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

Structure of a *Trypanosoma brucei* α/β-hydrolase fold protein with unknown function

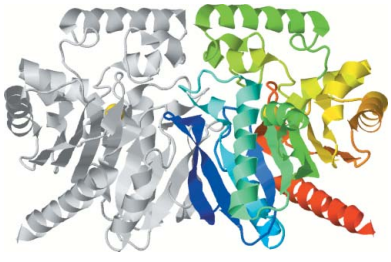
Ethan A. Merritt,[a,b]* Margaret Holmes,[a,b] Frederick S. Buckner,[a,c] Wesley C. Van Voorhis,[a,c] Erin Quartly,[a,d] Eric M. Phizicky,[a,d] Angela Lauricella,[a,e] Joseph Luft,[a,e] George DeTitta,[a,e] Helen Neely,[a,b] Frank Zucker[a,b] and Wim G. J. Hol[a,b]

[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 98195-7185, USA, [d]Department of Biochemistry and Biophysics, University of Rochester School of Medicine, Rochester, NY 14642, USA, and [e]Hauptman–Woodward Institute, Buffalo, NY 14203, USA

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

Received 7 April 2008
Accepted 21 April 2008

PDB Reference: Tbru020260AAA, 2q0x, r2q0xsf.



The structure of a structural genomics target protein, Tbru020260AAA from *Trypanosoma brucei*, has been determined to a resolution of 2.2 Å using multiple-wavelength anomalous diffraction at the Se *K* edge. This protein belongs to Pfam sequence family PF08538 and is only distantly related to previously studied members of the α/β-hydrolase fold family. Structural superposition onto representative α/β-hydrolase fold proteins of known function indicates that a possible catalytic nucleophile, Ser116 in the *T. brucei* protein, lies at the expected location. However, the present structure and by extension the other trypanosomatid members of this sequence family have neither sequence nor structural similarity at the location of other active-site residues typical for proteins with this fold. Together with the presence of an additional domain between strands $\beta6$ and $\beta7$ that is conserved in trypanosomatid genomes, this suggests that the function of these homologs has diverged from other members of the fold family.

1. Introduction

Sequence family Pfam PF08538 (Interpro IPR013744) consists of a set of coding sequences identified in the genomes of plant, fungal and protozoan species. The sequence family as a whole is recognizably related to other sequence families in the large α/β-hydrolase fold class; however, the pairwise sequence identity of individual members of PF0538 to any protein of known function is very low. We have determined the three-dimensional structure of a representative member of this sequence family from the eukaryotic parasite *Trypanosoma brucei*. This work was performed as part of a structural genomics project targeting medically important parasitic protozoa (Fan *et al.*, 2008).

The α/β-hydrolase fold class contains at least 35 sequence superfamilies representing a variety of biochemical activities. The enzymatically well characterized members of this fold class contain a catalytic residue at a highly conserved location connecting strand $\beta5$ of the central β-sheet to helix $\alpha5$. This catalytic nucleophile is variously a Ser, a Cys or an Asp depending on the specific enzymatic activity of the corresponding protein family (Ollis *et al.*, 1992; Heikinheimo *et al.*, 1999). In many of these families the active site contains a classic Ser/His/Asp catalytic triad, but others such as the epoxide hydrolases instead use an Asp/His/Asp triad, while some acetylcholinesterases use Ser/His/Glu. The His residue is always C-terminal to the other two residues of the triad, but is sometimes contributed by a structural element that is outside the core α/β topology (Heikinheimo *et al.*, 1999). Thus, the identity and precise location of the second and third members of the catalytic triad are less strongly conserved than the primary catalytic residue. The nature of the active site is unclear for some family members that are only known from genome sequencing, such as the *T. brucei* protein whose structure is presented here (Fig. 1).

2. Methods

2.1. Target selection and expression

T. brucei genomic sequence Tb10.6k15.0140 was selected as an SGPP structural genomics target based on sequence length, predicted

pI, predicted disorder, size of the Pfam sequence family and distance from mammalian homologs. The sequence was PCR-amplified from genomic DNA of *T. brucei* strain TREU927 GUTat 10.1 and cloned into *Escherichia coli* expression vector AVA421, which is derived from pET14b (Alexandrov *et al.*, 2004). The AVA421 vector contains a cleavable N-terminal His tag. The protein was purified using an Ni–NTA column and the bound protein was cleaved by protease 3C overnight at 277 K. The released protein was further purified by gel filtration on a HiLoad Superdex 200 column.

2.2. Protein crystallization

The purified protein was screened at the high-throughput facility at the Hauptman–Woodward Institute (Luft *et al.*, 2003) to identify initial crystallization conditions. The frozen sample (193 K) was rapidly thawed in a 303 K water bath prior to setup (Deng *et al.*, 2004). The sample was combined with 1536 distinct crystallization cocktails in 10 min in a single microassay plate (Greiner BioOne, 790801). After setup, each well of the plate held a unique microbatch-under-oil crystallization experiment (Chayen *et al.*, 1992) containing 200 nl of the sample combined with 200 nl crystallization cocktail under 5 µl USP-grade mineral oil (Sigma, M-1180). The experiment plate was stored at 277 K for one week and then imaged at 296 K. Images were manually reviewed; 51 of the 1536 crystallization experiments produced outcomes that were suitable for subsequent optimization trials.

Initial hits were optimized and crystals for data collection were grown by the sitting-drop method. The crystallization drop consisted

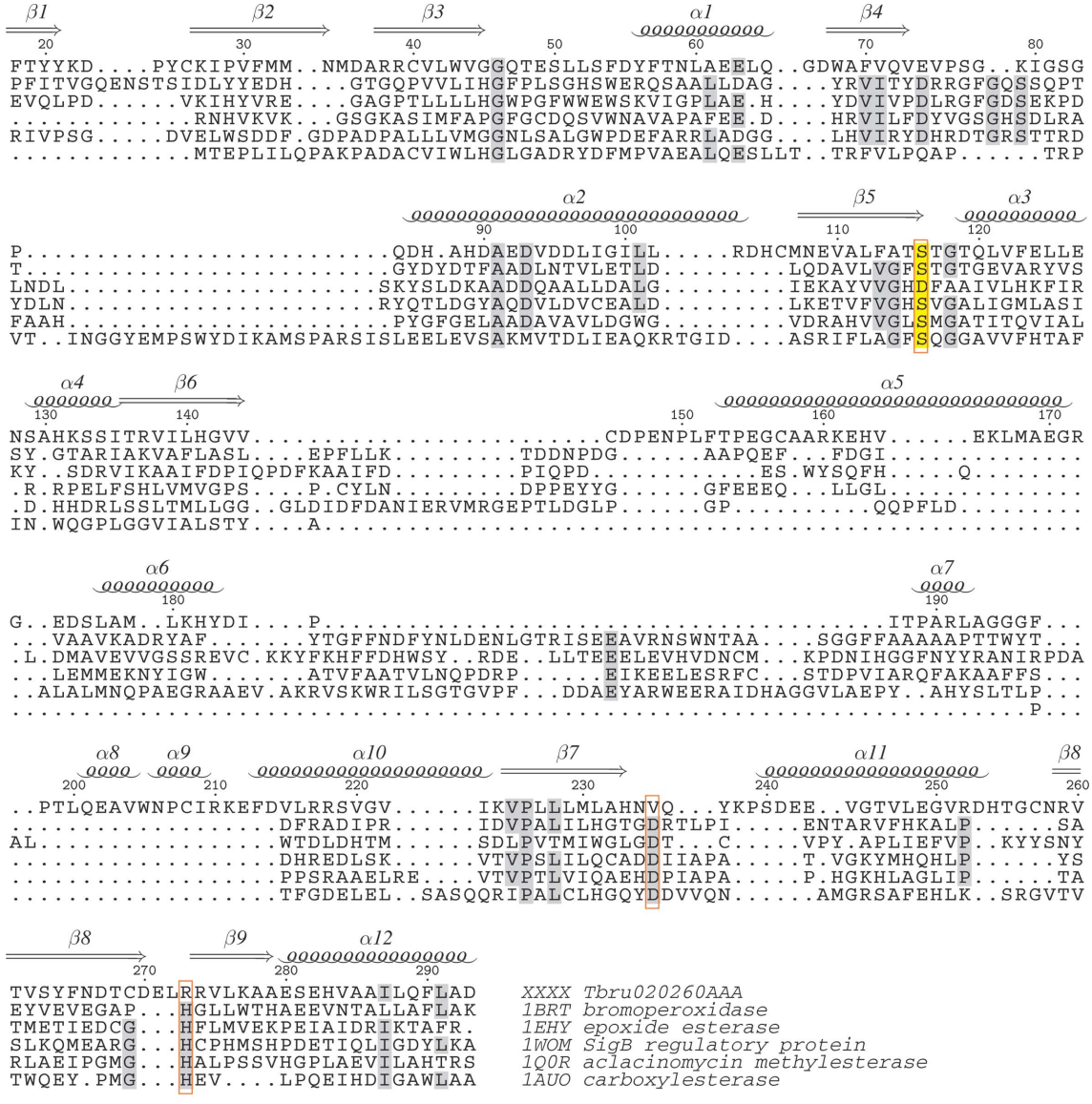


Figure 1
Structure-based sequence alignment of representative α/β-hydrolase fold proteins. Representative structures from five functionally distinct sequence superfamilies within the α/β-hydrolase fold class were jointly aligned using the *CE-MC* server (Guda *et al.*, 2004). These families were chosen on the basis of highest overall structural similarity to the *T. brucei* protein. Residues are highlighted in gray if they are identical in three or more of the sequences after structural superposition. Secondary-structural elements are shown corresponding to the *T. brucei* structure reported here. The spatial locations of the eight core β-strands are conserved across the fold class, but insertions between them are common. Except for 1au0, the sequences shown here have an insertion between β6 and β7. However, the *T. brucei* insertion is very different from the others in both sequence and structure. All characterized members of the fold class contain a catalytic residue at the end of strand β5, highlighted in yellow in this figure. In the *T. brucei* structure this corresponds to Ser116. The location of two other residues forming a catalytic triad is conserved in the four proteins of known function but not in the *T. brucei* protein, as indicated by the boxes.

Table 1
Data-collection and phasing statistics.

Values in parentheses are for the highest resolution shell.

	MAD λ_1	MAD λ_2	MAD λ_3	Native
Space group	$P6_1$			$P6_1$
Unit-cell parameters (Å)	a = 63.48, b = 63.48, c = 303.9			a = 63.56, b = 63.56, c = 303.2
Wavelength (Å)	0.9798	0.9797	0.9537	0.9795
Resolution (Å)	50–3.0 (3.11–3.0)	50–3.0 (3.11–3.0)	50–3.0 (3.11–3.0)	55–2.20 (2.32–2.20)
Total unique reflections	13767	13357	13695	37476
R_{merge}	0.089 (0.286)	0.106 (0.320)	0.096 (0.326)	0.057 (0.609)
Completeness (%)	91 (44)	98 (36)	90 (40)	100 (99)
$I/\sigma(I)$	11.6 (1.3)	10.2 (1.3)	11.0 (1.3)	16.1 (2.7)
Redundancy				2.9 (1.9)
Wilson B (Å^2)				49

of 0.4 µl protein solution (10.5 mg ml^{-1}) mixed with 0.4 µl reservoir solution containing 35% PEG 400, 0.1 *M* MgCl, 5 m*M* DTT, 0.1 *M* MES pH 6.0. The protein buffer contained 0.5 *M* NaCl, 2 m*M* BME, 0.025% NaN_3, 5% glycerol, 20 m*M* HEPES pH 7.5.

2.3. Data collection and structure determination

Data from a single crystal of native protein were collected to 2.2 Å resolution on SSRL beamline 9-2 and integrated using *MOSFLM* (Leslie, 1992) *via* the automated script package *ELVES* (Holton & Alber, 2004). Molecular-replacement attempts using the α/β-hydrolasefold core from various previously determined structures in the family as probes did not succeed. Three-wavelength data from a single crystal of SeMet-derivatized protein were collected on ALS beamline 8.2.2 and integrated using *HKL*-2000 (Otwinowski & Minor, 1997). *SOLVE* (Terwilliger & Berendzen, 1999) found nine Se sites and produced a set of initial phases to 3.2 Å resolution from the three-wavelength MAD data. These experimental phases were then merged with the observed F and $\sigma(F)$ values from the native data to 2.2 Å resolution and the merged data were fed into *RESOLVE* for phase extension, noncrystallographic symmetry density averaging, solvent flattening and initial autotracing (Terwilliger & Berendzen, 1999; Terwilliger, 2003). At this stage, *RESOLVE* was able to autotrace backbone segments corresponding to 405 residues in 36 fragments out of a total of 670 residues expected for the two monomers. This partial trace confirmed the presence of the core β-sheet of the expected α/β-hydrolase fold, but also indicated that the associated α-helices were sufficiently different from previously structures in this fold class to explain the failure of molecular replacement.

It is worth noting that the *RESOLVE* protocol 'resolve_build' was only able to identify 77 side-chain residues at this point and could not develop the model further without manual intervention. The model-building program *Coot* (Emsley & Cowtan, 2004) was used to edit out dubious trace fragments and to manually assign side chains for several fragments relative to Met residues clearly indicated by corresponding peaks in a Bijvoet difference Fourier map. We then fed this partial structural model back into the 'resolve_build' protocol along with the native data to 2.2 Å but with no experimental phases. From this starting point, *RESOLVE* autotraced 485 residues and assigned 302 side chains. The remaining residues were built by hand using *Coot* and refined in *REFMAC*5 (Murshudov *et al.*, 1997). In the final cycles of refinement, each monomer chain was described by six TLS groups identified by the *TLSMD* server (Painter & Merritt, 2006*b*) and TLS parameters were refined for each group. Crystallographic statistics are presented in Tables 1 and 2. Waters were added using *Coot*. Model quality was validated using *Coot* and *MolProbity* (Lovell *et al.*, 2003). The final model consists of residues 8–301 of chain *A* and residues 9–301 of chain *B*. One well ordered glycerol molecule was found to be associated with each monomer. The highest residual electron density after refinement lies in the region of residue

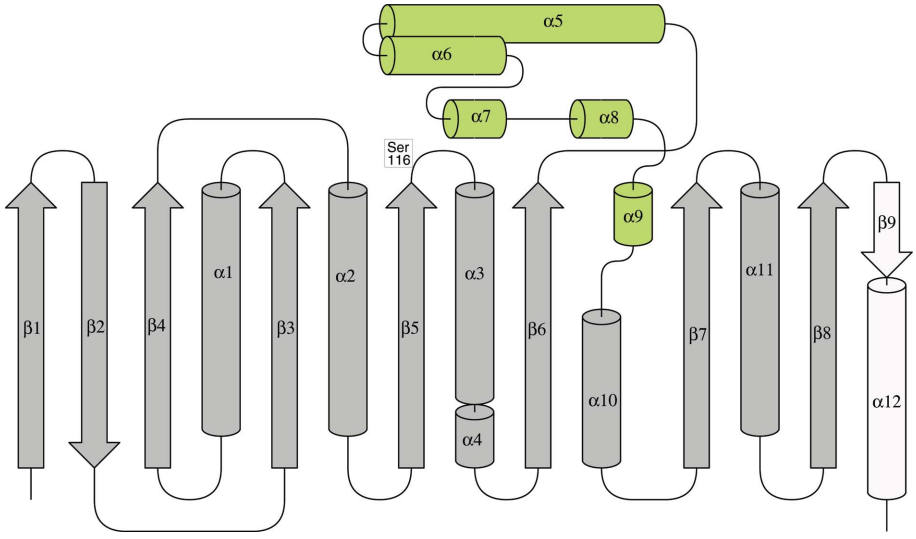


Figure 2
Topology of *T. brucei* protein Tbru020260AAA. Elements of secondary structure shown in gray correspond to the canonical α/β-hydrolase fold. Residues 145–211 of the *T. brucei* protein, shown in green, form a cap over one end of the dimer that is structurally divergent from other proteins with this overall fold.

Table 2
Refinement statistics.

Resolution (Å)	55–2.20
R	0.206
R_{free}	0.250
R.m.s.d. bonds (Å)	0.012
R.m.s.d. angles (°)	1.247
Protein atoms	4541
Nonprotein atoms	135
Residues in favored regions (%)	96
Residues in disallowed regions	0
TLS groups (residues)	*A*, 8–26, 27–107, 108–150, 151–211, 212–243, 244–301; *B*, 9–31, 32–73, 74–150, 151–211, 212–243, 244–301
Mean B_{iso} + B_{TLS}, protein atoms (Å^2)	53
Mean B_{iso}, nonprotein atoms (Å^2)	49

301 of each monomer and presumably corresponds to portions of the C-terminal remainder of the 335-residue chain; however, the density is not of sufficient quality to allow modeling of this region.

There is one notable site of deviation from typical φ/ψ backbone torsion angles in each monomer. This is the site of the conserved catalytic nucleophile, Ser116 in the present structure, which lies on a sharply bent nucleophile elbow that is characteristic of the α/β-hydrolase fold (Heikinheimo *et al.*, 1999).

3. Results

The present structure confirms that proteins in the PF08538 sequence family belong to the α/β-hydrolase fold class. The core β-sheet strands 1–8 follow the canonical fold very closely, as do helices α1, α2, α3 and α11. However, the secondary-structural elements between β6 and α10, comprised of residues 145–211, form an inserted capping domain which lies above the nucleophile elbow containing Ser116 (Figs. 2 and 3). The closest structural neighbors in the PDB are various haloalkane dehalogenases and epoxide esterases. Of these, the top hits are bromoperoxidase A2 (PDB code 1brt; sequence identity 16% and C$^\alpha$ r.m.s.d. 3.6 Å for 205 residues) and the epoxide hydrolase from *Agrobacterium radiobacter* AD1 (PDB code 1ehy; sequence identity 11% and C$^\alpha$ r.m.s.d. 3.6 Å for 205 residues). The haloalkane dehalogenases and epoxide esterases also have a helical insertion between strands β6 and β7, but these do not resemble that of the *T. brucei* protein in either sequence or structure.

The *T. brucei* protein is observed as a dimer whose monomer–monomer interface buries 2000 Å^2 on each monomer surface. The interface is formed by residues in the twofold-related loops between β3 and α1 and between α5 and α6 and by the formation of an inter-chain antiparallel β-sheet between residues 13–17 in strand β1 of each monomer. The dimer association creates an extended flat ridge along the top of the molecule whose surface is formed by juxta-position of helices α5 and α6 from each monomer (Fig. 3). This extended surface is entirely defined by residues from the subfamily-specific α-helical insertion between core strands β6 and β7.

4. Discussion

The α/β-hydrolase fold class as a whole is characterized by conserved secondary structure and by a catalytic triad in the sequence order nucleophile/histidine/acid (Heikinheimo *et al.*, 1999). Pfam sequence family PF08538 contains genomic sequences of unknown function identified primarily from fungal and plant genomes. These sequences are remote in sequence space from functionally characterized members of the α/β-hydrolase fold class, but are consistent with the defining characteristics of the fold. Many, but not all, of them also exhibit a sequence motif (Prosite PS00120) that is associated with serine lipases. The sequence family also contains a single proteobacterial sequence, from *Thiomicrospira crunogena*, that exhibits all of these features.

Genomic sequences from the protozoan parasites *Trypanosoma brucei*, *T. cruzi* and *Leishmania major* have also been assigned to this family. The *T. brucei* structure reported here supports the assignment of these protozoan sequences to the α/β-hydrolase fold class, with an insertion between the canonical strands β6 and β7. The *T. cruzi* and *L. major* sequences are 66% and 33% identical to the *T. brucei* homolog, respectively, and this level of similarity applies to the insertion region as well as to the core fold regions. The *T. brucei* structure contains a candidate catalytic residue, Ser116, at the expected location of the conserved catalytic nucleophile, protruding into a deep binding cleft and exhibiting unusual backbone torsion angles. This is often an indicator of functional importance. However, there are no obvious candidates for the other two expected members of the catalytic triad. The active site of other members of the fold class typically contains a His residue protruding from a long loop following canonical strand β8. The *T. brucei* structure contains only a short loop at this position and instead the corresponding region of the putative active site is occupied by Ile185 from the inserted helix α6. His141 is the sole His residue in the vicinity of Ser116. The observed conformation of these two residues is such that the histidine side chain extends away from the Ser O$^\gamma$ and it would require a substantial conformational change to yield reasonable geometry for a Ser/His interaction. Furthermore, the conserved Asp which forms the third member of the catalytic triad in other family members is not present in the trypanosomatid sequences. Thus, notwithstanding the presence of Ser116 as a candidate nucleophile, it seems unlikely that the putative active site in the *T. brucei* protein can act as a hydrolase. Furthermore, the *T. cruzi* and *L. major* sequences have lost even the candidate nucleophile, containing Gly rather than Ser at this position (GeneDB sequences Tc00.1047053508307.70 and LmjF36.4780). Therefore, it seems very likely that while these proteins have retained key features of the α/β-hydrolase fold, they are not in fact hydrolases.

What, then, is their biological function? The deep binding cleft leading to the active site characteristic of the fold family is still present in the current structure and may retain a binding specificity inherited from an enzymatically active ancestral protein. The inserted α-helical domain and in particular helices α5 and α6 may recognize a

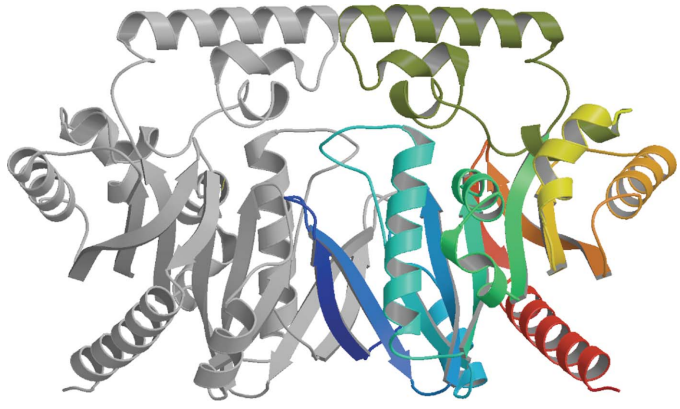


Figure 3
One monomer of the dimeric *T. brucei* protein is shown in gray. The other monomer is colored from blue at the N-terminus to red at the C-terminus. The eight strands making up the core β-sheet of the α/β-hydrolase fold in this monomer are arranged left to right as in Fig. 2. Secondary-structure elements between β6 and β7 (dark green in the figure) form a capping domain at the top of the structure.

specific partner protein, as is the case for analogous insertions in other proteins with this fold (Heikinheimo *et al.*, 1999). The inserted domain has no significant structural similarity to other structures in the PDB.

In the course of refining the crystal structure, we analyzed the vibrational modes inherent in the protein as observed in the crystal structure by fitting multi-group TLS models to the three-dimensional distribution of crystallographic *B* factors (Painter & Merritt, 2006*a*). *N*-group models with $N \geq 4$ consistently identified residues 151–211 ($\alpha5 \rightarrow \alpha9$) as making up a subdomain that undergoes concerted vibrational motion relative to the rest of the structure. This is consistent with the topology of the protein and with the hypothesis that this insertion constitutes a specific recognition domain that has been grafted onto the basic α/β-hydrolase core fold.

We conclude that the subfamily of Pfam PF08538 sequences represented by *T. brucei* sequence Tbru020260AAA and by the current structure constitutes an evolutionary offshoot of the larger α/β-hydrolase sequence family. The trypanosomatid homologs appear to have lost the residues necessary for hydrolytic activity, while adding a subfamily-specific domain between strands $\beta6$ and $\beta7$ of the canonical hydrolase fold. The biological function of this protein in trypanosomatids remains unknown.

Financial support from the Protein Structure Initiative award NIGMS GM64655 and from NIAID award AI067921 is gratefully acknowledged. Portions of this research were carried out at the Stanford Synchrotron Radiation Laboratory, 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. Other 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.

References

Alexandrov, A., Vignali, M., LaCount, D. J., Quartley, E., de Vries, C., Rosa, D. D., Babulski, J., Mitchell, S. F., Schoenfeld, L. W., Fields, S., Hol, W. G., Dumont, M. E., Phizicky, E. M. & Grayhack, E. J. (2004). *Mol. Cell. Proteomics*, **3**, 934–938.
Chayen, N. E., Shaw Stewart, P. D. & Blow, D. M. (1992). *J. Cryst. Growth*, **122**, 176–180.
Deng, J., Davies, D. R., Wisedchaisri, G., Wu, M., Hol, W. G. J. & Mehlin, C. (2004). *Acta Cryst.* D**60**, 203–204.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Fan, E. *et al.* (2008). *Methods Mol. Biol.* **426**, 497–513.
Guda, C., Lu, S., Scheeff, E. D., Bourne, P. E. & Shindyalov, I. N. (2004). *Nucleic Acids Res.* **32**, W100–W103.
Heikinheimo, P., Goldman, A., Jeffries, C. & Ollis, D. L. (1999). *Structure*, **7**, R141–R146.
Holton, J. & Alber, T. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 1537–1542.
Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **26**.
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.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Ollis, D. L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S. M., Harel, M., Remington, S. J., Silman, I. & Schrag, J. (1992). *Protein Eng.* **5**, 197–211.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Painter, J. & Merritt, E. A. (2006*a*). *Acta Cryst.* D**62**, 439–450.
Painter, J. & Merritt, E. A. (2006*b*). *J. Appl. Cryst.* **39**, 109–111.
Terwilliger, T. C. (2003). *Methods Enzymol.* **374**, 22–37.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.