Recent progress in rapid structure prediction without side-chain searching

Joe DeBartolo[1], Abhishek Jha[2], Andres Colubri[1], Karl F. Freed[2], Tobin R. Sosnick[1]*

[1]Department of Biochemistry and Molecular Biology, and the Institute for Biophysical Dynamics, University of Chicago, 929 E. 59th St., Chicago, IL 60637.

[2]James Franck Institute and Department of Chemistry, University of Chicago, 5735 S Ellis Ave, Chicago, IL 60637.

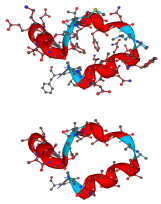
We predict protein structures for given amino acid sequences when the residues are pre-assigned to specific Ramachandran basins (RBs). The algorithm uses a protein model with a C_β side chain representation and a scoring function derived from an all-atom statistical potential that distinguishes between backbone atoms on different residue types. Structures are obtained by minimizing the scoring function with sampling of backbone torsional angles within their native RBs, termed "intra-basin folding". We also predict structures *ab initio,* which requires accurate determination of many of the native RBs. These RB assignments are obtained by "PsiBlast'ing" the PDB for fragments of other proteins that match our target's amino acid sequence. The fragments' RBs are used to construct a basin library. In many cases, the consensus from the basin library is sufficient to fold the protein using intra-basin folding. In other cases, a better basin prediction is generated using a new folding routine where RBs "hop" during the folding process according to their frequency in the basin library. This sampling is directed by profiling based on a statistical potential of pairwise interactions. Results are presented for several targets.

Statistical potential (DOPE-C_β)

DOPE (Discrete Optimized Protein Energy function) is an all-atom statistical potential from Shen & Sali, UCSF.

$$E_{DOPE} = \sum_{i,j} E_{ij}(r_{ij}) = -\sum_{i,j} \ln\left(\frac{f_{PDB}(r_{ij})}{f_{ref}(r_{ij})}\right)$$

Apply this "statistical potential function" to the simplified model

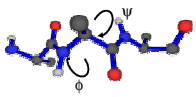


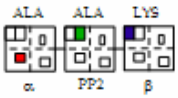
$E_{All\text{-}atom\ level} = E_{backbone+C\beta} + E_{side\text{-}chain}$

Throw away side-chain terms

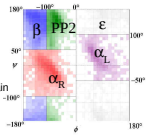
$E_{reduced\ representation} = E_{backbone+C\beta}$

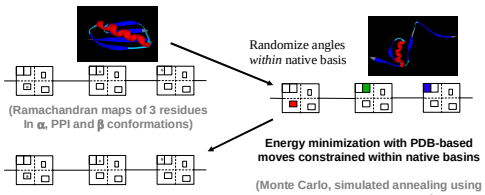
Sampling backbone dihedral angles



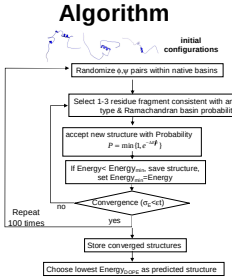


Sampling from a torsional library of trimers, dimers, and monomers remains within pre-assign Ramachandran basins.

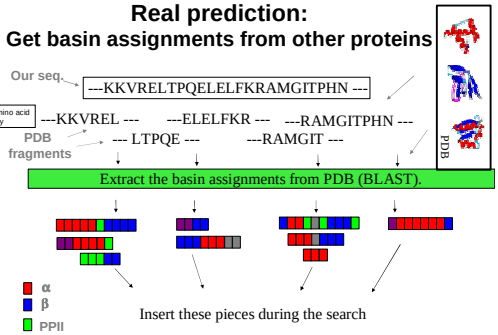




Algorithm



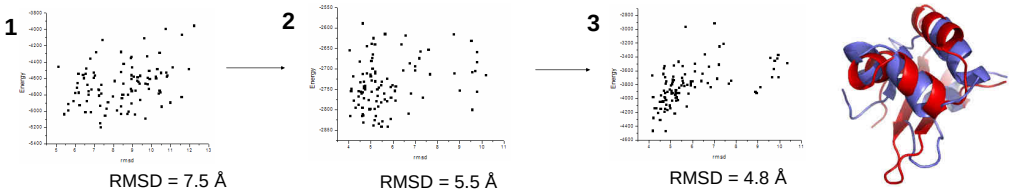
Real prediction:
Get basin assignments from other proteins



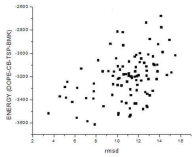
Examples of basin assignment improvements for *ab initio* routine

PDB ID	Percentage of correct basins from basin library consensus	Percentage of correct basins from ab initio consensus
2ICB	85	90
1MZM	77	96
1CTF	78	88
1AM3	92	100
1ORC	81	85
1AF7	84	87
1TIG	57	55

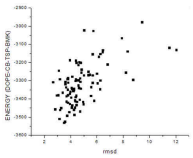
Ab initio structure prediction routine



Initial RB assignments obtained by Psi-blast'ing target sequence against the PDB for RB fragments. These fragments are the move-set when predicting 100 structures.



RB assignments obtained from the consensus values from the 100 predicted structures of the previous step. These RB assignments are used in the intra-basin algorithm



Add side chains to the 100 predicted structures of the previous step using SCWRL 3.0, and calculate energy with the all atom potential DOPE. Cluster using Cluster 3.0.

1UBQ native structure:
3.5 ab initio prediction from consensus

Predictions for folding within native basins.

PDB ID	Length	Secondary structure	DOPE-Cβ Prediction	DOPE Prediction	DOPE with clustering prediction	Bradley et al. (2005) prediction
1AF7	69	α	3.6	3.7	**3.4**	10.4
1B72	50	α	1.3	2	**1.3**	1.1
1CSP	67	β	7.7	8.9	**7.7**	4.7
1DCJ	72	αβ	2.2	3.2	**3.2**	2.5
1DI2	68	αβ	5.1	2.8	**2.8**	2.6
1MKY	77	αβ	10.8	9.1	**9.1**	6.3
1O2F	77	αβ	9.5	9.9	**7.9**	10.1
1TIF	57	αβ	2.5	1.9	**1.9**	4.1
1TIG	86	αβ	2.8	2	**2**	3.5
1R69	61	α	4.1	3.2	**4.5**	4.1
2REB	59	αβ	3.6	3.1	**3.1**	2.1
1SHF	59	β	8.7	8.7	**8.7**	10.8
			5.158333	4.875	**4.63333333**	5.191667

Canutescu, A. A., Shelenkov, A. A., Dunbrack, R. L., Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* 12:2001-2014, 2003.

Bradley, P., Misura, KM., Baker, D. Towards high resolution de novo structure prediction for small proteins. Science, 309, 1868-71, (2005)

Reference: M. J. L. de Hoon, S. Imoto, J. Nolan, and S. Miyano: Open Source Clustering Software. *Bioinformatics*, **20** (9): 1453--1454 (2004)

Predictions for several examples of folding within native basins: Sampling of the torsional library is restricted to pre-assigned Ramachandran basins, in this case those of the native structure. Side chains are added to the predictions using SCWRL 3.0, and structures are clustered using Cluster 3.0.