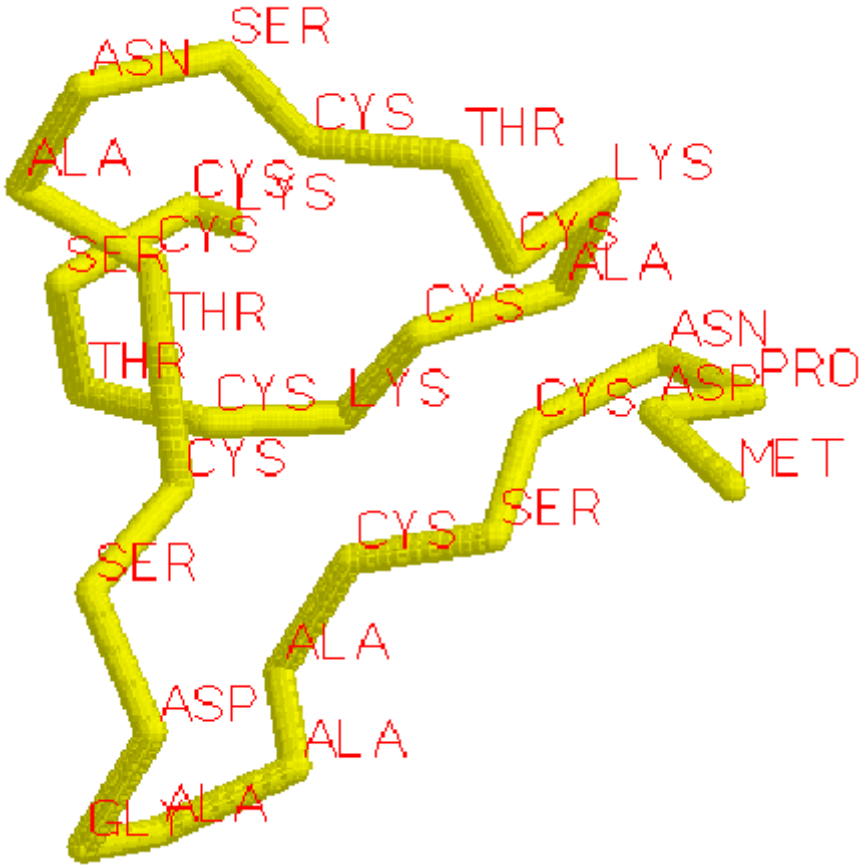
j3xu@ttic.edu
Toyota Technological Institute at Chicago

Protein Structure Prediction

The amino acid sequence, e.g.,

MDPNCSCAAAGDSCTCANSCTCLACKCTSCK,

folds into a 3D structure.



Experimental techniques

- time-consuming & expensive
- sometimes does not work

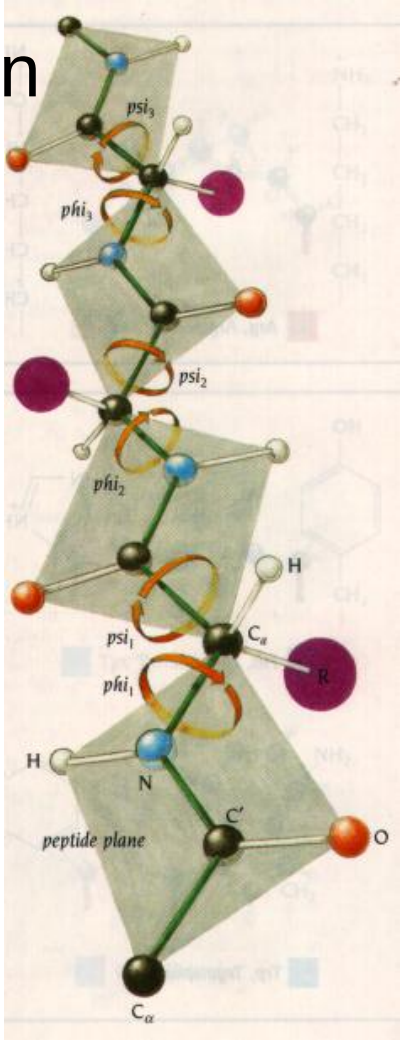
Computational methods

- fast & cheap, but challenging

Protein Structure Prediction

- The 3D structure of a protein encoded in its sequence
- A protein tends to stay at a minimum energy state
- Can be formulated as an optimization problem
  - the search space is enormous
  - the number of local minima increases exponentially

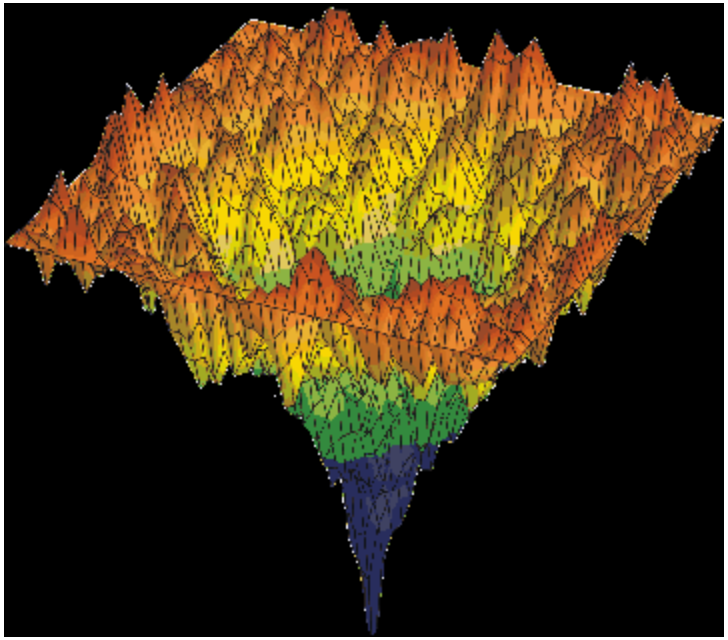
Computationally it is an exceedingly difficult problem.



Ab Initio Folding Challenge

- Thousands of atoms interacting with each other
- Protein folding in nature not fully understood
- Rugged energy landscape
- NP-complete even a simple HP model (Berger & Leighton, 1998)

Folding energy landscape



Two Major Components

- An efficient conformation sampling algorithm to explore the huge conformation space

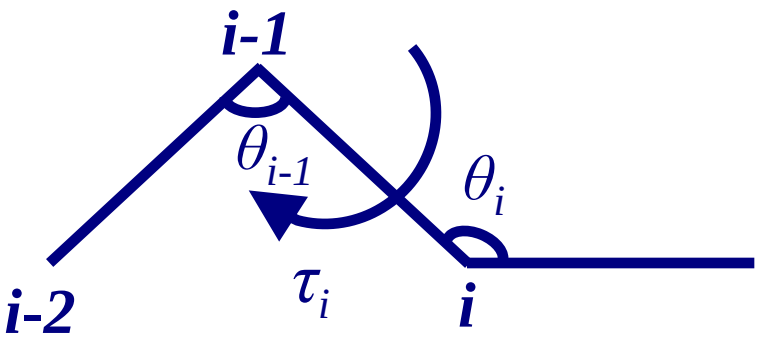
- An accurate energy function to differentiate native state from decoys

Our Work

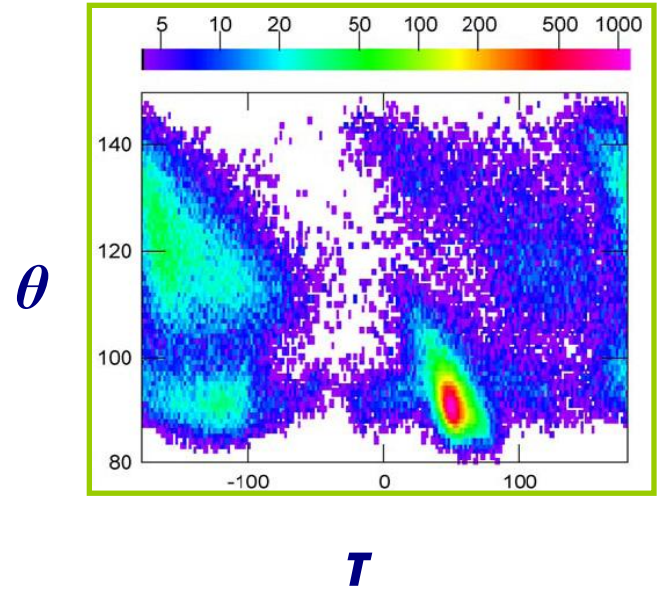
- Estimating the probability of a conformation using Conditional Random Fields (CRFs) from sequence information
- Sampling backbone angles using CRFs and directional statistics
- Minimizing energy by Simulated Annealing or Replica Exchange Monte Carlo

Modeling Backbone Angles Using Directional Statistics

(1) *C_α-trace Representation*

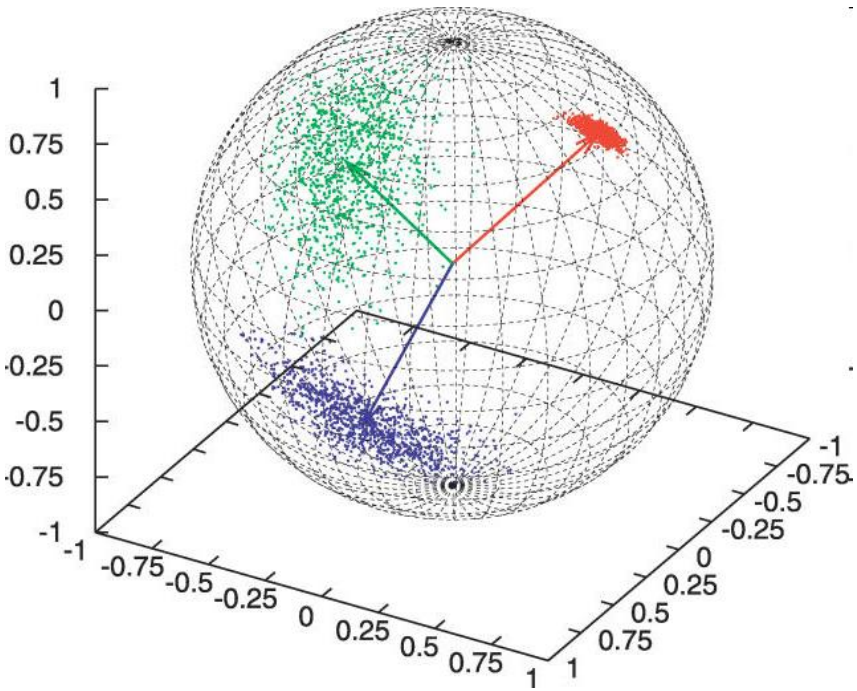


(2) *Distribution of Bond Angles*



FB5 Distribution

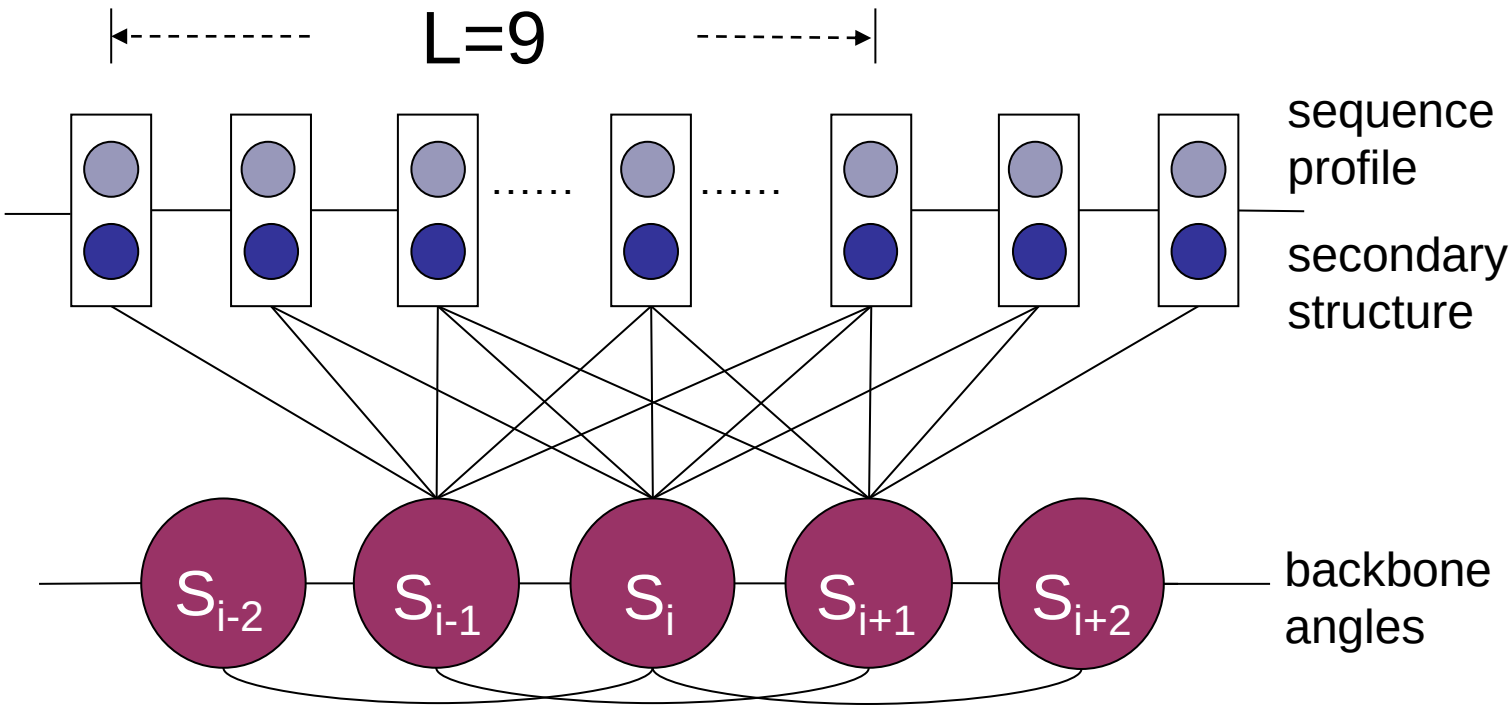
The angle space divided into 100 groups, each modeled by a 5-parameter Fisher-Bingham (FB5) distribution



Red: α-helix; blue: beta; green: loop

The picture is taken from Hamelryck et al., PLos Comp Biol 2006.

Modeling Sequence-Structure Relationship Using 2nd-order CRF



Each S_i represents a FB5 distribution at position i

Computational Challenge I: Model Training

- ~3000 protein structures for training (~800k data points), many more later
- ~1 million parameters in the CRF model
- MPI on ~100 fast CPUs, communication among MPI processes per iteration (every 15 minutes)
- Terminate within ~400 iterations (3-4 days)
- Can we use many CPUs through OSG?

Conformation Sampling

- Sample a protein segment with size in [1,15].
- Sample backbone angles of this segment by probability using the CRF model. This step samples the group the angles at a position belong to.
- Sample real-valued angles for each position using the FB5 distribution represented by the group.
- Rebuild coordinates for C_α atoms from angles.
- Rebuild other main chain atoms from C_α.

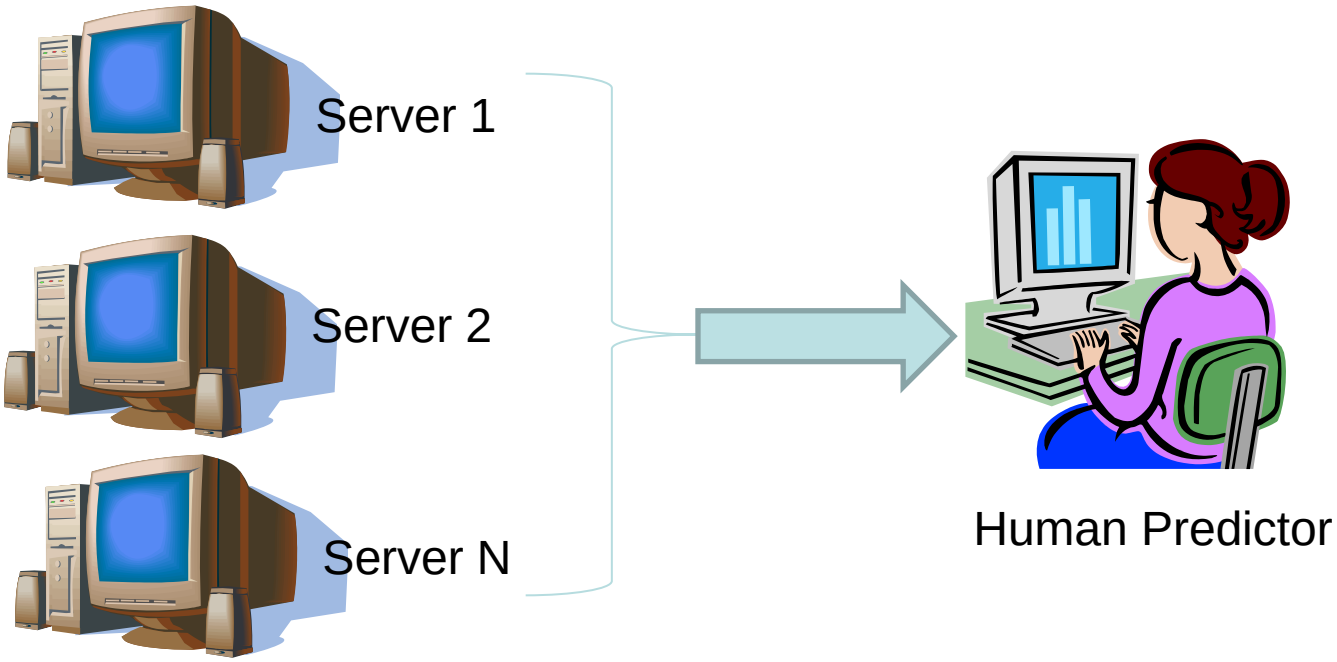
Computational Challenge II: Energy Minimization

- Two energy minimization methods:
  - Simulated Annealing (SA)
  - Replica Exchange Monte Carlo (REMC)
- For each protein, >10k decoys needed
- SA: ~30 mins for a single decoy, OSG ideal for SA
- REMC: ~8 hours for 10 decoys

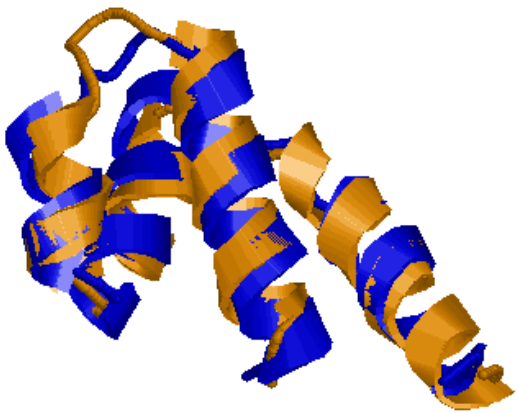
CASP

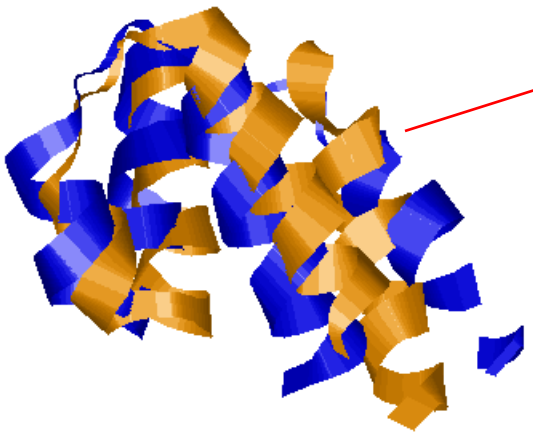
Critical Assessment of Structure Prediction

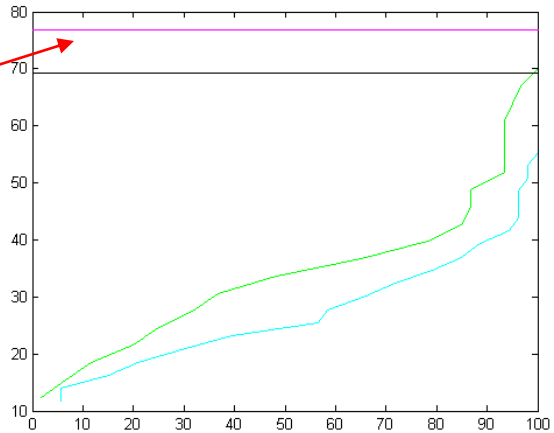
1. Public: organized by structure prediction community, evaluated by the third-party
2. Blind: experimental structures unknown



Ab Initio Folding Example T0416_D2







Best decoy RMSD 1.4Å

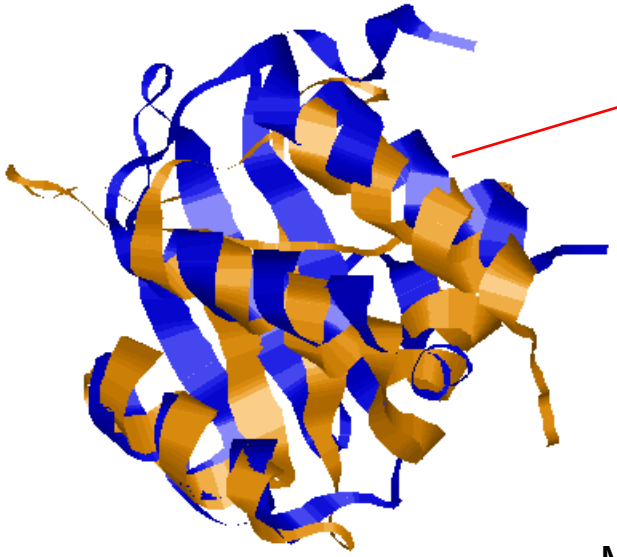
Best model RMSD 2.7Å

Better than all the 319 CASP8 models

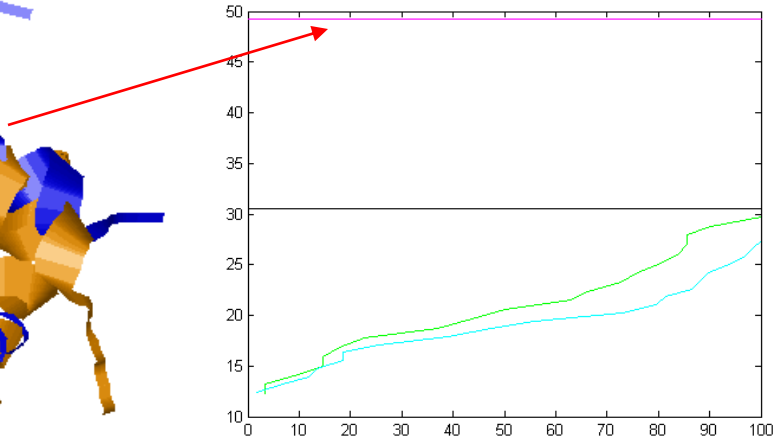
Native green, model yellow

Ab Initio Folding Example T0496_D1

Best decoy 5.0Å



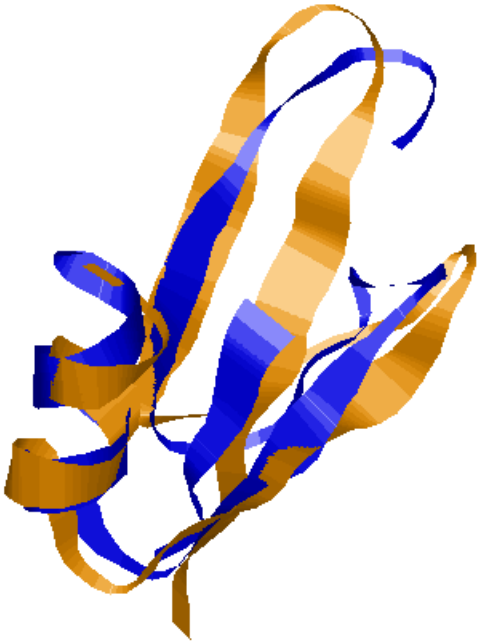
Best model 6.2Å



Much better than all the
483 CASP8 models
All CASP8 models RMSD >11Å

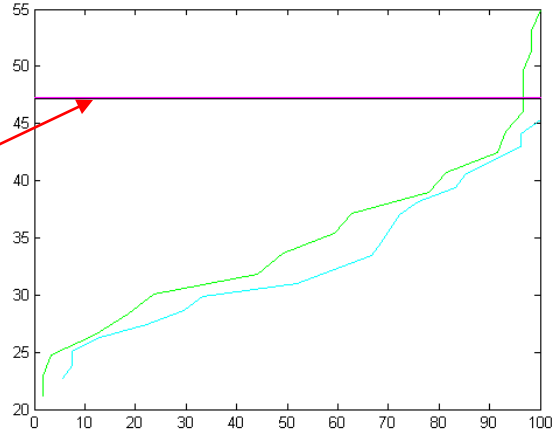
Native green, model yellow

Ab Initio Folding Example T0510_D3



Best decoy 3.0Å

Best model 6.9 Å



Only worse than 3 out of 321 CASP8 models

Native green, model yellow

Acknowledgements

Computational resources

- OSG (John McGee team)
- SHARCNet (www.sharcnet.ca)
- Teraport (U Chicago)
- TTIC cluster

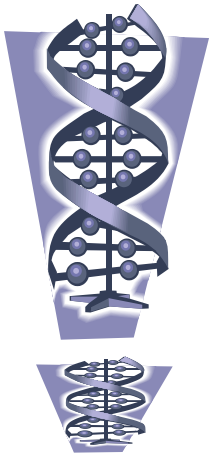
Students

- Jian Peng & Feng Zhao (TTI-C)
- Shuaicheng Li(Waterloo),Beckett Sterner (Chicago)

Collaborators

- Tobin Sosnick & Karl Freed group

Thank You



Second-order CRF Model (Cont'd)

Given sequence profile M and secondary structure X, the probability of backbone angles S is defined by

$$P_\lambda(S \mid M, X) = \frac{1}{Z(M, X)} \exp \sum_{i=1}^{N} F(S,M,X,i)$$

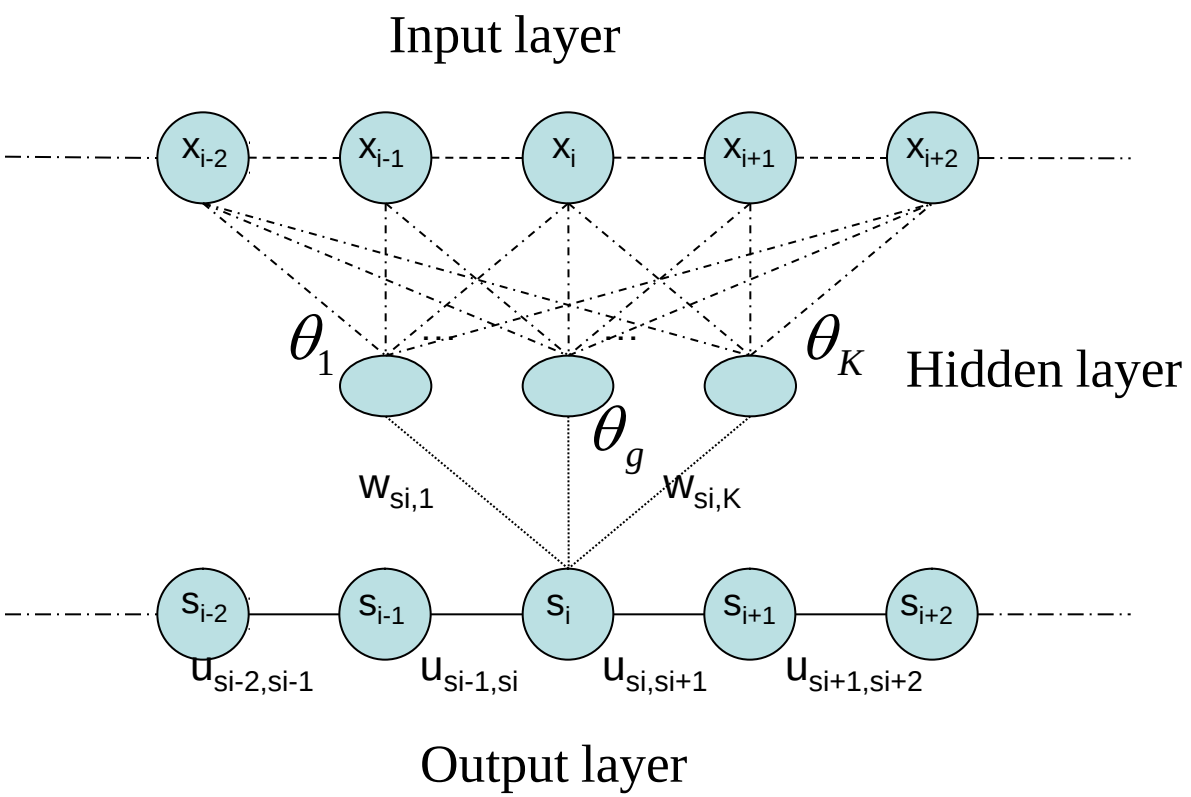
$$F(S, M, X, i) = e_1(s_{i-1}, s_i) + e_2(s_{i-1}, s_i, s_{i+1})$$ edge features

$$+ \sum_{j=i-w}^{i+w} (v_1(s_i, M_j, X_j) + v_2(s_{i-1}, s_i, M_j, X_j))$$ state features

Z(M,X): normalization factor;

F(S,M,X, i) denotes features at position i

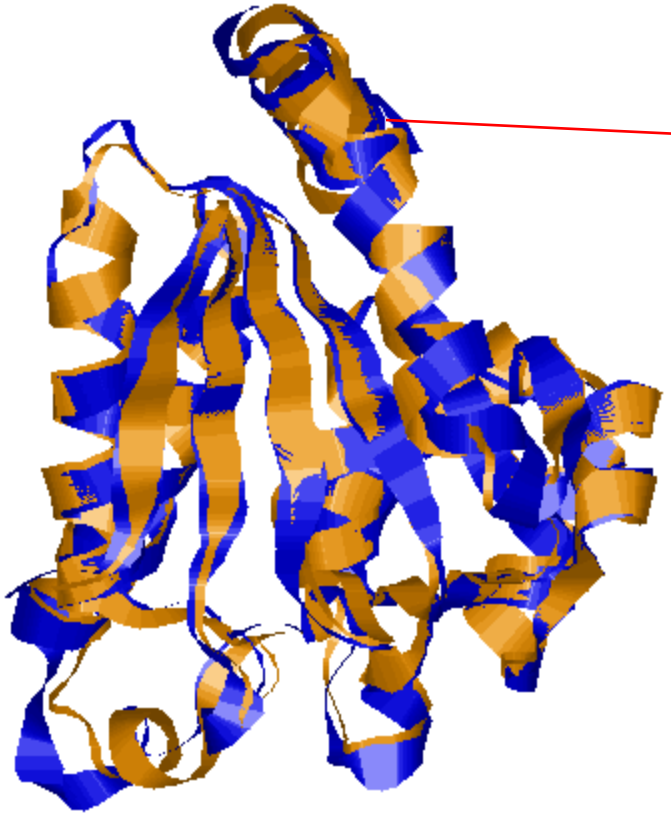
Generalize CRF to CNF: modeling nonlinear relationship

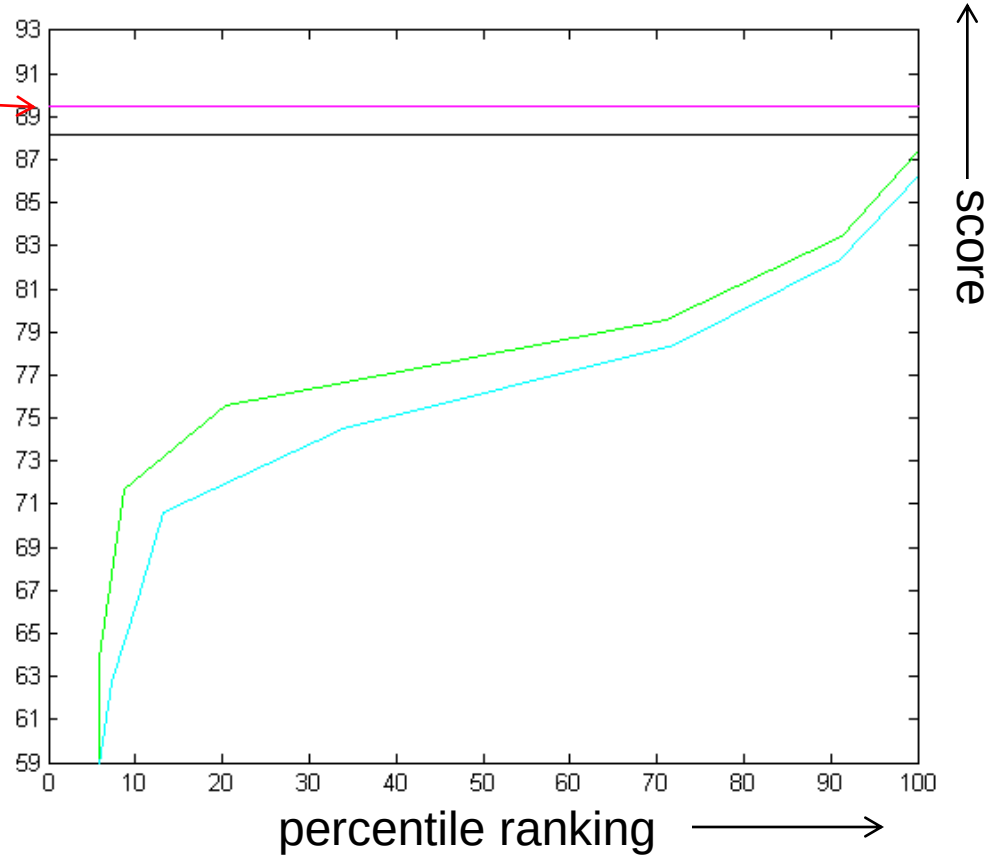


Energy Function

- DOPE: residue-specific distance-dependent pairwise statistical potential (Shen & Sali)
- ESP: simplified solvent-accessible surface area (SASA) potential
- KMBhbond: Baker’s hydrogen bonding energy for beta-sheet forming

Protein Threading Example: T0486





214 AAs, RMSD 1.34Å

Native green, model yellow

Seq ID with its templates is 20-30%
Better than 378 human and server models

Summary

- Our method comparable to Robetta and TOUCHSTONE-II, although using a simple energy function and sampling in continuous space
- CRF can accurately model protein sequence-structure relationship
- Our method works well on mainlyαproteins, but not very well on β-containing proteins

Comparison with TOUCHSTONE II

α **proteins**

PDB code	**Class**	**Length**	**TouchStone II**	**CRFFolder**
			Best Cluster	Best Cluster
1bw6A	α	56	4.79(2/3)	3.82(3/3)
1lea	α	72	5.69(5/5)	4.10(5/7)
2af8	α	86	11.07(5/6)	8.9(12/19)
256bA	α	106	3.61(2/3)	2.75(6/11)
1sra	α	151	10.71(3/12)	13.95(17/25)

Comparison with TOUCHSTONE II

αβ proteins

1gpt	αβ	47	6.30(1/25)	5.55(42/67)
1kp6A	αβ	79	10.01(8/14)	7.998(1/7)
1poh	αβ	85	9.10(5/9)	8.84(5/10)
1npsA	αβ	88	6.89(33/34)	9.91(41/57)
1t1dA	αβ	100	8.96(7/13)	9.22(10/13)

Comparison with TOUCHSTONE II

β proteins

1msi	β	66	7.72(19/28)	7.77(12/15)
1hoe	β	74	9.39(5/13)	9.87(16/35)
1ezgA	β	82	11.03(40/44)	10.42(42/66)
1sfp	β	111	7.48(2/18)	11.07(5/11)
1b2pA	β	119	12.52(31/56)	10.01(18/25)

Comparison w. Robetta in CASP8

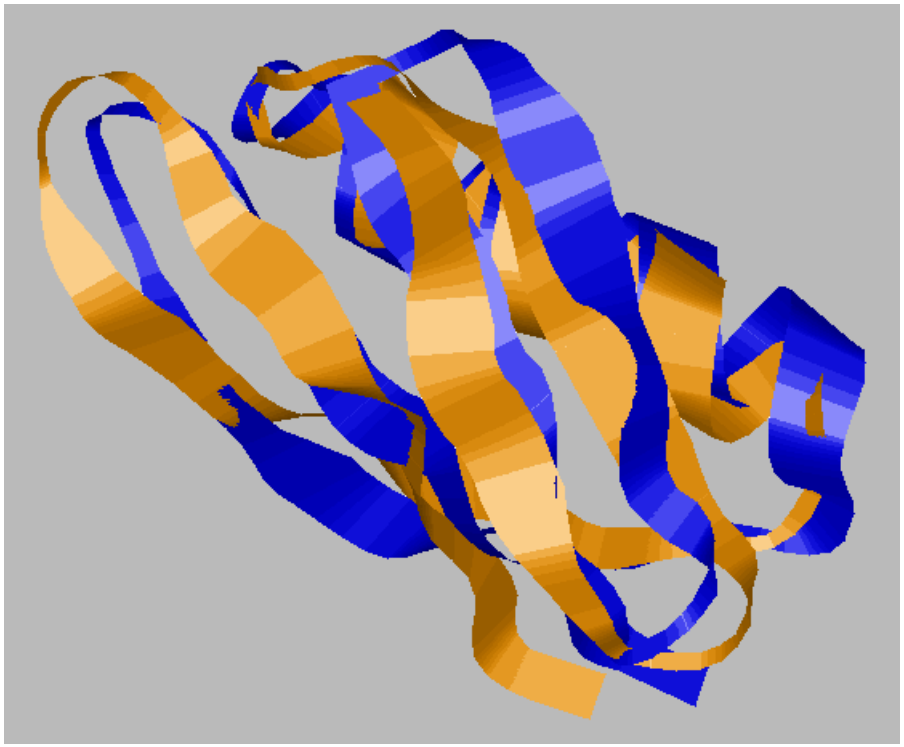
Target ID	Length	Class	Robetta	CRFFolder
T0397_D1	70	αβ	0.250	0.258
T0460	111	αβ	0.262	0.308
T0465	157	αβ	0.243	0.253
T0466	128	β	0.326	0.217
T0467	97	β	0.303	0.364
T0468	109	αβ	0.253	0.308
T0476	108	αβ	0.279	0.250
T0480	55	β	0.208	0.307

Comparison w. Robetta in CASP8 (Cont'd)

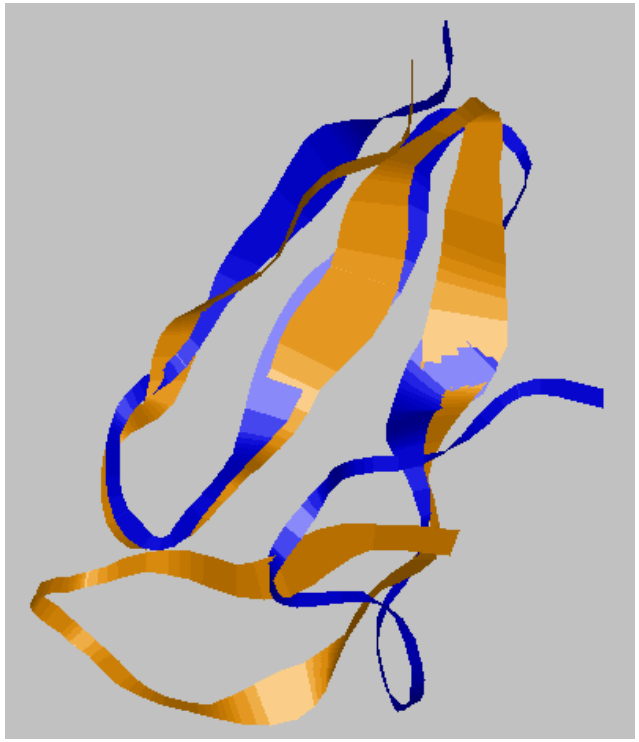
Target ID	Length	Class	Robetta	CRFFolder
T0482	120	αβ	0.352	0.223
T0484	62	α	0.253	0.249
T0495_D2	65	αβ	0.312	0.436
T0496_D1	110	αβ	0.235	0.293
T0496_D2	68	α	0.291	0.500
T0510_D4	43	αβ	0.147	0.352
T0513_D1	77	αβ	0.581	0.367
T0514	145	αβ	0.283	0.277
Average			0.286	0.310
Sum			4.578	4.960

Examples: small proteins

Yellow is predicted structure and blue native.



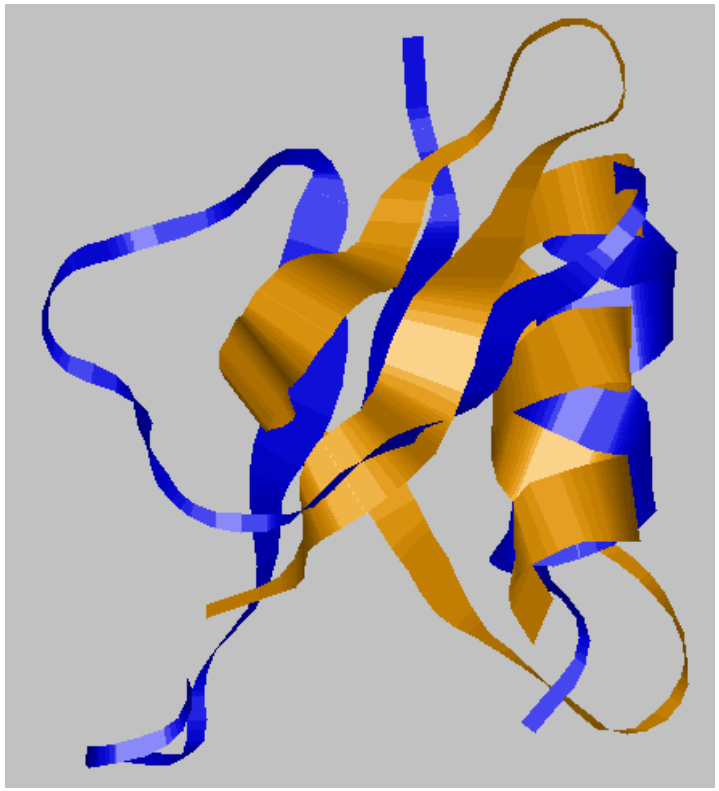
2GB1A, RMSD=1.85Å



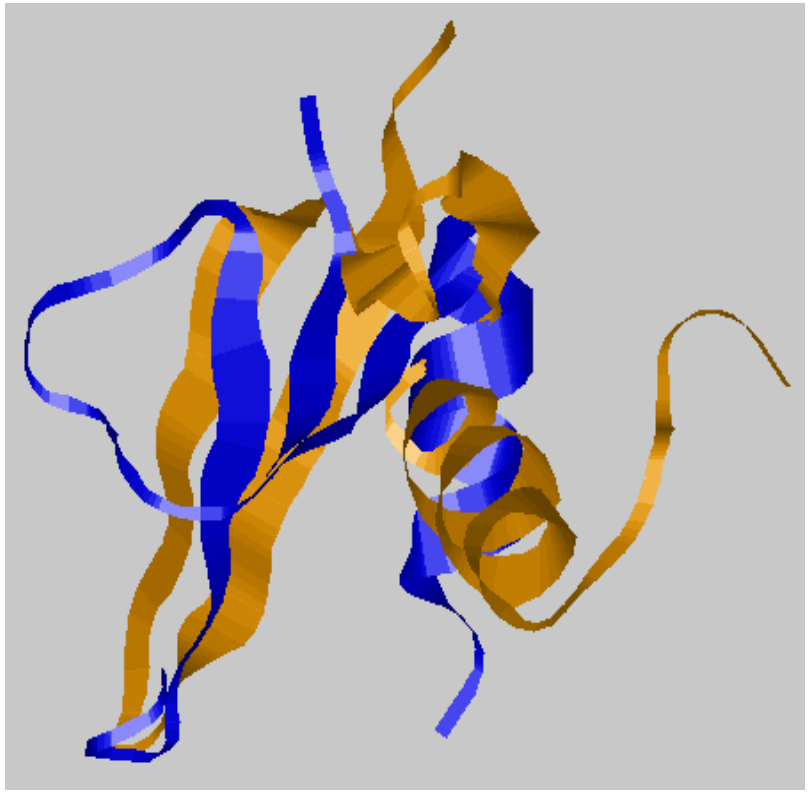
T0480, RMSD=2.86Å

Example: T0510_D3

Yellow is predicted structure and blue native.



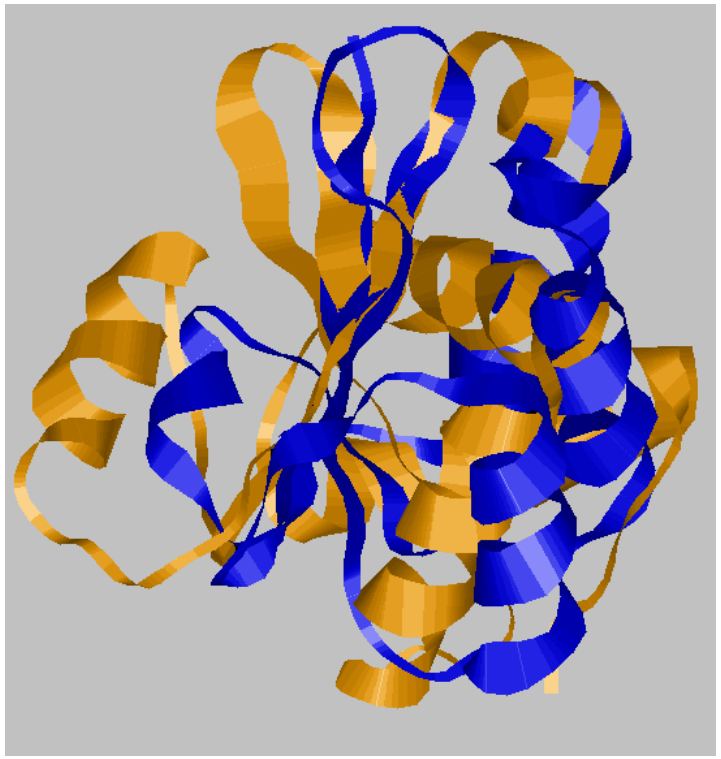
TM=0.352, RMSD=11.278
Our CASP8 submission



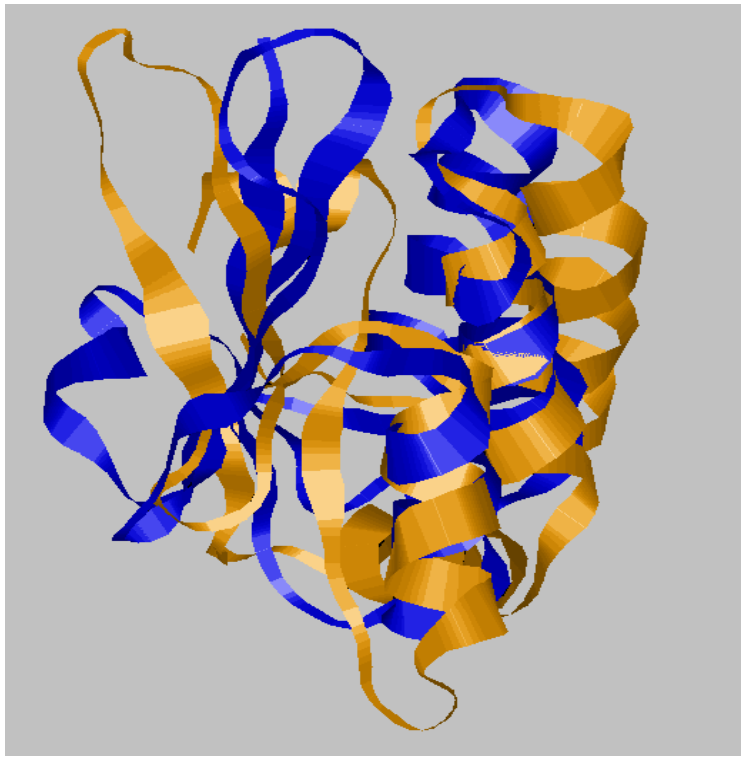
TM=0.348, RMSD=11.427
Second best decoy

Example: T0496_D1

Yellow is predicted structure and blue native.



TM=0.297,RMSD=11.457
Our CASP8 submission



TM=0.475, RMSD=6.592
Best decoy

Who Cares?

- Long history: more than 30 years
- Listed as a “grand challenge” problem
- Increasing gap between #sequences vs. #structures
- Competitions: CASP (1994-2008)
- Useful for
  - Drug design
  - Enzyme design
  - Function annotation
  - Target selection

Contents

1. Challenges

2. Our Solutions

3. Results

4. Conclusions

Conclusion(3)

Future Works:

1. Improve the sampling algorithm on beta regions

2. Develop better hydrogen-bonding energy items for the formation of beta sheets

Challenges

1. Express the complex protein sequence- structure relationship

2. Model the ab initio conformation search space in consistence with its continuous nature

3. Balance between time efficiency and accuracy

Our Solutions

1. Combining C_α-trace, FB5 distribution and BBQ method: a simplified and continuous representation

2. A 2nd-order CRF Model for protein sequence - structure relationship

3. Simple energy function: DOPE, ESP and KMBhbond

The Idea

- Build a graphical model in a continuous space
- Use the probability estimated from PSI-BLAST sequence profile and predicted secondary structure.
- Compare with the fragment assembly method and the lattice model.

Energy Function(1)

- **DOPE** distinguishes the amino acid identity and atomic identity of two interacting particles
- We only used the statistical potentials related to main-chain and C_β atoms

Energy Function(2)

- **ESP:** an approximation to the Ooi-Scheraga solvent-accessible surface area (SASA) potential
- Each residue is assigned with an environmental energy score, $ESP(aa,n)=-ln\{P(n|R,aa)/p(n|R)\}$
- $p(n|R)$ is the number of C_α atoms in an 8.5Å sphere for a given protein radius regardless of amino acid identity
- $p(n|R,aa)$ is the number of C_α atoms in an 8.5Å sphere for a given protein radius and amino acid identity.

Energy Function(3)

- **KMBhbond:** depends on the distance between the geometric centers of the N-H bond vector and the C=O bond vector

- Use three angles to describe the relative orientation of the bond vectors in the hydrogen bond:

  1) the bond angle between the N-H bond and the hydrogen bond
  2) the bond angle between the C=O bond and the hydrogen bond
  3) the dihedral angle about the acceptor-acceptor base bond.

Result 1. Decoy quality comparison between the first-order and second-order CRF samplers.

PDB code	Length	Class	Best		1%		2%		5%		10%	
			O-1	O-2	O-1	O-2	O-1	O-2	O-1	O-2	O-1	O-2
1aa2	108	α	7.3	7.3	9.3	9.5	9.8	9.9	10.4	10.4	10.9	10.8
1beo	98	α	6.4	5.8	8.4	8.1	8.8	8.6	9.5	9.3	10.1	9.9
1ctfA	68	αβ	3.7	3.7	5.1	4.6	5.4	4.9	5.9	5.3	6.5	5.7
1dktA	72	β	6.1	5.1	7.6	6.4	8.0	6.7	8.5	7.2	9.0	7.7
1enhA	54	α	2.3	2.2	3.1	2.6	3.3	2.6	3.7	2.8	4.1	2.9
1fc2C	43	α	1.9	2.3	2.7	2.6	2.8	2.7	3.1	2.9	3.4	3.0
1fca	55	β	4.9	5.0	6.4	6.2	6.7	6.5	7.3	6.9	7.7	7.2
1fgp	67	β	7.4	5.9	8.9	7.8	9.2	8.1	9.6	8.6	10.0	8.9
1jer	110	β	9.6	10.2	11.6	11.5	11.9	11.8	12.4	12.3	12.9	12.7
1nkl	78	α	3.6	3.1	4.7	3.8	5.0	3.9	5.4	4.3	5.8	4.6
1pgb	56	αβ	3.1	2.6	4.1	3.6	4.3	3.8	4.6	4.0	4.9	4.2
1sro	76	β	6.2	5.4	7.8	7.2	8.2	7.6	8.8	8.2	9.3	8.8
1trlA	62	α	3.5	3.7	4.4	4.5	4.6	4.6	4.9	4.8	5.2	5.0
2croA	65	α	2.8	2.6	3.6	3.2	3.8	3.4	4.2	3.6	4.6	3.9
2gb1A	56	β	2.9	2.0	4.0	3.5	4.2	3.6	4.6	3.9	4.9	4.1
4icbA	76	α	4.6	4.4	5.9	6.7	6.2	7.1	6.8	7.7	7.4	8.2
T052	98	β	7.6	8.4	10.6	10.1	11.0	10.5	11.6	11.1	12.1	12.6
T056	114	α	7.8	7.6	9.8	9.6	10.2	9.9	10.9	10.5	12.1	11.6
T059	71	β	6.3	6.2	8.5	8.0	8.8	8.2	9.3	8.7	9.6	9.0
T061	76	α	5.3	6.0	7.0	7.1	7.3	7.3	7.8	7.6	8.2	7.9
T064	103	α	7.2	7.5	9.4	9.0	9.9	9.5	10.7	10.4	11.4	11.1
T074	98	α	4.9	4.2	7.3	6.6	7.7	7.0	8.4	7.6	9.0	8.1
Average RMSD			5.3	5.1	6.8	6.4	7.2	6.7	7.7	7.2	8.1	7.6

Result 1. Decoy quality comparison with TOUCHSTONE II

PDB code	Class	Length	TouchStone II	CRFFolder			
			Best Cluster	Best Cluster	Best	1%	2%
1bw6A	α	56	4.79(2/3)	3.82(3/3)	2.75	3.38	3.54
1lea_q	α	72	5.69(5/5)	4.10(5/7)	3.41	4.20	4.48
2af8_q	α	86	11.07(5/6)	8.9(12/19)	7.07	8.53	8.97
256bAq	α	106	3.61(2/3)	2.75(6/11)	2.50	3.45	3.70
1sra_	α	151	10.71(3/12)	13.95(17/25)	10.82	13.76	14.24
1gpt_q	αβ	47	6.30(1/25)	5.55(42/67)	4.34	5.20	5.47
1kp6A	αβ	79	10.01(8/14)	7.998(1/7)	6.30	7.51	7.81
1poh	αβ	85	9.10(5/9)	8.84(5/10)	7.49	8.70	9.04
1npsA	αβ	88	6.89(33/34)	9.91(41/57)	7.87	9.19	9.66
1t1dA	αβ	100	8.96(7/13)	9.22(10/13)	6.51	9.51	9.94
1msi_	β	66	7.72(19/28)	7.77(12/15)	6.24	7.55	7.89
1hoe_	β	74	9.39(5/13)	9.87(16/35)	7.96	10.00	10.37
1ezgA	β	82	11.03(40/44)	10.42(42/66)	9.66	10.35	10.62
1sfp_	β	111	7.48(2/18)	11.07(5/11)	9.32	11.09	11.59
1b2pA	β	119	12.52(31/56)	10.01(18/25)	8.76	10.89	11.32

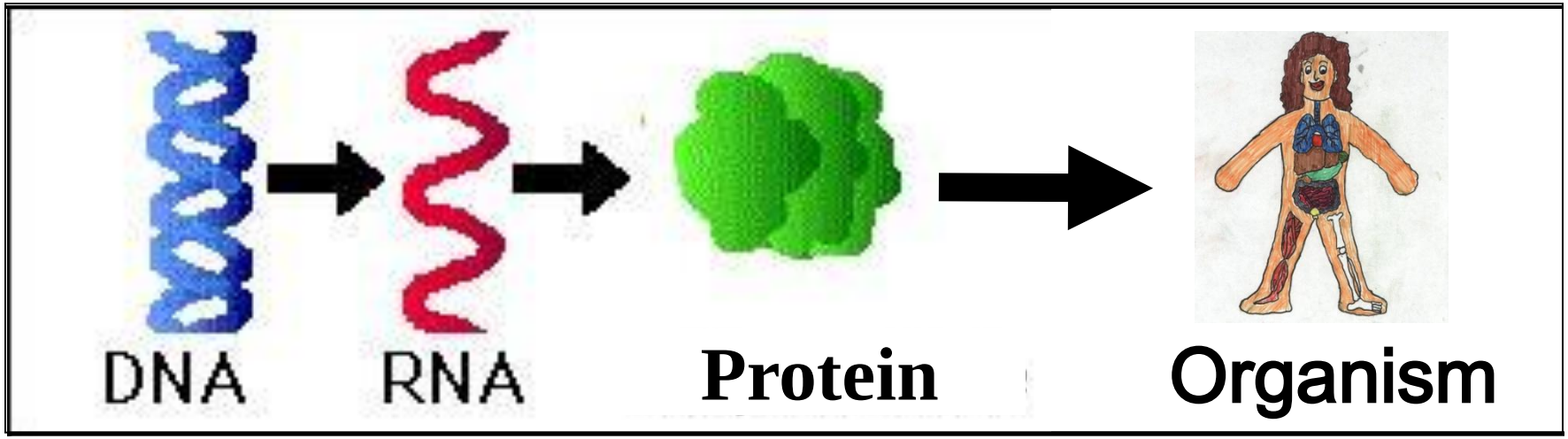
Comparison with TOUCHSTONE II

PDB code	Class	Length	TouchStone II	CRFFolder
			Best Cluster	Best Cluster
1bw6A	α	56	4.79(2/3)	3.82(3/3)
1lea	α	72	5.69(5/5)	4.10(5/7)
2af8	α	86	11.07(5/6)	8.9(12/19)
256bA	α	106	3.61(2/3)	2.75(6/11)
1sra	α	151	10.71(3/12)	13.95(17/25)
1gpt	αβ	47	6.30(1/25)	5.55(42/67)
1kp6A	αβ	79	10.01(8/14)	7.998(1/7)
1poh	αβ	85	9.10(5/9)	8.84(5/10)
1npsA	αβ	88	6.89(33/34)	9.91(41/57)
1t1dA	αβ	100	8.96(7/13)	9.22(10/13)
1msi	β	66	7.72(19/28)	7.77(12/15)
1hoe	β	74	9.39(5/13)	9.87(16/35)
1ezgA	β	82	11.03(40/44)	10.42(42/66)
1sfp	β	111	7.48(2/18)	11.07(5/11)
1b2pA	β	119	12.52(31/56)	10.01(18/25)

Comparison w. Robetta in CASP8

Target ID	Length	Class	Robetta	CRFFolder
T0397_D1	70	αβ	0.250	0.258
T0460	111	αβ	0.262	0.308
T0465	157	αβ	0.243	0.253
T0466	128	β	0.326	0.217
T0467	97	β	0.303	0.364
T0468	109	αβ	0.253	0.308
T0476	108	αβ	0.279	0.250
T0480	55	β	0.208	0.307
T0482	120	αβ	0.352	0.223
T0484	62	α	0.253	0.249
T0495_D2	65	αβ	0.312	0.436
T0496_D1	110	αβ	0.235	0.293
T0496_D2	68	α	0.291	0.500
T0510_D4	43	αβ	0.147	0.352
T0513_D1	77	αβ	0.581	0.367
T0514	145	αβ	0.283	0.277
Average			0.286	0.310
Sum			4.578	4.960

Biology in One Slide



Amino Acids

A protein is composed of a central backbone and a collection of (typically) 50-2000 amino acids (a.k.a. residues).

20 different amino acids, each consisting of up to 18 atoms, e.g.

Name	3-letter code	1-letter code
Leucine	Leu	L
Alanine	Ala	A
Serine	Ser	S
Glycine	Gly	G
Valine	Val	V
Glutamic acid	Glu	E
Threonine	Thr	T

Method
Continuous Representation(3)

(4) The backbone and C_β atoms are built using the extended BBQ method.

- BBQ = Backbone Building from Quadrilaterals
- Measure the average positions of C, O, and N atoms in local grids
- Extend the method to build coordinates for C_β
- RMSD is approximately 0.5A$^\circ$ on native structures.

Method
Energy Minimization(2)

- Decrease t using exponential cooling schedule $t_{k+1} = \alpha t_k$, (α= 0.9)
- At each t_k, the number of sampled conformations is set to $N_s=100 \times (1+N/100)$, N is the number of residues
- SA process terminates if (1) the temperature is low enough; or (2) the number of conformations generated in a single simulation process reaches a threshold (say 10,000).