Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The structure of the deacetylase domain of *Escherichia coli* PgaB, an enzyme required for biofilm formation: a circularly permuted member of the carbohydrate esterase 4 family

Takashi Nishiyama, Hiroki Noguchi, Hisashi Yoshida, Sam-Yong Park and Jeremy R. H. Tame*

Protein Design Laboratory, Graduate School of Nanobioscience, Yokohama City University, Suehiro 1-7-29, Yokohama, Kanagawa 230-0045, Japan

Correspondence e-mail: jtame@tsurumi.yokohama-cu.ac.jp

Received 5 July 2012
Accepted 8 October 2012

PDB Reference: N-terminal domain of PgaB, 3vus

Bacterial biofilm formation is an extremely widespread phenomenon involving the secretion of a protective exopolysaccharide matrix which helps the bacteria to attach to surfaces and to overcome a variety of stresses in different environments. This matrix may also include proteins, lipids, DNA and metal ions. Its composition depends on the bacterial species and growth conditions, but one of the most widely found components is polymeric β-1,6-*N*-acetyl-D-glucosamine (PGA). Several studies have suggested that PGA is an essential component of biofilm and it is produced by numerous bacteria, including *Escherichia coli*, *Staphylococcus epidermis*, *Yersinia pestis*, *Bordetella* spp. and *Actinobacillus* spp. In *E. coli*, PGA production and export are dependent on four genes that form a single operon, *pgaABCD*, which appears to have been transferred between various species. Biofilms themselves are recognized as environments in which such horizontal gene transfer may occur. The *pga* operon of *E. coli*, which is even found in innocuous laboratory strains, is highly homologous to that from the plague bacterium *Yersinia pestis*, and biofilm is believed to play an important role in the transmission of *Yersinia*. The crystal structure of the N-terminal domain of PgaB, which has deacetylase activity, is described and compared with models of other deacetylases.

1. Introduction

Bacteria may adopt either a free-living ‘planktonic’ or a sessile lifestyle. Sessile bacterial growth of a structured community of cells attached to a surface is known as a biofilm (Hall-Stoodley *et al.*, 2004). Biofilm formation is a general feature of microorganisms, and has been of intense interest for decades owing to its medical relevance; metabolically quiescent bacteria within biofilms are hidden from the immune system and are difficult to treat with antibiotics, and biofilms frequently form on catheters or other surgical implants (Costerton *et al.*, 1999). Polymeric β-1,6-*N*-acetyl-D-glucosamine (PGA) is a principal component of the exopolysaccharide matrix of many bacteria. Without the *pgaABCD* operon, *Escherichia coli* is unable to form biofilms (Wang *et al.*, 2004), and cleavage of PGA breaks up biofilms (Itoh *et al.*, 2005). The enzymes encoded by the operon produce a linear homopolymer of β-1,6-linked *N*-acetylglucosamine. This polymer was first described in studies of *Staphylococcus epidermis* and was referred to as polysaccharide intracellular adhesin (PIA; Mack *et al.*, 1996). As well as being involved in bacterial attachment to biofilm (Agladze *et al.*, 2005), PGA can play important roles in host–microbe interactions. It has been implicated in the colonization and virulence of both Gram-positive and Gram-negative bacteria (Cerca *et al.*, 2007; Vuong, Kocianova *et al.*, 2004; Vuong, Voyich *et al.*, 2004). Biofilm formation by *E. coli*, in particular the role of the *pgaABCD* operon, has been studied in detail by the group of Romeo (Itoh *et al.*, 2008; Wang *et al.*, 2004). Transcription of the *pgaABCD* operon is repressed by CsrA (carbon storage regulatory A), an RNA-binding protein that binds to the untranslated leader of target mRNA (Wang *et al.*, 2005). *pgaABCD* transcription requires NhaR (a LysR-family DNA-binding protein), which switches on PGA production in response to high pH and also to high concentrations of salt (Goller *et al.*, 2006).

PgaC and PgaD are inner membrane proteins, with the latter having no sequence similarity to any known structure; PgaC is a member of the GT-2 glycosyltransferase family. They consume UDP-GlcNAc from the cytoplasm and release PGA into the periplasm. PgaA and PgaB are not directly involved in PGA synthesis, but are required for its export (Itoh *et al.*, 2008). PgaA is believed to form a β-barrel in the outer membrane through which the polymer passes to exit the cell; a homologous structure from *Pseudomonas aeruginosa* has recently been crystallized and shown to form an 18-stranded β-barrel (Whitney *et al.*, 2011). PgaB is anchored to the outer membrane by attachment of lipid; it has a classical signal sequence followed by a cysteine (Cys21), which is the lipid-attachment site. Sequence analysis of PgaB suggests that the protein has an N-terminal deacetylase domain, a member of carbohydrate esterase family 4 (CE4) in the CAZy classification scheme (http://www.cazy.org; Cantarel *et al.*, 2009). Other members of this family act on carbohydrate polymers such as xylan or chitin, but only PgaB has been reported to be involved in PGA synthesis or export. The C-terminal domain of PgaB has an unknown function, but may bind to PgaA or to PGA itself. The deacetylase activity of PgaB appears to be necessary for PGA export, presumably by the exposure of amine groups, which acquire a positive charge (Itoh *et al.*, 2008). NMR analysis of PGA obtained by overexpression of the *pgaABCD* operon showed that the polymer has a high molecular weight (about 400 kDa) but contains fewer than 3% deacetylated residues (Wang *et al.*, 2004). PGA from other species may show a much higher proportion; that from *S. epidermis*, for example, shows 15–20% deacetylation, although the proportion may vary according to the growth conditions (Vuong, Kocianova *et al.*, 2004). Published studies therefore suggest that the deacetylase activity of PgaB is essential for biofilm formation but that it is only weakly active. The growing polymer chain presumably passes close to PgaB in order to be acted upon by it, yet fewer than one residue in 20 undergoes modification. The proportion of residues which must be deacetylated in order to promote secretion appears to be low, but the possibility exists that inhibition of this weak enzyme activity could be an effective route to prevent biofilm formation by *E. coli*. Biofilm formation by the plague bacterium *Yersinia pestis* involves an operon (*hmsHFRS*) that is highly homologous to *E. coli pgaABCD* and appears to play a significant role in bacterial transmission (Bobrov *et al.*, 2008; Darby, 2008). We have therefore solved the crystal structure of the catalytic domain of PgaB (PGABN), the first CE4 model from this organism, and compared it with known structures.

2. Materials and methods

2.1. Cloning

pgaB was amplified by PCR from *E. coli* JM109 chromosomal DNA and cloned into a modified pET28 expression vector using *Bam*HI and *Xho*I restriction sites. The primer sequences were CGGGATCCGCCCAGTCAAGAACATC-ATTTATACCG and CCGAGCCTCGAGTTACTGGAG-GTTTTCGTCATAAAC. The PCR product was digested with *Bam*HI and *Xho*I at 310 K for 2 h before purification using a QIAquick PCR Purification Kit (Qiagen). The purified PCR product was ligated into the cut vector using T4 DNA ligase (Wako) at room temperature for 1 h. The ligation mixture was used to transform *E. coli* DH5α, and pET28b-pgaBN was prepared from cultures using QIAprep (Qiagen). This construct directs expression of residues Gln24–Gln330 of PgaB with a hexahistidine tag at the N-terminus that is cleavable with TEV protease.

2.2. Expression and purification

pET28b-pgaBN was transformed into *E. coli* BL21 (DE3) and cells were grown at 310 K with shaking in 3 l LB medium containing kanamycin (50 µg ml^{-1}). When the OD$_{600}$ of the culture reached 0.5–0.6, PGABN expression was induced by adding IPTG to a final concentration of 0.5 m*M* and growth was continued overnight at 288 K. The cells were collected by centrifugation at 3000*g* at 277 K for 30 min. The pellet was suspended in 50 m*M* Tris–HCl pH 8.0, 0.1 *M* NaCl and was then lysed by sonication on ice. The lysate was centrifuged at 38 000*g* and 277 K for 30 min. The supernatant solution was loaded onto a 10 ml nickel Sepharose column (GE Healthcare) equilibrated with 50 m*M* Tris–HCl pH 8.0, 0.1 *M* NaCl, 10 m*M* imidazole; after washing, it was eluted with 50 m*M* Tris–HCl pH 8.0, 250 m*M* imidazole. The major protein fractions were collected and digested with TEV protease overnight at 277 K during dialysis into 50 m*M* Tris–HCl pH 8.0, 0.1 *M* NaCl. The protease:PgaB ratio was 1:50. The protein was reloaded onto the washed nickel Sepharose column and eluted with 50 m*M* Tris–HCl pH 8.0, 0.1 *M* NaCl. The pooled fractions containing PGABN were dialyzed into 50 m*M* Tris–HCl pH 8.0, 0.1 *M* NaCl before concentration to 30 mg ml^{-1} using Amicon centrifugal filter units (Millipore).

2.3. Crystallization and structure determination

Crystallization experiments were performed at 293 K using the hanging-drop vapour-diffusion method. Crystals grew in 16%(*w*/*v*) PEG 3350, 0.18 *M* sodium acetate, 0.1 *M* bis-Tris–HCl pH 6.5. Data were collected on beamline 17A of the Photon Factory, Tsukuba. The highest resolution data were collected from a crystal which had been soaked briefly in 1 m*M* mercury chloride but which appeared to be native in phasing trials. The data were used in the final refinement and

Table 1
Data-collection and refinement statistics.

Values in parentheses are for the outer shell.

Data set	Native	Hg, remote	Hg, inflection	Hg, peak	Pt
Space group	$P2_1$	$P2_1$			$P2_1$
Wavelength (Å)	1.00000	0.99321	1.00938	0.99957	1.00000
Unit-cell parameters (Å, °)	a = 39.6, b = 53.1, c = 144.2, β = 95.2	a = 39.0, b = 51.9, c = 144.7, β = 95.5			a = 39.3, b = 53.0, c = 144.4, β = 95.5
Resolution range (Å)	50.0–1.65 (1.68–1.65)	50.0–2.50 (2.54–2.50)			50.0–2.30 (2.34–2.30)
Reflections (measured/unique)	337723/63364	87449/19886	88848/19847	86805/19838	106033/25579
Completeness (%)	96.7/83.7	98.0/96.6	98.1/97.1	98.0/97.3	96.0/86.2
$R_{\rm merge}$† (%)	4.9/37.5	4.9/19.1	4.4/13.3	5.4/12.4	7.5/39.8
Multiplicity	4.9	4.4	4.5	4.4	4.2
$\langle I/\sigma(I)\rangle$	49.0	44.3	47.9	48.2	34.9
Refinement statistics					
Resolution range (Å)	25.0–1.65				
R factor/free R factor	0.205/0.257				
R.m.s.d. bond lengths (Å)	0.027				
R.m.s.d. bond angles (°)	2.29				
No. of water molecules	175				
Average B factor (protein/water) (Å^2)	33.4/35.2				
Ramachandran plot, residues in (%)					
Most favourable regions	87.6				
Additional allowed regions	11.1				
Generously allowed regions	0.9				
Disallowed regions	0.4				

† $R_{\rm merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$, where $I_i(hkl)$ is the intensity of an observation, $\langle I(hkl)\rangle$ in the mean value for that reflection and the summations are over all equivalents.

revealed a bound Hg atom with low occupancy on refinement. A total of 250 images of 1° oscillation were collected for each data set. Data processing and scaling were carried out with *HKL*-2000 and *SCALEPACK* (Otwinowski & Minor, 1997). The space group was found to be $P2_1$, with two molecules in the asymmetric unit. Data statistics are given in Table 1. Multiple-wavelength data were collected to 2.5 Å resolution on the same beamline using a crystal soaked in 2 mM mercury chloride for 14 h. A single data set to 2.3 Å resolution was also collected using a crystal soaked in 2 mM K_2PtCl_4 for 24 h. The native data set and the Pt-soak data set were collected using incident radiation of 1.000 Å wavelength. Phases were calculated using *PHENIX* (Adams *et al.*, 2010). Model building was carried out with *ARP*/*wARP* (Langer *et al.*, 2008; Morris *et al.*, 2003) and *Coot* (Emsley *et al.*, 2010; Emsley & Cowtan, 2004). Refinement was carried out with *REFMAC* (Murshudov *et al.*, 2011) and the *CCP*4 suite (Winn *et al.*, 2011). Noncrystallographic symmetry restraints and TLS group refinement were not applied. The Ramachandran plot showed several residues in unusual positions, but the agreement between the two copies of the molecule was very good. Slight main-chain disorder at residues 66 and 100 led to several pairs of equivalent residues in chains *A* and *B* lying at different positions in the Ramachandran plot. Two Ramachandran outliers were the active-site residues His55 and Asp115. Isotropic temperature factors were refined with default restraints. Figures were prepared with *PyMOL* (DeLano, 2002). Water molecules were checked manually. A single Hg atom was modelled into the structure with an occupancy of 20%. Density around the Zn atom of chain *B* showed less than full occupancy of the acetate, and a partially ordered water molecule was also modelled at this site. The final model and structure factors have been deposited in the Protein Data Bank as entry 3vus.

3. Results

3.1. Overall structure

Initial sequence analysis of the *pgaB* gene from *E. coli* suggested that the deacetylase activity resided in an N-terminal domain (PGABN) of roughly 300 residues in length. A threading search of known structures using the *Wurst* server (Torda *et al.*, 2004) suggested a strong match to aldolase from *Trypanosoma brucei* (PDB entry 1f2j), which has 14% sequence identity to the PgaB N-terminus (Chudzik *et al.*, 2000). This TIM β-barrel structure has no enzymatic activity in common with PgaB, but has eight $\alpha\beta$ structure repeats, whereas the CE4 enzymes have seven. PGABN shows sequence similarity to PdaA, an *N*-deacetylase from *Bacillus subtilis*, which is a CE4 member of known structure (Blair & van Aalten, 2004). On the basis of these results, a construct was made to express PgaB from Gln24 to Gln330 by PCR from *E. coli* genomic DNA, omitting the signal peptide and associated cysteine residue. This construct yielded 25 mg purified protein (with the N-terminal histidine tag removed) per litre of culture. Analytical ultracentrifugation showed the protein to exist as a monomer in solution (data not shown).

Crystals were grown that diffracted to almost 1.6 Å resolution and the structure (shown in Fig. 1*a*) was solved using two heavy-atom derivatives and multiple-wavelength anomalous measurements. Two copies of the molecule were found in the asymmetric unit. The ordered residues visible in the electron-density map began at Pro43 and ended at Val308 or

Gln309, so that roughly 20 disordered residues at each end of the polypeptide were not modelled. EXAFS analysis of a native crystal clearly showed the presence of zinc, a metal not purposely added to the protein at any stage of preparation or purification. In each copy of the molecule a single metal ion was readily located and modelled (Fig. 1*b*). A search for related models in the PDB using *DALI* (Holm & Rosenström, 2010) yielded the best match as SlCE4 (PDB entry 2cc0), an acetylxylan esterase from *Streptomyces lividans* involved in plant cell-wall degradation which has a structure similar to PdaA (Taylor *et al.*, 2006). Both enzymes are members of Pfam family PF01522. Overlaying the structures with *SSM* (Krissinel & Henrick, 2004) matched 128 residues of PdaA with PGABN with 18% sequence identity, giving an C$^\alpha$ r.m.s.d. of 2.3 Å. Similar results were obtained matching PGABN to SlCE4, giving an alignment of 138 residues (19.6% identity) and 2.4 Å r.m.s.d. Straightforward sequence searches against the PDB gave much poorer alignments, with the best fits being about 30% identical sequences of roughly 50 residues in length. Although most of these hits were to CE4 enzymes, they also included a subunit of the yeast ribosome and clearly were not all biologically relevant. Examination of the models quickly showed that PGABN is circularly permuted with respect to known CE4 members and that the last β-strand of their TIM-like barrel corresponds to the first strand of PGABN. Simple sequence alignments of PGABN with other members of the CE4 family therefore suggested quite different matches to those found from comparisons of the actual models and did not match conserved active-site residues. A sequence alignment between SlCE4 and PGABN is shown in Fig. 2(*a*), demonstrating the shift of one strand of the central barrel from the N-terminus of PGABN to the C-terminus of other CE4 enzymes. The result of fitting PGABN and SlCE4 by *SSM* is shown in Fig. 2(*b*).

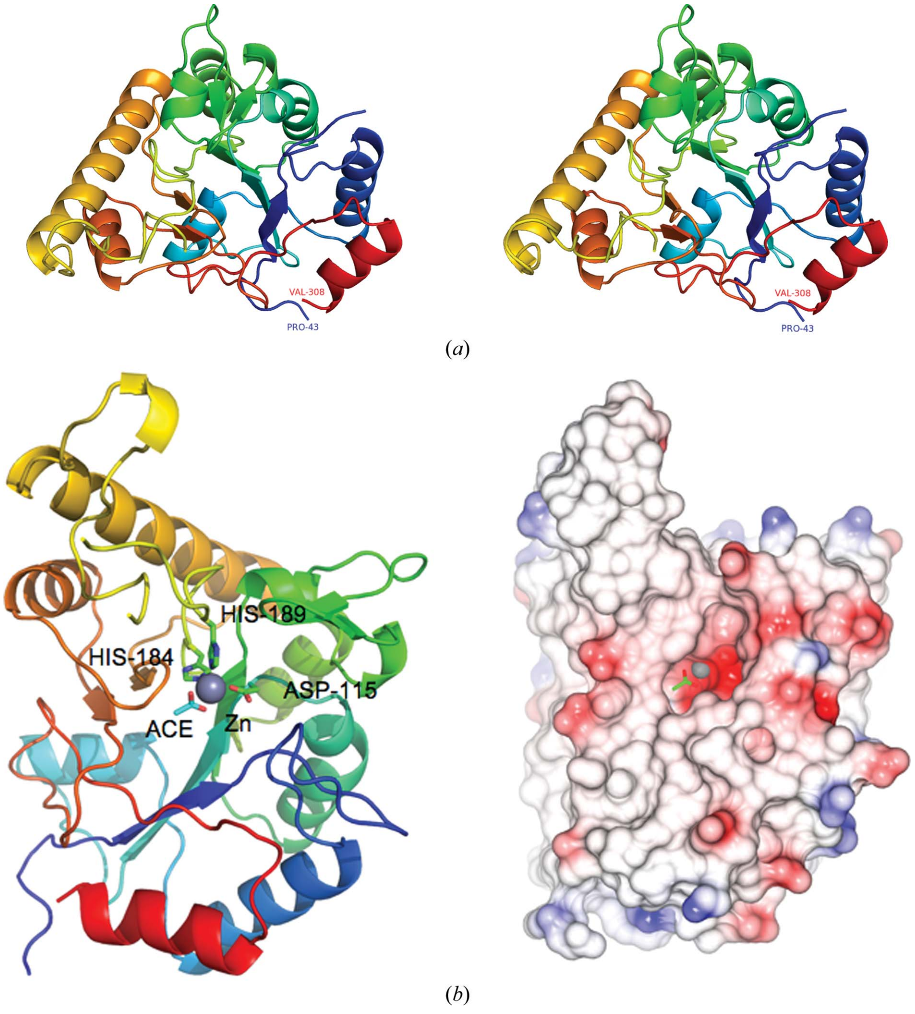
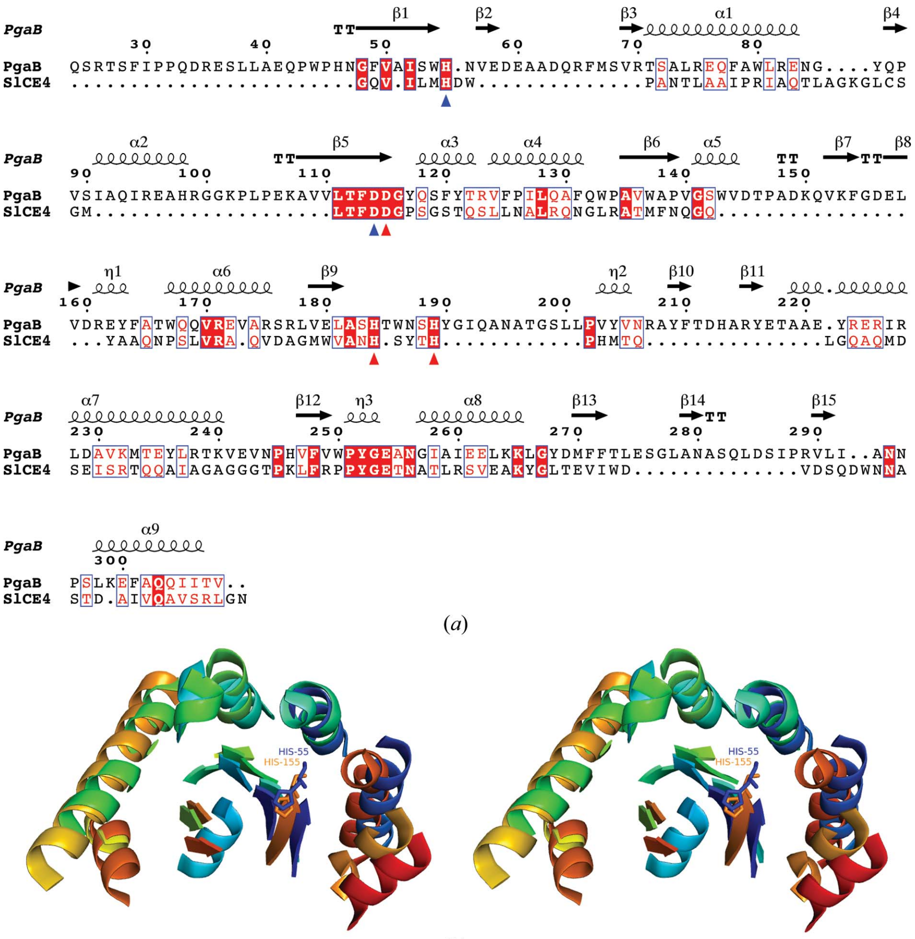


Figure 1
(*a*) A stereoview of the C$^\alpha$ trace of the PgaB N-terminal domain (PGABN), looking into the central barrel. The trace is coloured from blue (Pro43) to red (Val308), with α-helices shown as coils and β-strands as arrows. The figure was drawn using *PyMOL* (DeLano, 2002). Secondary structure was determined automatically. (*b*) The C$^\alpha$ trace of the PGABN monomer, coloured as in (*a*), is shown on the left, with the active-site zinc ion shown as a grey sphere. Residues coordinating the metal are shown as sticks, with O atoms coloured red and N atoms blue. The right-hand panel shows a surface representation of the monomer in the same orientation coloured by electrostatic potential. The zinc ion and acetate ligand were omitted from the calculation of the potential and are shown as ball-and-stick models over the protein surface. The strong negative charge of the active site is apparent. The remainder of the protein surface shows no distinct pattern of charge.

3.2. Active site

The active site of the CE4 family is found at the centre of the barrel-like β-sheet and involves a metal ion in some, but not all, cases. PdaA notably has no metal in the active site on purification and does not bind zinc ions even when these are added; cadmium ions do bind, but only at very high concentrations (Blair & van Aalten, 2004). Other members of the CE4 family have been reported to require cobalt for activity or to prefer cobalt to zinc (Taylor *et al.*, 2006). Nevertheless, two active-site residues (histidine and aspartic acid) were identified from a comparison of the PdaA and PgaB sequences; mutation of either residue in PgaB to alanine (D115A and H184A) blocked activity (Itoh *et al.*, 2008). PGABN shares the common coordination pattern of a zinc ion bound by two histidines and an aspartic acid residue (Fig. 3). The initial crystallization screen and subsequent optimization showed that the inclusion of acetate in the buffer greatly improved the crystal quality, and an acetate ion can be found coordinated to the metal ion of each monomer in the asymmetric unit. However, in one of these sites the occupancy seems to be less than 1.0, and a water molecule with partial occupancy was also modelled close to the zinc ion. Well diffracting


PgaB TT β1 β2 β3 α1 β4
30 40 50 60 70 80
PgaB QSRTSFIPPQDRESLLAEQPWPHNGFVAISWH.NVEDEAADQRFMSVRTSALREQFAWLRENG....YQP
SlCE4 ........................GQV.ILMHDW..............PANTLAAIPRIAQTLAGKGLCS
PgaB α2 TT β5 α3 α4 β6 α5 TT β7 TT β8
90 100 110 120 130 140 150
PgaB VSIAQIREAHRGGKPLPEKAVVLTFDDGYQSFYTRVFPILQAFQWPAVWAPVGSWVDTPADKQVKFGDEL
SlCE4 GM....................LTFDDGPSGSTQSLLNALRQNGLRATMFNQGQ................
PgaB η1 α6 β9 η2 β10 β11
160 170 180 190 200 210 220
PgaB VDREYFATWQQVREVARSRLVELASHTWNSHYGIQANATGSLLPVYVNRAYFTDHARYETAAE.YRERIR
SlCE4 ...YAAQNPSLVRA.QVDAGMWVANH.SYTH............PHMTQ................LGQAQMD
PgaB α7 β12 η3 α8 β13 β14 TT β15
230 240 250 260 270 280 290
PgaB LDAVKMTEYLRTKVEVNPHVFVWPYGEANGIAIEELKKLGYDMFFTLESGLANASQLDSIPRVLI..ANN
SlCE4 SEISRTQQAIAGAGGGTPKLFRPPYGETNATLRSVEAKYGLTEVIWD..............VDSQDWNNA
PgaB α9
300
PgaB PSLKEFAQQIITV..
SlCE4 STD.AIVQAVSRLGN
HIS-55
HIS-155
HIS-55
HIS-155

(*a*)

(*b*)

Figure 2
(*a*) A sequence alignment based on the crystal structures of PGABN and SlCE4 (PDB entry 2cc0). Conserved residues are shown in white on a red background. Red triangles show residues that coordinate the active-site zinc ion and blue triangles show nearby residues involved in catalysis. The residue numbering refers to PGABN. The first β-strand of PGABN (residues 48–55) matches the last strand of SlCE4, the sequence of which has been shifted from the C-terminus to the N-terminus in this alignment. His55 of PGABN matches His155 of SlCE4. The figure was produced by *ESPript* (Gouet *et al.*, 1999). (*b*) Overlay of the C$^{\alpha}$ traces of the PgaB N-domain and SlCE4 (PDB entry 2cc0), showing the helices and strands of both structures, but not coil. Alignment of the models was carried out with *SSM* (Krissinel & Henrick, 2004). Each structure is coloured from blue to red (N-terminus to C-terminus) as in (*a*), but circular permutation of the sequences leads to a marked difference in colouring, despite the close structural similarity. His55 of PGABN lies close to the active site, overlapping His155 of SlCE4, which is a conserved histidine in the CE4 family.

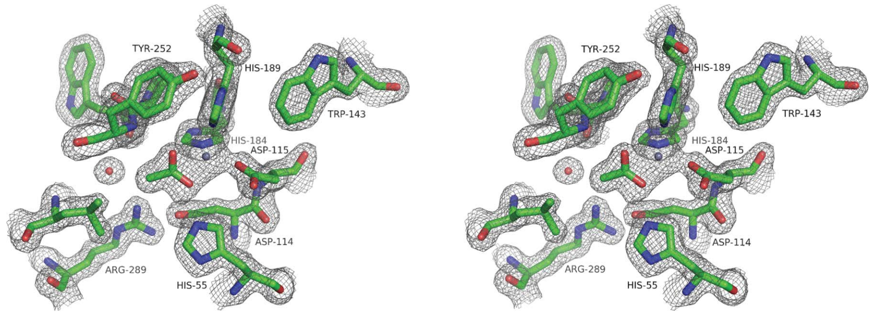


Figure 3
A stereo figure of the $2mF_{\rm o} - DF_{\rm c}$ electron-density map for the refined model covering the active site. Electron density is shown at a level of 1σ. The zinc ion is shown as a grey sphere and water molecules are shown as red spheres. The acetate ion coordinating the zinc ion is clearly visible in the density map and is found to hydrogen-bond to the conserved His55. The conserved LT*X*DDG motif of the NodB domain includes Asp114 and Asp115.

crystals could not be grown in the absence of acetate, suggesting that this ligand stabilized the protein. Attempts to remove acetate by soaking the crystals in acetate-free buffer limited the diffraction resolution.

Comparing PGABN with PdaA, it can be seen that many features of the active site are preserved. 128 C$^{\alpha}$ atoms from the core of the structures show roughly 17% sequence identity, and the zinc-binding residues His184 and His189 of PGABN closely overlap with His124 and His128 of PdaA. Loss of metal binding by PdaA is explained by the fact that it has no third coordinating residue; Asp115 of PGABN is replaced by an asparagine residue (Asn74) that points away from the active site (Fig. 4*a*). Other conserved residues of the active site of PdaA such as Phe98, Trp187 and Leu220 are not found in PGABN, the fold of which is quite different from PdaA in the region of the latter two residues. The active-site residue Asp73 of PdaA is preserved as Asp114 in PGABN, and in both enzymes this aspartic acid forms a salt bridge with an arginine side chain. This arginine residue (Arg163 in PdaA and Arg289 in PGABN) is found on a different β-strand of the barrel; this does not reflect the altered order of the secondary-structure elements but is a consequence of a quite different spatial geometry. Another conserved histidine residue (His55) lies close to the active site at the end of the first β-strand in PGABN, but its equivalent in PdaA (His222) lies at the end of the last β-strand of the barrel. In PdaA this histidine forms a salt bridge (His222–Asp193), but the aspartate has no equivalent in PGABN and His55 is only stabilized by a water molecule. However, His55 does hydrogen-bond to the acetate placed in density near the zinc ion (Fig. 3) and must contact the substrate in the active site. The ion pairs in this region of PdaA have been proposed to play a role in stabilizing charges during catalysis (Blair & van Aalten, 2004), but the details of any such mechanism presumably differ in members of the CE4 family with or without active-site metal ions.

Overlaying SlCE4 with PGABN gives a similar pattern to that with PdaA (Fig. 4*b*). Although SlCE4 is metal-dependent and PdaA has no metal centre, these two enzymes are not circularly permuted relative to one another, and the Asp73–Arg163 pair of PdaA is exactly mirrored by Asp12–Arg100 in SlCE4. His55 of PGABN is preserved as His155 in SlCE4. Thus, although the three enzymes share a common overall three-dimensional structure, each has an active site with characteristics not found in the others. PGABN and SlCE4 presumably share a common mechanism and the altered residues around the active site reflect the substrate preference.

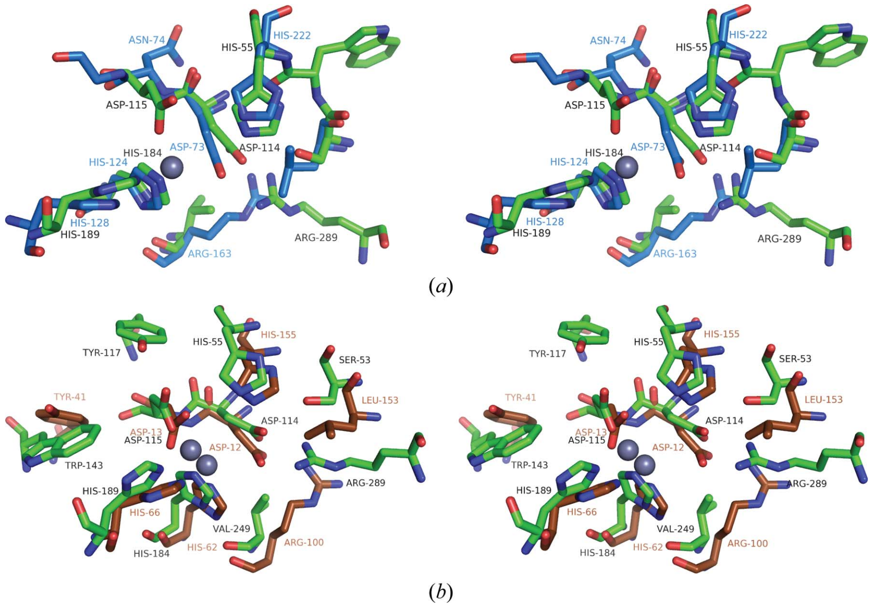


Figure 4
(*a*) A stereo overlay of PGABN (green) and PdaA (light blue) showing the active site. Labels in black indicate the residues in PGABN. O atoms are coloured red and N atoms blue. The models were fitted by overlapping 128 C$^\alpha$ atoms of the conserved core residues. The zinc ion of PGABN is shown as a grey sphere. PdaA (PDB entry 1w17) does not bind metal ions owing to the loss of the coordinating aspartate residue. Instead, Asn74 points away from the active site. Arg163 forms a salt bridge to Asp73, but the equivalent arginine in PGABN is Arg289, which sits on a different β-strand to Arg163 of PdaA. (*b*) Stereo overlay of the active sites of PGABN (green) and SlCE4 (brown). Both structures have a bound zinc ion with very similar coordination geometry. Other residues around the active site show low sequence conservation between the two structures, and Tyr117 of PGABN has no counterpart in SlCE4.

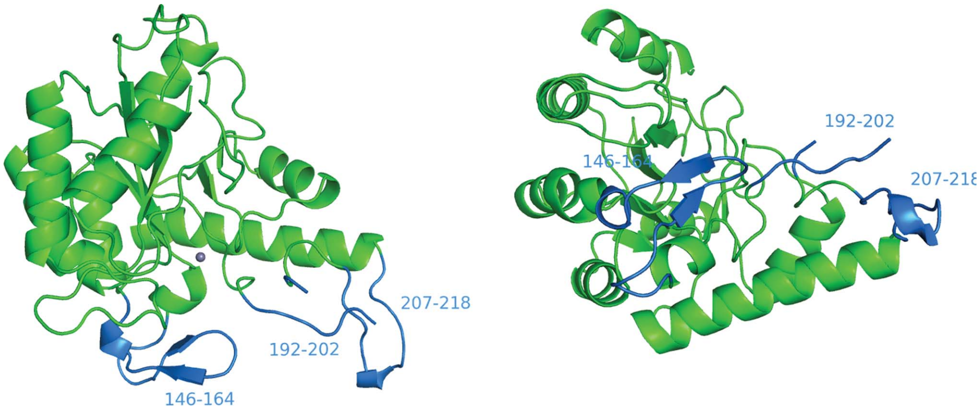


Figure 5
Two views of the C$^\alpha$ trace of PGABN rotated 90° relative to one another about a horizontal axis. The zinc ion is shown as a grey sphere. Surface loops near the active site are shown in blue. These regions are presumably flexible in solution and have no counterpart in SlCE4 or PdaA. They may bind to PgaA in the periplasm. Residues 195–200 are not visible in the electron-density map for the copy of the molecule shown.

3.3. Substrate binding

CE4 enzymes have a highly distorted barrel that lacks one of the $\alpha\beta$ repeats of regular TIM barrels, which creates a groove into which the extended polymer substrates of these enzymes can fit. PdaA has several positively charged residues lining this pocket (such as Lys34, Arg35 and Arg166) that are proposed to be involved in binding to the negatively charged peptidoglycan substrate. PGABN has no equivalent residues and acts upon an uncharged substrate; its enzyme activity is also presumably strongly affected by the insertion of two loops blocking one end of the substrate groove, apparently a feature unique to PGABN among the CE4 enzymes (Fig. 5). Residues Thr146–Phe164 face residues Tyr190–Pro202 on the opposite side of the groove. Neither loop appears to be stably fixed in position and residues 195–200 are not modelled in one copy of the molecule, but the presence of these loops may help to control access to the active site. A further loop (Arg207–Thr218) is more distant from the active site but may impede a long polymeric substrate from easy access to it. Overall, the structured domain of PGABN is rather longer than SlCE4 owing to these and other loops, which decorate its surface. Not all of these loops appear to be mobile; the surface loop from Ser276 to Pro288 appears to be fixed in place by hydrogen bonds formed between the side chain of Gln283 and the main-chain atoms of neighbouring residues.

Repeated attempts to grow well diffracting crystals of PGABN without acetate failed. The reason why acetate, the product of the reaction, strongly promotes crystallization is unclear from the structure, but the enzyme appears to be product-inhibited. As with previous structures from this family, we were therefore obliged to attempt to fit substrate analogues into the binding site by docking (Blair *et al.*, 2005). No clearly preferred binding sites were found which can be confidently predicted to bind single glucosamine residues, but the polymeric substrate may bind by weak interactions at several adjacent sites, just as a protease may require a minimum length of peptide in order to cut a single peptide bond. *In vivo*, it is also possible that the substrate is held in proximity to PgaB by interaction with other proteins in the synthetic machinery.

4. Discussion

A number of structures from the CE4 family have been solved by X-ray crystallography: the first of these was solved by a structural genomics consortium (PDB entry 1ny1; Northeast Structural Genomics Consortium, unpublished work), but no accompanying report has been published to date. The first enzyme from this family to be studied in detail was NodB, which removes the acetyl group from a GlcNAc residue in one step of the synthesis of Nod factors, which are molecular signals involved in the symbiosis of legumes and nitrifying bacteria (John *et al.*, 1993). A NodB homology domain was identified in this sequence (Kafetzopoulos *et al.*, 1993) and has been shown to be present in a variety of esterases with different substrates (Caufrier *et al.*, 2003). This conserved region begins roughly 20 residues upstream of the highly conserved LT*X*DDG motif, where *X* is F or Y (Fig. 2*a*). However, accurate sequence alignment of PGABN without the X-ray model was made difficult by the unique inserts that form loops near the substrate-binding site, and sequence analysis alone was unable to show that PGABN is a circularly permuted member of the CE4 family. The NodB domain forms part of a conserved tertiary structure but with associated secondary-structure elements attached to the N-terminus instead of the C-terminus, an arrangement not previously described. PGABN maintains commonly found features at the active site, including coordination to a catalytic zinc ion and nearby aspartic acid and histidine residues. The observed weak activity of the enzyme presumably arises from obstruction of the polymeric substrate by the unique surface loops of PGABN and/or a low intrinsic affinity of the catalytic domain for PGA. The crystallization of a fragment of PgaB has recently been reported. Little and coworkers attempted to crystallize full-length PgaB and identified a crystallizable fragment by proteolysis (Little, Whitney *et al.*, 2012). Our approach was to express both the N-terminal domain and the full-length protein directly, but so far only the N-terminal domain has yielded crystals.

Circular permutation of proteins has been known for some time and has become a route to artificial proteins of enhanced stability (Yu & Lutz, 2011). Hydrolases form one of the largest groups of naturally occurring circular permutants, although many types of protein are found to have such variants (Lo *et al.*, 2009). PGABN is clearly an offshoot of the CE4 family, as shown by the close similarity of the CE4 structures, and the fact that the N- and C-termini of PGABN are close in space, roughly 10 Å apart, a distance which can easily be bridged by a few amino-acid residues. It seems unlikely that a permuted variant of PGABN with the connectivity of the parent family would have markedly different properties from the wild-type protein. Circular permutation has also been noted to play a role in the evolution of calcium-dependent carbohydrate-binding modules involved in xylan recognition (Montanier *et al.*, 2010).

The very high level of conservation shown between PGABN and the equivalent domain of HmsF from *Yersinia* is unequivocal evidence of gene transfer between species. Codon-usage analysis implies that the gene has been introduced into *E. coli*. Whereas *Yersinia* apparently requires biofilm for its preferred method of infection, blocking the gut of fleas to cause them to expel the bacteria into the bloodstream of an animal, exopolysaccharide clearly plays a very different role in a bacterium found in the soil and the gut of higher animals. The gene used in the work described in this paper involved an attenuated laboratory strain of *E. coli*. The *hms* operon of *Yersinia* is named for haemin storage and is found within the *pgm* (pigmentation) locus, a 102 kb long region associated with colouration and iron uptake. The *hmsHFRS* operon was originally shown to be required for the haemin-storage phenotype, but subsequently it was found that an extra gene *hmsT* was also required which lies far from the *hmsHFRS* operon (outside the *pgm* locus) on the genome of *Y. pestis* (Jones *et al.*, 1999). This extra gene is not required for biofilm formation and is not found in *Y. enterolitica*. *E. coli* strain MG1655 has homologues for all five genes (*ycdSRQPT*), but does not show a *hms*$^{+}$ phenotype. Only *ycdQ* and *ycdP* from *E. coli* MG1655 (equivalent to *pgaC* and *pgaD* in *E. coli* K-12) complement mutations in *hmsR* and *hmsS* in *Y. pestis* (Jones *et al.*, 1999).

Biofilm research is an extremely active and growing field, driven by both medical and biotechnological goals, and the mucoid phenotype of *P. aeruginosa* is a much-studied model (Franklin *et al.*, 2011). This organism causes chronic lung infections in cystic fibrosis patients and produces an exopolysaccharide layer of alginate, a linear polymer of 1,4-linked β-D-mannouronic acid and its C5 epimer α-L-glucuronic acid. The structures of the alginate-export proteins clearly resemble those involved in PGA export in *E. coli* (Keiski *et al.*, 2010; Whitney *et al.*, 2011), raising the possibility that a similar strategy may be employed to tackle biofilm production by both microorganisms. After this paper was reviewed, the structure of PgaB was reported independently by the Howell group (Little, Poloczek *et al.*, 2012).

The authors are extremely grateful to the staff of the Photon Factory for the provision of beamtime in the aftermath of the tsunami disaster of 2011.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.

Agladze, K., Wang, X. & Romeo, T. (2005). *J. Bacteriol.* **187**, 8237–8246.

Blair, D. E., Schüttelkopf, A. W., MacRae, J. I. & van Aalten, D. M. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 15429–15434.

Blair, D. E. & van Aalten, D. M. (2004). *FEBS Lett.* **570**, 13–19.

Bobrov, A. G., Kirillina, O., Forman, S., Mack, D. & Perry, R. D. (2008). *Environ. Microbiol.* **10**, 1419–1432.

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. (2009). *Nucleic Acids Res.* **37**, D233–D238.

Caufrier, F., Martinou, A., Dupont, C. & Bouriotis, V. (2003). *Carbohydr. Res.* **338**, 687–692.

Cerca, N., Maira-Litrán, T., Jefferson, K. K., Grout, M., Goldmann, D. A. & Pier, G. B. (2007). *Proc. Natl Acad. Sci. USA*, **104**, 7528–7533.

Chudzik, D. M., Michels, P. A., de Walque, S. & Hol, W. G. J. (2000). *J. Mol. Biol.* **300**, 697–707.

Costerton, J. W., Stewart, P. S. & Greenberg, E. P. (1999). *Science*, **284**, 1318–1322.

Darby, C. (2008). *Trends Microbiol.* **16**, 158–164.

DeLano, W. L. (2002). *PyMOL*. http://www.pymol.org.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Franklin, M. J., Nivens, D. E., Weadge, J. T. & Howell, P. L. (2011). *Front. Microbiol.* **2**, 167.

Goller, C., Wang, X., Itoh, Y. & Romeo, T. (2006). *J. Bacteriol.* **188**, 8022–8032.

Gouet, P., Courcelle, E., Stuart, D. I. & Métoz, F. (1999). *Bioinformatics*, **15**, 305–308.

Hall-Stoodley, L., Costerton, J. W. & Stoodley, P. (2004). *Nature Rev. Microbiol.* **2**, 95–108.

Holm, L. & Rosenström, P. (2010). *Nucleic Acids Res.* **38**, W545–W549.

Itoh, Y., Rice, J. D., Goller, C., Pannuri, A., Taylor, J., Meisner, J., Beveridge, T. J., Preston, J. F. III & Romeo, T. (2008). *J. Bacteriol.* **190**, 3670–3680.

Itoh, Y., Wang, X., Hinnebusch, B. J., Preston, J. F. III & Romeo, T. (2005). *J. Bacteriol.* **187**, 382–387.

John, M., Röhrig, H., Schmidt, J., Wieneke, U. & Schell, J. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 625–629.

Jones, H. A., Lillard, J. W. & Perry, R. D. (1999). *Microbiology*, **145**, 2117–2128.

Kafetzopoulos, D., Thireos, G., Vournakis, J. N. & Bouriotis, V. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 8005–8008.

Keiski, C. L., Harwich, M., Jain, S., Neculai, A. M., Yip, P., Robinson, H., Whitney, J. C., Riley, L., Burrows, L. L., Ohman, D. E. & Howell, P. L. (2010). *Structure*, **18**, 265–273.

Krissinel, E. & Henrick, K. (2004). *Acta Cryst.* D**60**, 2256–2268.

Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Nature Protoc.* **3**, 1171–1179.

Little, D. J., Poloczek, J., Whitney, J. C., Robinson, H., Nitz, M. & Howell, P. L. (2012). *J. Biol. Chem.* **287**, 31126–31137.

Little, D. J., Whitney, J. C., Robinson, H., Yip, P., Nitz, M. & Howell, P. L. (2012). *Acta Cryst.* F**68**, 842–845.

Lo, W.-C., Lee, C.-C., Lee, C.-Y. & Lyu, P.-C. (2009). *Nucleic Acids Res.* **37**, D328–D332.

Mack, D., Fischer, W., Krokotsch, A., Leopold, K., Hartmann, R., Egge, H. & Laufs, R. (1996). *J. Bacteriol.* **178**, 175–183.

Montanier, C., Flint, J. E., Bolam, D. N., Xie, H., Liu, Z., Rogowski, A., Weiner, D. P., Ratnaparkhe, S., Nurizzo, D., Roberts, S. M., Turkenburg, J. P., Davies, G. J. & Gilbert, H. J. (2010). *J. Biol. Chem.* **285**, 31742–31754.

Morris, R. J., Perrakis, A. & Lamzin, V. S. (2003). *Methods Enzymol.* **374**, 229–244.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

Taylor, E. J., Gloster, T. M., Turkenburg, J. P., Vincent, F., Brzozowski, A. M., Dupont, C., Shareck, F., Centeno, M. S., Prates, J. A., Puchart, V., Ferreira, L. M., Fontes, C. M., Biely, P. & Davies, G. J. (2006). *J. Biol. Chem.* **281**, 10968–10975.

Torda, A. E., Procter, J. B. & Huber, T. (2004). *Nucleic Acids Res.* **32**, W532–W535.

Vuong, C., Kocianova, S., Voyich, J. M., Yao, Y., Fischer, E. R., DeLeo, F. R. & Otto, M. (2004). *J. Biol. Chem.* **279**, 54881–54886.

Vuong, C., Voyich, J. M., Fischer, E. R., Braughton, K. R., Whitney, A. R., DeLeo, F. R. & Otto, M. (2004). *Cell. Microbiol.* **6**, 269–275.

Wang, X., Dubey, A. K., Suzuki, K., Baker, C. S., Babitzke, P. & Romeo, T. (2005). *Mol. Microbiol.* **56**, 1648–1663.

Wang, X., Preston, J. F. III & Romeo, T. (2004). *J. Bacteriol.* **186**, 2724–2734.

Whitney, J. C., Hay, I. D., Li, C., Eckford, P. D., Robinson, H., Amaya, M. F., Wood, L. F., Ohman, D. E., Bear, C. E., Rehm, B. H. & Howell, P. L. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 13083–13088.

Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.

Yu, Y. & Lutz, S. (2011). *Trends Biotechnol.* **29**, 18–25.