doi:10.1017/S1431927611005721

De Novo Atomic-Detail Structure Prediction for Proteins Guided by Medium Resolution Density Maps

Jens Meiler*, Steffen Lindert,* and Phoebe L. Stewart**

* Department of Chemistry, Vanderbilt University, Nashville, TN 37212
** Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, TN 37232

EM-Fold is a software algorithm that folds proteins into medium resolution density maps obtained by cryo-electron microscopy (cryo-EM) or X-ray crystallography [1]. Models built by EM-Fold and refined by Rosetta generally have root mean square distance deviations (RMSDs) between 4 Å and 7 Å over the full length of the protein. These results demonstrate that it is possible to use computational methods to generate models for proteins that have the correct topology from a medium resolution density map and the primary sequence alone. In short, EM-Fold is capable to use the experimental data to deduct the topology of the protein by adding helix directionality and connectivity not visibly in the density map. While, side chain conformations in selected helix-helix interfaces were predicted correctly, the models were accurate at atomic detail only in limited regions. Often predicted helices lacked bends or differed in length from helices in the experimental structure. Furthermore, even though the true topology could be enriched by selection of low energy models, it could not be identified by virtue of score alone. On the other hand, the high-resolution experimental structures had considerably better scores than any of the models built using the EM-Fold protocol.

It was concluded that model refinement to atomic detail accuracy failed due to insufficient sampling: starting from correct topology models, refinement does not construct models sufficiently close to the native structure to stand out by score – possibly because refinement was not guided by the cryo-EM density map. Therefore larger scale deviations such as length or bending of SSEs cannot be corrected. However, accurate construction of loop regions and side chains depends on models with very high agreement of backbone coordinates within secondary structure elements. The present work demonstrates how atomic-detail not visible in the experimental data can be added when including the electron density map as a restraint in the Rosetta refinement step of EM-Fold [2]. This strength of the Rosetta algorithm was already demonstrated when combined with NMR and EPR experimental data [3-9].

The models created by EM-Fold for each of the seven successful cases from the benchmark in [1] were subjected to Rosetta loop building and refinement using the new density restraint functionality. Models from the EM-Fold refinement step were taken as start models for the loop construction in Rosetta. The loop building (round 1) was followed by two more rounds of identifying the regions of the models that agree least with the density map and then rebuilding these regions and relaxing the entire protein [2]. Table 1 summarizes the results of this work.

Table 1. Results of Rosetta refinement on seven successful benchmark proteins

protein	RMSD best start model [Å]	best RMSD model after round 1 [Å]	best RMSD after round 2 [Å]	best RMSD after round 3 [Å]
1IE9	(2.22)	3.88 (2.25)	3.05 (1.90)	2.55 (1.93)
1N83	(4.68)	5.21 (4.27)	4.31 (3.18)	3.41 (2.63)
1OUV	(2.21)	2.37 (2.01)	2.05 (1.80)	2.00 (1.79)
1QKM	(2.95)	3.72 (2.93)	2.87 (3.02)	2.82 (2.33)
1TBF	(1.93)	3.32 (2.26)	2.86 (2.19)	2.37 (1.96)
1Z1L	(2.70)	3.94 (3.05)	3.62 (3.12)	3.24 (2.66)
2AX6	(2.26)	4.22 (2.48)	3.61 (2.73)	2.99 (2.36)

RMSD values were determined over the backbone atoms N, C_α, C and O. Values in parentheses refer to RMSDs over secondary structure elements only.

After three rounds of Rosetta refinement, the RMSDs of the final models range from 2.0 Å to 3.4 Å over the full length of the proteins and between 1.8 Å and 2.6 Å over the helical residues. For proteins of size 250 to 350 residues, RMSD values below 2.5 Å generally mean that side chain conformations at least within the core of the protein are correctly recovered. These results are considerably better than the results obtained when refining with the version of Rosetta that does not use the density map as a restraint, where the RMSDs of the best RMSD models ranged from 3.9 Å to 7.1 Å over the full length of the protein. Fig. 1 shows the final model overlaid with the native structure for one of the seven benchmark cases.

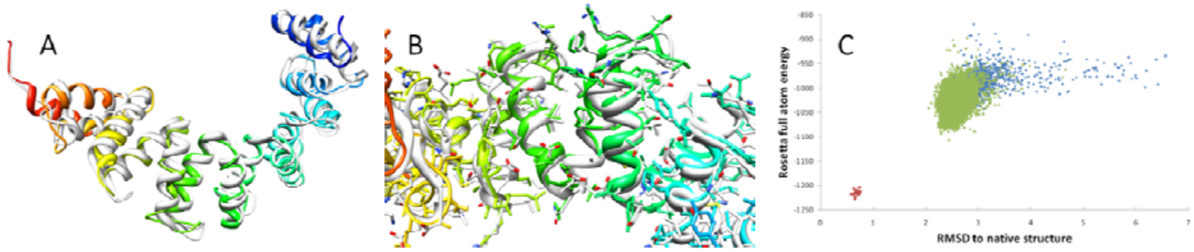

FIG. 1. Superimposition of one of the final models (colored in rainbow) of 1OUV after Rosetta refinement with original PDB structure (grey). (A) Comparison of the entire protein model with the native structure. (B) Side chain agreement of model with native structure in helical interface. (C) RMSD vs. Rosetta energy plot for the refinement of 1OUV.

[1] S. Lindert, et al., *Structure* 17 (2009) 990.
[2] F. DiMaio, et al., *J Mol Biol* 392 (2009)
[3] N. Alexander, et al., *Structure* 16 (2008)
[4] P. M. Bowers, et al., *J Biomol NMR* 18 (2000)
[5] S. M. Hanson, et al., *Structure* 16 (2008)
[6] J. Meiler and D. Baker, *Proc Natl Acad Sci U S A* 100 (2003)
[7] J. Meiler and D. Baker, *J Magn Reson* 173 (2005)
[8] B. Qian, et al., *Nature* 450 (2007)
[9] C. A. Rohl and D. Baker, *J Am Chem Soc* 124 (2002)
[10] This research was supported by NIH grants to PLS (R01 CA140538) and JM (NSF 0742762)