PROTEINS: Structure, Function, and Bioinformatics 63:711–715 (2006)

STRUCTURE NOTE

Crystal Structure of Intein Homing Endonuclease II Encoded in DNA Polymerase Gene From Hyperthermophilic Archaeon *Thermococcus kodakaraensis* Strain KOD1

Hiroyoshi Matsumura,[1] Hitomi Takahashi,[1] Tsuyoshi Inoue,[1] Takahiko Yamamoto,[1] Hiroshi Hashimoto,[2] Motomu Nishioka,[3] Shinsuke Fujiwara,[4] Masahiro Takagi,[5] Tadayuki Imanaka,[6] and Yasushi Kai[1]*

[1]*Department of Materials Chemistry, Graduate School of Engineering, Osaka University, Suita, Osaka, Japan*
[2]*Supramolecular Biology, International Graduate School of Arts and Sciences, Yokohama City University, Yokohama, Kanagawa, Japan*
[3]*Department of Chemical Science and Engineering, Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka, Japan*
[4]*Department of Bioscience, Nanobiotechnology Research Center, School of Science and Technology, Kwansei Gakuin University, Sanda, Hyogo, Japan*
[5]*Department of Chemical Materials Science, School of Materials Science, Japan Advanced Institute of Science and Technology, Tatsunokuchi, Ishikawa, Japan*
[6]*Department of Synthetic Chemistry and Biological Chemistry, Graduate School of Engineering, Kyoto University, Nishikyo-ku, Kyoto, Japan*

Introduction. Intein homing endonucleases are bifunctional proteins catalyzing both protein splicing and site-specific DNA double-strand cleavage.[1–3] Protein splicing is a posttranslational process involving precise excision of an intervening protein domain, termed an intein. An intein often exhibits site-specific endonuclease activity, which recognizes and cleaves the DNA sequence lacking its coding DNA sequence. The recognition sequences are usually asymmetrical and 12- to 40-bp long.[3]

Herein, we report the crystal structure of PI-TkoII, an intein endonuclease II from the hyperthermophilic archaeon *Thermococcus kodakaraensis* strain KOD1. PI-TkoII is a product of the polA gene: mature KOD DNA polymerase.[4] To date, crystal structures of intein homing endonucleases have only been described for PI-SceI from *Saccharomyces cerevisiae*,[5,6] GyrA intein in bacterial gyrase A subunit from *Mycobacterium xenopi*,[7] and PI-PfuI from *Pyroccocus furiosus*.[8] PI-TkoII shares a low sequence similarity to the other inteins including PI-SceI (21.7% identity in 411 amino acids overlap), GyrA intein (34.4% identity in 61 amino acids overlap), and PI-PfuI (22.7% in 322 amino acids overlap). Although the only structure of archaeal intein has been reported for PI-PfuI, molecular masses are very different between PI-TkoII and PI-PfuI, with 62 and 53 kDa, respectively. The minimal recognition sequence for PI-TkoII involves a 16-bp fragment (5′-CAGCTACTACGGTTAC-3′),[9] which is relatively short compared with other intein homing endonucleases. Structural information on PI-TkoII provides new insights for mechanisms involved in specific endonuclease activity. In the present study, we also discuss similarities and differences in domain architecture between PI-TkoII, PI-PfuI, and PI-SceI.

Results and Discussion. Crystal structures of native and SeMet PI-TkoII were solved at 2.7- and 2.5-Å resolution, respectively (Table I). Because no obvious structural differences between the native and SeMet PI-TkoII were observed, we describe here crystal structure of SeMet PI-TkoII at 2.5-Å resolution.

The final model includes 537 residues, 279 water molecules, six glycerols, and eight sulfates. According to the Structural Classification of Proteins database (SCOP),[10] PI-TkoII is a member of the "Hedgehog/intein (Hint) domain" fold. The molecule has an overall size of approximately 100 × 65 × 40 Å, and consists of four distinct domains [Fig. 1(A)]: the endonuclease domain (Endo: residues 273–432, green), the Hint domain (Hint: residues 1–127 and 496–537, dark blue), domain III (III: residues 433–495, light blue), and domain IV (IV: residues 128–272, pink). Structural comparisons showed that structures of endonuclease and Hint domains were mainly conserved among PI-TkoII, PI-PfuI, and PI-SceI, whereas domains III and IV were not [Fig. 1(A–D)]. Domain III is located at

Abbreviations: PI-TkoII, an intein endonuclease II from the hyperthermophilic archaeon *Thermococcus kodakaraensis* strain KOD1; PI-SceI, intein endonuclease from *S. cerevisiae*; PI-PfuI intein endonuclease from *Pyroccocus furiosus*; MAD, multiwavelength anomalous dispersion; SeMet, selenomethionine.

*Correspondence to: Yasushi Kai, Department of Materials Chemistry, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan. E-mail: kai@chem.eng.osaka-u.ac.jp

Received 14 July 2005; Accepted 20 October 2005

Published online 21 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20858

TABLE I. Summary of the Crystallographic Data

	Peak	Edge	Remote	Native
Data collection statistics				
Wavelength (Å)	0.9792	0.9794	0.9686	1.0
Unit cell (Å)		$a = 103.97$, $b = 150.54$, $c = 145.49$		$a = 107.70$, $b = 149.06$, $c = 146.13$
Space group	$C222_1$			
Resolution range (Å)	40–2.5 (2.59–2.5)	40–2.5 (2.59–2.5)	40–2.5 (2.59–2.5)	40–2.7 (2.8–2.7)
R_{merge} (%)[a,b]	6.2 (24.2)	6.4 (25.2)	4.8 (17.8)	4.6 (25.3)
Completeness (%)	91.5 (76.8)	91.7 (76.9)	88.5 (66.1)	93.2 (79.5)
Unique reflections	70,178	70,331	67,877	30,434
$I/\sigma(I)$	9.0 (2.8)	10.2 (2.9)	11.0 (2.9)	13.4 (1.7)
Refinement statistics				
Resolution range (Å)			40–2.5	40–2.7
No. of reflections			67,647	30,416
R_{cryst} (%)/R_{free} (%)[c,d]			21.7/25.2	20.6/24.3
RMSD bond length (Å)			0.007	0.008
RMSD bond angle (°)			1.4	1.3
Protein atoms			4,392	4,392
Water molecules			279	19
Glycerols			6	0
Sulfates			8	3
Ramachandran plot (%)[e]				
Favored			87.8	86.9
Allowed			10.3	12.2

[a]Values in parentheses-are for the highest resolution shell.
[b]Rmerge $= \Sigma|I - \langle I\rangle|/\Sigma I$, where I is the intensity of observation I and $\langle I\rangle$ is the mean intensity of the reflection.
[c]$R_{cryst} = \Sigma||F_0| - |/\Sigma|$ where Fo and Fc are the observed and calculated structure factor amplitudes, respectively.
[d]R_{free} was calculated using a randomly selected 5% of the data set that was omitted through all stages of refinement.
[e]Ramachandran plot was performed for all residues other than Gly and Pro.

the corresponding position of Stirrup domain in PI-PfuI, and at DRR (DNA recognition region) in PI-SceI. However, it shares no structural similarities with the Stirrup domain of PI-PfuI and DRR domain of PI-SceI [Fig. 1(A–C)]. Moreover, PI-TkoII contains an additional domain depicted as domain IV in Figure 1(A), which was missing in the structures of PI-PfuI and PI-SceI [Fig. 1(B and C)].

Despite its low sequence similarity, the structure of the endonuclease domain of PI-TkoII is similar to those of the other homing endonucleases. The size of the domain is approximately 60 × 30 × 25 Å. A search for protein folds using the coordinates of endonuclease domain and protein coordinates in the Protein Data Bank (PDB) with the program DALI[11] revealed that 96 proteins were structurally most similar (with a Z-score of 2.0 or higher) to the endonuclease domain of PI-TkoII. The three proteins with the highest structural similarity were: the intron-encoded I-DmoI from *Desulfurococcus mobilis*[12] [2.1 Å of root-mean-square deviation (RMSD) over 150 residues with 25% sequence identity; PDB code: 1B24], PI-SceI[5] (3.3 Å of RMSD over 153 residues with 20% sequence identity; PDB code: 1VDE), and PI-PfuI[8] (2.4 Å of RMSD over 103 residues with 17% sequence identity; PDB code: 1DQ3). The PI-TkoII endonuclease domain contains two subdomains, which are related to each other by a pseudo two-fold axis. Each subdomain contains an antiparalleled β-sheet motif named LAGLIDADG, depicted as Block C and Block E in Figure 1(D). The catalytically essential residues (Glu288 and Asp383, respectively) are conserved in these motifs [Fig. 1(A)]. The three β-sheets comprising residues 289–308, 330–338, and 417–429 protrude from the molecular surface to form a clustered negative charge surface made by Arg292, Arg295, Lys298, Lys306, Arg330, Arg333, Lys418, and Arg427. These protruded β-sheets are also found in PI-SceI and PI-PfuI, but they vary considerably in length, curvature, and amino acid sequences. These β-sheets, with nonsymmetrical structure may promote insertion into DNA grooves to recognize nonsymmetrical DNA duplexes.

The Hint domain is mainly composed of antiparalleled β-sheets, and is approximately 45 × 40 × 25 Å. According to DALI server,[11] five proteins were found to have a similar structure (with a Z-score of 2.0 or higher) in their Hint domain. Proteins with the highest structural similarity were PI-PfuI[8] (1.8 Å of RMSD over 164 residues with 23% sequence identity; PDB code: 1DQ3), DnaB intein from *Synechocystis sp*.[13] (1.7 Å of RMSD over 135 residues with 27% sequence identity; PDB code: 1MI8), GyrA intein[7] (2.7 Å of RMSD over 168 residues with 23% sequence identity; PDB code: 1AM2), the 17-kDa fragment of the Hedgehog domain[14] (2.7 Å of RMSD over 168 residues with 23% sequence identity; PDB code: 1AM2), and PI-SceI[5] (2.4 Å of RMSD over 129 residues with 13% sequence identity; PDB code: 1VDE). The Hint domain has a typical two subdomain structure, which exhibits a symmetrical architecture. The active site for protein splicing is located at the center of the horseshoe-shaped Hint domain, which contains highly invariant residues such as Ser1, His96, His536, and Asn537 [Fig. 1(A)]. These observations suggest that mechanisms of protein splicing are highly conserved among this family.

In contrast, domain III has no structural similarity to the other known protein structures. PI-PfuI and PI-SceI contain DNA binding domains, named Stirrup and DRR domains, respectively; and domain III is located at positions corresponding to Stirrup and DRR domains [Fig. 1(A–C)]. Domain III is likely to be involved in DNA recognition. The exposed concaved surface found in do-

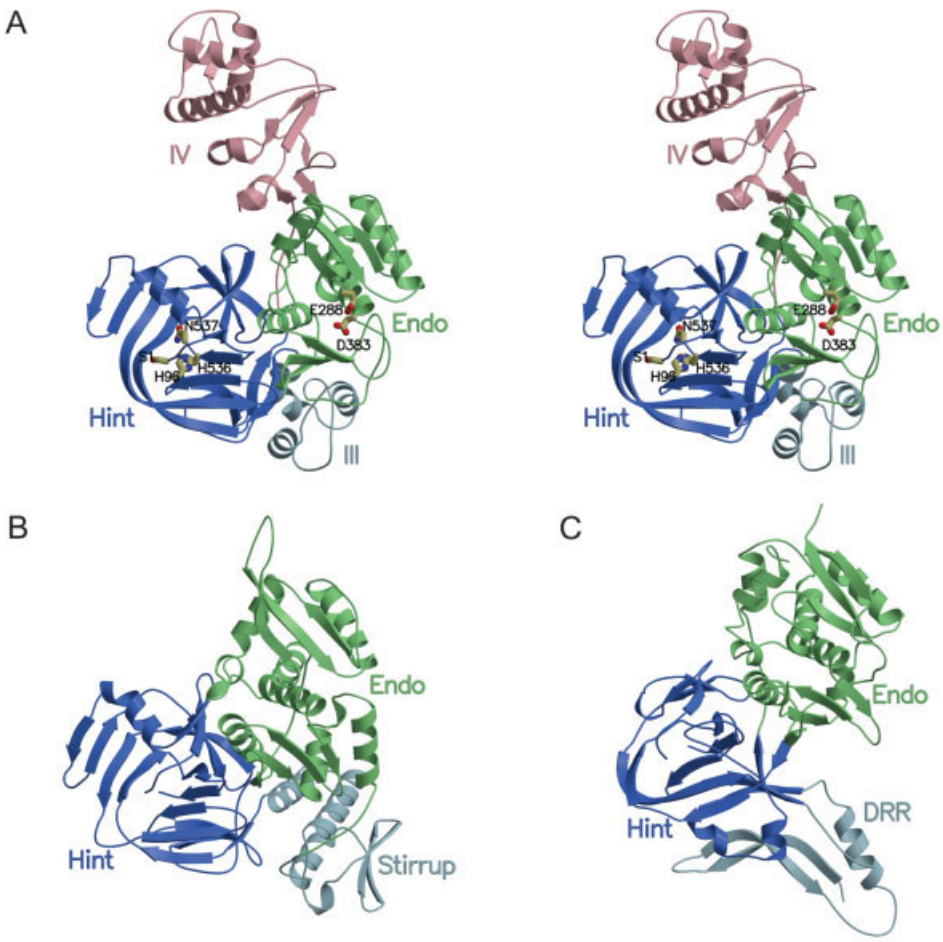
A
IV
Endo
Hint
III
N537
E288
D383
H536
H98
S1
B
Endo
Hint
Stirrup
C
Endo
Hint
DRR
D
Hint domain-N
Block A
Block B
Hint domain-C
Block F
Block G
Endonuclease domain
Block C
Block H'
Block D
Block E
Block H
Domain III
Domain IV

Figure 1

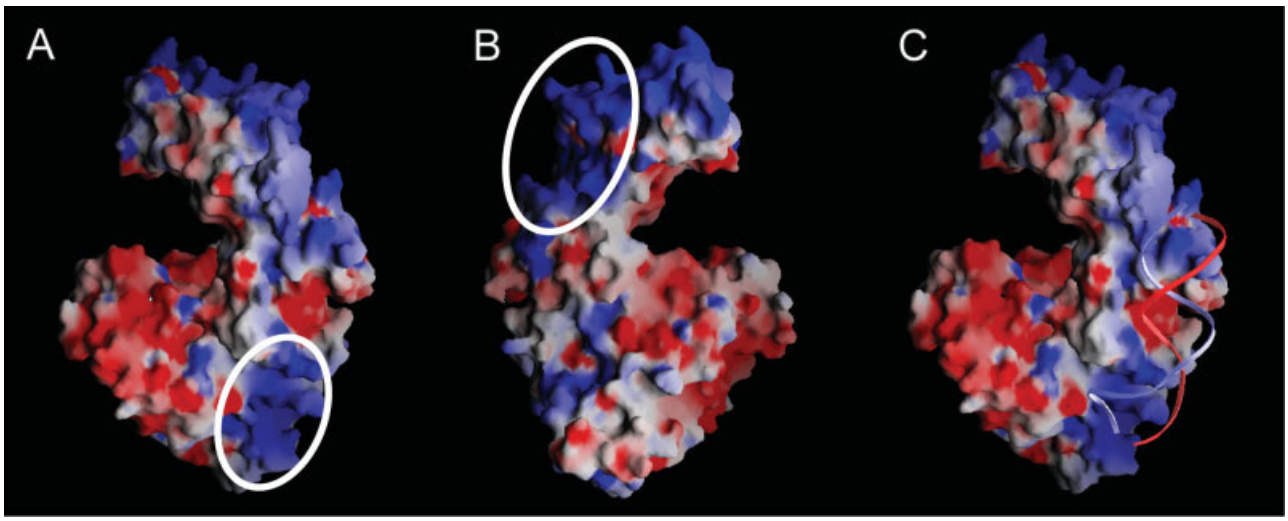

Fig. 2. **A:** Electrostatic charge distribution on the surface of PI-TkoII; the view is projected in the same orientation as Figure 1(A). **B:** View is rotated by 180° with respect to a vertical axis. **C:** Docking model for the protein–DNA complex. The minimal 16-bp DNA recognition sequence is shown as a ribbon.

main III is highly positively charged, as depicted by a circle in Figure 2(A). Side-chains of the charged residues such as Glu454, Lys453, and Lys466 protrude from the concave surface, which faces the same side of the adjacent endonuclease active site. A DNA binding model based on the structure of PI-SceI complexed with a DNA duplex[6] shows that a target DNA will fit well into the surface, with the minor groove of the DNA duplex [Fig. 2(C)]. Distance between the side-chain of Arg292 on the protruded loop of the endonuclease domain and the side-chain of Lys453 on the surface of domain III is 51 Å, which approximately corresponds to the 16-bp double-stranded DNA.

Domain IV is approximately 40 × 25 × 20 Å, and its surface is also highly positively charged. According to DALI server,[11] proteins with the closest structural homologies in domain IV are putative DNA binding proteins such as the predicted transcription regulator (3.9 Å of RMSD over 70 residues with 14% sequence identity; PDB code: 1XMA), Esa1 histone acetyltransferase fragment[15] (3.7 Å of RMSD over 64 residues with 3% sequence identity; PDB code: 1FY7), and the cell division control protein 6 (Cdc6)[16] (7.3 Å of RMSD over 75 residues with 12% sequence identity; PDB code: 1FNN). Because the corresponding domain of Cdc6 has been reported to be highly conserved, and to possibly be a DNA binding element,[16] domain IV of PI-TkoII might have the potential to bind to DNA. Many positive-charged residues such as Lys159, Lys161, Arg186, Arg187, Arg190, Arg199, Lys201, Lys202, Lys233, Lys256, Lys259, Arg267, and Arg269 are located on the surface of domain IV [depicted as a circle in Fig. 2(B)]. However, the surface is facing the other side of the endonuclease active site [depicted as a circle in Fig. 2(B)]. One possible hypothesis is that this positive-charged surface participates in binding to DNA in cases when target DNA length is much longer than that of the minimal recognition sequence. In that case, DNA bending would be induced by protein–DNA interactions similar to those observed in PI-SceI complexed with DNA.[6] Modeling analysis showed that DNA bending angle was greater than that in the PI-SceI and DNA complex; therefore, the DNA recognition mechanism of PI-TkoII might be different from that of PI-SceI.

Materials and Methods. X-ray diffraction data of the native crystal have been obtained previously.[17] To produce selenomethionine substituted (Se-Met) PI-TkoII, cells were cultivated in minimal medium containing Se-Met. Purification and crystallization procedures for Se-Met PI-TkoII were essentially the same as those used for the native protein. The microseeding method was used to optimize reproducibility for Se-Met PI-TkoII. Crystals of Se-Met PI-TkoII began to develop after a week, and growth was completed after a month at 293 K. X-ray diffraction data for SeMet crystals were collected with the synchrotron radiation source of SPring-8. SeMet PI-TkoII crystals belong to the orthogonal space group $C222_1$ with lattice constants of $a = 103.97$, $b = 150.54$, $c = 145.49$ Å (Table I). Multiple-wavelength anomalous diffraction (MAD) data were collected using three different wavelengths (0.9792, 0.9794, and 0.9686 Å). Reflections were indexed, integrated, and scaled DENZO and SCALEPACK.[18] Phase calculations were performed using CNS.[19] Model building

Fig. 1. Structures of PI-TkoII, PI-PfuI, and PI-SceI. **A:** Stereo ribbon diagram of PI-TkoII. PI-TkkoII contains the endonuclease, Hint, III, and IV domains colored in green, blue, light blue, and pink, respectively. Active site residues are shown. **B:** Ribbon diagram of PI-PfuI. **C:** Ribbon diagram of PI-SceI. **D:** Domain organization and multiple sequence alignment of PI-TkoII, PI-PfuI, and PI-SceI. The aligned sequences were displayed in ESPript.[24] Secondary structures were calculated by DSSP.[25] Secondary structure annotations and numberings on top correspond to PI-TkoII. α-Helices are represented by spirals and β-strand by arrows. White lettering boxed with a red background indicates residues that are conserved in all three sequences, and red lettering indicates similar residues. The conserved sequence blocks of inteins are indicated.

and refinement of native and SeMet PI-TkoII structure were done using the programs O[20] and CNS.[19] Refinement statistics are shown in Table I. The DNA docking model was prepared by manual docking based on the crystal structure of PI-SceI in complex with DNA[6] and then energy minimized with the program CNS.[19] Figures were generated by MOLSCRIPT,[21] RASTER3D,[22] GRASP,[23] and ESPRIPT.[24] Atomic coordinates and structure factors are available from the PDB under accession code 2CW7 for native and 2CW8 for SeMet PI-TkoII, respectively.

REFERENCES

1. Chong S, Shao Y, Paulus H, Benner J, Perler FB, Xu MQ. Protein splicing involving the *Saccharomyces cerevisiae* VMA intein. The steps in the splicing pathway, side reactions leading to protein cleavage, and establishment of an in vitro splicing system. J Biol Chem 1996;271(36):22159–22168.
2. Cooper AA, Stevens TH. Protein splicing: self-splicing of genetically mobile elements at the protein level. Trends Biochem Sci 1995;20(9):351–356.
3. Perler FB, Olsen GJ, Adam E. Compilation and analysis of intein sequences. Nucleic Acids Res 1997;25(6):1087–1093.
4. Takagi M, Nishioka M, Kakihara H, et al. Characterization of DNA polymerase from *Pyrococcus* sp. strain KOD1 and its application to PCR. Appl Environ Microbiol 1997;63(11):4504–4510.
5. Duan X, Gimble FS, Quiocho FA. Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. Cell 1997;89(4): 555–564.
6. Moure CM, Gimble FS, Quiocho FA. Crystal structure of the intein homing endonuclease PI-SceI bound to its recognition sequence. Nat Struct Biol 2002;9(10):764–770.
7. Klabunde T, Sharma S, Telenti A, Jacobs WR Jr, Sacchettini JC. Crystal structure of GyrA intein from *Mycobacterium xenopi* reveals structural basis of protein splicing. Nat Struct Biol 1998;5(1):31–36.
8. Ichiyanagi K, Ishino Y, Ariyoshi M, Komori K, Morikawa K. Crystal structure of an archaeal intein-encoded homing endonuclease PI-PfuI. J Mol Biol 2000;300(4):889–901.
9. Nishioka M, Fujiwara S, Takagi M, Imanaka T. Characterization of two intein homing endonucleases encoded in the DNA polymerase gene of *Pyrococcus kodakaraensis* strain KOD1. Nucleic Acids Res 1998;26(19):4409–4412.
10. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247(4):536–540.
11. Holm L, Sander C. Mapping the protein universe. Science 1996; 273(5275):595–603.
12. Silva GH, Dalgaard JZ, Belfort M, Van Roey P. Crystal structure of the thermostable archaeal intron-encoded endonuclease I-DmoI. J Mol Biol 1999;286(4):1123–1136.
13. Ding Y, Xu MQ, Ghosh I, et al. Crystal structure of a mini-intein reveals a conserved catalytic module involved in side chain cyclization of asparagine during protein splicing. J Biol Chem 2003;278(40):39133–39142.
14. Hall TM, Porter JA, Young KE, Koonin EV, Beachy PA, Leahy DJ. Crystal structure of a Hedgehog autoprocessing domain: homology between Hedgehog and self-splicing proteins. Cell 1997;91(1):85–97.
15. Yan Y, Barlev NA, Haley RH, Berger SL, Marmorstein R. Crystal structure of yeast Esa1 suggests a unified mechanism for catalysis and substrate binding by histone acetyltransferases. Mol Cell 2000;6(5):1195–1205.
16. Liu J, Smith CL, DeRyckere D, DeAngelis K, Martin GS, Berger JM. Structure and function of Cdc6/Cdc18: implications for origin recognition and checkpoint control. Mol Cell 2000;6(3):637–648.
17. Hashimoto H, Nishioka M, Inoue T, et al. Crystallization and preliminary X-ray crystallographic analysis of archaeal O6-methylguanine-DNA methyltransferase. Acta Crystallogr D Biol Crystallogr 1998;54(Pt 6 Pt 2):1395–1396.
18. Otwinowski Z. In: Sawyer L, Isaacs N, Bailey S, editors. Proceedings of the CCP4 study weekend: data collection and processing. Warrington: Daresbury Laboratory; 1993. p 56–62.
19. Adams PD, Pannu NS, Read RJ, Brunger AT. Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. Proc Natl Acad Sci USA 1997;94(10):5018–5023.
20. Jones TA. A graphics model building and refinement system for macromolecules. J Appl Crystallogr 1978;15:23–31.
21. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24: 946–950.
22. Merritt EA. Raster3D Version 2.0. A program for photorealistic molecular graphics. Acta Crystallogr D Biol Crystallogr 1994; 50(Pt 6):869–873.
23. Nicholls A, Sharp KA, Honig B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 1991;11(4):281–296.
24. Gouet P, Courcelle E, Stuart DI, Metoz F. ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics 1999; 15(4):305–308.
25. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–2637.