The Rosetta Method for Protein Structure Prediction

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Mark Craven
craven@biostat.wisc.edu
Spring 2010

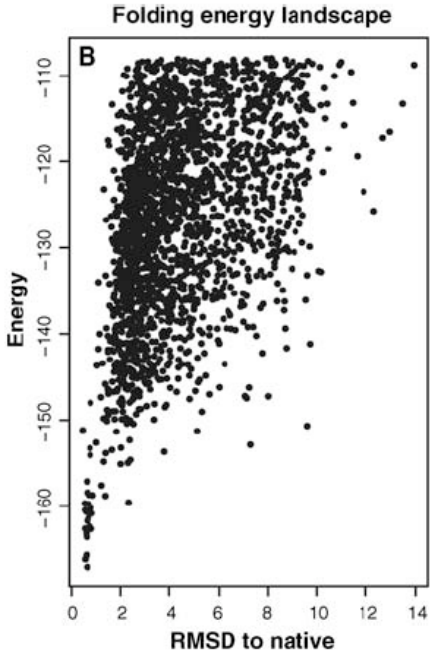
The Rosetta Approach

(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



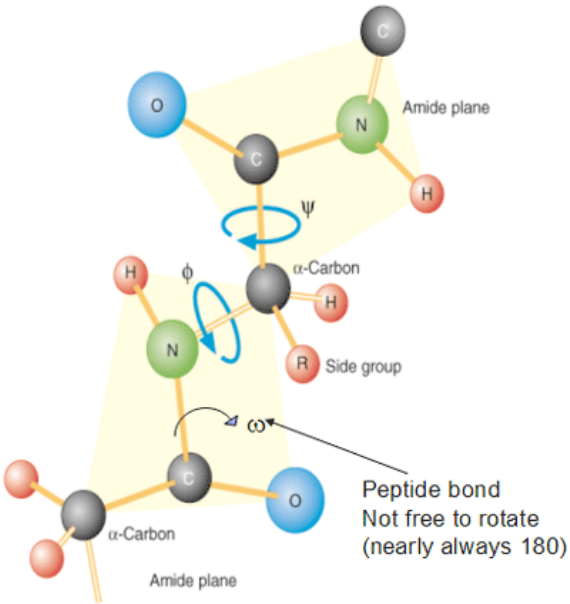
$$RMSD = \sqrt{\frac{\sum_n |x_n - \hat{x}_n|}{N}}$$

x_n coordinate of nth α carbon

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

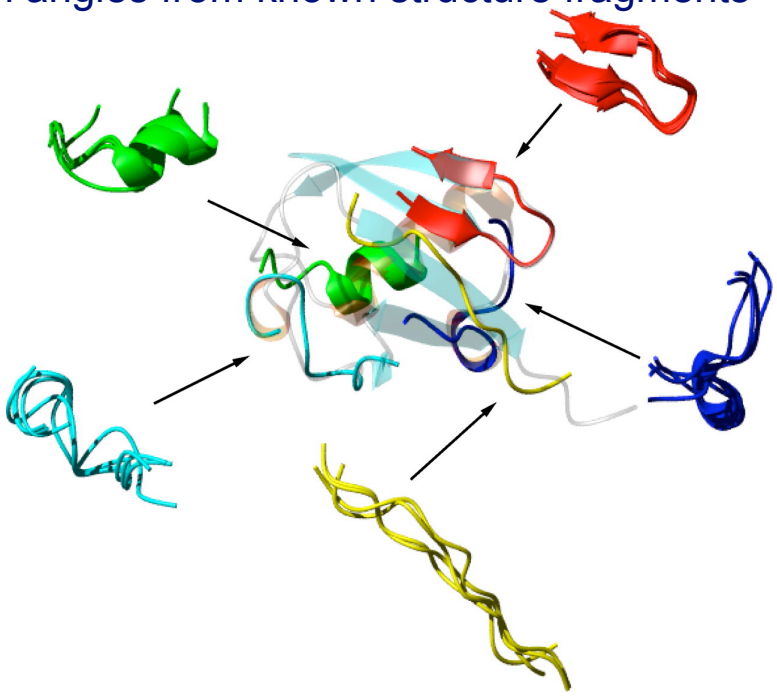
Representing Protein Structures

- 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

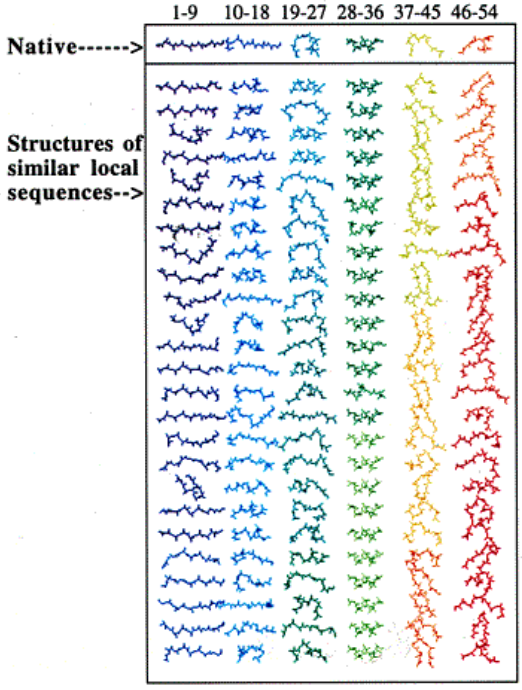
M = *M*'

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

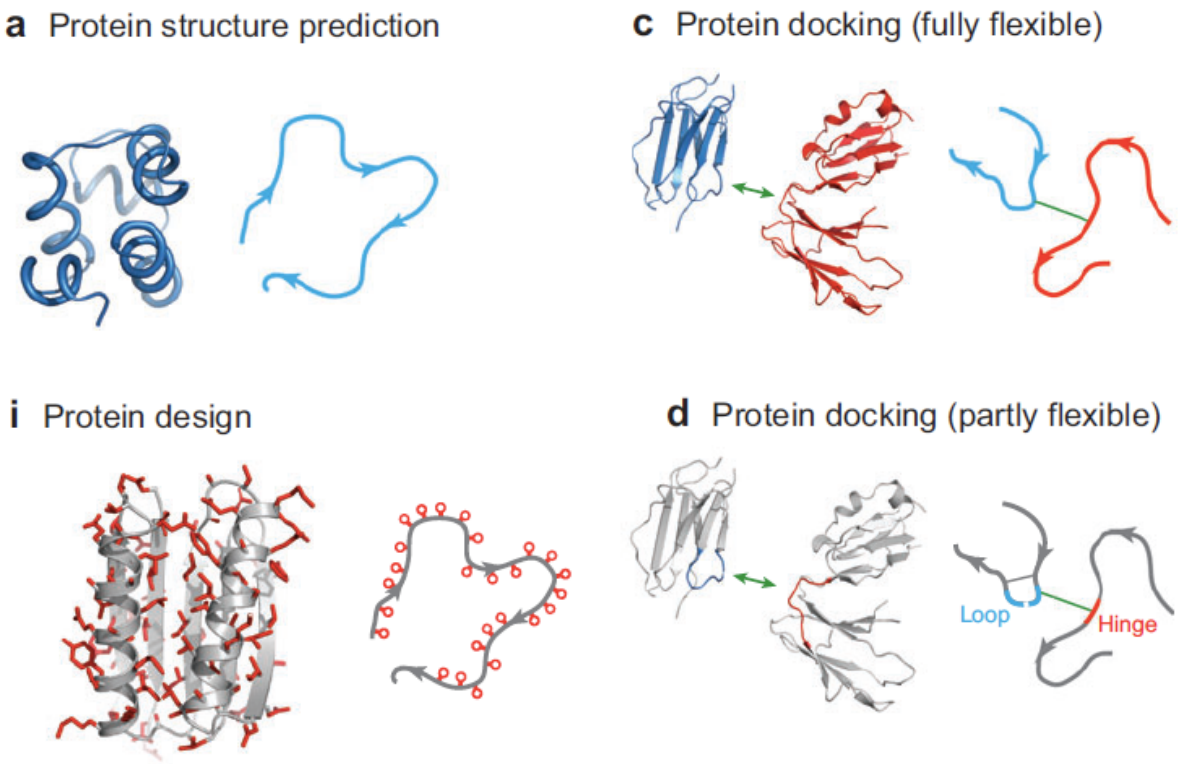
Scoring Function Takes Into Account

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

Some Details

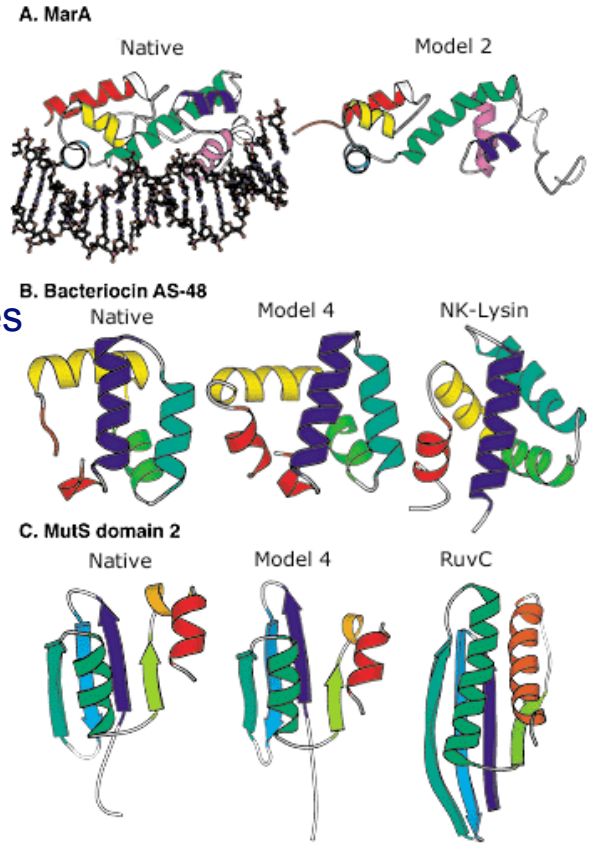
- scoring function search progressively adds terms during search
  - initially on the steric overlap term is used
  - then all but “compactness” terms are used
  - etc.
- search is initiated from different random seeds
- for some applications, an atomic-level scoring function is used

Applications of the Rosetta Approach



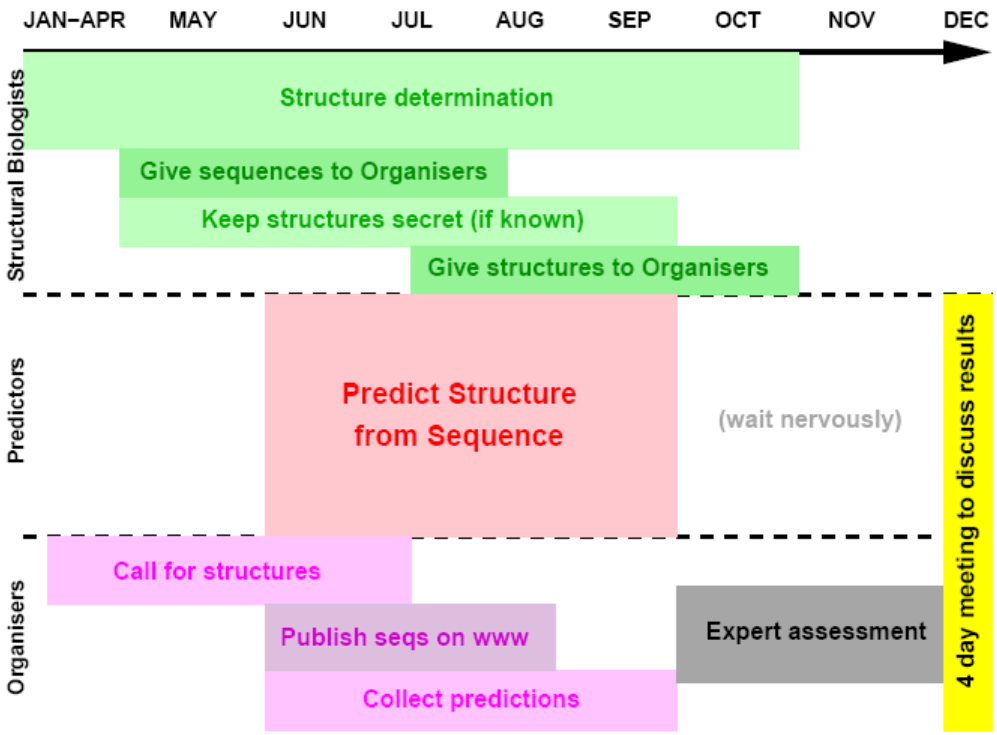
Some Rosetta-Predicted Structures

- *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



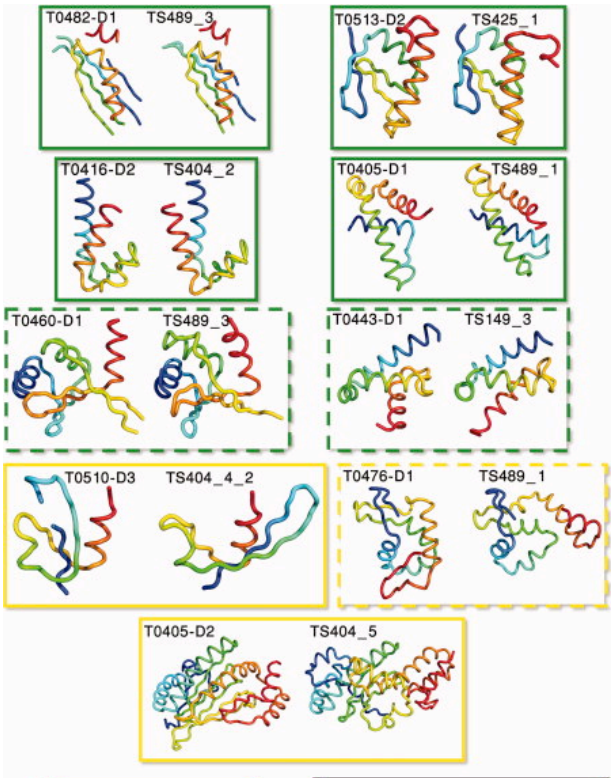
CASP

(Critical Assessment of Protein Structure Prediction)

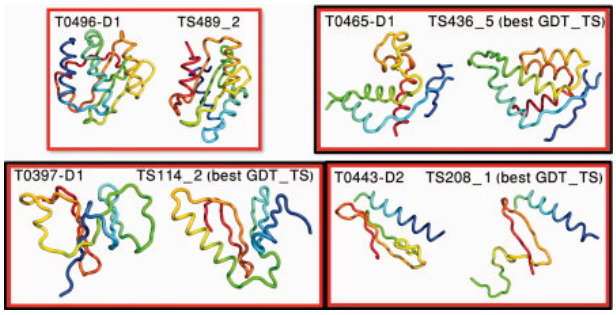


CASP 8 Best Models for *New Folds* Targets

excellent models



poor models



fair models

CASP8 *New Folds* Results

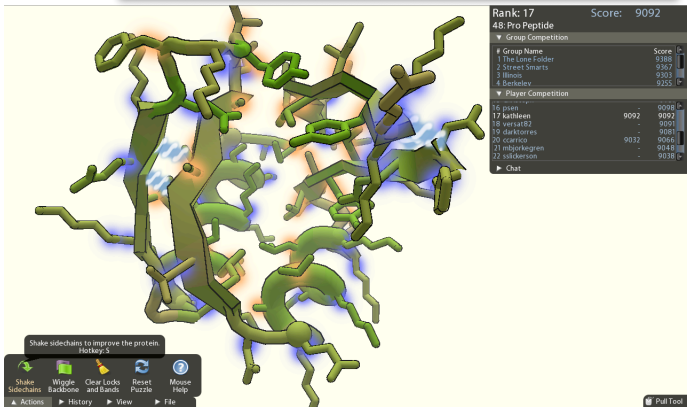
Table IV
Number of Best Models by Group

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

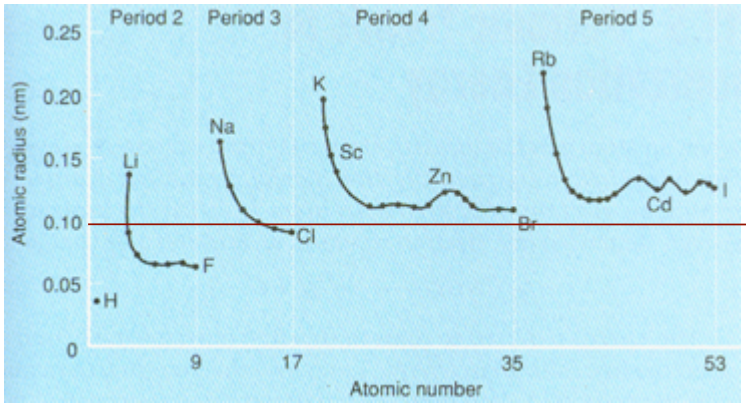
[a](s) indicates Server.

Want to Help Predict Structures?

- Rosetta@home
  http://bioinc.bakerlab.org/
- Foldit
  http://fold.it/portal/info/science



How Big is an Angstrom?



1 angstrom