Crystal structure of the conserved protein TT1542 from *Thermus thermophilus* HB8

NORIKO HANDA,¹ TAKAHO TERADA,^{1,2} YUKI KAMEWARI,¹ HIROAKI HAMANA,¹
JEREMY R.H. TAME,³ SAM-YONG PARK,³ KENGO KINOSHITA,^{1,3,4,9}
MOTONORI OTA,⁵ HARUKI NAKAMURA,^{1,6} SEIKI KURAMITSU,^{2,7}
MIKAKO SHIROUZU,^{1,2} AND SHIGEYUKI YOKOYAMA^{1,2,8}

¹RIKEN Genomic Sciences Center, Tsurumi, Yokohama 230-0045, Japan

²RIKEN Harima Institute at SPring-8, Mikazuki-cho, Sayo, Hyogo 679-5148, Japan

³Protein Design Laboratory, Yokohama City University, Tsurumi, Yokohama 230-0045, Japan

⁴Structure and Function of Biomolecules, PRESTO, Japan Science and Technology Corporation, Kawaguchi, Saitama 332-0012, Japan

⁵Global Scientific Information and Computing Center, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan

⁶Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan

⁷Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan

⁸Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan

⁹Structural Bioinformatics Laboratory, Yokohama City University, Tsurumi, Yokohama 230-0045, Japan

(RECEIVED April 9, 2003; FINAL REVISION May 15, 2003; ACCEPTED May 15, 2003)

Abstract

The TT1542 protein from *Thermus thermophilus* HB8 is annotated as a conserved hypothetical protein, and belongs to the DUF158 family in the Pfam database. A BLAST search revealed that homologs of TT1542 are present in a wide range of organisms. The TT1542 homologs in eukaryotes, PIG-L in mammals, and GPI12 in yeast and protozoa, have *N*-acetylglucosaminylphosphatidylinositol (GlcNAc-PI) de-*N*-acetylase activity. Although most of the homologs in prokaryotes are hypothetical and have no known function, Rv1082 and Rv1170 from *Mycobacterium tuberculosis* are enzymes involved in the mycothiol detoxification pathway. Here we report the crystal structure of the TT1542 protein at 2.0 Å resolution, which represents the first structure for this superfamily of proteins. The structure of the TT1542 monomer consists of a twisted β-sheet composed of six parallel β-strands and one antiparallel β-strand (with the strand order 3-2-1-4-5-7-6) sandwiched between six α-helices. The N-terminal five β-strands and four α-helices form an incomplete Rossmann fold-like structure. The structure shares some similarity to the sugar-processing enzymes with Rossmann fold-like domains, especially those of the GPGTF (glycogen phosphorylase/glycosyl transferase) superfamily, and also to the NAD(P)-binding Rossmann fold domains. TT1542 is a homohexamer in the crystal and in solution, the six monomers forming a cylindrical structure. Putative active sites are suggested by the structure and conserved amino acid residues.

Keywords: *Thermus thermophilus* HB8; conserved protein TT1542; GlcNAc-PI de-*N*-acetylase homolog; GlcNAc-Ins de-*N*-acetylase homolog; mycothiol *S*-conjugate amidase homolog; DUF158; crystallography

Reprint requests to: Shigeyuki Yokoyama, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan; e-mail: yokoyama@biochem.s.u-tokyo.ac.jp; fax: 81-45-503-9195.

Article and publication are at <http://www.genesdev.org/cgi/doi/10.1110/gad.03104003>.

An open reading frame (ORF) of *Thermus thermophilus* HB8 encodes a conserved hypothetical protein, TT1542 (GenBank accession no. AB101292), which consists of 227 amino acid residues. A PSI-BLAST search (Altschul et al. 1997) of the TT1542 sequence identified more than 100

proteins from prokaryotes and eukaryotes, with sequence identities ranging from 15% to 50% and E-values below 4×10^{-17} (Fig. 1). The TT1542 homologs from eukaryotes, PIG-L in mammals, and GPI12 in yeast and protozoa, are GlcNAc-PI de-*N*-acetylases (Nakamura et al. 1997; Watanabe et al. 1999; Chang et al. 2002), catalyzing the second step of glycosylphosphatidylinositol (GPI) membrane anchor biosynthesis (Stevens 1995; Kinoshita and Inoue 2000). The GPI anchor modification is widespread among eukaryotes and probably some archaea, but has not been found in bacteria (Eisenhaber et al. 2001). TT1542 is therefore unlikely to share a function with the eukaryotic homologs. Two TT1542 homologs from *Mycobacterium tuberculosis*, Rv1082 and Rv1170, have been shown to be a mycothiol *S*-conjugate amidase and a 1-*D*-myo-inositol-2-acetamido-2-deoxy- α -*D*-glucopyranoside (GlcNAc-Ins) de-*N*-acetylase, respectively (Newton et al. 2000a,b). They are involved in the mycothiol-dependent detoxification pathway found in most actinomycetes, but not in other microbes and eukaryotes (Newton et al. 1996; Fahey 2001; Newton and Fahey 2002), so TT1542 is also unlikely to function as a mycothiol *S*-conjugate amidase or a GlcNAc-Ins de-*N*-acetylase. All of the functionally annotated TT1542 homologs are hydrolases targeting the C2-amide bond of glucosamine in various disaccharides. TT1542 and most of its prokaryotic homologs are annotated as hypothetical proteins. In the Pfam database (Bateman et al. 2002), TT1542 and its homologs in many prokaryotes and fission yeast belong to the PF02585 or DUF158 family, which is equivalent to COG2120 in the National Center for Biotechnology Information database of Clusters of Orthologous Groups (Tatusov et al. 2001). In the present study, we solved the crystal structure of the conserved hypothetical protein, TT1542, and determined its molecular weight in solution by analytical ultracentrifugation. This is the first reported structure of a member of this protein superfamily, and here we discuss its structural characteristics.

Results

Structure determination

The crystals belong to the space group *P321*, with unit cell constants of $a = b = 107.13$ Å, $c = 98.62$ Å. There are

two monomers in an asymmetric unit. Two models, A and B, were built using TURBO-FRODO (Fig. 2). These are essentially identical, with a root-mean-square deviation (RMSD) of only 0.23 Å for the main chain atoms. Crystallographic statistics are given in Table 1.

Monomer structure

The monomer structure consists of a mixed β -sheet sandwiched between a total of six α -helices (Fig. 3A). The topology is shown in Figure 3B. The β -sheet contains six parallel β -strands ($\beta 1$ – $\beta 6$) and one antiparallel β -strand ($\beta 7$), in the order $\beta 3$ – $\beta 2$ – $\beta 1$ – $\beta 4$ – $\beta 5$ – $\beta 7$ – $\beta 6$. The topology of the first five β -strands ($\beta 1$ – $\beta 5$) and the first four α -helices ($\alpha 1$ – $\alpha 4$) is similar to the incomplete Rossmann fold (Rossmann et al. 1975). The typical Rossmann fold consists of six parallel β -strands in the order $\beta 3$ – $\beta 2$ – $\beta 1$ – $\beta 4$ – $\beta 5$ – $\beta 6$, but variants lacking $\beta 6$ are common. The remaining C-terminal region of TT1542 forms an additional structure that consists of two β -strands ($\beta 6$ and $\beta 7$) and two α -helices ($\alpha 5$ and $\alpha 6$), in the order $\beta 6$ – $\alpha 5$ – $\alpha 6$ – $\beta 7$, followed by a hook-like tail, and is attached to the incomplete Rossmann fold structure (Fig. 3A,B). Hydrophobic amino acids are clustered in the C-terminal tail (Figs. 1, 3A). Most of the enzymes with Rossmann fold-like structures have a ligand-binding site along the C-terminal edge of the β -sheet, mainly in the loops between the β -strands and the α -helices (Baker et al. 1992). TT1542 also has a hydrophilic cavity at this site (Fig. 3A).

A search of the Protein Data Bank with the program DALI (Holm and Sander 1997) revealed that TT1542 exhibits the highest structure similarity to two proteins, UDP-*N*-acetylglucosamine (UDP-GlcNAc) 2-epimerase (1f6d-A, DALI Z-score = 7.1, RMSD = 3.8 over 138 C α residues; Fig. 3C; Campbell et al. 2000) and *Escherichia coli* MurG, which is a GlcNAc transferase (1f0k-A, DALI Z-score = 6.8, RMSD = 3.3 over 127 C α residues; Ha et al. 2000). However, their amino acid sequences share only weak similarity with that of TT1542 (identities of 9% and 13%, respectively). These proteins belong to a large superfamily, termed GPGTF, of diverse sugar processing enzymes (Wrabl and Grishin 2001). They have two Rossmann fold-like domains in the monomer proteins (Fig. 3C; Wrabl and Grishin 2001). The proteins with the next highest simi-

Figure 1. Sequence comparison between TT1542 and its homologs. The figure was constructed by PSI-BLAST followed by CLUSTAL W (Thompson et al. 1994). The first column shows the protein identifier. The first 13 proteins are annotated as hypothetical or uncharacterized proteins: Ba, *Bacillus anthracis*; Dr, *Deinococcus radiodurans*; Pa, *Pyrococcus abyssi*; Ch, *Cytophaga hutchinsonii*; Ma, *Methanosarcina acetivorans*; Ct, *Chlorobium tepidum*; Tt, *Thermoanaerobacter tengcongensis*; Mk, *Methanopyrus kandleri*; Ca, *Chloroflexus aurantiacus*; Sm, *Sinorhizobium meliloti*; Yp, *Yersinia pestis*; Sr, *Streptomyces rochei*; Np, *Nostoc punctiforme*. Rv1082 and Rv1170 from *Mycobacterium tuberculosis* (Mt) are a mycothiol *S*-conjugate amidase and a GlcNAc-Ins deacetylase, respectively. The GPI12 proteins from *Saccharomyces cerevisiae* (Sc), *Leishmania major* (Lm), and *Trypanosoma brucei* (Tb), and the PIG-L proteins from *Rattus norvegicus* (Rn) and *Homo sapiens* (Hs) are GlcNAc-PI de-*N*-acetylases. “I,” “H,” and “E” represent % identity, % homology, and E-values from the PSI-BLAST results, respectively. Strictly conserved and similar residues are represented within a red box and with red letters, respectively. Residues predicted as “catalytic” by the computational analysis (Ota et al. 2003) are indicated by red circles.

Figure 2. σ_A -weighted $2F_o - F_c$ map showing the electron density for residues in the putative active site (stereoview). The electron density (blue) is contoured at 1.3σ .

larity are the RCK domain of the K^+ channel (1id1-A, DALI Z-score = 6.2, RMSD = 3.5 over 112 C α residues, 9% identity; Fig. 3D; Jiang et al. 2001) and UDP-galactose 4-epimerase (1xel, DALI Z-score = 6.1, RMSD = 3.7 over 141 C α residues, 11% identity; Thoden et al. 1996). They belong to a superfamily of NAD(P)-binding Rossmann fold domains (Baker et al. 1992). All of these Rossmann folds of the GPGTF and NAD(P)-binding superfamilies are the typical Rossmann fold with six parallel β -strands sandwiched by five α -helices. In contrast, the Rossmann fold-like region of TT1542 lacks the sixth β -strand and the fifth α -helix, while the rest of the region shows high similarity to the typical Rossmann folds (Fig. 3A,C,D). Furthermore, the C-terminal $\beta 6$ - $\alpha 5$ - $\alpha 6$ - $\beta 7$ -tail structure of TT1542 is completely different from the structures following the typical Rossmann folds (Fig. 3A,C,D). No protein is known to have an incomplete five-stranded Rossmann fold-like structure followed by β - α - α - β . The topology of TT1542 is therefore unique. The C-terminal region is involved in the monomer-monomer interactions in the hexamer (see below).

Quaternary structure

The two monomers (A and B) in an asymmetric unit are related by a noncrystallographic twofold axis perpendicular to the crystallographic threefold axis (Fig. 4A,C). The tails intertwine with each other, and the tail of one monomer contacts the $\beta 6$ and $\beta 7$ strands, and the $\alpha 6$ helix of the other monomer in the asymmetric unit (Fig. 4C). These dimer contacts are mainly hydrophobic and bury $1460 \text{ \AA}^2$ (13.1% of the surface area) per monomer.

In the crystal, TT1542 forms a homohexamer and can be viewed as a trimer of dimers (Fig. 4A,B,C). The A'B' and A''B'' dimers are related to the AB dimer by crystallographic

threefold symmetry (Fig. 4A). The interactions between the dimers bury a total surface area of $3640 \text{ \AA}^2$, 18.7% of the dimer surface area. The A-B', A'-B'', and A''-B interactions are formed by the $\alpha 6$ helices and the $\beta 5$ - $\beta 6$ loops. The A-A', A'-A'', A''-A, B-B', B'-B'', and B''-B interactions are formed by the packing of the $\alpha 3$ helix, the $\alpha 4$ helix, the 3_{10} helix, the $\alpha 4$ - $\beta 5$ loop, and the C-terminal tail in one monomer and the $\beta 2$ - $\alpha 2$ loop, the $\beta 3$ - $\alpha 3$ loop, the $\beta 4$ - $\alpha 4$ loop, the $\beta 5$ strand, and the $\beta 5$ - $\beta 6$ loop in the other monomer. Most of these contacts are hydrophobic. The hexamer forms a cylinder, and the hydrophilic cavity mentioned above is found at the side of this structure (Fig. 5A,B).

To determine the predominant species in solution, the molecular weight of TT1542 was measured using analytical ultracentrifugation. Both sedimentation velocity and equilibrium experiments were carried out, using both UV absorption and light interference. The sedimentation velocity data showed TT1542 sedimented as a single species, with molecular weight estimated to be 139,500 D. Sedimentation equilibrium yielded molecular weight values of 139,430 and 141,645 D using UV absorption and light interference, respectively. The expected molecular weight of the hexamer is 148,722 D. The low experimental values (with errors of about 7% and 5%) may be due to errors in the calculated partial specific volume due to the large buried surface area. The fit of the data to a single ideal species model is shown in Figure 6.

Location of the active site

More than 100 amino acid sequences similar to TT1542 were found using PSI-BLAST. These included hypothetical proteins in bacteria and archaea, PIG-L in mammals, GPI12 in yeast and protozoa, and Rv1170 and Rv1082 from *M. tuberculosis* (Fig. 1). The aligned sequences are much less conserved in the C-terminal half than in the N-terminal half. The sequence alignment reveals several highly conserved amino acids. The P⁹-H¹⁰-P¹¹-D¹²-D¹³ sequence of TT1542, which is highly conserved with the consensus (P/A)-H-(P/A)-D-D (where P/A can be either Pro or Ala), is found at the region of the $\beta 1$ - $\alpha 1$ loop and the start of the $\alpha 1$ helix (Fig. 1). In this sequence, Pro 9 and Pro 11 are involved in the $\beta 1$ - $\alpha 1$ turn, and the side chains of His 10, Asp 12, and Asp 13 point into the hydrophilic cavity (Figs. 2, 4B, 5B, 7A). The $\alpha 4$ - $\beta 5$ loop in the adjacent monomer participates in forming this cavity (Figs. 4B, 5B, 7B). In addition, Thr 38, Arg 51, Glu 54, Ala 58, Asp 74, His 108, Asp 110, and His 111 are highly conserved (Fig. 1). All of these conserved residues are clustered in the region, and most of their side chains face the hydrophilic cavity (Figs. 2, 7A), suggesting that it is likely to be the active site.

To consider this idea further, we performed a computational analysis to search for functionally important residues among the conserved residues (Ota et al. 2003).

Figure 3. (A) Ribbon diagram of the TT1542 monomer (stereoview). The α -helices are red, the β -strands are cyan, the 3_{10} helix is green, and the random coils are gray. The disordered region (between Thr 178 and Val 186) is represented by a dotted line. This figure was drawn using MOLSCRIPT (Kraulis 1991) and RASTER3D (Merritt and Bacon 1997). (B) Topology diagram of the TT1542 monomer. The α -helices are represented by pale orange cylinders, the β -strands are cyan arrows, and the 3_{10} -helix is a green cylinder. The TT1542 structure can be considered to consist of two parts: an N-terminal incomplete Rossmann fold consisting of a five-stranded parallel β -sheet sandwiched by a total of four helices, and a smaller C-terminal β - α - β fold attached to the N-terminal Rossmann fold. A total of seven β -strands make a twisted β -sheet. (C) Ribbon diagram of UDP-N-acetylglucosamine (UDP-GlcNAc) 2-epimerase. (D) Ribbon diagram of the RCK domain of the K^+ channel.

Figure 4. Ribbon diagrams of the hexameric structures of TT1542 (stereoview). (A) TT1542 hexamer with the threefold crystallographic symmetry axis (perpendicular to the plane of the paper). Monomers A, B, A', B', A'', and B'' are colored differently. (B) Different view of the TT1542 hexamer. The figure represents after a 90° rotation of (A) versus the *x* axis. (C) The figure represents after a 90° rotation of (B) versus the *y* axis. The twofold noncrystallographic symmetry axis, which relates monomer B to monomer A, is perpendicular to the plane of the paper.

As shown in Table 2, His 10, Asp 12, Asp 13, Glu 15, Arg 51, Glu 54, and Asp 74 are predicted to be catalytically important. To find possible ligands, we searched for complex structures with similar ligand binding sites in

the PDB (Kinoshita and Nakamura 2003). Several ligand binding sites have structures and electrostatic potentials similar to those of the TT1542 cavity. The ligands include thiazine- and thiazepine-based matrix metalloprote-

Figure 5. Distribution of the electrostatic potential on the solvent-accessible surface of the TT1542 hexamer (stereoview). Blue and red surfaces represent positive and negative potentials, respectively. (A) The figure is the same orientation as in Figure 4A. (B) The figure is the same orientation as in Figure 4B. The putative active sites are indicated by arrows.

ase (MMP) inhibitors, *N*-hydroxy-4-[(4-methoxyphenyl) sulfonyl]-2,2-dimethyl-hexahydro-1,4-thiazepine-3(*S*)-carboxamide (**4a**; PDB: 1d5j), *N*-hydroxy 1*N*-(4-methoxyphenyl) sulfonyl-4-(*Z,E*-*N*-methoxyimino)pyrrolidine-(2*R*)-carboxamide (**14**; PDB: 1d7x), *N*-[[2-methyl-4-hydroxycarbonyl]but-4-yl-*N*]-benzyl-*P*-[phenyl]-*P*-[methyl] phosphinamid (RGV-25727; PDB: 1b3d); *S*-adenosylmethionine (PDB: 1mjg, 1mjl, and 1mj2); factor Xa inhibitor, thieno[3,2-*B*]pyridine-2-sulfonic acid [1-(1-amino-isoquinolin-7-ylmethyl)-2-oxo-pyrrolidin-3-yl]-amide (RPR208815; PDB: 1f0r); diphenylether sulfone inhibitor, diphenyl-ether sulfone based hydroxamic acid (RS-130830; PDB: 966c); and glucose (PDB: 1jdd). Figure 7B shows the MMP inhibitor **4a** docked in the putative active site of TT1542.

Discussion

The PSI-BLAST search of the Swiss-Prot database using TT1542 found more than 100 related sequences in a wide

range of organisms, including prokaryotes and eukaryotes. Although most of the proteins are hypothetical and have no known function, several are hydrolases that cleave the C2-amide bond of glucosamine in various disaccharides. The three-dimensional structure of the TT1542 monomer shows that the conserved N-terminal region forms an incomplete Rossmann fold, which is similar to the structures of several sugar-processing enzymes. The putative active site of the TT1542 monomer is at the expected position for Rossmann-fold enzymes. Unfortunately, the ligand of a sugar-processing enzyme is not easy to identify from either sequence or structural information. The similarity search against the binding sites found in the PDB yielded several putative ligands, most with some similarities to disaccharides. Our findings suggest that TT1542 is a hydrolase of some disaccharide.

The highest scoring structural homologs, UDP-GlcNAc 2-epimerase and the GlcNAc transferase MurG, belong to the GPGTF superfamily of enzymes (Campbell et al. 2000; Ha et al. 2000; Wrabl and Grishin 2001), which have two Rossmann fold-like structures in each monomer. The structure of TT1542 is similar to the N-terminal domain in these proteins, but less so to the C-terminal domain. The structure of UDP-GlcNAc 2-epimerase complexed with UDP shows that the C-terminal domain contains the UDP binding site, and the N-terminal domain is expected to bind the GlcNAc portion of its ligand, UDP-GlcNAc (Campbell et al. 2000).

Figure 6. A plot of the sedimentation equilibrium data with the residuals from the best fit to a single ideal species. This plot shows the data using protein at 0.4 mg mL^{-1} and a speed of 8000 rpm. The estimated partial specific volume of the protein is 0.741 , and the solvent density was calculated to be 1.011 g mL^{-1} . All nine data sets (three speeds, three concentrations) were fitted together.

Figure 7. Detailed views of the putative active site (stereoview). (A) The conserved acidic residues, Arg, Thr, and His, are colored red, purple, pink, and green, respectively. Most of their side chains face this putative active site. The orientation is the same as in Figures 4B and 5B. (B) The conformations of the conserved amino acids with the docked MMP inhibitor **4a**. The orientation is the same as in (A). The carbon atoms of one monomer, the adjacent monomer, and the MMP inhibitor **4a** are colored green, cyan, and yellow, respectively. Atoms of nitrogen, oxygen, and sulfur are colored blue, red, and orange, respectively.

The next highest scoring TT1542 homologs, the RCK domain of the K^+ channel and UDP-galactose 4-epimerase, belong to a superfamily of NAD(P)-binding Rossmann fold domains, which have one Rossmann fold structure in each monomer (Thoden et al. 1996; Jiang et al. 2001). These two proteins are known to be homodimers, and in UDP-galactose 4-epimerase, each monomer binds one NAD^+ or NADH (Thoden et al. 1996). TT1542 also has two Rossmann fold-like structures in the dimeric form, but the domain orientations are different from those of these homologs.

Between the end of the $\beta 1$ strand and the beginning of the $\alpha 1$ helix, the highly conserved sequence $P^9-H^{10}-P^{11}-D^{12}-D^{13}$ forms a turn in TT1542 (Figs. 1, 2, 3A). Computational analysis suggests that His 10, Asp 12, and Asp 13 in this sequence, and Glu 15 may be catalytic residues (Table 2). Highly conserved residues are also clustered within this region in UDP-GlcNAc 2-epimerase, the GlcNAc transferase MurG, and the NAD(P)-binding Rossmann fold domains (Baker et al. 1992; Campbell et al. 2000; Ha et al. 2000). In the case of the NAD(P)-binding Rossmann fold

domains, a conserved glycine-rich (G-rich) motif (consensus sequence G-X-G-X-X-G/A) forms a tight turn, which hydrogen bonds to the adenine ribose ring, either directly or indirectly (Baker et al. 1992). In addition, in many classical Rossmann domains, an acidic residue at the C-terminal end of the $\beta 2$ strand is commonly used to hydrogen bond to the adenine ribose (Baker et al. 1992). TT1542 has a threonine (Thr 38) at this position, and serine or threonine are found in the TT1542 sequence homologs (Fig. 1). Thr 38 is separated from the putative active site by Asp 74 however (Fig. 7A), and computational analysis (Ota et al. 2003) predicts Asp74 (but not Thr 38) to be catalytic. Arg 51 and Glu 54 in the $\alpha 2$ helix are also highly conserved, and these side chains are close enough to hydrogen bond with each other. Both Arg 51 and Glu 54 are predicted to be catalytically important residues (Ota et al. 2003). All of these predicted catalytic residues in TT1542 are located in the putative active site, and their polar side chains face this cavity. Further studies are necessary to identify the substrate and the enzyme activity of TT1542.

The eukaryotic sequence homologs of TT1542, PIG-L, and GPI12, are GlcNAc-PI de-*N*-acetylases that catalyze the second step of GPI biosynthesis (Stevens 1995; Nakamura et al. 1997; Watanabe et al. 1999; Kinoshita and Inoue 2000; Chang et al. 2002). Many GPI-anchored proteins have important biological activities, and they include membrane-binding enzymes, receptors, and antigens (Ikezawa 2002). Furthermore, the GPI-bound proteins of parasitic protozoa are believed to be the dominant parasite toxins (Delorenzi et al. 2002). The GPI12 protein from the African sleeping sickness parasite *Trypanosoma brucei* is essential to this pathogenic organism, and is considered to be a potential drug target (Chang et al. 2002). GPI-bound proteins may play an important role in other parasitic diseases, such as malaria, the leishmaniasis, and Chagas' disease. The two TT1542 homologs from *M. tuberculosis* involved in mycothiol metabolism are also potential drug targets. Most actinomycetes have a mycothiol-dependent detoxification pathway (Newton and Fahey 2002), and mycothiol-deficient mutants in *Mycobacterium smegmatis* are sensitive to many antibiotics (Newton et al. 1999; Rawat et al. 2002).

The TT1542 structure presented here is the first three-dimensional structure of this superfamily of proteins. This structure will help to model other members of this superfamily, several of which are important drug targets.

Materials and methods

Cloning, protein expression, purification, and crystallization

The gene for *T. thermophilus* HB8 TT1542 was amplified by PCR from *T. thermophilus* HB8 genomic DNA. The PCR product was cloned into the pET11a expression vector (Novagen).

Selenomethionine (SeMet)-substituted protein was expressed in the *E. coli* methionine auxotroph strain B834(DE3). Cells were grown in M9 minimal medium supplemented with SeMet, and protein expression was induced by isopropyl- β -D-thiogalactopyranoside (IPTG). The cells were disrupted by sonication and then were incubated at 70°C for 30 min. The cell lysate was loaded on a Q Sepharose Fast Flow (Amersham Biosciences) column (50 mL) previously equilibrated with 20 mM Tris-HCl buffer (pH 8.0) containing 2 mM DTT. The protein was eluted with a linear gradient of 0–1.0 M NaCl in 20 mM Tris-HCl buffer (pH 8.0) containing 2 mM DTT. Next, the protein sample was loaded on a Resource ISO (Amersham Biosciences) column (1 mL) previously equilibrated with 20 mM Tris-HCl buffer (pH 8.0) containing 1.2 M $(\text{NH}_4)_2\text{SO}_4$ and 2 mM DTT. The protein was eluted with a linear gradient of 1.2–0 M $(\text{NH}_4)_2\text{SO}_4$ in 20 mM Tris-HCl buffer (pH 8.0) with 2 mM DTT. Finally the protein sample was loaded on a Superdex 75 HR column (Amersham Biosciences) previously equilibrated with 20 mM Tris-HCl buffer (pH 8.0) containing 300 mM NaCl and 2 mM DTT, and eluted with this buffer. The yield of purified TT1542 was 0.068 mg per 1 g wet cells.

The crystals were grown at 20°C by the hanging-drop vapor-diffusion method (protein at 3.0 mg mL⁻¹) against a reservoir solution containing 1.6 M $(\text{NH}_4)_2\text{SO}_4$, 100 mM HEPES HCl at pH

7.1, and 100 mM NaCl. Crystals with a size of 0.2 × 0.2 × 0.1 mm grew within a week.

Data collection and processing

Data were collected at RIKEN beamline BL44B2 of SPring-8, Harima, Japan (Adachi et al. 2001) at three wavelengths (Table 1). All data were processed using the HKL2000 and SCALEPACK programs (Otwinowski and Minor 1997). General handling of the scaled data was carried out with programs from the CCP4 suite (CCP4 1994). The positions of the Se atoms and the initial multiwavelength anomalous dispersion (MAD) phases were determined using the program SOLVE (Terwilliger and Berendzen 1999), and the MAD phases were improved with RESOLVE (Terwilliger 2001). The resulting electron density map was extremely clear.

Model building and structural refinement

The data collected at the remote wavelength (0.9752 Å) were used to refine the model. Phase extension to 2.0 Å and automated model

Table 1. Data collection and refinement statistics

Data set	Edge	Peak	Remote
Data collection and processing			
Wavelength (Å)	0.9808	0.9803	0.9752
Resolution range (Å)	20–2.0	20–2.0	20–2.0
Unique reflections (total)	44178 (406551)	44478 (411655)	44495 (419342)
Completeness (%)	99.3 (94.2) ^a	99.2 (95.8)	99.8 (99.3)
R_{sym}^b (%)	3.8 (20.2) ^a	4.0 (22.3)	3.9 (21.2)
I/σ (I)	45.8 (5.6) ^a	46.0 (7.3)	58.3 (6.8)
Phasing statistics			
Resolution range (Å)	20–2.5		
Se sites/monomer	3		
FOM _{MAD} ^c	0.38		
FOM _{Resolve} ^d	0.60		
Model refinement			
Resolution range (Å)	20–2.0		
No. of reflections	42222		
No. of protein atoms	3409		
No. of water molecules	156		
ESU _{positional} ^e	0.142		
ESU _{thermal} ^f	5.19		
$R_{\text{work}}/R_{\text{free}}^g$ (%)	20.1/24.3		
Stereochemistry			
RMSD for bond length (Å)	0.022		
RMSD for bond angles (°)	1.8		
Residues in the Ramachandran plot			
Most favored region (%)	95.5		
Additional allowed regions (%)	4.0		
Generously allowed regions (%)	0.6		

^a Statistics for the highest resolution shell are given in parentheses.

^b $R_{\text{sym}} = (\sum_h \sum_i |I_{hi} - \langle I_h \rangle|) / (\sum_h \sum_i I_{hi})$ where h indicates unique reflection indices, and i indicates symmetry equivalent indices.

^c Figure of merit after SOLVE phasing.

^d Figure of merit after RESOLVE.

^e Estimated standard uncertainty of the positional parameters.

^f Estimated standard uncertainty of the thermal parameters.

^g $R_{\text{work}} = \sum |F_{\text{obs}} - F_{\text{calc}}| / \sum F_{\text{obs}}$ for all reflections and R_{free} was calculated using randomly selected reflections (5%).

Table 2. Prediction of the catalytic residues in TT1542

Residues	CN ^a	C _{local} ^b	C _{space} ^c	S ^d	R ^e	D ^f	L ^g	p _{cat} ^h	Decision ⁱ
His 10	1.00	0.90	0.78	-0.91	3	0.25	2	0.81	✓
Asp 12	1.00	0.90	0.79	1.13	13	2.60	3	0.68	✓
Asp 13	1.00	0.88	0.83	0.86	14	2.17	3	0.68	✓
Glu 15	1.00	0.93	0.78	0.89	17	2.37	0	0.50	✓
Thr 38	1.00	0.73	0.68	-1.23	1	0.00	0	0.09	
Arg 51	1.00	0.77	0.65	-0.25	9	0.73	3	0.61	✓
Glu 54	1.00	0.60	0.61	0.10	13	0.54	2	0.50	✓
Asp 74	1.00	0.65	0.68	-0.20	9	0.95	1	0.78	✓
His 108	1.00	0.57	0.53	-0.74	6	0.40	1	0.10	
Tyr 145	1.00	0.70	0.65	-1.77	1	0.00	0	0.09	

^a Conservation number (CN) defined using the amino-acid grouping scheme by Taylor (Taylor 1986; Zvelebil et al. 1987).

^b Averaged conservation numbers estimated within a local window along the sequence.

^c Averaged conservation numbers estimated over the contact sites in the 3D space.

^d Score, ^eRank, and ^fScore difference, which indicate the stability of the mutant protein.

^g Location number assigned to each category: 3 for hole, 2 for cleft, 1 for surface, and 0 for inside the protein structure according to the preference for the catalytic residues.

^h Probability of the catalytic residues.

ⁱ Catalytic residues are predicted if $p_{cat} \geq 0.5$ (✓).

building were performed with ARP/wARP (Perrakis et al. 2001). The remaining residues were built with the program TURBO-FRODO. A starting model was refined using X-PLOR v3.851 (Brunger 1996) with bulk solvent correction. Conventional positional refinement and individual B-factor refinement were alternated with manual fitting to the electron density using TURBO-FRODO. REFMAC (Murshudov et al. 1999) was used for the final refinement stage. The final model has very good geometry, as examined by PROCHECK (Laskowski et al. 1992): 95.5% of the residues have ϕ/ψ angles in the “most favored region” of the Ramachandran plot and 100% are in the “allowed regions.”

The electron density for the model is very good, as shown in Figure 2. Exceptions include the residues between Thr 178 and Val 186; the side chains of Lys 45, Arg 174, Thr 178, Val 186, and Lys 189 in monomer A; Glu 50, Arg 51, Arg 52, Arg 107, Arg 128, Glu 133, Val 186, and Lys 189 in monomer B. The data collection and refinement statistics are listed in Table 1.

Atomic coordinates have been deposited into the Protein Data Bank with PDB code 1UAN.

Analytical ultracentrifugation

All analytical ultracentrifuge experiments were carried out with a Beckman Optima XL-I analytical ultracentrifuge. The sample buffer was 20 mM Tris-HCl pH 8.0, 300 mM sodium chloride, and 5 mM β -mercaptoethanol, and all experiments were performed out at 20°C. The solvent density and protein partial specific volume ($\bar{v}$) were estimated with SEDNTERP (Laue et al. 1992). Sedimentation velocity data were obtained at 40,000 rpm using an Epon two channel centerpiece, with a loading concentration of 0.27 mg mL⁻¹. The data were analyzed with the program SEDFIT (Schuck 1998), and the polydispersity of the solution was checked by the van Holde-Weischet method. The molecular weight was estimated by fitting the data to the Svedberg equation. The error in the calculated molecular weight was estimated by the Monte Carlo

routine in SEDFIT. Sedimentation equilibrium experiments were carried out with six-channel centerpieces with loading concentrations of 0.9, 0.4, and 0.2 mg mL⁻¹. Data were obtained at 6, 8, and 10 krpm. A total equilibration time of 14 h was used for each speed, with a scan taken at 12 h to ensure equilibrium had been reached. The absorbance wavelength was 280 nm, and the optical baseline was determined by overspeeding at 35 krpm at the end of data collection. For the interference data, an initial scan taken at 3000 rpm was collected after 10 min, and subtracted from the data collected at equilibrium. The equilibrium data were fitted using the manufacturer's software.

Prediction of catalytic residues

In this method (Ota et al. 2003), the conserved sites identified by multiple sequence alignments were classified into functionally important sites and others, based on the assumption that the amino acid residues in functionally important sites could contribute toward destabilizing the protein. In this study, a multiple alignment was constructed using CLUSTAL W (Thompson et al. 1994) for a set of sequences detected by a Blast search against the Swiss-Prot database with E-values $>10^{-5}$. The advantages of this method are that a sophisticated 3D profile (Ota et al. 2001) is used to evaluate the fitness of the amino acid residues in each site, and that it combines other structural features specifically found in the active sites of known proteins; that is, the catalytic residues tend to exist on a partially buried site, to be involved in the β or coil structures, and to reside in holes or clefts in the protein structure.

Ligand search and docking model production

To find binding sites in the PDB similar to that of the hypothetical protein TT1542, we used the search method originally developed by Kinoshita et al. (2002) with some modifications. The details of the modifications will be described elsewhere (Kinoshita and Nakamura 2003). The main improvement is a change in the descriptors of the surface geometry. A molecular surface was generated by Connolly's algorithm (Connolly 1983), and the electrostatic potential at the surface was calculated by solving the Poisson-Boltzmann equation numerically for the precise continuum model with the self-consistent boundary algorithm, to eliminate the effect of the boundary in the finite-difference method with the 1.0 Å grid size (Nakamura and Nishida 1989). Dielectric constants of 2 and 80 were used for the protein and the solvent region, respectively, and an ionic strength of 0.1 M was assumed for every case. The molecular surface was represented by a set of triangular meshes with a normal vector at each vertex on the surface, and the similarity of the surface geometry was assessed by the similarity of the spatial arrangement of the normal vectors. Then the corresponding vertices with similar spatial arrangements and similar electrostatic potentials were searched (Kinoshita et al. 2002). Curvatures at the surface points are now used to describe the surface geometry. This change significantly enhances the performance of the method to find subtle similarities. The similarity search was carried out with all 22,747 binding sites found in the PDB, excluding the low resolution (>2.5 Å) models, NMR entries, and metal binding sites. The query was the entire surface of TT1542, and the database contained the number of subsurfaces around the binding sites. Vertices in each subsurface were selected within 5 Å of the nearest atom in the ligand. The similarity to each binding site was measured by the number of corresponding vertices, and then it was converted to two indices, the Z-score and the coverage. The Z-score is calculated by $(n - \text{mean})/\text{SD}$, where n is the number of

corresponding vertices, and mean and SD are the mean value and standard deviation of the numbers of corresponding vertices obtained by the similarity search against the database, respectively. The coverage is the number of corresponding vertices divided by the number of vertices in each binding site. The Z-score can evaluate the significance of the number of vertices within the database, but it tends to become large when the binding site is large, because large binding sites contain the large numbers of vertices. On the other hand, the coverage will be small for a large ligand. Similarity is more significant with higher scores. The best match found for TT1542 was MMP inhibitor **4a** in PDB entry 1d5j; 308 vertices (Z-score = 3.8, coverage = 0.51) were found to correspond to the vertices in the putative active site of TT1542. The putative complexes were generated by superposing the ligand onto TT1542 by matching corresponding vertices. Figure 7B shows the MMP inhibitor **4a** docked in the putative active site of TT1542. The RMSD of the superimposition was 0.7 Å.

Acknowledgments

This work was supported by the National Project on Protein Structural and Functional Analyses, Ministry of Education, Culture, Sports, Science and Technology. We thank Dr. Nobuo Kamiya, Mr. Taiji Matsu, and Dr. Hisashi Naitow for help in data collection at RIKEN beamline BL44B2 of SPring-8. We thank Dr. David Scott for help in data analysis of analytical ultracentrifuge experiments.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

- Adachi, S., Oguchi, T., Tanida, H., Park, S.-Y., Shimizu, H., Miyatake, H., Kamiya, N., Shiro, Y., Inoue, Y., Ueki, T., et al. 2001. The RIKEN structural biology beamline II (BL44B2) at the SPring-8. *Nucl. Instrum. Methods Phys. Res. A* **467**–**468**: 711–714.
- Altschul, S.F., Madden, T.L., Schaffer, 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**: 3387–3402.
- Baker, P.J., Britton, K.L., Rice, D.W., Rob, A., and Stillman, T.J. 1992. Structural consequences of sequence patterns in the fingerprint region of the nucleotide binding fold. Implications for nucleotide specificity. *J. Mol. Biol.* **228**: 662–671.
- Bateman, A., Birney, E., Cerruti, L., Durbin, R., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30**: 276–280.
- Brunger, A.T. 1996. *X-PLOR, Version 3.85*. Yale University Press, New Haven, CT.
- Campbell, R.E., Mosimann, S.C., Tanner, M.E., and Strynadka, N.C. 2000. The structure of UDP-N-acetylglucosamine 2-epimerase reveals homology to phosphoglycosyl transferases. *Biochemistry* **39**: 14993–15001.
- CCP4 (Collaborative Computational Project Number 4). 1994. The CCP4 suite: Programs for protein crystallography. *Acta. Crystallogr. D* **50**: 760–763.
- Chang, T., Milne, K.G., Guthrie, M.L., Smith, T.K., and Ferguson, M.A. 2002. Cloning of *Trypanosoma brucei* and *Leishmania major* genes encoding the GlcNAc-phosphatidylinositol de-N-acetylase of glycosylphosphatidylinositol biosynthesis that is essential to the African sleeping sickness parasite. *J. Biol. Chem.* **277**: 50176–50182.
- Connolly, M.L. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**: 709–713.
- Delorenzi, M., Sexton, A., Shams-Eldin, H., Schwarz, R.T., Speed, T., and Schofield, L. 2002. Genes for glycosylphosphatidylinositol toxin biosynthesis in *Plasmodium falciparum*. *Infect. Immun.* **70**: 4510–4522.
- Eisenhaber, B., Bork, P., and Eisenhaber, F. 2001. Post-translational GPI lipid anchor modification of proteins in kingdoms of life: Analysis of protein sequence data from complete genomes. *Protein Eng.* **14**: 17–25.
- Fahey, R.C. 2001. Novel thiols of prokaryotes. *Annu. Rev. Microbiol.* **55**: 333–356.
- Ha, S., Walker, D., Shi, Y., and Walker, S. 2000. The 1.9 Å crystal structure of *Escherichia coli* MurG, a membrane-associated glycosyltransferase involved in peptidoglycan biosynthesis. *Protein Sci.* **9**: 1045–1052.
- Holm, L. and Sander, C. 1997. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **25**: 231–234.
- Ikezawa, H. 2002. Glycosylphosphatidylinositol (GPI)-anchored proteins. *Biol. Pharm. Bull.* **25**: 409–417.
- Jiang, Y., Pico, A., Cadene, M., Chait, B.T., and MacKinnon, R. 2001. Structure of the RCK domain from the *E. coli* K⁺ channel and demonstration of its presence in the human BK channel. *Neuron* **29**: 593–601.
- Kinoshita, K., Furui, J., and Nakamura, H. 2002. Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics* **2**: 9–22.
- Kinoshita, K. and Nakamura, H. 2003. Identification of protein biochemical functions by similarity search using the molecular surface database, eF-site. *Protein Sci.* (this issue)
- Kinoshita, T. and Inoue, N. 2000. Dissecting and manipulating the pathway for glycosylphosphatidylinositol-anchor biosynthesis. *Curr. Opin. Chem. Biol.* **4**: 632–638.
- Kraulis, P.J. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**: 946–950.
- Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. 1992. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**: 283–291.
- Laue, T.M., Shah, B., Ridgeway, T.M., and Pelletier, S.L. 1992. SEDNTERP. In *Analytical ultracentrifugation in biochemistry and polymer science* (eds. S.E. Harding et al.), pp. 90–125. Royal Society of Chemistry, UK.
- Merritt, E.A. and Bacon, D.J. 1997. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* **277**: 505–524.
- Murshudov, G.N., Lebedev, A., Vagin, A.A., Wilson, K.S., and Dodson, E.J. 1999. Efficient anisotropic refinement of macromolecular structures using FFT. *Acta Crystallogr. D* **55**: 247–255.
- Nakamura, H. and Nishida, S. 1989. Numerical calculations of electrostatic potentials of protein-solvent systems by the self consistent boundary method. *J. Phys. Soc. Jpn.* **56**: 1609–1622.
- Nakamura, N., Inoue, N., Watanabe, R., Takahashi, M., Takeda, J., Stevens, V.L., and Kinoshita, T. 1997. Expression cloning of PIG-L, a candidate N-acetylglucosaminyl-phosphatidylinositol deacetylase. *J. Biol. Chem.* **272**: 15834–15840.
- Newton, G.L. and Fahey, R.C. 2002. Mycothiol biochemistry. *Arch. Microbiol.* **178**: 388–394.
- Newton, G.L., Arnold, K., Price, M.S., Sherrill, C., Delcardayre, S.B., Aharonowitz, Y., Cohen, G., Davies, J., Fahey, R.C., and Davis, C. 1996. Distribution of thiols in microorganisms: Mycothiol is a major thiol in most actinomycetes. *J. Bacteriol.* **178**: 1990–1995.
- Newton, G.L., Unson, M.D., Anderberg, S.J., Aguilera, J.A., Oh, N.N., delCardayre, S.B., Av-Gay, Y., and Fahey, R.C. 1999. Characterization of *Mycobacterium smegmatis* mutants defective in 1-D-myo-Inositol-2-amino-2-deoxy-α-D-glucopyranoside and mycothiol biosynthesis. *Biochem. Biophys. Res. Commun.* **255**: 239–244.
- Newton, G.L., Av-Gay, Y., and Fahey, R.C. 2000a. N-Acetyl-1-D-myo-inositol-2-amino-2-deoxy-α-D-glucopyranoside deacetylase (MshB) is a key enzyme in mycothiol biosynthesis. *J. Bacteriol.* **182**: 6958–6963.
- . 2000b. A novel mycothiol-dependent detoxification pathway in mycobacteria involving mycothiol S-conjugate amidase. *Biochemistry* **39**: 10739–10746.
- Ota, M., Isogai, Y., and Nishikawa, K. 2001. Knowledge-based potential defined for a rotamer library to design protein sequences. *Protein Eng.* **14**: 557–564.
- Ota, M., Kinoshita, K., and Nishikawa, K. 2003. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.* **327**: 1053–1064.
- Otwinowski, Z. and Minor, W. 1997. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**: 307–326.
- Perrakis, A., Harkiolaki, M., Wilson, K.S., and Lamzin, V.S. 2001. ARP/wARP and molecular replacement. *Acta Crystallogr. D Biol. Crystallogr.* **57**: 1445–1450.
- Rawat, M., Newton, G.L., Ko, M., Martinez, G.J., Fahey, R.C., and Av-Gay, Y. 2002. Mycothiol-deficient *Mycobacterium smegmatis* mutants are hypersensitive to alkylating agents, free radicals, and antibiotics. *Antimicrob. Agents Chemother.* **46**: 3348–3355.

- Rossmann, M.G., Liljas, A., Branden, C.-I., and Banaszak, L.J. 1975. In *The enzymes* (ed. P.D. Boyer), pp. 61–102. Academic Press, New York.
- Schuck, P. 1998. Sedimentation analysis of non-interacting and self-associating solutes using numerical solutions to the Lamm equation. *Biophys. J.* **75**: 1503–1512.
- Stevens, V.L. 1995. Biosynthesis of glycosylphosphatidylinositol membrane anchors. *Biochem. J.* **310**: 361–370.
- Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**: 22–28.
- Taylor, W. 1986. The classification of amino acid conservation. *J. Theor. Biol.* **119**: 205–218.
- Terwilliger, T.C. 2001. Map-likelihood phasing. *Acta Crystallogr. D* **57**: 1763–1775.
- Terwilliger, T.C. and Berendzen, J. 1999. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**: 849–861.
- Thoden, J.B., Frey, P.A., and Holden, H.M. 1996. Molecular structure of the NADH/UDP-glucose abortive complex of UDP-galactose 4-epimerase from *Escherichia coli*: Implications for the catalytic mechanism. *Biochemistry* **35**: 5137–5144.
- Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**: 4673–4680.
- Watanabe, R., Ohishi, K., Maeda, Y., Nakamura, N., and Kinoshita, T. 1999. Mammalian PIG-L and its yeast homolog Gpi12p are *N*-acetylglucosaminylphosphatidylinositol de-*N*-acetylases essential in glycosylphosphatidylinositol biosynthesis. *Biochem. J.* **339**: 185–192.
- Wrabl, J.O. and Grishin, N.V. 2001. Homology between O-linked GlcNAc transferases and proteins of the glycogen phosphorylase superfamily. *J. Mol. Biol.* **314**: 365–374.
- Zvelebil, M.J., Barton, G.J., Taylor, W.R., and Sternberg, M.J. 1987. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**: 957–961.