PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

STRUCTURE NOTE

Crystal structure of uncharacterized protein TTHA1756 from *Thermus thermophilus* HB8: Structural variety in UPF0150 family proteins

Akio Ebihara,[1] Miho Manzoku,[1] Hitoshi Iino,[1] Mayumi Kanagawa,[1] Akeo Shinkai,[1] Shigeyuki Yokoyama,[1,2,3] and Seiki Kuramitsu[1,3,4]*

[1] RIKEN SPring-8 Center, Harima Institute, 1-1-1 Kouto, Sayo-cho, Sayo-gun, Hyogo 679-5148, Japan

[2] Department of Biophysics and Biochemistry, Graduate School of Science, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

[3] RIKEN Genomic Sciences Center, Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan

[4] Department of Biological Sciences, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan

Key words: uncharacterized protein; hypothetical protein; UPF0150; HicB; double-stranded RNA-binding domain; RNase H; domain swapping; DUF1902; structural genomics.

INTRODUCTION

The genome sequences of diverse organisms commonly contain a large fraction of genes encoding functionally uncharacterized proteins (PEDANT database; http://pedant.gsf.de/index.jsp). Determination of the three-dimensional structures of such proteins can provide valuable functional clues that may not be revealed by sequence data alone and may contribute to expansion of the protein folding pattern.[1,2] The UPF0150 family protein in the Pfam database[3] is a small structural domain of about 70 amino acids that is functionally uncharacterized. To date, the UPF0150 family consists of 342 bacterial, 50 archeal, and 5 viral proteins. Like many genomes containing multiple copies of UPF0150 family protein genes,[4] that of an extremely thermophilic bacterium, *Thermus thermophilus* HB8, has the genes of five proteins of this family: TTHA1756, TTHA0933, TTHA1912, TTHA0231, and TTHA0281 [Fig. 1(A)]. We previously reported the crystal structure of TTHA0281.[7] In this study, we have solved the crystal structure of the TTHA1756 protein, another UPF0150 family protein from *T. thermophilus* HB8. Compared with structural homologs, we found that the UPF0150 protein family has two types of quaternary structure with a double-stranded RNA-binding domain (dsRBD) as a common fold, and a less-conserved region among the family proteins probably contributes to the structural variety.

METHODS

Protein production, crystallization, and data collection

The *TTHA1756* gene was cloned into pET-11a (Novagen). Selenomethionine (SeMet)-labeled TTHA1756 was produced in methionine auxotroph *Escherichia coli* B834(DE3) cells. The cell lysate was heated at 70°C for 10 min. The clear supernatant was applied to a TOYOPEARL SuperQ-650M column (Tosoh) pre-equilibrated with 20 m*M* Tris-HCl buffer (pH 8.0), and then the proteins were eluted with a linear gradient of 0 to 1 *M* NaCl. The TTHA1756-containing fraction was purified

This work was performed at RIKEN SPring-8 Center, Harima Institute.
Grant sponsors: RIKEN Structural Genomics/Proteomics Initiative and the National Project on Protein Structural and Functional Analyses, Ministry of Education, Culture, Sports, Science and Technology of Japan.
*Correspondence to: Seiki Kuramitsu, RIKEN SPring-8 Center, Harima Institute, 1-1-1 Kouto, Sayo-cho, Sayo-gun, Hyogo 679-5148, Japan.
E-mail: kuramitu@spring8.or.jp
Received 26 December 2007; Revised 21 January 2008; Accepted 29 January 2008
Published online 12 March 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22018

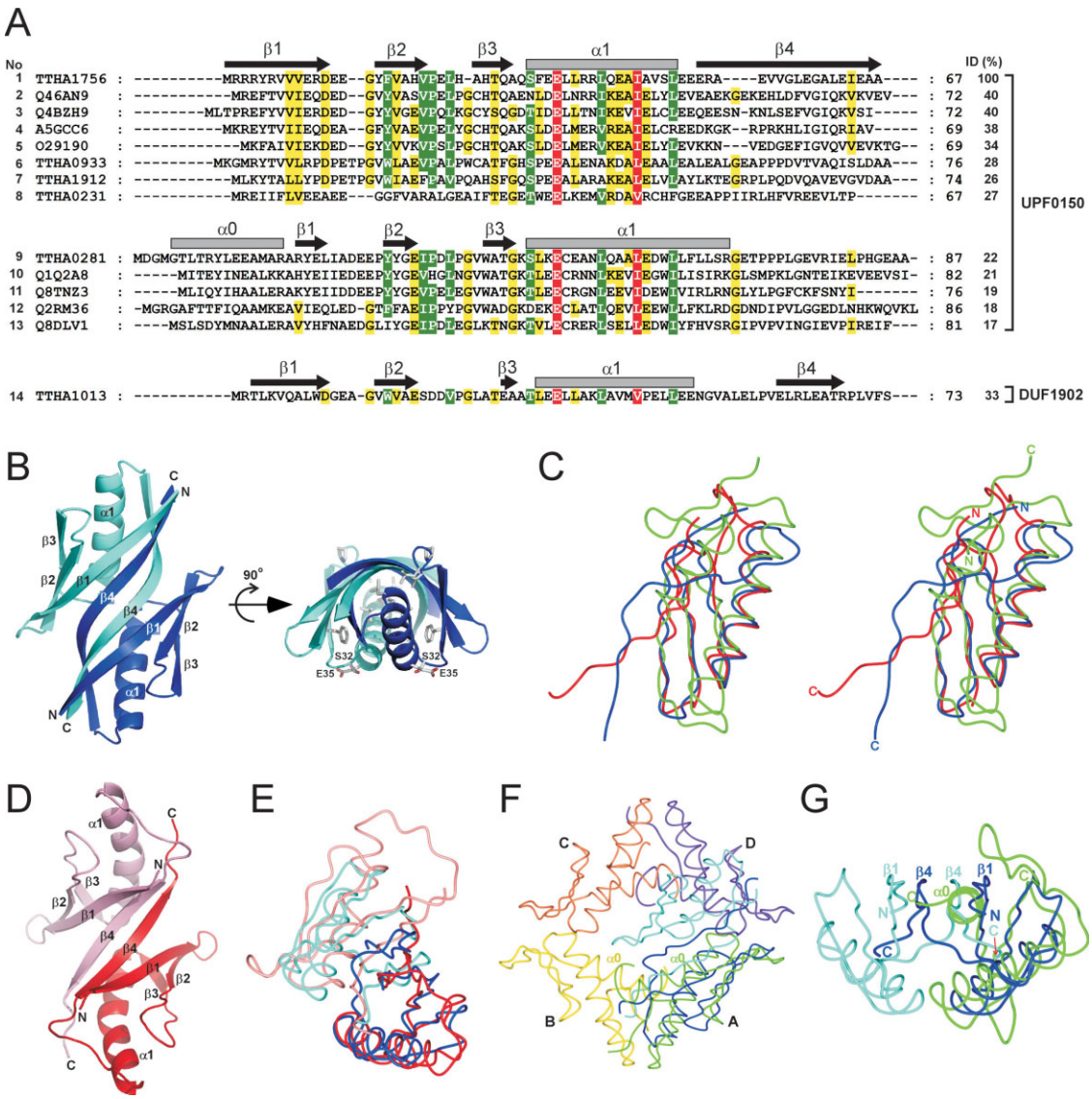

Figure 1

*(**A**) Sequence alignment of TTHA1756 and sequence homologs. Using PSI-BLAST search,[5] Nos. 2 to 7 and Nos. 10 to 13 were detected using TTHA1756 (No. 1) and TTHA0281 (No. 9) as a query, respectively. No. 8 is a member of the UPF0150 family. No. 14 was identified through the structure-based similarity search. All sequences were aligned using CLUSTAL W.[6] Except for six sequences from T. thermophilus HB8, each sequence is identified by the UniProtKB/TrEMBL entry number and the species names are as follows: Q46AN9, Methanosarcina barkeri; Q4BZH9, Crocosphaera watsonii; A5GCC6, Geobacter uraniumreducens; Q29190, Archaeoglobus fulgidus; Q1Q2A8, Candidatus Kuenenia stuttgartiensis; Q8TNZ3, Methanosarcina acetivorans; Q2RM36, Moorella thermoacetica; and Q8DLV1, Synechococcus elongatus. Residues are colored in accordance with sequence conservation: 100% (white on red), 80% (white on green), and 60% (shaded in yellow). Secondary structural elements of TTHA1756, TTHA0281, and TTHA1013 are shown above the respective sequences as grey rectangles (α-helices) and arrows (β-strands). "ID" represents percentage identity to TTHA1756. (**B**) Ribbon diagram of TTHA1756 dimer. The N- and C-termini are labeled. Each subunit is colored differently. The right panel was drawn after rotation of 90° about the horizontal axis of the dimer. Highly conserved residues (white on red and white on green, panel A) are shown as a stick-model. (**C**) Stereo view of superposing of TTHA1756 (blue) and its structural homologs, TTHA1013 (red) and TTHA0281 (green). Their N- and C-termini are labeled with the same color-coding. (**D**) Ribbon diagram of TTHA1013 dimer. Each subunit is colored differently. (**E**) Structure comparison of the TTHA1756 dimer (blue and cyan) and the TTHA1013 dimer (red and pink). One of the subunits of these dimers (blue and red) is superimposed. (**F** and **G**) Structure comparison of the TTHA1756 dimer (blue and cyan) with the TTHA0281 tetramer (green, yellow, tan, and purple, panel F) and TTHA0281 monomer (green, panel G). One of the subunits of the TTHA1756 dimer (blue) is superimposed with one subunit of the TTHA0281 tetramer (green). The helices α0 of subunits A and B are labeled.*

with a gradient of 0 to 500 m*M* NaCl on a Resource Q column (GE Healthcare Biosciences). The TTHA1756-containing fraction was applied to a hydroxyapatite column (Bio-Scale CHT10-I, Bio-Rad) pre-equilibrated with 10 m*M* sodium phosphate, 150 m*M* NaCl (pH 7.0), and eluted with a sodium phosphate gradient of 10 to 250 m*M*. TTHA1756 was applied to a HiLoad 16/60 Superdex 75 pg column (GE Healthcare Biosciences) pre-equili-

Table I
X-ray Data Collection and Refinement Statistics

Data set	Se-Remote	Se-Peak	Se-Edge
Wavelength (Å)	0.9000	0.9789	0.9794
Space group		$P3_121$	
Unit cell parameters (Å)		$a = b = 50.15$, $c = 46.82$, $\gamma = 120^\circ$	
Molecules per asymmetric unit		1	
Resolution range (Å)	50–1.8	50–1.75	50–1.8
Reflections observed/unique	65,371/6,606	70,232/7,163	64,763/6,590
Completeness (%)	99.9 (99.4)	99.8 (98.4)	99.7 (98.0)
Redundancy	9.9	9.8	9.8
Average $I/\sigma I$	39.6 (9.2)	36.2 (7.5)	40.3 (8.4)
R_{merge} (%)[a]	5.1 (19.7)	5.6 (19.9)	5.0 (18.3)
Refinement			
Resolution range (Å)	43–1.8		
Total reflections	6,575		
R_{work} (%)[b]/R_{free} (%)[c]	24.2/27.1		
Number of protein atoms/water atoms	545/33		
RMSD[d] bond length (Å)	0.005		
RMSD angle (°)	1.20		
Average B-factor (Å^2)	23.5		
Ramachandran plot			
Most favored (%)	95.1		
Additional (%)	4.9		
Disallowed (%)	0.0		

Values in parentheses are for the highest resolution shell.
[a] $R_{merge} = \Sigma_h\Sigma_i|I_{h,i} - \langle I_h\rangle|/\Sigma_h\Sigma_i I_{h,i}$ where $I_{h,i}$ is the ith observation of reflection h and $\langle I_h\rangle$ is the mean intensity of reflection h.
[b] $R_{work} = \Sigma||F_{obs}| - |F_{calc}||/\Sigma|F_{obs}|$ where F_{obs} and F_{calc} are the observed and calculated structure factors, respectively.
[c] R_{free} is calculated in the same manner as R_{work}, except that F_{obs} corresponds to the 10% of reflections not used in the refinement.
[d] RMSD, root mean square deviation.

brated with 20 m*M* Tris-HCl, 200 m*M* NaCl (pH 8.0), and eluted with the same buffer. TTHA1756 was applied to a HiPrep 26/10 Desalting column (GE Healthcare Biosciences) pre-equilibrated with 20 m*M* Tris-HCl (pH 8.0). The final sample comprised of 18 mg/mL protein in 20 m*M* Tris-HCl and 1 m*M* dithiothreitol (pH 8.0).

Crystals of TTHA1756 were obtained by the hanging-drop vapor diffusion method at 20°C. Drops were assembled with 2 μL of a 10 mg/mL protein solution mixed with an equal volume of the precipitant solution comprising 15% polyethylene glycol 3350, 0.2 *M* diammonium hydrogen citrate, 10 m*M* phenol, and 0.1 *M* sodium citrate (pH 5.0). Crystals were soaked in the precipitant solution containing 20% trehalose and then flash-frozen in a liquid nitrogen stream (−173°C). Multiple-wavelength anomalous dispersion (MAD) data sets for SeMet-labeled crystals were collected at the RIKEN Structural Genomics Beamline II BL26B2, SPring-8 (Hyogo, Japan).[8] The data sets were processed and scaled with the program package HKL2000.[9]

Structure determination and analysis

Selenium sites were determined with the program SOLVE[10] using the MAD data sets. The resulting phases were improved with the program RESOLVE,[11] followed by automatic model tracing with the program ARP/wARP.[12] Model refinement was performed with the programs XtalView/X-fit[13] and CNS.[14] The final model was validated using the program PROCHECK in the CCP4 suite.[15] The refinement statistics are summarized in Table I. The solvent-accessible surface was analyzed with AREAIMOL in the CCP4 suite.[15] Figures were drawn using the programs CCP4mg[16] and PyMOL.[17] The atomic coordinates and structure factors have been deposited in PDB (http://www.rcsb.org/) under accession code 2YZT.

RESULTS AND DISCUSSION

The crystal structure of TTHA1756 was determined by the MAD method and refined to 1.8-Å resolution (Table I). The asymmetric unit comprises one molecule of 67 amino acid residues. The crystal structure adopts an α + β structure consisting of a four-stranded antiparallel β-sheet packed on one side by one α-helix and forms a β-β-β-α-β fold [Fig. 1(B)]. TTHA1756 forms a dimer around a crystallographic two-fold axis, which results in an eight-stranded β-sheet with β4 of one subunit intervening between β1 and β4 of the other subunit [Fig. 1(B)]. An accessible surface area of 2,234 Å^2 (∼30%) is buried at the interface. The molecular weight estimated by gel filtration was 13 k, whereas its subunit molecular mass is 7.7 kDa. These results indicate that TTHA1756 intrinsically forms a dimer. Most conserved hydrophobic residues are located in the interior of the central region

(β2, β3, and α1), probably to stabilize the fold, whereas conserved hydrophilic residues, that is, Ser-32 and Glu-35, are located on the surface [Fig. 1(B, right)].

We searched for structural homologs of TTHA1756 using the DALI server.[18] Representative hits with a root mean square deviation (RMSD) of less than 3 Å for more than 40 paired C_α atoms were TTHA1013 hypothetical protein[19] (PDB code: 1WV8, Z-score = 5.1, RMSD = 2.9 Å), TTHA0281 hypothetical protein[7] (PDB code: 2DSY, Z-score = 4.4, RMSD = 2.5 Å), dsRBD[20] (PDB code: 1DI2, Z-score = 3.9, RMSD = 2.8 Å), and the integrase fragment[21] (PDB code: 2BB8, Z-score = 3.0, RMSD = 1.8 Å). Based on its structural similarity, TTHA1756 constitutes a member of the dsRBD-like fold family in the Structural Classification of Proteins database.[22] Superposing of TTHA1756, TTHA1013, and TTHA0281 revealed the good structural similarity with the dsRBD-like fold of β-β-β-α topology as a common fold [Fig. 1(C)]. However, the conformations of the C-terminal residues following the common α-helix (α1) are quite different in these proteins, and the α-helix (α0) preceding the dsRBD-like fold is specific to TTHA0281.

TTHA1013, which belongs to the DUF1902 protein family,[3] also forms a dimer around a crystallographic two-fold axis, which generates an eight-stranded β-sheet [Fig. 1(D)].[19] The dimer interface consists of only the β4 strands of the two subunits, indicating that domain swapping has occurred in TTHA1756. The TTHA1013 dimer is not superimposable on the TTHA1756 one [Fig. 1(E)], since the β-sheet platform of TTHA1013 is concave whereas that of TTHA1756 is convex. On the other hand, TTHA0281, a member of the UPF0150 family,[3] forms a tetramer with a back-to-back association of two dimers, that is, AB and CD [Fig. 1(F)].[7] The dimer assembly of TTHA0281 is totally different from that of TTHA1756 [Fig. 1(F)], and is mediated by the N-terminal α0 helices.[7] When the TTHA0281 monomer is superimposed on the TTHA1756 dimer [Fig. 1(G)], α0 of TTHA0281 lies along with β-strands of the TTHA1756 dimer, which probably hinders dimer formation, as observed in TTHA1756. Collectively, the UPF0150 protein family has two types of quaternary structure.

Apart from the central region (β2, β3, and α1), the N- and C-terminal regions consist of less-conserved residues [Fig. 1(A)], and the conformations of both regions of TTHA1756 and TTHA0281 differ significantly [Fig. 1(C)]. These findings indicate that a less-conserved region in the UPF0150 family proteins contributes to the variety in the quaternary structure in the same protein family and to expansion of the protein folding pattern. In the N-terminal region, TTHA0933, TTHA1912, and TTHA0231 are more similar to TTHA1756 than TTHA0281. The UPF0150 family proteins from *T. thermophilus* HB8 except for TTHA0281 may each form a dimer.

Recently, Makarova et al.[4] predicted, from comparative analysis of protein sequence and structure, that the HicB protein (UPF0150 family protein) is involved in RNA-binding and cleavage. Notably, Glu-35 of TTHA1756, which corresponds to one of the catalytic residues of nuclease families with the RNase H-fold,[4] is highly conserved in UPF0150 family proteins and is located on the surface [Fig. 1(B, right)]. Binding of either RNA or DNA using the dsRBD-like fold is mediated by a different contact surface of the same scaffold.[20,21] The variation in the quaternary structure of the UPF0150 family proteins may affect the preference of nucleic acid polymers, which underlies the biological role of this family protein.

ACKNOWLEDGMENTS

The authors thank Y. Nakamura, R. Hirose, and K. Hamada for the assistance during the data collection at SPring-8, and N. Takahashi for the structural refinement.

REFERENCES

1. Zhang C, Kim SH. Overview of structural genomics: from structure to function. Curr Opin Chem Biol 2003;7:28–32.
2. Yakunin AF, Yee AA, Savchenko A, Edwards AM, Arrowsmith CH. Structural proteomics: a tool for genome annotation. Curr Opin Chem Biol 2004;8:42–48.
3. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32:D138–D141.
4. Makarova KS, Grishin NV, Koonin EV. The HicAB cassette, a putative novel. RNA-targeting toxin-antitoxin system in archaea and bacteria. Bioinformatics 2006;22:2581–2584.
5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
6. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
7. Okazaki N, Kumei M, Manzoku M, Kuramitsu S, Shirouzu M, Shinkai A, Yokoyama S. Structure of a UPF0150-family protein from *Thermus thermophilus* HB8. Acta Crystallogr F 2007;63:173–177.
8. Ueno G, Kanda H, Hirose R, Ida K, Kumasaka T, Yamamoto M. RIKEN structural genomics beamlines at the SPring-8: high throughput protein crystallography with automated beamline operation. J Struct Funct Genomics 2006;7:15–22.
9. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol 1997;276:307–326.
10. Terwilliger TC, Berendzen J. Automated MAD and MIR structure solution. Acta Crystallogr D 1999;55:849–861.
11. Terwilliger TC. Maximum-likelihood density modification. Acta Crystallogr D 2000;56:965–972.
12. Perrakis A, Morris R, Lamzin VS. Automated protein model building combined with iterative structure refinement. Nat Struct Biol 1999;6:458–463.
13. McRee DE. XtalView/Xfit-A versatile program for manipulating atomic coordinates and electron density. J Struct Biol 1999;125:156–165.
14. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ,

Rice LM, Simonson T, Warren GL. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D 1998;54:905–921.
15. Collaborative Computational Project N. The CCP4 suite: programs for protein crystallography. Acta Crystallogr D 1994;50:760–763.
16. Potterton L, McNicholas S, Krissinel E, Gruber J, Cowtan K, Emsley P, Murshudov GN, Cohen S, Perrakis A, Noble M. Developments in the CCP4 molecular-graphics project. Acta Crystallogr D 2004;60: 2288–2294.
17. DeLano WL. The PyMOL Molecular Graphics System. Palo Alto, CA: DeLano Scientific. Available at: http://wwwpymolorg 2002.
18. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res 1998;26:316–319.
19. Hattori M, Mizohata E, Manzoku M, Bessho Y, Murayama K, Terada T, Kuramitsu S, Shirouzu M, Yokoyama S. Crystal structure of the hypothetical protein TTHA1013 from *Thermus thermophilus* HB8. Proteins 2005;61:1117–1120.
20. Ryter JM, Schultz SC. Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. EMBO J 1998;17:7505–7513.
21. Connolly KM, Wojciak JM, Clubb RT. Site-specific DNA binding using a variation of the double stranded RNA binding motif. Nat Struct Biol 1998;5:546–550.
22. 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;247:536–540.