doi:10.1017/S1431927611001504
Microsc. Microanal. 17 (Suppl 2), 2011

Predicting Models for a Domain of DNA-PKcs

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

* 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 cryoEM or X-ray crystallography [1]. In this work EM-Fold was applied to build a model for a region of DNA dependent protein kinase catalytic subunit (DNA-PKcs). Both a medium resolution cryoEM density map [2] and a medium resolution crystal structure [3] of the molecule have been determined recently. Neither was at sufficient resolution to trace the backbone of the molecule. The catalytic subunit contains 4128 residues and has about 135 predicted helices (68% of the sequence) which is about one order of magnitude too large for direct application of EM-Fold.

However, the density maps clearly identify an extended heat repeat motive of 25 density rods. To identify the sequence that corresponds to this part of the map, the entire sequence was submitted to Pfam [4]. Four matches to the target sequence were identified with significant score: NUC194 domain (alignment to residues 1815 - 2210), FAT domain (alignment to residues 3023 - 3470), Phosphatidylinositol 3- and 4-kinase (alignment to residues 3748 - 4014) and FATC domain (alignment to residues 4097 - 4128). Closer inspection of the results revealed that the FAT domain is a member of the Tetratrico peptide repeat superfamily (TPR), many of which are heat repeats. Also a visual inspection of the secondary structure prediction for the entire DNA-PKcs revealed a region consisting of 31 α-helices of similar length between residues 2700 and 3540. This segment underwent fold recognition using Phyre [5]. Several of the fold recognition results were significant (E-values smaller than 1.0e-06) and are heat repeats very close in overall shape and size to the density map. Examples include Karyopherin β2 (SCOPE: d1qbkb, PDB: 1qbk, E-value: 2.5e-06) and Importin β (SCOPE: d1qgra, PDB: 1qgr, E-value: 3.5e-06). The sequence identity of the significant hits ranges from 5 to 10%. These results corroborate that region 2900 - 3540 in sequence corresponds likely to the heat repeat region in the density maps. The structures of the ten most significant hits were fitted into the heat repeat regions of the density map. Six of them including Karyopherin β2 and Importin β are good fits in terms of size and overall shape of the molecule. However, only about 20% of the density rods are filled with an accurately placed α-helix. Panel D in Fig 1 actually shows the fit of Karyopherin β2 into the density map.

The programs jufo, psipred and profPhD were used to predict secondary structure for the heat repeat domain. The predictions among those methods agree very well. A total of 31 helices of ten or more residues were predicted. The density map used for input to EM-Fold was generated from the crystallographic structure factors [3]. In the heat repeat region about 25 density rods of at least 13.5 Å in length are observed. The density map that was originally calculated from the structure factors with the crystallographic CCP4 software package does not contain perpendicular axes, rather it has cell axes of 90°, 105°, and 90°. A function OrthorgonalizeMap was implemented to convert a density map with non-orthogonal axes into a map with orthogonal axes. Then EM-Fold assembly and refinement steps were performed in a similar manner to that

described for previous applications of EM-Fold. 200 top scoring topologies from the assembly step were transferred to the refinement step and the top scoring 100 refined topologies were transferred into Rosetta.

Evaluating the top 200 scoring models after the assembly step showed that models had been built into the density map in both possible orientations for the N-terminal end of the sequence region. However the majority of models (167/200) have their N-terminal end in the lower part of the density toward the "base" region of the molecule. Of the top 100 models after refinement step, 75 have their N-terminal end in the lower part of the density. The top 100 scoring topologies after EM-Fold refinement served as input for the first round of Rosetta refinement. The top scoring 30 topologies after the first round were carried over into a second round of Rosetta refinement and finally the top 20 topologies from the second round went into a third round of Rosetta refinement. Of the top 20 models after the third round of Rosetta refinement, 17 have their N-terminal end in the lower part of the density. A closer look at the average Rosetta Energy Unit (REU) per residue revealed that the top scoring DNA-PKcs models have 1.8 REU/ helical residue. This compares to 2.6 REU/ helical residue for the top scoring models of helical proteins in a published benchmark [1]. The somewhat less favorable average REU values for the DNA-PKcs models may be related to difficulty in modeling such a large protein as accurately as the benchmark proteins which had an average size of about 200 residues.

We show models of the top three scoring topologies after the third round of Rosetta refinement (Fig. 1). They give a good indication of how the sequence could fit into the density map and also allow for targeted experiments designed to test our predictions. Based on this work we hypothesize that region 2900 - 3540 of the sequence corresponds to a heat repeat region in the density map and that the N-terminus of this domain points to the "base" region of the molecule [6].

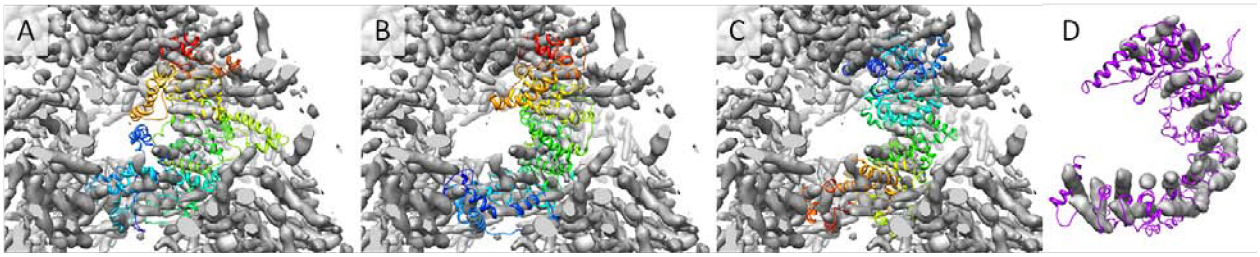

FIG. 1. Models representing the top scoring three topologies after EM-Fold and Rosetta refinement. (A, B) The top two scoring topologies have their N-terminal end in the lower part of the density toward the region referred to as the base. (C) The third best scoring topology has its N-terminal end in the upper part of the density. (D) Fit of Karyopherin β2 into density map.

[1] S. Lindert, et al., *Structure* 17 (2009) 990.
[2] D. R. Williams, et al., *Structure* 16 (2008) 468.
[3] B. L. Sibanda, et al., *Nature* 463 (2010) 118.
[4] R. D. Finn, et al., *Nucleic Acids Res* 38 (2010) D211.
[5] L. A. Kelley and M. J. Sternberg, *Nat Protoc* 4 (2009) 363.
[6] This research was supported by NIH grants to PLS (R01 CA140538) and JM (NSF 0742762)

https://doi.org/10.1017/S1431927611001504 Published online by Cambridge University Press