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

Structure of the conserved hypothetical protein MAL13P1.257 from *Plasmodium falciparum*

Margaret A. Holmes,[a,b] Frederick S. Buckner,[a,c] Wesley C. Van Voorhis,[a,c] Christopher Mehlin,[a,b] Erica Boni,[a,b] Thomas N. Earnest,[a,d] George DeTitta,[a,e] Joseph Luft,[a,e] Angela Lauricella,[a,e] Lori Anderson,[a,b] Oleksandr Kalyuzhniy,[a,b] Frank Zucker,[a,b] Lori W. Schoenfeld,[a,b] Wim G. J. Hol[a,b,f] and Ethan A. Merritt[a,b]*

[a]Structural Genomics of Pathogenic Protozoa (SGPP) Consortium, USA, [b]Department of Biochemistry, University of Washington, Seattle, WA 98195-7742, USA, [c]Department of Medicine, University of Washington, Seattle, WA 98195, USA, [d]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [e]Hauptman–Woodward Institute, Buffalo, NY 14203, USA, and [f]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

Correspondence e-mail: merritt@u.washington.edu

Received 27 December 2005
Accepted 16 February 2006

PDB Reference: MAL13P1.257, 1zso, r1zsosf.

The structure of a conserved hypothetical protein, PlasmoDB sequence MAL13P1.257 from *Plasmodium falciparum*, Pfam sequence family PF05907, has been determined as part of the structural genomics effort of the Structural Genomics of Pathogenic Protozoa consortium. The structure was determined by multiple-wavelength anomalous dispersion at 2.17 Å resolution. The structure is almost entirely β-sheet; it consists of 15 β-strands and one short 3_{10}-helix and represents a new protein fold. The packing of the two monomers in the asymmetric unit indicates that the biological unit may be a dimer.

1. Introduction

The present structure determination of *Plasmodium falciparum* protein MAL13P1.257 (Kissinger *et al.*, 2002) was undertaken as part of the Structural Genomics of Pathogenic Protozoa (SGPP) consortium effort targeting proteins from eukaryotic tropical pathogens. One goal of structural genomics is to determine the structures of proteins that are members of sequence families with unknown folds. This protein was selected for structure determination because it is a conserved hypothetical protein belonging to Pfam family PF05907 (DUF866) (Bateman *et al.*, 2004), whose members have no significant sequence homology to any structure in the PDB. This sequence family contains eukaryotic proteins of unknown function. The *P. falciparum* protein, whose SGPP identifier is Pfal004331AAA, is 156 amino acids long, has a molecular weight of 18.7 kDa and has a theoretical pI of 4.6. Here, we report the structure of Pfal004331AAA at 2.17 Å resolution determined by multiple-wavelength anomalous dispersion and refined to an R value of 0.183 and a free R value of 0.233. The structure contains 15 β-strands and one 3_{10}-helix. This constitutes a three-dimensional fold not seen previously and provides the first structural basis for homology modeling of other members of the PF05907 sequence family. There are two closely associated molecules in the crystallographic asymmetric unit; their large buried surface area suggests that the biologically active unit may be a dimer.

2. Materials and methods

Ligase-independent cloning (LIC) was used to append a His tag to the N-terminus and a TAA stop codon to the C-terminus of the MAL13P1.257 gene of *P. falciparum*, giving MAHHHHHH-orf-TAA (SGPP identifier Pfal004331AAA). The vector has a T7 promoter for growth in *Escherichia coli* with auto-induction media (without the use of IPTG). Selenomethionine protein was produced according to the protocol of Studier (2005) in BL21 DE3 Star *E. coli*. Cells were lysed by sonication in 25 m*M* HEPES buffer with 500 m*M* NaCl, 0.2%(*w*/*v*) cholate, 0.1 mg ml^{-1} lysozyme, 1 m*M* β-mercaptoethanol and Roche EDTA-free protease inhibitors. Lysates were centrifuged to remove insoluble cell debris and the cleared lysates were tumbled with Qiagen Ni–NTA superflow resin. Purification followed the protocol reported by Mehlin *et al.* (2006). The resin was washed once with 10 m*M* imidazole and twice with 20 m*M* imidazole and eluted with 15 ml 250 m*M* imidazole in SGPP standard buffer (500 m*M* NaCl, 25 m*M* HEPES pH 7.25, 0.025% sodium azide, 5% glycerol). The

Table 1
Data-collection statistics.

Redundancy, completeness and R_{merge} are as reported by *HKL*2000; $I/\sigma(I)$ is as reported by *TRUNCATE*. Values in parentheses are for the highest resolution shell. Data beyond 2.17 Å were not used in refinement.

	Se peak (phasing)	Se inflection	Se remote	Refinement
Space group	$P2_12_12_1$			$P2_12_12_1$
Unit-cell parameters (Å)	a = 62.6, b = 71.0, c = 78.5			a = 62.2, b = 71.1, c = 78.6
V_M (Å^3 Da^{-1})	2.2			2.2
Wavelength (Å)	0.97957	0.97972	0.95372	0.9795
Resolution range (Å)	50–2.50 (2.59–2.50)	50–2.50 (2.59–2.50)	50–2.50 (2.59–2.50)	50–2.15 (2.23–2.15)
No. of unique reflections	11846	11883	12763	18151
Redundancy	6.9 (5.2)	6.8 (5.0)	6.8 (5.5)	5.9 (3.3)
Completeness (%)	92.9 (28.1)	93.8 (37.2)	99.9 (99.5)	92.8 (57.7)
R_{merge}	0.049 (0.128)	0.064 (0.278)	0.072 (0.341)	0.093 (0.365)
Mean $I/\sigma(I)$	53 (14)	40 (6.9)	35 (5.8)	27 (3.7)
Mean FOM for phasing	0.63			

Table 2
Refinement and model statistics.

Target ideal geometry is that of *REFMAC* v.5.2.0005. φ/ψ categorization is that of *PROCHECK*. Wilson *B* factor is as reported by *TRUNCATE*. Values of (B_{iso} + B_{TLS}) were generated by expanding the TLS description and individual ADP values of the refined model into an equivalent anisotropic description using the *CCP*4 programs *TLSEXTRACT* and *TLSANL* (Collaborative Computational Project, Number 4, 1994). We report the resulting equivalent isotropic ADP as B_{eq} = (B_{iso} + B_{TLS}). Values in parentheses are for the highest resolution shell.

Resolution range (Å)	20–2.17 (2.22–2.17)
R_{work}, 18076 (899) reflections	0.183 (0.209)
R_{free}, 926 (51) reflections	0.233 (0.314)
R.m.s.d. bonds (Å)	0.017
R.m.s.d. angles (°)	1.58
Residues in most favored region of φ/ψ	260 [88%]
Residues in additional allowed region of φ/ψ	36 [12%]
Residues in generously allowed region of φ/ψ	1 [0.3%]
Residues in disallowed region of φ/ψ	0
No. of protein atoms	2647
No. of water molecules	162
Wilson *B* factor (Å^2)	32
TLS groups chain *A*	9–27, 28–84, 85–118, 119–156
TLS groups chain *B*	6–31, 32–83, 84–118, 119–156
Mean [e.s.d.] of (B_{iso} + B_{TLS}) for atoms in chain *A* (Å^2)	37 [9]
Mean [e.s.d.] of (B_{iso} + B_{TLS}) for atoms in chain *B* (Å^2)	46 [13]

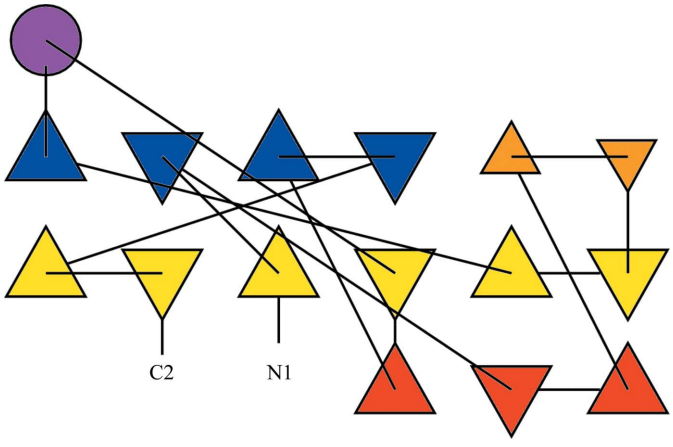

Figure 1
A topology diagram of the *A* chain of Pfal004331AAA, produced by the *TOPS* topology cartoon server (Michalopoulos *et al.*, 2004). The triangles represent β-strands and the circle represents the 3_{10}-helix. The upward-pointing triangles represent strands that come out of the plane of the diagram and downward-pointing triangles represent strands that go into the plane. The largest, six-stranded, β-sheet is colored yellow, the four-stranded sheet is blue, the three-stranded sheet is red, the two-stranded sheet is orange and the 3_{10}-helix is purple.

eluate was dialyzed overnight against standard buffer and aggregates were separated by size-exclusion chromatography the next day. Fractions were analyzed by SDS–PAGE and then pooled for concentration to 12.4 mg ml^{-1}. Dithiothreitol was added to 2 m*M* before proteins were flash-frozen for shipment to the crystal-screening and crystal-growth laboratories.

Pfal004331AAA was screened for crystallization at the Hauptman–Woodward Institute (Luft *et al.*, 2003). The protein was combined with 1536 different crystallization cocktail solutions in a single plate under mineral oil to prevent dehydration. Experiments were set up using standard commercially available liquid-handling systems. Plates were imaged over a four-week time course. Images were reviewed and crystallization conditions were forwarded to the SGPP crystal-growth laboratory in Seattle for optimization. There, crystallization conditions found in the initial large-scale screen (1.0 *M* LiCl, 20% PEG 6000, 0.1 *M* Tris pH 8) were optimized for pH, major precipitant and additive concentrations using a vapor-diffusion sitting-drop method. The crystallization conditions for the crystal used for structure solution and initial phasing were 1.0 *M* LiCl, 30% PEG 6000, 0.1 *M* Tris pH 8, 298 K; conditions for the crystal used for structure refinement were 1.0 *M* LiCl, 25 m*M* $Mg(NO_3)_2$, 25% PEG 6000, 0.1 *M* Tris pH 8, 298 K. The two crystals were cryoprotected in solutions that contained 21% xylitol and 30% glycerol, respectively, and flash-frozen in liquid nitrogen prior to shipping for data collection.

X-ray diffraction data were collected at the Advanced Light Source on beamline 8.2.1 for the crystal used in MAD phasing and on beamline 8.2.2 for the crystal used in refinement. All data were integrated and scaled using *HKL*2000 (Otwinowski & Minor, 1997). The asymmetric unit contains two monomers of Pfal004331AAA, which contain a total of eight SeMet residues, including those in the

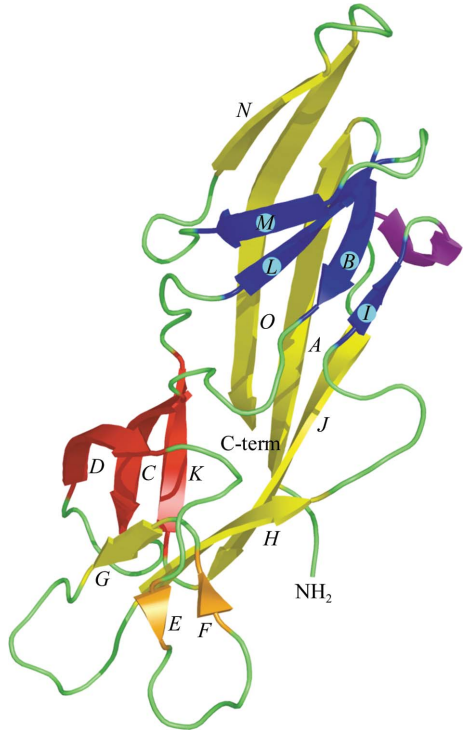

Figure 2
A ribbon representation of the *A* chain of Pfal004331AAA. The secondary-structure elements are colored as in Fig. 1. Each sheet is labeled alphabetically, denoting its position in the sequence. The figure was generated using *PyMOL* (DeLano, 2002).

N-terminal tag. The program *SOLVE* (Terwilliger, 2003) was used to locate the Se atoms and to calculate phases. The program *RESOLVE* was used for density modification, twofold NCS map averaging and automatic tracing. *SOLVE* and *RESOLVE* were run a number of times with both 3.0 and 2.5 Å outer resolution limits, each time finding four Se atoms, two in each of the two chains. A composite model was assembled from the resultant autotraced models, using the density-modified maps, with the program *XFIT* (McRee, 1999). Refinement of this model was then carried out against the peak data from the higher resolution data set using the program *REFMAC*5 (Murshudov *et al.*, 1997) *via* the *ccp*4*i* interface (Potterton *et al.*, 2004), alternating with rounds of building in *XFIT*. No NCS restraints were used in refinement. Data-collection statistics are shown in Table 1 and refinement statistics are shown in Table 2.

In the last three cycles of refinement, each chain was described by four TLS groups identified by the *TLSMD* server (Painter & Merritt, 2006) and TLS parameters were refined for each group (Table 2). Eight N-terminal residues of chain *A* and five N-terminal residues of chain *B*, including three well ordered residues from the His tag, were not assigned to any TLS group. Refinement was monitored using 5% of the data reserved for $R_{\rm free}$. Final structure validation was performed with *PROCHECK* (Laskowski *et al.*, 1993) and *MolProbity* (Lovell *et al.*, 2003). The model of the *A* chain consists of all 156 residues in the open reading frame, while the model of the *B* chain additionally contains the last three residues of the His tag.

3. Results

The structure of Pfal004331AAA has been solved to 2.17 Å resolution. The structure contains 15 β-strands, which form four β-sheets. The largest sheet is comprised of six strands and runs the length of the molecule. There are two smaller sheets, one of four strands and one of three strands, which flank the large sheet. A short two-stranded sheet and a short 3_{10}-helix complete the secondary structure. Prior to determining the structure of Pfal004331AAA, an iterative *PSI-BLAST* search (Altschul *et al.*, 1997) for protein sequences with significant similarity to Pfal004331AAA found none with an annotated function or with a known structure. After the structure had been determined, it was submitted to the *DALI* server (Holm & Sander, 1993) and to the *VAST* server (Gibrat *et al.*, 1996; Madej *et al.*, 1995) in order to search for structures with similar topology. No significant match to any structure in the PDB was reported by either server. Therefore, we conclude that the structure of Pfal004331AAA constitutes a previously unobserved protein fold.

Fig. 1 shows a topology diagram of the monomer. The largest sheet is composed of six strands [βA (Thr4–Glu13), βG (Phe61–Ile63), βH (Ser72–Val77), βJ (Arg94–Arg102), βN (Trp135–Asn139) and βO (Met144–Asn156)] arranged in an antiparallel fashion, which account for almost 30% of the total residues. This largest sheet extends the length of the molecule and is flanked by three smaller β-sheets (Fig. 2). One of these is an antiparallel four-stranded sheet [βB (Val15–Phe19), βI (Ser84–Glu86), βL (Phe115–Asp119) and βM (Leu124–Val128)], which is positioned at one end of the large β-sheet. A three-stranded β-sheet [βC (Trp26–Asp32), βD (Thr38–Phe44) and βK (Ile104–Phe109)] is positioned at the other end of the large sheet. A small antiparallel two-stranded β-sheet [βE (Leu50–Glu51) and βF (Thr58–Ala59)] lies across the large sheet from the three-stranded sheet. There is a short 3_{10}-helix (Val88–Asn90) between one edge of the large sheet and the four-stranded sheet.

The distribution of electrostatic potential at the protein surface is unremarkable and there are no extensive hydrophobic patches. The surface of the chain *A* monomer has one obvious cavity; it is the largest pocket found by the *CASTp* server (Binkowski *et al.*, 2003; Liang *et al.*, 1998) and has a volume of 310 Å^3. This pocket is formed mainly by adjacent edges of the large sheet and the four-stranded sheet and by the 3_{10}-helix which connects them (atoms from residues 5, 8, 79–86, 89 and 94–98). The cavity is diametrically opposite the proposed dimer interface (see below). It is lined with both non-polar and polar residues and is occupied by five water molecules (Fig. 3). The equivalent cavity in chain *B* is smaller owing to different side-chain conformations of Asn89 and Arg94 and contains only two well ordered water molecules.

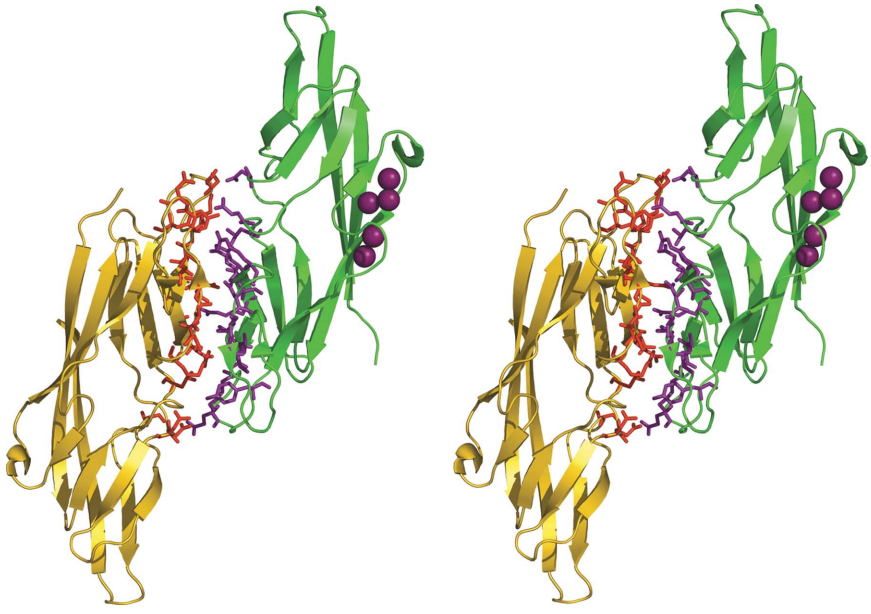

Figure 3
Stereoview of the proposed Pfal004331AAA dimer. The *A* chain is on the right in green in roughly the same orientation as in Fig. 2; the *B* chain is on the left in gold. Residues from the *A* chain that are within 4 Å of the *B* chain are shown as purple sticks and residues from the *B* chain that are within 4 Å of the *A* chain are shown in red. The five water molecules occupying the single significant cavity in monomer *A* are shown as purple spheres. The figure was generated using *PyMOL*.

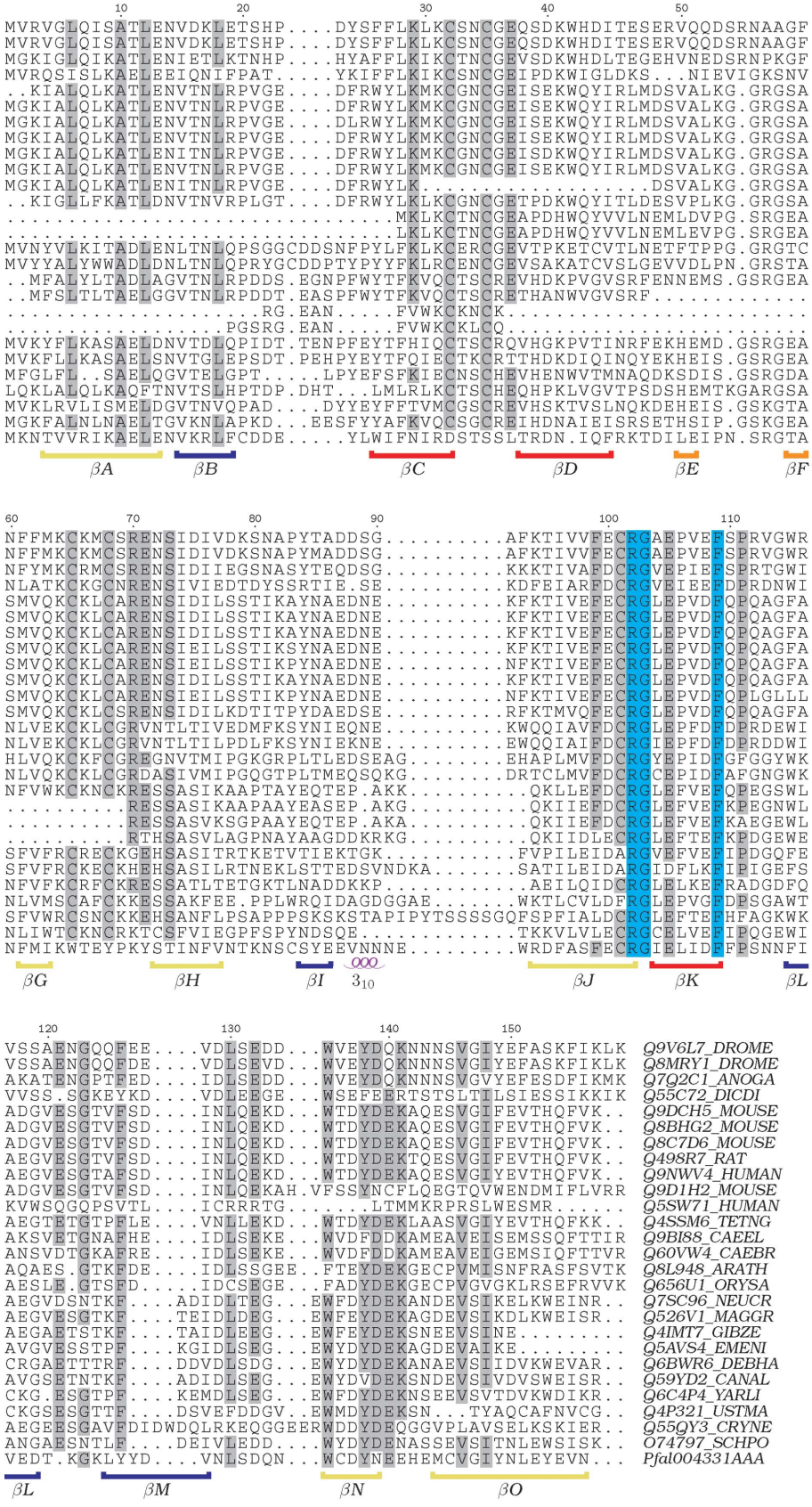

Figure 4
CLUSTALW multiple sequence alignment for representative sequences from Pfam family PF05907 (DUF866). Residue positions with sequence identity greater than 60% are shaded gray and residue positions that are 100% conserved are shaded cyan. Sequences are identified by their TrEMBL entry names. Residue numbers correspond to those of the *P. falciparum* structure reported here. Bars indicating individual β-strands are colored as in Figs. 1 and 2. Figure generated using T_EX*shade* (Beitz, 2000).

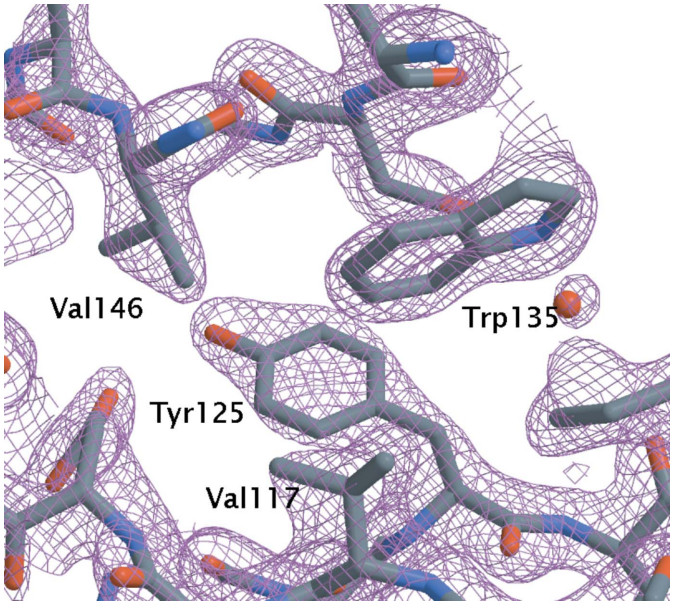

Figure 5
Electron density for a highly conserved hydrophobic core region. Residues Tyr125 (Phe in other sequence family members), Trp135 and Val146 are highly conserved in the PF05907 sequence family. They form a hydrophobic core region anchoring β-strands *M*, *N* and *O*. The density shown is contoured at 2.5σ in a σ_A-weighted ($2mF_o - DF_c$) electron-density map.

The crystallographic asymmetric unit contains two closely associated monomers. Their conformation is essentially identical, with an r.m.s.d. of 0.49 Å for 152 C^{α} positions after superposition. Deviations were observed only at the termini (residues 1–3 and 156) and at two short loops involved in crystal-packing contacts (residues 54–57 and 79–81). The monomer–monomer interaction was characterized using the *Protein–Protein Interaction Server* (Jones & Thornton, 1996). The total interface accessible surface area of chain *A* is 890 Å^2 and that for chain *B* is 860 Å^2. These areas and the associated gap volume index of 3.0 are consistent with a weak homodimer. Fig. 3 shows the proposed dimer. The two chains interact directly *via* 11 hydrogen bonds and indirectly *via* a network of nine water molecules that are trapped between the two chains, reducing the surface complementarity. The chromatogram from the size-exclusion step of protein purification showed that about 20% of the protein ran as a dimer (not shown); however, only protein from the major (80%) peak was used for crystallization trials. We conclude that Pfal004331AAA most likely acts as a weak dimer; the physiological dimerization state is not certain.

4. Discussion

The structure of *P. falciparum* protein MAL13P1.257 is the first structure to be determined of a member of Pfam sequence family PF05907. The biological function of these proteins is unknown. This family is currently represented by 48 sequences indexed by *Interpro* 12.0 (Mulder *et al.*, 2005) with an average length of 165 residues; the domain structure of 33 of these has been characterized (Bateman *et al.*, 2004). These are single-domain proteins, with certain exceptions. A complete DUF866 domain appears at the N-terminus of a large (1248 residues) coding sequence from *Ustilago maydis*, UM05492.1, containing no other annotated domain structure. There is also strong similarity to a 68-residue stretch of *Aspergillus nidulans* ORF AN7606.2 (Q5AVS4_EMENI in Fig. 4), which contains in addition an FAD-dependent phenol hydroxylase domain and an LSM domain (Pfam PF01423). A somewhat longer stretch of sequence similarity is found in *Giberella zeae* (Fig. 4; Q4IMT7_GIBZE), which contains a second domain annotated as a mitochondrial RNA-processing domain (Pfam PF08296).

There are two regions of high sequence conservation between the present protein and the other PF05907 family members. One region is a hydrophobic core formed by residues Tyr125, Trp135 and Val146 (Fig. 5). The second region spans β-strands *J* and *K* and consists of residues 99–111. Curiously, several residues which are highly conserved across other family members are not conserved in the present *P. falciparum* protein. For example, it is striking that two C*XX*C motifs are strongly conserved in all other family members (Fig. 4 and additional sequences not shown) but are not present in *P. falciparum*. These correspond to residues 32–35 and 65–68 in the present structure. Neither the residues surrounding the surface cavity nor the residues involved in the dimer interface are maintained across the sequence family. These observations may indicate that specific biological function is not conserved throughout the entire sequence family. Nevertheless, the present structure provides a basis for structural modeling of a previously uncharacterized eukaryotic sequence family, one with a three-dimensional fold that has not previously been observed.

We are grateful for the contributions of other SGPP consortium members, including Peter Myler, Elizabeth Worthey, Tracy Arakaki, Jürgen Bosch, Jonathan Caruthers, Mark Robien, Christophe Verlinde, Larry de Soto and Martin Criminale. Portions of this work were carried out at the Advanced Light Source, which is supported by the Director, Office of Science, Office of Basic Energy Sciences of the US Department of Energy under Contract No. DE-AC02-05CH11231. This work was supported by NIH awards GM64655 and GM62617.

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.
Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C. & Eddy, S. R. (2004). *Nucleic Acids Res.* **32**, D138–D141.
Beitz, E. (2000). *Bioinformatics*, **16**, 135–139.
Binkowski, T. A., Naghibzadeh, S. & Liang, J. (2003). *Nucleic Acids Res.* **31**, 3352–3355.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
DeLano, W. L. (2002). *The PyMOL Molecular Graphics System.* http://www.pymol.org.
Gibrat, J. F., Madej, T. & Bryant, S. H. (1996). *Curr. Opin. Struct. Biol.* **6**, 377–385.
Holm, L. & Sander, C. (1993). *J. Mol. Biol.* **233**, 123–138.
Jones, S. & Thornton, J. M. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
Kissinger, J. C. *et al.* (2002). *Nature (London)*, **419**, 490–492.
Laskowski, R., MacArthur, M., Moss, D. & Thornton, J. (1993). *J. Appl. Cryst.* **26**, 283–291.
Liang, J., Edelsbrunner, H. & Woodward, C. (1998). *Protein Sci.* **7**, 1884–1897.
Lovell, S., Davis, I., Arendall, W. B. III, de Bakker, P., Word, J., Prisant, M., Richardson, J. & Richardson, D. (2003). *Proteins*, **50**, 437–450.
Luft, J. R., Collins, R. J., Fehrman, N. A., Lauricella, A. M., Veatch, C. K. & DeTitta, G. T. (2003). *J. Struct. Biol.* **142**, 170–179.
McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.
Madej, T., Gibrat, J. F. & Bryant, S. H. (1995). *Proteins*, **23**, 356–369.
Mehlin, C., Boni, E., Buckner, F. S., Engel, L., Feist, T., Gelb, M., Haji, L., Kim, D., Liu, C., Mueller, N., Myler, P. J., Reddy, J. T., Sampson, J. N., Subramanian, E., Van Voorhis, W. C., Worthey, E., Zucker, F. & Hol, W. G. J. (2006). *Mol. Biochem. Parasitol.* In the press..

Michalopoulos, I., Torrance, G. M., Gilbert, D. R. & Westhead, D. R. (2004). *Nucleic Acids Res.* **32**, D251–D254.
Mulder, N. J. *et al.* (2005). *Nucleic Acids Res.* **33**, D201–D205.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Painter, J. & Merritt, E. A. (2006). *J. Appl. Cryst.* **39**, 109–111.
Potterton, L., McNicholas, S., Krissinel, E., Gruber, J., Cowtan, K., Emsley, P., Murshudov, G. N., Cohen, S., Perrakis, A. & Noble, M. (2004). *Acta Cryst.* D**60**, 2288–2294.
Studier, F. W. (2005). *Protein Expr. Purif.* **41**, 207–234.
Terwilliger, T. C. (2003). *Methods Enzymol.* **374**, 22–37.