Structural Bioinformatics

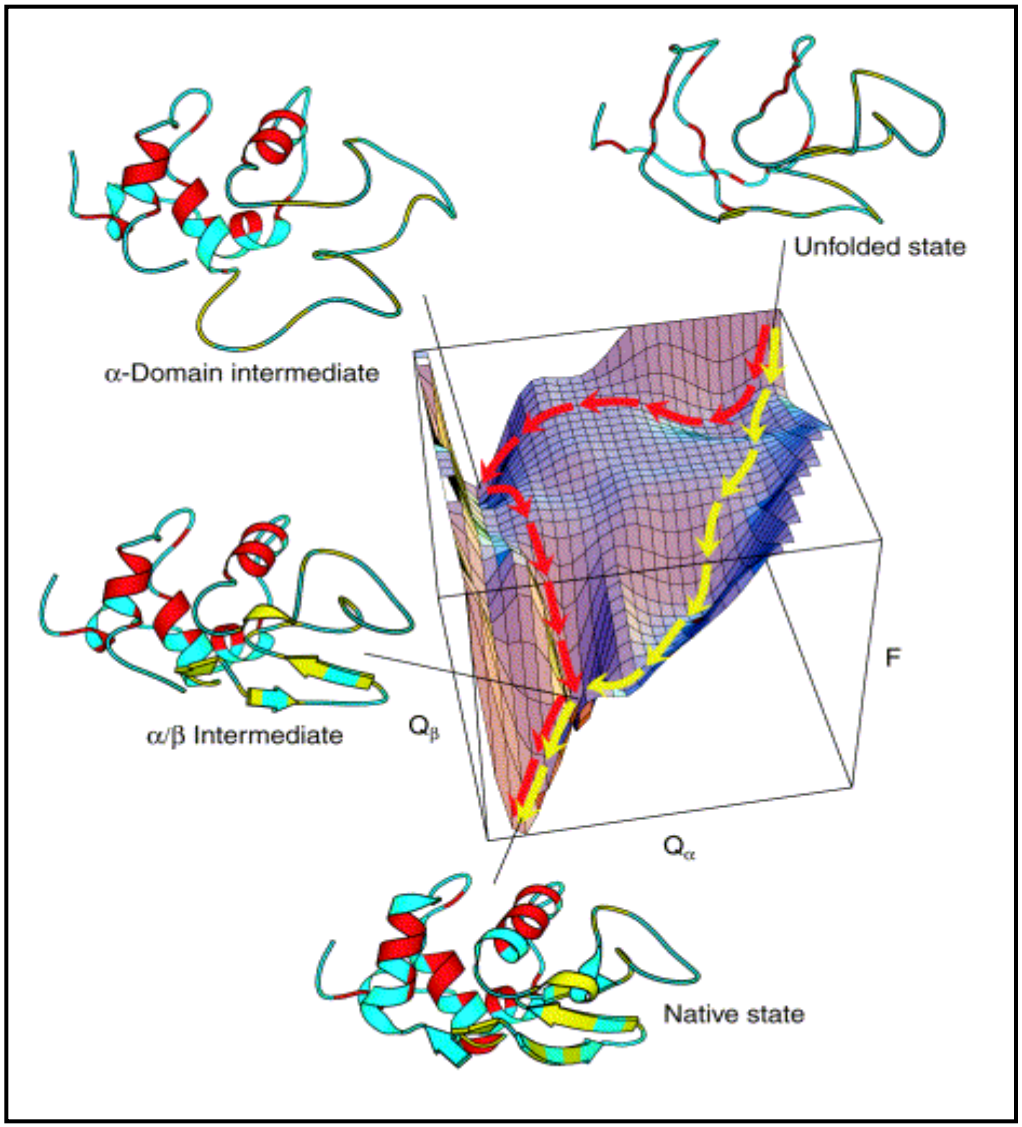
GENOME 541

Spring 2020

Lecture 3: Protein Structure Prediction

Frank DiMaio (dimaio@uw.edu)

Last lecture

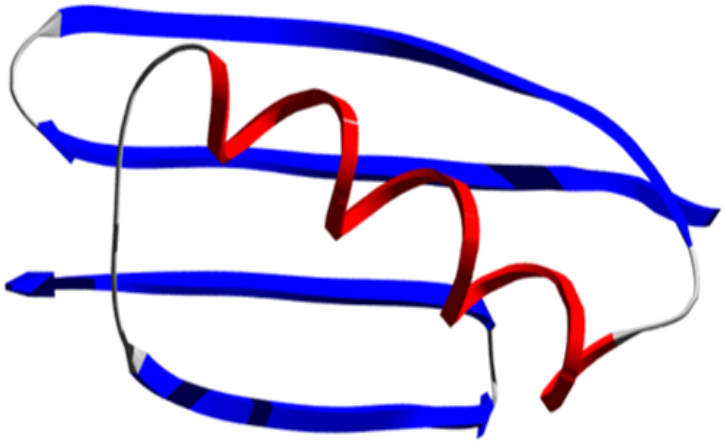


Typically, proteins fold by progressive formation of native-like structures.

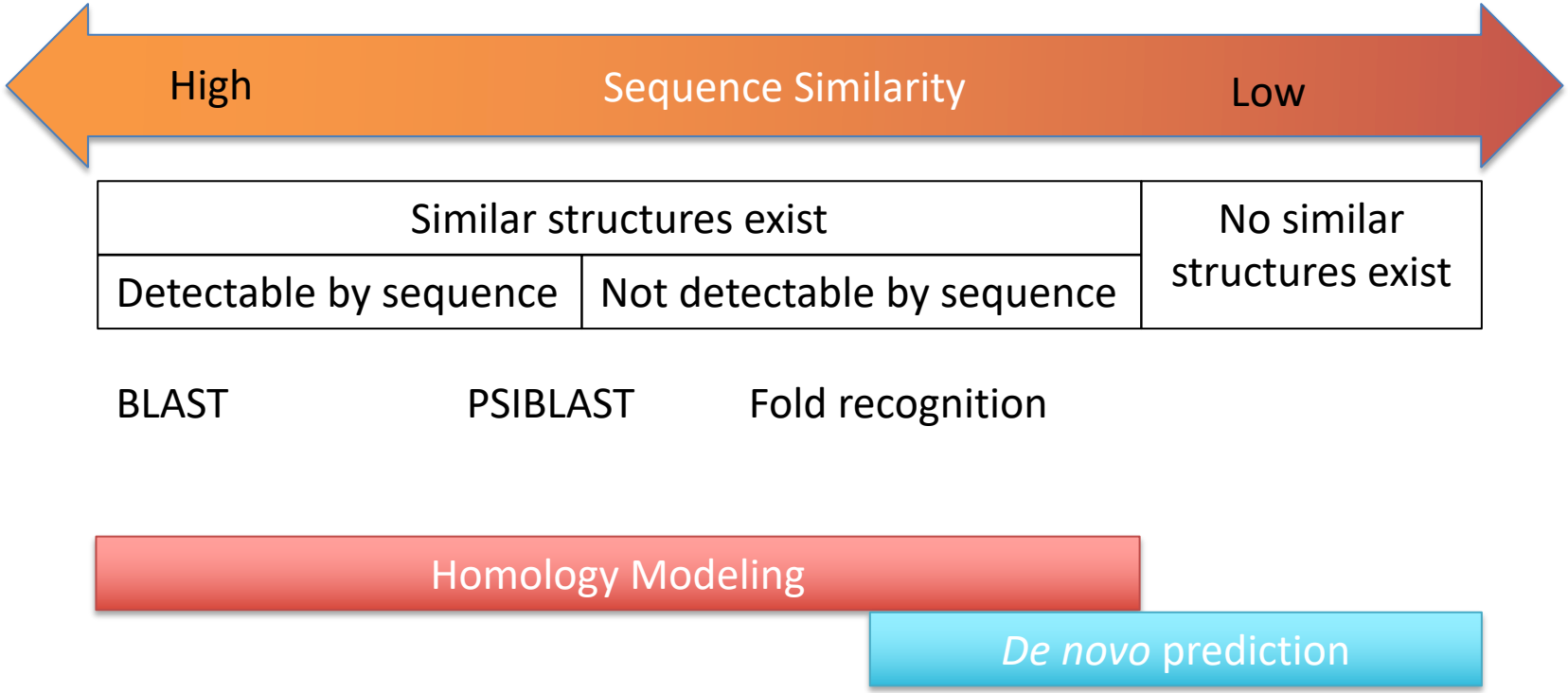
Folding energy surface is highly connected with many different routes to final folded state.

Structure Prediction

DEIVKMSPIIRFYSSGNAGLRTYIGDHKSCVMCTYWQNLLTYESGILLPQRSRTSR



Targets and Methods



Prediction Strategies

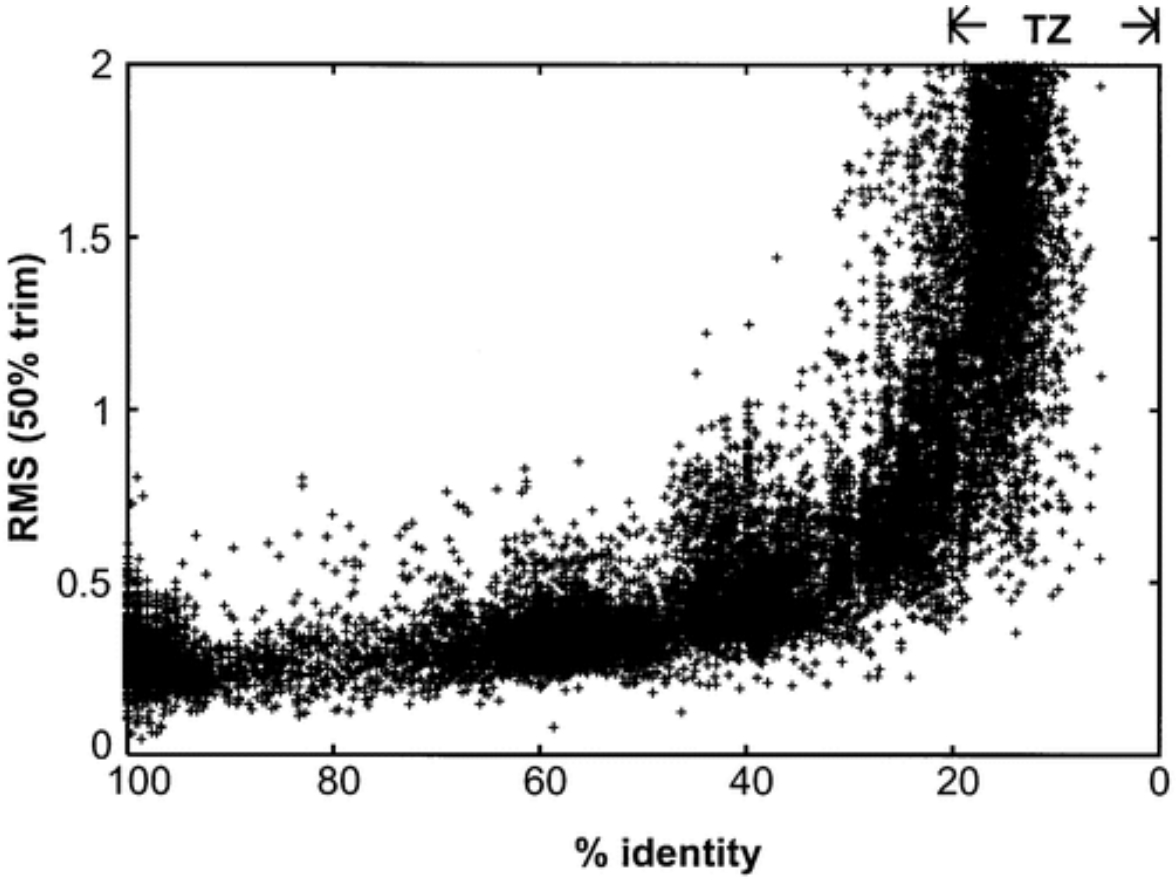
Homology Modeling

- Proteins that share similar sequences share similar folds.
- Use known structures as the starting point for model building.
- Can not be used to predict structure of new folds.

De Novo Structure Prediction

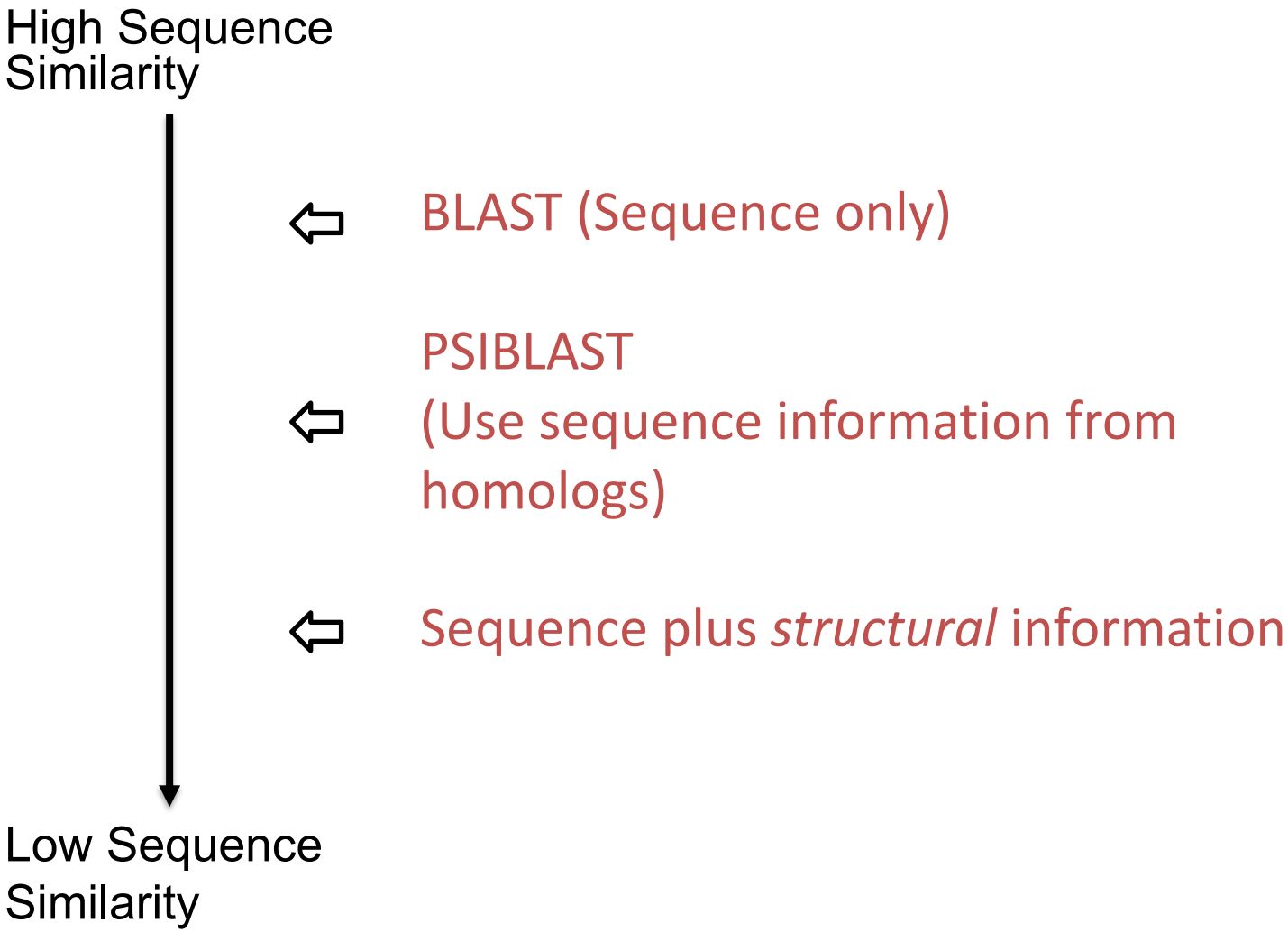
- Do not rely on global similarity with proteins of known structure
- Folds the protein from the unfolded state.
- Very difficult problem, search space is gigantic

Similar Sequences Share Similar Structures



Wilson, Kreychman, Gerstein (2000)

Fold Recognition as a Function of Sequence Similarity



BLAST (Basic Local Alignment Search Tool)

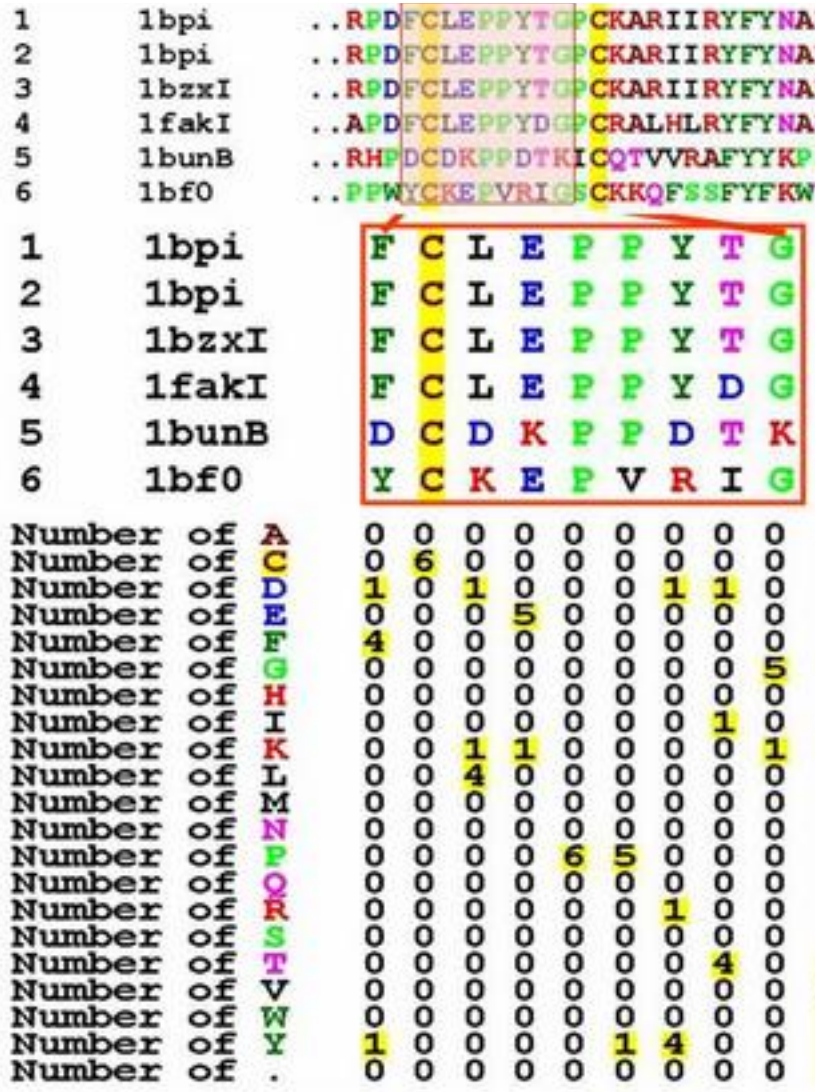
BLAST is a fast sequence alignment algorithm that identifies high-scoring local alignments by finding short exact matches (seeds) and extending outward. BLAST uses the BLOSUM62 aa substitution matrix by default.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

PSI-BLAST

- Position-Specific Iterated BLAST
- Allows more distantly related sequences to be identified
- Steps
  1. Use BLAST to identify related sequences
  2. Create a profile from related sequences
  3. Search for related sequences using this profile

Sequence Profile

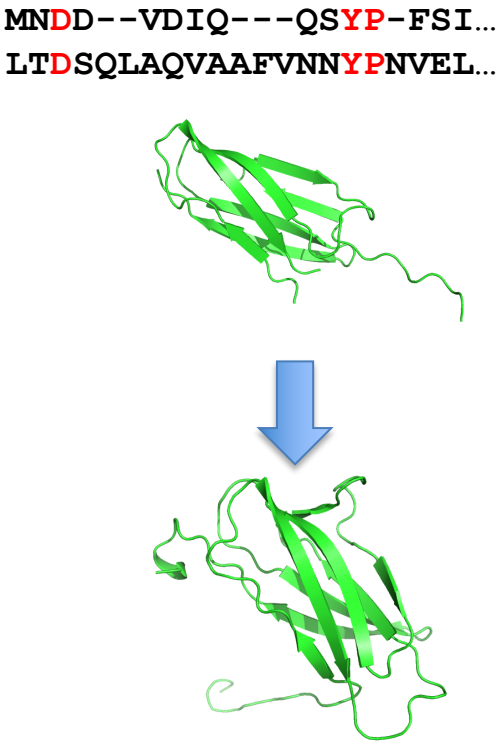


- For each column in a MSA count how often each amino acid occurs
- Combine with prior information about substitution frequencies (ie. BLOSUM62)
- Convert counts to log odds scores. End product is a Position-Specific Scoring Matrix (PSSM)

Homology Modeling

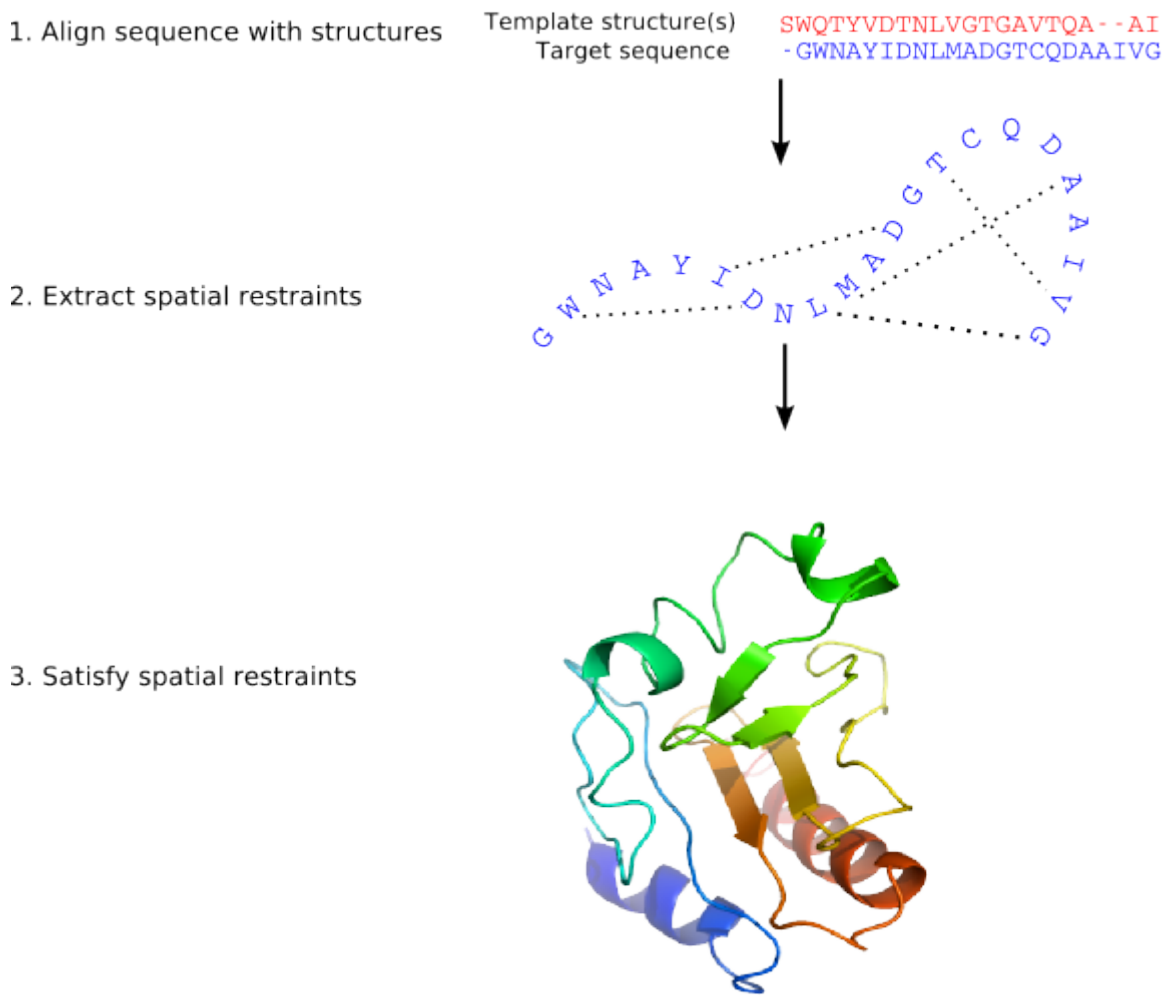
- Identify homologous protein sequences

- Build model by
  1. “Threading” residues in corresponding positions of homologous structure
  2. Sampling conformations of unaligned residues
  3. All-atom refinement



Homology Modeling

Alternately, extract conserved atom-pair distances and fold de novo



Sali and Blundell 1993

Fold recognition

Target Sequence

ALKKGF…HFDTSE

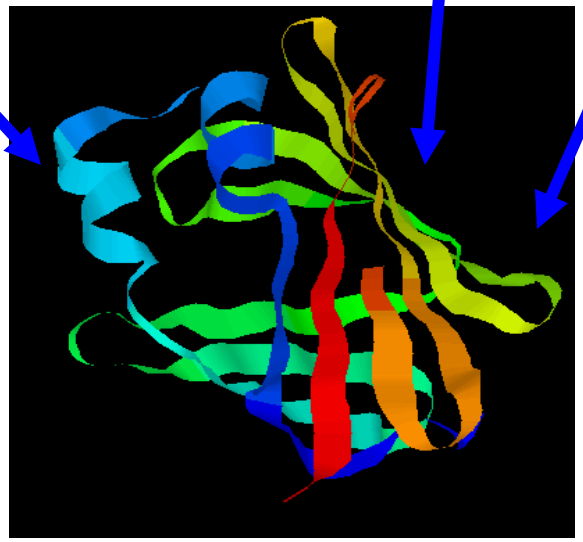
Structure Templates

1. Align target **sequence** with template **structure** from the Protein Data Bank (PDB)
2. Calculate energy score to evaluate goodness of fit between target sequence & template structure
3. Rank models based on energy scores

Fold recognition

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

What is probability that two specific residues are in contact?



How well does a specific residue fit structural environment?

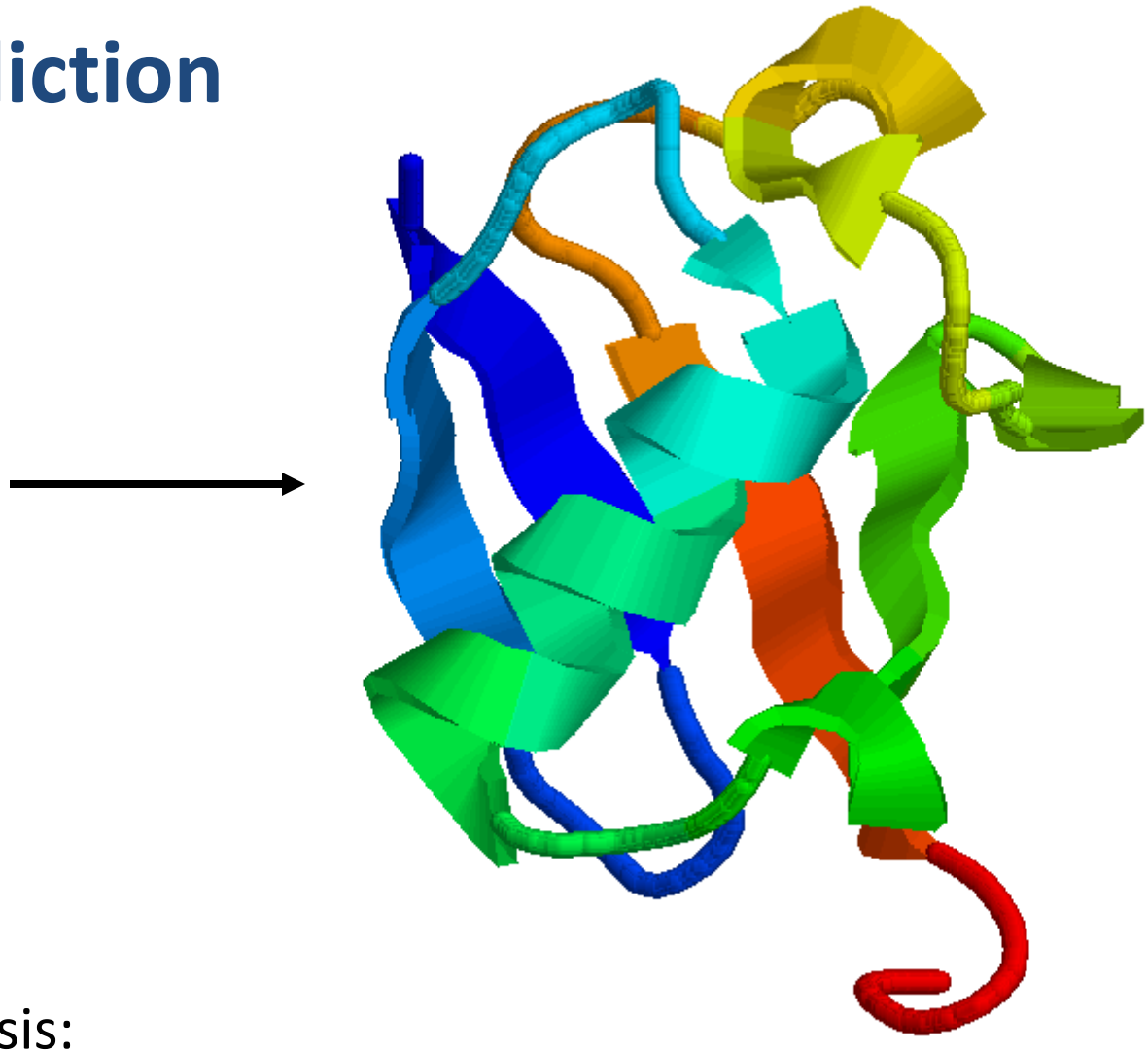
Alignment gap penalty?

Total energy: $E_p + E_s + E_g$

Find a sequence-structure alignment that minimizes the energy function

De novo protein structure prediction

```
MQIFVKTLTGKTIT
LEVEPSDTIENVKA
KIQDKEGIPPDQQR
LIFAGKQLEDGRTL
SDYNIQKESTLHLV
LRLRGG
```

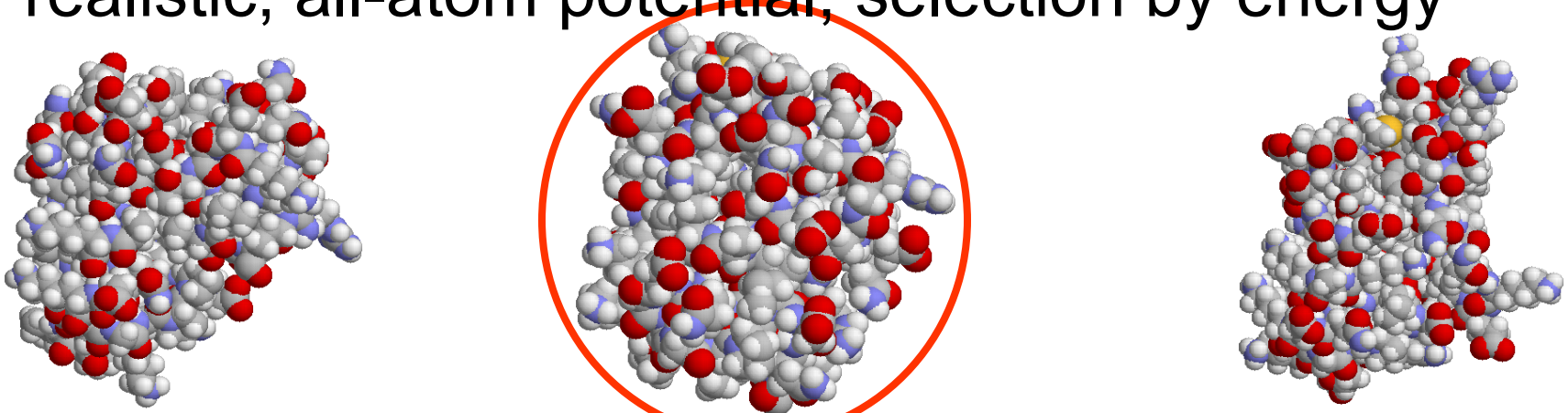


Thermodynamic hypothesis:
The native state is the lowest-energy conformation.

Structure Prediction Protocol

- Large-scale search of conformational space using a low-resolution potential

- Refinement of candidate models in a physically realistic, all-atom potential; selection by energy



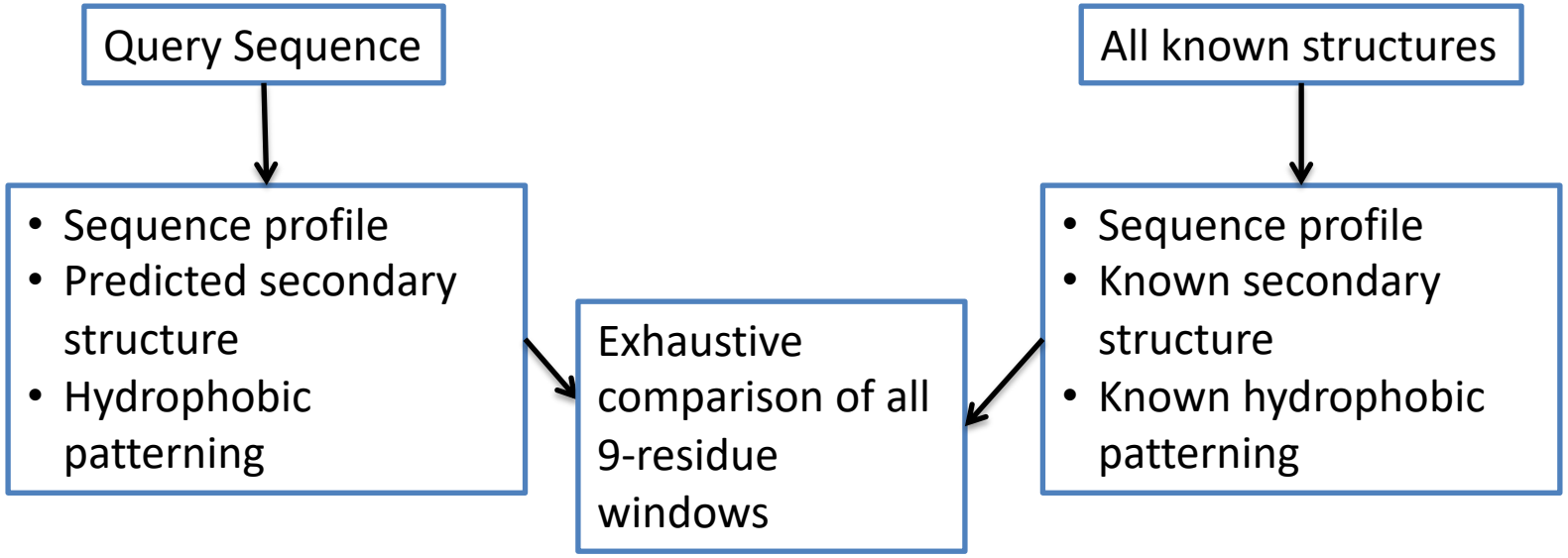
Insights from Folding Studies

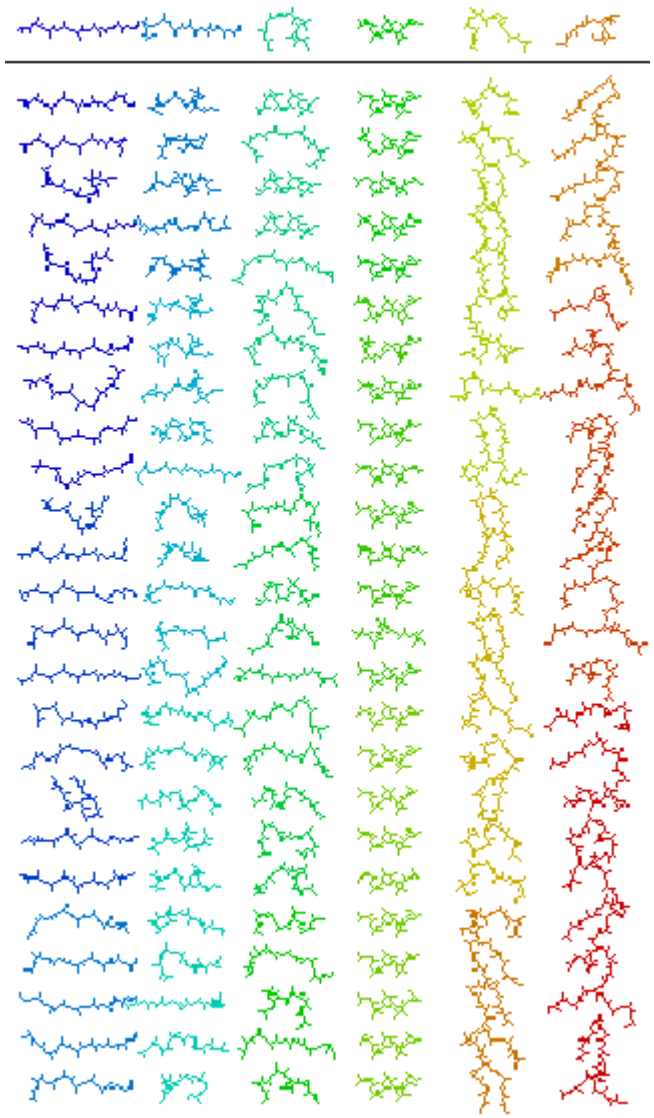
1. Local (*sequence-specific)* interactions strongly bias conformational sampling.
2. Folding is guided by hydrophobic burial, assembly of secondary structure, excluded volume.
3. Native interactions on average stronger / more consistent than non-native interactions => native minima broader than non-native minima.

Fragment-based Methods (Rosetta)

- **Hypothesis:** the PDB database contains all the possible conformations that a short region of a protein chain might adopt

- How do we choose fragments that are most likely to correctly represent the query sequence?



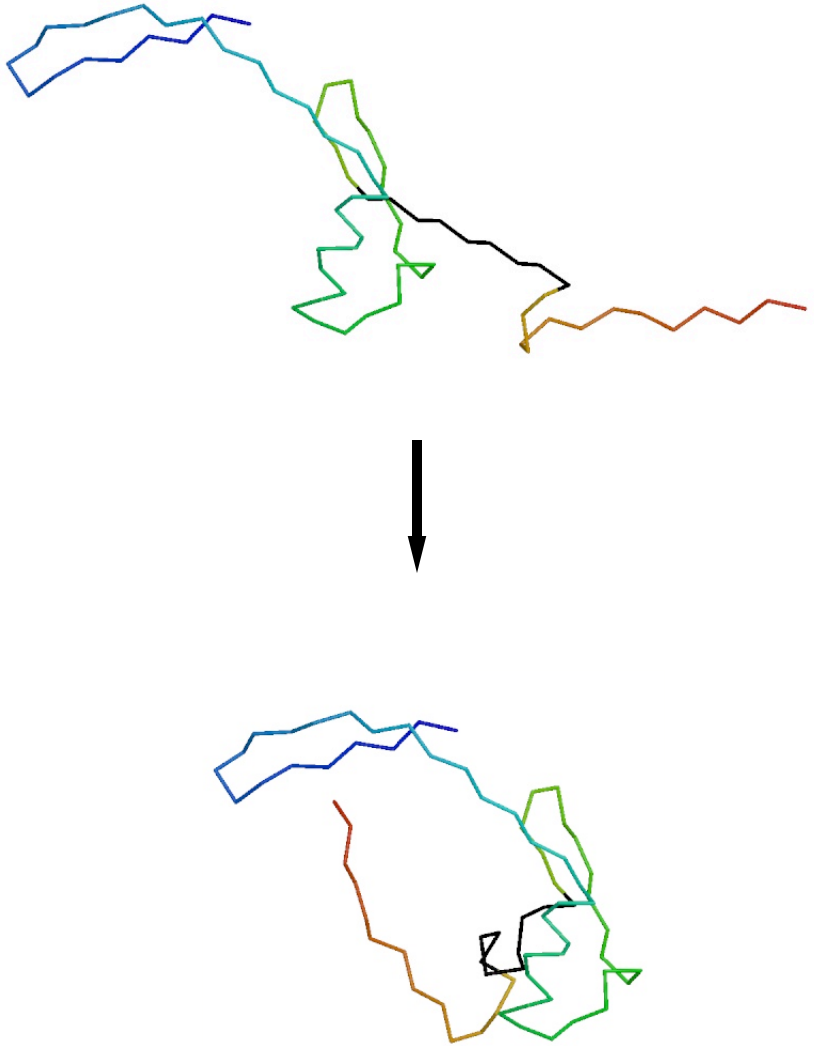


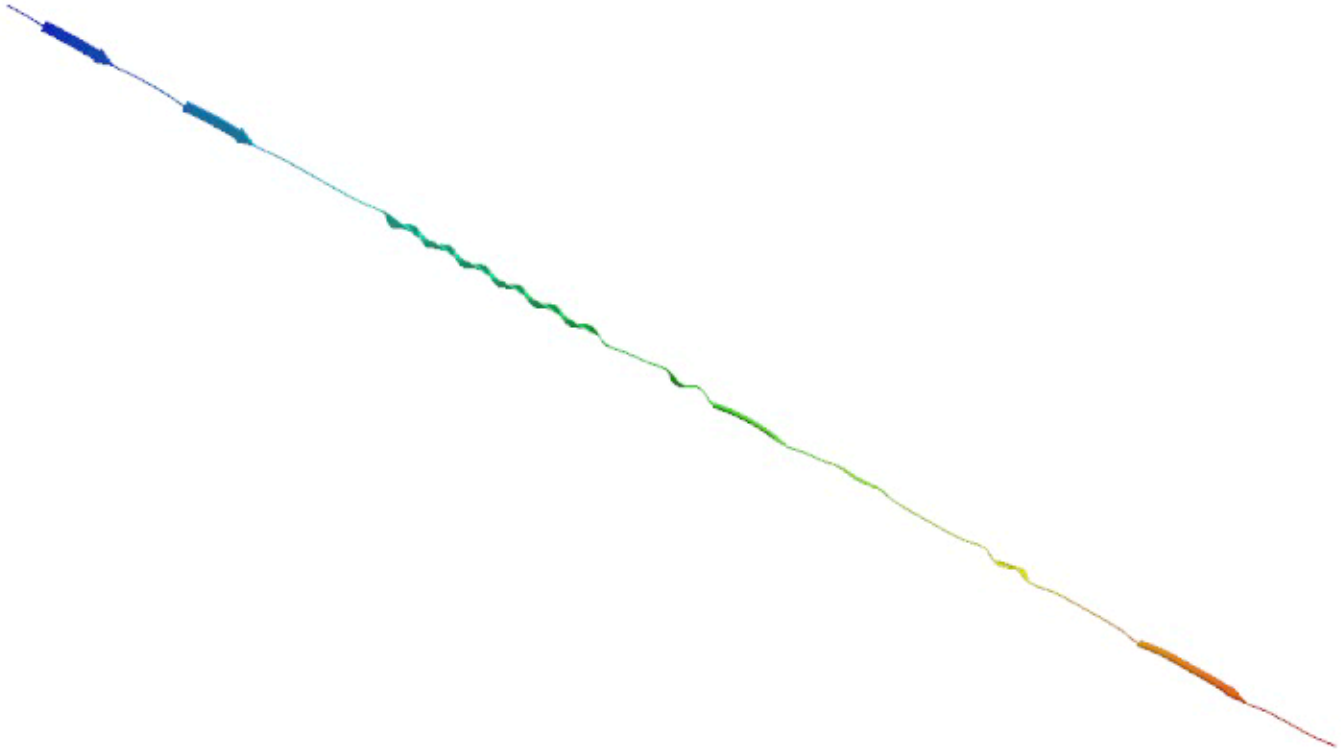
Fragment Libraries

- A unique library of fragments is generated for each 9-residue window in the query sequence.
- Assume that the distributions of conformations in each window reflects conformations this segment would actually sample.
- Regions with very strong local preferences will not have a lot of diversity in the library. Regions with weak local preferences will have more diversity in the library.

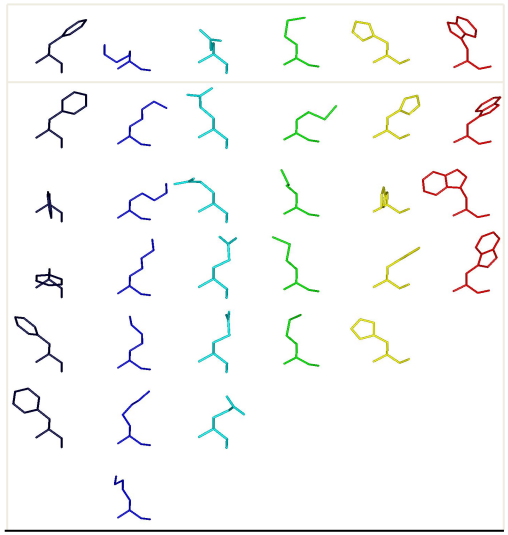
Generating Structures from Fragments

- Low resolution energy function used in initial search through conformational space
- Side chains represented by single "centroid" pseudoatom
- Major contributions from
  - Hydrophobic burial
  - Beta-strand pairing
  - Steric overlap
  - Specific residue pair interactions



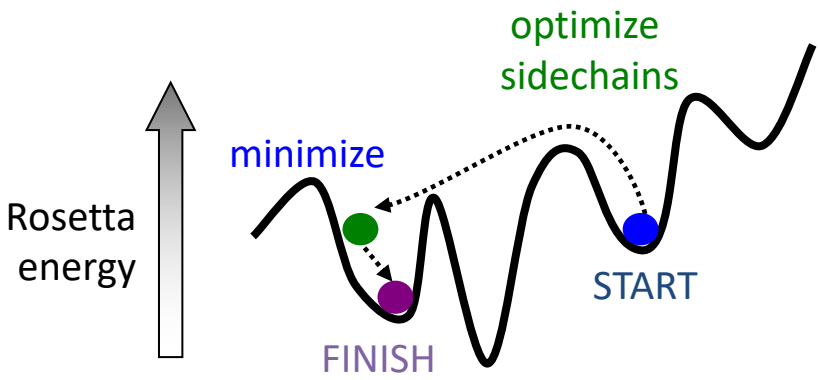


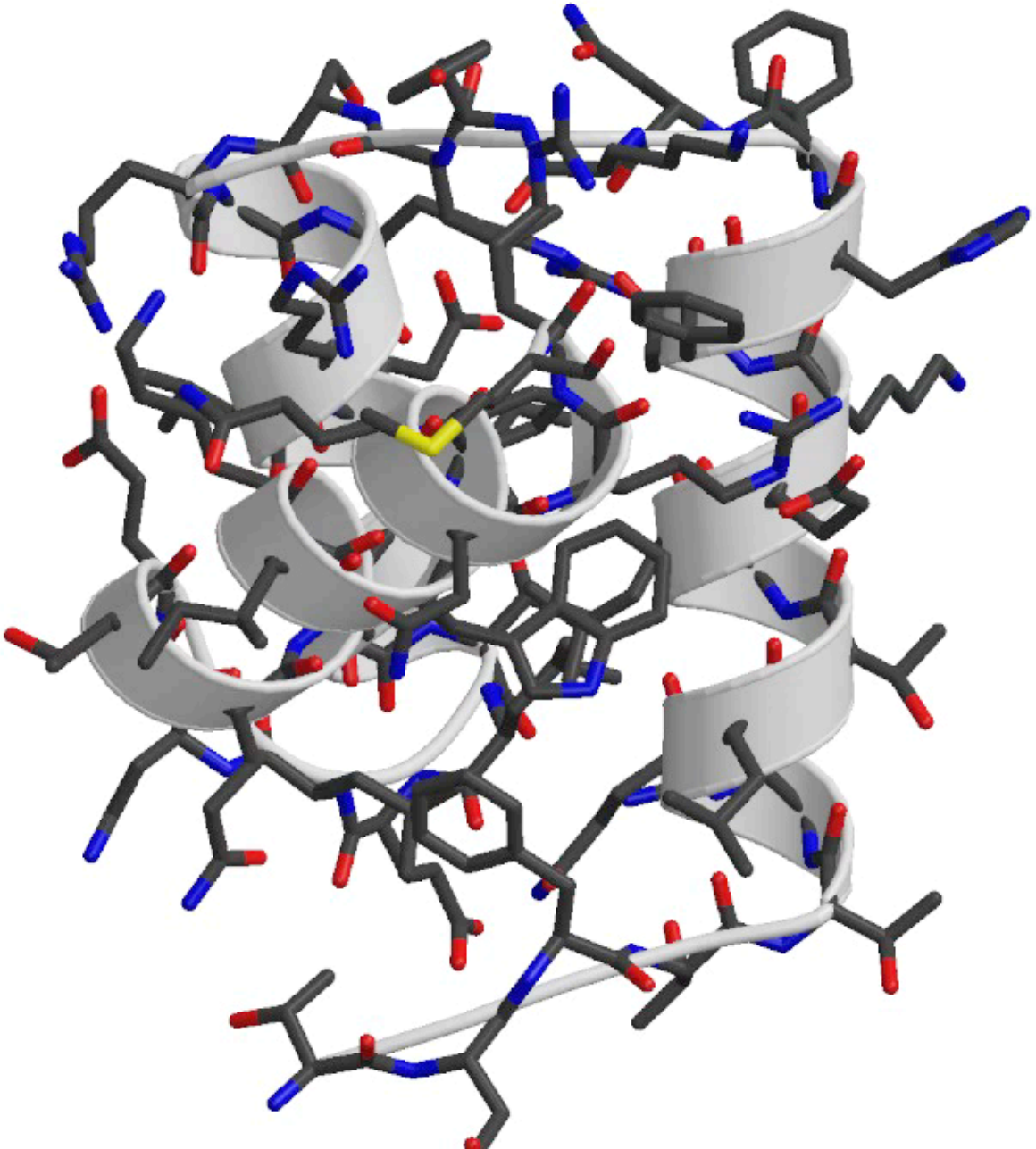
High-resolution model refinement



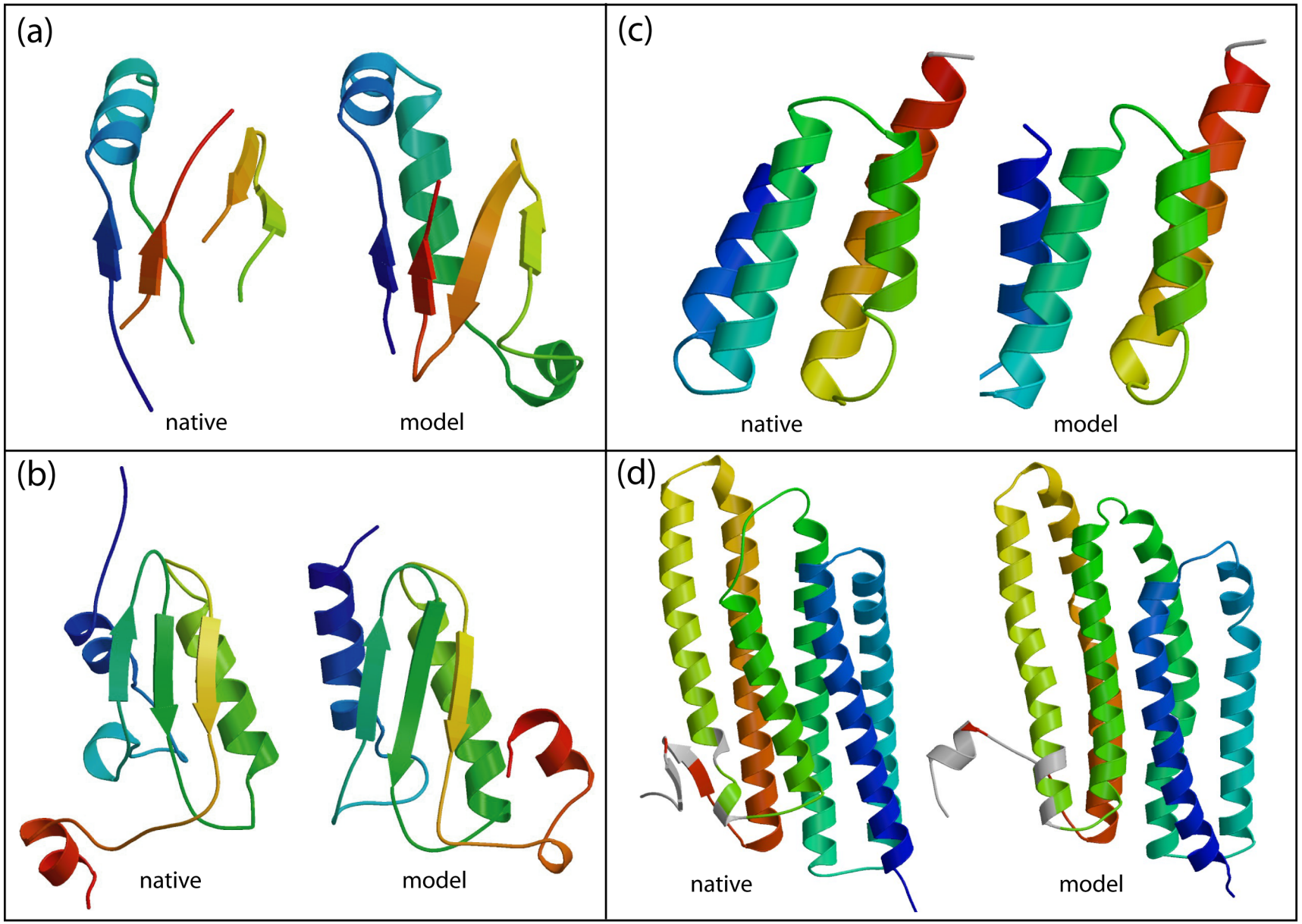
sidechain rotamers

- Rosetta “relax” structure refinement:
  - Discrete sidechain optimization via Simulated Annealing Monte Carlo
  - Gradient-based minimization of energy with respect to *torsion angles*
- Potential function: Rosetta all-atom energy
  - Lennard-Jones,
  - LK implicit solvation,
  - Coloumb electrostatics
  - orientation-dependent hydrogen bonding,
  - PDB derived torsional potential

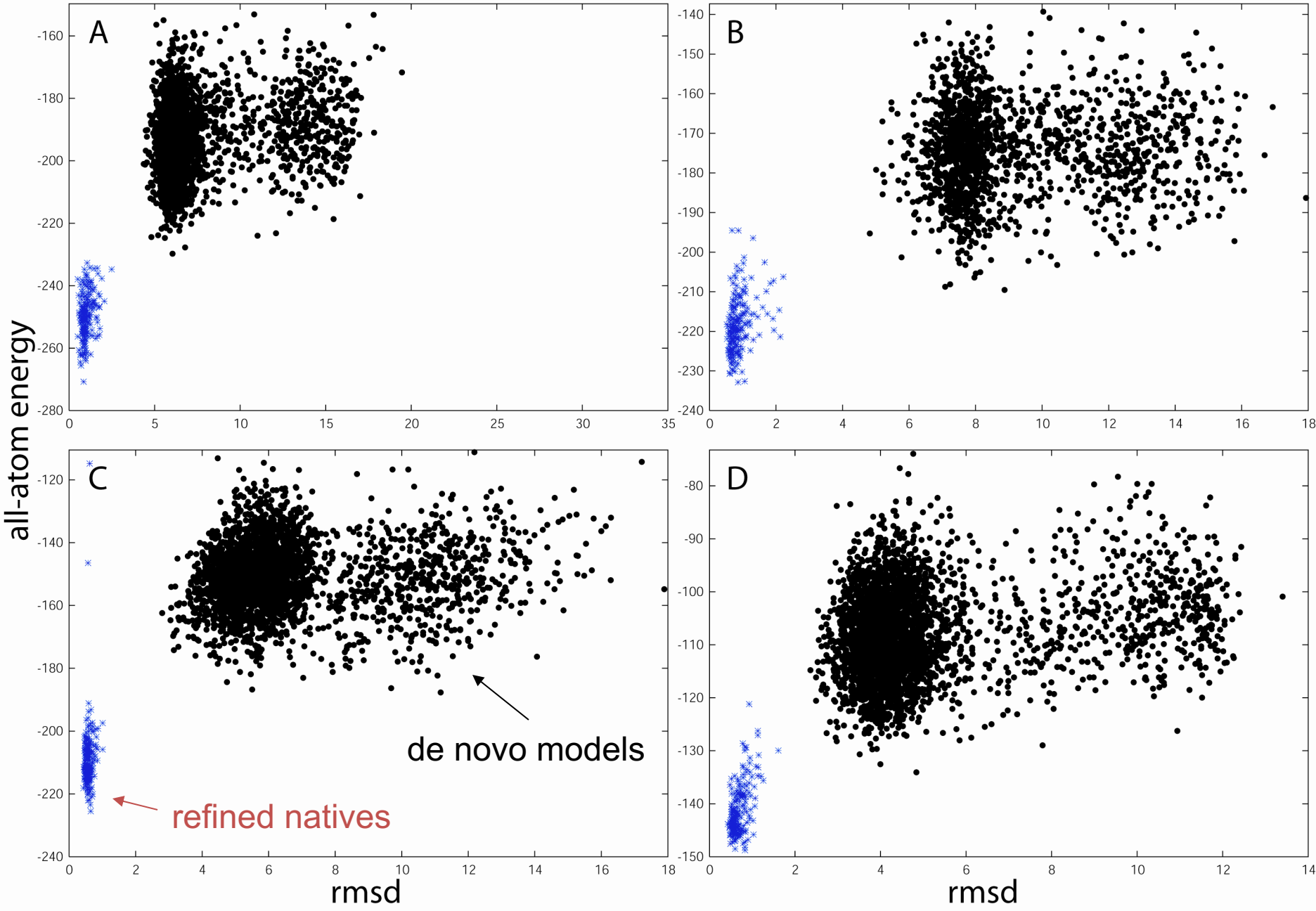




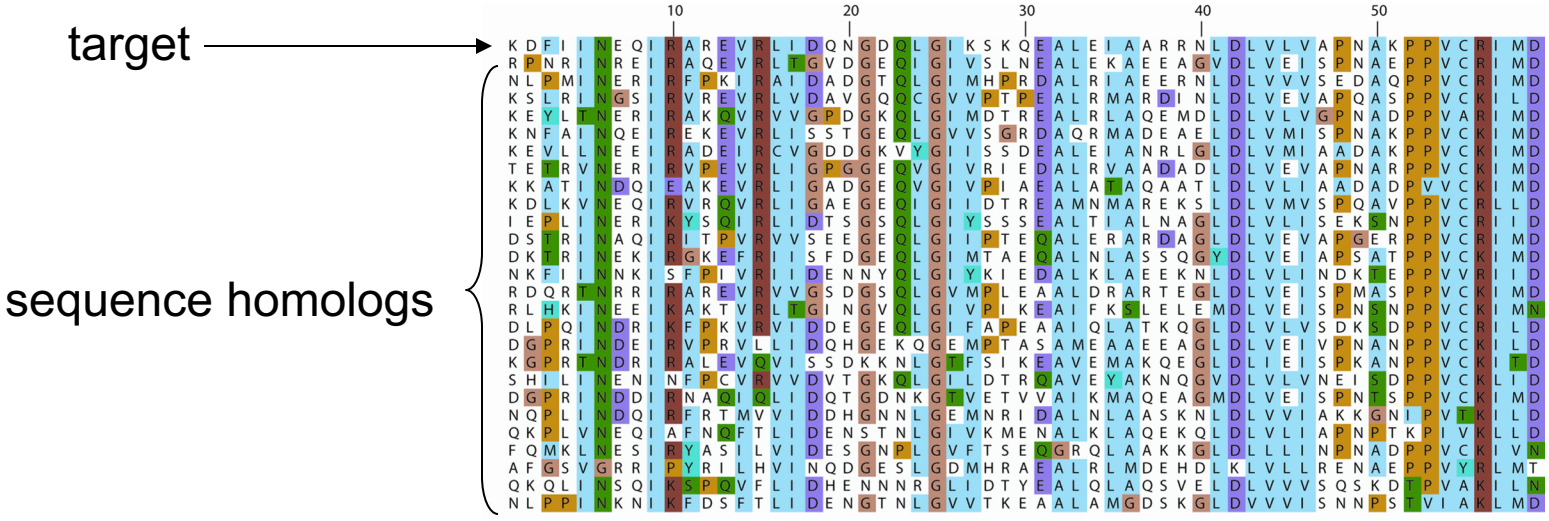
CASP predictions



Near-native models have lower all-atom energies

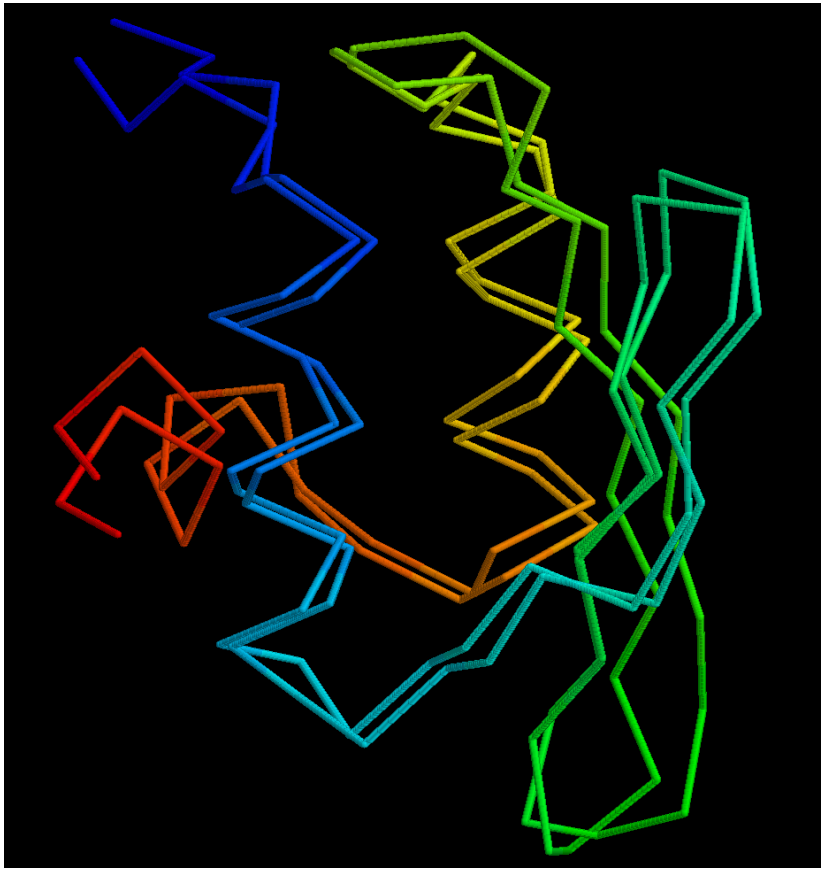


Use evolutionary information to improve sampling

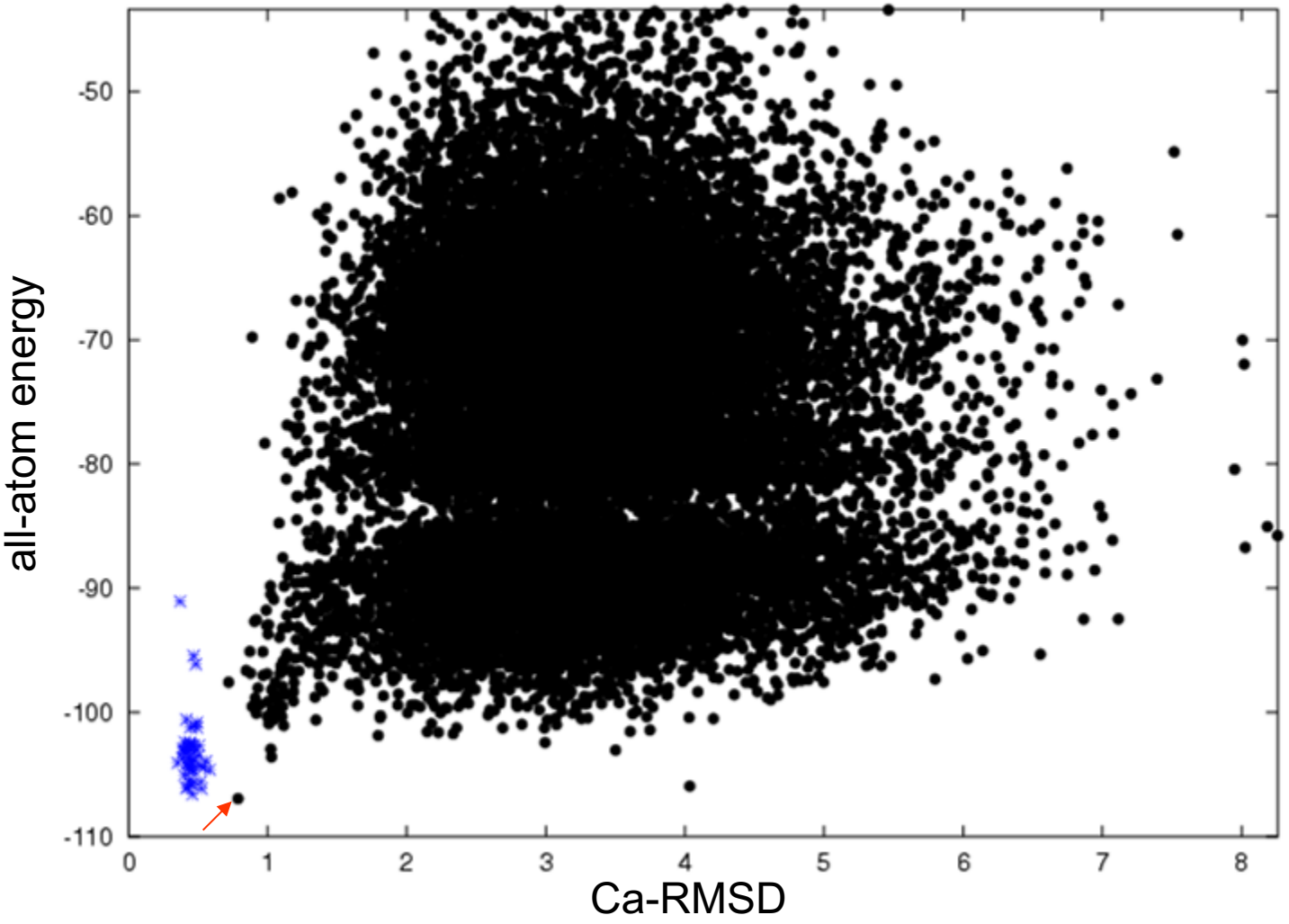


- Generate low-resolution models for all sequence homologs of the protein of interest. Thread the target sequence onto these models to generate a diverse population of structures for all-atom refinement.

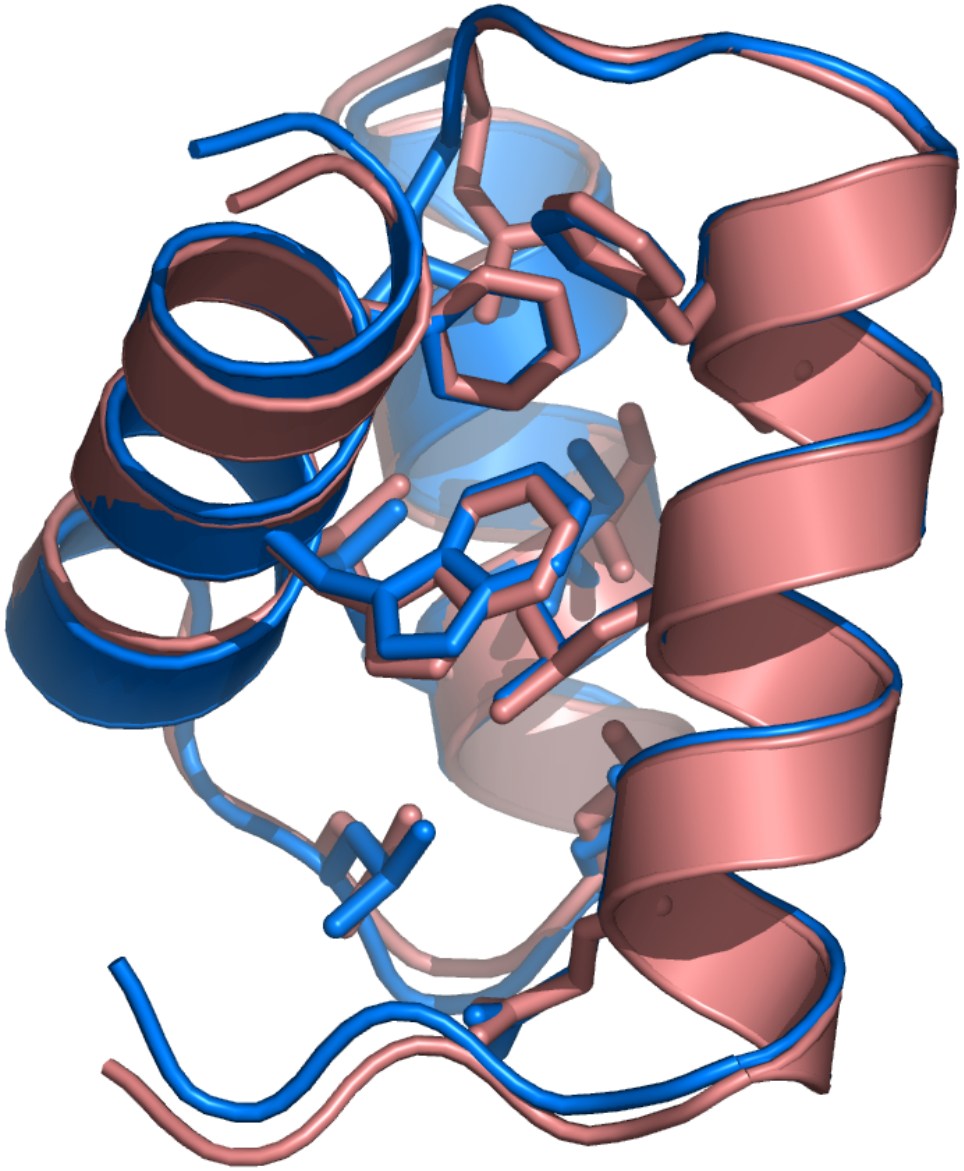
CASP6 *ab initio* prediction (1.59Å)



16 protein benchmark

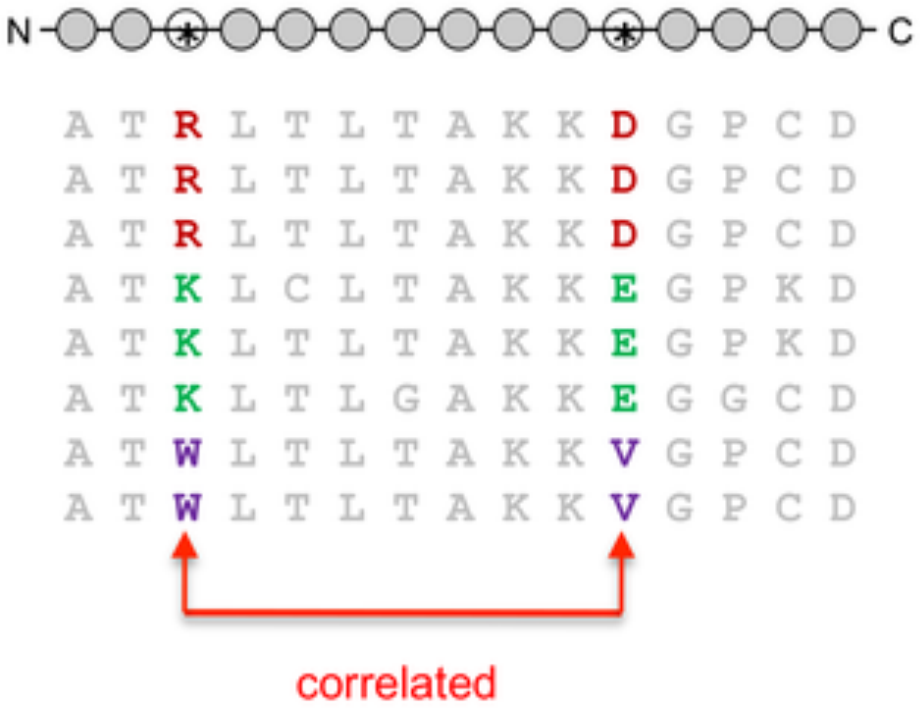


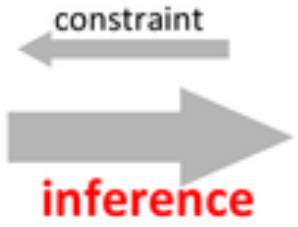
Hox-B1: 0.8Å (0.8Å core sidechains)

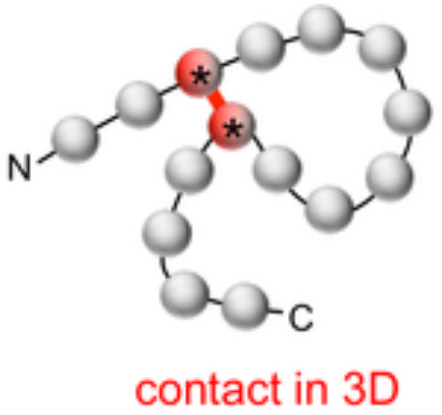


Hox-B1: 0.8Å (0.8Å core sidechains)

Correlated mutations carry information about distance relationships in protein structure.

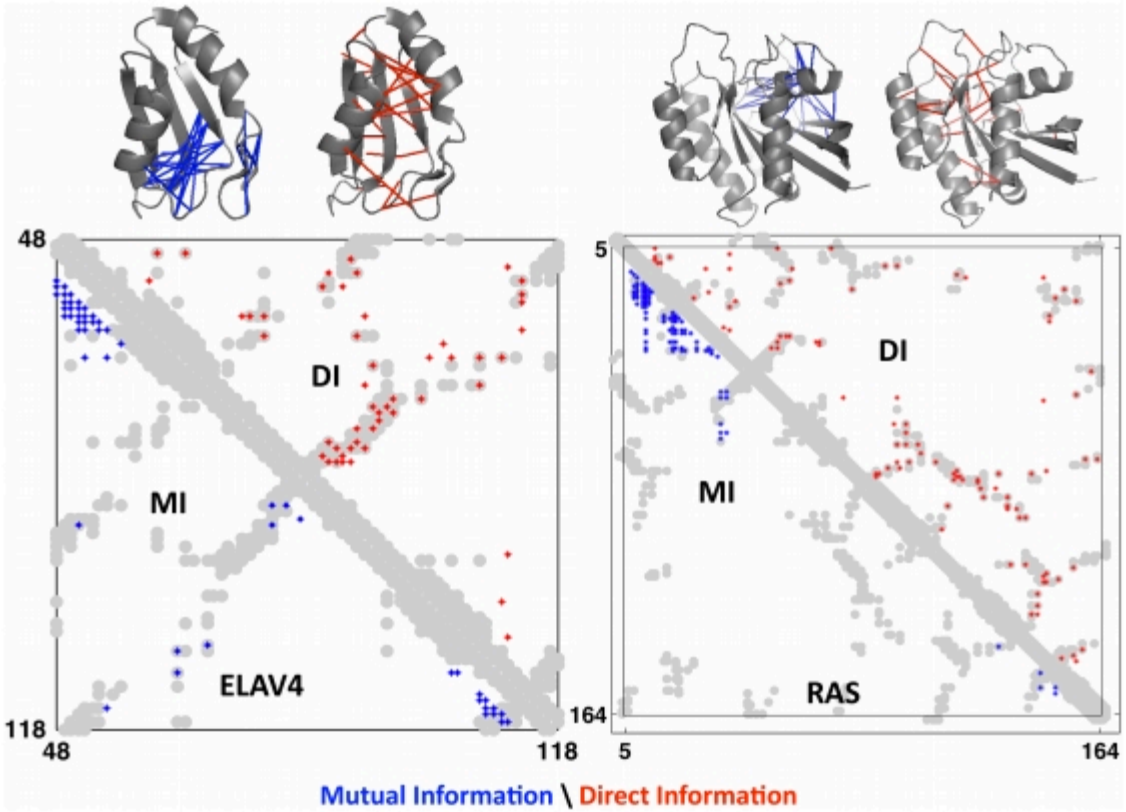






Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, et al. (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS ONE 6(12): e28766. https://doi.org/10.1371/journal.pone.0028766
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028766

Correlated mutations carry information about distance relationships in protein structure.



$$P(\mathbf{X}=\mathbf{x})=\frac{1}{Z}\exp\left(\sum_{i=1}^{L}\left[\mathbf{v}_i(x_i)+\sum_{j>i}^{L}\mathbf{w}_{ij}(x_i,x_j)\right]\right).$$

Learning the DCA (direct coupling analysis) matrix

The essence of DCA is then to assume that the rows, i.e. our aligned homologous proteins, are independent events drawn from a Potts-model probability distribution,

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} h_i(\sigma_i) + \frac{1}{2} \sum_{i,j=1}^{N} J_{ij}(\sigma_i, \sigma_j) \right), \tag{1}$$

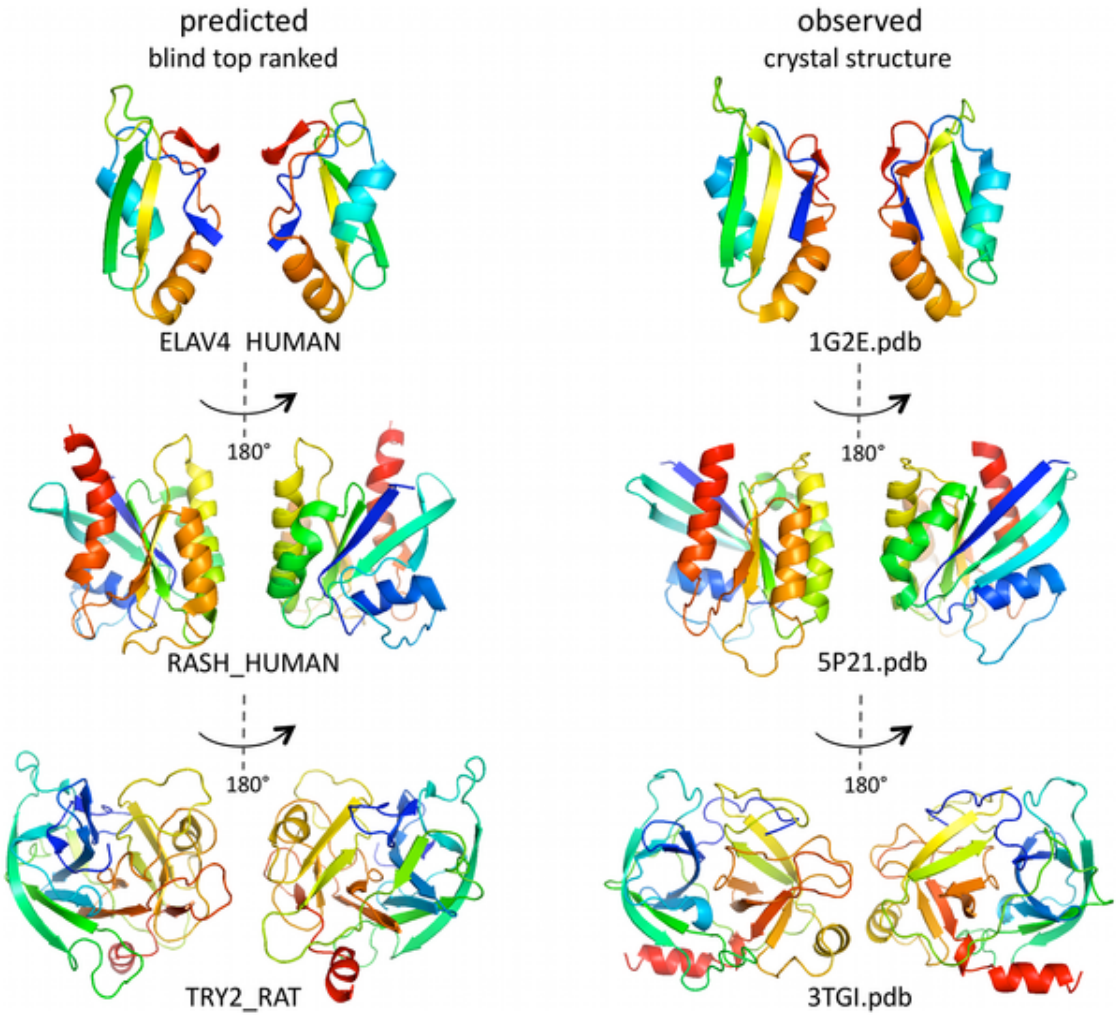
and to use the interaction parameters $\mathbf{J}_{ij}$ as predictions of spatial proximity among amino-acid pairs in the protein structure.

Problem: **Z** cannot be tractably computed

Solutions:

- Mean-field approach (mfDCA) (https://www.pnas.org/content/108/49/E1293)
- Pseudo-likelihood (plmDCA) (https://journals.aps.org/pre/abstract/10.1103/PhysRevE.87.012707)

Predicted 3D structures for three representative proteins.



Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, et al. (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS ONE 6(12): e28766. https://doi.org/10.1371/journal.pone.0028766
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028766

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model

Sheng Wang[☯], **Siqi Sun**[☯], **Zhen Li, Renyu Zhang, Jinbo Xu***

Toyota Technological Institute at Chicago, Chicago, Illinois, United States of America

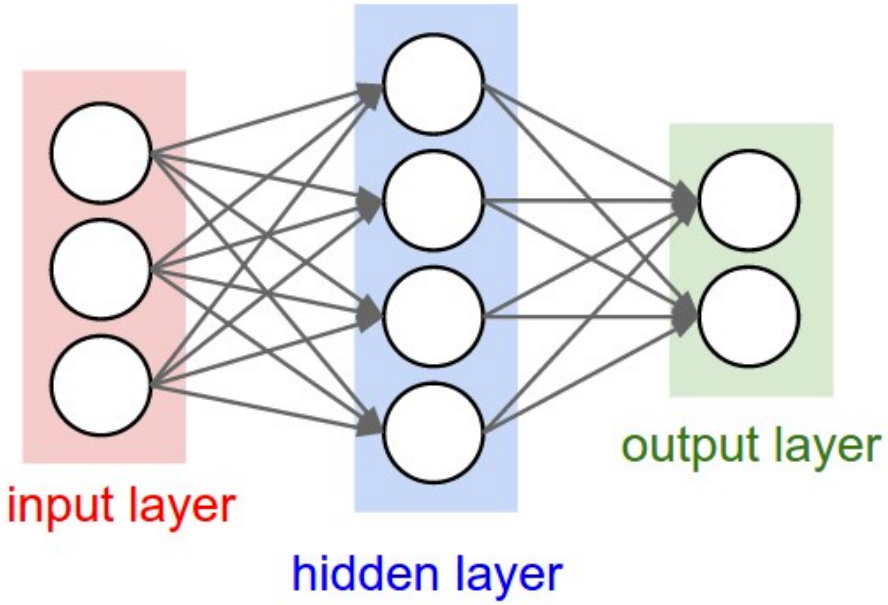
☯ These authors contributed equally to this work.
* jinboxu@gmail.com

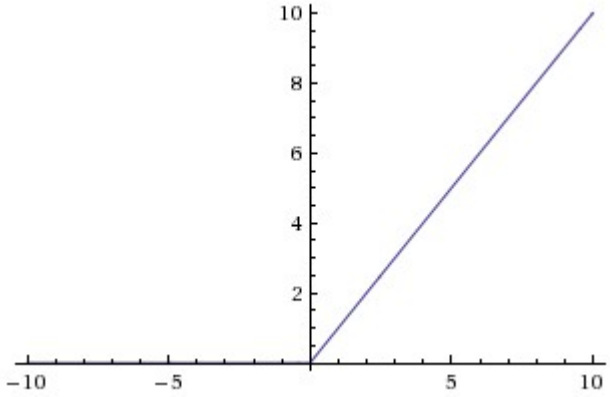
Abstract

Motivation

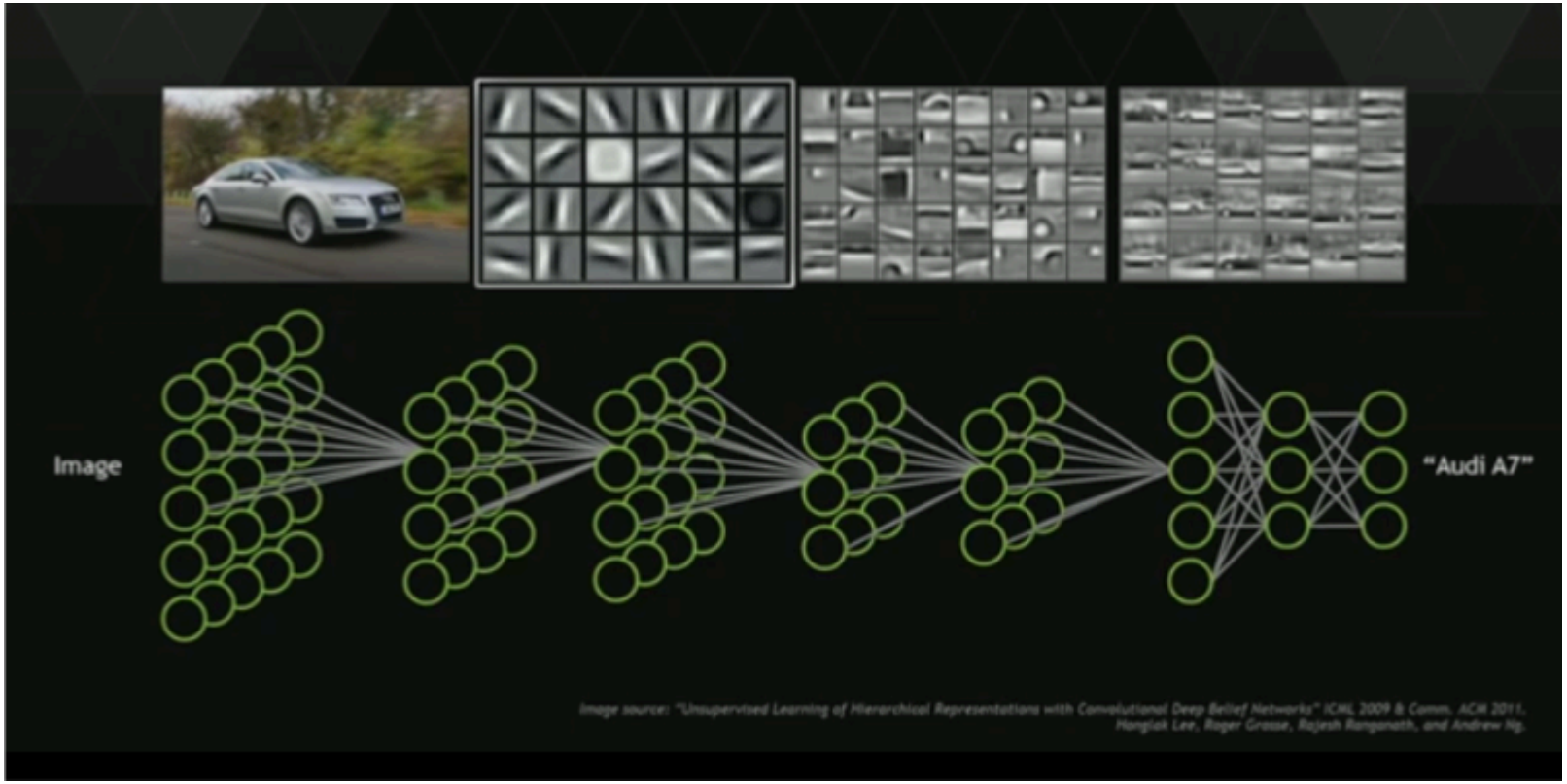
Protein contacts contain key information for the understanding of protein structure and function and thus, contact prediction from sequence is an important problem. Recently exciting progress has been made on this problem, but the predicted contacts for proteins without many sequence homologs is still of low quality and not very useful for de novo structure prediction.

Neural networks

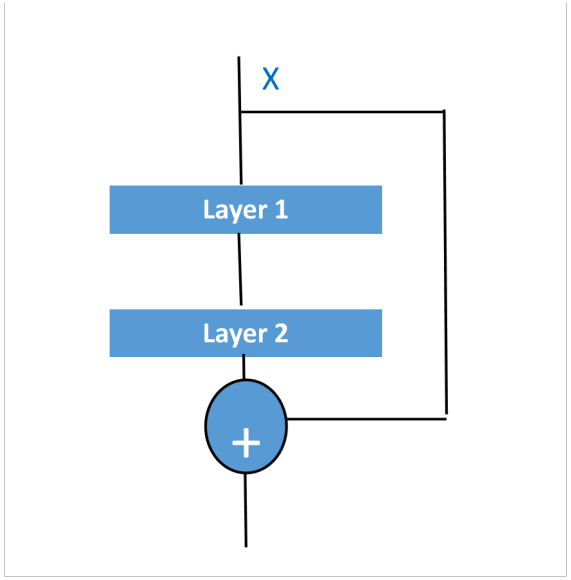


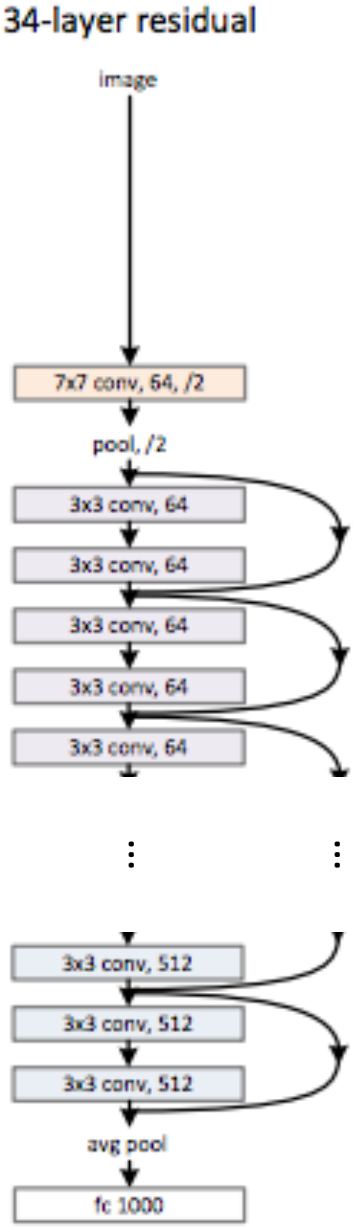


Convolutional neural networks

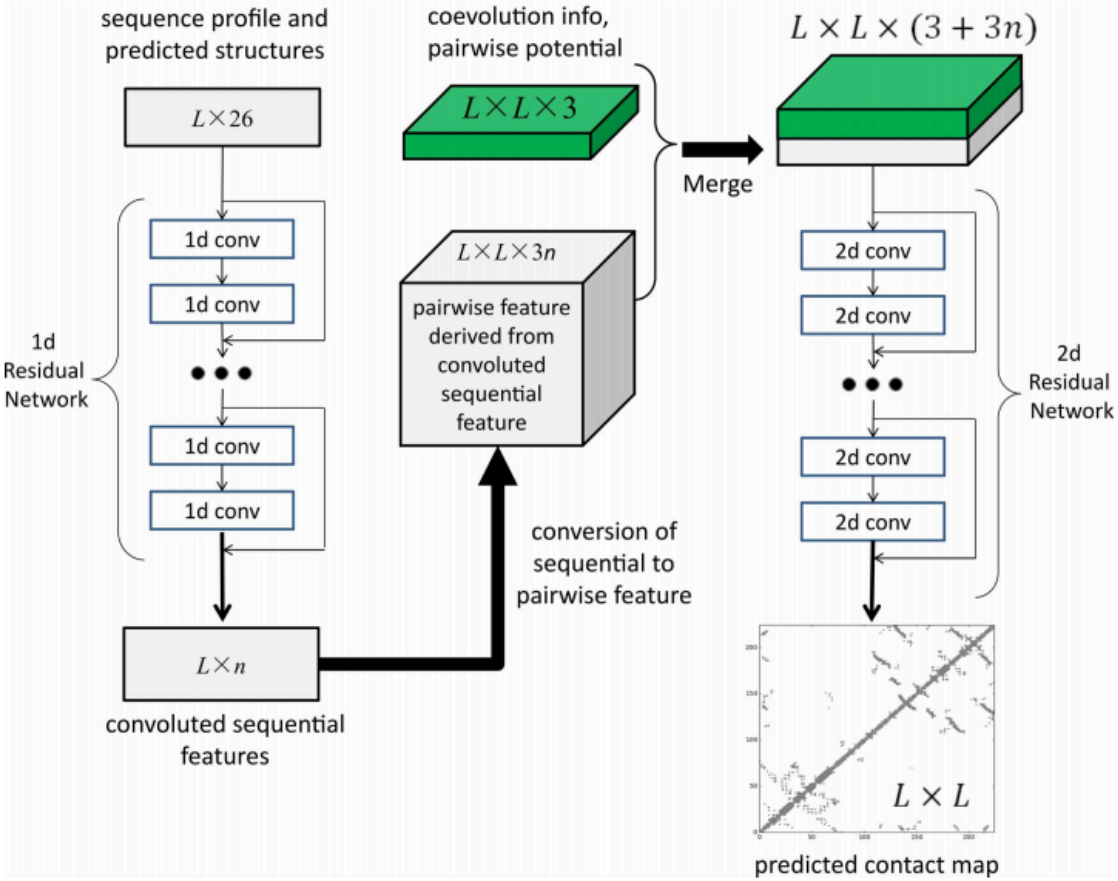


Deep residual neural networks

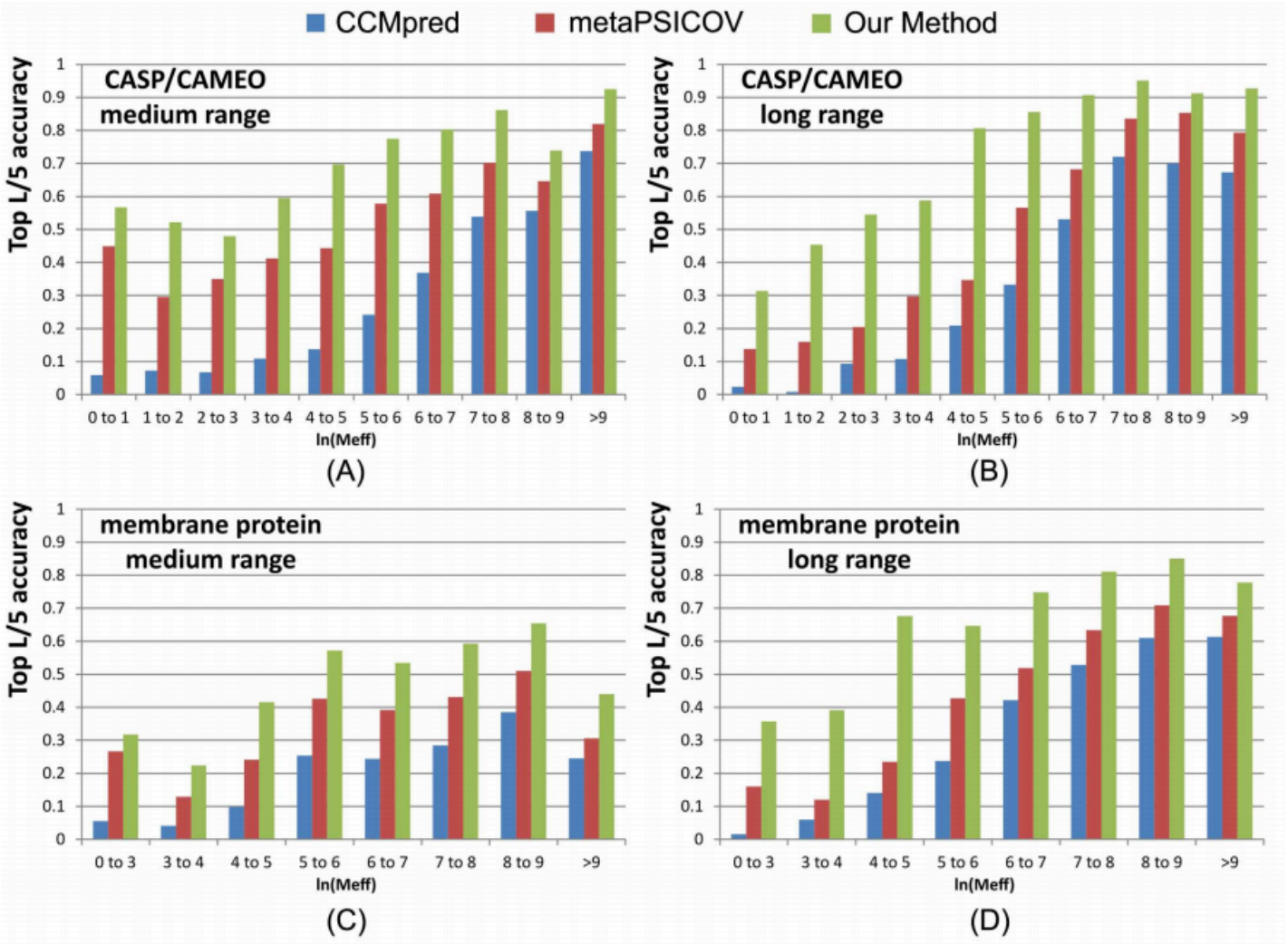




Learning a contact map from co-evolving residues



Contact prediction accuracy



Structure prediction accuracy

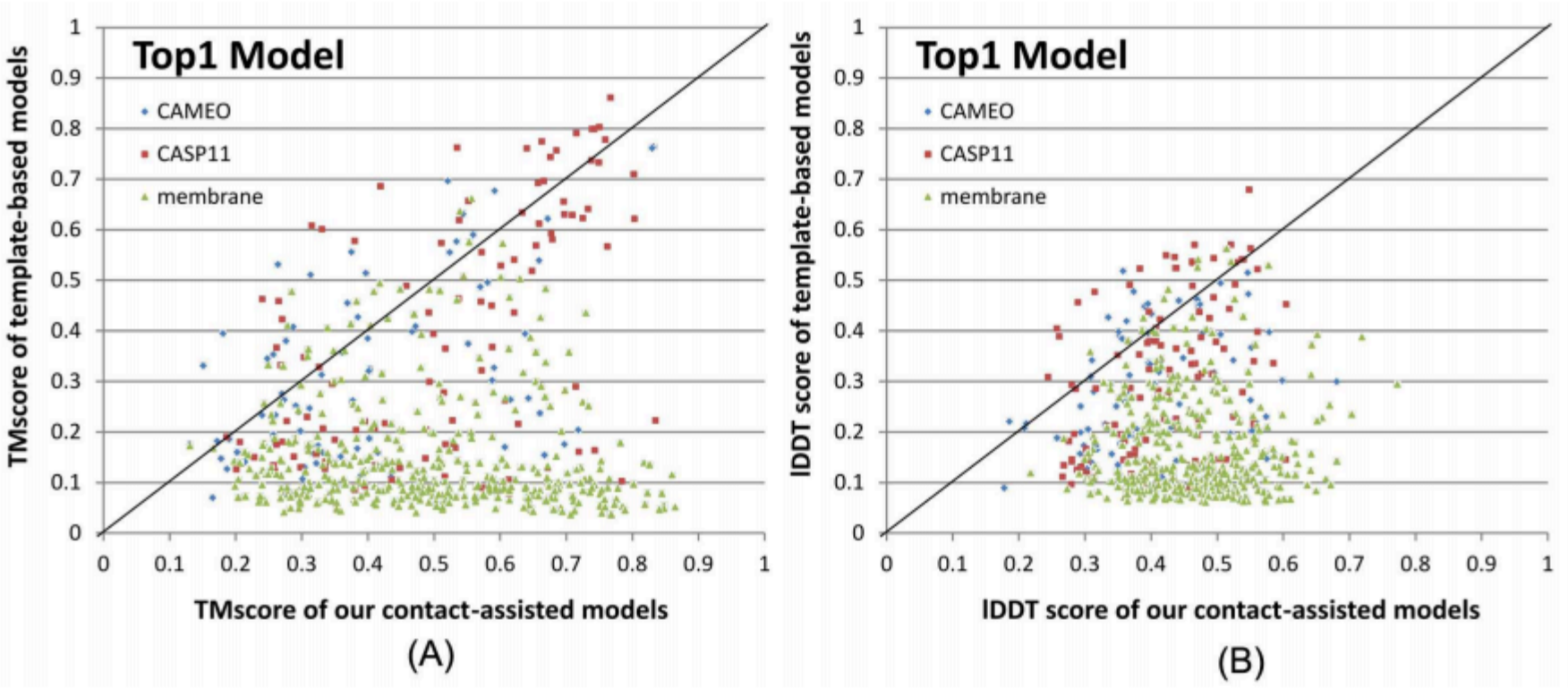
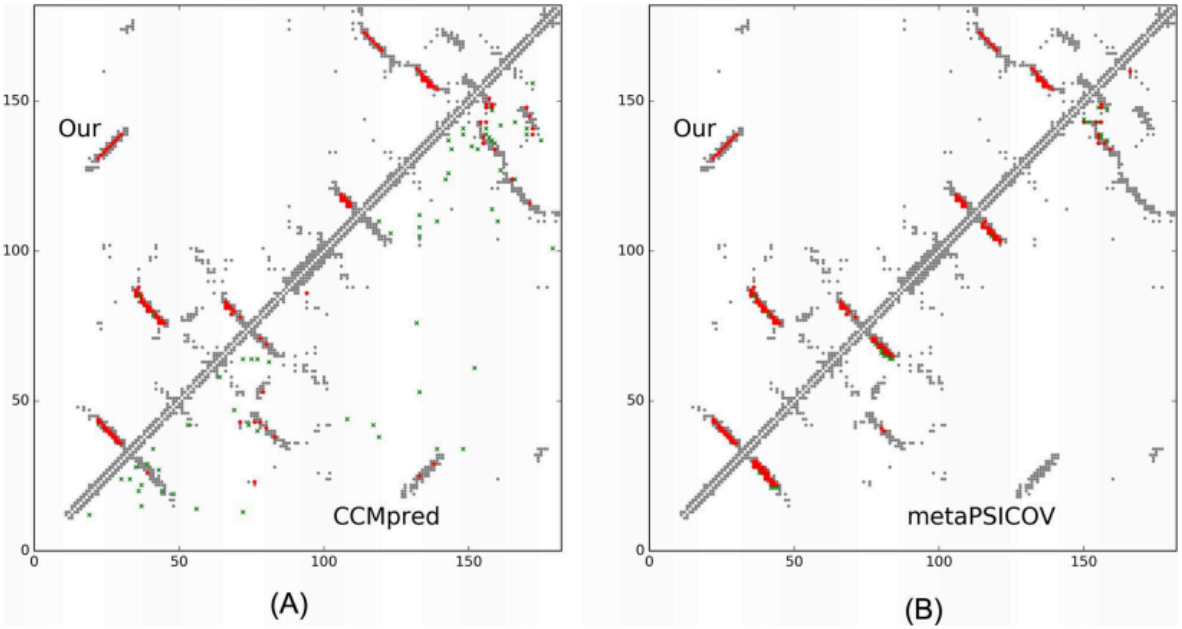


Fig 4. Comparison between our contact-assisted models of the three test sets and their template-based models in terms of (A) TMscore and (B) lDDT score. The top 1 models are evaluated.

Inferring better contact maps (I)



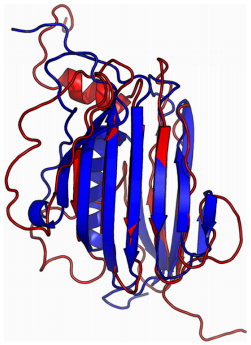
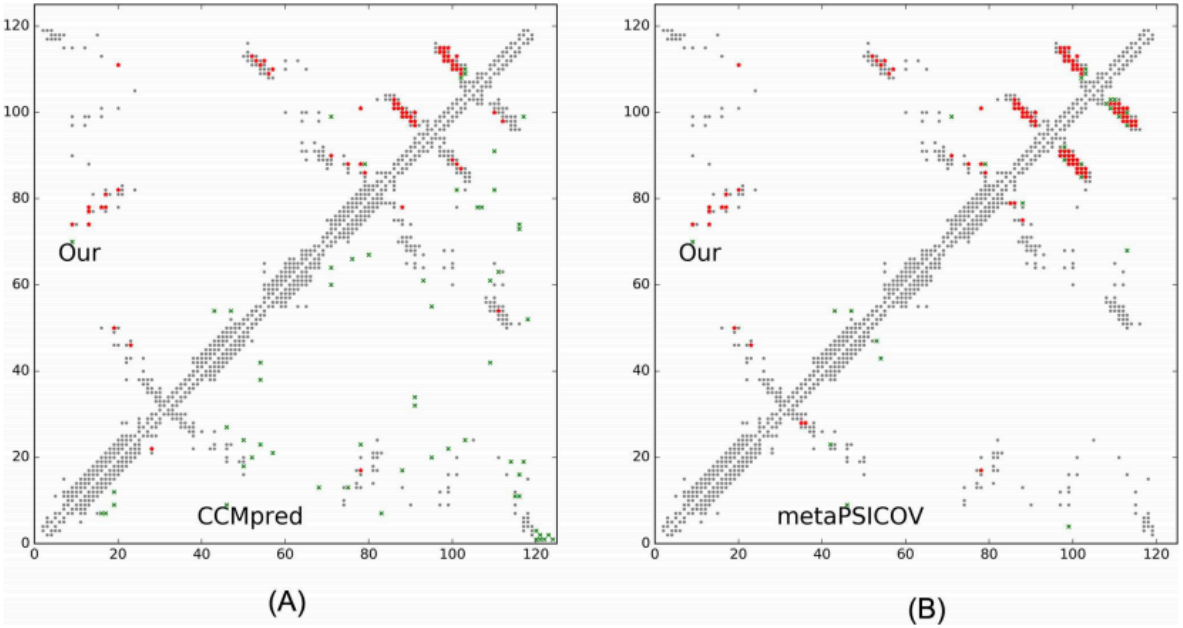


Fig 6. Overlap between top L/2 predicted contacts (in red or green) and the native contact map (in grey) for CAMEO target 2nc8A. Red (green) dots indicate correct (incorrect) prediction. (A) The comparison between our prediction (in upper-left triangle) and CCMpred (in lower-right triangle). (B) The comparison between our prediction (in upper-left triangle) and MetaPSICOV (in lower-right triangle).

Inferring better contact maps (II)



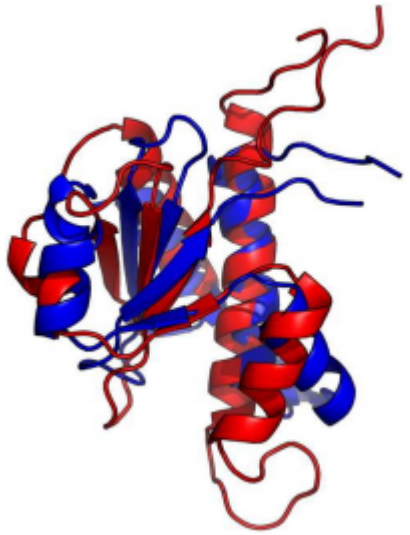


Fig 9. Overlap between top L/2 predicted contacts (in red or green) and the native contact map (in grey) for CAMEO target 5dcjA. Red (green) dots indicate correct (incorrect) prediction. (A) The comparison between our prediction (in upper-left triangle) and CCMpred (in lower-right triangle). (B) The comparison between our prediction (in upper-left triangle) and MetaPSICOV (in lower-right triangle).