Protein Cell 2010, 1(5): 453–458
DOI 10.1007/s13238-010-0045-7

COMMUNICATION

Crystal structure of a novel non-Pfam protein PF2046 solved using low resolution B-factor sharpening and multi-crystal averaging methods

Jing Su[1*], Yang Li[1*], Neil Shaw[1*], Weihong Zhou[2], Min Zhang[3], Hao Xu[4], Bi-Cheng Wang[4], Zhi-Jie Liu[1] ✉

[1] National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China
[2] College of Life Sciences, Nankai University, Tianjin 300071, China
[3] School of Life Sciences, Anhui University, Hefei 230039, China
[4] Department of Biochemistry and Molecular Biology, University of Georgia, Atlanta, GA 30605, USA
✉ Correspondence: zjliu@ibp.ac.cn
Received February 5, 2010 Accepted March 18, 2010

ABSTRACT

Sometimes crystals cannot diffract X-rays beyond 3.0 Å resolution due to the intrinsic flexibility associated with the protein. Low resolution diffraction data not only pose a challenge to structure determination, but also hamper interpretation of mechanistic details. Crystals of a 25.6 kDa non-Pfam, hypothetical protein, PF2046, diffracted X-rays to 3.38 Å resolution. A combination of Se-Met derived heavy atom positions with multiple cycles of B-factor sharpening, multi-crystal averaging, restrained refinement followed by manual inspection of electron density and model building resulted in a final model with a R value of 23.5 (R_{free}= 24.7). The asymmetric unit was large and consisted of six molecules arranged as a homodimer of trimers. Analysis of the structure revealed the presence of a RNA binding domain suggesting a role for PF2046 in the processing of nucleic acids.

KEYWORDS low resolution diffraction, PF2046, B-factor sharpening, a homodimer of trimers

INTRODUCTION

A number of proteins from different genomes are annotated as hypothetical proteins (Brenner, 2000). The primary amino acid sequence of such proteins does not contain any information that could potentially provide clues for their classification into Pfam families (Bateman et al., 2004) or their function. One way of determining the function of such proteins is to solve the 3-dimensional structure of the protein and compare it with structures of proteins with known function deposited in Protein Data Bank (PDB) (Berman et al., 2000). Depending on the match, the structure is then searched for motifs or signature catalytic residues in order to narrow down on the function. Such an approach provides a good starting point for designing functional experiments to support the structural evidence in order to determine the function of hypothetical proteins (Brenner and Levitt, 2000; Chandonia and Brenner, 2005).

Pyrococcus furiosus is an anaerobic hyperthermophile, which has an optimal temperature of growth at 100°C. ORF PF2046 is a 25.6 kDa non-Pfam, hypothetical protein from *Pyrococcus furiosus*. The primary amino acid sequence does not provide any clue to its function. We decided to determine the structure of this protein to obtain clues about its function. The structure solved at 3.4 Å resolution using the anomalous signal of selenium reveals that PF2046 has a RNA binding domain and could be involved in the processing of nucleic acids.

RESULTS

The amino acid sequence of PF2046, when subjected to a Wu-Blast (Lopez et al., 2003) search of PDB for structural homologs, did not yield any structures with significant

*These authors contributed equally to this work.

homology. Therefore, a Se-Met derivative of the protein was used to solve the phases. All the three methionines of PF2046 could be replaced by Se-Met and the anomalous signal from 18 selenium atoms in ASU could be detected. The ASU consists of six molecules of PF2046 with a solvent content of 67%. Interestingly, the protein seems to have crystallized as a homodimer of trimers, with the trimers within the dimer related by 42.3 Å translation along NCS 3-fold axis of the trimer and 15.8° rotation around the same axis. To find out the oligomeric state of the protein in solution, analytical ultracentrifugation studies were carried out. The results suggest that PF2046 assembles into a trimer in solution. Initial attempts to solve the structure of PF2046 failed because the resolution of the diffraction data was low and the ASU was large with 1422 amino acids. Extensive purification, chemical modification (Rayment, 1997; Walter et al., 2006; Shaw et al., 2007) and coverage of a larger crystallization space during screening and optimization could not improve the resolution beyond 3.38 Å. Incorporation of Se-Met in the protein not only helped obtain the phase, but the heavy atom positions also served as useful guides during the tracing of the map. Initially only half of the total residues could be traced automatically by Autobuild (Perrakis et al., 1999), and in addition, the quality of the map was poor. The Wilson plot indicated a B factor of 89.0 for the original Se-Met data. Therefore, B-factor sharpening was used to improve the map (Pannu et al., 1998; Bass et al., 2002; DeLaBarre and Brunger, 2003). The program CAD of the CCP4 package was employed for sharpening the B-factors. An artificial minus B-factor was imposed on the original mtz file resulting in a B factor of 20 and enhancement of the electron-density map (Fig. 1). In addition to B-factor sharpening, multi-crystal averaging (Chen et al., 2005) with another data set collected at the Se edge gave a 12-fold averaging, which resulted in a significant improvement of the map and the model. A Dmmulti (Cowtan and Zhang, 1999) script run with 1000 cycles helped trace a number of side chains. Coot was used to inspect the electron density and manually build parts of the model (Emsley and Cowtan, 2004). A combination of iterative cycles of refinement (Phenix.refine) (Adams et al., 2002), B-factor sharpening, multi-crystal averaging and manual model building resulted in a final model with a R value of 23.5 (R_{free}= 24.7) and refinement statistics listed in Table 1.

PF2046 is annotated as a conserved hypothetical protein. A PSI-Blast (Altschul et al., 1997) search using the amino acid sequence failed to retrieve any homologous proteins with known function. In addition, the primary sequence of PF2046 does not contain any known conserved domains. To gain clues about the function, we analyzed the structure of PF2046 using DALI (Holm et al., 2008) for structural homology with structures deposited in PDB (Berman et al., 2000). The list of top 20 structural matches consisted exclusively of ribonucleases, in particular RNase H. The best match was RNase HI from *Sulfolobus* (PDB code 2EHG) with a Z score of 7.6 for an overlap of 124 amino acids out of 149, with an r.m.s.d. value of 3.6 Å and a sequence identity of 12%. A ProFunc analysis (Laskowski et al., 2005) gave similar results—the secondary structural elements of PF2046 are similar to RNase H. Therefore, PF2046 could possibly function as a RNase H. However, this hypothesis needs to be verified experimentally with functional assays.

CATH server classified the protein as alpha beta with a 2-layer sandwich architecture and topology similar to double stranded RNA binding domain (Pearl et al., 2004). The overall structure consists of 3 helices and 3 sheets. Strands β1, β2, β3 and β4 of sheet 1 and strands β7, β8, β9, β10, β11, and β12 from sheet 3 seem to be encircling helix α3 (Fig. 2A). Interestingly, helix α2 is bent by 110° at the center with one end protruding out of the protein. Within the trimer, the protrusions look like a clamp covering a positively charged cavity at the center (Fig. 2B and 2C).

DISCUSSION

Although the primary amino acid sequence did not provide any clue to the function of PF2046, analysis of the structure of PF2046 by DALI, Profunc and CATH indicated that the secondary structural elements of PF2046 are similar to a RNase HI (Katayanagi et al., 1990; Davies et al., 1991; Ohtani et al., 2004). We analyzed the structure further for signature motifs and catalytic residues necessary for the processing of nucleic acids. PF2046 has a basic protrusion, which is used

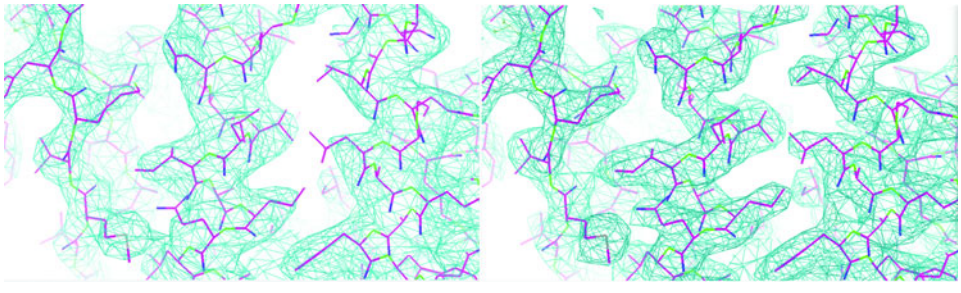

Figure 1. Improvement of electron density by B factor sharpening. Protein:solvent boundaries were clear and there was an obvious improvement in the density for chains after B factor sharpening. Map before (left panel) and after (right panel) B factor sharpening.

Table 1 Data collection and refinement statistics

	2046_Peak	2046_INF
data collection	SER-CAT, APS	SER-CAT, APS
wavelength (Å)	0.9794	0.9794
space group	C 2	C 2
cell dimensions		
a, *b*, *c* (Å)	216.43, 125.10, 89.98	216.52, 125.17, 90.18
α, β, γ (º)	90.0, 101.88, 90.0	90.0, 101.90, 90.0
resolution (Å)	3.38	3.50
R_{sym} or R_{merge}	0.104 (0.441)	0.094 (0.475)
$I/\sigma I$	15.94 (2.83)	15.33 (2.57)
completeness (%)	99.3 (99.3)	99.4(99.3)
redundancy	4.8 (4.8)	4.8 (4.8)
refinement		
resolution (Å)	3.38	
No. reflections	34164	
R_{work}/R_{free}	23.51/ 24.76	
No. atoms		
protein	10149	
water	0	
R.m.s deviations		
bond lengths (Å)	0.038	
bond angles (º)	3.250	
mean B value (Å^2)	98.82	
ramachandran analysis		
favored region (%)	83.64	
allowed region (%)	16.13	
outliers (%)	0.23	

Numbers in parentheses are statistics for the highest resolution shell.

by most of the RNase HI enzymes to bind nucleic acids (Fig. 3A). In addition, the quaternary structure of PF2046 is similar to the RNase H domain of HIV reverse transcriptase (Davies et al., 1991). Especially, the tunnel observed in the tertiary structure of PF2046 is similar in architecture to the one seen in HIV reverse transcriptase used for binding double stranded nucleic acids (Fig. 3). Furthermore, RNase HI catalyzes hydrolysis using three aspartic acids and a glutamic acid. The tertiary structure of PF2046 reveals an identical motif formed by an aspartate (Asp197) contributed by each monomer within a trimer and a glutamate residue, Glu200, located at the center of the tunnel within a highly basic patch (Fig. 3A). The carboxylic acids are known to bind metal ions in RNase HI (Katayanagi et al., 1990; Davies et al., 1991; Ohtani et al., 2004). In addition to the carboxylic acids, a histidine residue participates in the catalysis. This histidine residue is conserved in the human and *E. coli* RNase HI and is also found at a similar position in HIV-1 RNase H. The histidine is replaced by an arginine in the archaeal RNase HI from *Solfolobus tokodaii* (Ohtani et al., 2004). Interestingly, an arginine, Arg227, occupies the position of the histidine in the structure of PF2046, suggesting that the archaeal RNase HI prefers an arginine at this position (Fig. 3A and 3B). Structure based evidences suggest that PF2046 has a catalytic machinery similar to RNase HI (Fig. 3A and 3C). A number of proteins like integrases (Dyda et al., 1994), transposases (Rice and Mizuuchi, 1995), and resolvases (Ariyoshi et al., 1994) carry out functions using similar catalytic residues. Further functional studies are warranted to elucidate the exact function and physiologic role of PF2046.

MATERIALS AND METHODS

Protein production

The gene for *ORF PF2046* (NCBI gene ID 1469931) was amplified from the genomic DNA of *Pyrococcus furiosus* and cloned into pET28b vector. *E. coli* BL21 (DE3) was freshly transformed with

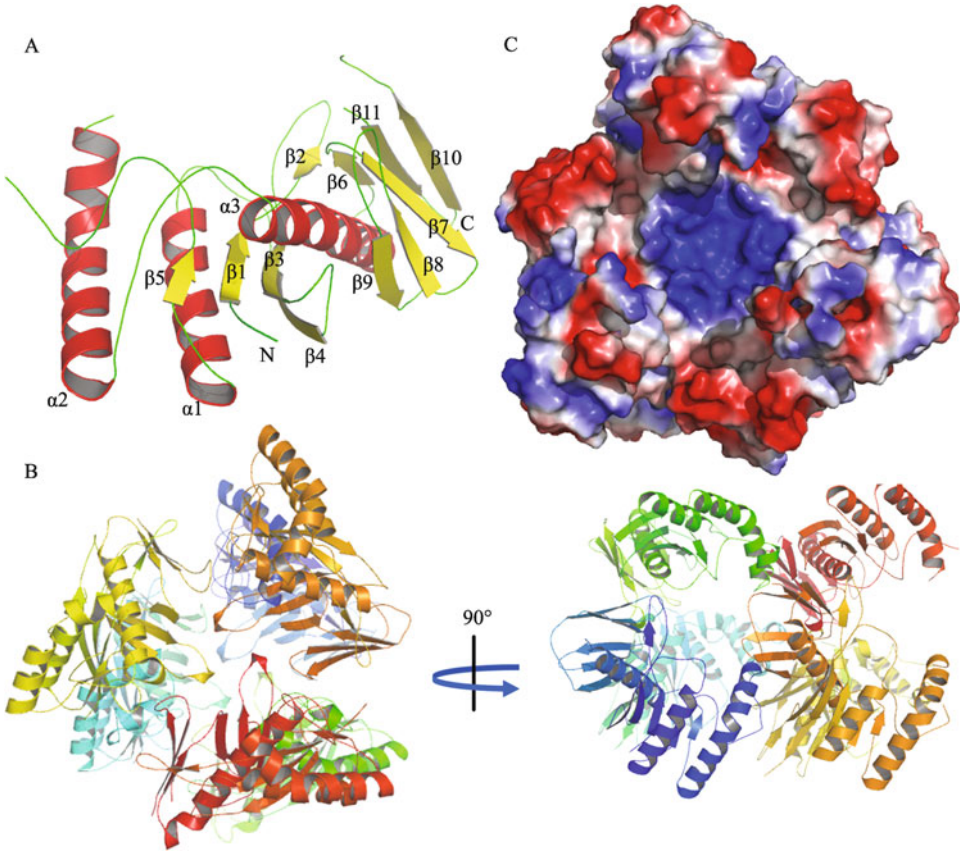

Figure 2. Overall structure of PF2046. (A) A cartoon representation of the monomer of PF2046. The N and C terminals of the protein are labeled as N and C, respectively. (B) Quaternary structure of PF2046 showing the secondary structural elements arranged as a homodimer of trimers. Each monomer is represented in a different color. (C) A surface electrostatic potential representation of the PF2046 structure. Positive potential is colored blue, negative potentials are colored red.

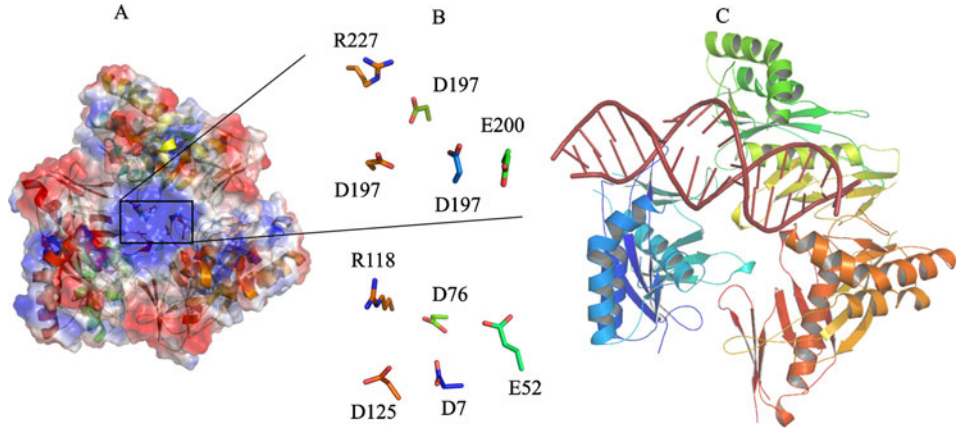

Figure 3. Putative active site. (A) A basic cavity at the centre of PF2046 contains amino acids critical for the processing of nucleic acids similar to those seen in RNase HI. Putative catalytic amino acids of PF2046 (shown in surface representation) have been enlarged and depicted as sticks. (B) Catalytic amino acids of RNase HI from *Sulfolobus* (PDB code 2EHG) shown as sticks. (C) The double stranded nucleic acid ligand (salmon color) of HIV-1 RNase H (PDB code 1HYS) was superimposed over the putative active site of PF2046 using the CCP4 molecular graphics software. A trimer of PF2046 is shown as cartoon.

plasmid containing a PF2046 gene. N-terminal hexa-histidine tagged protein was produced by growing cells at 37°C until culture density reached OD_{600nm} 0.8. The culture was cooled down to 16°C and induced with 0.2 mM IPTG for 20 h. Seleno-methionine labeled

protein was produced by transforming the plasmid into *E. coli* B834 (DE3) and growing the cells in M9 media supplemented with 40 mg/L of seleno-L-methionine. Cells were harvested by centrifugation and lysed by sonication. Unbroken cells and debris were removed by centrifugation and the clarified supernatant was subjected to a heat treatment at 70°C for 1 h during which most of the *E. coli* protein precipitated. After centrifugation, the recombinant PF2046 was purified from the supernatant by Ni-affinity chromatography. His-tag was cleaved by treating the protein with thrombin at 30°C for 2 h. Uncut protein was removed by a second round of Ni-affinity chromatography. The protein was further purified by gel filtration using a Superdex G75 HR column equilibrated with 20 mM Tris-HCl, pH 8.0, 200 mM NaCl, and 1 mM DTT. Fractions containing the protein were pooled and concentrated to 10 mg/mL before setting up crystallization drops.

Crystallization and data collection

PF2046 was set up for crystallization immediately after purification. Crystallization was carried out in hanging drop vapor diffusion. Commercially available sparse matrix screens were used to screen crystallization space. Two microliter crystallization drops containing 1 μL protein mixed with 1 μL mother liquor were equilibrated over 300 μL reservoir solution and incubated at 16°C. Crystals were obtained at a number of conditions within a week. Tetragonal crystals were formed in a crystallization solution containing 20% PEG 4000, 20% (*v/v*) 2-propanol, 0.1 M sodium citrate. 0.1 M sodium acetate pH 5.0, 0.1 M sodium chloride, 10% (*w/v*) MPD.

Crystals were frozen in liquid nitrogen prior to diffraction testing and data collection. Diffraction data of the selenium-labeled crystal were collected at beamline 22-ID, APS, Argonne National Laboratory. Data were indexed and scaled to 3.38 Å resolution using HKL2000 (Otwinowski and Minor, 1997). The statistics of the data are listed in Table 1.

Phasing and refinement

The phases were determined using Se-Met peak data by SAD method at 3.38 Å resolution. Due to the limitation of low resolution and data quality, the phases and electron density map are suboptimal. The phases were improved by the combination of B-factor sharpening and multi-crystal averaging between Se-Met peak data and Se-Met edge data. Several cycles of B-factor sharpening, multi-crystal averaging, restrained refinement, and manual model building resulted in a final model with an R value of 23.5 (R_{free}= 24.7). The asymmetric unit was large and consisted of six molecules (1422 amino acids) arranged as a homodimer of trimers. Details of data collection and refinement statistics are listed in Table 1.

Analytical ultracentrifugation

Analytical sedimentation velocity experiments were carried out using a ProteomeLab™ XL-I protein characterization system (Beckman Coulter). An-60Ti rotor was used to centrifuge a 1 mg/mL protein sample suspended in 20 mM Tris-HCl, pH 7.5, 150 mM NaCl, at 60,000 rpm. Absorbance was read at 280 nm. A set of 93 scans were collected at 1 min intervals. Data were analyzed using Sedfit software.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yong Xiong at Yale University for the help and suggestions on the B-factor sharpening method. This work was funded by the Ministry of Science and Technology of China (Grant Nos. 2006AA02A316, 2009DFB30310 and 2006CB910901), the National Natural Science Foundation of China (Grants Nos. 30670427 and 30721003), the Ministry of Health of China (Grant No. 2008ZX10404), CAS Research Grant (No. KSCX2-YW-R-127 and INFO-115-D01-2009). Crystallographic data were collected at SECCAT beamline 22-ID at the Advanced Photon Source, Argonne National Laboratory.

ABBREVIATIONS

APS, Advanced Photon Source; ASU, asymmetric unit; MPD, 2-methyl- 2,4-pentanediol; NCS, noncrystallographic symmetry; PF, *Pyrococcus furiosus*; Se-Met, seleno-methionine

REFERENCES

Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., Sauter, N.K., and Terwilliger, T.C. (2002). PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr D Biol Crystallogr 58, 1948–1954.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402.

Ariyoshi, M., Vassylyev, D.G., Iwasaki, H., Nakamura, H., Shinagawa, H., and Morikawa, K. (1994). Atomic structure of the RuvC resolvase: a holliday junction-specific endonuclease from E. coli. Cell 78, 1063–1072.

Bass, R.B., Strop, P., Barclay, M., and Rees, D.C. (2002). Crystal structure of Escherichia coli MscS, a voltage-modulated and mechanosensitive channel. Science 298, 1582–1587.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., *et al*. (2004). The Pfam protein families database. Nucleic Acids Res 32, D138–141.

Berman, H.M., Bhat, T.N., Bourne, P.E., Feng, Z., Gilliland, G., Weissig, H., and Westbrook, J. (2000). The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol 7 Suppl, 957–959.

Brenner, S.E. (2000). Target selection for structural genomics. Nat Struct Biol 7, 967–969.

Brenner, S.E., and Levitt, M. (2000). Expectations from structural genomics. Protein Sci 9, 197–200.

Chandonia, J.M., and Brenner, S.E. (2005). Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. Proteins 58, 166–179.

Chen, B., Vogan, E.M., Gong, H., Skehel, J.J., Wiley, D.C., and Harrison, S.C. (2005). Determining the structure of an unliganded and fully glycosylated SIV gp120 envelope glycoprotein. Structure 13, 197–211.

Cowtan, K.D., and Zhang, K.Y. (1999). Density modification for macromolecular phase improvement. Prog Biophys Mol Biol 72, 245–270.

Davies, J.F. 2nd, Hostomska, Z., Hostomsky, Z., Jordan, S.R., and Matthews, D.A. (1991). Crystal structure of the ribonuclease H domain of HIV-1 reverse transcriptase. Science 252, 88–95.

DeLaBarre, B., and Brunger, A.T. (2003). Complete structure of p97/valosin-containing protein reveals communication between nucleotide domains. Nat Struct Biol 10, 856–863.

Dyda, F., Hickman, A.B., Jenkins, T.M., Engelman, A., Craigie, R., and Davies, D.R. (1994). Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. Science 266, 1981–1986.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 60, 2126–2132.

Holm, L., Kääriäinen, S., Rosenström, P., and Schenkel, A. (2008). Searching protein structure databases with DaliLite v.3. Bioinformatics 24, 2780–2781.

Katayanagi, K., Miyagawa, M., Matsushima, M., Ishikawa, M., Kanaya, S., Ikehara, M., Matsuzaki, T., and Morikawa, K. (1990). Three-dimensional structure of ribonuclease H from *E. coli*. Nature 347, 306–309.

Laskowski, R.A., Watson, J.D., and Thornton, J.M. (2005). ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33, W89–93.

Lopez, R., Silventoinen, V., Robinson, S., Kibria, A., and Gish, W. (2003). WU-Blast2 server at the European Bioinformatics Institute. Nucleic Acids Res 31, 3795–3798.

Ohtani, N., Yanagawa, H., Tomita, M., and Itaya, M. (2004). Cleavage of double-stranded RNA by RNase HI from a thermoacidophilic archaeon, Sulfolobus tokodaii 7. Nucleic Acids Res 32, 5809–5819.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol 276, 307–326.

Pannu, N.S., Murshudov, G.N., Dodson, E.J., and Read, R.J. (1998). Incorporation of prior phase information strengthens maximum-likelihood structure refinement. Acta Crystallogr D Biol Crystallogr 54, 1285–1294.

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., *et al*. (2004). The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res 33, D247–251.

Perrakis, A., Morris, R., and Lamzin, V.S. (1999). Automated protein model building combined with iterative structure refinement. Nat Struct Biol 6, 458–463.

Rayment, I. (1997). Reductive alkylation of lysine residues to alter crystallization properties of proteins. Methods Enzymol 276, 171–179.

Rice, P., and Mizuuchi, K. (1995). Structure of the bacteriophage Mu transposase core: a common structural motif for DNA transposition and retroviral integration. Cell 82, 209–220.

Shaw, N., Cheng, C., Tempel, W., Chang, J., Ng, J., Wang, X.-Y., Perrett, S., Rose, J., Rao, Z., Wang, B.-C., *et al*. (2007). (NZ)CH...O contacts assist crystallization of a ParB-like nuclease. BMC Struct Biol 7, 46–58.

Walter, T.S., Meier, C., Assenberg, R., Au, K.-F., Ren, J., Verma, A., Nettleship, J.E., Owens, R.J., Stuart, D.I., and Grimes, J.M. (2006). Lysine methylation as a routine rescue strategy for protein crystallization. Structure 14, 1617–1622.

