Protein Structure Analysis with Sequential Monte Carlo Method

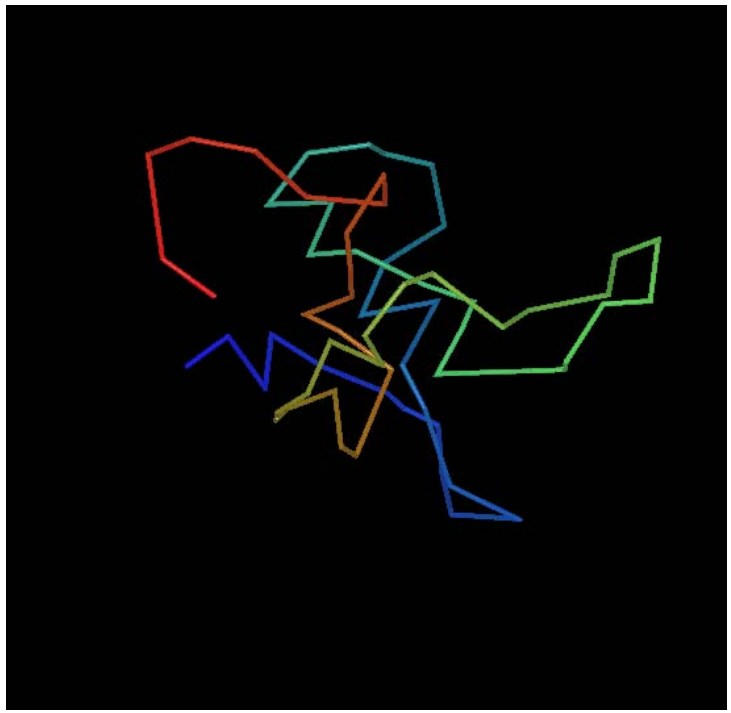
Jinfeng Zhang
Computational Biology Lab
Department of Statistics
Harvard University

Introduction

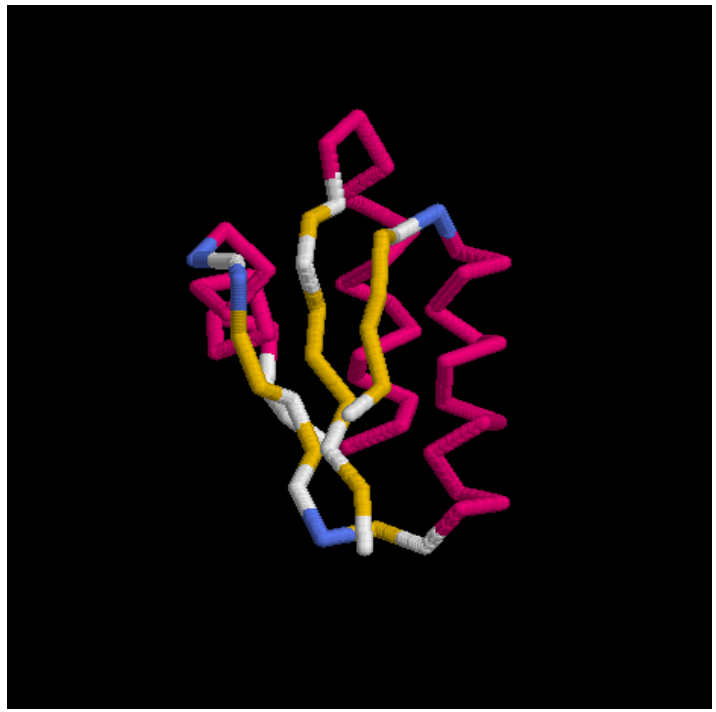
- Structure → Function & Interaction
  - Protein structure initiative (PSI) is speeding up the information flow from sequence to structures.
  - Information does not readily flow from structures to structures.
  - Neither does it readily flow from structures to applications.
- What are the bottle necks?
  - Sampling method.
  - Potential function.

Sampling Methods
-- Folding & Growth

Folding Method

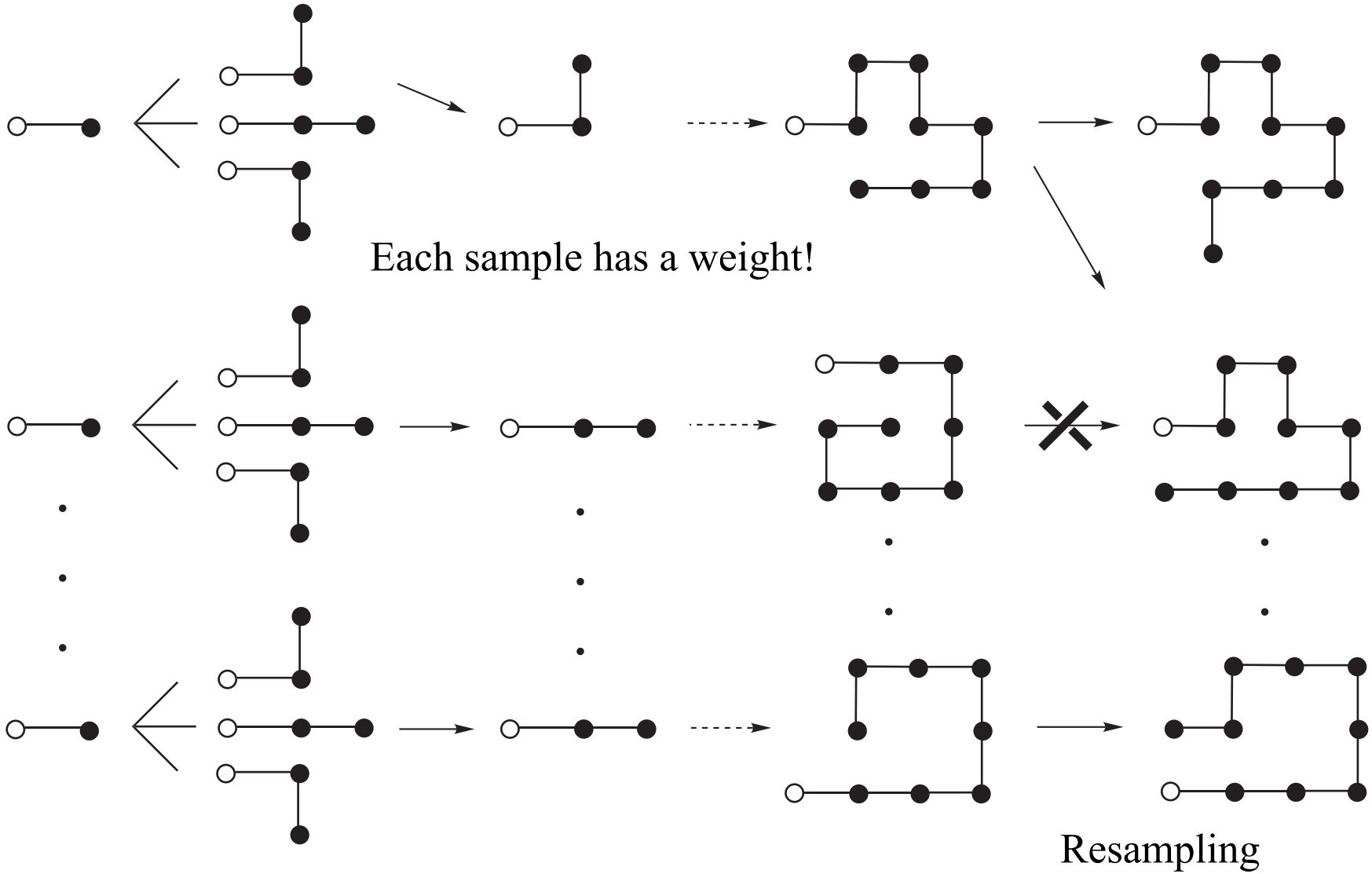


Growth Method



From http://www.bioinformatics.buffalo.edu/

Sequential Monte Carlo (SMC) -- Step by Step



SMC
-- Summary

- Short chains:
  - Exhaustive enumeration, useful for evaluation of SMC performance.
- Long chains:
  - Sequential Monte Carlo, estimating interesting properties.
- The main ingredients of SMC are:
  - Sequence of distributions "approaching" the target distribution $\pi(x_1,\ldots,x_n)$.
  - Sampling distribution $g_{t+1}(x_{t+1}|x_1,\ldots,x_t)$.
  - Resampling scheme.

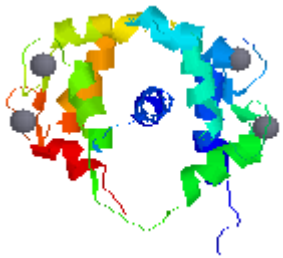
$$\hat{\mu}_h = \frac{\sum_{j=1}^{m} h(x_1^{(j)},\ldots,x_n^{(j)}) \cdot w_n^{(j)}}{\sum_{j=1}^{m} w_n^{(j)}}$$

Reference for SMC

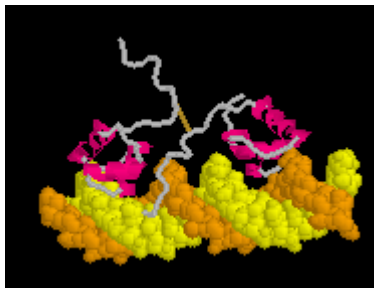
- J.S. Liu and R. Chen (1998). SMC for dynamic systems. *J Amer Statist Assoc* **93**, 1032-45.
- J.S. Liu (2001). *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag.
- J. Liang, J. Zhang, R. Chen, (2002). *J. Chem. Phys.* 117:7, 3511-3521.
- J. Zhang, R. Chen, C. Tang, and J. Liang, (2003). *J. Chem. Phys.* 118:12, 6102-6109.
- J. Zhang, Y. Chen, R. Chen, and J. Liang, (2004). *J. Chem. Phys.* 121:1, 592-603.

Near Native Structures of Proteins

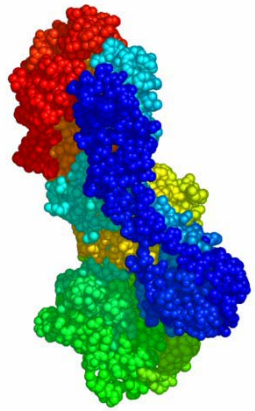
Native State is an Ensemble of Structures



2BBN



Lac repressor



Ca^{2+} ATPase pump

- Protein functions and interactions are determined by the near native structures.

Biological Problems

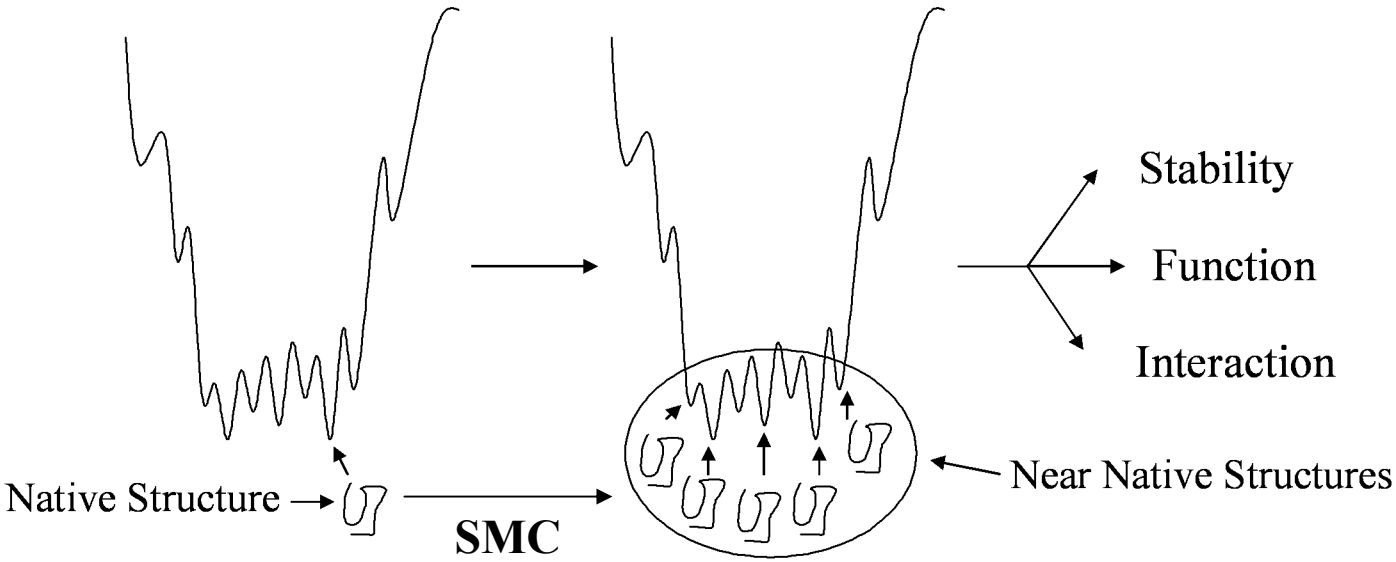
- Stability
  - Probability of NNS under Boltzmann distribution.
- Function
  - Analysis of NNS to detect correlated structural changes.
- Interaction
  - Near native structures with diversified interfaces.
- Difficulty of protein structure prediction
  - Probability of NNS under uniform distribution.

Methods for Studying NNS

- Experimental method, such as NMR
  - Study one protein at a time. Limited to protein types.
- MD simulation
  - Computationally expensive. Applicable for small proteins.
- MCMC
  - Folding around the constrained native structure template is not efficient.
- NMR combined with MD
  - Vendruscolo M, *et. al. Nature* (2005), **433**:128-32

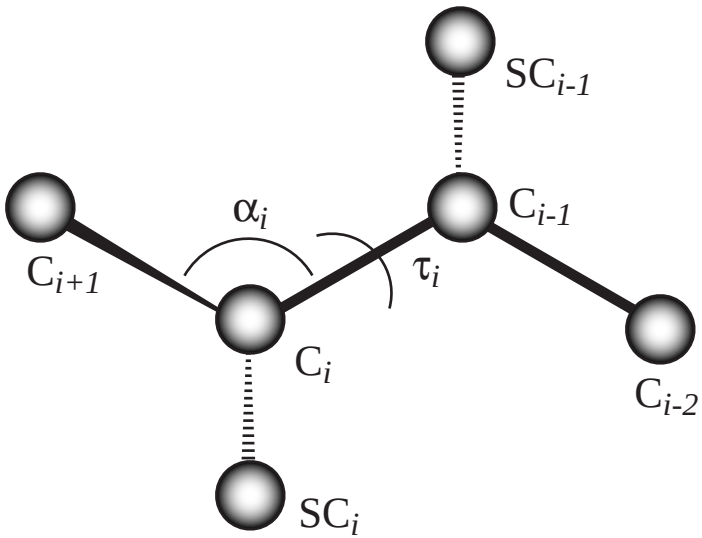
Near Native Structures

-- Connecting Experimental Structures and Applications

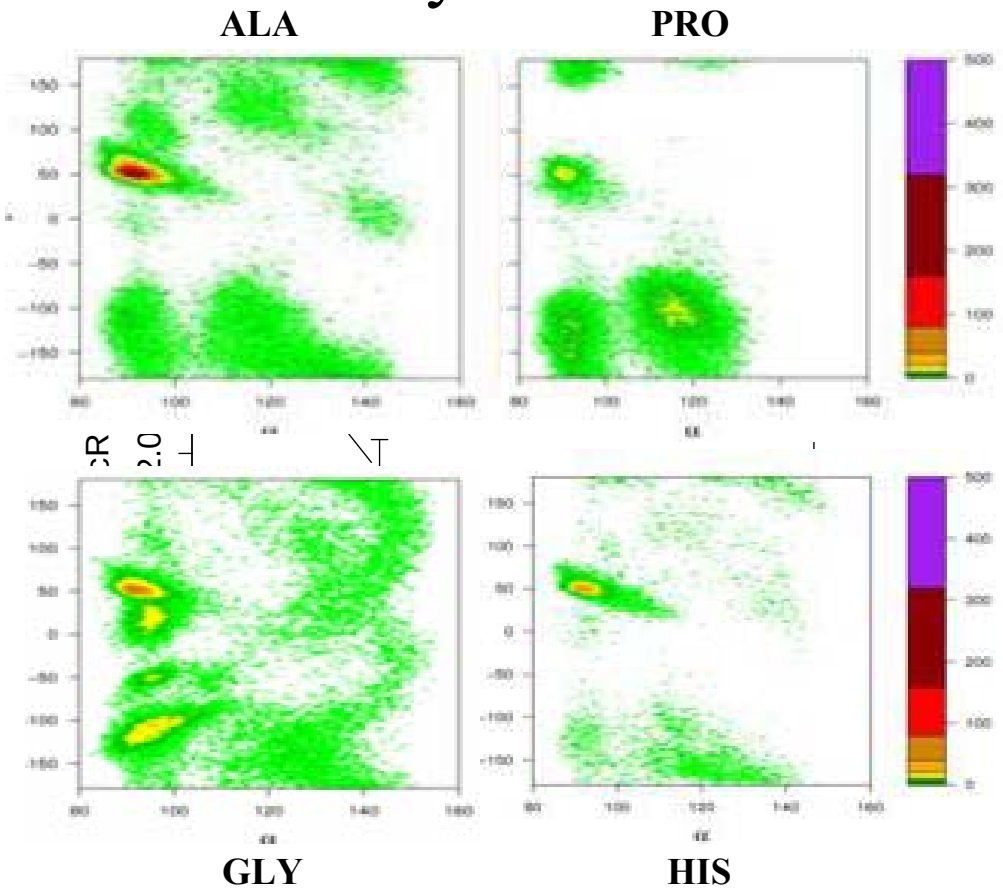


Representation of Protein Structures

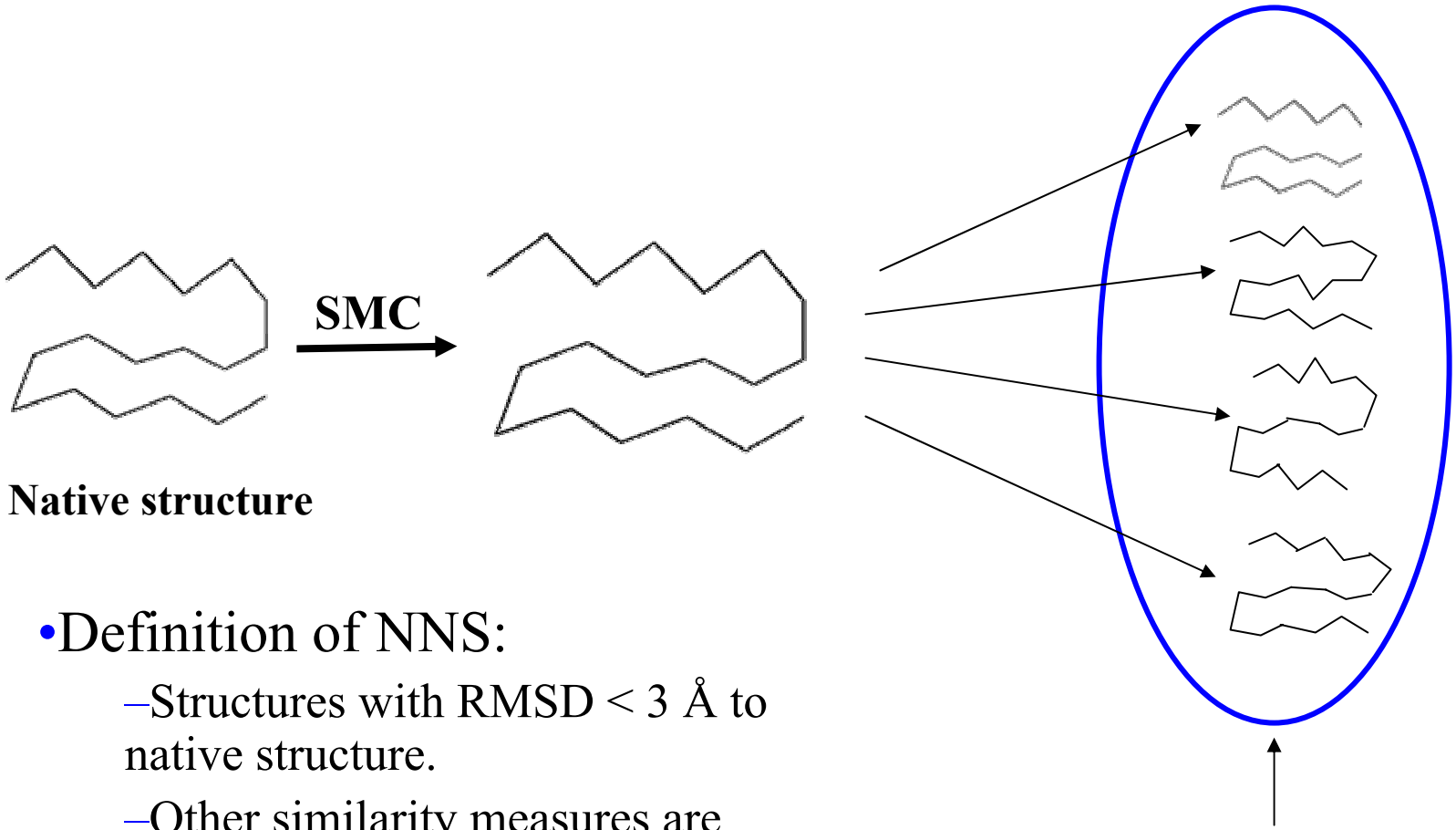
- Optimized discrete state model (ODSM).



- Accuracy of ODSM.



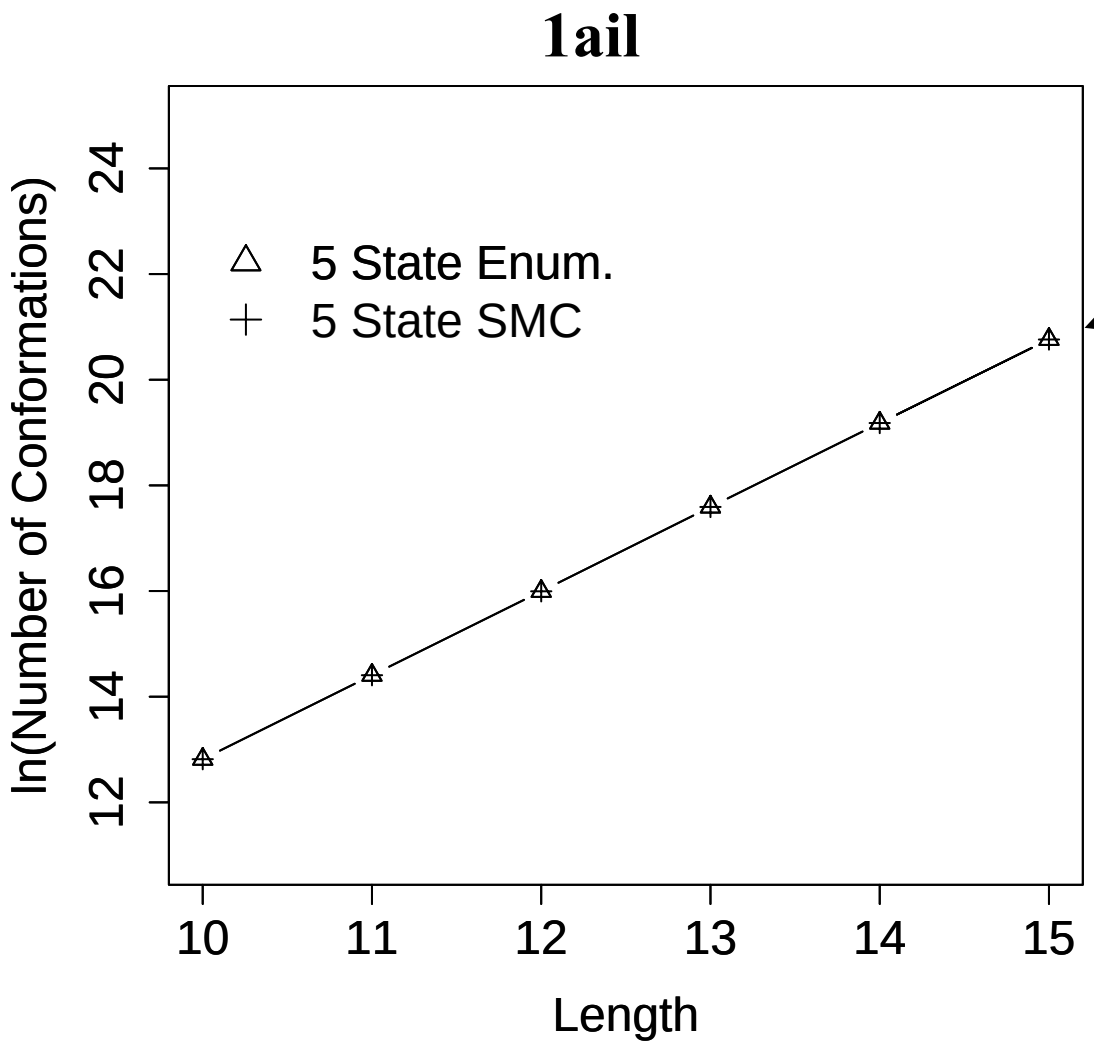
Sequential Monte Carlo for Sampling NNS



- Definition of NNS:
  - –Structures with RMSD < 3 Å to native structure.
  - –Other similarity measures are possible.

Comparison with Enumeration I.

-- Estimation of Number of Conformations



1.042×10^9

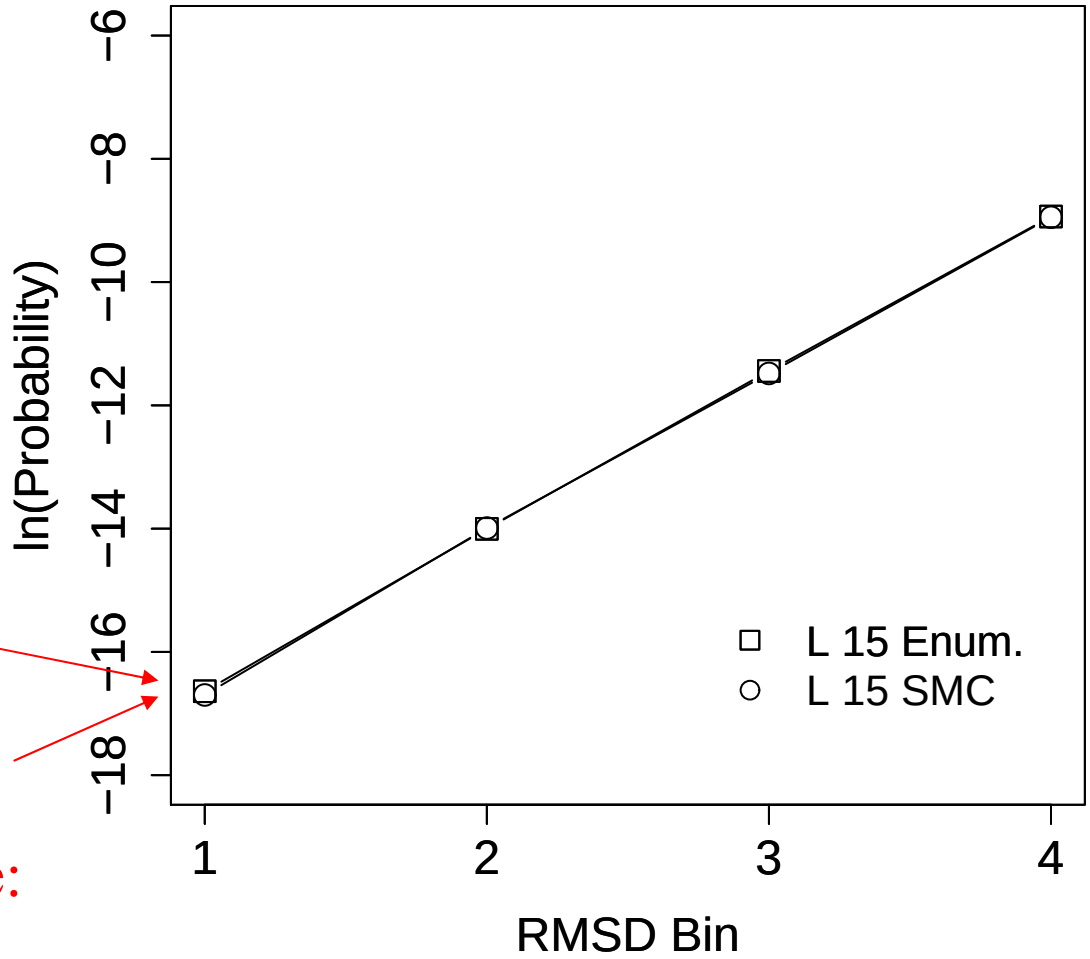
1.039×10^9

Sample size:
10,000.

Comparison with Enumeration II.

-- Estimation of NNS

RMSD Bin:
1: 1.0 Å - 1.5 Å;
2: 1.5 Å - 2.0 Å;
3: 2.0 Å - 2.5 Å;
4: 2.5 Å - 3.0 Å;



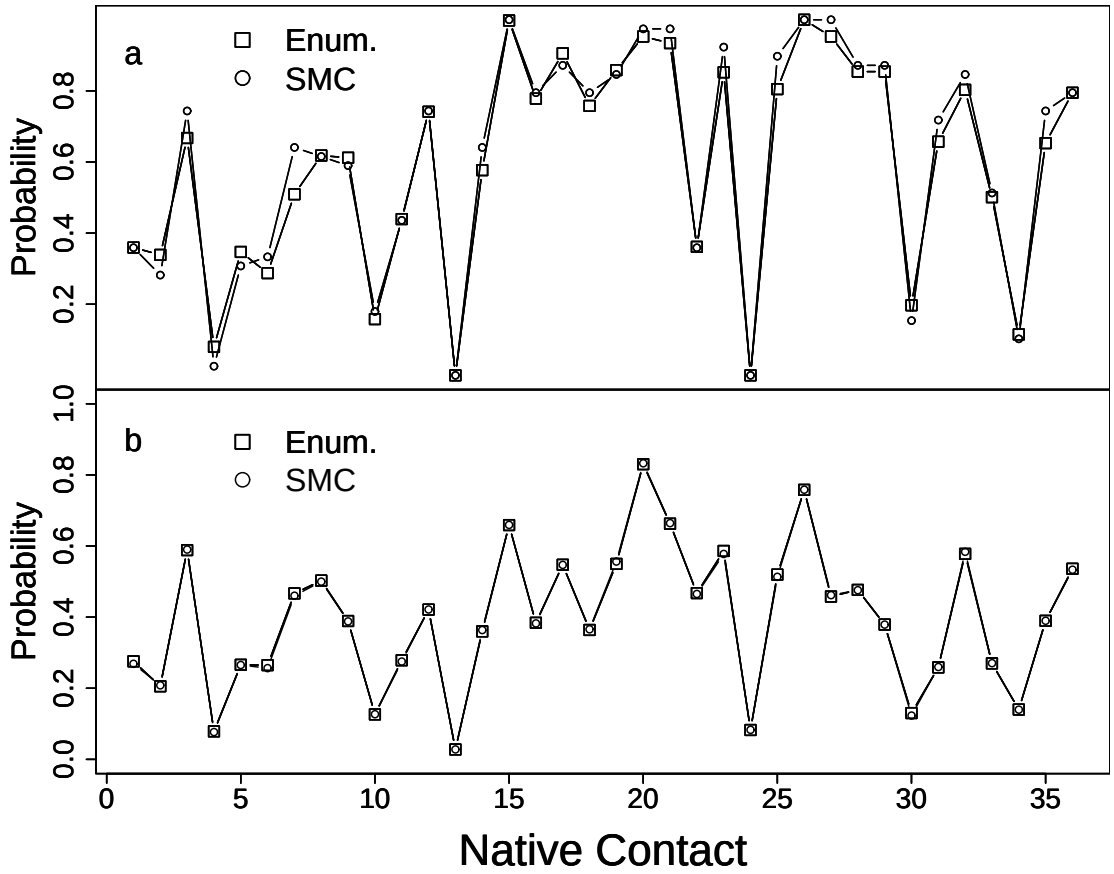
5.94×10^{-8}

5.60×10^{-8}

Sample size:
10,000.

Comparison with Enumeration III.

-- Estimation of Native Contacts



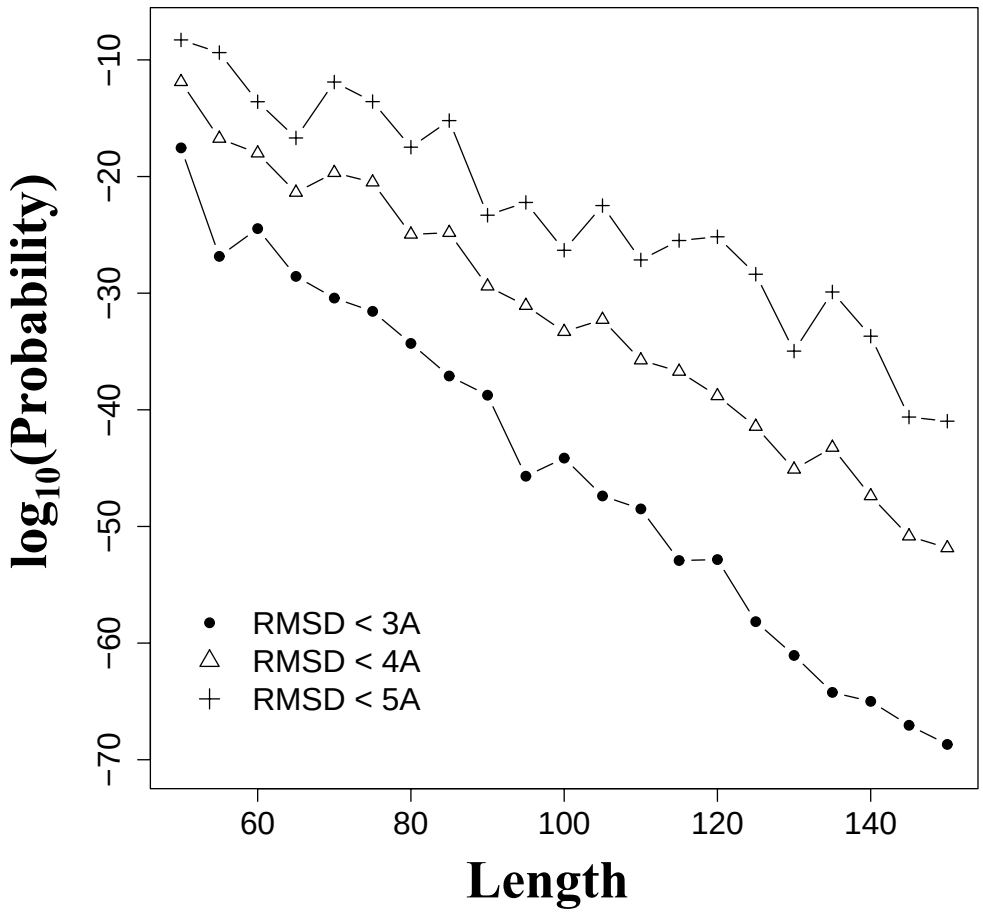
1nkd, RMSD Bin-2: 1.5 Å - 2.0 Å;

1nkd, RMSD Bin-4: 2.5 Å - 3.0 Å;

Probability of NNS

-- How Difficult Protein Structure Prediction is?

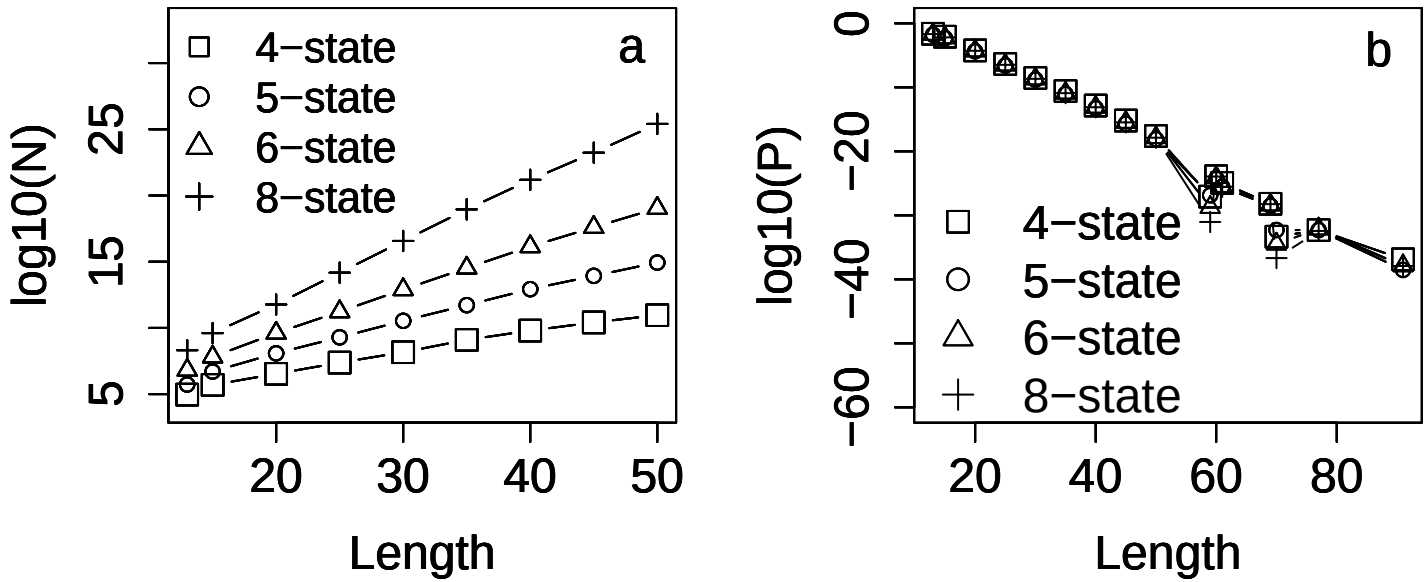
Probability of NNS for 70 non-homologous proteins grouped by their length with 5 residues per interval.



Probability of NNS

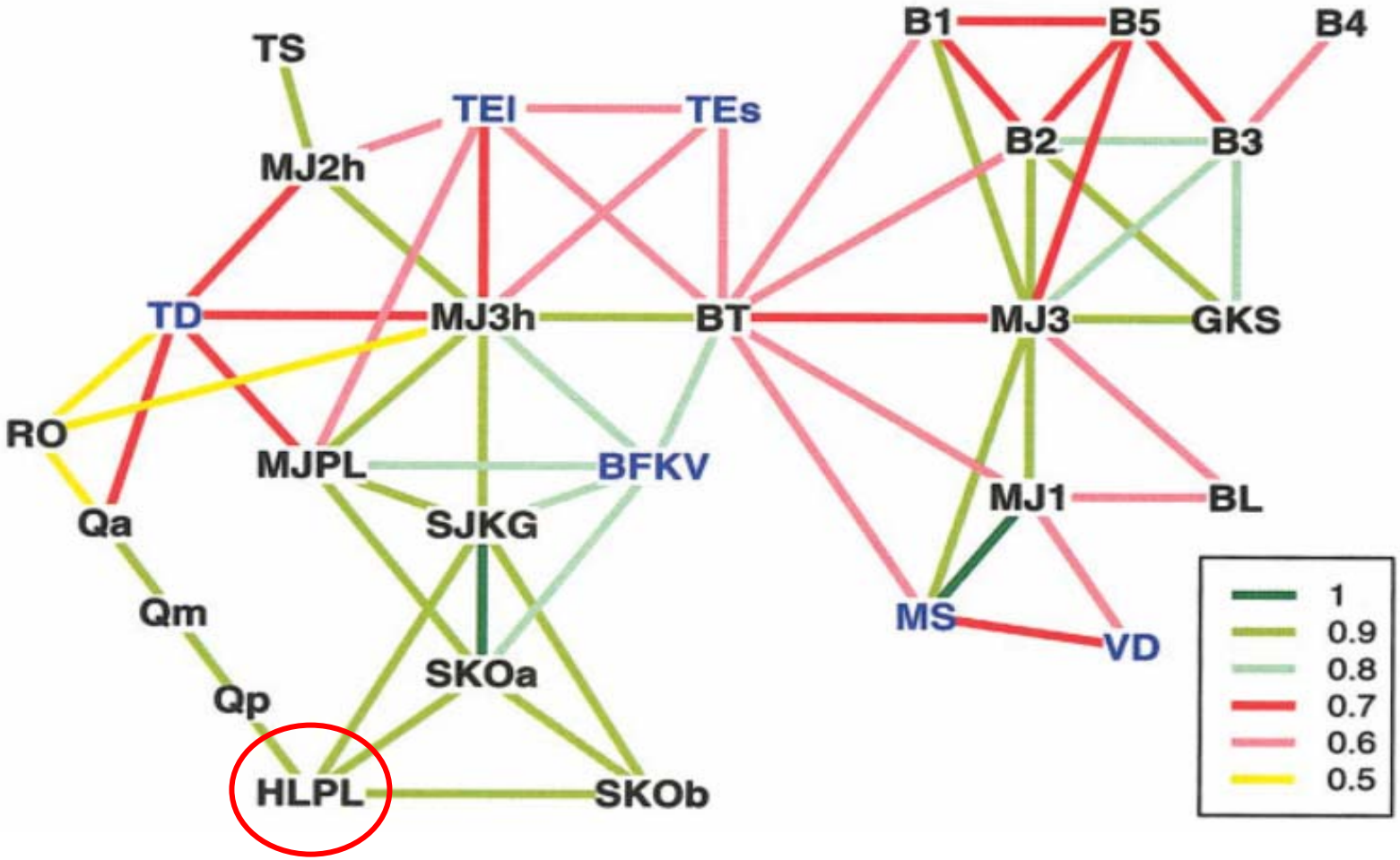
-- Effect of Model Complexity

Average probability of NNS for 8 proteins at partial length and full length.



- 4,5,6,8-state models all have same probability of NNS.

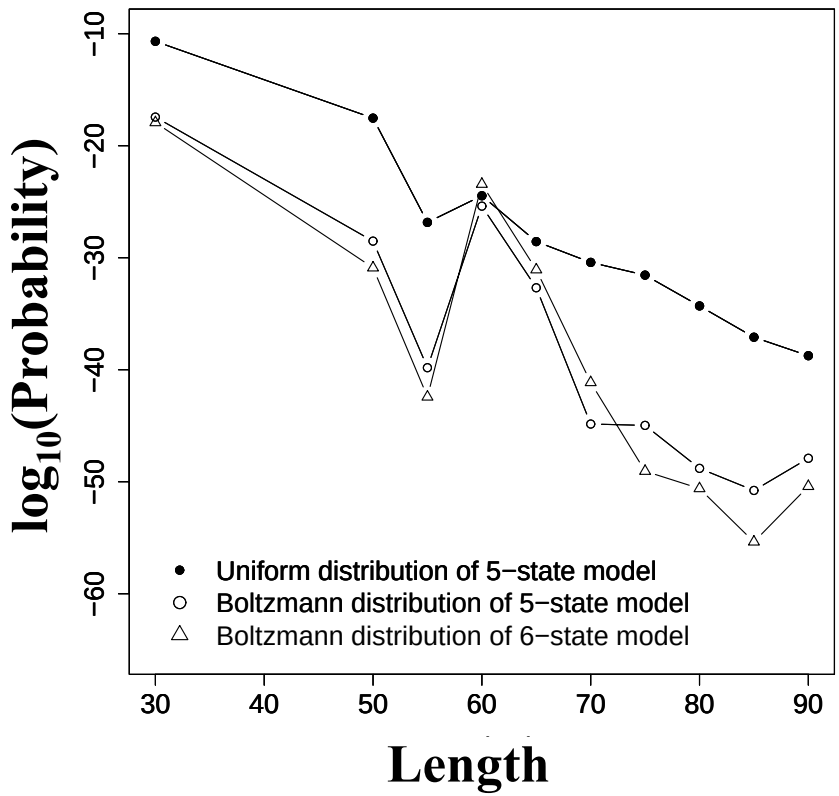
Probability Under Boltzmann Distribution -- Contact Potentials



Piotr Pokarowski *et. al.,* PROTEINS, 59:49–57 (2005)

Probability of NNS Under Boltzmann Distributions

- Probability of NNS for 32 proteins with length from 31 to 90.



- Pair-wise contact potential function stabilize NNS poorly.

Summary for NNS

- Sequential Monte Carlo (SMC) for studying near native structures (NNS).
- Probability of NNS is estimated for proteins up to length 150.
- Models with different complexities have same probability of NNS.
- Rigorous evaluation criterion for potential functions. Contact potentials do not stabilize native structures.

Side Chain Modeling

Introduction

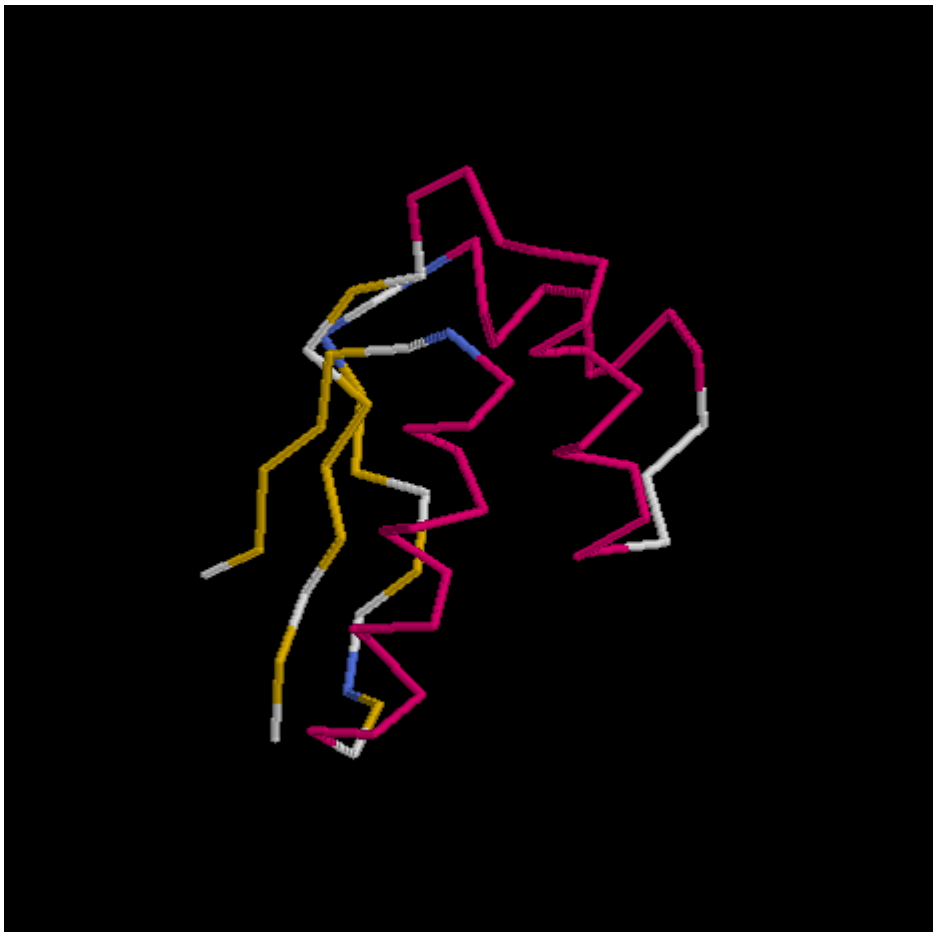
- Side chain modeling is important for protein structure prediction, protein interaction, and protein design.
- Most current methods are looking for single conformation with minimum potential energy.
- In structure prediction, the energy of a conformation is normally calculated ignoring the side chain conformational entropy.

Questions

- Do structures with similar compactness have similar side chain conformational entropy?
- Do structures with similar fold have similar side chain conformational entropy?
- Do native structures have higher side chain entropy than random structures with similar compactness or similar fold?

We address these questions with our new side chain modeling method.

SMC for Side Chain Modeling



- Number of side chain conformations, N_{sc}.

↓

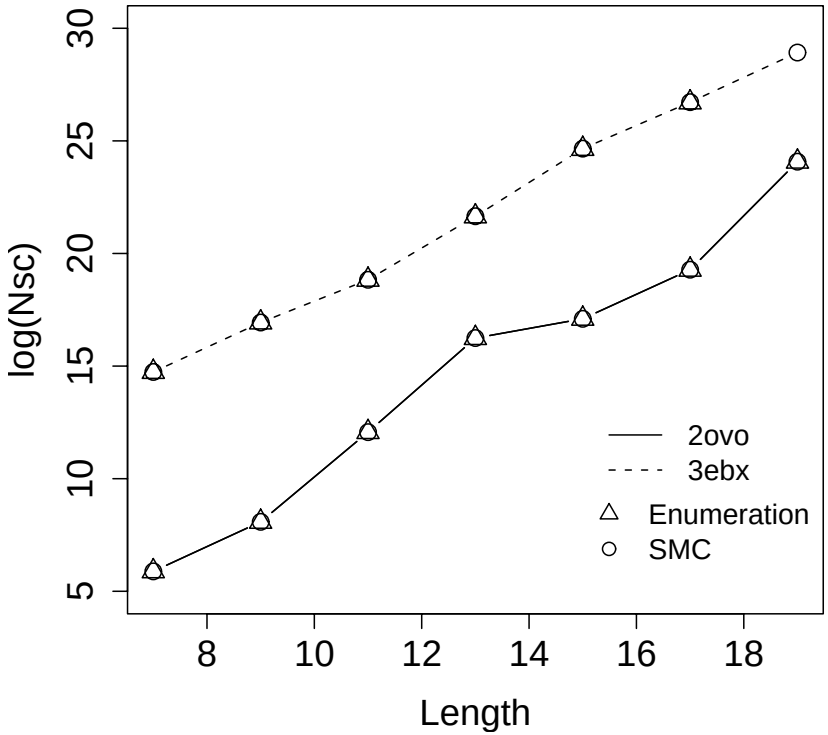
- Side chain conformational entropy.

$$S_{sc} = k_B \ln(N_{sc})$$

↓

- Stability.
- Folding and Packing.

Validation of SMC
-- Comparison with Enumeration



The total SAW side chain conformation for a fragment of 3ebx, residue 1-17, is 396,325,923,840 (3.96×10^{11}).

The estimated number is 4.01×10^{11} with a sample size of 1,000 for 10 runs.

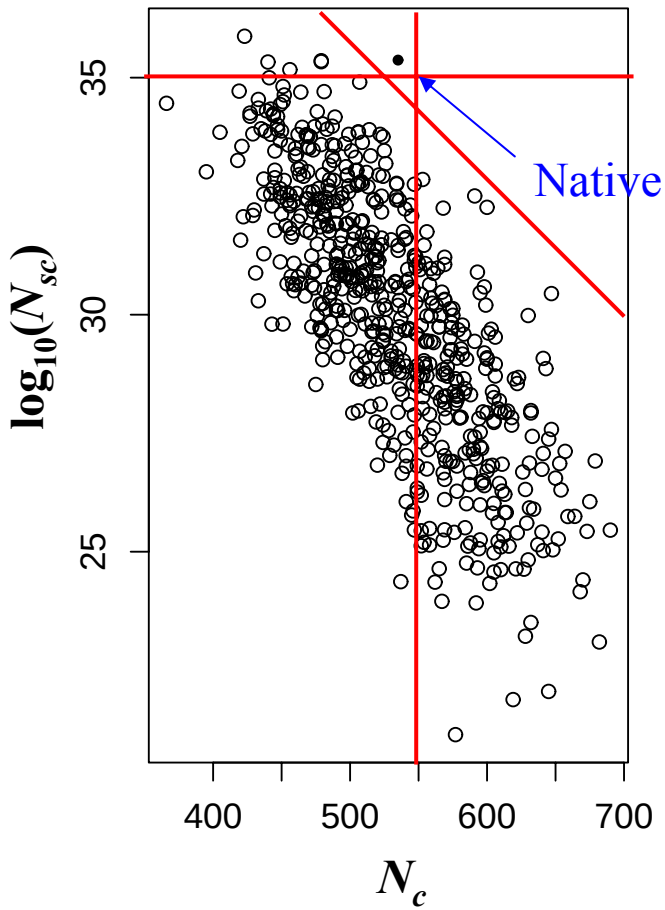
Do structures with similar compactness have similar side chain conformational entropy?

- Structures satisfying:
  - same sequence,
  - similar compactness,
  - different backbone conformations.

Decoys Structures

- Decoys are generated to fool potential functions.
- 24 decoy proteins are selected from 5 decoy sets in Decoys ‘R’ Us database.
  - 4state_reduced: 7 proteins (about 600 structures each protein).
  - fisa: 3 proteins (500 decoys).
  - fisa_casp3: 4 proteins (1000-2500 decoys).
  - lattice_ssfit: 5 proteins (2000 decoys).
  - lmds: 5 proteins (300-500 decoys).
- Compactness are measured by one of the two parameters: radius of gyration (R_g) or number of residue contact (N_c).

Side Chain Entropy of Native and Decoys Structures



1ctf

On average, the number of side chain conformations for native 1ctf is 10^5 times more than a decoy structure!

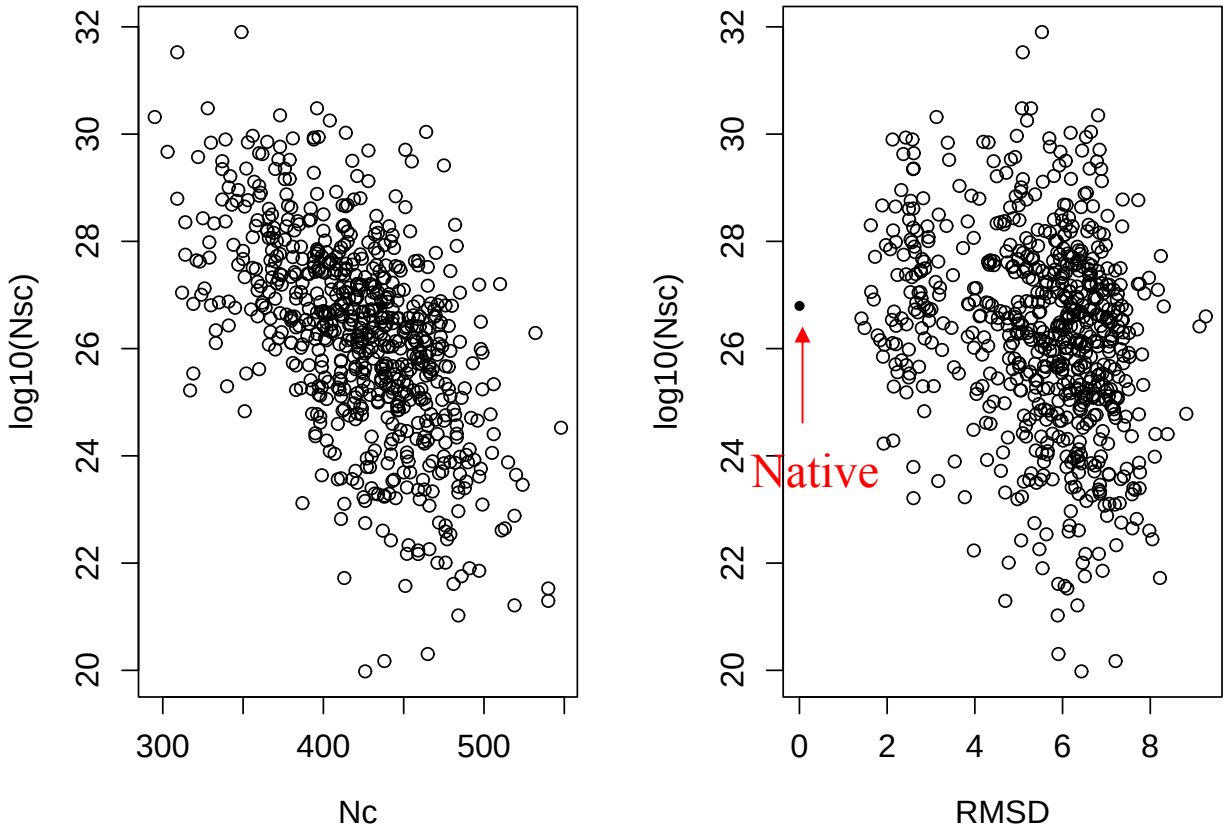
Native vs. Decoys

Protein	Nsc	Type	DecoySet	Protein	Nsc	Type	DecoySet
1ctf	Y		4state	1r69	Y		4state
1sn3	Y		4state	2cro	Y		4state
3icb	N	M	4state	4pti	N	S	4state
4rxn	N	M	4state	1fc2	N	I	fisa
1hdd-C	N	I	fisa	4icb	N	M	fisa
1bg8-A	N	S	fisa_casp3	1bl0	Y		fisa_casp3
1eh2	N	M	fisa_casp3	smd3	Y		fisa_casp3
1beo	N	S	lattice	1dkt-A	N	I	lattice
1fca	Y		lattice	1nkl	Y		lattice
1pgb	Y		lattice	1b0n	N	M	lmds
1bba	N	NMR	lmds	1igd	Y		lmds
1shf	Y		lmds	2ovo	N	S	lmds

Y: Proteins for which side chain entropy is maximized.
N: Proteins for which side chain entropy is not maximized.
M: Metal binding protein
S: Disulfide protein
I: Involved in Interaction

Proteins with Disulfide Bonds

4pti



Structures with similar compactness can have very different side chain conformational entropy.

Native structures tend to maximize side chain conformational entropy.

Do structures with similar conformation have similar side chain conformational entropy?

- Structures satisfying:
  - same sequence,
  - similar (but not the same) conformations.

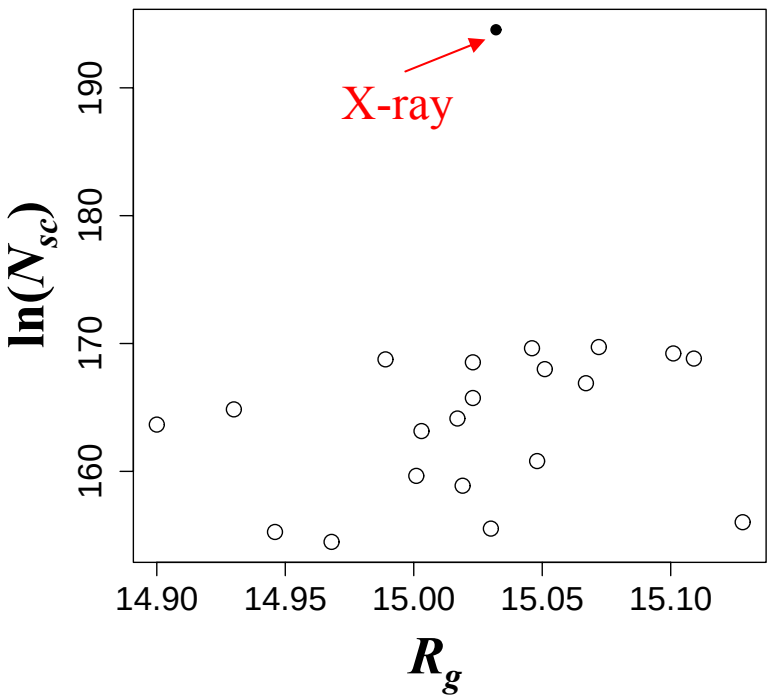
X-ray and NMR Structures

- Experimental X-ray structure vs. NMR structures
  - Very similar backbone folds.
  - Differ in details, such as packing of loop and contacts.
  - Potential derived from X-ray structures fails to recognize NMR structures and *vice versa*. Why?

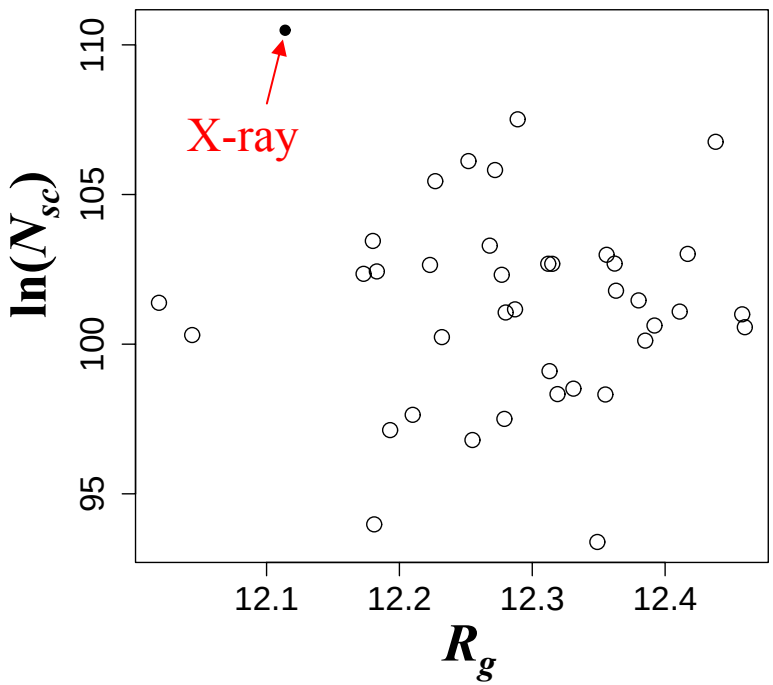
Sergiy O. Garbuzynskiy *et. al.*, Proteins, 60:139–147 (2005)

Side Chain Entropy of X-ray and NMR Structures

1eq0 (NMR) : 1hka (X-ray)

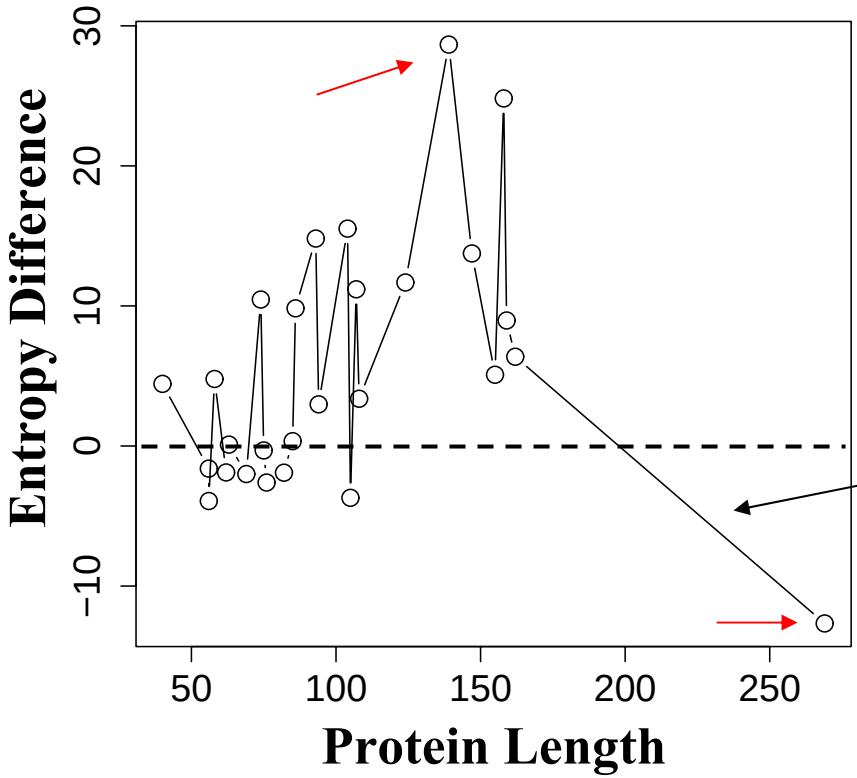


1bmw (NMR) : 1who (X-ray)



X-ray structures have similar fold and compactness as NMR structures, but higher side chain entropy.

Side Chain Entropy Difference between X-ray and NMR Structures



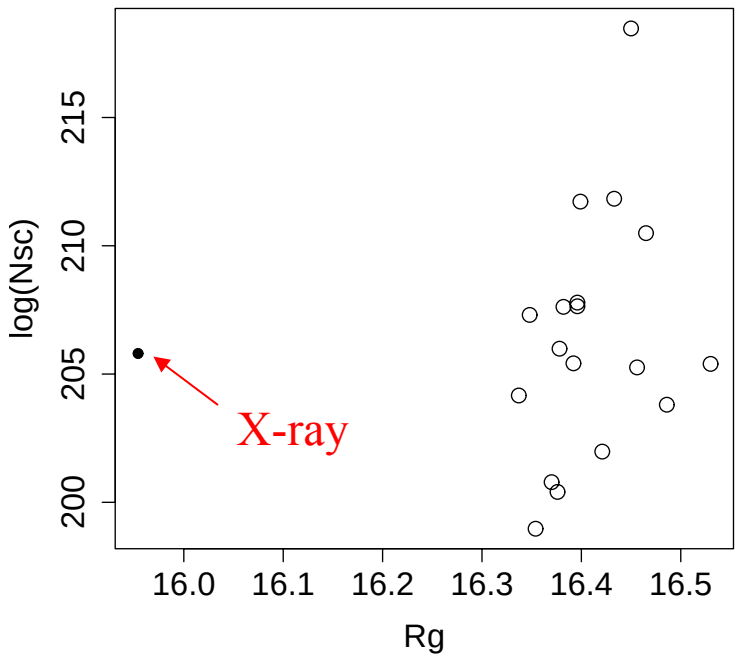
$$\ln \frac{N_{sc}(X-ray)}{\max\{N_{sc}(NMR)\}}$$

In general, X-ray structure has higher side chain entropy than NMR structures of the same protein.

Two Packing Modes
-- Balance between Enthalpy and Entropy

1ah2 (NMR) : 1svn (X-ray)

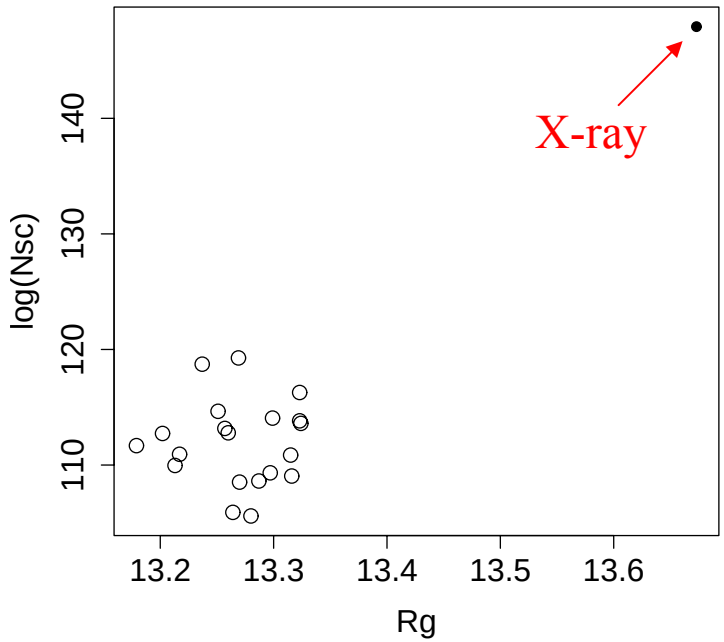
RMSD: 1.76 Å



Higher compactness, comparable side chain entropy.

1pfl (NMR) : 1fil (X-ray)

RMSD: 1.65 Å



Lower compactness, much higher side chain entropy.

Summary for Side Chain Modeling

- Protein folding is a subtle balance between enthalpy and entropy, not simply minimizing enthalpy to compensate the lose of entropy.
- Side chain entropy plays very important role in protein stability, and can be used in discrimination of native and decoy structures, especially similar structures.
- Packing of NMR structures are sub-optimal compared to X-ray structures.

Acknowledgement

Prof. Jun Liu	Computational Biology Lab Department of Statistics Harvard University
Prof. Jie Liang	Bioengineering Department University of Illinois at Chicago
Prof. Rong Chen	Department of Information and Decision Science University of Illinois at Chicago
Dr. Ming Lin	Department of IDS, UIC
NIH	