The Rosetta Method for Protein Structure Prediction

The Rosetta Approach

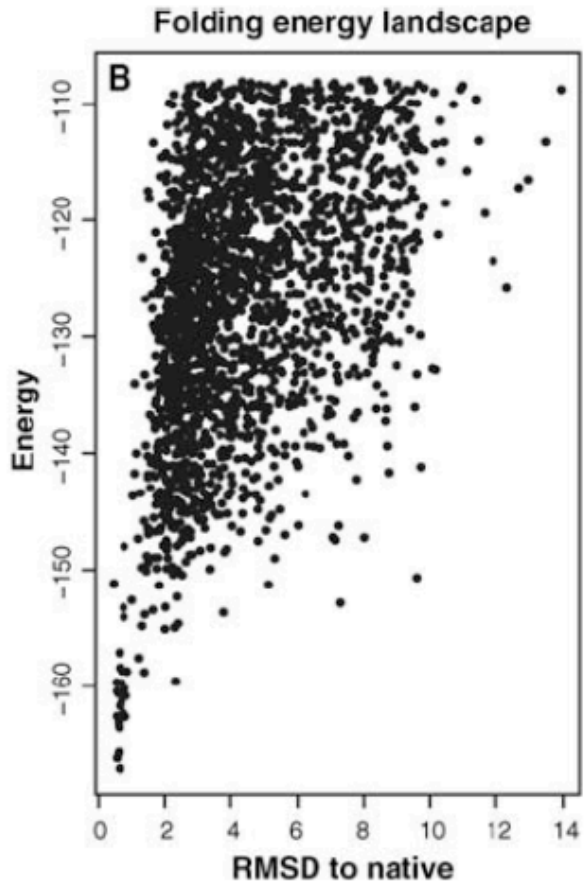
(David Baker lab, Univ. of Washington)

- In contrast to threading, Rosetta does *de novo* prediction – doesn't use templates/homologous structures

- instead performs Monte Carlo search through space of conformations to find minimal energy conformation

The Folding Energy Landscape

- energies of conformations considered in Rosetta's Monte Carlo minimization procedure for a given protein



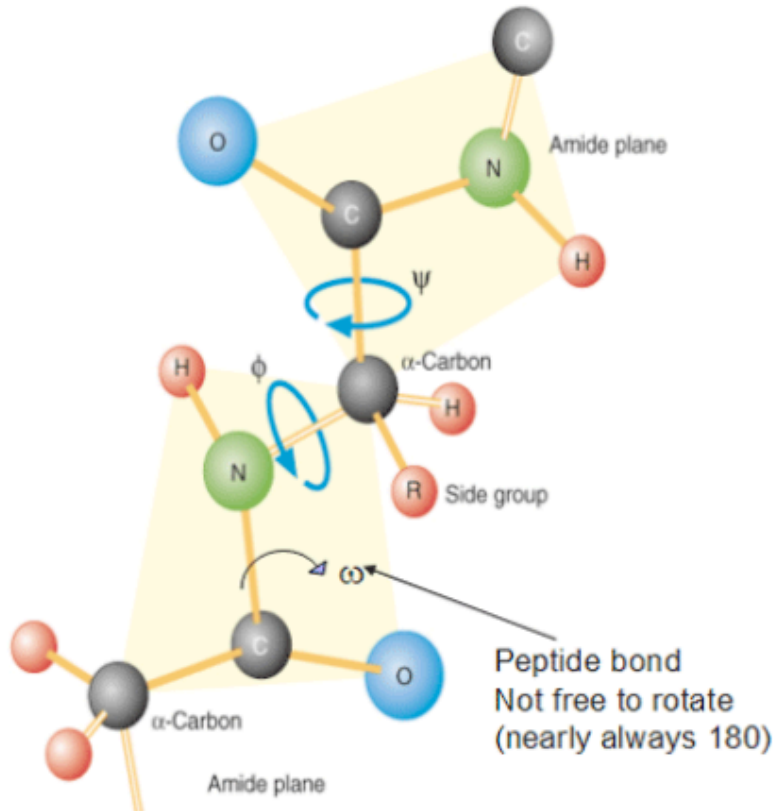
$$RMSD = \sqrt{\frac{\sum_{n} \left| x_n - \hat{x}_n \right|}{N}}$$

x_n coordinate of nth α carbon

$\hat{x}_n$ *predicted* coordinate of nth α carbon

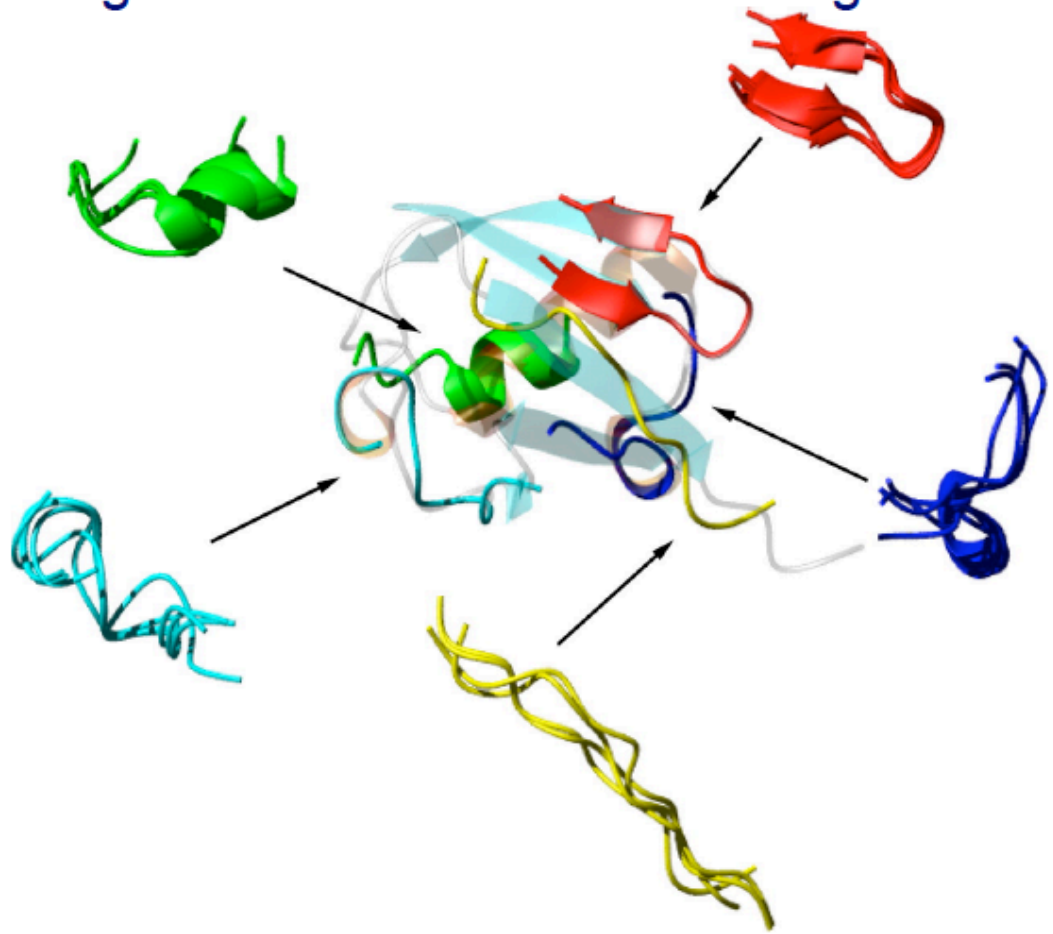
Representing Protein Structures

- the predicted structure of a protein is represented in terms of the *torsion angles* of the polypeptide backbone



Overview of the Rosetta Approach

- Rosetta searches structure space by replacing the *torsion angles* of a fragment in the current model with torsion angles from known structure fragments



The Rosetta Approach

Given: protein sequence *P*

for each window of length 9 in *P* assemble a set of structure fragments

M = initial structure model of *P* (fully extended conformation)

S = score(*M*)

while stopping criteria not met

randomly select a fixed width "window" of amino acids from *P*

randomly select a fragment from the list for this window

M' = *M* with torsion angles in window replaced by angles from fragment

S' = score(*M'*)

if Metropolis criterion(*S*, *S'*) satisfied

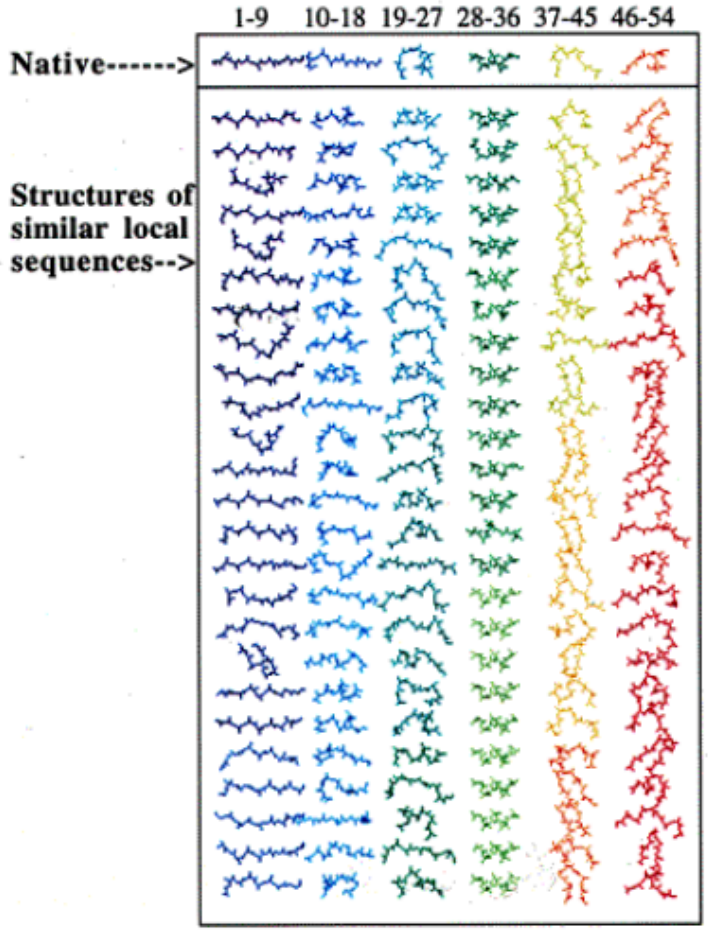
M = *M'*

S = *S'*

Return: predicted structure *M*

Fragment Selection

- fragments are selected from known structures
- the window-fragment matches are calculated using
  - PSI-BLAST to build a profile model of the sequence
  - the predicted secondary structure of the sequence



Metropolis Criterion

- given the previous structure model with score S and the new one with score S', accept the new one with probability

$$\min\left(1,\ \exp\left(-\frac{S' - S}{T}\right)\right)$$

"temperature" parameter that is varied during the search

Scoring Function Takes Into Account

- residue environment (solvation)
- residue pair interactions (electrostatics, disulfides)
- strand pairing (hydrogen bonding)
- strand arrangement into sheets
- helix-strand packing
- steric repulsion
- etc.

Some Details

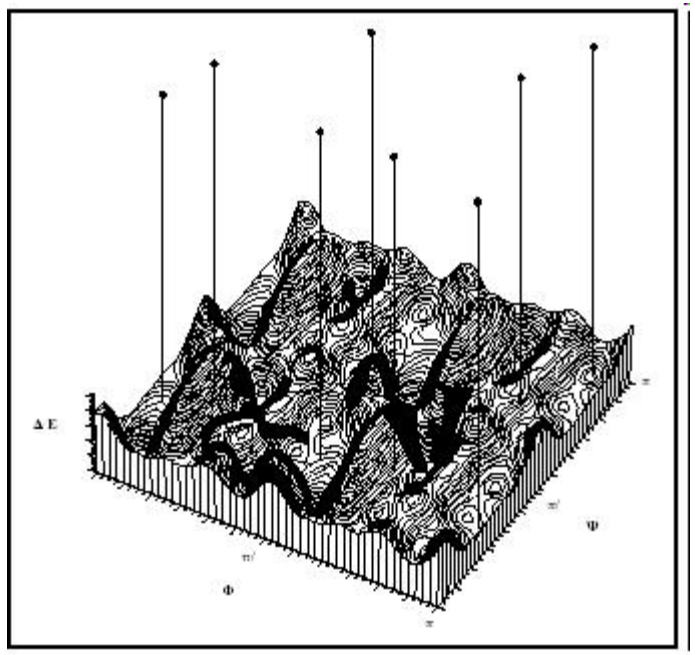
- scoring function search progressively adds terms during search
  - initially on the steric overlap term is used
  - then all but “compactness” terms are used
  - etc.
- search is initiated from different random seeds
- for some applications, an atomic-level scoring function is used

CS-ROSETTA

The ROSETTA Procedure

- Monte Carlo fragment replacement
- Monte Carlo side chain packing
- Monte Carlo minimization
- As t goes to infinity (cubed? more?), it converges to the answer!

Monte Carlo (Random Sampling)



http://www.chemistryexplained.com/images/chfa_03_img0571.jpg

- Randomly (or pseudorandomly) pick a configuration and evaluate its energy.
- If acceptably low, store result.
- If not, move a distance away from that point as a function of the energy (Metropolis criterion, a.k.a. simulated annealing) and evaluate again
- When some convergence threshold or time limit is met, stop and return stored results.

Advantages of Monte Carlo

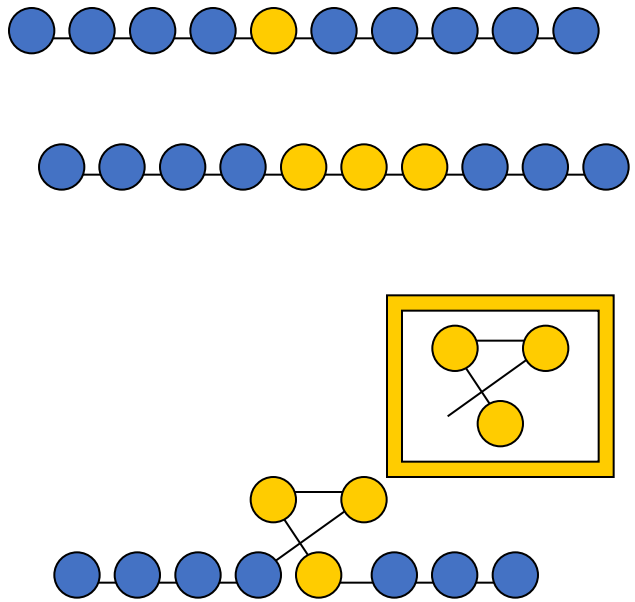
- Individual computations are cheap
  - Exponential search spaces are slow to search exhaustively
  - Probabilistic worst case is identical to simple brute-force
- Can be done as an empirical black box
  - Can approximate molecular dynamics with empirical energy functions

When Should Monte Carlo Be Used?

- No provable bounds on running time
  - Monte Carlo linear algebra?
  - Monte Carlo comparison sort? (Bozo Sort)
- No provable bounds on accuracy
  - Convergence != global minimum
- Only sample what you can’t reasonably deterministically predict

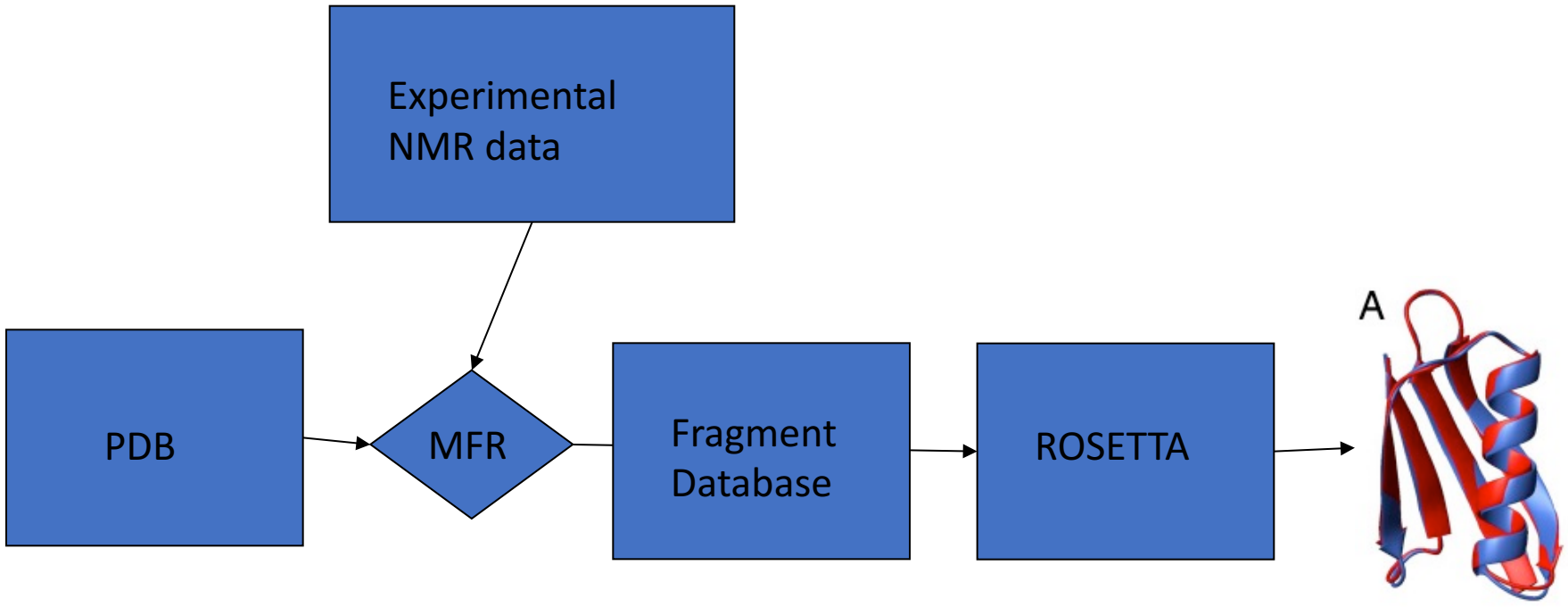
Application to ROSETTA

- Monte Carlo fragment replacement
  - Randomly select a position, and the 8 residues following it
  - Randomly select a 9 residue fragment from database, and match the fragment's bond angles



Chemical-Shift Rosetta

- Use NMR data as an additional criterion in fragment selection phase.

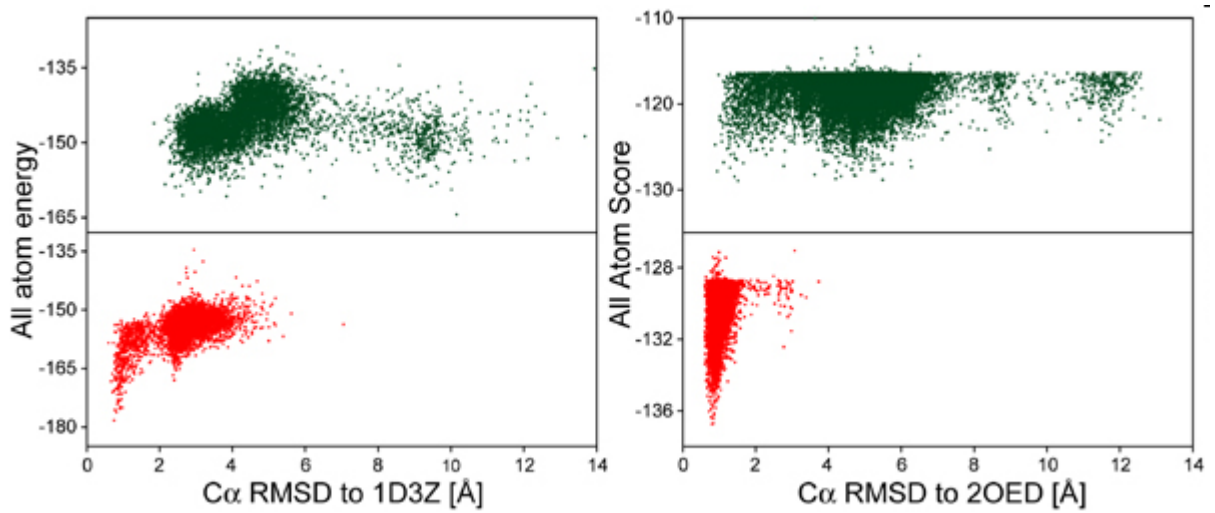


From fig. (3)

Molecular Fragment Replacement (MFR)

- Given AA sequence (from genomic data or otherwise) search PDB for best possible matches.
- Find fragments of known proteins that best match the sequence and predicted chemical shift best fit experimental data.
  - Chemical shifts predicted via SPARTA, which was trained on 200 proteins and is 10% more accurate than SHIFTX

Results



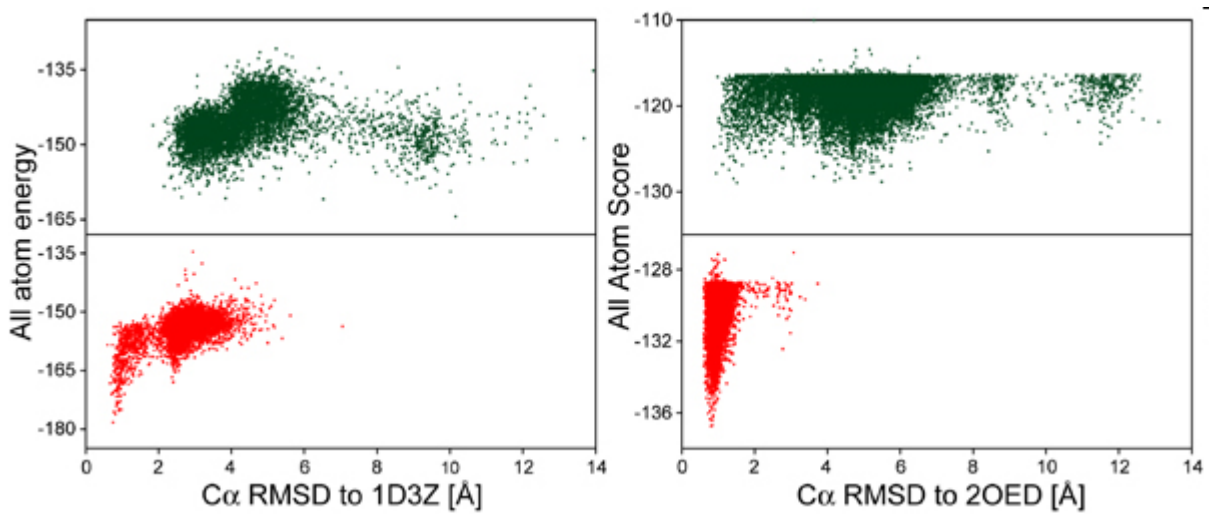
- MFR-selected fragments generate lower energy structures than standard ROSETTA fragments
- Lowest-energy conformations for C^{α} deviated 1~2 Å from reference structure
- Some exceptions, but ROSETTA doesn't consider the chemical shifts, and adding it to the empirical energy function improved results

Robustness

- When backbone chemical shift assignments are incomplete, CS-ROSETTA is still better at picking fragments than ROSETTA
- If a whole section of the protein's chemical data is missing then it's like that part is just being run with vanilla ROSETTA

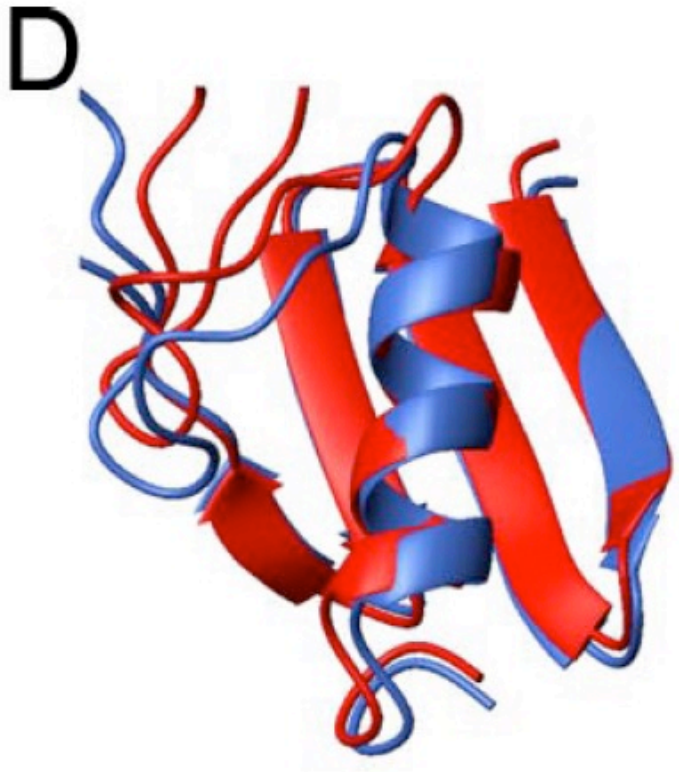
Convergence

- Convergence is concluded when the newly derived structure has rmsd approx. 2Å from the lowest energy structure so far.
- Baker et al. suggest identifying a “funneling phenomenon”



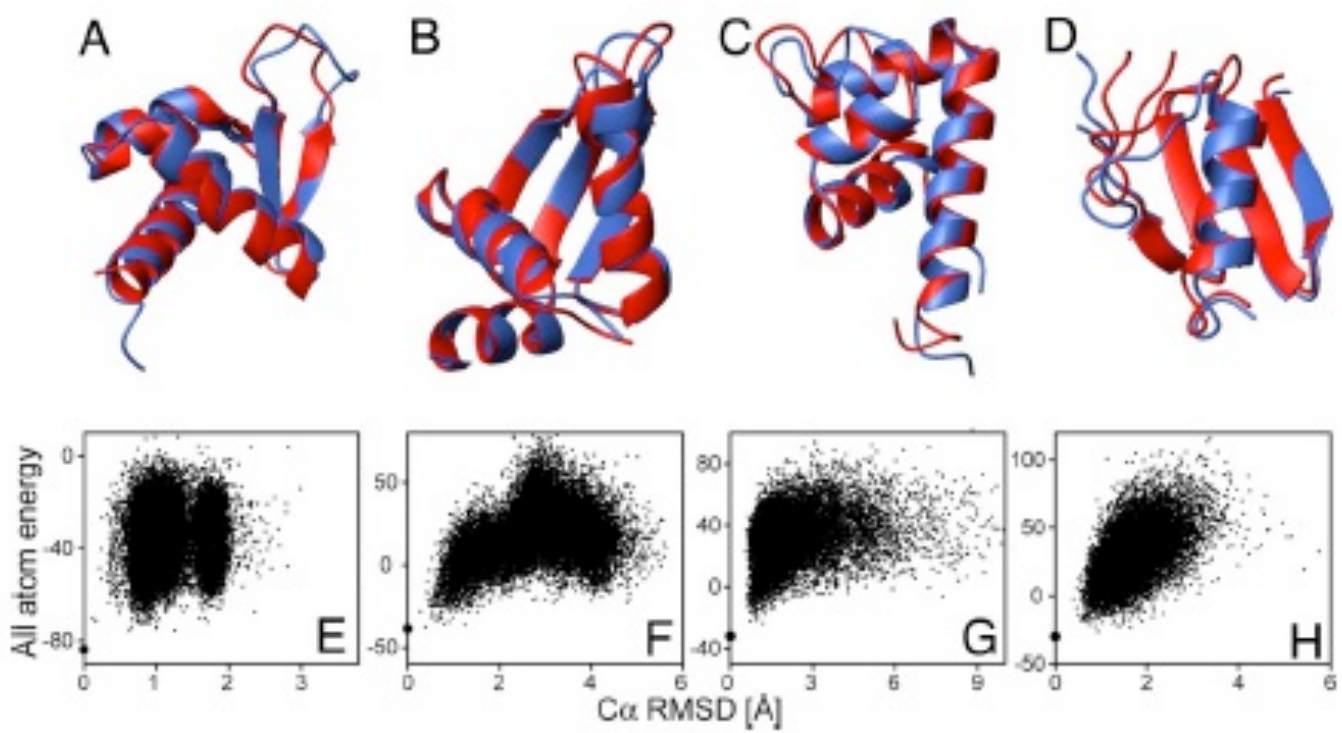
Convergence

- Convergence rapidly decreases with increasing protein size, and CS-ROSETTA begins to fail at around 130 residues.
- Convergence is also adversely affected by long, disordered loops in the reference structure



From fig(5)

Blind Prediction



- The ordered portions have remarkably good rmsd, values <1 Å for 6 and less than approx. 2 for the other 3

Blind Prediction

- Structures are strikingly similar:
  - ROSETTA's energy model favors hydrogen bonds, which results in extended secondary structure by a few residues
  - Disordered sections can be detected by chemical shifts with Random Coil Index and thus prohibited from contributing to secondary structure
  - Core side-chain packing was also less accurate

Conclusions

- CS-ROSETTA is faster and thus able to handle bigger problems than traditional ROSETTA.
- CS-ROSETTA is 50% faster than traditional triple-NMR structure determination
- CS-ROSETTA is perhaps better able to determine the structure of systems not stable enough for conventional NMR…?

CS-ROSETTA?

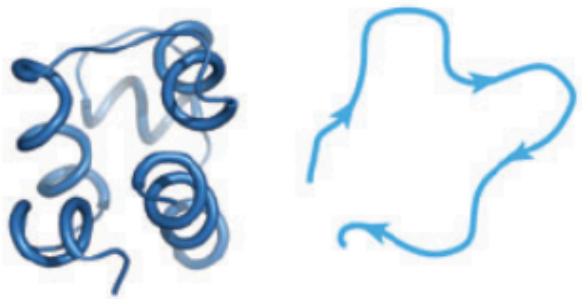
- Is there a mathematically derived limit on how big a protein can be?
  - ROSETTA runs 28,000 iterations, so if the search space of a protein exceeds 28000*n* for some *n* it is most likely going to fail?
- Each additional sample gives us more information. Is it possible to identify the “statistically significant global minimum?”
- Given assignments, Chemical shifts should also tell us more about secondary structure (guided side chain packing and minimization?)

TABLE I
COMPONENTS OF ROSETTA ENERGY FUNCTION[a]

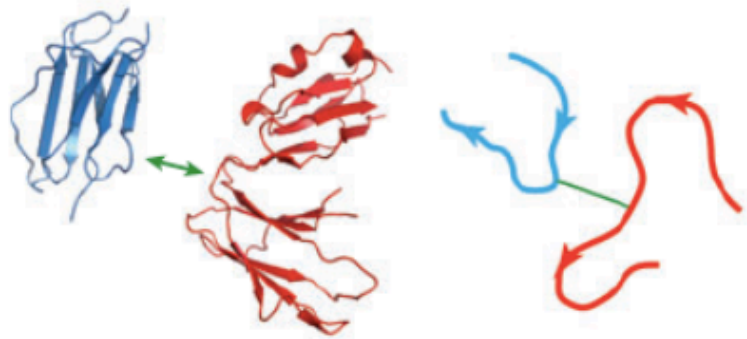
Name	Description (putative physical origin)	Functional form	Parameters (values)
env[b]	Residue environment (solvation)	$\sum_i -\ln\left[P(\mathrm{aa}_i\|\mathrm{nb}_i)\right]$	i = residue index aa = amino acid type nb = number of neighboring residues[c] (0, 1, 2. . . 30, >30)
pair[b]	Residue pair interactions (electrostatics, disulfides)	$\sum_i \sum_{j>i} -\ln\left[\frac{P(\mathrm{aa}_i, \mathrm{aa}_j\|s_{ij}d_{ij})}{P(\mathrm{aa}_i\|s_{ij}d_{ij})P(\mathrm{aa}_i\|s_{ij}d_{ij})}\right]$	i, j = residue indices aa = amino acid type d = centroid–centroid distance (10–12, 7.5–10, 5–7.5, <5 Å) s = sequence separation (>8 residues)
SS[d]	Strand pairing (hydrogen bonding)	$\mathrm{SchemeA}: \mathrm{SS}_{\phi,\theta} + \mathrm{SS}_{hb} + \mathrm{SS}_d$ $\mathrm{SchemeB}: \mathrm{SS}_{\phi,\theta} + \mathrm{SS}_{hb} + \mathrm{SS}_{d\sigma}$ where $\mathrm{SS}_{\phi,\theta} = \sum_m \sum_{n>m} -\ln\left[P(\phi_{mn}, \theta_{mn}\|d_{mn}, \mathrm{sp}_{mn}, s_{mn})\right]$ $\mathrm{SS}_{hb} = \sum_m \sum_{n>m} -\ln\left[P(\mathrm{hb}_{mn}\|d_{mn}, s_{mn})\right]$ $\mathrm{SS}_d = \sum_m \sum_{n>m} -\ln\left[P(d_{mn}\|s_{mn})\right]$ $\mathrm{SS}_{d\sigma} = \sum_m \sum_{n>m} -\ln\left[\mathrm{P}(d_{mn}\sigma_{mn}\|\rho_m, \rho_n)\right]$	m, n = strand dimer indices; dimer is two consecutive strand residues V = vector between first N atom and last C atom of dimer m = unit vector between $\hat{V}_m$ and $\hat{V}_n$ midpoints x = unit vector along carbon–oxygen bond of first dimer residue y = unit vector along oxygen–carbon bond of second dimer residue ϕ, θ = polar angles between $\hat{V}_m$ and $\hat{V}_n$ (36° bins) hb = dimer twist, $\sum_{k=m,n} 0.5(\lvert\hat{\boldsymbol{m}}\cdot\hat{\boldsymbol{x}}_k\rvert + \lvert\hat{\boldsymbol{m}}\cdot\hat{\boldsymbol{y}}_k\rvert)$ (< 0.33, 0.33–0.66, 0.66–1.0, 1.0–1.33, 1.33–1.6, 1.6–1.8, 1.8–2.0) d = distance between $\hat{V}_m$ and $\hat{V}_n$ midpoints (< 6.5 Å) σ = angle between $\hat{V}_m$ and $\hat{M}$ (18° bins) sp = sequence separation between dimer-containing strands (< 2, 2–10, > 10 residues) s = sequence separation between dimers (>5 or >10) ρ = mean angle between vectors $\hat{m}$, $\hat{x}$ and $\hat{m}$, $\hat{y}$ (180° bins)

Applications of the Rosetta Approach

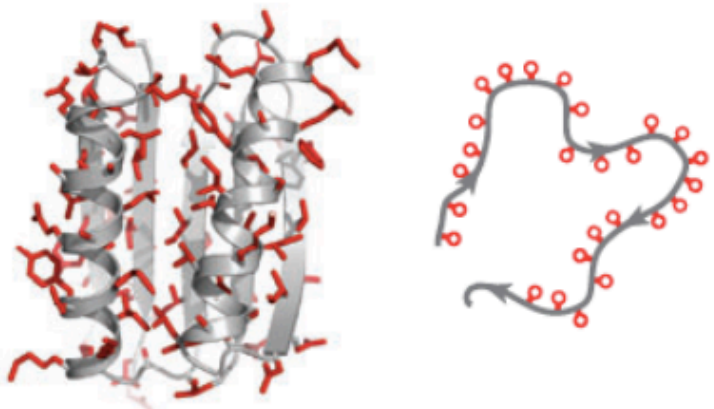
a Protein structure prediction



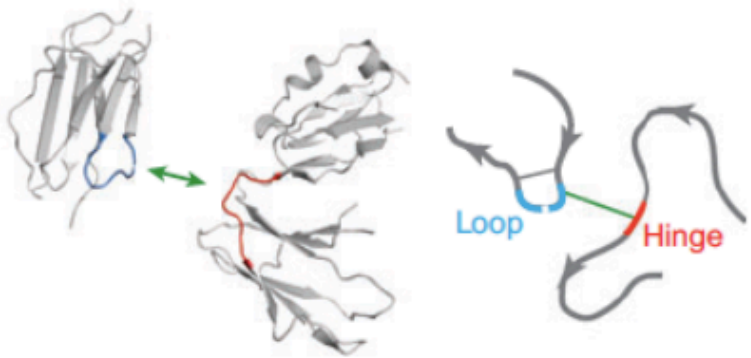
c Protein docking (fully flexible)



i Protein design

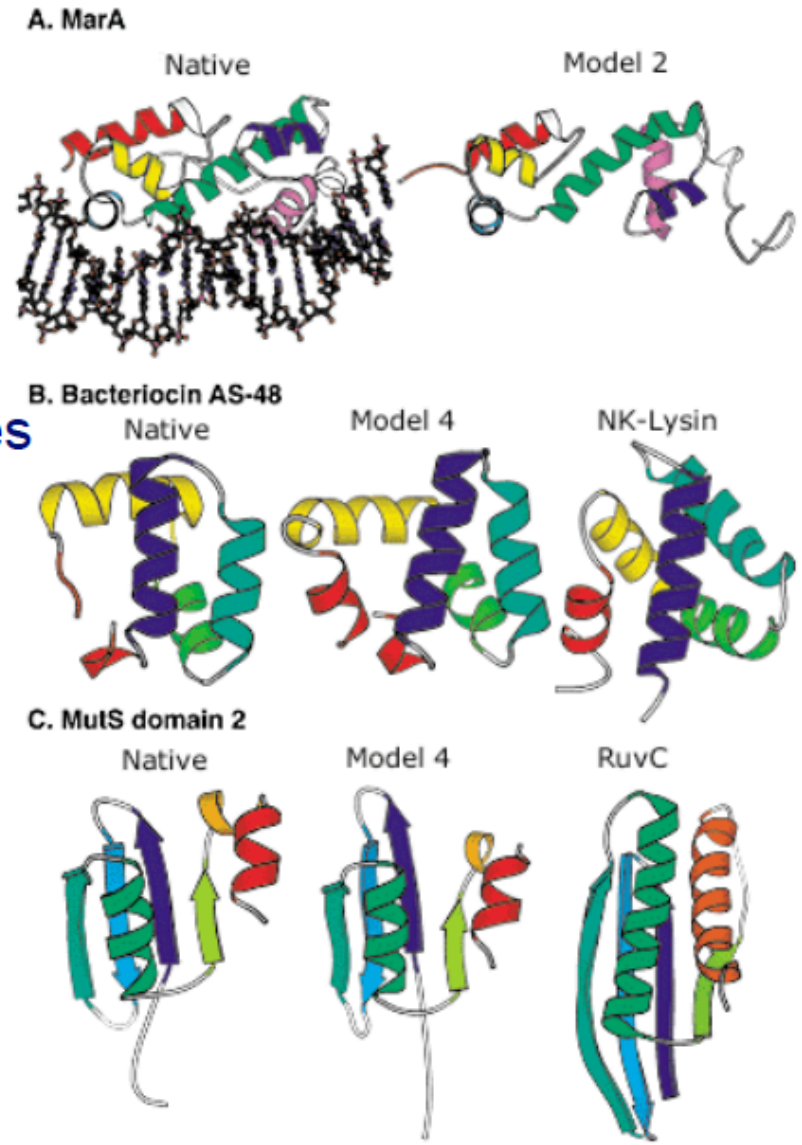


d Protein docking (partly flexible)



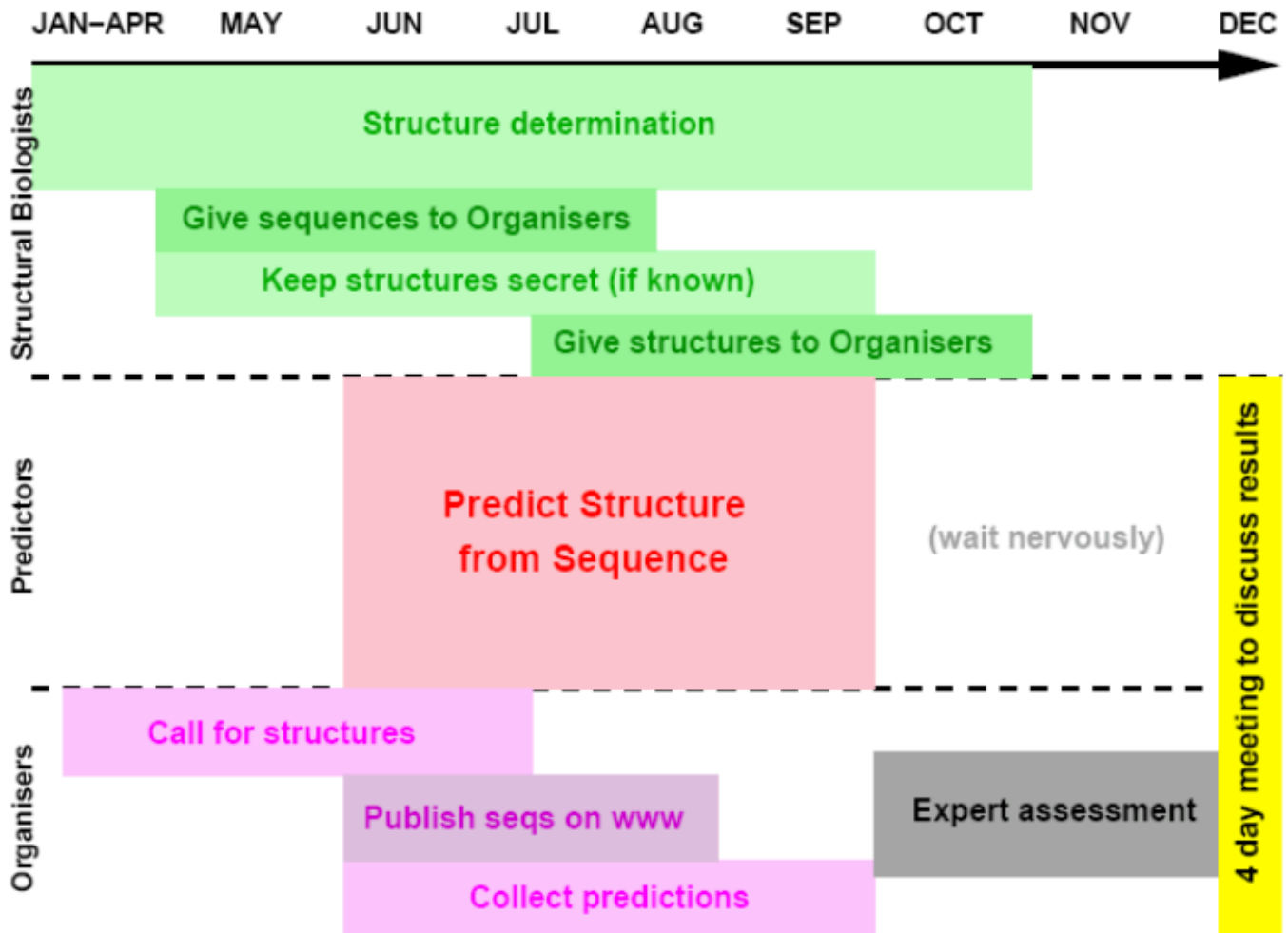
Some Rosetta-Predicted Structures

- *Native* indicates the real structure
- *Model* indicates the predicted structure
- the rightmost structures in cases B. and C. show similar structures identified by searching a structure database with the model



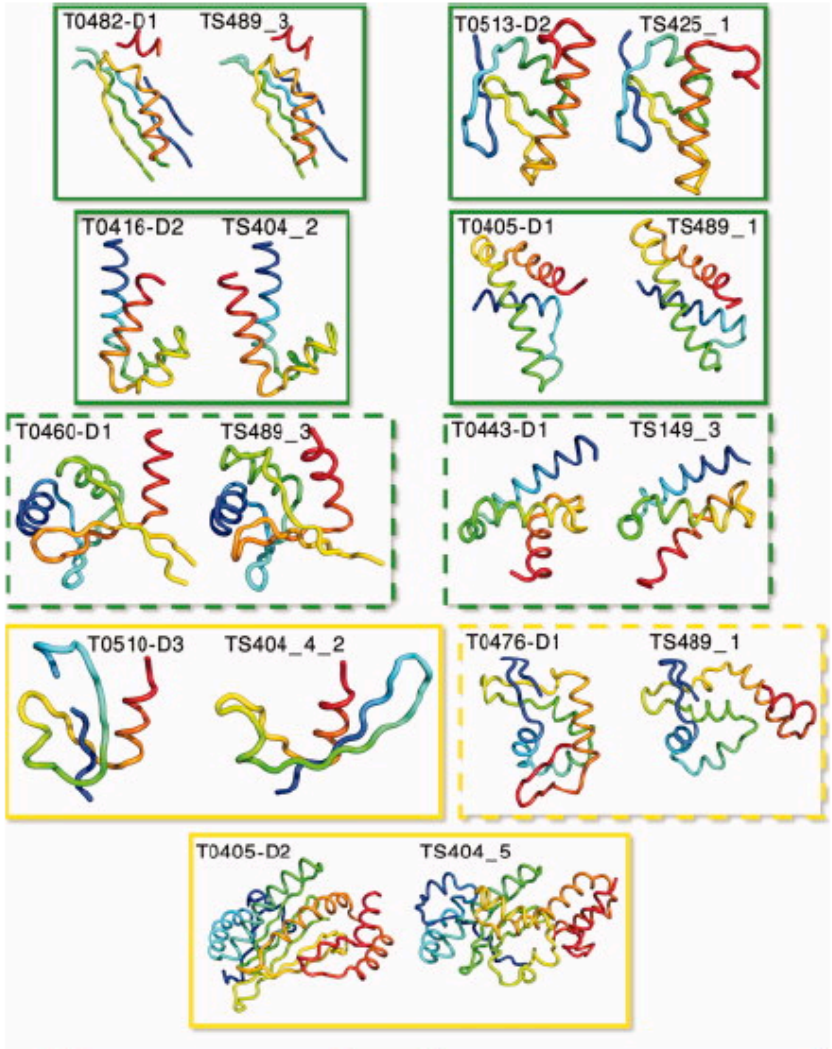
CASP

(Critical Assessment of Protein Structure Prediction)

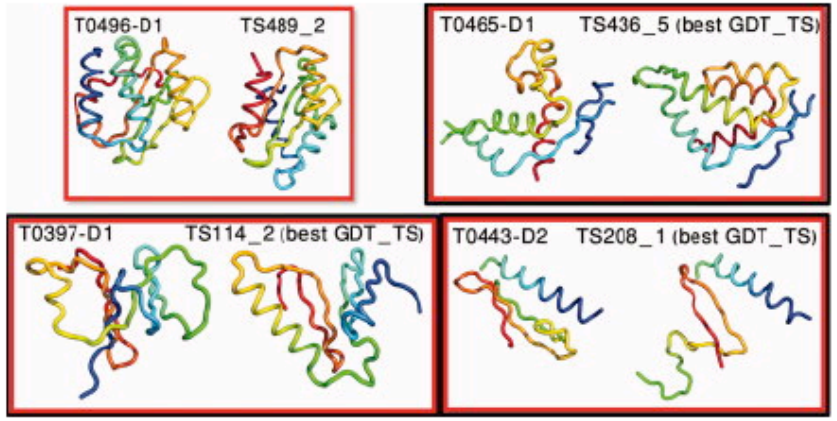


CASP 8 Best Models for *New Folds* Targets

excellent models



poor models



fair models

CASP8 *New Folds* Results

Table IV
Number of Best Models by Group

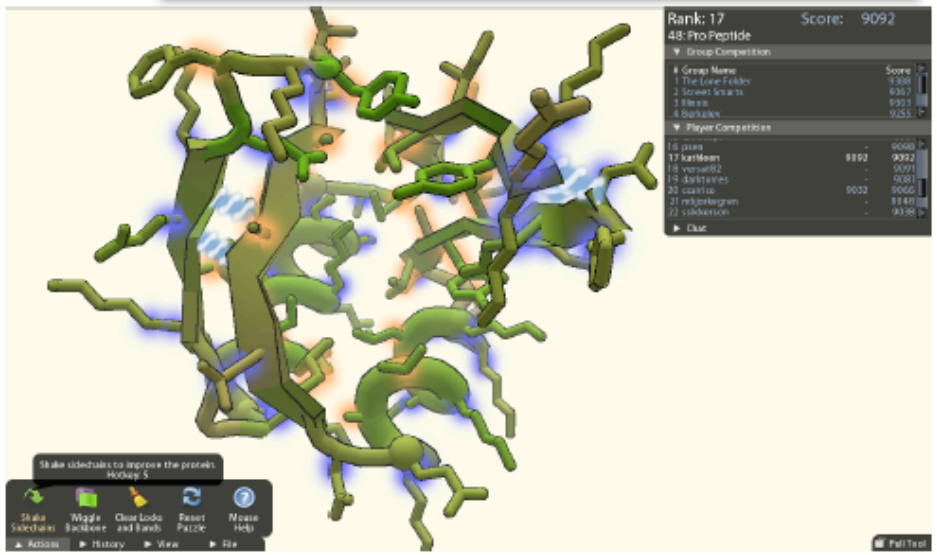
Group	Number of BEST models
DBAKER	5
MUFOLD-MD (s)[a]	3
BAKER-ROBETTA (s)	1
Keasar	1
A-TASSER	1
MidWayFolding	1
GS-KudlatyPred (s)	1
MULTICOM	1
Pcons_dot_net (s)	1
Zico	1
ZicoFullSTP	1
ZicoFullSTPFullData	1

[a](s) indicates Server.

Want to Help Predict Structures?

- Rosetta@home
  http://bioinc.bakerlab.org/

- Foldit
  http://fold.it/portal/info/science



How Big is an Angstrom?