Acta Crystallographica Section F
Structural Biology and Crystallization Communications
ISSN 1744-3091

The first crystal structure of an archaeal helical repeat protein

Kazunari Yoneda,[a] Haruhiko Sakuraba,[a] Hideaki Tsuge,[b] Nobuhiko Katunuma,[b] Seiki Kuramitsu,[c] Takeshi Kawabata[d] and Toshihisa Ohshima[a]*

[a]Department of Biological Science and Technology, Faculty of Engineering, University of Tokushima, Tokushima 770-8506, Japan, [b]Institute for Health Sciences, Tokushima Bunri University, Tokushima 770-8514, Japan, [c]Department of Biology, Graduate School of Science, Osaka University, Japan, and [d]Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan.

Correspondence e-mail: ohshima@bio.tokushima-u.ac.jp

Received 30 April 2005
Accepted 17 June 2005
Online 30 June 2005

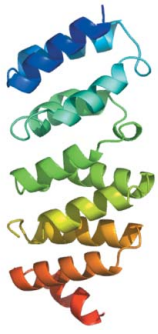
PDB Reference: ST1625p, 1wy6, r1wy6sf.

The crystal structure of ST1625p, a protein encoded by a hypothetical open reading frame ST1625 in the genome of the hyperthermophilic archaeon *Sulfolobus tokodaii*, was determined at 2.2 Å resolution. The only sequence similarity exhibited by the amino-acid sequence of ST1625p was a 33% identity with the sequence of SSO0983p from *S. solfataricus*. The 19 kDa monomeric protein was observed to consist of a right-handed superhelix assembled from a tandem repeat of ten α-helices. A structural homology search using the *DALI* and *MATRAS* algorithms indicates that this protein can be classified as a helical repeat protein.

1. Introduction

A large amount of genomic sequence information has been provided by completed and ongoing genome-sequencing projects. However, the function of many proteins can still not be deduced from comparative sequence analyses (Altschul *et al.*, 1997) owing to the absence of reliable sequence similarities to proteins with known functions. Structural proteomics aims to determine their three-dimensional structures. Although it is difficult for us to infer molecular function when the protein structure has an uncharacterized fold (Yang *et al.*, 1998; Colovos *et al.*, 1998; Shin *et al.*, 2002), accumulation of three-dimensional structural information about functionally unknown proteins is an essential and important step in the effort to predict their function.

ST1625p is a functionally unidentified protein encoded by a hypothetical open reading frame (ORF ID ST1625) in the genome of the hyperthermophilic archaeon *Sulfolobus tokodaii* (Kawarabayasi *et al.*, 2001). The only sequence similarity exhibited by ST1625p was a 33% amino-acid identity with the sequence of SSO0983p from the hyperthermophilic archaeon *S. solfatalicus* (Fig. 1). In addition, we have found that the ORFs of ST1625p and SSO0983p consist of similar gene clusters to the genes ST1624 and SSO0985, respectively. The ST1624 and SSO0985 genes encode homologues of the membrane-bound dye-linked D-proline dehydrogenase, which was originally found in the hyperthermophilic archaeon *Pyrobaculum*



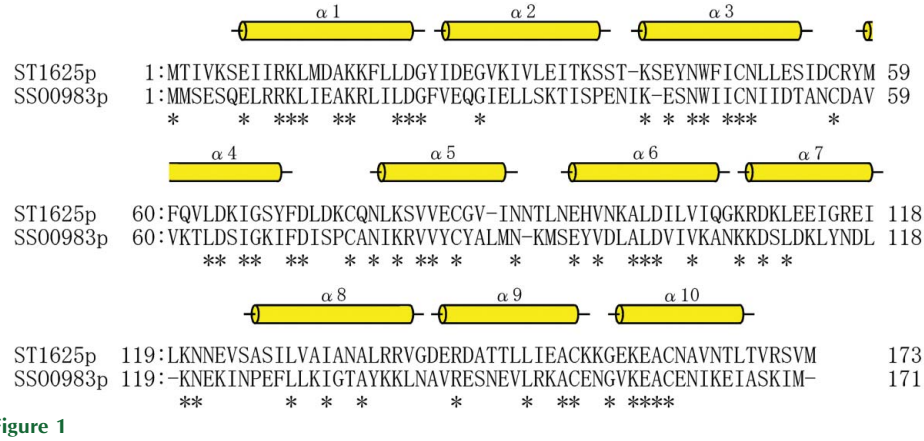


Figure 1
Amino-acid sequence alignment of *S. tokodaii* ST1625p and *S. solfataricus* SSO0983p hypothetical proteins. The two homologues were identified by an aligned *PSI-BLAST* search. α-Helices α1–α10 are shown. Asterisks represent the conserved residues in the proteins.

islandicum and characterized by members of our laboratory (Satomura *et al.*, 2002). We have identified that ST1624p functions as a novel membrane-bound dye-linked D-amino-acid dehydrogenase similar to the *P. islandicum* dye-linked D-proline dehydrogenase (unpublished data). Although the functions of ST1625p and SSO0983p are still unknown, their gene localization suggests that ST1625p and SSO0983p have a function related to the membrane-bound dye-dependent D-amino-acid dehydrogenases. In the present study, we determined the crystal structure of the hypothetical protein ST1625p at 2.2 Å resolution. As a result, it was revealed that the protein consists of a unique superhelix with a low-level structure resemblance to domains from other proteins with known three-dimensional structures.

2. Materials and methods

2.1. Cloning and expression

The following set of oligonucleotide primers was used to amplify the ST1625 gene fragment by PCR: 5′-ATAT**CATATG**ACCATCGTAAAAAGCGAAATAATTCGCAAA-3′, containing a unique *Nde*I restriction site overlapping the 5′ initial codon, and 5′-ATAT**GGATCC**GCGGCCGCTTATTACATGACACTTCTTACTGT-3′, containing a unique *Bam*HI restriction site proximal to the 3′-end of the termination codon. The chromosomal *S. tokodaii* DNA was isolated as described by Yamagishi & Oshima (1990) and used as the template. The amplified 0.5 kbp fragment was digested with *Nde*I and *Bam*HI and ligated with the expression vector pET-11a linearized with *Nde*I and *Bam*HI to generate pEST1625. *Escherichia coli* strain Rosetta-gami(DE3) was transformed with pEST1625. The transformants were cultivated in 3 l medium containing 15 g polypeptone, 30 g yeast extract, 60 g glycerol, 30 g lactose, 15 g NaCl and 50 µg ml^{-1} ampicillin for 24 h at 310 K.

2.2. Purification of protein

E. coli cells (∼91 g wet weight from a 3 l culture) were harvested by centrifugation, suspended in 20 m*M* Tris–HCl buffer pH 8.0 containing 5 m*M* β-mercaptoethanol and disrupted by ultrasonication. The entire operation was performed at room temperature (∼298 K) and the fractions containing ST1625p were checked by SDS–PAGE in all purification steps. The crude extract was heated at 358 K for 20 min and the denatured protein was then removed by centrifugation (100 000*g*, 2 h). Solid $(NH_4)_2SO_4$ was added to the protein solution to 1.2 *M*. The protein solution was loaded onto a ToyoScreen PPG 600M column (14.6 × 30 mm; Tosoh) equilibrated with 20 m*M* Tris–HCl buffer pH 8.0 containing 5 m*M* β-mercaptoethanol and 1.2 *M* $(NH_4)_2SO_4$. After the column had been washed with the same buffer (around three column bed volumes), the protein was eluted with a linear gradient of 1.2–0 *M* $(NH_4)_2SO_4$ in the same buffer. The fractions containing ST1625p were collected and dialyzed against 20 m*M* Tris–HCl buffer pH 8.0 containing 5 m*M* β-mercaptoethanol. Solid $(NH_4)_2SO_4$ was added to the protein solution to 1.5 *M*. The protein solution was loaded onto a ToyoScreen PPG 600M column (14.6 × 30 mm; Tosoh) equilibrated with 20 m*M* Tris–HCl buffer pH 8.0 containing 5 m*M* β-mercaptoethanol and 1.5 *M* $(NH_4)_2SO_4$. After the column had been washed with the same buffer (around three column bed volumes), the protein was eluted with a linear gradient of 1.5–0 *M* $(NH_4)_2SO_4$ in the same buffer. The ST1625p-containing fractions were collected and dialyzed against 20 m*M* Tris–HCl buffer pH 8.0 containing 5 m*M* β-mercaptoethanol. The protein solution was loaded onto a Resource Q column (16 × 30 mm Amersham Biosciences) equilibrated with 20 m*M* Tris–HCl buffer pH 8.0 containing 5 m*M* β-mercaptoethanol. After the column had been washed with the same buffer (around three column bed volumes), the protein was eluted with a linear gradient of 0–1 *M* NaCl in the same buffer. The fractions containing ST1625p were collected and dialyzed against 20 m*M* potassium phosphate buffer pH 7.0 containing 5 m*M* β-mercaptoethanol. The protein solution was loaded onto a Bioscale column (12 × 88 mm; Bio-Rad) equilibrated with 10 m*M* potassium phosphate buffer pH 7.0 containing 5 m*M* β-mercaptoethanol. After the column had been washed with the same buffer (around three column bed volumes), the protein was eluted with a linear gradient of 10–250 m*M* potassium phosphate in the same buffer. The fractions containing ST1625p were collected and dialyzed against 20 m*M* Tris–HCl buffer pH 8.0. The protein solution was loaded onto a Superdex 200 column (10 × 300 mm; Amersham Biosciences) equilibrated with 20 m*M* Tris–HCl buffer pH 8.0 containing 150 m*M* NaCl and 5 m*M* β-mercaptoethanol and the protein was eluted with the same buffer. The ST1625p fractions were collected, concentrated with an Amicon Ultra PL-5 (Millipore) and used for crystallization.

Table 1
Summary of crystal parameters, data-collection and refinement statistics.

Values in parentheses are for the last resolution shell.

	Native	Au
Crystal characteristics		
Space group	$P2_12_12_1$	
Unit-cell parameters (Å)	a = 36.40, b = 50.35, c = 82.36	
MIRAS data		
Wavelength (Å)	1.5418	1.5418
Resolution range (Å)	43.0–2.2 (2.28–2.20)	43.0–2.38 (2.48–2.38)
Total reflections	127292	154273
Unique reflections	8157	7211
Completeness (%)	95.0 (92.4)	99.0 (100)
R_{sym}†	0.038 (0.096)	0.089 (0.20)
FOM‡		0.46
DM-FOM§		0.61
Refinement statistics		
Resolution range (Å)	20.0–2.2	
Unique reflections	7378	
R factor/free R factor¶	0.222/0.275	
No. of protein atoms	1254	
No. of water molecules	71	
R.m.s. deviations from ideal geometry		
Bond lengths (Å)	0.006	
Bond angles (°)	1.0	
Average isotropic B value (Å^2)	21.8	

† $R_{sym} = \sum_h \sum_i |I_i(h) - \langle I(h)\rangle| / \sum_h \sum_i |I_i(h)|$, where $I_i(h)$ is the intensity measurement for a reflection h and $\langle I(h)\rangle$ is the mean intensity for this reflection. ‡ FOM is the mean figure of merit after SIRAS phasing. § DM-FOM is the mean figure of merit after density modification. ¶ R factor $= \sum_h ||F_{obs}| - |F_{calc}|| / \sum_h |F_{obs}|$. The free R factor was calculated with randomly selected reflections (10%).

2.3. Crystallization and data collection

ST1625p was crystallized at room temperature by the sitting-drop vapour-diffusion method. 1 µl protein solution (7.3 mg ml^{-1}) in buffer containing 20 m*M* Tris–HCl pH 8.0, 150 m*M* NaCl and 5 m*M* β-mercaptoethanol was mixed with 1 µl mother liquor containing 100 m*M* phosphate/citrate buffer pH 4.2, 100 m*M* NaCl and 20% PEG 8000. Needle-shaped crystals appeared within 3 d and grew in a week to approximate dimensions of 0.1 × 0.1 × 1 mm.

The crystals were coated with a layer of viscous oil (1:1 mixture of Paratone-N and mineral oil) and transferred into a stream of nitrogen gas for data collection at 100 K. Diffraction data were collected at a 2.2 Å resolution on an R-AXIS VII imaging-plate detector using a rotating copper-anode in-house generator (Rigaku MicroMax-007) operating at 40 kV and 20 mA. The oscillation angle per image was set to 1°. The data were processed using *HKL*2000 v.0.97.647

(Otwinowski & Minor, 1997). The ST1625p crystals belonged to the orthorhombic space group $P2_12_12_1$. Assuming the presence of one molecule in the asymmetric unit, the value of the Matthews coefficient (V_M; Matthews, 1968) and the solvent content were calculated to be 1.9 Å^3 Da^{-1} and 35.9%, respectively. These values are within the range frequently observed for protein crystals. A heavy-atom derivative was prepared by soaking the crystals in a mother liquor containing 0.1 m*M* $HAuCl_4$ for 14 h. Data were collected at a 2.38 Å resolution by the same method as described above.

2.4. Structure determination and refinement

Native and gold data sets were used for phase calculation (Table 1) by the SIRAS (single isomorphous replacement with an anomalous scattering) method using *SOLVE* v.2.02 (Terwilliger & Berendzen, 1999). The SIRAS map at 2.2 Å was subjected to maximum-likelihood density modification, followed by autotracing using *RESOLVE* v.2.02 (Terwilliger, 1999). An initial model was built using *XtalView* v.4.0 (McRee, 1999). Several cycles of rigid-body refinement, positional refinement and simulated annealing were performed at 2.2 Å resolution with *CNS* v.1.1 (Brünger *et al.*, 1998). The model was adjusted in *XtalView* using both $|F_o| - |F_c|$ and $2|F_o| - |F_c|$ maps. The current model contains residues 7–119, 122–167 and 71 water molecules, with an *R* factor of 22.2% and a free *R* factor of 27.5%. The model geometry was analyzed with *PROCHECK* v.3.5.4 (Laskowski *et al.*, 1993) and 91.8% of the non-glycine residues were in the most favoured region of the Ramachandran plot and 8.2% were in the additionally allowed region.

2.5. Structure comparison

The structure of ST1625p was compared with the structures in the PDB (as of 9 November 2004) using *DALI* v.2.0 (Holm & Sander, 1998) and *MATRAS* v.1.2 (Kawabata, 2003). *DALI* identified five structural homologues of ST1625p. The r.m.s. deviations are 7.8 (for 114 pairs of aligned C^α atoms for clathrin-assembly protein short form fragment; 1hg5 chain *A*; *Z* = 6.2), 5.0 (for 90 pairs of aligned C^α atoms for protein farnesyltransferase; 1ft1 chain *A*; *Z* = 6.2), 5.2 (for 139 pairs of aligned C^α atoms for adaptor-related protein complex 2 α2 subunit; 1gw5 chain *B*; *Z* = 5.8), 4.7 (for 139 pairs of aligned C^α atoms for β-catenin fragment; 3bct; *Z* = 5.7) and 3.9 (for 85 pairs of aligned C^α atoms for leucine-rich repeat variant; 1lrv; *Z* = 4.5). *MATRAS* also identified five structural homologues. The r.m.s. deviations are 4.6 (for 149 pairs of aligned C^α atoms for B-cell lymphoma 3-encoded protein; 1k1a chain *A*; *Z* = 23.4), 4.7 (for 155 pairs of aligned C^α atoms for mo25 protein; 1upk chain *A*; *Z* = 21.3), 5.2 (for 153 pairs of aligned C^α atoms for adaptor-related protein complex 2 α2 subunit; 1gw5 chain *A*; *Z* = 20.8), 4.6 (for 150 pairs of aligned C^α atoms for 26S proteasome non-ATPase regulatory subunit 10; 1uoh chain *A*; *Z* = 20.3) and 4.7 (for 151 pairs of aligned C^α atoms for I-κ-B-α; 1ikn chain *D*; *Z* = 20.1) (Fig. 2).

3. Results and discussion

3.1. Structure and fold

The overall structure of ST1625p exhibited a pill bug-like shape with approximate dimensions of 14 × 24 × 59 Å (Fig. 2*a*). The overall fold generated by the right-handed superhelix consisted of ten α-helices, with the complete absence of any β-sheet structure. The secondary-structure elements were as follows: α1, residues 7–21; α2, residues 24–37; α3, residues 40–53; α4, residues 59–69; α5, residues 78–87; α6, residues 93–105; α7, residues 108–118; α8, residues 126–139; α9, residues 142–153; α10, residues 157–167. In addition, we found that the three disulfide bonds were formed at Cys48–Cys75, Cys56–Cys84 and Cys153–Cys161.

3.2. Comparison with known protein structures

The structural homologues obtained by *MATRAS* do not always coincide with those found by *DALI*. In addition, the structural homologues found by *MATRAS* and *DALI* show no similarity to each other in amino-acid sequence. All of these proteins are classified as helical repeat proteins assembled from tandem repeats of an α-helical structural unit (Murzin *et al.*, 1995). In general, helical repeat proteins contain a repeating amino-acid motif. For example, the armadillo repeat contains a 42-amino-acid repeat (ARM motif; Matthew & David, 1999) and the leucine-rich repeat contains an L*XX*L*X*L repeat (LRV motif; Kajava & Kobe, 2002). Thus, we analyzed the repeating amino-acid motif of ST1625p with the

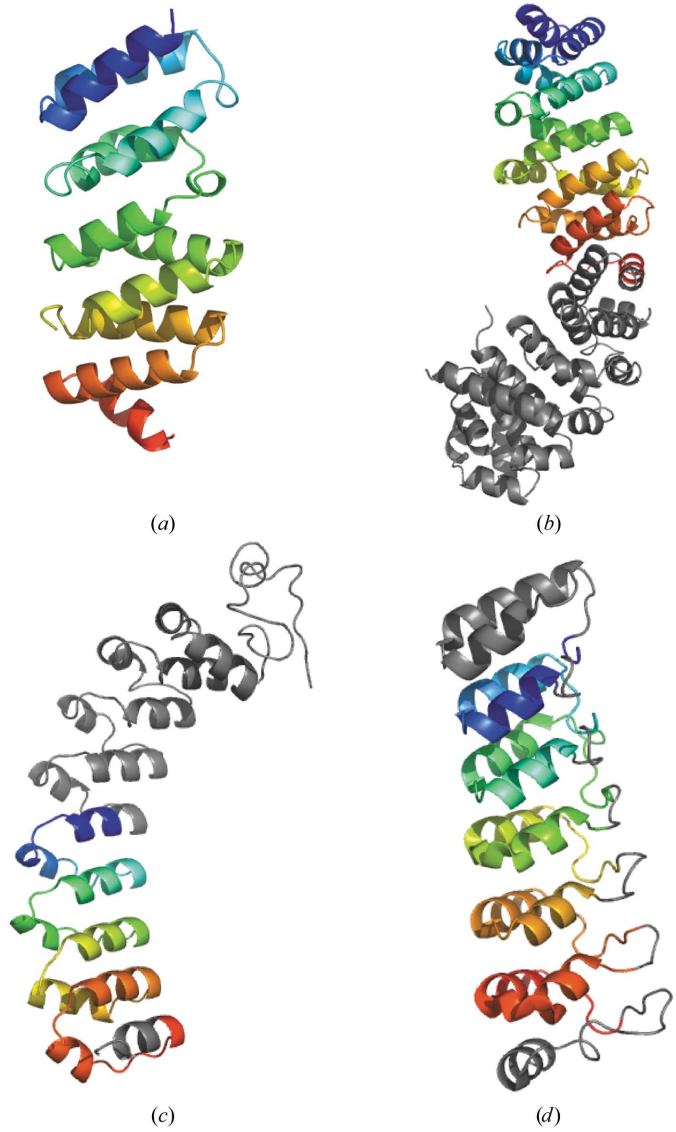


Figure 2
Crystal structure of ST1625p and structural homologues identified by *DALI* (Holm & Sander, 1998) and *MATRAS* (Kawabata, 2003). The rainbow colouring identifies the parts sharing similar structure with ST1625p. (*a*) ST1625p, (*b*) β-catenin fragment (PDB code 3bct; Huber *et al.*, 1997), (*c*) leucine-rich repeat variant (PDB code 1lrv; Peters *et al.*, 1996) and (*d*) B-cell lymphoma 3-encoded protein (PDB code 1k1a; Michel *et al.*, 2001).

REPALIGN (http://act.jst.go.jp/) program. No repeating amino-acid motif was observed in the amino-acid sequence. The presence of disulfide bonds has not yet been identified in these structural homologues. These results show that the α–α superhelix fold of ST1625p is distinct from those of the common helical repeat proteins.

Many types of crystal structures of eukaryotic helical repeat proteins have already been reported, but only limited crystal structure information is known for proteins from prokaryotes (Matthew & David, 1999). The presence of such a helical repeat protein has not yet been described in the archaea, the third domain of life. Although the molecular function of the protein is currently unknown, this structure could be a novel template for a helical repeat protein family.

This work was supported in part by a grant from the 'National Project on Protein Structural and Functional Analysis' promoted by the Ministry of Education, Science, Sports and Culture of Japan.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Colovos, C., Cascio, D. & Yeates, T. O. (1998). *Structure*, **6**, 1329–1337.
Holm, L. & Sander, C. (1998). *Nucleic Acids Res.* **26**, 316–319.
Huber, A. H., Nelson, W. J. & Weis, W. I. (1997). *Cell*, **90**, 871–882.
Kajava, A. V. & Kobe, B. (2002). *Protein Sci.* **11**, 1082–1090.
Kawabata, T. (2003). *Nucleic Acids Res.* **31**, 3367–3369.
Kawarabayasi, Y. *et al.* (2001). *DNA Res.* **8**, 123–140.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.
Matthew, R. G. & David, B. (1999). *Curr. Opin. Struct. Biol.* **9**, 383–389.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Michel, F., Soler-Lopez, M., Petosa, C., Cramer, P., Siebenlist, U. & Muller, C. W. (2001). *EMBO J.* **20**, 6180–6190.
Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Peters, J. W., Stowell, M. H. & Rees, D. C. (1996). *Nature Struct. Biol.* **3**, 991–994.
Satomura, T., Kawakami, R., Sakuraba, H. & Ohshima, T. (2002). *J. Biol. Chem.* **277**, 12861–12867.
Shin, D. H., Yokota, H., Kim, R. & Kim, S.-H. (2002). *J. Struct. Funct. Genomics*, **2**, 53–66.
Terwilliger, T. C. (1999). *Acta Cryst.* D**55**, 1863–1871.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Yamagishi, A. & Oshima, T. (1990). *Nucleic Acids Res.* **18**, 1133–1136.
Yang, F., Gustafson, K. R., Boyd, M. R. & Wlodawer, A. (1998). *Nature Struct. Biol.* **5**, 763–764.