

# The structure of the first representative of Pfam family PF06475 reveals a new fold with possible involvement in glycolipid metabolism

Constantina Bakolitsa,<sup>a,b</sup> Abhinav Kumar,<sup>a,c</sup> Daniel McMullan,<sup>a,d</sup> S. Sri Krishna,<sup>a,b,e</sup> Mitchell D. Miller,<sup>a,c</sup> Dennis Carlton,<sup>a,f</sup> Rafael Najmanovich,<sup>g</sup> Polat Abdubek,<sup>a,d</sup> Tamara Astakhova,<sup>a,e</sup> Hsiu-Ju Chiu,<sup>a,c</sup> Thomas Clayton,<sup>a,f</sup> Marc C. Deller,<sup>a,f</sup> Lian Duan,<sup>a,e</sup> Ylva Elias,<sup>a,f</sup> Julie Feuerhelm,<sup>a,d</sup> Joanna C. Grant,<sup>a,d</sup> Slawomir K. Grzechnik,<sup>a,e</sup> Gye Won Han,<sup>a,f</sup> Lukasz Jaroszewski,<sup>a,b,e</sup> Kevin K. Jin,<sup>a,c</sup> Heath E. Klock,<sup>a,d</sup> Mark W. Knuth,<sup>a,d</sup> Piotr Kozbial,<sup>a,b</sup> David Marciano,<sup>a,f</sup> Andrew T. Morse,<sup>a,e</sup> Edward Nigoghossian,<sup>a,d</sup> Linda Okach,<sup>a,d</sup> Silvy Oommachen,<sup>a,c</sup> Jessica Paulsen,<sup>a,d</sup> Ron Reyes,<sup>a,c</sup> Christopher L. Rife,<sup>a,c</sup> Christina V. Trout,<sup>a,f</sup> Henry van den Bedem,<sup>a,c</sup> Dana Weekes,<sup>a,b</sup> Aprilfawn White,<sup>a,d</sup> Qingping Xu,<sup>a,c</sup> Keith O. Hodgson,<sup>a,h</sup> John Wooley,<sup>a,c</sup> Marc-André Elslinger,<sup>a,f</sup> Ashley M. Deacon,<sup>a,c</sup> Adam Godzik,<sup>a,b,e</sup> Scott A. Lesley<sup>a,d,f</sup> and Ian A. Wilson<sup>a,f,\*</sup>

<sup>a</sup>Joint Center for Structural Genomics, <http://www.jcsg.org>, USA, <sup>b</sup>Program on Bioinformatics and Systems Biology, Burnham Institute for Medical Research, La Jolla, CA, USA, <sup>c</sup>Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA, USA, <sup>d</sup>Protein Sciences Department, Genomics Institute of the Novartis Research Foundation, San Diego, CA, USA, <sup>e</sup>Center for Research in Biological Systems, University of California, San Diego, La Jolla, CA, USA, <sup>f</sup>Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA, USA, <sup>g</sup>Département de Biochimie, Université de Sherbrooke, Québec, Canada, and <sup>h</sup>Photon Science, SLAC National Accelerator Laboratory, Menlo Park, CA, USA

Correspondence e-mail: [wilson@scripps.edu](mailto:wilson@scripps.edu)

Received 16 April 2009

Accepted 12 June 2009

**PDB Reference:** PA1994 from *P. aeruginosa*, 2h1t, r2h1tsf.

The crystal structure of PA1994 from *Pseudomonas aeruginosa*, a member of the Pfam PF06475 family classified as a domain of unknown function (DUF1089), reveals a novel fold comprising a 15-stranded  $\beta$ -sheet wrapped around a single  $\alpha$ -helix that assembles into a tight dimeric arrangement. The remote structural similarity to lipoprotein localization factors, in addition to the presence of an acidic pocket that is conserved in DUF1089 homologs, phospholipid-binding and sugar-binding proteins, indicate a role for PA1994 and the DUF1089 family in glycolipid metabolism. Genome-context analysis lends further support to the involvement of this family of proteins in glycolipid metabolism and indicates possible activation of DUF1089 homologs under conditions of bacterial cell-wall stress or host–pathogen interactions.

## 1. Introduction

In an effort to extend the structural coverage of proteins for which the biological function is unknown and cannot be deduced by homology (*i.e.* domains of unknown function; DUFs), targets were selected from Pfam protein family PF06745 (DUF1089). DUF1089 homologs are present in pathogenic actinobacteria, burkholderia, firmicutes and lactobacilli. Here, we report the crystal structure of PA1994, the first structural representative of this family, which was determined using the semi-automated high-throughput pipeline of the Joint Center for Structural Genomics (JCSG; <http://www.jcsg.org>; Lesley *et al.*, 2002) as part of the NIGMS Protein Structure Initiative (PSI; <http://www.nigms.nih.gov/Initiatives/PSI/>). The *PA1994* gene of *Pseudomonas aeruginosa*, an opportunistic human pathogen (Gomez & Prince, 2007), encodes a protein with a molecular weight of 21.6 kDa (residues 1–187) and a calculated isoelectric point of 4.9.

We show that global and local structural and chemical similarities to lipid-binding proteins suggest the involvement of PA1994 with the bacterial membrane, while genome-context analysis supports a role for the DUF1089 family in glycolipid metabolism that is likely to be triggered under conditions of osmotic stress or host–pathogen interactions. These structural insights should help to guide future functional studies.

## 2. Materials and methods

### 2.1. Protein production and crystallization

Clones were generated using the Polymerase Incomplete Primer Extension (PIPE) cloning method (Klock *et al.*, 2008). The gene encoding PA1994 (GenBank NP\_250684; gi:15597190; Swiss-Prot Q912B5) was amplified by polymerase chain reaction (PCR) from *P. aeruginosa* PA01-LAC genomic DNA using *PfuTurbo* DNA polymerase (Stratagene) and I-PIPE (Insert) primers (forward primer, 5'-ctgtactccaggcATGAGTCGCGACCGTCTGTACACCTGGG-3'; reverse primer, 5'-aattaagtcgcttaGAGACGCTGGAAGAGACCCGGGTAATCG-3'; target sequence in upper case) that included sequences for the predicted 5' and 3' ends. The expression vector pSpeedET, which encodes an amino-terminal tobacco etch virus (TEV) protease-cleavable expression and purification tag (MGSDKIHHHHHHENLYFQ/G), was PCR-amplified with V-PIPE

(Vector) primers (forward primer, 5'-taacgcgacttaattaactcgtttaacgg-tctccagc-3'; reverse primer, 5'-gccctggaagtacaggtttctgatgatgatgatg-atg-3'). V-PIPE and I-PIPE PCR products were mixed to anneal the amplified DNA fragments. *Escherichia coli* GeneHogs (Invitrogen) competent cells were transformed with the V-PIPE/I-PIPE mixture and dispensed onto selective LB-agar plates. The cloning junctions were confirmed by DNA sequencing. Expression was performed in selenomethionine-containing medium with suppression of normal methionine synthesis. At the end of fermentation, lysozyme was added to the culture to a final concentration of 250  $\mu\text{g ml}^{-1}$  and the cells were harvested and frozen. After one freeze-thaw cycle, the cells were sonicated in lysis buffer [50 mM HEPES pH 8.0, 50 mM NaCl, 10 mM imidazole, 1 mM tris(2-carboxyethyl)phosphine-HCl (TCEP)] and the lysate was clarified by centrifugation at 32 500g for 30 min. The soluble fraction was passed over nickel-chelating resin (GE Healthcare) pre-equilibrated with lysis buffer, the resin was washed with wash buffer [50 mM HEPES pH 8.0, 300 mM NaCl, 40 mM imidazole, 10% (v/v) glycerol, 1 mM TCEP] and the protein was eluted with elution buffer [20 mM HEPES pH 8.0, 300 mM imidazole, 10% (v/v) glycerol, 1 mM TCEP]. The eluate was buffer-exchanged with HEPES crystallization buffer (20 mM HEPES pH 8.0, 200 mM NaCl, 40 mM imidazole, 1 mM TCEP) using a PD-10 column (GE Healthcare) and incubated with 1 mg TEV protease per 15 mg eluted protein. The protease-treated eluate was passed over nickel-chelating resin (GE Healthcare) pre-equilibrated with HEPES crystallization buffer and the resin was washed with the same buffer. The flowthrough and wash fractions were combined and concentrated to 11.2 mg ml<sup>-1</sup> by centrifugal ultrafiltration (Millipore) for crystallization trials. PA1994 was crystallized using the nanodroplet vapor-diffusion method (Santarsiero *et al.*, 2002) with standard JCSG crystallization protocols (Lesley *et al.*, 2002). Sitting drops composed of 200 nl protein solution mixed with 200 nl crystallization solution were equilibrated against a 50  $\mu\text{l}$  reservoir at 277 K for 40 d prior to harvesting. Initial screening for diffraction was carried out using the Stanford Automated Mounting system (SAM; <http://smb.slac.stanford.edu/facilities/hardware/SAM/UserInfo>; Cohen *et al.*, 2002) at the Stanford Synchrotron Radiation Lightsource (SSRL; Menlo Park, California, USA). The crystallization reagent that produced the PA1994 crystal used for the structure solution contained 5% (v/v) 2-methyl-2,4-pentanediol (MPD; racemic mixture), 10% (w/v) PEG 6000 and 0.1 M HEPES pH 7.5. Ethylene glycol was added to the crystal as a cryoprotectant to a final concentration of 15% (v/v). A rod-shaped crystal with approximate dimensions of 200  $\times$  20  $\times$  20  $\mu\text{m}$  was mounted in a nylon loop. The diffraction data were indexed in the monoclinic space group C2 (Table 1). The molecular weight and oligomeric state of PA1994 were determined using a 0.8  $\times$  30 cm Shodex Protein KW-803 column (Thomson Instruments) pre-calibrated with gel-filtration standards (Bio-Rad).

## 2.2. Data collection, structure solution and refinement

Multiple-wavelength anomalous diffraction (MAD) data were collected at SSRL on beamline BL11-1 at wavelengths corresponding to the inflection ( $\lambda_1$ ), peak ( $\lambda_2$ ) and high-energy remote ( $\lambda_3$ ) of a selenium MAD experiment. The data sets were collected at 100 K with an ADSC Q315 CCD detector using the *Blu-Ice* data-collection environment (McPhillips *et al.*, 2002). The MAD data were integrated and reduced using *XDS* and then scaled with the program *XSCALE* (Kabsch, 1993). Phasing was performed with *SHELX* (Sheldrick, 2008) and *AutoSHARP* (Bricogne *et al.*, 2003), which resulted in a mean figure of merit of 0.15 with four selenium positions. Two were high occupancy, corresponding to the main selenium positions at

**Table 1**

Summary of crystal parameters, data collection and refinement statistics for PA1994 (PDB code 2h1t).

Values in parentheses are for the highest resolution shell.

	$\lambda_1$ MADSe	$\lambda_2$ MADSe	$\lambda_3$ MADSe
Space group	C2		
Unit-cell parameters ( $\text{\AA}$ , $^\circ$ )	$a = 130.03$ , $b = 41.90$ , $c = 78.65$ , $\beta = 91.2$		
Data collection			
Wavelength ( $\text{\AA}$ )	0.9793	0.9789	0.9116
Resolution range ( $\text{\AA}$ )	28.3–1.80	28.3–1.91	28.3–1.80
	(1.85–1.80)	(1.96–1.91)	(1.85–1.80)
No. of observations	136388	121791	146173
No. of unique reflections	38719	33103	39473
Completeness (%)	98.0 (83.9)	99.6 (97.2)	99.7 (98.3)
Mean $I/\sigma(I)$	9.9 (1.9)	10.6 (3.4)	10.3 (2.6)
$R_{\text{merge}}$ on $I^\dagger$ (%)	9.9 (51.4)	10.5 (35.1)	9.9 (51.7)
Model and refinement statistics			
Resolution range ( $\text{\AA}$ )	28.3–1.80		
No. of reflections (total)	35699 $\ddagger$		
No. of reflections (test)	1772		
Completeness (%)	90.2		
Data set used in refinement	$\lambda_1$ MADSe		
Cutoff criterion	$ F  > 0$		
$R_{\text{cryst}}^\S$	0.170		
$R_{\text{free}}^\P$	0.213		
Stereochemical parameters			
Restraints (r.m.s.d. observed)			
Bond angles ( $^\circ$ )	1.58		
Bond lengths ( $\text{\AA}$ )	0.015		
Average isotropic $B$ value ( $\text{\AA}^2$ )	20.5 $\dagger\dagger$		
ESU $\ddagger\ddagger$ based on $R_{\text{free}}$ ( $\text{\AA}$ )	0.13		
Protein residues/atoms	370/3051		
Waters/other solvent molecules	367/11		

$\dagger R_{\text{merge}} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$ .  $\ddagger$  Owing to ice rings, a total of 3016 reflections were omitted from the resolution ranges 1.91–1.93, 2.02–2.04 and 2.23–2.27  $\text{\AA}$ . Typically, a few reflections were also excluded owing to negative intensities and rounding errors in the resolution limits and unit-cell parameters.  $\S R_{\text{cryst}} = \sum_{hkl} ||F_{\text{obs}}| - |F_{\text{calc}}|| / \sum_{hkl} |F_{\text{obs}}|$ , where  $F_{\text{calc}}$  and  $F_{\text{obs}}$  are the calculated and observed structure-factor amplitudes, respectively.  $\P R_{\text{free}}$  is the same as  $R_{\text{cryst}}$  but for 5.0% of the total reflections chosen at random and omitted from refinement.  $\dagger\dagger$  This represents the total  $B$  including both the TLS and residual  $B$  components.  $\ddagger\ddagger$  Estimated overall coordinate error (Collaborative Computational Project, Number 4, 1994; Cruickshank, 1999).

residues A143 and B143, whereas the others were low occupancy (20% relative to the primary site), corresponding to an alternate conformation of residue 143 in each monomer (<4.7  $\text{\AA}$  from the primary site). It should be noted that the presence of only one ordered SeMet site (two conformations) per 188 residues in the protein chain sufficed for successful phasing and model building. Automated model building was performed with *ARP/wARP* (Cohen *et al.*, 2004) and model completion and refinement were performed with *Coot* (Emsley & Cowtan, 2004) and *REFMAC 5.2* (Winn *et al.*, 2003). Refinement included phase restraints from *AutoSHARP* and TLS refinement with two TLS groups per chain as suggested by the *TLSMD* server (Painter & Merritt, 2006). Data reduction and refinement statistics are summarized in Table 1.

## 2.3. Validation and deposition

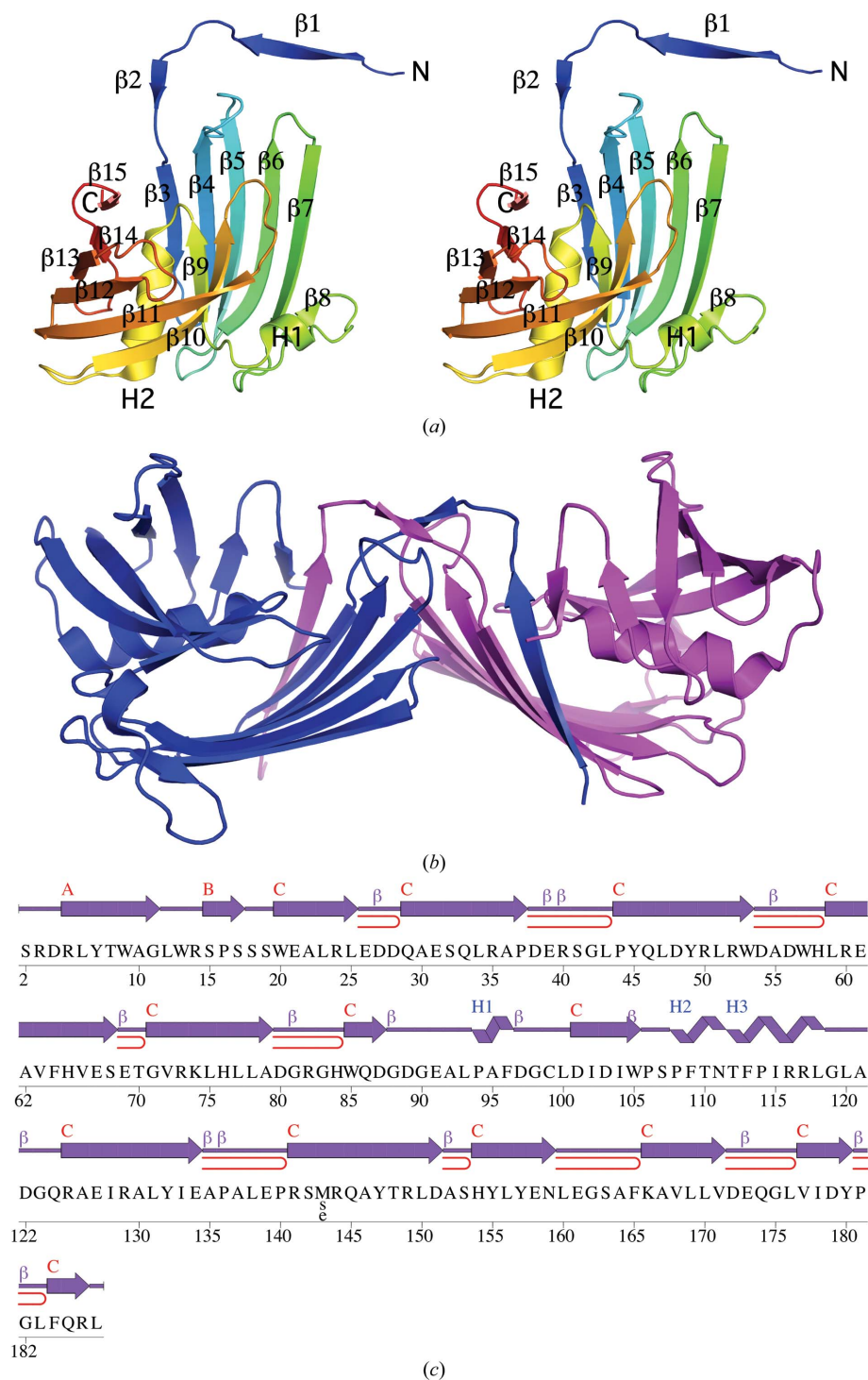
Analysis of the stereochemical quality of the model was accomplished using *AutoDepInputTool* (Yang *et al.*, 2004), *MolProbity* (Davis *et al.*, 2004), *SFCHECK* 4.0 (Collaborative Computational Project, Number 4, 1994) and *WHATIF* 5.0 (Vriend, 1990). Protein quaternary-structure analysis was performed using the *PISA* server (Krissinel & Henrick, 2007). Fig. 1(c) was adapted from an analysis using *PDBsum* (Laskowski *et al.*, 2005) and all other figures were prepared with *PyMOL* (DeLano Scientific). Atomic coordinates and experimental structure factors for PA1994 at 1.80  $\text{\AA}$  resolution have been deposited in the PDB under accession code 2h1t.

### 3. Results and discussion

#### 3.1. Overall structure

The crystal structure of PA1994 (Fig. 1*a*) was determined to 1.80 Å resolution using the multiple-wavelength anomalous dispersion

(MAD) method. Refinement statistics are summarized in Table 1. The final model includes 370 residues (residues 2–187 of chain *A* and residues 4–187 of chain *B*), nine ethylene glycol molecules, two MPD molecules and 367 water molecules in the asymmetric unit. No electron density was observed for the N-terminal glycine (0)



**Figure 1**

Crystal structure of PA1994 from *P. aeruginosa*. (a) Stereo ribbon diagram of the PA1994 monomer color coded from the N-terminus (blue) to the C-terminus (red). Helices (H1–H2) and  $\beta$ -strands ( $\beta$ 1– $\beta$ 15) are indicated. (b) Ribbon representation of the PA1994 dimer showing domain swapping of the N-terminal  $\beta$ -strands. Monomers are depicted in blue and magenta. (c) Diagram showing the secondary-structure elements of PA1994 superimposed on its primary sequence. The labeling of secondary-structure elements is in accord with *PDBsum* (<http://www.ebi.ac.uk/pdbsum>), where  $\alpha$ -helices are sequentially labeled (H1, H2, H3 etc.),  $\beta$ -strands are labeled (A, B, C etc.) according to the  $\beta$ -sheets to which they are assigned,  $\beta$ -turns and  $\gamma$ -turns are designated by Greek letters ( $\beta$ ,  $\gamma$ ) and  $\beta$ -hairpins are designated by red loops. For PA1994, the  $\alpha$ -helix (H2),  $3_{10}$ -helix (H1),  $\beta$ -strands in  $\beta$ -sheets (A–C),  $\beta$ -turns ( $\beta$ ) and  $\beta$ -hairpins are indicated.

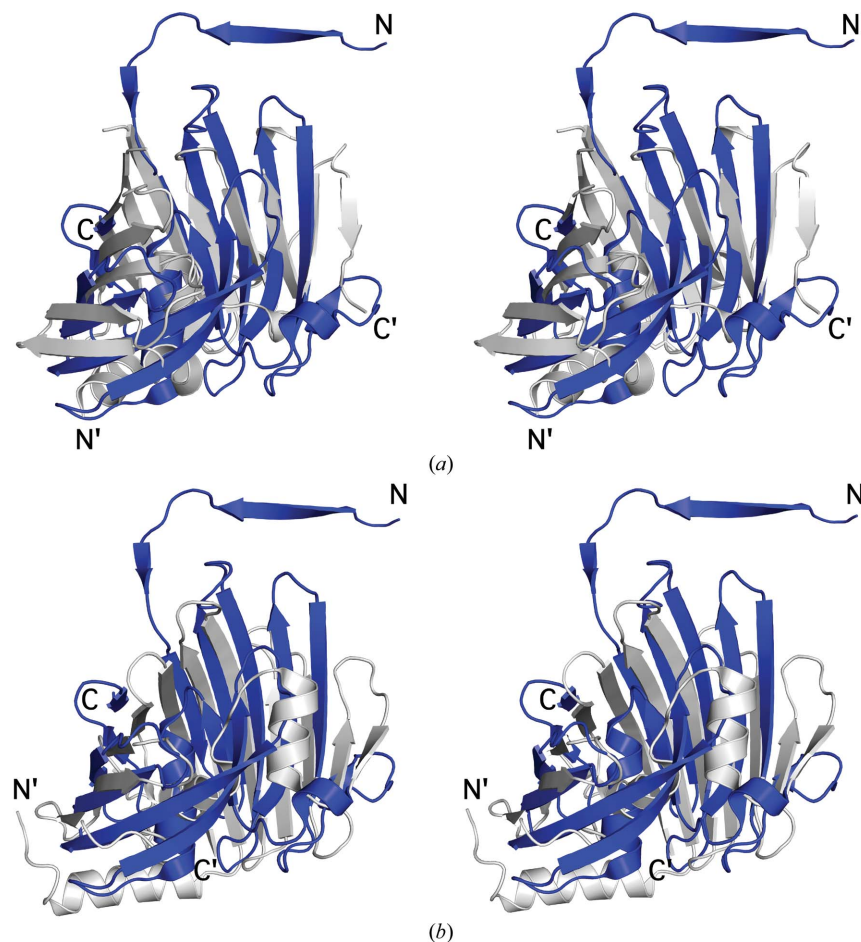
remaining after cleavage of the expression and purification tag, for the terminal selenomethionine (residue 1) of chains *A* and *B* or for Ser2 and Arg3 in chain *B*. The side chains of Arg5 and Glu91 in chain *B* were omitted owing to weak electron density. The Matthews coefficient ( $V_M$ ; Matthews, 1968) was  $2.5 \text{ \AA}^3 \text{ Da}^{-1}$  and the estimated solvent content was 50.1%. A Ramachandran plot produced by *MolProbity* (Davis *et al.*, 2004) showed that 99.2% of the residues are in favored regions. The two outliers, Pro106 in chains *A* and *B*, are actually found in a *cis* conformation in both chains and have clear electron density.

*SCOP* (release 1.75) classifies PA1994 as a single-domain protein with a novel fold termed a spiral  $\beta$ -roll (<http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.c.bdb.b.b.b.html>), with a 15-stranded  $\beta$ -sheet wrapped around a central helix (Fig. 1). The N-terminal half of the sheet is formed by strands  $\beta_3$ – $\beta_7$  supplemented by a  $\beta_1$ -strand exchange from the other monomer in the asymmetric unit (Fig. 1*b*) that hydrogen bonds extensively to the  $\beta_3$  and the shorter  $\beta_{15}$  strands (Figs. 1*a* and 1*b*). This swapping additionally involves strand  $\beta_2$  and results in a large buried dimerization interface of  $\sim 3000 \text{ \AA}^2$  per monomer. A short  $\beta$ -strand ( $\beta_8$ ) and  $3_{10}$ -helix H1 separate the first half of the  $\beta$ -sheet from the more tightly curved C-terminal region (strands  $\beta_{10}$ – $\beta_{15}$ ). Helix H2 and strand  $\beta_9$  are sandwiched between the two halves of the  $\beta$ -sheet in the center of the molecule.

PA1994 can be viewed as consisting of two subdomains: the first half of the  $\beta$ -sheet ( $\beta_{1'}$ ,  $\beta_3$ – $\beta_8$ ) and helix H1 (residues 1–98) compose the first domain, which packs against the other subdomain consisting

of the second half of the  $\beta$ -sheet ( $\beta_9$ – $\beta_{15}$ ) and helix H2 (residues 99–187). Both subdomains are present in DUF1089-family members and a sequence analysis of the family indicates a high degree of conservation in the residues that are implicated in stabilizing both regions of the molecule. Stacking interactions, both intermolecular (Trp9–Pro108') and intramolecular (Trp57–Phe113), show strict or high conservation. Additionally, conserved stacking interactions are observed in residue pairs involving the H2 helix and both the N-terminal (Trp57–Phe113) and the C-terminal (Pro114–Tyr147) halves of the  $\beta$ -sheet, as well as the conserved binding-pocket residues (Trp9–Pro108', Trp57–Phe113 and Pro106/Pro108–Phe184; see below).

A search with *FATCAT* (Ye & Godzik, 2004) identified that the highest structural similarity is with outer membrane proteins (SH3-like barrel fold), NTF2-like proteins (cystatin-like fold) and fatty acid-binding proteins (lipocalin fold). *DALI* (Holm & Sander, 1995) showed significant hits with a number of different folds, including  $\beta$ -galactosidase (immunoglobulin-like  $\beta$ -sandwich fold), iron-transport proteins (transmembrane  $\beta$ -barrel fold), lipovitellin (lipovitellin–phosvitin complex/ $\beta$ -sheet shell regions fold), tail-associated lysozyme (phage-tail protein fold) and lipoprotein localization factors (prokaryotic lipoprotein localization factor fold). A search using secondary-structure matching (*SSM*; Krissinel & Henrick, 2004) identified the lipoprotein localization factor LolA (PDB code 1iwl) as the top hit ( $Z$  score 2.5,  $P$  score 0.0), although the  $P$  score indicates a statistically insignificant match.

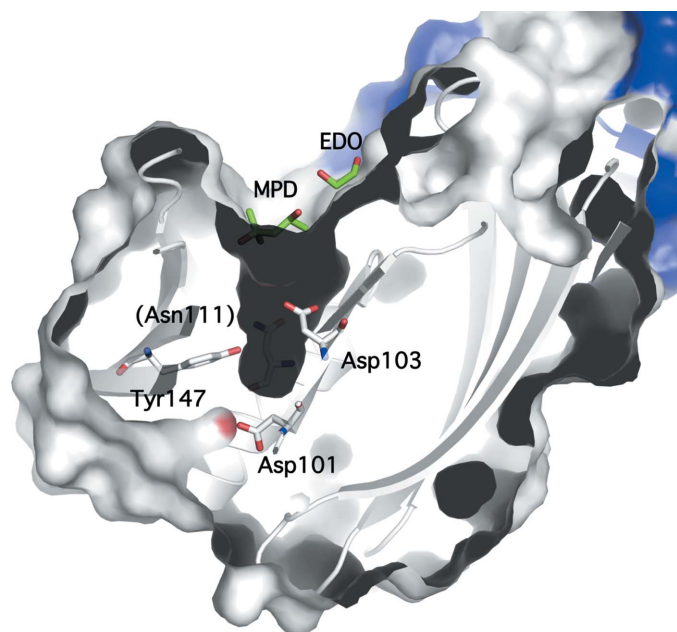


**Figure 2** PA1994 exhibits structural similarity to the lipoprotein chaperones LolA and LolB. (a) Stereoview of the structural superposition of PA1994 (PDB code 2h1t, residues 2–187, blue) and LolA (PDB code 1iwl, residues 1–182, gray). (b) Stereoview of the structural superposition of PA1994 (PDB code 2h1t, residues 2–187, blue) and LolB (PDB code 1iwn, residues 10–186, gray).

Although PA1994 appears to constitute a new fold, we decided to investigate subfold similarities in an attempt to identify shared structural features that could provide insight into the origin and function of PA1994. The highest structural similarity identified by visual inspection was with lipoprotein localization factors A and B (LolA and LolB) from *E. coli*, which are highly conserved bacterial proteins that are implicated in lipoprotein sorting and membrane localization (Takeda *et al.*, 2003). Superimposition of PA1994 onto LolA, with an r.m.s.d. of 3.1 Å, reveals that these proteins share the same fold and topology over the 11  $\beta$ -strands and the central helix, although the sequence identity over 104 aligned residues is not significant at only 5% (Fig. 2*a*). Differences within the barrel include PA1994 strands  $\beta$ 9– $\beta$ 10, which are absent in both lipoprotein localization factors, strand  $\beta$ 8 (absent in LolA) and the orientation of the central helix in LolB (Figs. 2*a* and 2*b*). Outside the barrel, the main differences involve an additional N-terminal helix in LolA located at the bottom of the  $\beta$ -barrel and the LolA C-terminal  $3_{10}$ -helix and  $\beta$ -strand (Figs. 2*a* and 2*b*). Both of these C-terminal structural elements, which are absent in PA1994, are involved in the specific membrane localization of lipoproteins by LolA (Okuda *et al.*, 2008). No strand-swapping is observed in either LolA or LolB, although the N-terminal  $\beta$ -strand is present in both cases and overlaps with the swapped strand from the PA1994 dimer.

### 3.2. Analysis of a conserved cavity

An analysis of PA1994 using the *CastP* server (Binkowski *et al.*, 2003) revealed a deep pocket (15  $\times$  6  $\times$  7 Å) enclosed mainly by helix H2 and strand  $\beta$ 7, with additional contributions made by strands  $\beta$ 10– $\beta$ 12 and the loop between strands  $\beta$ 14 and  $\beta$ 15. This pocket is lined with conserved hydrophilic residues (Ser107, Thr110, Asn111, Thr112 and Gln145) and contains the hydroxyl group of the invariant Tyr147 in addition to an acidic pocket formed by two invariant aspartates



**Figure 3**

An acidic pocket conserved in the DUF1089 family suggests a ligand-binding site. The PA1994 monomers, colored white and blue, are shown as a ribbon diagram and as a surface representation. Invariant residues (Asp101, Asp103 and Tyr147) are indicated, with the conserved Asn111 located behind the pocket labeled in parentheses. The ethylene glycol (EDO) and MPD molecules that line the entrance to the acidic pocket in the crystal are shown in green.

(Asp101 and Asp103; Fig. 3). The pocket is in a similar location to the cavity in LolA that has been shown to bind lipids (Watanabe *et al.*, 2006). However, the binding pocket is hydrophobic in LolA, whereas the PA1994 pocket is acidic, suggesting a more hydrophilic ligand. The entrance to the pocket in PA1994 forms a long and narrow groove (20  $\times$  7 Å) composed of strictly or highly conserved hydrophobic residues (Ile102, Pro106, Pro108, Phe165, Leu170 and Ile178) and also involves the dimerization interface (Trp13), suggesting a hydrophobic component of the ligand and the likely requirement of dimerization for binding. Analytical size-exclusion chromatography in combination with static light scattering indicates that PA1994 is a dimer in solution. Two crystallization-reagent molecules (ethylene glycol and MPD) line both the groove and the pocket, indicating that both regions could be implicated in ligand binding (Fig. 3*b*). Both LolA and PA1994 contain a *cis*-proline (Pro89 in LolA and Pro106 in PA1994) at the N-terminal end of the central helix. Because of the relatively low energy barrier between *trans* and *cis* conformations, *cis*-prolines are often involved in function and have been implicated in both protein stabilization (Truckses *et al.*, 1996) and catalysis (Charbonnier *et al.*, 1999), suggesting that this residue might serve a similar purpose in LolA and PA1994.

A search against a database of nonredundant cognate binding sites using *IsoCleft* (Najmanovich *et al.*, 2008), a graph-matching algorithm that searches for both geometrical and chemical composition similarities, identified shared features between the PA1994 pocket and the binding sites of proteins implicated in bacterial cell-wall biosynthesis, with alanine racemase from *P. aeruginosa* (PDB code 1rcq; 21 atoms in common, Tanimoto similarity score 0.39, Z score 4.26, *P* value  $7.54 \times 10^{-3}$ ; LeMagueres *et al.*, 2003) and hyaluronate lyase from *Streptococcus pneumoniae* (PDB code 1loh; 21 atoms in common, Tanimoto similarity score 0.38, Z score 4.01, *P* value  $1.03 \times 10^{-2}$ ; Jedrzejewski *et al.*, 2002) as the top hits. Additional similarities include the binding of sugars, with galactose mutarotase (PDB code 1so0; 25 atoms in common, Tanimoto similarity score 0.38, Z score 4.08, *P* value  $9.44 \times 10^{-3}$ ; Thoden *et al.*, 2004) and *meso*-2,3-butanediol dehydrogenase (PDB code 1geg; 20 atoms in common, Tanimoto similarity score 0.36, Z score 3.82, *P* value  $1.32 \times 10^{-2}$ ; Otagiri *et al.*, 2001) as the closest matches, in addition to an inorganic pyrophosphatase (PDB code 1wpm; 25 atoms in common, Tanimoto similarity score 0.37, Z score 3.89, *P* value  $1.21 \times 10^{-2}$ ; Fabrichny *et al.*, 2004). *IsoCleft* also identified similarities between the hydrophobic groove along the PA1994 pocket entrance and dimerization interface and the lipid-binding site in *Candida rugosa* lipase (PDB code 1lpn; 31 atoms in common, Tanimoto similarity score 0.20, Z score 3.98, *P* value  $1.08 \times 10^{-2}$ ; Grochulski, Bouthillier *et al.*, 1994).

Taken together, these structural and chemical similarities support a role for PA1994 and the DUF1089 family in glycolipid binding. The extensive dimerization interface observed in the structure, in addition to the SEC/SLS data, suggest that a dimer is likely to be the biologically relevant oligomeric state of PA1994. The swapped  $\beta$ -strands appear to participate in stabilizing the conserved cavity. Substrate binding might induce large-scale conformational changes, as is the case for the lipid-binding proteins that share structural similarities with PA1994 (Marland *et al.*, 2006; Oguchi *et al.*, 2008; Grochulski, Li *et al.*, 1994).

### 3.3. Genome-context analysis

Glycophospholipids, which are implicated in the synthesis of complex cell-wall structures that enable some pathogens to modulate the response by the host immune system, have been suggested to bind to similar-sized acidic pockets as that observed in PA1994 (Marland *et*

*al.*, 2006). Glycolipids serve as key immunomodulatory molecules in host–pathogen interactions (Nigou *et al.*, 2008) and lipases have been known to act as virulence factors (Smoot, 1997). In addition to their role in pathogenicity, bacterial cell-wall glycolipids are modified in response to variations in temperature, pH and other environmental stressors (Mykytczuk *et al.*, 2007), with changes affecting both the lipid and sugar composition of the membrane (Bengoechea *et al.*, 2002; Tymczyszyn *et al.*, 2005).

The genome context (<http://string.embl.de>) of DUF1089-family members additionally supports a role in glycolipid biosynthesis which is likely to be induced under conditions of cell-wall stress or host–pathogen interactions. PA1994 is predicted with a high degree of confidence to be in functional association with a peptidyl prolyl *cis*–*trans* isomerase (PA1996), an enzyme that functions as a chaperone and is up-regulated under conditions of cell-wall stress (Muthaiyan *et al.*, 2008). The prolyl *cis*–*trans* isomerase could also assist in the folding of PA1994, as Pro106 appears to be involved in stabilization of both the hydrophobic core and the acidic pocket. Similarly, R02764, a DUF1089 homologue from *Sinorhizobium meliloti*, is predicted to be functionally linked to a glyceraldehyde 3-phosphate dehydrogenase [R02763, normally a cytosolic enzyme involved in energy metabolism that shows pH-dependent association with bacterial cell walls (Antikainen *et al.*, 2007), where it becomes involved in host–pathogen interactions (Schaumburg *et al.*, 2004)], a transketolase (R02762, an enzyme implicated in lipopolysaccharide metabolism; Eidels & Osborn, 1971) and a taurine-uptake ABC transporter (RB0965; taurine is a constituent of the bacterial cell wall that has been implicated in membrane stabilization and recovery from osmotic shock; Yancey, 2005). MT3862, a DUF1089 homologue from *Mycobacterium tuberculosis*, is also predicted with high confidence to be in functional association with two osmoprotectant proteins (MT3863 and MT3864) implicated in glycine betaine-dependent transport. In addition to its role in maintaining membrane fluidity, glycine betaine acts as a chemical chaperone (Diamant *et al.*, 2001), stabilizing proteins under conditions of environmental stress.

Availability of more DUF1089-member sequences and structures might shed light on the evolutionary history of this intriguing protein family. The information presented here, in combination with further biochemical and biophysical studies, should yield valuable insights into the functional role of PA1994. Models of PA1994 homologs can be accessed at [http://www1.jcsg.org/cgi-bin/models/get\\_mor.pl?key=2hltA](http://www1.jcsg.org/cgi-bin/models/get_mor.pl?key=2hltA).

Additional information about PA1994 is available from TOPSAN (Krishna *et al.*, 2010) <http://www.topsan.org/explore?PDBid=2h1t>.

## 4. Conclusions

The first structural representative of the DUF1089 family reveals a novel fold. Remote global and local similarities to lipid-binding and glycan-binding proteins along with genome-context analysis support a role for PA1994 in glycolipid metabolism that is likely to be induced under conditions of cell-wall stress or host–pathogen interactions.

This work was supported by the National Institute of General Medical Sciences Protein Structure Initiative grant Nos. P50 GM62411 and U54 GM074898. Portions of this research were carried out at the Stanford Synchrotron Radiation Lightsource (SSRL). The SSRL is a national user facility operated by Stanford University on behalf of the US Department of Energy, Office of Basic Energy Sciences. The SSRL Structural Molecular Biology Program is supported by the Department of Energy, Office of Biological and

Environmental Research and by the National Institutes of Health (National Center for Research Resources, Biomedical Technology Program and the National Institute of General Medical Sciences). Genomic DNA from *P. aeruginosa* PA01-LAC (ATCC No. 47085D) was obtained from the American Type Culture Collection (ATCC). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

## References

- Antikainen, J., Kuparinen, V., Lahteenmaki, K. & Korhonen, T. K. (2007). *J. Bacteriol.* **189**, 4539–4543.
- Bengoechea, J. A., Zhang, L., Toivanen, P. & Skurnik, M. (2002). *Mol. Microbiol.* **44**, 1045–1062.
- Binkowski, T. A., Naghibzadeh, S. & Liang, J. (2003). *Nucleic Acids Res.* **31**, 3352–3355.
- Bricogne, G., Vornrhein, C., Flensburg, C., Schiltz, M. & Paciorek, W. (2003). *Acta Cryst.* **D59**, 2023–2030.
- Charbonnier, J. B., Belin, P., Moutiez, M., Stura, E. A. & Quemeneur, E. (1999). *Protein Sci.* **8**, 96–105.
- Cohen, A. E., Ellis, P. J., Miller, M. D., Deacon, A. M. & Phizackerley, R. P. (2002). *J. Appl. Cryst.* **35**, 720–726.
- Cohen, S. X., Morris, R. J., Fernandez, F. J., Ben Jelloul, M., Kakaris, M., Parthasarathy, V., Lamzin, V. S., Kleywegt, G. J. & Perrakis, A. (2004). *Acta Cryst.* **D60**, 2222–2229.
- Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* **D50**, 760–763.
- Cruickshank, D. W. J. (1999). *Acta Cryst.* **D55**, 583–601.
- Davis, I. W., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2004). *Nucleic Acids Res.* **32**, W615–W619.
- Diamant, S., Eliahu, N., Rosenthal, D. & Goloubinoff, P. (2001). *J. Biol. Chem.* **276**, 39586–39591.
- Eidels, L. & Osborn, M. J. (1971). *Proc. Natl Acad. Sci. USA*, **68**, 1673–1677.
- Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* **D60**, 2126–2132.
- Fabrighini, I. P., Lehtio, L., Salminen, A., Zyryanov, A. B., Baykov, A. A., Lahti, R. & Goldman, A. (2004). *Biochemistry*, **43**, 14403–14411.
- Gomez, M. I. & Prince, A. (2007). *Curr. Opin. Pharmacol.* **7**, 244–251.
- Grochulski, P., Bouthillier, F., Kazlauskas, R. J., Serreqi, A. N., Schrag, J. D., Ziomek, E. & Cygler, M. (1994). *Biochemistry (Mosc.)*, **33**, 3494–3500.
- Grochulski, P., Li, Y., Schrag, J. D. & Cygler, M. (1994). *Protein Sci.* **3**, 82–91.
- Holm, L. & Sander, C. (1995). *Trends Biochem. Sci.* **20**, 478–480.
- Jedrzejas, M. J., Mello, L. V., de Groot, B. L. & Li, S. (2002). *J. Biol. Chem.* **277**, 28287–28297.
- Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795–800.
- Klock, H. E., Koesema, E. J., Knuth, M. W. & Lesley, S. A. (2008). *Proteins*, **71**, 982–994.
- Krishna, S. S., Weekes, D., Bakolitsa, C., Elsliger, M.-A., Wilson, I. A., Godzik, A. & Wooley, J. (2010). *Acta Cryst.* **F66**, 1143–1147.
- Krissinel, E. & Henrick, K. (2004). *Acta Cryst.* **D60**, 2256–2268.
- Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.
- Laskowski, R. A., Chistyakov, V. V. & Thornton, J. M. (2005). *Nucleic Acids Res.* **33**, D266–D268.
- LeMagueres, P., Im, H., Dvorak, A., Strych, U., Benedik, M. & Krause, K. L. (2003). *Biochemistry*, **42**, 14752–14761.
- Lesley, S. A. *et al.* (2002). *Proc. Natl Acad. Sci. USA*, **99**, 11664–11669.
- Marland, Z., Beddoe, T., Zaker-Tabrizi, L., Lucet, I. S., Brammananth, R., Whisstock, J. C., Wilce, M. C., Coppel, R. L., Crellin, P. K. & Rossjohn, J. (2006). *J. Mol. Biol.* **359**, 983–997.
- Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
- McPhillips, T. M., McPhillips, S. E., Chiu, H.-J., Cohen, A. E., Deacon, A. M., Ellis, P. J., Garman, E., Gonzalez, A., Sauter, N. K., Phizackerley, R. P., Soltis, S. M. & Kuhn, P. (2002). *J. Synchrotron Rad.* **9**, 401–406.
- Muthaiyan, A., Silverman, J. A., Jayaswal, R. K. & Wilkinson, B. J. (2008). *Antimicrob. Agents Chemother.* **52**, 980–990.
- Mykytczuk, N. C., Trevors, J. T., Leduc, L. G. & Ferroni, G. D. (2007). *Prog. Biophys. Mol. Biol.* **95**, 60–82.
- Najmanovich, R., Kurbatova, N. & Thornton, J. (2008). *Bioinformatics*, **24**, i105–i111.
- Nigou, J., Vasselon, T., Ray, A., Constant, P., Gilleron, M., Besra, G. S., Sutcliffe, I., Tiraby, G. & Puzo, G. (2008). *J. Immunol.* **180**, 6696–6702.
- Oguchi, Y., Takeda, K., Watanabe, S., Yokota, N., Miki, K. & Tokuda, H. (2008). *J. Biol. Chem.* **283**, 25414–25420.

- Okuda, S., Watanabe, S. & Tokuda, H. (2008). *FEBS Lett.* **582**, 2247–2251.
- Otagiri, M., Kurisu, G., Ui, S., Takusagawa, Y., Ohkuma, M., Kudo, T. & Kusunoki, M. (2001). *J. Biochem. (Tokyo)*, **129**, 205–208.
- Painter, J. & Merritt, E. A. (2006). *Acta Cryst. D* **62**, 439–450.
- Santarsiero, B. D., Yegian, D. T., Lee, C. C., Spraggon, G., Gu, J., Scheibe, D., Uber, D. C., Cornell, E. W., Nordmeyer, R. A., Kolbe, W. F., Jin, J., Jones, A. L., Jaklevic, J. M., Schultz, P. G. & Stevens, R. C. (2002). *J. Appl. Cryst.* **35**, 278–281.
- Schaumburg, J., Diekmann, O., Hagendorff, P., Bergmann, S., Rohde, M., Hammerschmidt, S., Jansch, L., Wehland, J. & Karst, U. (2004). *Proteomics*, **4**, 2991–3006.
- Sheldrick, G. M. (2008). *Acta Cryst. A* **64**, 112–122.
- Smoot, D. T. (1997). *Gastroenterology*, **113**, S31–S34.
- Takeda, K., Miyatake, H., Yokota, N., Matsuyama, S., Tokuda, H. & Miki, K. (2003). *EMBO J.* **22**, 3199–3209.
- Thoden, J. B., Timson, D. J., Reece, R. J. & Holden, H. M. (2004). *J. Biol. Chem.* **279**, 23431–23437.
- Truckses, D. M., Somoza, J. R., Prehoda, K. E., Miller, S. C. & Markley, J. L. (1996). *Protein Sci.* **5**, 1907–1916.
- Tymoczyszyn, E. E., Gomez-Zavaglia, A. & Disalvo, E. A. (2005). *Arch. Biochem. Biophys.* **443**, 66–73.
- Vriend, G. (1990). *J. Mol. Graph.* **8**, 52–56.
- Watanabe, S., Matsuyama, S. & Tokuda, H. (2006). *J. Biol. Chem.* **281**, 3335–3342.
- Winn, M. D., Murshudov, G. N. & Papiz, M. Z. (2003). *Methods Enzymol.* **374**, 300–321.
- Yancey, P. H. (2005). *J. Exp. Biol.* **208**, 2819–2830.
- Yang, H., Guranovic, V., Dutta, S., Feng, Z., Berman, H. M. & Westbrook, J. D. (2004). *Acta Cryst. D* **60**, 1833–1839.
- Ye, Y. & Godzik, A. (2004). *Nucleic Acids Res.* **32**, W582–W585.