PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

STRUCTURE NOTE

Crystal structure of a novel non-Pfam protein AF1514 from *Archeoglobus fulgidus* DSM 4304 solved by S-SAD using a Cr X-ray source

Yang Li,[1†] Pazilat Bahti,[2†] Neil Shaw,[1] Gaojie Song,[1] Shunmei Chen,[1] Xuejun Zhang,[3] Min Zhang,[4] Chongyun Cheng,[1] Jie Yin,[1] Jin-Yi Zhu,[5] Hua Zhang,[5] Dongsheng Che,[5] Hao Xu,[5] Abdulla Abbas,[2] Bi-Cheng Wang,[5] and Zhi-Jie Liu[1*]

[1] National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

[2] College of Life Science and Technology, Xinjiang University, Urumqi 830046, China

[3] Department of Immunology, Tianjin Medical University, Tianjin 300070, China

[4] Life Sciences College, Anhui University, Hefei 230039, China

[5] SECSG, Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30605

***Key words:** sulfur phasing; non-Pfam; methylation; X ray.*

INTRODUCTION

Many computational tools have been developed recently to accurately predict the structure of a protein from its amino acid sequence.[1–3] In general, if a query protein sequence shares at least 30% sequence identity with a protein sequence whose 3D structure has been determined, the structure of this query sequence can be modeled based on the template structure, using MODELLER software for example.[3,4] Computational software, however, cannot guarantee accurate prediction for those new proteins that share low sequence similarity in PDB. Therefore, experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) are still the main approaches for a protein structural study.[5,6]

The current target selection strategy of most structural genomics centers[7–9] mainly focuses on the representatives of manually curated protein families (Pfam),[10–12] that is, the selected protein sequence shares at least one conserved domain with other members within a family. In this way, the solved representative structures can be used as structural templates to predict structures of the remaining protein sequences in the same family using computational tools. It has been shown that this "Pfam" target selection strategy increases not only the number of novel structures, but also the number of new folds.[13] However, over-emphasis of Pfam and ignoring non-Pfam sequences (i.e., not sharing any conserved domain in Pfam) in target selection might lead to biased distribution of Pfam and non-Pfam structures in PDB, and possibly slow the growth rate of new structures and folds.

Our analysis on 150 microbial genomes showed that non-Pfam sequences account for 25–30% of all Open Reading Frames (ORFs) for most genomes, and some could reach to 60% (unpublished data). The high percentage of non-Pfams over all ORFs reminds us non-Pfams should not be neglected while devising a target selection strategy. On the other hand, these non-Pfam sequences for each genome are either paralogous non-Pfam (in which sequences have homologous partners within the same organism), or orthologous non-Pfam (in which sequences have orthologous partners in the closely

Grant sponsor: Ministry of Science and Technology of China; Grant numbers: 2006AA02A316, 2006CB910903; Grant sponsor: National Natural Science Foundation of China; Grant numbers: 30670427, 30700123, 30770432; Grant sponsors: US National Institutes of Health; Grant number: 1P50 GM62407; Grant sponsor: CAS-KIST Collaborative Research Grant, University of Georgia Research Foundation, Georgia Research Alliance.

[†] Yang Li and Pazilat Bahti contributed equally to this work.

*Correspondence to: Zhi-Jie Liu, National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China.
E-mail: zjliu@ibp.ac.cn

Received 29 December 2007; Revised 24 January 2008; Accepted 25 January 2008
Published online 24 March 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22025

related organisms), or singleton non-Pfam (in which sequences have only one copy in the organism). These three non-Pfam kinds are either organism-specific or genus-specific, implying the possible existence of undiscovered unique features of non-Pfams, such as unique SCOP fold or CATH topology.

Many Pfam sequences have significant biological meaning, while the functions of most non-Pfam sequences are unknown to date. However, this does not mean that non-Pfam sequences are biologically less meaningful. Some non-Pfam sequences and structures are predicted to play important roles for the uniqueness of these organisms. Therefore, a non-Pfam selection strategy will not only accelerate the expansion of SCOP fold space, but also help biologists understand the functions of these proteins based on structures. The present work describes the structure of AF1514, a non-Pfam protein from *Archeoglobus fulgidus* with unknown function, solved at 1.8 Å resolution by using anomalous signal of sulfur generated by chromium X rays (wavelength = 2.29 Å).

METHODS

Protein production

E coli BL21 was freshly transformed with plasmid containing AF1514 gene. Cells were grown at 37°C until culture density reached $OD_{600\ nm} - 0.8$. The culture was cooled down to 12°C and induced with 0.2 m*M* IPTG for 40 h. Cells were harvested by centrifugation and lysed by sonication. Cell debris was removed by centrifugation and the clarified supernatant subjected to Ni-affinity chromatography. The protein was further purified using size exclusion chromatography. The purified AF1514 protein was divided into two equal aliquots. One aliquot was methylated as described previously,[14–16] while the other aliquot was directly concentrated without any chemical modification. Both, methylated and non-methylated protein samples were concentrated to ~18 mg mL^{-1} in 20 m*M* Tris-HCl, pH 8.0, 200 m*M* NaCl, and 1 m*M* DTT before setting up crystallization drops.

Crystallization and data collection

Crystallization screening was carried out in hanging drop vapor diffusion method using TTP Lab Tech mosquito robot. Commercially available sparse matrix screens (Hampton Research - Crystal Screen 1 and 2, Index and PEG/Ion Screen and Emerald Biosystems' Wizard I and II) were used to screen crystallization space. Crystallization optimization was carried out in 2 μL hanging drops containing 1 μL protein mixed with 1 μL mother liquor. The drops were equilibrated over 300 μL reservoir solution and incubated at 16°C. Tetragonal crystals appeared in 5 days for both non-methylated and methylated proteins in a crystallization solution containing 0.1*M* sodium acetate pH 5.0, 0.1*M* sodium chloride, 10% (w/v) MPD. Crystals were flash frozen in liquid nitrogen prior to mounting and data were collected at cryogenic temperature (100 K). The diffraction quality of the non-methylated protein crystals mounted directly from the mother liquor was poor and could diffract X rays to only 3.5 Å. The resolution improved after optimizing the cryoprotectant concentration and the best crystal diffracted X rays to 2.4 Å. The sulfur anomalous diffraction data for the non-methylated protein crystal was collected using a chromium rotating anode X-ray source and R-AXIS IV^{++} detector (Rigaku) with 102 mm crystal to detector distance and 240 s exposure time per image. To improve the signal-to-noise ratio of the anomalous signal of sulfur, the crystal was scanned 2 × 360° and a 2.4 Å resolution dataset for the non-methylated AF1514 crystal was collected. The crystal for methylated AF1514 protein was obtained under the same crystallization condition as the non-methylated protein. Methylated protein crystal dataset consisting of a single axis ϕ scan with 308 half-a-degree oscillation images was recorded on a cupper rotating anode X-ray source and R-AXIS IV^{++} detector (Rigaku) using a crystal-to-detector distance of 160 mm and 240 s exposure time per image. A 1.8 Å resolution dataset for the methylated AF1514 crystal was collected. Both methylated and non-methylated data sets were indexed, integrated and scaled using HKL2000.[17] The crystals belong to space group $P4_12_12$ (identified during structure determination) with unit-cell parameters a = 49.24 Å, b = 49.24 Å, c = 106.46 Å. The data collection statistics are listed in Table I.

Phasing and refinement

The structure was solved by the sulfur SAD method.[18] The anomalous signal of sulfur atoms from two cysteine residues and a methionine was located by SHELXD[19] using a 2.44 Å resolution dataset collected in-house using a chromium X-ray source. Initial phasing was done using program Sharp.[20] Most of the polypeptide backbone could be traced automatically by Arp/Warp.[21] Minor revisions to the model were done manually using COOT.[22] Refinement was carried out using REFMAC[23] against a higher resolution dataset of methylated protein crystal collected using copper X-ray source. The refinement converged to give the statistics presented in Table I. The final model was validated using MolProbity[24] and PROCHECK[25] prior to submission to the Protein Data Bank.[26]

RESULTS AND DISCUSSION

A Wu-Blast search of PDB for structural homologues failed to retrieve any structure similar to AF1514 (*E* value > 0.50). The protein has two methionine (Met1 and Met4) and three cysteine residues (Cys53, Cys54, and Cys64) that could be used for sulfur phasing. Met1 and Cys54 were disordered and could not be used for the phasing. Anomalous signal of sulfurs from two cysteine

Table I
Data Collection and Refinement Statistics

Dataset	Methylated	Non-methylated
X-ray Generator	FR-E+ SuperBrightTM	MicroMax-007 HF +VariMax^{TM}
Detector	R-AXIS IV++	R-AXIS IV++
Crystal-to-detector distance (mm)	160	102
Wavelength (Å)	1.54	2.29
Number of images	308	720
Oscillation width (°)	0.5	1.0
Space group	$P4_12_12$	$P4_12_12$
Unit cell parameters		
a (Å)	49.74	49.27
b (Å)	49.74	49.27
c (Å)	106.46	106.60
Resolution range (Å)	50.00–1.80 (1.86–1.80)	50.00–2.44 (2.53–2.44)
Completeness (%)	91.4 (58.0)	99.7 (100.0)
Total measured reflections	110162	266772
Unique reflections	11982 (735)	5332 (502)
Redundancy	9.2 (2.5)	50 (41.1)
Rsym (%)	4.2 (21.5)	9.5 (47.3)
I/σ	46.2 (4.8)	67.9 (13.6)
Refinement		
Resolution limits	45.08–1.80	
Reflections used	11347	
Number of refined atoms	807	
Rwork (Rfree, %)	20.5 (22.6)	
Bond RMSD lengths (Å)/angles (°)	0.007/1.009	
Mean B value (Å^2)	19.37	
All-atom clash score	7.09	
Ramachandran favored	83/85 (no outliers)	

The numbers in parentheses represent values for the highest resolution shell.

[a] $R_{\text{sym}} = \sum|I_i - \langle I \rangle|/\sum I$ where I_i is the intensity of the ith observation and $\langle I \rangle$ is the mean intensity of the reflections.

[b] $R_{\text{work}} = \sum||F_{\text{obs}}| - |F_{\text{calc}}||/\sum|F_{\text{obs}}|$ where F_{calc} and F_{obs} are the calculated and observed structure factor amplitude, respectively.

[c] $R_{\text{free}} = \sum||F_{\text{obs}}| - |F_{\text{calc}}||/\sum|F_{\text{obs}}|$ where all reflections belong to a test set of randomly selected data.

residues (Cys53 and Cys64) and a methionine (Met4) was used for determination of the phases. A longer wavelength chromium X-ray source was used to collect a 2.44 Å dataset. An initial experimental electron density map calculated at 2.5 Å was of very good quality and more than 90% of the residues could be fit in automatically using ARP/wARP. The final model refined to 1.8 Å resolution had an *R* value of 20.5% (*R* free 22.6%). The overall geometry of the model was excellent with no residues lying in the disallowed region of the Ramchandran plot.

The asymmetric unit consists of one monomer of the protein based on the calculated solvent content of 62.0%. However, the size exclusion chromatography profile of AF1514 suggested that the protein could exist as a dimer. Further analysis of the region around the asymmetric unit revealed that two molecules of the protein associate to form a homodimer. Interestingly, the two molecules in the homodimer are related by a crystallographic symmetry twofold axis. There are eight hydrogen bonds between the two beta sheets of the two adjacent molecules with the secondary structural elements running anti-parallel to each other. There are 167 water molecules in the final model. Electron density for amino acid residues 3–87 was clearly visible.

The overall structure consists of 2 α helices, 5 β sheets, and 8 loops [Fig. 1(A)]. Two monomers of AF1514 sit side by side with the secondary structural elements of one monomer running anti parallel to the other within the dimer [Fig. 1(B)]. Surface electrostatic potential map of the protein showed an uneven distribution of charge on the protein surface [Fig. 1(C)]. CATH server[27] classified the structure as mixed alpha beta with a topology similar to thiol ester dehydrase. Superimposition of the structure of AF1514 over a thiol ester hydrolase from *Arthrobacter* (PDB code 1Q4S) showed 57 of the 85 main chain carbons overlapped with an RMSD of 2.7 Å. The secondary structural elements of AF1514 are arranged in a "hot dog" fold similar to the thiol ester hydrolase.[28] While the quaternary structure of the *Arthrobacter* thiol esterase represents a tetramer, the AF1514 protein exists as a dimer. The substrate is seen sitting in a wedge between the two subunits of the dimer pair.[28] An equivalent binding site for the substrate in AF1514 could not be identified successfully by superimposing the structure

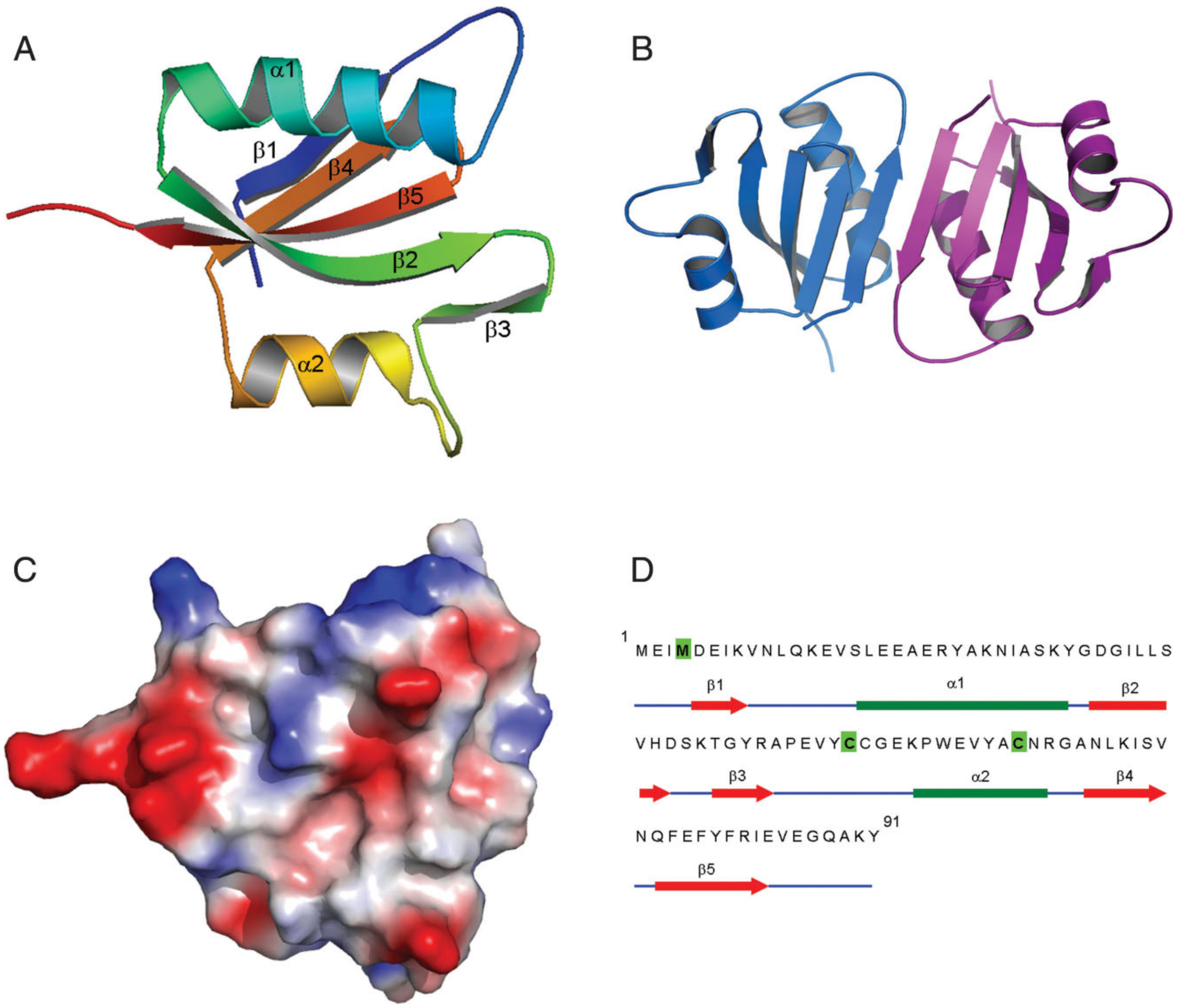
A
α1
β1
β4
β5
β2
β3
α2
B
C
D
1
MEIMDEIKVNLQKEVSLEEAERYAKNIASKYGDGILLS
β1
α1
β2
VHDSKTGYRAPEVYCCGEKPWEVYACNRGANLKISV
β3
α2
β4
NQFEFYFRIEVEGQAKY
91
β5

Figure 1

*Overall structure of AF1514. **A:** A cartoon representation of the AF1514 structure. AF1514 is made up of two helices, five beta sheets and numerous loops. **B:** A homodimer of AF1514 showing two monomers sitting next to each other with the secondary structural elements running anti parallel. **C:** A surface electrostatic potential representation of the AF1514 structure. Positive potential is colored blue, negative potentials are colored red. **D:** Primary sequence of AF1514 annotated with secondary structural elements. Residues highlighted in green contributed the anomalous signal of sulfur.*

of the *Arthrobacter* thiol esterase over AF1514. AF1514 therefore is less likely to function as a thiol esterase. DALI analysis of AF1514 structure failed to identify any structural neighbor of known function with significant similarity (Z score > 2.5).

In future, as more structures with known function are deposited in PDB, clues about the function of AF1514 based on structure could be obtained. Since no homologues with known function could be identified based on primary sequence and structural similarity, it is quite likely AF1514 may be carrying out a unique function. Further functional studies are required in order to determine the exact role of AF1514 in *Archeoglobus fulgidus*.

ACKNOWLEDGMENTS

Accession numbers:

Atomic co-ordinates and structure factors for the AF1514 structure have been deposited in the Protein Data Bank (accession code 3C0F).

REFERENCES

1. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
2. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences

using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.

3. Sali A. Comparative protein modeling by satisfaction of spatial restraints. Mol Med Today 1995;1:270–277.
4. Sánchez R, Pieper U, Melo F, Eswar N, Martí-Renom MA, Madhusudhan MS, Mirković N, Sali A. Protein structure modeling for structural genomics. Nat Struct Biol 2000;7 (Suppl):986–990.
5. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 1958;181:662–666.
6. Wuthrich K. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. Science 1989;243:45–50.
7. Brenner SE. Target selection for structural genomics. Nat Struct Biol 2000;7 (Suppl):967–969.
8. Brenner SE, Levitt M. Expectations from structural genomics. Protein Sci 2000;9:197–200.
9. Chandonia JM, Brenner SE. Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. Proteins 2005;58:166–179.
10. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 1998;26:320–322.
11. Sonnhammer EL, Eddy SR, Durbin R. Pfam: A comprehensive database of protein families based on seed alignments. Proteins 1997;28:405–420.
12. Bateman A, et al. The Pfam protein families database. Nucleic Acids Res 2004;32:D138–D141.
13. Murzin AG, Brenner SE, Hubbard T, Chothia C. A structural classification of proteins for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
14. Rayment I. Reductive alkylation of lysine residues to alter crystallization properties of proteins. Methods Enzymol 1997;276:171–179.
15. Walter T, Meier C, Assenberg R, Au K-F, Ren J, Verma A, Nettleship J, Owens R, Stuart D, Grimes J. Lysine methylation as a routine rescue strategy for protein crystallization. Structure 2006;14:1617–1622.
16. Shaw N, Cheng C, Tempel W, Chang J, Ng J, Wang X-Y, Perrett S, Rose J, Rao Z, Wang B-C, Liu Z-J. (NZ)CH.O Contacts assist crystallization of a ParB-like nuclease. BMC Struct Biol 2007;7:46–58
17. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol 1997;276:307–326.
18. Wang BC. Resolution of phase ambiguity in macromolecular crystallography. Methods Enzymol 1985;115:90–112.
19. Uson I, Sheldrick GM. Advances in direct methods for protein crystallography. Curr Opin Struct Biol 1999;9:643–648.
20. de la Fortelle E, Bricognne G. Maximum-likelihood heavy-atom Parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. Methods Enzymol 1997;276:472–494.
21. Perrakis A, Morris R, Lamzin VS. Automated protein model building combined with iterative structure refinement. Nat Struct Biol 1999;6:458–463.
22. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D 1997;53:240–255.
23. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta D 2004;60:2126–2132.
24. Davis IW, Murray LW, Richardson JS, Richardson DC. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res 2004;32:W615–W619.
25. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.
26. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, Westbrook J. The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol Suppl 2000;7:957–959.
27. Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res 2005;33:D247–D251.
28. Thoden JB, Zhuang Z, Dunaway-Mariano D, Holden H. The structure of 4-hydroxybenzoyl-CoA thioesterase from *Arthrobacter sp. strain SU.* J Biol Chem 2003;278:43709–43716.