Computational Biology and Chemistry 42 (2013) 1–4

Contents lists available at SciVerse ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research article

# Computational determination of the orientation of a heat repeat-like domain of DNA-PKcs

Steffen Lindert[a,1], Phoebe L. Stewart[b,2], Jens Meiler[a,*]

[a] Department of Chemistry, Vanderbilt University, Nashville, TN 37212, USA
[b] Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, TN 37232, USA

ARTICLE INFO

*Article history:*
Received 10 July 2012
Accepted 7 November 2012

*Keywords:*
Computational structure prediction
Medium resolution density maps

ABSTRACT

DNA dependent protein kinase catalytic subunit (DNA-PKcs) is an important regulatory protein in non-homologous end joining a process used to repair DNA double strand breaks. Medium resolution structures both from cryoEM and X-ray crystallography show the general topology of the protein and positions of helices in parts of DNA-PKcs. EM-Fold, an algorithm developed for building protein models into medium resolution density maps has been used to generate models for the heat repeat-like "Ring structure" of the molecule. We were able to computationally corroborate placement of the N-terminus of the domain that supports a previously published hypothesis. Targeted experiments are suggested to test the model.


## 1. Introduction

Both endogenous and exogenous sources can lead to DNA double strand breaks which in turn can cause chromosome translocations and deletions (Biedermann et al., 1991; Kemp et al., 1984; Zdzienicka et al., 1988). Left unrepaired this can lead to cell death. DNA double strand breaks can be repaired through two mechanisms of which non-homologous end joining is the prevalent one in mammalian cells (Critchlow and Jackson, 1998). DNA dependent protein kinase (DNA-PK) is a central player in regulating non-homologous end joining. It is a heterotrimer holoenzyme built up of the DNA dependent protein kinase catalytic subunit (DNA-PKcs) and the heterodimer Ku70/Ku80. DNA-PKcs is a serine/threonine protein kinase and belongs to the phosphatidylinositol-3 (PI-3) kinase-like kinase (PIKK) superfamily (Hartley et al., 1995). The main purpose of DNA-PKcs is sensing and transmitting DNA damage signals (Anderson, 1993; Hoekstra, 1997). Structure determination of DNA-PKcs is crucial to understand its function and has proven difficult for many decades. Until recently only low resolution structural information based on cryoEM or electron crystallography has been available (Chiu et al., 1998; Leuther et al., 1999; Rivera-Calzada et al., 2005). Both a medium resolution cryoEM density map (Williams et al., 2008) and a medium resolution crystal structure (Sibanda et al., 2010) of the molecule have been determined within the past few years. While valuable structural information could be gleaned from these medium resolution structures neither was at sufficient resolution to trace the backbone of the molecule. Obtaining atomic detail structural information for DNA-PKcs remains a major challenge in the field.

EM-Fold is a software algorithm that folds proteins into medium resolution density maps obtained by cryoEM or X-ray crystallography (Lindert et al., 2009). It has been shown to be particularly efficient when density maps of highly helical proteins show clear density rods for helical sections of the protein. To use EM-Fold secondary structure elements (SSEs) have to be predicted from the proteins primary sequence and positions of density rods have to be identified. EM-Fold then uses a two-step protocol where predicted SSEs are placed into the density rods (assembly step) and the best assembled models are subsequently refined inside the density map. At this stage structures are transitioned into Rosetta to build missing loops and side chain and to further refine the models. In previous benchmarks EM-Fold has been demonstrated to work best on all helical proteins of sizes up to 350 amino acids (Lindert et al., 2009). Among the many benchmark cases one protein, 1OUV (Lüthy et al., 2004), had a classical heat repeat fold. Despite 1OUV being the largest protein in the benchmark set, the EM-Fold protocol was able to identify the correct fold and build a low RMSD model. We speculated that this is due at least in part to the heat repeat fold which translates into short loop lengths and a relatively low contact order compared to all other benchmark proteins. Further benchmarks (Lindert et al., 2012a,b) corroborated the notion that folding success increases with increased secondary structure content (i.e. short loop segments).

* Corresponding author at: Center for Structural Biology, 465 21st Ave South, BIOSCI/MRBIII, Room 5144B, Nashville, TN 37232-8725, USA. Tel.: +1 615 936 5662; fax: +1 615 936 2211.
*E-mail address:* jens.meiler@vanderbilt.edu (J. Meiler).
[1] Current address: Department of Pharmacology, UCSD, La Jolla, CA 92093, USA.
[2] Current address: Department of Pharmacology and Cleveland Center for Membrane and Structural Biology, Case Western Reserve University, Cleveland, OH 44106, USA.

1476-9271/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2012.11.001

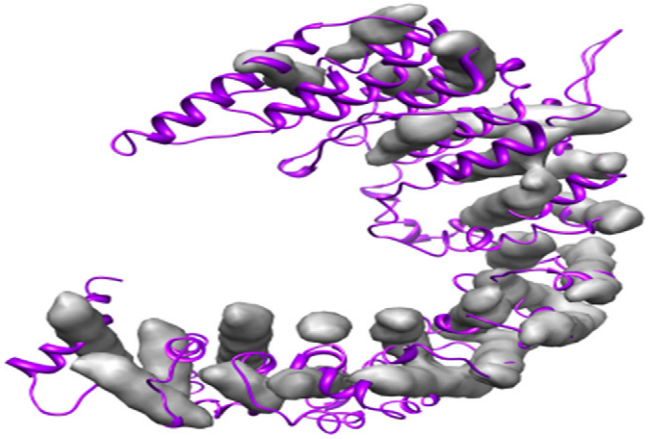

**Fig. 1.** Fit of Karyopherin β2 into density map.

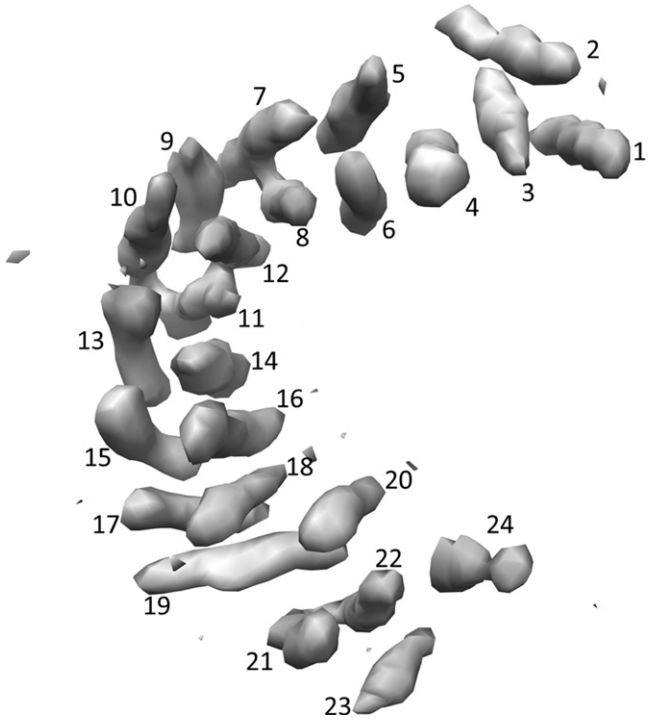


**Fig. 2.** Numbering of density rods of minimum size 13.5 Å from crystallographic density map.

## 2. Results and discussion

In this work EM-Fold's preference for heat repeat proteins was exploited to build a model for a region of DNA dependent protein kinase catalytic subunit (DNA-PKcs). The entire catalytic subunit contains 4128 residues and has about 135 predicted helices (68% of the sequence) which is approximately one order of magnitude too large for direct application of EM-Fold. However, the density maps clearly identify an extended heat repeat motive of 24 density rods, a region that was described as "Ring structure" in Sibanda et al. (2010). While previous benchmark only tested performance up to a size of 15 helices we are confident that heat repeat sequences of this size can be predicted using EM-Fold. The fact that it is unknown which exact part of the sequence corresponds to the ring structure poses a formidable challenge to the application of EM-Fold. To identify the sequence that corresponds to this part of the map, the entire sequence was submitted to Pfam (Finn et al., 2010). 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 (Kelley and Sternberg, 2009). 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. Fig. 1 shows the fit of Karyopherin β2 into the density map. Overall size, shape and curvature are identical while actual positions of helices differ.

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 (Sibanda et al., 2010). In the heat repeat region about 24 density rods of at least 13.5 Å in length are observed. These density rods are shown in Fig. 2 along with their sequential numbering in agreement with the crystal structure numbering. The density map that was originally calculated from the structure factors with the crystallographic CCP4 software package (Winn et al., 2011) does not contain perpendicular axes, rather it has cell axes of 90°, 105°, and 90°. While this is common for density maps derived from crystallographic data all cryoEM density maps have mutually perpendicular axes. Thus the cryoEM map readers of both EM-Fold and Rosetta would incorrectly process the map. A function OrthorgonalizeMap was implemented into the BCL 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 EM-Fold refinement step and the top scoring 100 refined topologies were transferred into Rosetta. These numbers are slightly higher than the benchmark number owing to the increased protein size.

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

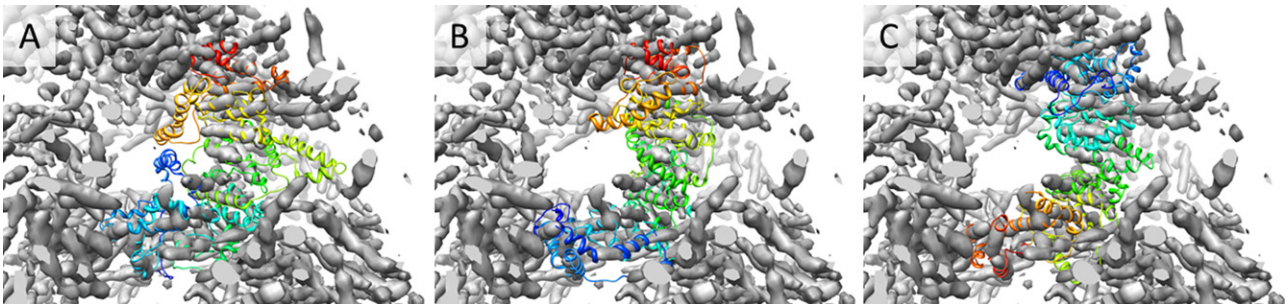


**Fig. 3.** Models representing the top scoring three topologies after EM-Fold and Rosetta refinement. (A and B) The top two scoring topologies have their N-terminal end in the lower part of the density toward the region referred to as the base. (C) The third best scoring topology has its N-terminal end in the upper part of the density.

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 (Lindert et al., 2009). 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.

The best scoring 21 models after Rosetta refinement fall into two topologies (Fig. 3A and B). These models predict that the N-terminus of this domain points to the "base" region of the molecule. The 22nd model shows an alternate placement of the N-terminus (Fig. 3C). While the superior score already favors the first orientation, we employed a confidence analysis that relies on a receiver operating characteristics (ROC) of repeated helix placement developed in (Lindert et al., 2009) to conclusively distinguish between the two orientations. The results indicate that the repeated placement of specific helices into specific density rods translates into a greater than 90% confidence that the N-terminus of this domain points to the "base" region of the molecule. For the three top scoring topologies Table 1 lists the placement of the predicted sequence into the density rods identified in Fig. 2. The atomic coordinates of the top three scoring topologies are provided as supplementary information. Fig. 4 shows a close-up view of the N-terminal part of the best scoring model within the density map. Length and shape of predicted helices and observed density rods are in good agreement. The overall agreement of model and map can be specified by a cross correlation coefficient of 0.69 between density map and a 5 Å resolution density map simulated from the model. This correlation coefficient was calculated using UCSF Chimera's 'Fit in Map' tool and represents good agreement between model and density map. Also the correlation coefficient of the top scoring topology is about 2.2% higher than that of the model with the reversed topology.

The structural models provided as supplementary information allow for the design of targeted experiments to test the predictions. We include the coordinates of the alternative placement only as it might help to design experiments that distinguish between the opposite directionalities. The results of the fold recognition along with the EM-Fold/Rosetta models suggest that sequence region 2900–3540 of DNA-PKcs could be amenable to expression, purification, and characterization as a separate domain. Then the structure

**Table 1**
Summary of placement of predicted helices into the density rods in the top three scoring topologies. Density rods are labeled according to Fig.2. The first residue number corresponds to placement of that part of the helix into the part of the rod that was identified in the crystal structure as N-terminal. Approximate sequence information for the helices placed in the rods is given. All numbers are based on folding residues 2700 and 3540 as a separate protein and can thus range from 1 to 841. To identify the residue information within the whole protein, 2699 residues have to be added to the numbers.

| Density rod | Topology 1 | Topology 2 | Topology 3 |
|---|---|---|---|
| 1 | 127–145 | 20–32 | 742–755 |
| 2 | 152–167 | 86–100 | 788–777 |
| 3 | 172–185 | 152–167 | 773–760 |
| 4 | 86–100 | 127–145 | 727–705 |
| 5 | 307–318 | 172–185 | 663–645 |
| 6 | 20–32 | 189–201 | 619–630 |
| 7 | 306–287 | 220–233 | 605–593 |
| 8 | 220–233 | 307–318 | 642–632 |
| 9 | 341–353 | 382–394 | 565–551 |
| 10 | 264–280 | 264–280 | 497–477 |
| 11 | 370–382 | 358–370 | 464–453 |
| 12 | 476–497 | 395–414 | 587–571 |
| 13 | 395–414 | 476–497 | 414–395 |
| 14 | 464–453 | 353–341 | 341–353 |
| 15 | 551–565 | 551–565 | 382–394 |
| 16 | 605–593 | 571–587 | 280–264 |
| 17 | 587–571 | 464–453 | 307–318 |
| 18 | 612–627 | 593–605 | 246–234 |
| 19 | 727–705 | 727–705 | 101–123 |
| 20 | 642–632 | 632–642 | 201–189 |
| 21 | 742–755 | 755–742 | 127–145 |
| 22 | 760–773 | 773–760 | 185–172 |
| 23 | 788–777 | 788–777 | 152–167 |
| 24 | 645–663 | 645–663 | 32–20 |

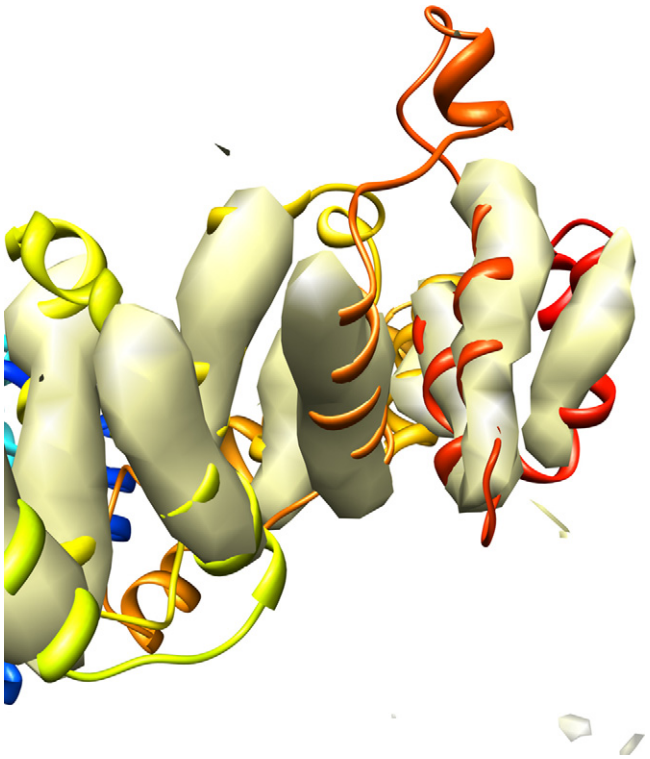

**Fig. 4.** Close-up view of C-terminal part of top scoring predicted model. Good agreement between the placed helices and the density rods is apparent.

of this smaller domain could be probed with circular dichroism to determine alpha-helical content and potentially with site-directed spin labeling electron paramagnetic resonance to confirm residues at the helix–helix interfaces. Possibly this separate domain can be crystallized to confirm the structures predicted with our folding protocol.

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. While being far from structure determination of the entire molecule these results underline that the medium resolution density map provided useful guidance during the modeling and that the generated models provide important testable hypotheses which may advance our structural understanding of the DNA dependent protein kinase catalytic subunit as well yield an improvement of computational methods for interpreting moderate resolution density maps.

## Acknowledgements

This research was supported by NIH grants to PLS (R01 CA140538) and JM (NSF 0742762, R01 GM080403).

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compbiolchem.2012.11.001.

## References

Anderson, C.W., 1993. DNA damage and the DNA-activated protein kinase. Trends in Biochemical Sciences 18, 433–437.

Biedermann, K.A., Sun, J.R., Giaccia, A.J., Tosto, L.M., Brown, J.M., 1991. Scid mutation in mice confers hypersensitivity to ionizing radiation and a deficiency in DNA double-strand break repair. Proceedings of the National Academy of Sciences of the United States of America 88, 1394–1397.

Chiu, C.Y., Cary, R.B., Chen, D.J., Peterson, S.R., Stewart, P.L., 1998. Cryo-EM imaging of the catalytic subunit of the DNA-dependent protein kinase. Journal of Molecular Biology 284, 1075–1081.

Critchlow, S.E., Jackson, S.P., 1998. DNA end-joining: from yeast to man. Trends in Biochemical Sciences 23, 394–398.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., et al., 2010. The Pfam protein families database. Nucleic Acids Research 38, D211–D222.

Hartley, K.O., Gell, D., Smith, G.C., Zhang, H., Divecha, N., et al., 1995. DNA-dependent protein kinase catalytic subunit: a relative of phosphatidylinositol 3-kinase and the ataxia telangiectasia gene product. Cell 82, 849–856.

Hoekstra, M.F., 1997. Responses to DNA damage and regulation of cell cycle checkpoints by the ATM protein kinase family. Current Opinion in Genetics and Development 7, 170–175.

Kelley, L.A., Sternberg, M.J., 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nature Protocols 4, 363–371.

Kemp, L.M., Sedgwick, S.G., Jeggo, P.A., 1984. X-ray sensitive mutants of Chinese hamster ovary cells defective in double-strand break rejoining. Mutation Research 132, 189–196.

Leuther, K.K., Hammarsten, O., Kornberg, R.D., Chu, G., 1999. Structure of DNA-dependent protein kinase: implications for its regulation by DNA. The EMBO Journal 18, 1114–1123.

Lindert, S., Staritzbichler, R., Wotzel, N., Karakas, M., Stewart, P.L., et al., 2009. EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. Structure 17, 990–1003.

Lindert, S., Alexander, N., Wotzel, N., Karakas, M., Stewart, P.L., et al., 2012a. EM-Fold: De Novo atomic-detail protein structure determination from medium-resolution density maps. Structure 20, 464–478.

Lindert, S., Hofmann, T., Wotzel, N., Karakas, M., Stewart, P.L., et al., 2012b. Ab initio protein modeling into cryoEM density maps using EM-Fold. Biopolymers.

Lüthy, L., Grütter, M.G., Mittl, P.R.E., 2004. The crystal structure of helicobacter cysteine-rich protein C at 2.0 Å resolution: similar peptide-binding sites in TPR and SEL1-like repeat proteins. Journal of Molecular Biology 340, 829–841.

Rivera-Calzada, A., Maman, J.D., Spagnolo, L., Pearl, L.H., Llorca, O., 2005. Three-dimensional structure and regulation of the DNA-dependent protein kinase catalytic subunit (DNA-PKcs). Structure 13, 243–255.

Sibanda, B.L., Chirgadze, D.Y., Blundell, T.L., 2010. Crystal structure of DNA-PKcs reveals a large open-ring cradle comprised of HEAT repeats. Nature 463, 118–121.

Williams, D.R., Lee, K.J., Shi, J., Chen, D.J., Stewart, P.L., 2008. Cryo-EM structure of the DNA-dependent protein kinase catalytic subunit at subnanometer resolution reveals alpha helices and insight into DNA binding. Structure 16, 468–477.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., et al., 2011. Overview of the CCP4 suite and current developments. Acta Crystallographica. Section D: Biological Crystallography 67, 235–242.

Zdzienicka, M.Z., Tran, Q., van der Schans, G.P., Simons, J.W., 1988. Characterization of an X-ray-hypersensitive mutant of V79 Chinese hamster cells. Mutation Research 194, 239–249.