Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structure of the C-terminal domain of the catalase–peroxidase KatG from *Escherichia coli*

Xavier Carpena,[a] William Melik-Adamyan,[b] Peter C. Loewen[a]* and Ignacio Fita[c]

[a]Department of Microbiology, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada, [b]Institute of Crystallography of Russian Academy of Sciences, 117333 Moscow, Russia, and [c]Consejo Superior de Investigaciones Científicas, Parc Cientific, Josep Samitier 1-5, 08028 Barcelona, Spain

Correspondence e-mail: peter_loewen@umanitoba.ca

Received 19 July 2004
Accepted 19 August 2004

PDB References: KatG, space group $I4_1$, 1u2k, r1u2ksf; space group $P1$, 1u2l, r1u2lsf; space group $P2_12_12_1$, 1u2j, r1u2jsf.

Catalase–peroxidases or KatGs, the apparent *in vivo* activators of the anti-tubercular pro-drug isoniazid, are active as homodimers, each subunit having two distinct but sequence- and structure-related domains. The N-terminal domain contains the haem group and is catalytically active, while the C-terminal domain lacks the cofactor. The C-terminal domain of KatG from *Escherichia coli* is expressed as a soluble protein which has been crystallized in triclinic, orthorhombic and tetragonal crystal forms. Packing in the orthorhombic crystals, with eight molecules in the asymmetric unit, follows the pattern of commensurate modulated structures, which explains the diversity of pseudo-origin peaks observed in the native Patterson map. The different crystal forms arise from variations in the length and sequence of the N-terminal extensions in the different constructs. Despite the variability in the N-terminal region, the overall domain conformations beginning with Pro437 are very similar both to each other and to the C-terminal domains within the native structures of the KatGs from *Haloarcula marismortui* and *Burkholderia pseudomallei*. Some structural reorganization in the C-terminal domain relative to the N-terminal domain has evolved to compensate for the absence of the haem group. A high percentage of the residues in the C-terminal domains of KatG proteins from different sources are highly conserved and these residues are spread uniformly throughout the domain. The easily folded nature and retention of structure in the C-terminal domain suggests that it may serve as a platform for the folding of the N-terminal domain and for stabilization of the molecular dimer.

1. Introduction

Catalase–peroxidases are capable of degrading hydrogen peroxide as either a catalase ($2H_2O_2 \rightarrow 2H_2O + O_2$) or a peroxidase ($2AH_{red} + H_2O_2 \rightarrow 2H_2O + 2A_{ox}$). The removal of H_2O_2 prevents potential damage to cellular components, giving both activities a protective role. Potentially, the peroxidase reaction could have a second role of modifying a substrate to produce a specific organic product, but this is uncertain because the *in vivo* substrate (AH_{red}) is unknown. The first catalase–peroxidase to be isolated was hydroperoxidase I (HPI) from *Escherichia coli* (Claiborne *et al.*, 1979) and this was followed by the cloning (Loewen *et al.*, 1983), genetic mapping (Loewen *et al.*, 1985), characterization (Triggs-Raine & Loewen, 1987) and sequencing (Triggs-Raine *et al.*, 1988) of the encoding gene *katG*. Subsequently, KatG analogues were identified in a variety of bacteria and fungi, with over 58 sequences being available for phylogenetic analysis and comparison (Klotz & Loewen, 2003).

The catalase–peroxidase or KatG family has garnered notoriety through its role as the *in vivo* activator of the widely

Table 1
Data-collection and structural refinement statistics and structure analysis.

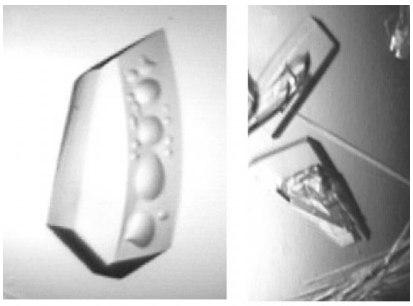
Values in parentheses are for the highest resolution shell.

	Native†	Ta_6Br_{14}†	Native‡	SeMet‡
Data-collection statistics				
Space group	$P2_12_12_1$	$P2_12_12_1$	$I4_1$	$P1$
Unit-cell parameters				
a (Å)	84.2	84.5	59.2	52.9
b (Å)	98.7	101.4	59.2	52.8
c (Å)	302.8	302.8	160.4	56.0
α (°)	90	90	90	93.8
β (°)	90	90	90	90.7
γ (°)	90	90	90	103.3
Resolution (Å)	20–2.1 (2.2–2.1)	20–4.0 (4.1–4.0)	17.3–2.0 (2.1–2.0)	29.0–2.3 (2.4–2.3)
Unique reflections	135504 (17775)	19826 (1123)	17719 (2406)	25540 (3043)
Completeness (%)	91.4 (92.8)	88.0 (74.3)	95.4 (93.5)	97.9 (96.6)
$\langle\sigma F\rangle/\langle F\rangle$ (%)	8.7 (18.5)	6.4 (11.4)	4.1 (15.3)	8.0 (18.3)
Model refinement statistics				
No. reflections	92771		16817	24228
R_{cryst} (%)	23.3		18.9	17.5
R_{free} (%)	32.7		24.9	25.1
Non-H atoms	18136		2233	4720
Water molecules	—		93	211
Average B factor (Å^2)				
Protein	26.2		22.7	26.4
Water	—		31.3	26.5
R.m.s. deviations				
Bonds (Å)	0.012		0.019	0.008
Angles (°)	1.359		1.751	0.942

† EcKatG C-terminal domain with His tag attached. ‡ EcKatG C-terminal domain with His tag removed.

$P2_12_12_1$ mgsshhhhhhssglvprgshmyigpevpkedLIWQD$_{436}$...L$_{726}$
$P1$ ----------gsHMYIGPEVPKEDLIWQD$_{436}$...L$_{726}$
$I4_1$ ----------gshmyigpevpkedliwQD$_{436}$...L$_{726}$

Figure 1
Sequences of EcKatG C-terminal domain constructs found in the different crystal types. The residues shown in lower case are present in the protein but are disordered and not present in the structures.



(*a*) (*b*)

Figure 2
Orthorhombic crystals from the His-tagged EcKatG C-terminal domain Tyr422–Leu726 (*a*). After removal of the His-tag by thrombin proteolysis, two new crystal forms were obtained: tetragonal crystals and thin but long needles (*b*).

used anti-tubercular pro-drug isoniazid. A mutation in *katG* commonly results in *Mycobacterium tuberculosis* being isoniazid-resistant and this is one reason for the worldwide resurgence of tuberculosis (Deretic *et al.*, 1996). Attempts to crystallize *M. tuberculosis* KatG (MtKatG) had been unsuccessful until very recently (Bertrand *et al.*, 2004), but crystals of KatG from three other organisms, *Haloarcula marismortui* (HmKatG; Yamada *et al.*, 2001, 2002), *Burkholderia pseudomallei* (BpKatG; Carpena *et al.*, 2002, 2003) and *Synechococcus* sp. (SyKatG; Wada *et al.*, 2002) were reported in the meantime, providing valuable insights into the structure and function of the enzyme.

Catalase–peroxidases are normally active as homodimers, although homotetrameric associations do occur, as in the enzyme from *E. coli* (EcKatG or HPI). The individual subunits are relatively large at approximately 80 000 Da and have two distinct sequence-related domains, possibly fused following a gene-duplication event (Welinder, 1991). During subsequent evolution, the N-terminal domain retained (or acquired) the ability to bind haem and the C-terminal domain lost (or never acquired) this ability. Despite this functional divergence, considerable structural homology has been conserved, with a root-mean-square deviation (r.m.s.d.) of 2.19 Å for the C$^\alpha$ atoms of 133 residues in the ten α-helical sections that are topologically equivalent in the two domains of BpKatG. The two subunits in the molecule are associated mainly through reciprocal interactions between the N- and C-terminal domains, resulting in there being few N-terminal–N-terminal and no C-terminal–C-terminal domain interactions. A role for the C-terminal domain, as other than an agent for dimer formation and stabilization, has not been defined.

Partial proteolysis of EcKatG produces two fragments of 32 and 34 kDa, both beginning at Tyr422, that contain much of the C-terminal domain, while the N-terminal domain is completely proteolysed. A truncated *katG* gene encoding the C-terminal domain expresses soluble protein which crystallized (Carpena *et al.*, 2002) and in this paper we report its structure determination in three different crystal forms.

2. Materials and methods

2.1. Crystallization and structure determination

A construct from the C-terminal domain of EcKatG, corresponding to the Tyr422–Leu726 region (Fig. 1) plus an N-terminal extension of 21 residues that included a hexa-His tag and a thrombin target, gave plate-like crystals belonging to the orthorhombic $P2_12_12_1$ space group (Carpena *et al.*, 2002) (Fig. 2*a*). These orthorhombic crystals were soaked for 48 h in a 20 mM solution of Ta_6Br_{14} to produce a heavy-atom derivative. Removal of the vector-encoded N-terminal extension by overnight thrombin proteolysis at 293 K, leaving only four non-KatG residues ahead of Tyr422 (Fig. 1), allowed two

further crystal morphologies to be obtained from 34% PEG 4000, 50 m*M* sodium acetate and 0.1 *M* Tris pH 8.5 (Fig. 2*b*). Crystals with a box-like morphology of about 0.15 × 0.15 × 0.05 mm in size were tetragonal, space group $I4_1$, and needle-like crystals of 0.05 × 0.05 × 0.4 mm in size were triclinic, space group *P*1. Native data sets were collected from the orthorhombic and the tetragonal crystal forms using synchrotron beamlines BM14, ID13 and ID14eh4 at the ESRF (Table 1). Data sets were also collected at the ESRF for the tantalum derivative and, at a wavelength above the selenium absorption edge (λ = 0.97860 Å), for triclinic crystals containing selenomethionine. All data sets were indexed, reduced, scaled and merged with *DENZO* and *SCALEPACK* (Otwinowski & Minor, 1996) (Table 1). For the orthorhombic crystal, special care was taken in data collection to avoid overlaps by orientating the longest axis parallel with the spindle axis and using a small rotation range. Originally processed as *P*222, it was later scaled to $P2_12_12_1$ after analysis of the reflections in the pseudo-precession images (Carpena *et al.*, 2002). A portion (5%) of the measured reflections in every data set was reserved for $R_{\rm free}$ monitoring during refinement.

The crystal structures of HmKatG (Yamada *et al.*, 2002) and of BpKatG (Carpena *et al.*, 2003) were reported while this work was in progress, making molecular-replacement solution of the EcKatG C-terminal domain structure possible. Appropriate changes were introduced in the BpKatG C-terminal structure to reflect the sequence of EcKatG and the modified model was used for a molecular-replacement search using *AMoRe* (Navaza, 1994). The initial solution for the tetragonal data set with one molecule in the asymmetric unit gave a map that showed clear continuity from Asp436 to Leu726. Refinement was started with programs in the *CNS* suite (Brünger *et al.*, 1998) and completed with the program *REFMAC* (Murshudov *et al.*, 1997). Solvent molecules were added to the model with the programs *ARP*/*wARP* (Lamzin & Wilson, 1997) and *WATPEAK* (Collaborative Computational Project, Number 4, 1994) and manually with the graphics program *O* (Jones *et al.*, 1991). During the final rounds of refinement bulk-solvent correction was applied and the whole resolution range available was used (Table 1). This refined model was used as the search model to find solutions for the triclinic and orthorhombic data sets with two and eight molecules in their asymmetric units, respectively. Refinement of these crystal forms proceeded similarly to the protocol followed with the tetragonal crystals (Table 1). For the orthorhombic crystals, data from the tantalum derivative were also considered in an attempt to determine the organization of the N-terminal extensions. Tantalum clusters were located by anomalous difference Fourier using the phases available from the molecular-replacement solution.

3. Results and discussion

3.1. Structure of the EcKatG C-terminal domain

The tetragonal crystals of the EcKatG C-terminal domain were generated from protein with only four non-KatG residues ahead of Tyr422 and contained one molecule in the asymmetric unit. The final electron-density map was continuous for main- and side-chain atoms from Gln435 to Leu726, leaving 17 N-terminal residues, Gly418–Trp434, that were not visible and not included in the model. The electron-density maps of the triclinic crystals, also obtained using the short C-terminal domain, showed clear density from His420 to Leu726 for both molecules in the asymmetric unit, with only Gly418 and Ser419 not being traced. The long construct of the C-terminal domain, which contained the complete His-tag extension, produced only orthorhombic crystals for which electron-density maps were continuous from Leu432 to Leu726 for the eight molecules in the asymmetric unit, leaving 21 residues from the His-tag extension and ten from EcKatG

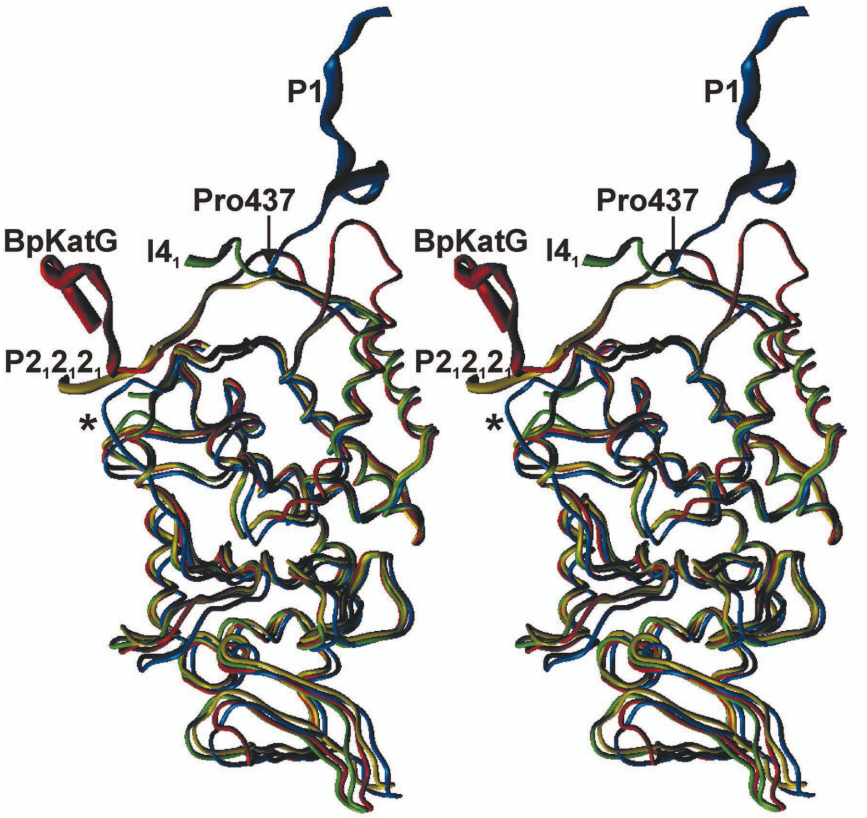


Figure 3
Stereo image of the superimposition of a single subunit from each of the three EcKatG C-terminal domain space groups ($P2_12_12_1$ in gold, *P*1 in blue and $I4_1$ in green) and the C-terminal domain of BpKatG (in red). Pro437, the apparent beginning of the globular domain, is indicated. The asterisk indicates the loop that differs in the *P*1 subunit compared with all the others.

untraced. All residues in the 11 C-terminal domain molecules from the three crystal forms fall within the Ramachandran energetically favourable regions and only Pro496 presents a *cis*-peptide bond.

Despite variations in the N-terminal regions, the C-terminal domain structures were very similar, beginning at Pro437 and extending to Leu726 (Fig. 3). The averaged root-mean-square deviation (r.m.s.d.) for 278 C^{α} atoms is 0.72 Å between the tetragonal and triclinic structures and 0.61 Å between the triclinic and orthorhombic structures. The most significant departures from the averaged C-domain structure occur in the N-terminal regions before residue 437, but other deviations are evident between residues 566 and 570 in the triclinic crystals, around residue 588 in the orthorhombic crystals and in residues 724–726 for the tetragonal crystals. The EcKatG C-terminal domain structures are also very similar to the C-terminal domains of HmKatG (Yamada *et al.*, 2002) and BpKatG (Carpena *et al.*, 2003), which begin at Pro437 (Fig. 3). The most significant deviations between the EcKatG and BpKatG C-terminal domains occur between residues 504 and 509, 520 and 522 (including a six-residue deletion in EcKatG relative to BpKatG) and 657 and 661.

The conformational variability in the N-terminal region between residues 422 and 436 of the various crystals of the C-terminal domain of EcKatG can be rationalized in terms of packing differences that are evident in the different residues involved in crystal contacts in each of the three crystal types (Fig. 4). In the tetragonal crystals the N-terminus is exposed to the solvent, allowing a disordered structure, while in the triclinic crystals the N-terminal region is stabilized by interactions with neighbouring molecules, resulting in an ordered terminus. In fact, the two molecules in the asymmetric unit of the triclinic crystals are related by a quasi-twofold transformation corresponding to a rotation by 179.5° that results in a pseudo-monoclinic packing, space group *C*2, with only one molecule in the asymmetric unit (Fig. 5). This pseudo-symmetry implies similar crystal contacts for the two molecules in the asymmetric unit, explaining the high structural similarity between the two molecules, in particular in the protruding N-terminal region, with small differences existing only in the loop 666–672.

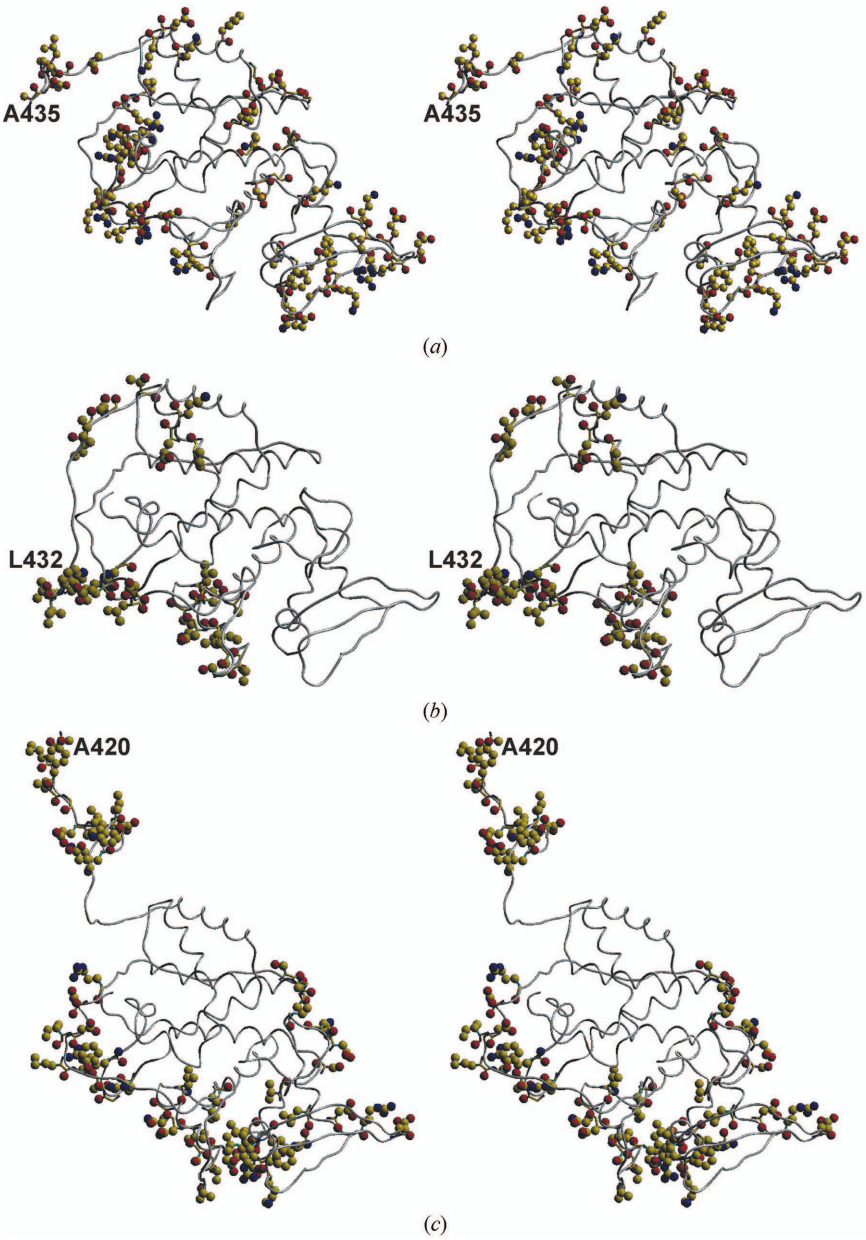


Figure 4
Stereo image of subunits from each of the three crystal packings, with the crystal contact points indicate by spheres. (*a*) $I4_1$, (*b*) $P2_12_12_1$, (*c*) *P*1. The N-terminal residue in each subunit is indicated.

3.2. Packing in the orthorhombic crystals

The orthorhombic crystals of the C-terminal domain, the first crystal form obtained, produced peculiar diffraction patterns that often had overlapping spots along the longest (*c*) crystal axis (Carpena *et al.*, 2002). While no prominent peaks were found in the self-rotation function, the native Patterson map contained three pseudo-origin peaks at *u*, *v*, *w* positions (0.053, 0.5, 0.25), (0.0, 0.0, 0.5) and (0.106, 0.0, 0.5). These features are compatible with three molecules in the asymmetric unit (about 80% solvent content) related by non-crystallographic translations (Carpena *et al.*, 2002). However, the molecular-replacement solution contains eight molecules in the asymmetric unit (44% solvent content) clustered into two groups of four molecules, labelled *A*, *B*, *C*, *D* and *E*, *F*, *G*, *H* (Fig. 6). Transformations relating molecules inside

Table 2
Transformation relating the various subunits, *A*–*H*, in the orthogonal $P2_12_12_1$ crystal packing.

For each transformation, the first line gives the rotation (°), the second the direction cosines and the third the fractional translation.

	A	*B*	*C*	*D*	*E*	*F*	*G*
B	0 0.000 1.000 0.000 0.050 0.487 0.256						
C	0 0.003 0.003 0.999 0.106 0.000 0.503	1 −1.000 0.000 0.000 0.050 0.517 0.246					
D	0 0.001 0.000 −1.000 0.050 0.495 0.246	2 −1.000 0.000 0.000 0.001 0.035 0.501	0 0.000 −0.999 0.008 0.059 0.488 0.245				
E	72 0.253 0.017 0.967 0.208 0.531 −0.018	179 0.566 −0.813 −0.140 0.065 0.843 0.744	111 −0.012 0.175 −0.984 0.434 0.003 0.006	111 −0.016 0.185 −0.983 0.034 0.951 0.288			
F	110 0.001 −0.174 −0.985 0.358 0.468 −0.019	165 0.816 0.579 −0.002 −0.430 0.151 0.749	72 −0.236 −0.002 0.972 0.143 −0.006 0.007	72 −0.249 −0.008 0.969 0.565 0.053 0.286	1 0.000 −0.001 −1.000 0.060 0.002 0.503		
G	72 0.260 0.017 0.965 0.214 0.041 −0.268	179 0.566 −0.812 −0.145 0.084 0.334 0.494	111 −0.012 0.181 −0.983 0.440 0.512 −0.243	111 −0.016 0.191 −0.982 0.045 0.452 0.039	0 0.000 1.000 0.000 0.062 0.491 0.251	1 0.014 −0.003 0.999 0.070 0.504 0.246	
H	110 0.003 −0.170 −0.985 0.371 −0.024 −0.268	165 0.816 0.578 −0.001 −0.415 −0.357 0.499	72 −0.232 −0.005 0.973 0.156 −0.499 −0.243	72 −0.244 −0.011 0.970 0.579 −0.445 0.036	1 −0.012 0.002 −0.999 0.002 0.489 0.249	0 0.000 −1.000 0.000 0.060 0.493 0.250	1 −0.014 0.006 −0.999 0.020 0.024 0.492

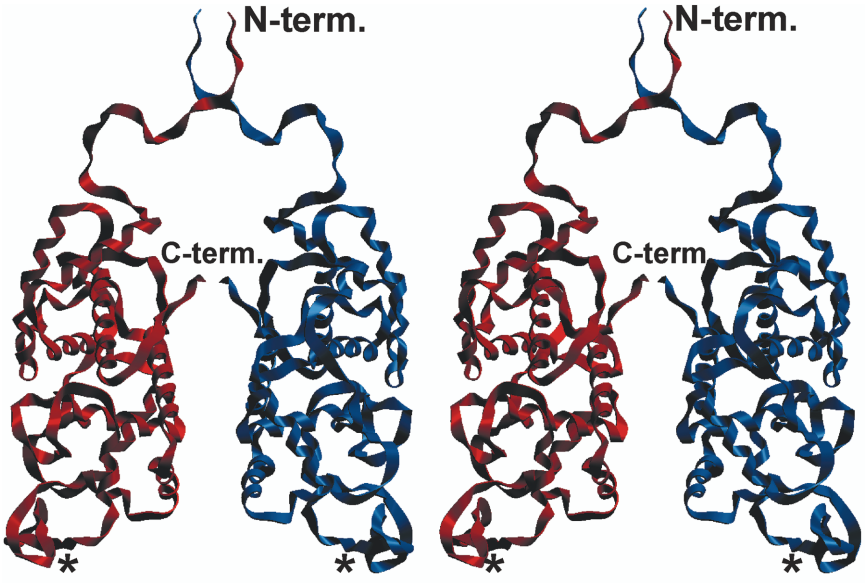


Figure 5
Stereo image of two molecules from the triclinic *P*1 crystal, illustrating the quasi-*C*2 packing for the molecule.

the two groups consist of almost pure non-crystallographic translations, with the rotational components being close to zero (Table 2). In turn, transformations relating molecules between the two different groups consist of rotations around axes parallel to the crystal axis *c* (Table 2).

Having two clusters with the four molecules inside each cluster related by non-crystallographic translations should result in 12 peaks in the native Patterson maps [$2 \times 4 \times (4 - 1)/2$], but only three peaks are experimentally observed. Why? Translations between molecules from the same cluster coincide in the two clusters, which should leave six [$4 \times (4 - 1)/2$] observable peaks in the Patterson maps. However, three of the translations between molecules from the same group coincide because of the disposition of the non-crystallographic translations and the crystal twofold screw axis along *c*, leaving the experimentally observed three peaks (Table 2, Fig. 6). This peculiar packing originates from a displacement of about 0.053 (in fractional coordinates) along cell axis *a* between contiguous molecules situated at around 0.5 along axis *b* and 0.25 along axis *c*. Displacements along axis *a* increase to a maximum value of 0.106 and then decrease, producing a wave appearance for the four molecules along the *c* axis. If displacement along the *a* axis were nil, the unit-cell *c* axis could be halved and the space group would become $C2_122$ (or the standard $C222_1$ after exchanging the *a* and *c* axes) with two molecules,

one from each cluster, in the asymmetric unit. This type of packing organization has not, to our knowledge, been described in protein crystals and would correspond to a commensurate modulated structure (*International Tables for Crystallography*, 1999, Vol. *C*, pp. 797–835). Such packing adds a degree of freedom (displacement along axis *a*), but allows molecules related by translation to retain almost identical packing interactions. A critical factor in this packing organization appears to be the interaction of protruding N-terminal extensions with neighbouring molecules (Fig. 4), a conclusion supported by the appearance of alternative crystal forms only when the N-terminal extension is absent. In the asymmetric unit of the $P2_12_12_1$ crystals there are four regions of unexplained bulky extra density that probably correspond to some of the 31 missing residues from the N-termini of the eight molecules. The presence of this unexplained density contributes to the relatively high agreement factors obtained for this crystal form (Table 2).

3.3. C-domain interactions

KatG structures have a conserved cluster of hydrophobic residues that normally make a major contribution to the intrasubunit interactions between the N- and C-domains, but

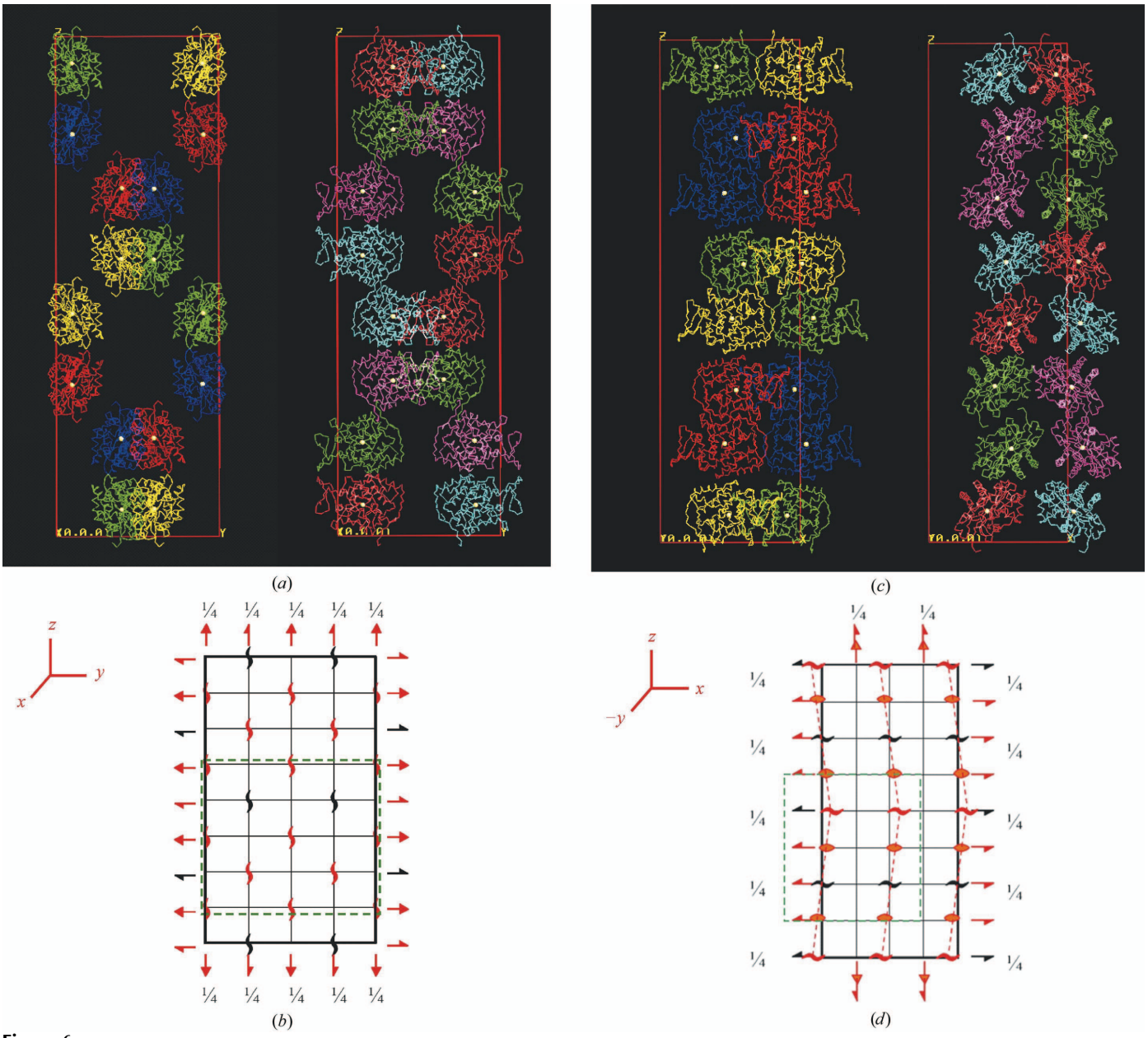


Figure 6
Packing representation of the orthogonal $P2_12_12_1$ form. Molecules *A*–*D* (*A*, yellow; *B*, red; *C*, bright green; *D*, blue) representation in the *zy* section (*a*, left) and in the *zx* section (*c*, left). Molecules *E*–*H* (*E*, orange; *F*, cyan; *G*, pale green; *H*, violet) representation in the *zy* section (*a*, right) and in the *zx* section (*c*, right). Their centre of mass has been drawn as a spherical white dot for better interpretation. Symmetry-operation schemes have been included for both sections (*b* and *d*). Non-crystallographic screw and binary axes are depicted in orange and red. The $P2_12_12_1$ crystallographic operators are shown in black. The $C2_122$ crystal cell is represented as a dashed green box.

which become accessible to the solvent when the N-domain is omitted. A portion of the cluster, including EcKatG residues Met416, Tyr422, Val427, Leu432 and Trp434, is outside the ordered core of the C-domain starting at Pro437 and may prevent the C-terminal domain crystals from reproducing interactions equivalent to the intrasubunit interactions seen between the N- and C-domains in intact KatGs. In addition, there are two hydrophobic patches that participate in intersubunit interactions in the dimer which would be exposed to the solvent in the intact KatG structures. These conserved patches include EcKatG residues Leu595, Met609, Leu646, Leu647, Met649, Trp653, Phe663, Leu682 and Leu690, respectively. A different subset of these residues participate in the packing in each of the different crystal types. For example, residues 653 and 663 are involved in $I4_1$ packing, 595, 609 and 649 are involved in $P2_12_12_1$ packing, and 609, 646, 647, 649, 663 and 690 are involved in $P1$ packing.

It is quite remarkable that two of the three crystal forms obtained with the C-domain constructs present pseudo-symmetry in their crystal packings. A reason for this could be related to the protruding and loosely structured N-terminal region in the constructs. To avoid contacts in the crystal, these flexible regions can be placed fully exposed to the solvent, as in the tetragonal crystals, an organization that presents progressively more stringent packing requirements as the relative size of the region increases. Alternatively, when these flexible regions are involved directly in packing, the conformational variability may be used to preserve quasi-symmetrical interactions with different molecules while increasing the degrees of freedom, as in the triclinic and orthorhombic crystal forms and in some other structures (Guarné *et al.*, 1998).

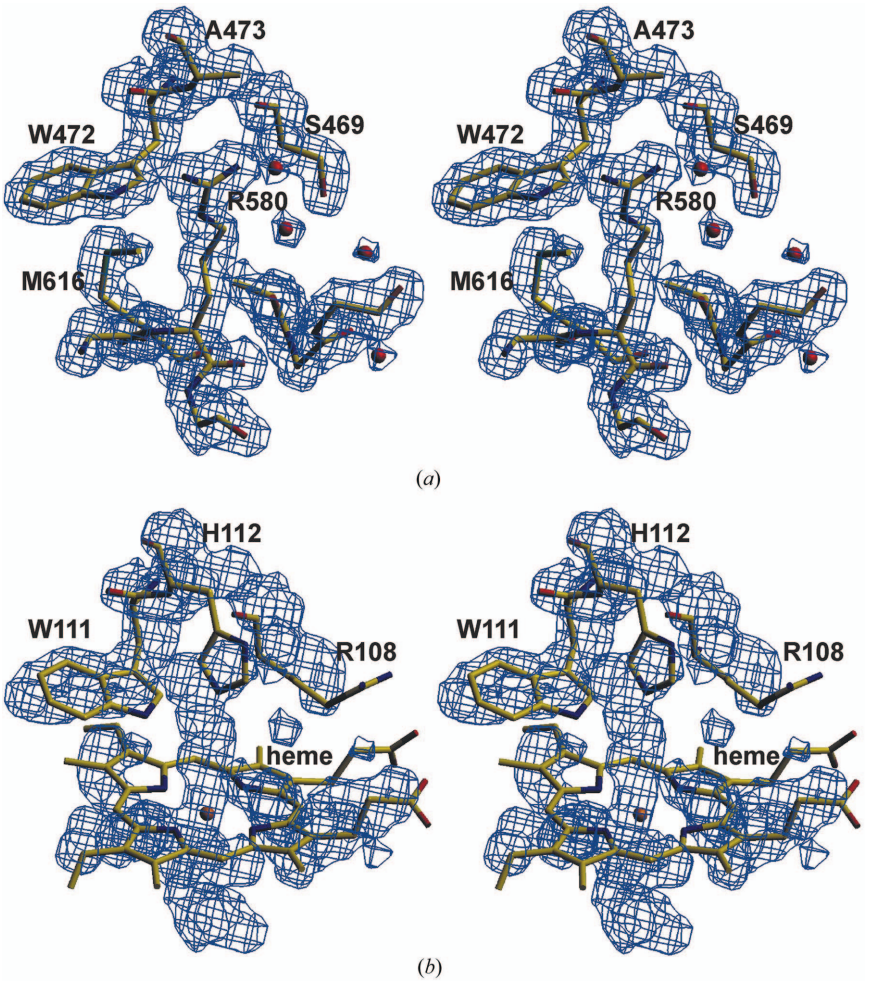


Figure 7
Stereo images showing the organization in the region of EcKatG C-terminal domain that is equivalent to the haem cavity in the N-terminal domain. In (*a*), the electron density in the region has the EcKatG C-domain model as refined superimposed on it. In (*b*), the distal-side catalytic residues and haem from the N-terminal domain of BpKatG are superimposed on the density to illustrate the differences between the two cavities, but also to illustrate how some of the C-terminal residues are situated to compensate for the absence of the haem or the catalytic residues. In comparing the three distal-side catalytic residues, the arginine is replaced by serine (Ser469) and the active-site histidine is replaced by Ala473. Only the active-site tryptophan is retained as Trp472.

3.4. Features of the C-terminal domain

Sequence similarity between the N- and C-terminal domains of catalase–peroxidases has suggested that the two-domain structure arose from the fusion of two adjacent genes following a gene-duplication event (Welinder, 1991). It seems that both domains originally had a haem-binding site that was lost from the C-terminal domain during subsequent evolution. The alternative possibility that neither had a haem-binding site and it evolved in the N-terminal domain is considered to be less likely. Superimposition of the two domains identifies the location in the C-terminal domain where haem may once have been bound and a comparison with the N-terminal region provides insights into differences related to the presence or absence of the haem. Two deletions have occurred in the C-terminal domain, including the equivalent of N-terminal domain residues 203–235 from the sequence between 593 and 596, and the equivalent of residues 283–328 from the sequence between 642 and 652. In the absence of the deleted residues, the intervening and adjacent sequences adopt new paths that extend into the putative haem cavity. Remarkably, the main-chain atoms of the intervening protein segments seem to mimic portions of the haem structure, lying in the plane where the haem would have been bound, including a portion of the main chain that seemingly replaces one of the haem propionates (Fig. 7). Despite retention of the helical structure in the C-terminal domain, the catalytic residues are not conserved except for the equivalent of the active-site tryptophan (Fig. 7).

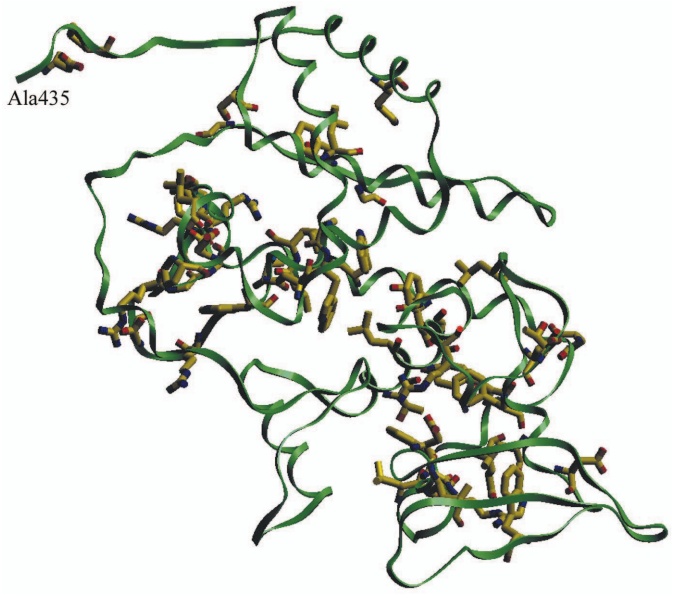


Figure 8
Locations of the residues in the single subunit of the $I4_1$ form that are identical in 95% of 52 KatG sequences. The sequences used are described in Klotz & Loewen (2003).

Determination of the locations of highly conserved residues in KatG proteins reveals an unusually high proportion of conserved residues compared with other proteins and striking differences in their distribution between the N- and C-terminal domains. The proportion of residues conserved at greater than 95% identity in the 53 of the 58 KatG sequences that are complete over more than 650 residues is lower in the C-terminal domain (52 of 322 residues or 18.3%) than in the N-terminal domain (107 of 390 residues or 27.3%), but the proportion of highly conserved residues in both domains is significantly higher than in other proteins. For comparison, the proportion of residues conserved with over 95% identity is 14.2% in 22 Cu-Zn superoxide dismutase sequences, 9.2% in 25 Fe-Mn superoxide dismutase sequences, 8.2% in 24 plant peroxidase sequences, 11.6% in 228 catalase sequences and 12.3% in 28 pyruvate kinase sequences. The highly conserved residues in EcKatG-C are arrayed to span the length of the domain (Fig. 8) from its junction with the N-terminal domain to where it forms an interface with the N-terminal domain of the second subunit. By contrast, the conserved residues in the N-terminal domains are spread at higher density and more generally throughout the core of the domain roughly surrounding the haem pocket.

The C-terminal domain was originally isolated from EcKatG as 32 and 34 kDa fragments following proteolysis (Carpena *et al.*, 2002), a property consistent with a certain amount of inflexibility preventing its access to the protease active site. In addition, while expression of the C-terminal domain from a truncated *E. coli katG* gene produces soluble C-terminal domain (Carpena *et al.*, 2002), expression of a truncated *katG* encoding the N-domain produces only insoluble protein (Baker *et al.*, 2004). Taken together, the enhanced resistance to proteolytic degradation, the relatively facile folding and the almost linear array of conserved residues stretching from one end to the other suggest an important structural role for the C-terminal domain of KatGs. Rapid folding of the C-terminal domain may create the nucleus or platform that facilitates or stabilizes both the folding of the N-terminal domain and the association of monomers into the dimer. Catalase–peroxidases are found mainly in bacteria and a few fungi and the harsher or more varied environment experienced by such microorganisms may provide an explanation for the evolution of the two-domain fusion in which the C-terminal domain is dedicated to stabilizing the N-terminal or catalytic domain. By contrast, the more stable environment of a plant cell may have made the extra protein domain and a dimeric structure unnecessary.

4. Conclusions

The C-terminal domain of the catalase-peroxidase KatG of *E. coli* (HPI) is protease-tolerant and can be expressed as soluble protein from a truncated *katG* gene, whereas the N-terminal domain is expressed as insoluble protein.

Crystals of the C-terminal domain were produced in three different space groups, $P2_12_12_1$, $I4_1$ and $P1$, depending on crystallization conditions and N-terminal sequence.

The crystal packing in the $P2_12_12_1$ space group presents a crystallographic novelty, with three pseudo-origin peaks reflecting a commensurate modulated packing.

The basic subunit structure, involving ten α-helical segments, is conserved in all three space groups of the catalase–peroxidase C-terminal domain, the catalase–peroxidase N-terminal domain and the subunit of plant peroxidases.

The facile folding, protease-resistance and pattern of conserved residues of the C-terminal domain suggests that it may serve as a platform to facilitate and stabilize the folding of the active-site-containing N-terminal domain.

This work was supported by grant BIO2002-04419 from Ministerio de Ciencia y Technologia (MCYT) (to IF), by grant OGP9600 from the Natural Sciences and Engineering Research Council of Canada (NSERC) and by the Canadian Research Chair Program (to PCL) and by grant BIO2002-04419 (to IF) and fellowship EX-2003-0866 (to XC) from the Ministerio de Educación, Cultura y Deporte of Spain.

References

Baker, R. D., Cook, C. O. & Goodwin, D. C. (2004). *Biochem. Biophys. Res. Commun.* **320**, 833–839.
Bertrand, T., Eady, N. A. J., Jones, J. N., Nagy, J. M., Jamart-Grégoire, B., Raven, E. L. & Brown, K. A. (2004). In the press.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Carpena, X., Loprasert, S., Mongkolsuk, S., Switala, J., Loewen, P. C. & Fita, I. (2003). *J. Mol. Biol.* **327**, 475–489.
Carpena, X., Switala, J., Loprasert, S., Mongkolsuk, S., Fita, I. & Loewen, P. C. (2002). *Acta Cryst.* D**58**, 2184–2186.

Claiborne, A., Malinowski, D. P. & Fridovich, I. (1979). *J. Biol. Chem.* **254**, 11664–11668.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Deretic, V., Pagán-Ramos, E., Zhang, Y., Dhandayuthapani, S. & Via, L. E. (1996). *Nature Biotechnol.* **14**, 1557–1561.
Guarné, A., Tormo, J., Kirchweger, K., Pfistermueller, D., Fita, I. & Skern, T. (1998). *EMBO J.* **17**, 7469–7479.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Klotz, M. G. & Loewen, P. C. (2003). *Mol. Biol. Evol.* **20**, 1098–1112.
Lamzin, V. S. & Wilson, K. S. (1997). *Methods Enzymol.* **277**, 269–305.
Loewen, P. C., Triggs, B. L., George, C. S. & Hrabarchuk, B. E. (1985). *J. Bacteriol.* **162**, 661–667.
Loewen, P. C., Triggs, B. L., Klassen, G. R. & Weiner, J. H. (1983). *Can. J. Biochem.* **61**, 1315–1321.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Otwinowski, Z. & Minor, W. (1996). *Methods Enzymol.* **276**, 307–326.
Triggs-Raine, B. L., Doble, B. W., Mulvey, M. R., Sorby, P. A. & Loewen, P. C. (1988). *J. Bacteriol.* **170**, 4415–4419.
Triggs-Raine, B. L. & Loewen, P. C. (1987). *Gene*, **52**, 121–128.
Wada, K., Tada, T., Nakamura, Y., Kinoshita, T., Tamoi, M., Shigeoka, S. & Nishimura, K. (2002). *Acta Cryst.* D**58**, 157–159.
Welinder, K. (1991). *Biochim. Biophys. Acta*, **1080**, 215–220.
Yamada, Y., Fujiwara, T., Sato, T., Igarashi, N. & Tanaka, N. (2002). *Nature Struct. Biol.* **9**, 691–695.
Yamada, Y., Saijo, S., Sato, T., Igarashi, I., Usui, H., Fujiwara, T. & Tanaka, N. (2001). *Acta Cryst.* D**57**, 1157–1158.