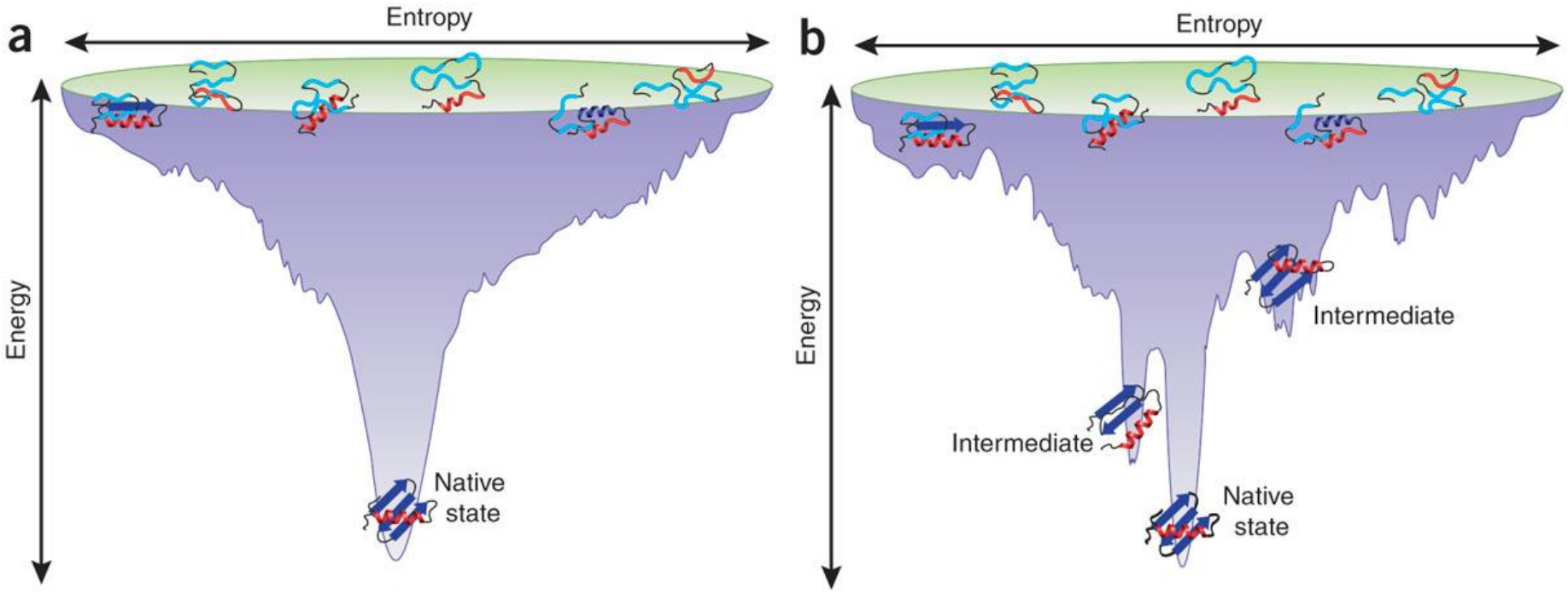
Metropolis Monte Carlo

Folding funnel



Alice I Bartlett & Sheena E Radford (2009). *Nature Structural & Molecular Biology* 16:582-588

MC Simulation - Method

Procedure `MCsearch`$(\phi,\ c,\ \nu)$

Input: ϕ – the number of search steps to perform, c – the current conformation, and ν the search neigbourhood

Output: c' – the modified conformation

for $i \leftarrow 1 \ldots \phi$ **do**
  $c' \leftarrow c$;
  $k \leftarrow \widehat{\mathcal{U}}(1, n)$;
  $c' \leftarrow \mathcal{M}(c', k, \nu)$;
  $\Delta E \leftarrow E(c') - E(c)$;
  if $\Delta E \leq 0$ **then**
    $c \leftarrow c'$;
  else
    $q \leftarrow \mathcal{U}(0, 1)$;
    if $q > e^{\frac{-\Delta E}{T}}$ **then**
      $c \leftarrow c'$;
    endif
  endif
endfor

Replica Exchange Monte Carlo

REMC Simulation - Method

Procedure `REMCSimulation`$(\mathbf{c}, E^*, \phi, \nu)$

Input: $\mathbf{c}$ – the state of the extended ensemble, E^* – the optimal energy, ϕ – the number of local steps, ν – the search neighbourhood

Output: $\mathbf{c}'$ – the modified state of the extended ensemble

```
E' ← 0;
offset ← 0;
while E' > E* do
    foreach replica i in M do
        MCsearch (φ,c_i,ν);
        if E(c_i) < E' then
            E' ← E(c_i);
        endif
    endfch
    i ← offset + 1;
    while i + 1 ≤ M do
        j ← i + 1;
        Δ ← (β_j − β_i)(E(c_i) − E(c_j));
        if Δ ≤ 0 then
            swapLabels (c_i,c_j);
        else
            q ← U(0,1);
            if q ≤ e^{−Δ} then
                swapLabels (c_i,c_j);
            endif
        endif
        i ← i + 2;
    endw
    offset ← 1 − offset;
endw
```

Simulated Annealing

- Simulated annealing.
  - T large $\Rightarrow$ probability of accepting an uphill move is large.
  - T small $\Rightarrow$ uphill moves are almost never accepted.
  - Idea: turn knob to control T.
  - Cooling schedule: T = T(i) at iteration i.

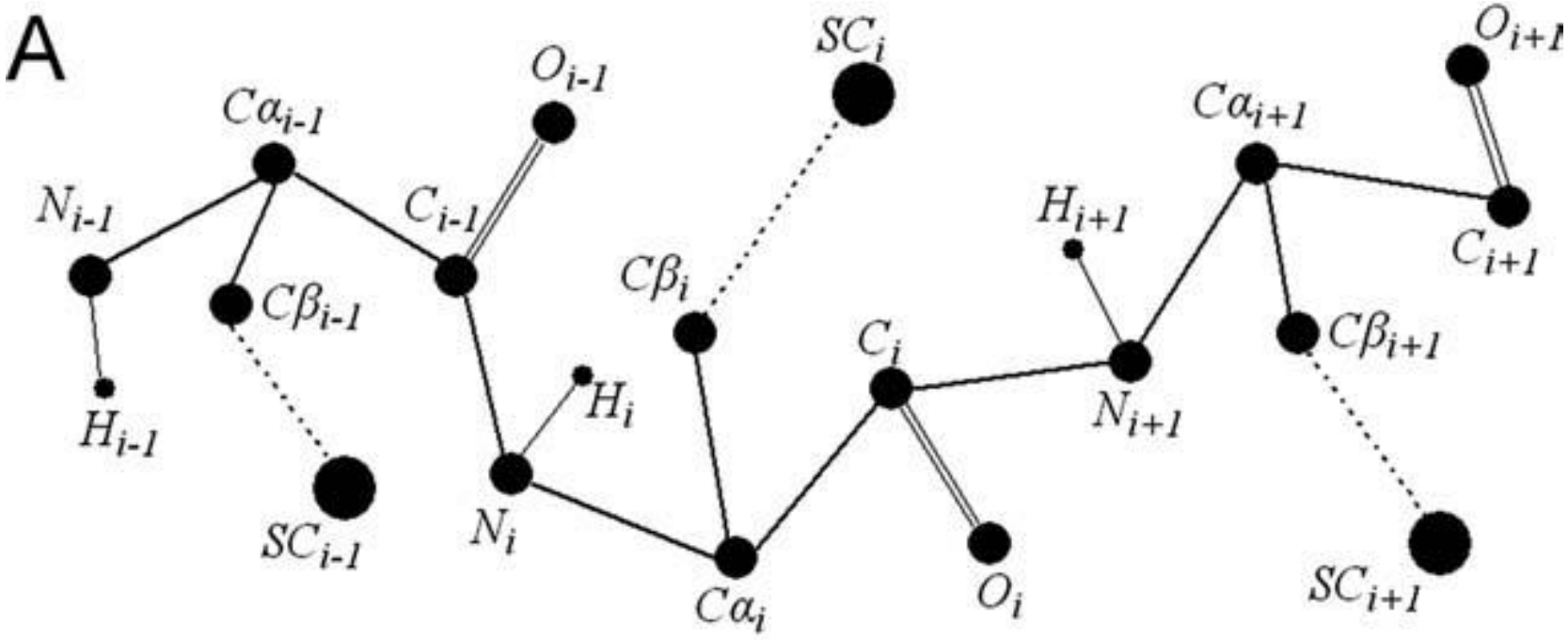
- Physical analog.
  - Take solid and raise it to high temperature, we do not expect it to maintain a nice crystal structure.
  - Take a molten solid and freeze it very abruptly, we do not expect to get a perfect crystal either.
  - Annealing: cool material gradually from high temperature, allowing it to reach equilibrium at succession of intermediate lower temperatures.

Protein Folding
Ab initio Method

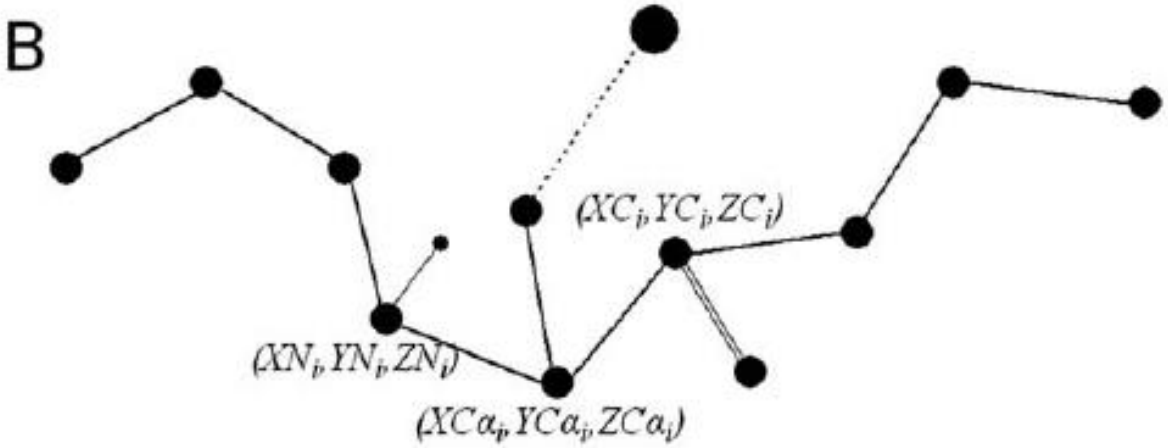
Ab initio/De novo protein modeling

1) An accurate energy function that corresponds the most thermodynamically stable state to the native structure of a protein.
2) An efficient search method capable of quickly identify low-energy states through conformational search.
3) The ability to select native-like models from a collection of decoy structures.

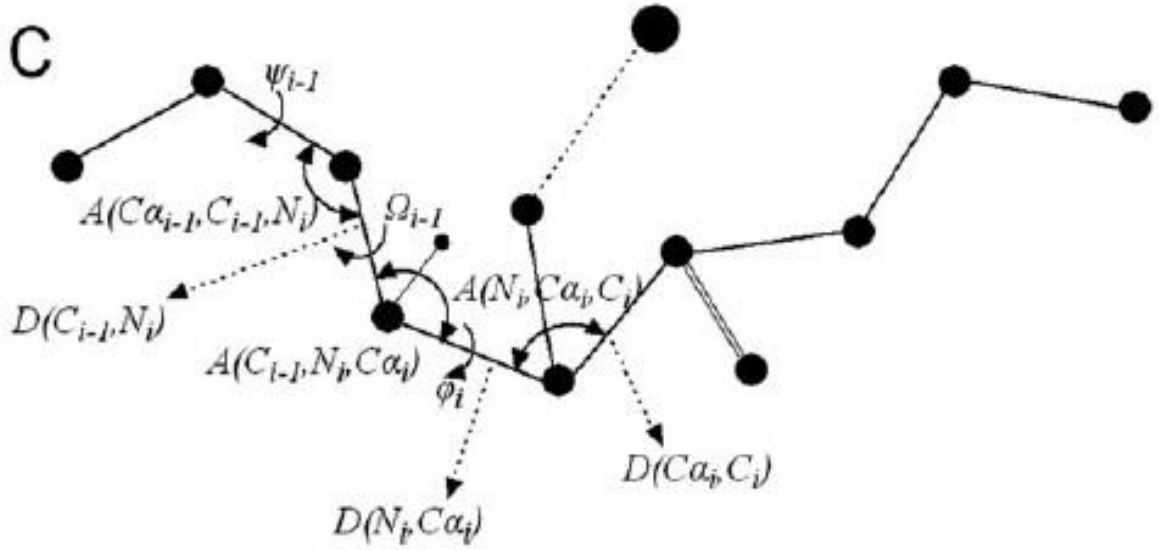
Model representation



Model representation



Cartesian Co-ordinate

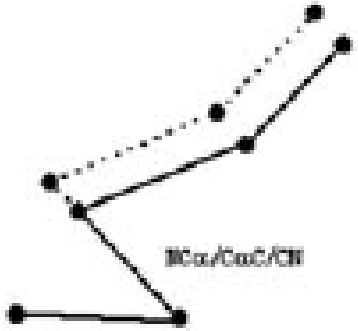


Torsional Angle System

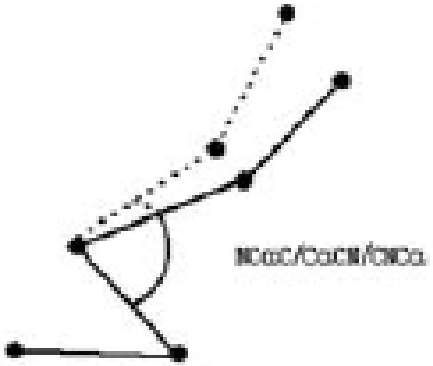
Conformational movements

- Residue-level movements
- Segment-level movements
- Topology-level movements
- Global movements

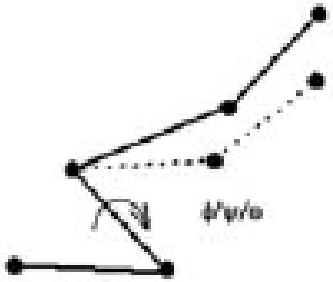
Residue-level movements



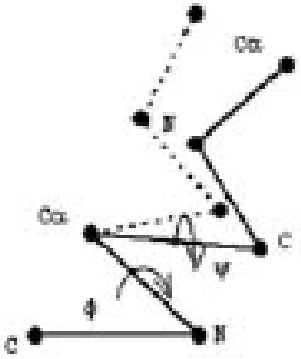
M1



M2

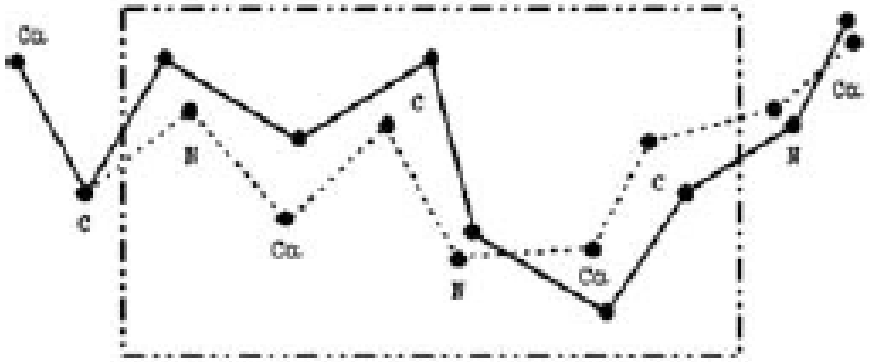


M3

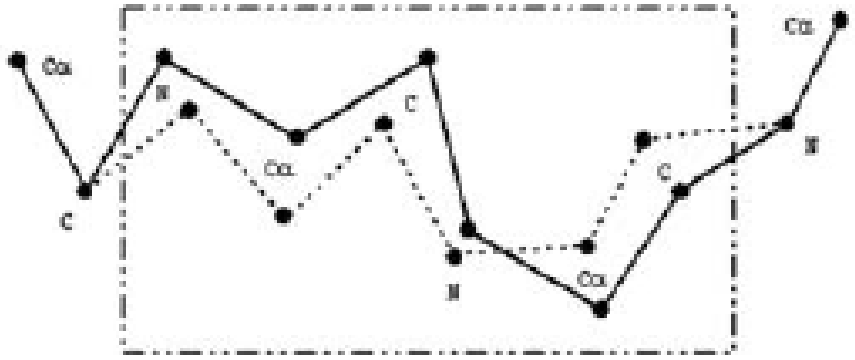


M4

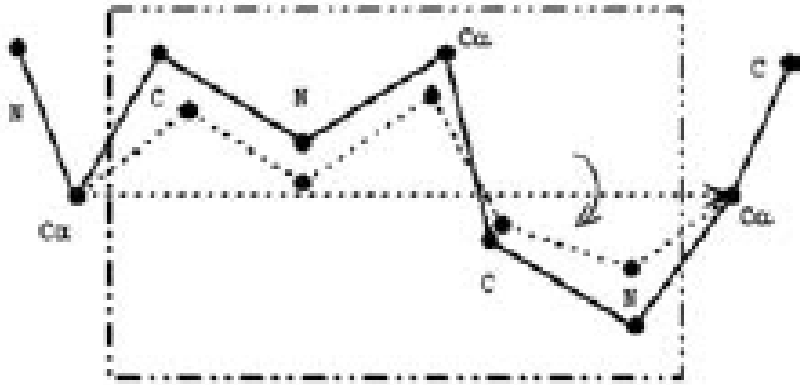
Segment-level movements



M5

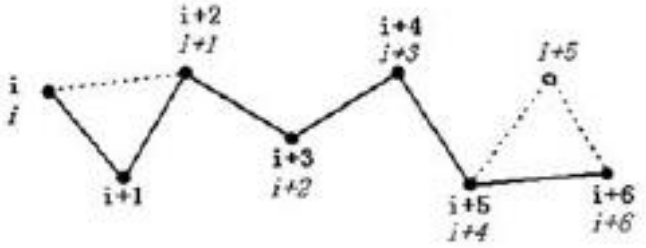


M6

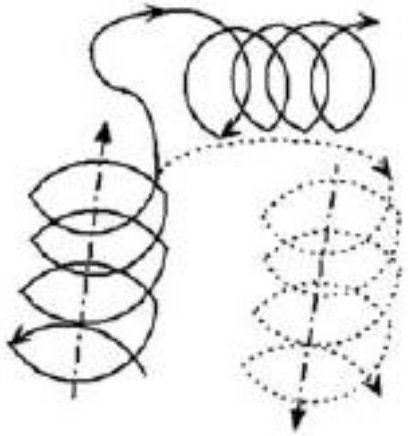


M7

Topology-level movements

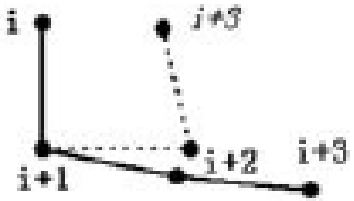


M8



M9

M10



M11

Global movements

- $30L^{0.5}$ movements will be attempted using Metropolish criteria.
- Temperature of each replica is determined by

$$T_i = T_{\min}\left(\frac{T_{\max}}{T_{\min}}\right)^{\frac{40-i}{39}}, \qquad 1 \le i \le 40$$

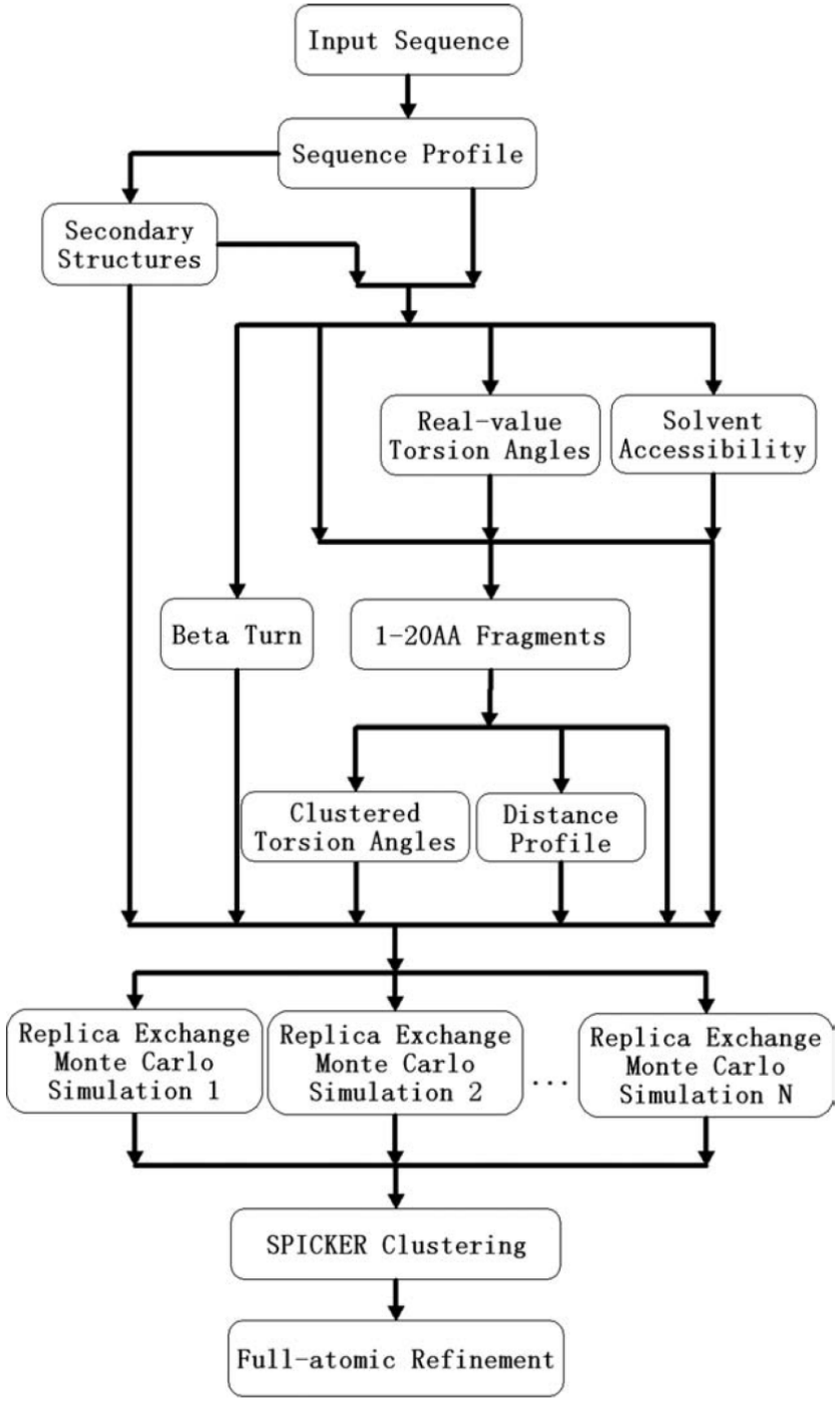
where T_{max}=2.4+0.016L, T_{min}=0.6+0.00067L are the temperatures for the first and last replicas.

Method features

- First torsional angle pairs at each residue position are clustered by standard clustering algorithms to select at most 30 torsion-angle pairs for each residue position.
- The reduced number of torsion angles along with their associated bond lengths and bond angles constitute a look-up table to speed up local movement during simulation.
- Distance profile (a histogram of pair-wise distances extracted from unrelated experimental structures based on the occurrence of fragments at different positions but from the same templates) is extracted from fragments.
- The predicted SA is used in the energy term. The predicted three-state SS types will guide the simulation to generate decoy structures with the similar SS types.
- If one template fragment is successfully placed into the decoy by the fragment substitution movement, this segment will have the same SS types as the fragment structure.
- The predicted probabilities of β-turn positions will be used to guide one movement for b-turn formation.

Ab initio protein folding

- QUARK



Xu and Zhang(2012) Proteins 1715:1735

REMC simulation

- 40 replicas
- 200 cycles are run for each protein by default
- 10 different REMC simulations with different starting random numbers are initiated.
- The Lehmer random number generator is used for random number generation (256 different streams with a long period 2.15×10^9 in each stream).
- In total, 5000 decoys randomly selected from the last 150 cycles of the 10 low-temperature replicas from the 10 simulations are gathered and clustered (using SPICKER).

Energy parameters

1. Backbone atomic pair-wise potential
2. Side-chain center pair-wise potentials
3. Excluded volume
4. Hydrogen Bonding
5. Solvent Accessibility
6. Backbone torsion potential
7. Fragment-based distance profile
8. Radius of gyration
9. Strand–helix–strand packing
10. Helix packing
11. Strand packing

Simulation Energy

$$E_{\text{tot}} = E_{\text{prm}} + w_1 E_{\text{prs}} + w_2 E_{\text{ev}} + w_3 E_{\text{hb}} + w_4 E_{\text{sa}} + w_5 E_{\text{dh}} + w_6 E_{\text{dp}} + w_7 E_{\text{rg}} + w_8 E_{\text{bab}} + w_9 E_{\text{hp}} + w_{10} E_{\text{bp}}$$

w_1=0.1; w_2=0.03; w_3=0.03; w_4=4; w_5=0.4; w_6=0.6; w_7=1.0; w_8=1.0; w_9=0.05; w_{10}=0.1

atomic-level terms (Eprm, Eprs, and Eev),
residue-level terms (Ehb, Esa, Edh, and Edp),
topology-level terms (Erg, Ebab, Ehp, and Ebp)

Atomic-level energy terms

Backbone atomic pair-wise potential

$$E_{\mathrm{prm}}(i,j,r_{ij}) = -RT\log\left(\frac{N_{\mathrm{obs}}(i,j,r_{ij})}{r_{ij}^{\alpha}N_{\mathrm{obs}}(i,j,r_{\mathrm{cut}})}\right)$$

r_{ij} = distance(ith and jth types of atoms); r_{cut} =15 Å; α=1.61; $N_{obs}(i,j,r_{ij})$ observed number of pairs between atoms i & j with distance r_{ij}.

Side-chain center pair-wise potentials

$$E_{\mathrm{prs}}(i,j,r_{ij}) = -RT\log\left(\frac{N'_{\mathrm{obs}}(i,j,r_{ij})}{r_{ij}^{\alpha'}N'_{\mathrm{obs}}(i,j,r_{\mathrm{cut}})}\right)$$

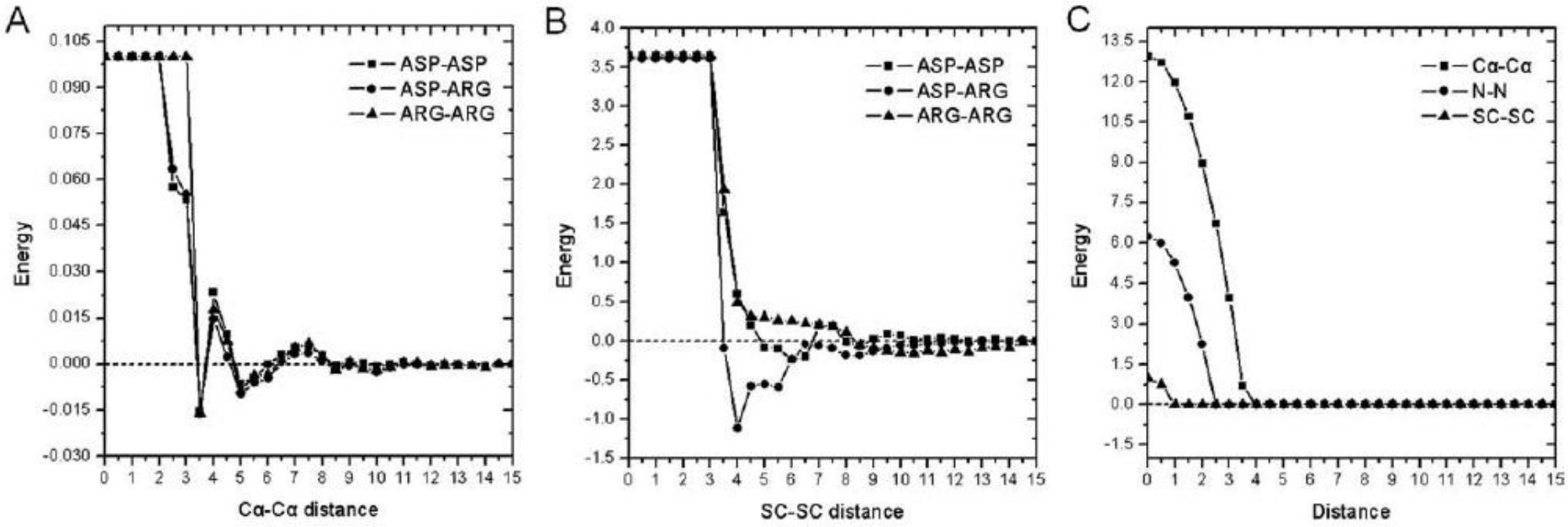
α'=1.40; $N'_{obs}(i,j,r_{ij})$ observed number between SC and SC or other backbone atoms i & j with distance r_{ij}.

Excluded volume

$$E_{\mathrm{ev}}(i,j,r_{ij}) = \begin{cases} (vdw(i)+vdw(j))^2 - r_{ij}^2 & \text{if } r_{ij} < vdw(i)+vdw(j) \\ 0 & \text{else} \end{cases}$$

vdw(i) is the van der Waals radius of the ith atom

Distance specific contact potential



Residue-level energy terms

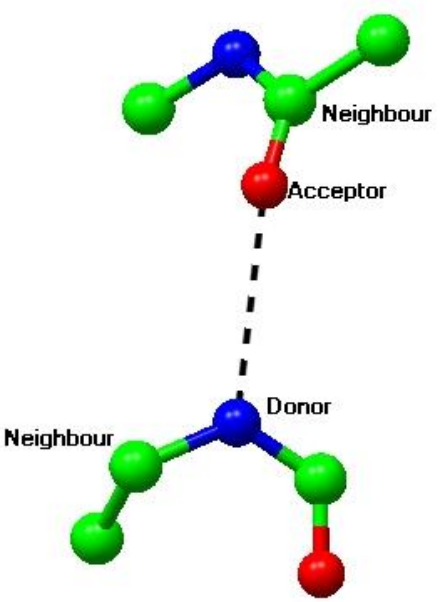
Hydrogen Bonding

$$E_{\mathrm{hb}}(i,j,T_k)=\sum_{l=1}^{n_k}\frac{(f_l(i,j)-\mu_{kl})^2}{2\delta_{kl}^2},\qquad n_k=\begin{cases}4 & k=1,2\\ 3 & k=3,4\end{cases}$$

T_k denotes the kth type of H-bond restraint;
$f_l(i,j)$ is the lth feature calculated from decoy structures;
μ_{kl} and δ_{kl} are the mean and standard deviation of the lth feature of type k H-bond in Table I.

Table I
Mean and Standard Deviation of Four H-Bond Features in α-Helix and β-Sheet Structures

	Acceptor i, *donor j*	$D(O_i,H_j)$ (Å)	$A(C_i,O_i,H_j)$ (°)	$A(O_i,H_j,N_j)$ (°)	$T(C_i,O_i,H_j,N_j)$ (°)
T_1	Helix, $j=i+4$	2.00/0.53	147/10.58	159/11.25	160/25.36
T_2	Helix, $j=i+3$	2.85/0.32	89/7.70	111/8.98	−160/7.93
T_3	Parallel	2.00/0.30	155/11.77	164/11.29	180/68.96
T_4	Antiparallel	2.00/0.26	151/12.38	163/11.02	−168/69.17



Residue-level energy terms

Solvent Accessibility

$$E_{sa} = \sum_{i=1}^{L} |s_i - s_i^E|$$

$$s_i = 1 - w \sum_{d(G_i,G_j)<9\text{Å}} \frac{A_{aa(j)}}{d^2(G_i, G_j)}$$

L is the sequence length; s_i^E is the expected SA of the ith residue, which is predicted by the back-propagation NN trained from the checkpoint file by PSI-BLAST and SS types by DSSP.

A_{aa} is the precalculated maximum solvent accessible surface area for amino acid aa.
w = 0.007.
G_i is the geometric center of the ith residue, calculated from the coordinates of N, Ca, C, O, Cb, and SC atoms.
$d(G_i,G_j)$ is the distance between the ith and jth residue centers.

Residue-level energy terms

Backbone torsion potential

$$E_{\mathrm{dh}} = -\sum_{i=2}^{L-1} \log(P(\phi_i, \psi_i | aa(i), ss(i)))$$

ϕ_i and ψ_i are torsion-angles of ith aa; $P(\phi, \psi|aa,ss)$ is the conditional probability of ϕ and ψ at the residue type aa and the SS type ss, which are calculated from the high-resolution experimental structures.

Fragment-based distance profile

$$E_{\mathrm{dp}} = -\sum_{(i,j)\in S_{\mathrm{dp}}} \log(N_{i,j}(d_{ij}))$$

d_{ij} is the distance between the ith and jth Ca atoms in the decoy structure.
$N_{i,j}(d)$ is the distance profile for residue i and j extracted from the 10-mer fragment structures, with d divided from 0 to 9 Å in the interval 0.5 Å.
S_{dp} is the set of residue pairs with distance profiles.

Topology-level energy terms

Radius of gyration

$$E_{\mathrm{rg}} = \begin{cases} 0 & r_{\min} \le r \le r_{\max} \\ (r_{\min} - r)^2 & r < r_{\min} \\ (r - r_{\max})^2 & r > r_{\max} \end{cases}$$

r is the radius of gyration of the simulated decoy structure, r_{min} and r_{max} are the minimum and maximum of estimated radius of gyration

Strand–helix–strand packing

$$E_{\mathrm{bab}} = \begin{cases} E_{\mathrm{pen}} & \mathrm{left-handed} \\ 0 & \mathrm{else} \end{cases}$$

the penalty energy E_{pen} equals to the negative value of the total hydrogen bonding energy between the two β-strands in the motif

Topology-level energy terms

Helix packing

$$E_{\mathrm{hp}}(i,j) = -\log(P(d_{ij},\ \phi_{ij}))$$

d_{ij} is the distance between the medial axis of the ith helix and that of the jth helix;
ϕ_{ij} is the torsion angle of the axis vectors which are oriented from N- to C-terminal;
$P(d_{ij}, \phi_{ij})$ is the probability distribution calculated from the nonredundant experimental structures, where d_{ij} is split into 30 bins in [0, 15 Å] and ϕ_{ij} into 36 bins in [-180°, 180°].

Strand packing

$$E_{\mathrm{bp}}(i,j) = -\log(P(aa(i), aa(j), T_{ij}))$$

P(A, B, T) is the probability for amino acids A and B in the sheet type T (parallel or antiparallel), calculated from the high-resolution experimental structures.

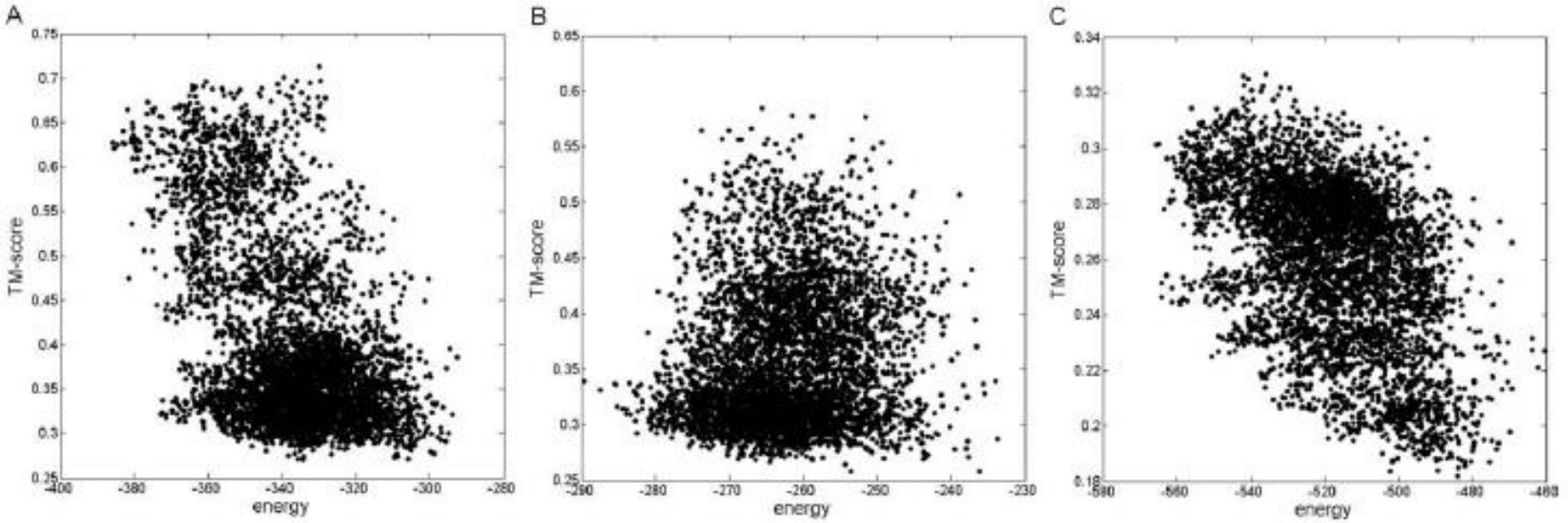
Dataset

Goal is to predict the structure of proteins which have no homologous templates hit by threading algorithms and are inaccurate to predict by template based methods.

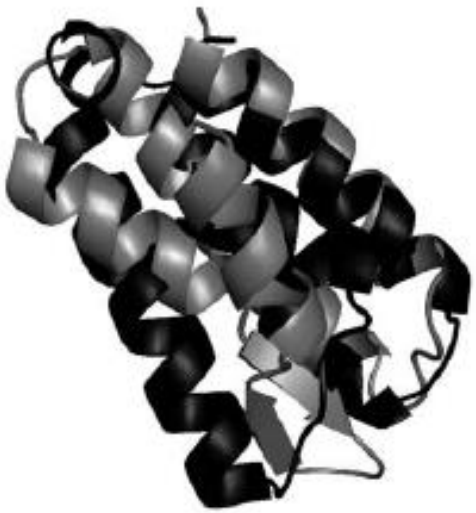
- A nonredundant set of 6023 proteins from PDB with sequence identity <25%, resolution<1.8 Angstrom, and R-factor<0.25.
- Run LOMETS to identify ‘‘Hard’’ targets (means none of the threading algorithms detects a template with the Z-score higher than the given cutoff). In total, 665 sequences are identified.
- Manually check and exclude the targets which have obvious broken chains or incompact shapes. The remaining list contains 413 proteins.
- Training set consists of randomly picked 88 small globular proteins as the training set (length 70-100 aa).
- Test set consists of randomly selected 145 globular proteins (51 small: 70–100 aa and 94 medium-sized: 100–150 aa).

QUARK: Result

Xu and Zhang(2012) Proteins 1715:1735



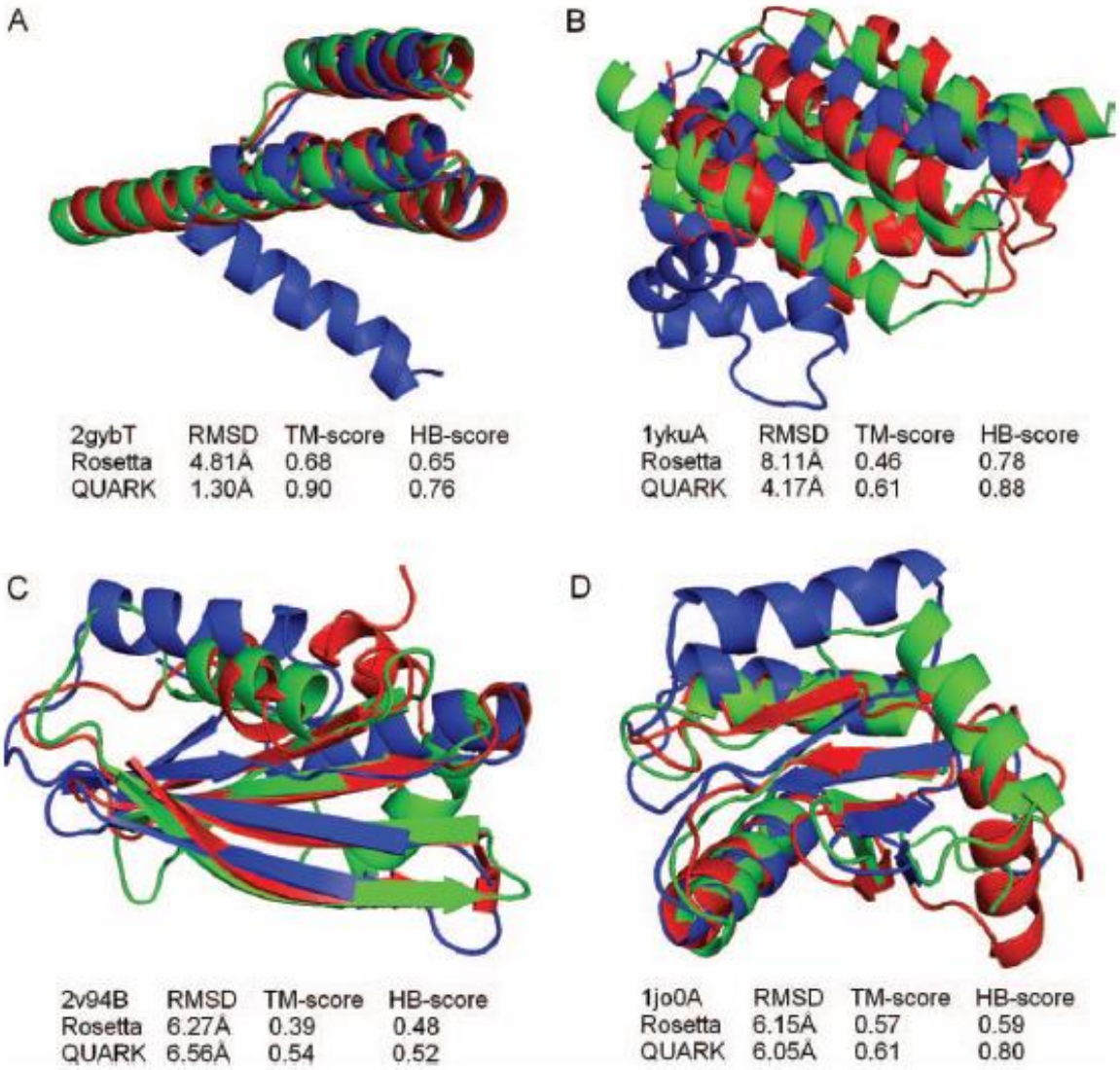
1b4bA, PCC=-0.475, TM-score=0.624



2bl7A, PCC=-0.116, TM-score=0.527

2o42A, PCC=-0.489, TM-score=0.229

QUARK: Result



2gybT	RMSD	TM-score	HB-score
Rosetta	4.81Å	0.68	0.65
QUARK	1.30Å	0.90	0.76

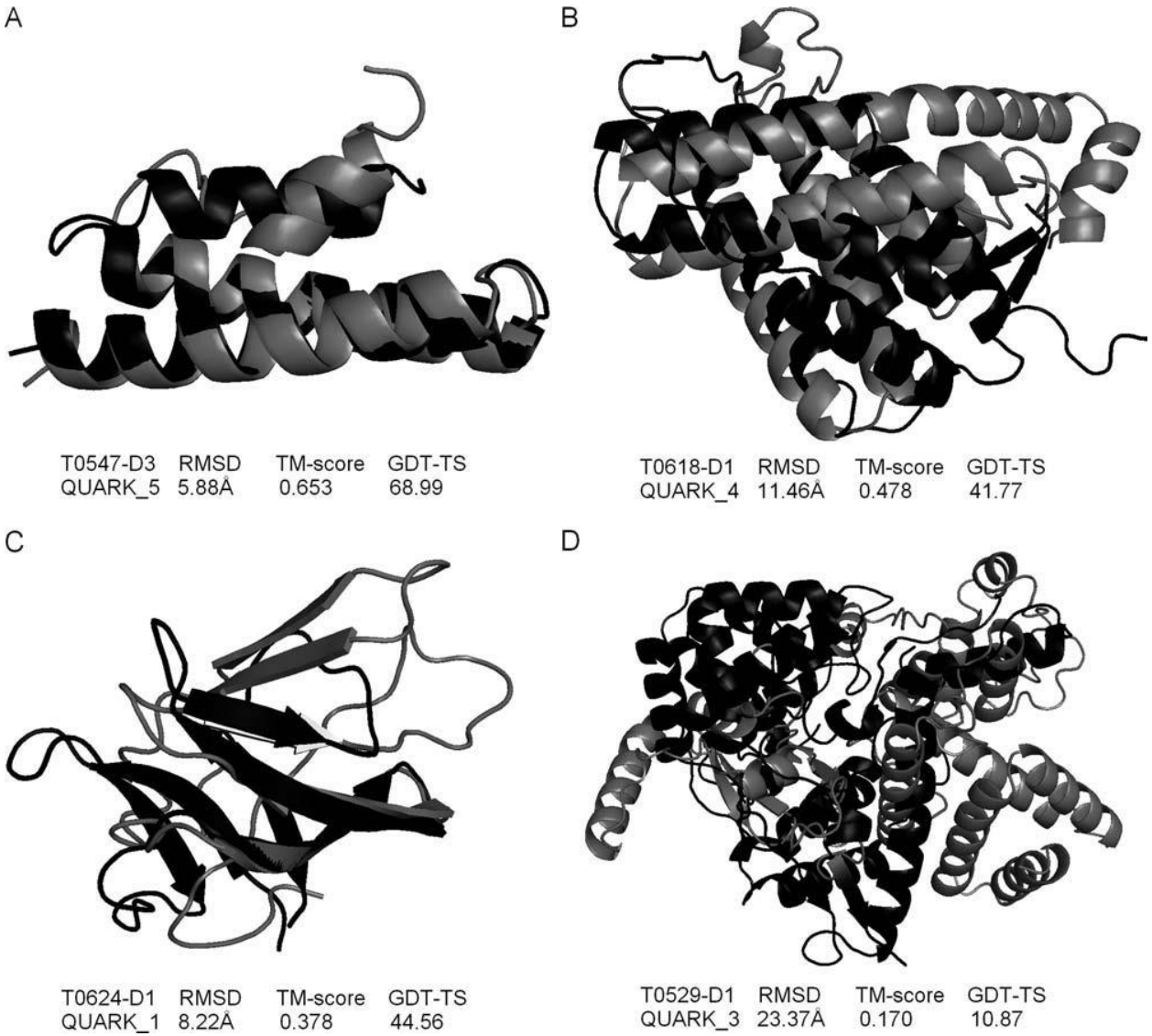
1ykuA	RMSD	TM-score	HB-score
Rosetta	8.11Å	0.46	0.78
QUARK	4.17Å	0.61	0.88

2v94B	RMSD	TM-score	HB-score
Rosetta	6.27Å	0.39	0.48
QUARK	6.56Å	0.54	0.52

1jo0A	RMSD	TM-score	HB-score
Rosetta	6.15Å	0.57	0.59
QUARK	6.05Å	0.61	0.80

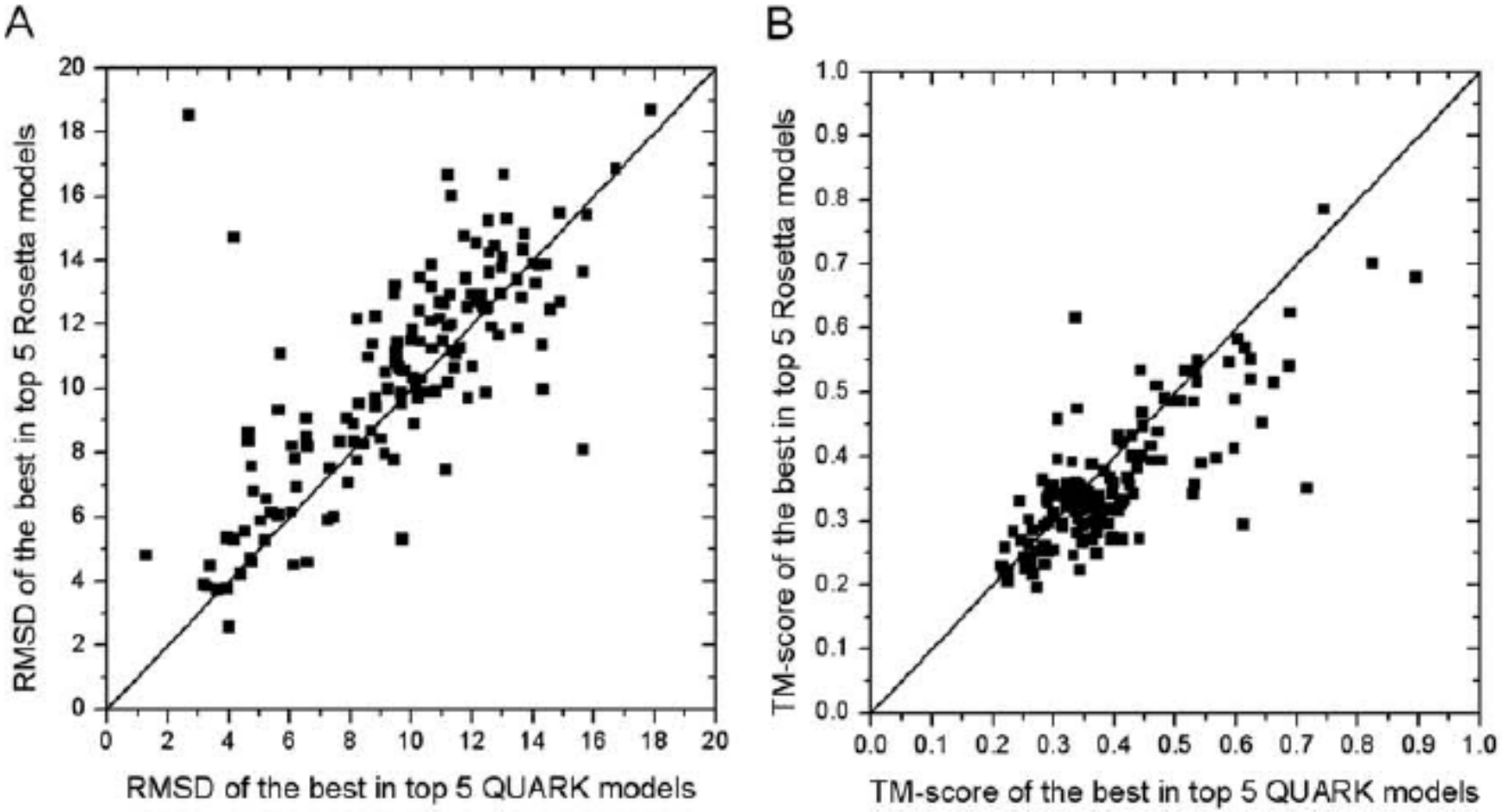
Xu and Zhang(2012) Proteins 1715:1735

QUARK: Result



Xu and Zhang(2012) Proteins 1715:1735

QUARK and Rosetta



Server and Metaserver

- Server – database or web
- Metaserver
  - A unified view of multiple servers/resources
  - Normalization of data across different sources
  - Generate a consensus
  - May or may not have any conclusion from the multiple resources

LOMETS: A local meta-threading-server for protein structure prediction

Wu and Zhang (2007) *NAR*, 35(10)3375-82

Component threading programs in LOMETS

1. FUGUE
2. HHSEARCH
3. PROSPECT2
4. SAM-T02
5. SPARKS2
6. SP3
7. PAINT
8. PPA-I
9. PPA-II

FUGUE

FUGUE is developed at the Blindell Lab. It aligns target sequence profile against template structural profile collected from HOMSTRAD.

Dynamic programming algorithm is used to find the best sequence–structure match.

PROSPECT2

PROSPECT2 is developed at the Xu Lab, which uses a score function including residue mutations, secondary structure propensity, solvent accessibility and pairwise contact potential.

A divide-and-conquer searching approach is exploited to generate the global optimization of alignments.

SPARKS2 and SP3

Both methods have been developed at the Zhou lab. In SPARKS2, the autours exploit a sequence profile–profile alignement combined with a single-body knowledge-based statistical potential; in SP3, they use a residue depth dependent structure profile to replace the single-body potential in the SPARKS2.

Both methods use dynamic programming for the sequence–structure alignment search.

SAM-T02

SAM-T02 is developed at the Karplus lab, which starts from the PSI-Blast sequence database search. Based on the PSI-Blast multiple sequence alignment, a hidden Markov model (HMM) will be constructed in an iterative way, which is then exploited to search through the whole template library by the Viterbi algorithm.

HHSEARCH

HHSEARCH is developed at the Soding lab, which aligns the profile HMM of target with the profile HMM of templates by maximizing the log-sum-of-odds score.

Z-score cutoff

- Z-score (the energy in standard deviation units relative to mean)

- For PPA-I, SP3, PPA-II, SPARKS2, PROSPECT2, FUGUE, HHSEARCH, PAINT and SAM-T02, the Z-score cut are 8.2, 8.0, 7.0, 8.8, 4.0, 6.0, 11.0, 0.5 and 9.5, respectively.

Threading Model Selection

- Models in LOMETS are selected from individual servers purely based on consensus
  - 30 models are taken from the top predictions of the nine servers sequentially from PPA-I, SP3, PPA-II, SPARKS, PROSPECT, FUGUE, HHSEARCH, PAINT and SAM-T02
  - Select first of all, then second of all …..
  - the order of the servers are based on their performance on independent test runs.

Consensus scoring and ranking

- Each (ith) of the 30 models is calculated by the average TM-score

$$\langle \text{TM-score}_i \rangle = \frac{1}{29} \sum_{j=1}^{29} \text{TM-score}_{ij}.$$

Normalization?

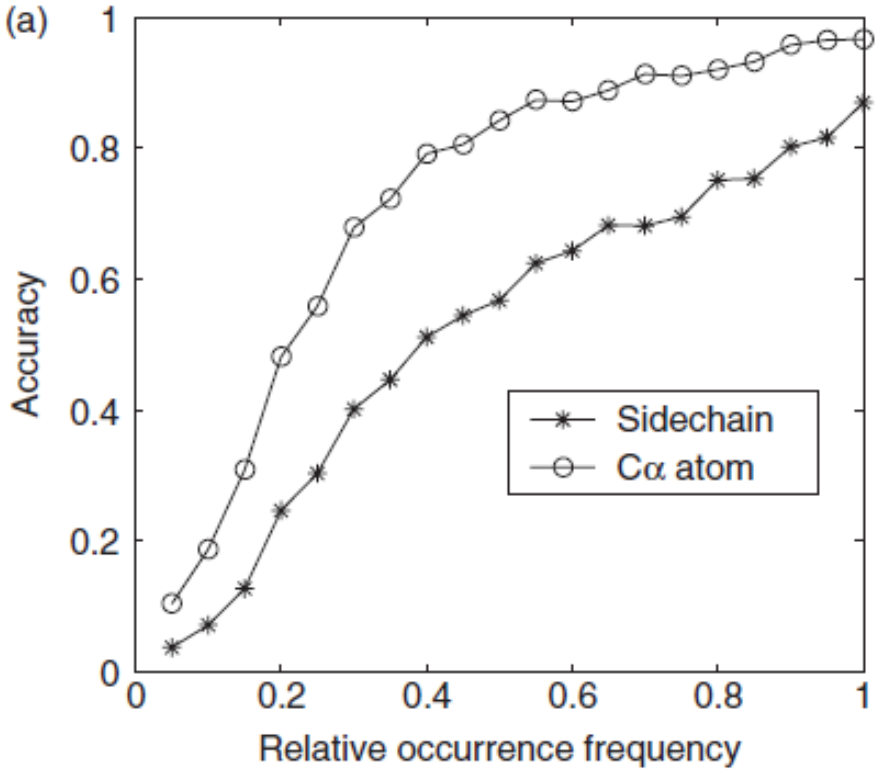
$$\text{TM-score} = \text{Max}\left[\frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{1 + \left(\frac{d_i}{d_0} \right)^2} \right]$$

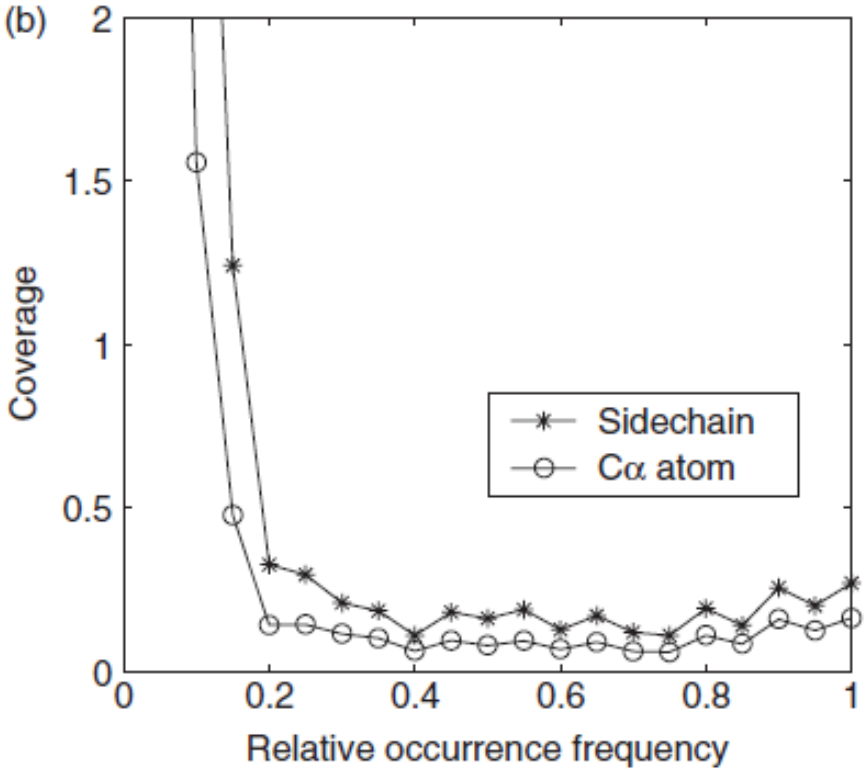
Ranked by their TM-score.

Test Dataset

- 620 non-homologous proteins (<25% sequence identity with lengths from 50 to 600) from PDBSELECT (2006 March).

Result


(a)
Accuracy
Relative occurrence frequency
Sidechain
Cα atom
0
0.2
0.4
0.6
0.8
1


(b)
Coverage
Relative occurrence frequency
Sidechain
Cα atom
0
0.5
1
1.5
2
0.2
0.4
0.6
0.8
1

Summary

- LOMETS servers is at least 7% more accurate than all the individual servers.
- The difference is also statistically meaningful with a t-test at 0.1% of significance level.
- The average CPU time for a medium size protein (200 residues) is ≤20 min when runs parallel on nine nodes of a cluster.

QUARK server

- http://zhanglab.ccmb.med.umich.edu/QUARK/

- http://zhanglab.ccmb.med.umich.edu/QUARK/Q24927/

CASP

Critical Assessment of protein Structure Prediction

http://predictioncenter.org/

http://www.predictioncenter.org/index.cgi?page=public_serv