Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

A new family of β-helix proteins with similarities to the polysaccharide lyases

Devin W. Close,[a]* Sara D'Angelo[b] and Andrew R. M. Bradbury[a]*

[a]Bioscience Division, Los Alamos National Laboratory, MS888, Los Alamos, NM 87545, USA, and [b]New Mexico Consortium, Los Alamos, NM 87544, USA

Correspondence e-mail: devinclose@gmail.com , amb@lanl.gov

Received 7 May 2014
Accepted 8 July 2014

PDB references: Cthe_2159, gadolinium derivative, 4phb; native form with calcium bound, 4peu

Microorganisms that degrade biomass produce diverse assortments of carbohydrate-active enzymes and binding modules. Despite tremendous advances in the genomic sequencing of these organisms, many genes do not have an ascribed function owing to low sequence identity to genes that have been annotated. Consequently, biochemical and structural characterization of genes with unknown function is required to complement the rapidly growing pool of genomic sequencing data. A protein with previously unknown function (Cthe_2159) was recently isolated in a genome-wide screen using phage display to identify cellulose-binding protein domains from the biomass-degrading bacterium *Clostridium thermocellum*. Here, the crystal structure of Cthe_2159 is presented and it is shown that it is a unique right-handed parallel β-helix protein. Despite very low sequence identity to known β-helix or carbohydrate-active proteins, Cthe_2159 displays structural features that are very similar to those of polysaccharide lyase (PL) families 1, 3, 6 and 9. Cthe_2159 is conserved across bacteria and some archaea and is a member of the domain of unknown function family DUF4353. This suggests that Cthe_2159 is the first representative of a previously unknown family of cellulose and/or acid-sugar binding β-helix proteins that share structural similarities with PLs. Importantly, these results demonstrate how functional annotation by biochemical and structural analysis remains a critical tool in the characterization of new gene products.

1. Introduction

Cellulolytic microorganisms have evolved fascinatingly complex systems for metabolizing organic biomass. Certain members of the bacterial class Clostridia, including the thermophilic anaerobe *Clostridium thermocellum*, are of particular interest owing to their ability to act on many different substrates. These organisms can break down nearly every form of simple and complex plant carbohydrate, including cellulose, hemicellulose and pectin, and are found in almost every ecosystem from soil to the human gut (Demain *et al.*, 2005; Tracy *et al.*, 2012). Microorganisms that have such broad metabolic capability deploy an extensive collection of carbohydrate-active enzymes (CAZymes; Cantarel *et al.*, 2009), carbohydrate-binding modules (CBMs) and other accessory proteins that are either secreted or displayed on extracellular structures such as cellulosomes (Bayer *et al.*, 2004; Demain *et al.*, 2005; Hyeon *et al.*, 2010).

The complexity of CAZymes produced by carbohydrate-active organisms is astounding, with an example being the human gut, where even a relatively small sampling of 177

doi:10.1107/S1399004714015934

different bacterial species was shown to encode >10 000 different CAZyme genes (Kaoutari *et al.*, 2013; Lombard *et al.*, 2014). Despite the number and the diversity of CAZymes, the chemistry used to cleave glycosidic bonds is remarkably well conserved, with the vast majority falling into one of two mechanistic classes (Cantarel *et al.*, 2009; Kaoutari *et al.*, 2013): glycoside hydrolases (GHs), which use a hydrolytic mechanism (Henrissat, 1991), and polysaccharide lyases (PLs), which use nonhydrolytic β-eliminative catalysis (Charnock *et al.*, 2002; Garron & Cygler, 2010; Lombard *et al.*, 2010). Currently, there are 133 families of GHs and 23 families of PLs, which are segregated based on sequence conservation that typically results in conserved catalytic machinery but not necessarily substrate specificity (Henrissat, 1991; Lombard *et al.*, 2010; Kaoutari *et al.*, 2013). PLs are less well characterized than GHs, with most characterization carried out on enzymes isolated from Enterobacteriaceae, especially *Dickeya dadantii* (formerly known as *Erwinia chrysanthemi*; Abbott & Boraston, 2008; Creze *et al.*, 2008). Studies on PLs from other bacteria have been limited, with the best characterized being PLs from *Bacillus* sp. (Akita *et al.*, 2001; Zheng *et al.*, 2012). Relatively little is known about PLs from the class Clostridia, with limited biochemical characterization (Tamaru & Doi, 2001; Hla *et al.*, 2005) and no clostridial PL structures annotated to date.

CAZyme and CBM sequences are often evolutionarily conserved and observed as domains within the context of larger proteins, with each domain contributing a distinct function (*e.g.* substrate binding, membrane anchoring, catalysis *etc.*). Complete genomes of cellulolytic bacteria are continually being drafted (Hemme *et al.*, 2010; Lombard *et al.*, 2014) using bioinformatic and proteomic approaches (Gold & Martin, 2007; Yang *et al.*, 2012), resulting in the annotation of many of the domains involved in biomass metabolism. Despite these advances, the function of many genes remains enigmatic owing to limited experimental characterization (Lombard *et al.*, 2014) and low sequence similarity with protein domains of known function. In fact, a recent survey of the Pfam database (Finn *et al.*, 2013) lists over 3000 domain of unknown function (DUF) families, representing more than 20% of all domains currently in this database (Goodacre *et al.*, 2013). Further, it has been estimated that upwards of 30–49% of genes in public databases that have an ascribed function may be annotated incorrectly (Jones *et al.*, 2007; Bell *et al.*, 2013). Therefore, for organisms such as *C. thermocellum* that dedicate a significant fraction of their genomes to biomass conversion many genes are likely to be erroneously annotated or encode new domain families that cannot currently be characterized by sequence alone.

In an attempt to functionally annotate domains and genes involved in biomass conversion, we recently generated a highly diverse protein domain library from *C. thermocellum*. By means of open reading frame (ORF) filtering, the library consists mostly of functionally folded domains with sufficient diversity to represent the entire *C. thermocellum* genome (D'Angelo *et al.*, 2011). The *C. thermocellum* protein domain library was used to isolate cellulose-binding domains using selection by phage display (data not shown; manuscript in preparation). In doing so, we identified several previously identified CBMs, as well as one gene called Cthe_2159 that does not align with any characterized sequence but which aligns well (*E*-value of 1.2×10^{-72}) with the domain of unknown function 4353 (DUF4353) family in the Pfam database (Finn *et al.*, 2013). Despite being represented by ~568 sequences from ~299 different bacterial and archaeal species, DUF4353 has no known function or structural representatives. This level of conservation, together with our identification of Cthe_2159 as a cellulose-binding domain, implies a protein class or family that is likely to be important for carbohydrate binding or breakdown, warranting further characterization.

To better characterize and potentially annotate the function of Cthe_2159, we determined its structure using X-ray crystallography. The structure was determined using single-wavelength anomalous dispersion (SAD) with gadolinium for phasing on a home-source (Cu *K*α) instrument. We found that the protein folds into a right-handed parallel β-helix structure, a fold common to the carbohydrate esterase (CE), GH and particularly PL enzyme families. Despite very low sequence similarity to any protein from these or other β-helix protein families, the structure reveals functional clues, including Ca^{2+} ions, one of which is coordinated in a highly similar manner to Ca^{2+}-ion coordination within the active site of the PL9 family. We also demonstrate that the protein binds cellulosic and pectic substrates, but were unable to show enzymatic activity. Cthe_2159 is the first structurally characterized representative of a previously unknown family of β-helix proteins that, based on the common structural motif, metal binding and carbohydrate interaction, are similar to PL enzymes.

2. Materials and methods

2.1. Cloning, protein expression and purification

Full-length Cthe_2159 (including the predicted signal sequence) and Cthe_3077 (a gene that encodes a CBM to serve as a positive control in binding assays) were amplified from *C. thermocellum* ATCC 27405 genomic DNA and cloned into a pET-based expression plasmid for expression and purification from *Escherichia coli* BL21(DE3) cells. The oligonucleotide sequences used for cloning the full-length Cthe_2159 gene from the *C. thermocellum* genomic DNA were 5′-TCGAGCGCGCATGCCGTGCAGCCAAGTGGAGTTTC-3′ (Cthe_2159full*bssHII*) and 5′-ATCGGCTAGCTTCCTCCAGCTTACCAAGACAG-3′ (Cthe_2159full*nheI*). A Cthe_3077 construct containing the CBM domain was amplified using oligonucleotides designed based on previous work (Berdichevsky *et al.*, 1999): 5′-CACCGAGCGCGCATGCCGCAAATACACCGGTATCAG-3′ (Cthe_3077*bssHII*) and 5′-CTGTGTGCTAGCTACTACACTGCCACCGGGTTCTTT-3′ (Cthe_3077*nheI*). The Cthe_2159 fragment 124–225 (YP_001038554.1), which was identified through cellulose-binding phage-display screening, was expressed as a recombinant protein. Two point mutations, D199G and E213G, were inadvertently introduced during the Cthe_2159 cloning process. The full-length Cthe_2159 protein was expressed by growing bacteria in autoinduction medium (Studier, 2005) at 37°C for 4 h followed by 20°C for an additional 20–24 h with constant shaking at 250 rev min^{-1}. Bacteria were pelleted and frozen and then stored at −80°C. Protein was prepared by lysis using an Avestin Emusiflex cell homogenizer (Avestin) followed by affinity chromatography using nickel agarose (Qiagen). Protein eluted from the nickel resin was concentrated and further purified on a 320 ml XK 26/60 Sephadex 200 size-exclusion column using an ÄKTAprime liquid-chromatography system (GE Healthcare). The column was pre-equilibrated and run in a buffer consisting of 15 m*M* HEPES pH 7.3, 50 m*M* NaCl. Elution volumes were correlated to approximate molecular weights using extrapolation from size-exclusion standards (Bio-Rad). Fractions from the size-exclusion column were pooled, resulting in protein with >95% purity.

2.2. Carbohydrate-binding assays

Binding assays were performed on the following substrates: regenerated cellulose magnetic beads (RGC; Iontosorp), Avicel PH-101 (Fluka), microcrystalline cellulose (MCC; BCR), xylan (Sigma) and polygalacturonic acid (PGA) sodium salt (Sigma). All substrates were resuspended in water prior to use. Binding assays were performed using filter plates (Millipore) that were pre-blocked with 2% blocking buffer [1%(*w*/*v*) BSA, 1%(*w*/*v*) fish gelatin in HBS] to prevent nonspecific binding to the filter. Binding was performed in a 100 µl solution of 1% blocking buffer with 20 µg of each substrate and 7 µ*M* solutions of either full-length (FL) Cthe_2159, Cthe_2159 124–225, Cthe_3077 (a cellulose-binding positive control) or a nonbinding protein control (a variant of the fluorescent protein eCGP123; Kiss *et al.*, 2009) with continuous shaking for 1 h at room temperature. All proteins were purified as described above and produced in fusion with an SV5 affinity tag. After six washes with HBST (25 m*M* HEPES pH 7.4, 150 m*M* NaCl, 0.1% Tween 20) and HBS (25 m*M* HEPES pH 7.4, 150 m*M* NaCl) under suction using a vacuum filter apparatus, an α-SV5 phycoerythrin (PE)-conjugated monoclonal antibody was added and mixed with continuous shaking for 1 h at room temperature. The washes were repeated and the substrate and bound protein were resuspended in a total of 200 µl HBS and transferred to a black plate. PE fluorescence was recorded using an Infinite M200 plate reader (Tecan) at 540 and 595 nm excitation and emission wavelengths, respectively. Each binding assay value is reported as the mean of 3–4 independent replicates, with error bars showing a standard deviation above and below the mean in the associated figures.

2.3. Crystallization, data collection and processing

Pooled fractions from size-exclusion chromatography were concentrated to ~16 mg ml^{-1} and tested for crystallization using hanging-drop vapour diffusion. Initial hits were obtained in a mixture consisting of 2 µl protein solution and 2 µl well solution, in which the well solution consisted of 20% PEG 8000, 0.1 *M* MES pH 6.0, 0.2 *M* calcium acetate, at room temperature. Single crystals suitable for mounting and diffraction of X-rays were obtained after optimization of the conditions to 19%(*w*/*v*) PEG 8000, 0.1 *M* MES pH 6.0, 0.1 *M* calcium acetate. Data for both native and Gd-soaked crystals were collected from relatively small crystals with approximate dimensions of ~0.25 mm that were not cryoprotected prior to freezing. Gadolinium soaking was accomplished by adding 2 µl of a 0.5 *M* $GdCl_3$ solution to 8 µl well solution and then allowing the crystals to soak at room temperature for 15 min prior to cooling and data collection. Crystals were looped from solution directly into liquid nitrogen and then mounted on the goniometer at 100 K. Data were collected using an in-house system consisting of a MicroMax-007 HF generator (Rigaku), an R-AXIS IV^{++} detector (Rigaku), a liquid-nitrogen cryostream (Oxford Cryosytems) and the *HKL*-3000 software for data collection (Minor *et al.*, 2006). Images from native and Gd-derivative data sets were indexed, refined and integrated using *iMosflm* (Battye *et al.*, 2011) followed by scaling and merging using *SCALA* within the *CCP*4 suite (Winn *et al.*, 2011) with the statistics listed in Table 1. A single molecule and ~46–50% solvent content were expected based on Matthews coefficient analysis.

Table 1
Data-collection and refinement statistics.

Crystallographic statistics were generated by *phenix.table_one* and *phenix.cc_star* (Adams *et al.*, 2010). Values in parentheses are for the highest resolution shell.

	Cthe_2159, native	Cthe_2159, Gd derivative
Wavelength (Å)	1.542	1.542
Resolution range (Å)	26.80–1.80 (1.86–1.80)	34.51–2.18 (2.26–2.18)
Space group	$P2_12_12$	$P2_12_12$
Unit-cell parameters (Å)	a = 70.7, b = 123.4, c = 34.5	a = 68.8, b = 122.8, c = 34.5
Total No. of reflections	55471 (4162)	59475 (5776)
No. of unique reflections	28069 (2197)	15967 (1554)
Multiplicity	2.0 (1.9)	3.7 (3.7)
Completeness (%)	97.2 (77.9)	99.7 (99.6)
$\langle I/\sigma(I)\rangle$	17.2 (5.7)	8.0 (2.8)
Wilson B factor (Å^2)	13.0	22.8
$R_{\rm merge}$ (%)	2.6 (10.0)	13.4 (45.0)
$CC_{1/2}$	0.995 (0.968)	0.984 (0.803)
CC*	0.999 (0.992)	0.996 (0.944)
$R_{\rm work}/R_{\rm free}$ (%)	14.6/17.1	18.4/23.4
No. of non-H atoms		
Total	2275	2070
Protein	1869	1847
Ligands	3	12
Solvent	403	211
No. of protein residues	250	248
R.m.s. deviations		
Bond lengths (Å)	0.006	0.008
Bond angles (°)	1.080	1.080
Ramachandran favoured (%)	94	92
Ramachandran outliers (%)	0.4	0.4
Clashscore	1.61	2.16
Average B factor (Å^2)		
Overall	17.7	27.8
Protein	15.1	27.0
Ligands	20.9	68.6
Solvent	29.3	31.9
PDB code	4peu	4phb

2.4. Structure solution, refinement and analysis

Gd sites were identified and the solution was determined by the SAD phasing method using *phenix.autosol* (Adams *et al.*, 2010). 15 Gd sites were found and used for phasing (*Phaser*), with an initial figure of merit (FOM) of 0.502. Density modification and initial model building using *RESOLVE* resulted in 225 residues being built with an $R_{\rm work}$ and $R_{\rm free}$ of 27.4 and 31.8%, respectively. The Gd-derivative protein model was placed into the native data set for further refinement and model building. Test-set reflections (used in $R_{\rm free}$ calculation) from the Gd data were copied and extended to higher resolution in the native data set. Refinement was performed using *phenix.refine*. The refined models were validated using *MolProbity* (Chen *et al.*, 2010) and final refinement statistics were calculated using *phenix.table_one* (Adams *et al.*, 2010) and are reported in Table 1. Electron-density maps including anomalous difference maps were generated using *phenix.maps* or *CCP*4. Alignment with other structures was performed using *FATCAT* (Ye & Godzik, 2004) or *PDBeFold* (Krissinel & Henrick, 2004). Graphics were produced using *PyMOL* (Schrödinger), *APBS* (Baker *et al.*, 2001) and *Inkscape* (http://www.inkscape.org). Structural data for the native and Gd-derivative data sets were deposited in the PDB with accession codes 4peu and 4phb, respectively.

3. Results and discussion

3.1. Binding to cellulosic substrates

Cthe_2159 encodes a 286-amino-acid protein (30 131 Da) with the first 26 residues predicted to be a signal peptide sequence. A truncated version of the protein consisting of residues 124–225 was originally selected from the filtered *C. thermocellum* genome displayed on phage (manuscript in preparation), leading to the study described here. The full-length protein (Cthe_2159 FL), the truncated version (Cthe_2159 124–225), a positive control (Cthe_3077; a CBM from *C. thermocellum*) and a nonbinding negative-control protein were all tested for binding to regenerated cellulose (RGC), Avicel, microcrystalline cellulose (MCC) and xylan. Fig. 1 shows that both versions of Cthe_2159 bind to cellulosic substrates and xylan at levels above background. Cthe_2159 124–225 bound to RGC at levels comparable to the positive CBM control but does not bind as well to Avicel, RGC or MCC. The full-length protein also bound to the different forms of cellulose, but not as well as the selected domain. The interaction of Cthe_2159 with cellulose substrates and xylan suggests a role in carbohydrate metabolism.

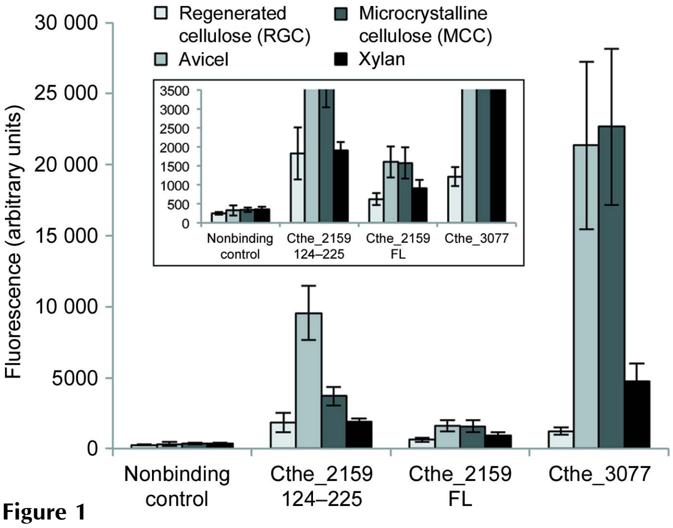

Figure 1
Full-length (FL) Cthe_2159 (residues 1–286) was tested for binding along with a truncated version (124–225) that was isolated in a genome-wide screen based on its ability to bind regenerated cellulose (RGC). Cthe_3077, a conserved cellulose-binding module (CBM), was included as a positive control and an unrelated protein was used as a nonbinding control. Binding was assessed using a filter plate binding ELISA (see §2.2). Error bars represent the standard deviation of three replicates. The inset shows an enlarged view to show differences at lower fluorescence values.

3.2. Crystallization and structure determination using gadolinium

Given its interaction with cellulose but its lack of homology to proteins with known carbohydrate-binding functionality, we determined the structure of Cthe_2159 to better understand its function. Full-length Cthe_2159 protein was used to perform crystallization trials and diffraction-quality crystals were obtained. A 1.8 Å resolution native data set was collected using home-source Cu *Kα* rotating-anode radiation. With no obvious sequence homology to structures in the Protein Data Bank (PDB), molecular replacement was not possible and the minimal number of methionine residues in the sequence posed a problem for selenomethionine (SeMet) substitution and phasing. We therefore soaked crystals with a gadolinium (Gd) salt to obtain a Gd derivative that could be used to directly obtain phase information. Gadolinium was chosen primarily because it has one of the highest f'' signals (~12 e) for any heavy atom using 1.54 Å wavelength Cu *Kα* X-rays (Girard *et al.*, 2003; Molina *et al.*, 2009), enabling anomalous phasing using home-source radiation. A 2.2 Å resolution Gd-derivative data set was collected and single-wavelength anomalous dispersion (SAD) was used to generate high-quality density-modified maps (Figs. 2*a* and 2*b*). The unit-cell parameters varied only slightly between the native and the Gd-derivative crystals (Table 1) and an initial model was built into the Gd-derivative maps and used to refine the structure of Cthe_2159 from both the Gd-derivative and the native data sets.

3.3. Structure and overall fold

The overall structure of Cthe_2159 is that of a right-handed parallel *β*-helix (Fig. 2*c*), a structurally conserved fold found in viruses, bacteria, archaea and eukaryotes (Jenkins *et al.*, 1998; Kajava & Steven, 2006). Continuous electron density was modelled for Cthe_2159 residues 36–285, revealing 29 *β*-strands and no helices. The first five N-terminal strands form two sandwiched antiparallel *β*-sheets, with the rest of the structure (strands *β*6–*β*29) forming eight helical turns of right-handed *β*-helix. Following the standard nomenclature used for

β-helices (Yoder & Jurnak, 1995; Jenkins *et al.*, 1998), each helical turn is comprised of three β-strands (PB1, PB2 and PB3) with intervening turns T1 (between PB1 and PB2), T2 (between PB2 and PB3) and T3 (between PB3 and PB1 of the following turn) (Fig. 2*d*). Several of the turns form an extended sequence between strands, including smaller loops in T1 between β6 and β7 (T1$_{6–7}$) and β21 and β22 (T1$_{21–22}$), with the T3 turn between β11 and β12 (T3$_{11–12}$) being the largest intervening loop structure (Figs. 2*c*, 2*d* and 2*e*). Looking down the axis of the β-helix (Fig. 2*e*), the helical turns are superimposed in a heart shape, with a continuous external cleft formed at the junction between T3 and PB1. This view also reveals that the interior of the β-helix consists almost entirely of stacked isoleucine, valine, glycine, leucine and alanine side chains and is devoid of solvent. The β-helix fold is found in some GH and CE enzymes and is widely observed in PL families, which in combination with the observed cellulose and xylan binding discussed above supports a role for Cthe_2159 in carbohydrate binding or cleavage.

3.4. Calcium binding

A defining feature of most β-helix PL enzymes *versus* GHs and other hydrolytic enzymes is the utilization of metal-assisted β-elimination, where a metal, frequently Ca^{2+}, serves to neutralize acidic groups on the substrate prior to proton abstraction by an adjacent Brønsted base and ultimately glycosidic bond cleavage (Garron & Cygler, 2010; Seyedarabi *et al.*, 2010). In addition, calcium ions can also mediate interactions of sugars with CBMs and even stabilize protein folds (Boraston *et al.*, 2004; Ghosh *et al.*, 2013). The structure of Cthe_2159, which was crystallized in the presence of calcium, reveals three well ordered Ca^{2+} ions (Ca1, Ca2 and Ca3; Figs. 2 and 3). Assignment of Ca^{2+} in the native data set was established using an anomalous difference map (Ca^{2+} f'' = 1.3 e at 1.54 Å) that showed obvious peaks for calcium at levels comparable to, if not larger than, those of cysteine and methionine S atoms (Fig. 2*c*; S f'' = 0.56 e at 1.54 Å). Ca1 is extensively coordinated with octahedral geometry by four aspartate residues (Asp215, Asp243, Asp244 and Asp247; Fig. 3*b*) and two waters in the cleft formed between T3 and PB1 of helical turns 8 and 9 at the C-terminal end of the protein and the β-helix structure (Figs. 2 and 3). Ca2 and Ca3 are coordinated between T1$_{6–7}$ and T1$_{21–22}$ (Figs. 3*a* and 3*c*) within an obvious cleft that is contiguous with the Ca1 site. In the Gd-derivative data set, well ordered Gd^{3+} ions are observed in equivalent positions to Ca1, Ca2 and Ca3, implying high-affinity but not entirely specific metal-binding sites. Precedents for lanthanides occupying Ca^{2+}-binding sites

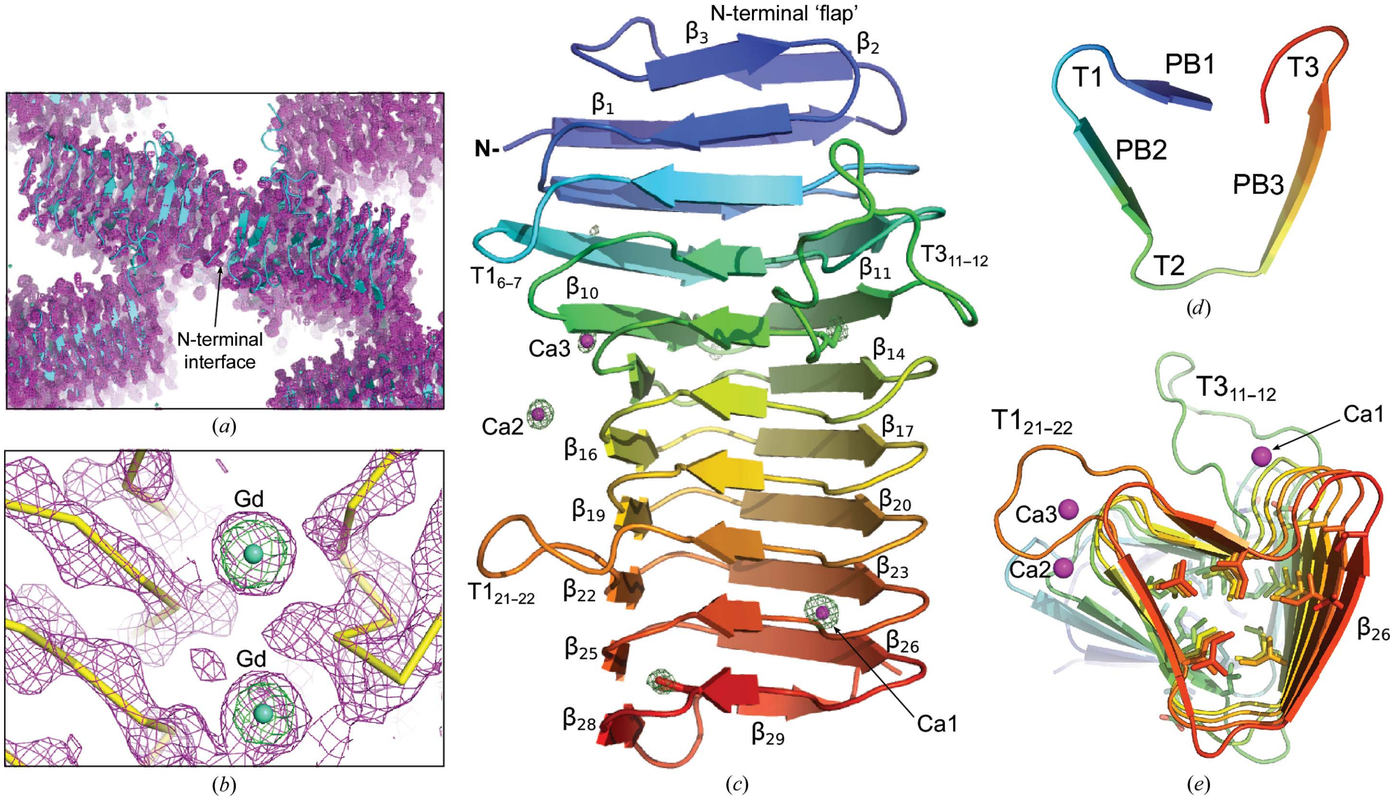

Figure 2
The crystal structure of Cthe_2159. (*a*, *b*) The structure was determined using gadolinium (Gd) soaking and SAD phasing with the experimental density-modified *RESOLVE* map shown in magenta (1.5σ); anomalous difference map peaks (5.0σ) are shown in green for two of the Gd sites in the Gd-derivative data set. (*c*) Cartoon depiction of the native Cthe_2159 structure from the N-terminus (blue; residue 36) to the C-terminus (red; residue 285) with secondary structure and calcium ions (Ca1, Ca2 and Ca3) indicated as magenta spheres. An anomalous difference map (green mesh; 5.0σ) was used to verify the identity of calcium. Other observed peaks in the anomalous difference map are owing to cysteine and methionine S atoms. (*d*) A single turn of β-helix is shown with standard nomenclature indicated. (*e*) A view down the axis of the β-helix showing the highly superimposable stacking of Ile, Val and Leu side chains within the interior of the structure.

within catalytic centres exist, including substitution in an *E. chrysanthemi* PL PelC structure (Colman *et al.*, 1972; Yoder & Jurnak, 1995).

3.5. Insights into carbohydrate binding

In the original phage-display experiments, two fragments of Cthe_2159 comprising residues 124–225 and 172–275 were selected for their ability to bind cellulose, suggesting that residues important for cellulose interactions are found in the common region within residues 172–225 (shown in green in Fig. 3*a*). These residues coincide with a region of the protein structure adjacent to Ca1 and are enriched in lysine and asparagine side chains. This side-chain composition results in an obvious pocket (colored blue), which we term the K/N pocket, with high positive surface potential (Fig. 3*d*). This pocket resembles a common feature of CBMs and PLs, where polar side chains, namely asparagine and lysine, form electrostatic and hydrogen-bonding interactions with negatively charged substrate (Abbott & Boraston, 2008). The region of the protein immediately adjacent to the K/N pocket and around $T3_{11-12}$ may also be important as it is highly enriched in solvent-exposed aromatic phenylalanine residues (Fig. 3*c*). Solvent-exposed aromatics such as phenylalanine and tyrosine are common constituents of substrate-binding clefts in both CBMs and CAZymes (Boraston *et al.*, 2004; Abbott & Boraston, 2008).

Cthe_2159 124–225 consists of both the K/N pocket and the phenylalanine-enriched region and binds cellulosic substrates (Fig. 1). Since many β-helix proteins are PLs and bind or break down pectic substrates, we tested binding to polygalacturonic acid (PGA), a model substrate for pectate lyases. As shown in Fig. 3(*e*), Cthe_2159 124–225 binds very robustly to PGA, with an approximately tenfold tighter binding relative to RGC or MCC (see Fig. 1). In our assay, the level of binding of Cthe_2159 124–225 to PGA is comparable to the binding of Cthe_3077 to MCC (Fig. 1), inferring that the Cthe_2159–PGA interaction is biologically important. Full-length Cthe_2159 binds to PGA, but not to the same extent as the truncated version.

Based on these results, we tested Cthe_2159 for activity against a variety of pectic substrates, including pectin and PGA from different sources and at various levels of purity from 75 to 95%. We used the standard activity assay in which the accumulation of 4,5-unsaturated product is monitored based on the increase in absorbance at 232 nm (Collmer *et al.*, 1988; Herron *et al.*, 2003; Seyedarabi *et al.*, 2010; Hassan *et al.*,

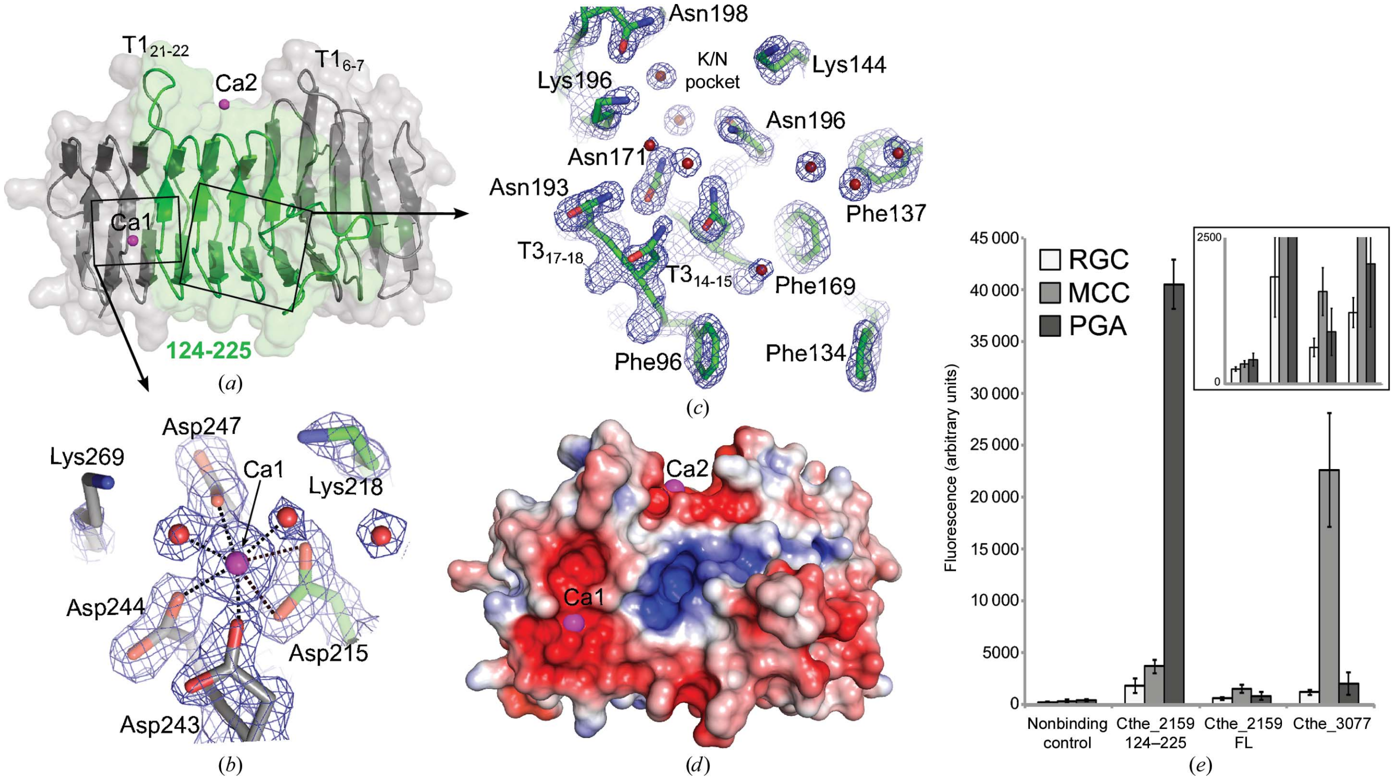

Figure 3
(*a*) Cartoon depiction of the structure with residues 124–225 shown in green (see text). Ca1 and Ca2 are shown for reference. (*b*) Enlarged view of the Ca1 active site showing extensive coordination of the Ca1 ion and potentially important lysine side chains. Well ordered water molecules in the region are shown as red spheres. The $2mF_o - DF_c$ map contoured at 2.0σ is shown as a blue mesh. Dashed lines indicate distances of between 2.35 and 2.5 Å. (*c*) Enlarged view of the K/N and phenylalanine-rich binding regions. (*d*) Electrostatic surface potential calculated using *APBS* [$-4kT$/e (red) to $+4kT$/e (blue); Baker *et al.*, 2001] in the same orientation as in (*a*). An obvious positively charged pocket (the K/N pocket) is observed immediately adjacent to the Ca1 binding site. (*e*) Comparative binding of different proteins and constructs to polygalacturonic acid (PGA), regenerated cellulose (RGC) and microcrystalline cellulose (MCC) in filter plate binding ELISA (see §2.2), with error bars representing the standard deviation of three replicate experiments. PGA is a model substrate for pectate lyases. An enlarged view scaled to lower fluorescence is shown in the box. Cthe_2159 124–225 consisting of the K/N pocket and phenylalanine-enriched region binds to PGA at levels 100-fold above the nonbinding control.

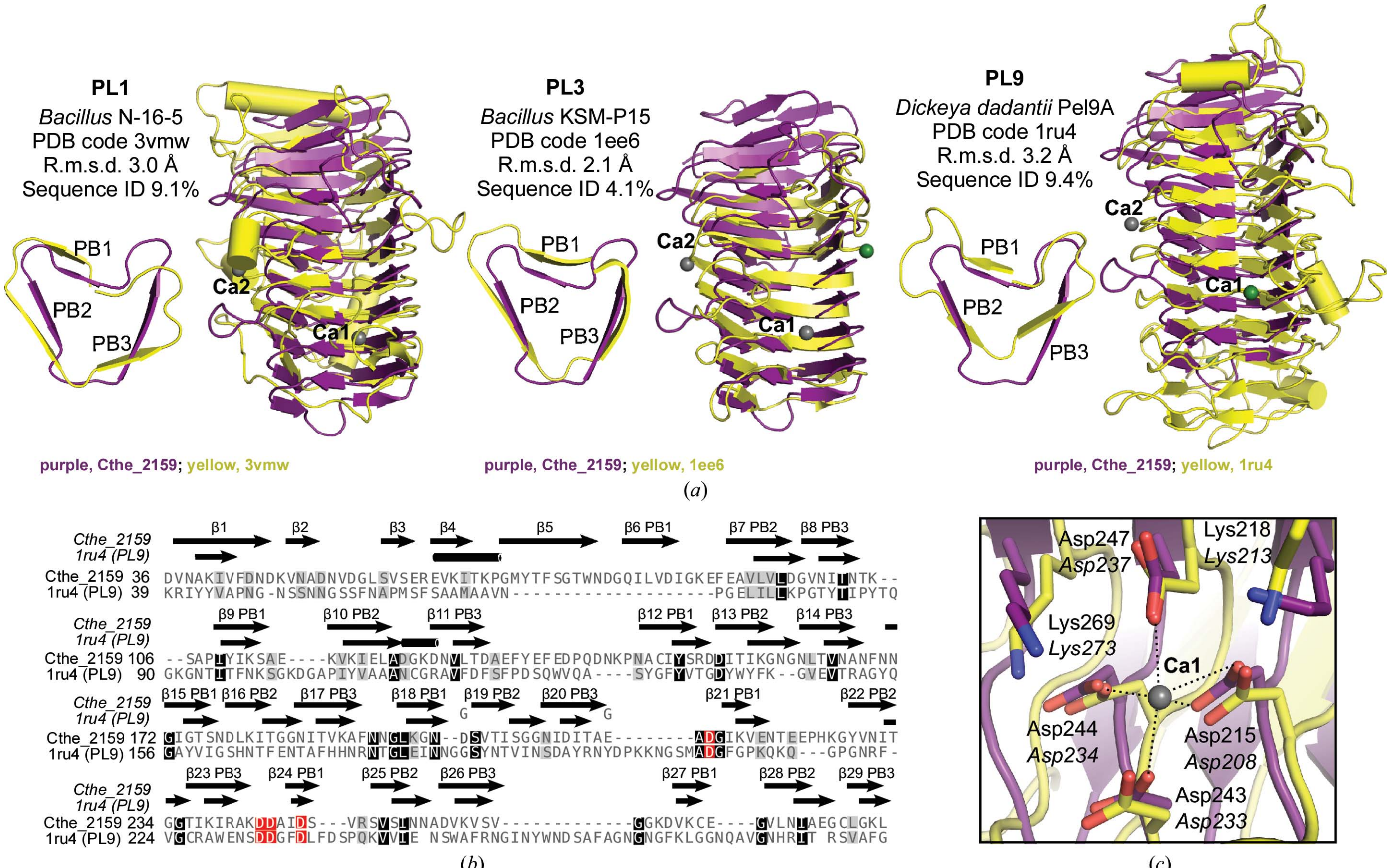
PL1
Bacillus N-16-5
PDB code 3vmw
R.m.s.d. 3.0 Å
Sequence ID 9.1%
PB1
PB2
PB3
Ca2
Ca1
purple, Cthe_2159; yellow, 3vmw
PL3
Bacillus KSM-P15
PDB code 1ee6
R.m.s.d. 2.1 Å
Sequence ID 4.1%
PB1
PB2
PB3
Ca2
Ca1
purple, Cthe_2159; yellow, 1ee6
PL9
Dickeya dadantii Pel9A
PDB code 1ru4
R.m.s.d. 3.2 Å
Sequence ID 9.4%
PB1
PB2
PB3
Ca2
Ca1
purple, Cthe_2159; yellow, 1ru4
(*a*)
β1 β2 β3 β4 β5 β6 PB1 β7 PB2 β8 PB3
Cthe_2159
1ru4 (PL9)
Cthe_2159 36 DVNAKIVFDNDKVNADNVDGLSVSEREVKITKPGMYTFSGTWNDGQILVDIGKEFEAVLVLDGVNIINTK--
1ru4 (PL9) 39 KRIYYVAPNG-NSSNNGSSFNAPMSFSAAMAAVN-------------------PGELILLKPGTYTIPYTQ
β9 PB1 β10 PB2 β11 PB3 β12 PB1 β13 PB2 β14 PB3
Cthe_2159 106 --SAPIYIKSAE----KVKIELADGKDNVLTDAEFYEFEDPQDNKPNACIYSRDDITIKGNGNLTVNANFNN
1ru4 (PL9) 90 GKGNTITFNKSGKDGAPIYVAAANCGRAVFDFSFPDSQWVQA----SYGFYVTGDYWYFK--GVEVTRAGYQ
β15 PB1 β16 PB2 β17 PB3 β18 PB1 β19 PB2 β20 PB3 β21 PB1 β22 PB2
G
G
Cthe_2159 172 GIGTSNDLKITGGNITVKAFNNGLKGN--DSVTISGGNIDITAE--------ADGIKVENTEEPHKGYVNIT
1ru4 (PL9) 156 GAYVIGSHNTFENTAFHHNRNTGLEINNGGSYNTVINSDAYRNYDPKKNGSMADGFGPKQKQ---GPGNRF-
β23 PB3 β24 PB1 β25 PB2 β26 PB3 β27 PB1 β28 PB2 β29 PB3
Cthe_2159 234 GGTIKIRAKDDAIDS---VRSVSINNADVKVSV-------------GGKDVKCE----GVLNIAEGCLGKL
1ru4 (PL9) 224 VGCRAWENSDDGFDLFDSPQKVVIE NSWAFRNGINYWNDSAFAGNGNGFKLGGNQAVGNHRIT RSVAFG
(*b*)
Asp247
Asp237
Lys218
Lys213
Lys269
Lys273
Ca1
Asp244
Asp234
Asp215
Asp208
Asp243
Asp233
(*c*)

Figure 4
Comparing Cthe_2159 with β-helix fold polysaccharide lysase (PL) proteins. (*a*) Three-dimensional structure alignments of Cthe_2159 with representative structures from three PL families (PL1, PL3 and PL9) performed using the *FATCAT* alignment program (Ye & Godzik, 2004). A single representative turn from each alignment is shown as viewed down the axis of the β-helix. A second view of the entire aligned proteins is shown perpendicular to the axis of the β-helix. The overall r.m.s.d. for C^{α} atoms and sequence identity for equivalent residues is indicated. Calcium observed in Cthe_2159 is indicated as grey spheres, and green spheres represent the positions of calcium in the aligned PL structures. (*b*) Structure-based alignment of Cthe_2159 with *D. dadantii* Pel9A showing alignment of secondary-structure elements (arrows, β-strands; cylinders, α-helices) and amino acids at equivalent positions in the respective structures. The D199G and E213G mutations found in the protein used to determine the crystal structure are indicated with a G above the alignment. The alignment is colored using a BLOSUM62 scoring matrix. Red boxes are side chains that coordinate the primary calcium ions in the respective structures. (*c*) Despite very low identity for the rest of the protein, residues coordinating calcium in the primary calcium-binding site (Ca1) for Cthe_2159 (bold) and Pel9A (italicized; PDB entry 1ru4) and potential catalytic lysine side chains (Lys273 for Pel9A; Jenkins *et al.*, 2004) are nearly superimposable.

2013). We tested activity at varying pHs from 7.3 to 9.5, at temperatures ranging from 20 to 60°C and with various metals, including Ca^{2+}, Mn^{2+}, Mg^{2+}, Fe^{3+}, Ni^{2+} and Co^{2+}, but were unable to observe an obvious increase in absorbance at 232 nm under any of these conditions. Since the protein shows binding to cellulosic substrates, we also tested activity against the model cellulose substrates 4-methylumbelliferyl-β-D-cellobiose (4MUL; Chernoglazov *et al.*, 1989; Boschker & Cappenberg, 1994) and carboxymethylcellulose (CMC; Miller, 1959; Wood & Bhat, 1988) but did not observe activity against these either. Altogether, the structure and binding data indicate that Cthe_2159 is similar to PL enzymes, but a specific substrate and reaction conditions have not been identified and therefore Cthe_2159 cannot yet be classified as an enzyme.

3.6. Cthe_2159 is the first representative of a new β-helix family resembling polysaccharide lyases

The β-helix fold is evolutionarily conserved and ubiquitous, and although it is utilized as a structural fold it is most frequently observed in enzymatic domains (Jenkins *et al.*, 1998). The GH, CE and PL classes of CAZYymes in particular utilize the fold, but the presence of a Ca^{2+} catalytic centre is distinctive of PLs, suggesting that Cthe_2159 is most similar to the PL enzyme families PL1, PL3, PL6 and PL9. Despite overall fold (three-dimensional structural alignments of Cthe_2159 with three different PLs are shown in Fig. 4*a*) and Ca^{2+}-binding site similarity (Fig. 4*c*), Cthe_2159 shows very low sequence homology to known PLs or to any other β-helix proteins. When performing structure-based alignments against the PDB, the 20 protein structures that align most closely with Cthe_2159 are almost exclusively PL, GH or CE enzymes and have an average C^{α} r.m.s.d. of ~3.1 Å but an average sequence identity of only ~7% (Supplementary Table S1[1]). Interestingly, while Ca^{2+} has been observed in certain CBM inter-

[1] Supporting information has been deposited in the IUCr electronic archive (Reference: DW5104).

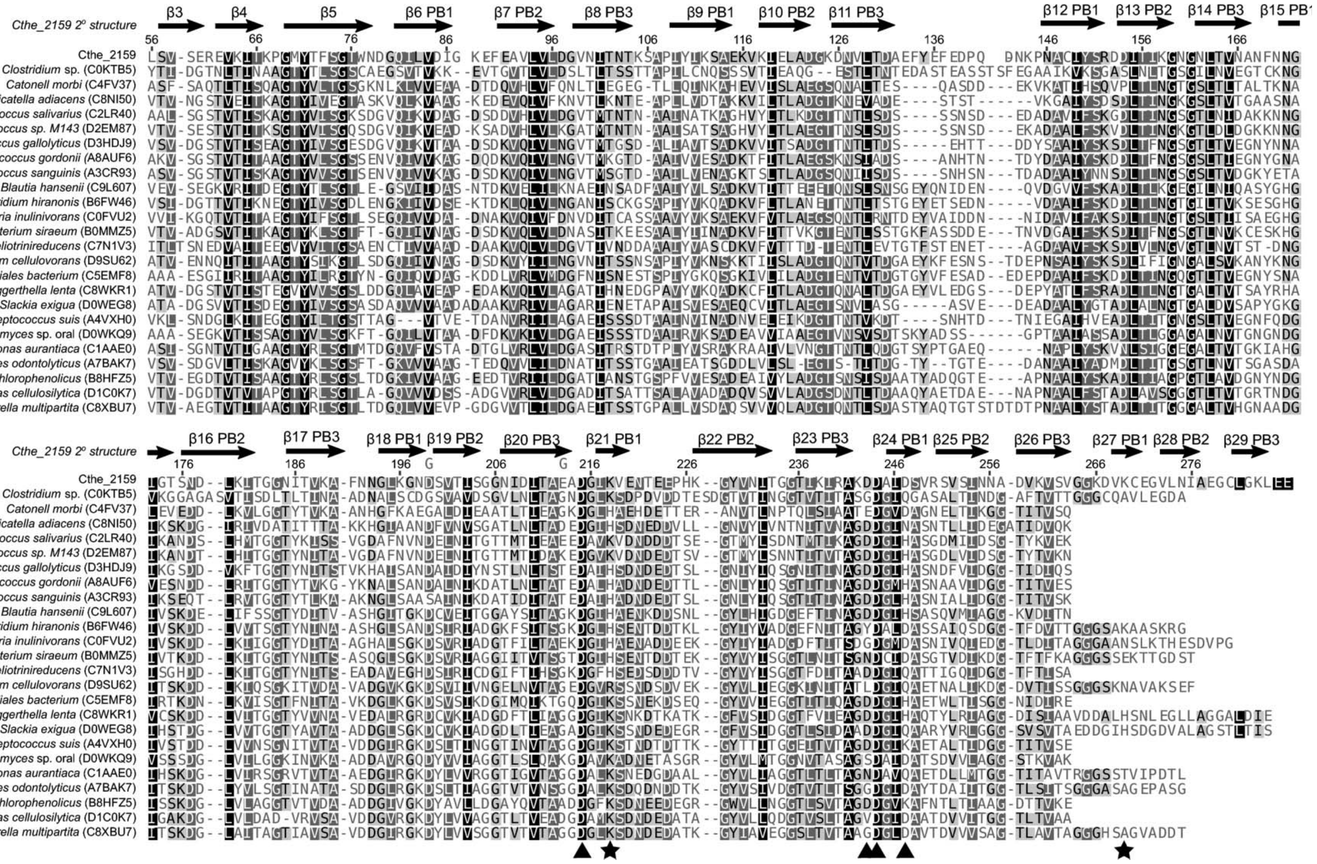
Cthe_2159 2° structure
β3
β4
β5
β6 PB1
β7 PB2
β8 PB3
β9 PB1
β10 PB2
β11 PB3
β12 PB1
β13 PB2
β14 PB3
β15 PB
Cthe_2159
Clostridium sp. (C0KTB5)
Catonell morbi (C4FV37)
Granulicatella adiacens (C8NI50)
Streptococcus salivarius (C2LR40)
Streptococcus sp. M143 (D2EM87)
Streptococcus gallolyticus (D3HDJ9)
Streptococcus gordonii (A8AUF6)
Streptococcus sanguinis (A3CR93)
Blautia hansenii (C9L607)
Clostridium hiranonis (B6FW46)
Roseburia inulinivorans (C0FVU2)
Eubacterium siraeum (B0MMZ5)
Slackia heliotrinireducens (C7N1V3)
Clostridium cellulovorans (D9SU62)
Clostridiales bacterium (C5EMF8)
Eggerthella lenta (C8WKR1)
Slackia exigua (D0WEG8)
Streptococcus suis (A4VXH0)
Actinomyces sp. oral (D0WKQ9)
Gemmatimonas aurantiaca (C1AAE0)
Actinomyces odontolyticus (A7BAK7)
Arthrobacter chlorophenolicus (B8HFZ5)
Xylanimonas cellulosilytica (D1C0K7)
Nakamurella multipartita (C8XBU7)
β16 PB2
β17 PB3
β18 PB1
β19 PB2
β20 PB3
β21 PB1
β22 PB2
β23 PB3
β24 PB1
β25 PB2
β26 PB3
β27 PB1
β28 PB2
β29 PB3
G
G
56 66 76 86 96 106 116 126 136 146 156 166
176 186 196 206 216 226 236 246 256 266 276

Figure 5

Amino-acid sequence alignment of Cthe_2159 with a representative selection from a DUF4353 seed alignment produced by Pfam (Finn *et al.*, 2013). Organism and UniProt ID are indicated. Identical and similar residues are colored using a BLOSUM62 scoring matrix and a threshold of 1. Conservation of aliphatic side chains at PB positions, Gly residues at turns and insertions at loop sites (*e.g* T$3_{11–12}$) predict a conserved overall fold. The D199G and E213G mutations found in the protein used to determine the crystal structure are indicated with a G above the alignment. Positions important for Ca1 coordination (Asp215, Asp243, Asp244 and Asp247) are indicated by solid triangles, and lysines in proximity to the Ca1 site are indicated with solid stars.

actions (Boraston *et al.*, 2004), to our knowledge no CBMs to date are known to adopt the *β*-helix fold.

Although Cthe_2159 is structurally similar to PLs in several respects, it does possess unique features relative to other PLs. For one, the interior of the *β*-helix is devoid of the canonical asparagine ladder, aromatic stacks and disulfide bonds that are ubiquitous across the PL1, PL3 and PL9 families (Jenkins *et al.*, 2004; Creze *et al.*, 2008; Garron & Cygler, 2010). Instead, the interior of the helix is comprised almost entirely of stacked aliphatic residues that are highly superimposable when viewed down the axis of the *β*-helix (Fig. 2*e*) and results in shape differences of the cross-section when viewed along this axis (Fig. 4*a*). Cthe_2159 is similar to PL3 enzymes in that it is relatively short, consisting of eight helical turns and lacking the N-terminal ‘capping’ helix observed for other *β*-helix proteins (Garron & Cygler, 2010). In place of a capping helix, strands *β*1–*β*5 form a unique motif in which *β*1 intervenes between *β*2 and *β*5 in an *α*-parallel arrangement with *β*2 and parallel to *β*5 (Fig. 2*c*). This arrangement results in *β*2 and *β*3 forming a ‘flap’ that folds back over the protein only after *β*1 forms a sheet with *β*5. This N-terminal ‘flap’ demonstrates a third, and previously unknown, structural feature for terminating the continuity of *β*-helix structures. Whereas this feature terminates the solenoid of a single chain, the crystal structure reveals that strands *β*2 and *β*3 can form contiguous sheets with an adjacent monomer (N-terminal interface) along a twofold axis (Fig. 2*a*). Similar dimerization and *β*-helix extension has been observed for a GH28 *β*-helix from *T. maritima*, with multimerization contributing to overall stability (Pijning *et al.*, 2009). While the buried surface area is relatively small for this interface, and we do not observe obvious dimerization after gel filtration (not shown), it is a striking feature that may be functionally important in the presence of longer, polymeric substrates.

Despite the lack of demonstrated enzymatic activity, the presence of three well ordered Ca^{2+} ions is a feature that suggests similarity to PL proteins, in particular the PL9 family. A structural alignment of the primary Ca^{2+} ions and coordinating residues shows that the Ca1 site is nearly superimposable with the only PL9 structure determined to date: Pel9A from *E. chrysanthemi* (*D. dadantii*; PDB entry 1ru4; Jenkins *et al.*, 2004; Fig. 4*c*). Lysine is the Brønsted base in the PL3 and PL9 families, whereas arginine fulfils the role in PL1 enzymes (Jenkins *et al.*, 2004; Creze *et al.*, 2008; Garron & Cygler, 2010). As in Pel9A, there are no arginine side chains in proximity to Ca1, but there are two lysines, Lys218 and

Lys269. Lys269 of Cthe_2159 is in the same proximity relative to the active-site Ca^{2+} as Lys273 of Pel9A (Fig. 4*c*), which is reported to be the Brønsted base (Jenkins *et al.*, 2004). Even though the Ca^{2+}-coordinating and potentially catalytic side chains are essentially superimposable, the overall sequence identity of the two proteins is low (∼9%; Figs. 4*a* and 4*b*), the number of turns varies and the Ca1 site of Cthe_2159 is near the C-terminus of the protein, whereas it is more centrally located in the Pel9A and PL1 enzymes.

When comparing Cthe_2159 with other DUF4353 family members in a sequence alignment, important structural features are conserved (Fig. 5). This is especially true of the residues that dictate the overall fold, with high conservation of the characteristic aliphatic 'stacking' residues found within PB2, PB3 and turn residues in T2, regions that are generally considered to be the most conserved within β-helix proteins (Garron & Cygler, 2010). The Ca1-coordinating residues are nearly ubiquitous, with the positions equivalent to Asp215 and Asp244 essentially invariant and one or other of Asp243 or Asp247 being present (Fig. 5). This implies a conserved Ca^{2+}-binding site, although it is not clear whether Ca^{2+} coordination is important for metal-assisted lyase activity, carbohydrate binding or an as yet to be determined role. Although their positions relative to Ca1 are suggestive of a possible role as Brønsted bases, conservation of Lys218 and Lys269 is not obvious between family members in this alignment. Interestingly, a histidine is the most common substitute for lysine at position 218 (Fig. 5), a side chain that has not been observed as a Brønsted base in PLs, but its role as the base, if in fact the family has an enzymatic function, is chemically feasible.

4. Conclusions

We identified Cthe_2159, a protein of unknown function listed as a DUF4353 family member, in a genome-wide functional annotation screen for *C. thermocellum* domains that interact with cellulose. Our goal was to demonstrate a high-throughput method for identifying carbohydrate-interacting domains that could not be categorized by sequence alone. We determined the structure of Cthe_2159 to gain functional insight into the protein and the DUF4353 family because the sequence alone was not informative.

To determine the structure, we used Gd soaking, which proved to be a facile and rapid method for anomalous phasing using home-source Cu $K\alpha$ radiation. Despite its potential, very few examples exist of Gd phasing, with fewer than 40 of the nearly 100 000 crystal structures currently in the PDB even containing a Gd ligand and only a small fraction of these using Gd for phasing (Girard *et al.*, 2003; Lagartera *et al.*, 2005; Molina *et al.*, 2009; Barends *et al.*, 2014). We add to this by showing that Gd^{3+} can be used to obtain very robust phases by X-ray diffraction with Cu $K\alpha$ radiation using $GdCl_3$ without the need to complex the ion (Girard *et al.*, 2003). In our case, Gd^{3+} ions were tightly coordinated at equivalent positions to the calcium-binding sites observed in the native model, but a total of 13 Gd^{3+} ions were modelled in the Gd-derivative structure, suggesting that Gd^{3+} can be used for phasing even in the absence of Ca^{2+}-binding or other metal-binding sites.

The structural analysis reported here shows that Cthe_2159, and in turn the DUF4353 family, are a new family of β-helix proteins with features similar to carbohydrate-active and carbohydrate-binding proteins. Robust binding to cellulose and PGA provides functional evidence for a role in carbohydrate interaction. Further support for this role comes from the observation that DUF4353 family members are modular and often in fusion with transmembrane, dockerin and S-layer homology domains, fusions that are common to PLs and other carbohydrate-active or carbohydrate-binding domains (Lombard *et al.*, 2010; Finn *et al.*, 2013). The observed Ca^{2+} coordination is highly similar to PLs (in particular Pel9A; Jenkins *et al.*, 2004) that utilize metal-assisted β-elimination and typically act on pectate or pectin (Garron & Cygler, 2010), but enzymatic activity has yet to be established for Cthe_2159. Indeed, Cthe_2159 may not have enzymatic activity and may instead utilize Ca^{2+} for binding or fold stabilization, as is the case for a number of CBMs and lectin proteins (Boraston *et al.*, 2004; Jamal-Talabani *et al.*, 2004; Ghosh *et al.*, 2013). The lack of demonstrated enzymatic activity may be owing to the limitations of our particular assays, owing to incorrect substrates and/or conditions or because the protein is not an enzyme despite strong similarities to PLs. In future studies, we will focus on more exhaustive and rigorous activity assays using other substrates (including pectate, pectin, hyaluronan, chondroitin, heparin, xanthan and alginate), with the possibility that it may act on a substrate previously not observed for PLs.

The work described here illustrates the power of structural determination as a means of gene annotation. Although structural determination is a relatively low-throughput technique, the identification of Cthe_2159 as a member of a new family of β-helix proteins with structural similarities to the polysaccharide lyases, coupled with its ability to bind cellulosic and xylan substrates, would not have been possible without this structural determination. This is particularly true given the very low sequence homology of Cthe_2159 to the carbohydrate-binding and carbohydrate-active β-helix proteins that it structurally resembles. This new insight into the binding activity and structural similarity of Cthe_2159 will allow putative annotation of the DUF4353 family as β-helix proteins with similarities to the polysaccharide lyases.

We would like to acknowledge Andrey Kovalevsky and the Los Alamos National Laboratory Protein Crystallography Station (PCS) for crystallization materials and data-collection facilities. We would like to thank Tom Terwilliger for helpful discussion and comments on the manuscript. We would also like to thank the anonymous reviewers for insightful comments and suggestions. Funding was provided by a DOE GTL grant awarded to ARMB.

References

Abbott, D. W. & Boraston, A. B. (2008). *Microbiol. Mol. Biol. Rev.* **72**, 301–316.

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

Akita, M., Suzuki, A., Kobayashi, T., Ito, S. & Yamane, T. (2001). *Acta Cryst.* D**57**, 1786–1792.

Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 10037–10041.

Barends, T. R. M., Foucar, L., Botha, S., Doak, R. B., Shoeman, R. L., Nass, K., Koglin, J. E., Williams, G. J., Boutet, S., Messerschmidt, M. & Schlichting, I. (2014). *Nature (London)*, **505**, 244–247.

Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. W. (2011). *Acta Cryst.* D**67**, 271–281.

Bayer, E. A., Belaich, J.-P., Shoham, Y. & Lamed, R. (2004). *Annu. Rev. Microbiol.* **58**, 521–554.

Bell, M. J., Collison, M. & Lord, P. (2013). *PLoS One*, **8**, e75541.

Berdichevsky, Y., Ben-Zeev, E., Lamed, R. & Benhar, I. (1999). *J. Immunol. Methods*, **228**, 151–162.

Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. (2004). *Biochem. J.* **382**, 769–781.

Boschker, H. T. S. & Cappenberg, T. E. (1994). *Appl. Environ. Microbiol.* **60**, 3592–3596.

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

Charnock, S. J., Brown, I. E., Turkenburg, J. P., Black, G. W. & Davies, G. J. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 12067–12072.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.

Chernoglazov, V. M., Jafarova, A. N. & Klyosov, A. A. (1989). *Anal. Biochem.* **179**, 186–189.

Collmer, A., Ried, J. L. & Mount, M. S. (1988). *Methods Enzymol.* **161**, 329–335.

Colman, P. M., Weaver, L. H. & Matthews, B. W. (1972). *Biochem. Biophys. Res. Commun.* **46**, 1999–2005.

Creze, C., Castang, S., Derivery, E., Haser, R., Hugouvieux-Cotte-Pattat, N., Shevchik, V. E. & Gouet, P. (2008). *J. Biol. Chem.* **283**, 18260–18268.

D'Angelo, S., Velappan, N., Mignone, F., Santoro, C., Sblattero, D., Kiss, C. & Bradbury, A. R. M. (2011). *BMC Genomics*, **12**, S5.

Demain, A. L., Newcomb, M. & Wu, J. H. D. (2005). *Microbiol. Mol. Biol. Rev.* **69**, 124–154.

El Kaoutari, A., Armougom, F., Gordon, J. I., Raoult, D. & Henrissat, B. (2013). *Nature Rev. Microbiol.* **11**, 497–504.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L., Tate, J. & Punta, M. (2013). *Nucleic Acids Res.* **42**, D222–D230.

Garron, M. L. & Cygler, M. (2010). *Glycobiology*, **20**, 1547–1573.

Ghosh, A., Luís, A. S., Brás, J. L. A., Pathaw, N., Chrungoo, N. K., Fontes, C. M. G. A. & Goyal, A. (2013). *PLoS One*, **8**, e80415.

Girard, É., Stelter, M., Vicat, J. & Kahn, R. (2003). *Acta Cryst.* D**59**, 1914–1922.

Gold, N. D. & Martin, V. J. J. (2007). *J. Bacteriol.* **189**, 6787–6795.

Goodacre, N. F., Gerloff, D. L. & Uetz, P. (2013). *mBio*, **5**, e00744-13.

Hassan, S., Shevchik, V. E., Robert, X. & Hugouvieux-Cotte-Pattat, N. (2013). *J. Bacteriol.* **195**, 2197–2206.

Hemme, C. L. *et al.* (2010). *J. Bacteriol.* **192**, 6494–6496.

Henrissat, B. (1991). *Biochem. J.* **280**, 309–316.

Herron, S. R., Scavetta, R. D., Garrett, M., Legner, M. & Jurnak, F. (2003). *J. Biol. Chem.* **278**, 12271–12277.

Hla, S. S., Kurokawa, J., Suryani, Kimura, T., Ohmiya, K. & Sakka, K. (2005). *Biosci. Biotechnol. Biochem.* **69**, 2138–2145.

Hyeon, J.-E., Yu, K.-O., Suh, D. J., Suh, Y.-W., Lee, S. E., Lee, J. & Han, S. O. (2010). *FEMS Microbiol. Lett.* **310**, 39–47.

Jamal-Talabani, S., Boraston, A. B., Turkenburg, J. P., Tarbouriech, N., Ducros, V. M.-A. & Davies, G. J. (2004). *Structure*, **12**, 1177–1187.

Jenkins, J., Mayans, O. & Pickersgill, R. (1998). *J. Struct. Biol.* **122**, 236–246.

Jenkins, J., Shevchik, V. E., Hugouvieux-Cotte-Pattat, N. & Pickersgill, R. W. (2004). *J. Biol. Chem.* **279**, 9139–9145.

Jones, C. E., Brown, A. L. & Baumann, U. (2007). *BMC Bioinformatics*, **8**, 170.

Kajava, A. V. & Steven, A. C. (2006). *Fibrous Proteins: Amyloids, Prions and Beta Proteins*, pp. 55–96. New York: Academic Press.

Kiss, C., Temirov, J., Chasteen, L., Waldo, G. S. & Bradbury, A. R. M. (2009). *Protein Eng. Des. Sel.* **22**, 313–323.

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

Lagartera, L., González, A., Stelter, M., García, P., Kahn, R., Menéndez, M. & Hermoso, J. A. (2005). *Acta Cryst.* F**61**, 221–224.

Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P. M. & Henrissat, B. (2010). *Biochem. J.* **432**, 437–444.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. (2014). *Nucleic Acids Res.* **42**, D490–D495.

Miller, G. L. (1959). *Anal. Chem.* **31**, 426–428.

Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. (2006). *Acta Cryst.* D**62**, 859–866.

Molina, R., Stelter, M., Kahn, R. & Hermoso, J. A. (2009). *Acta Cryst.* D**65**, 823–831.

Pijning, T., van Pouderoyen, G., Kluskens, L., van der Oost, J. & Dijkstra, B. W. (2009). *FEBS Lett.* **583**, 3665–3670.

Seyedarabi, A., To, T. T., Ali, S., Hussain, S., Fries, M., Madsen, R., Clausen, M. H., Teixteira, S., Brocklehurst, K. & Pickersgill, R. W. (2010). *Biochemistry*, **49**, 539–546.

Studier, F. W. (2005). *Protein Expr. Purif.* **41**, 207–234.

Tamaru, Y. & Doi, R. H. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 4125–4129.

Tracy, B. P., Jones, S. W., Fast, A. G., Indurthi, D. C. & Papoutsakis, E. T. (2012). *Curr. Opin. Biotechnol.* **23**, 364–381.

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

Wood, T. M. & Bhat, K. M. (1988). *Methods Enzymol.* **160**, 87–112.

Yang, S., Giannone, R. J., Dice, L., Yang, Z. K., Engle, N. L., Tschaplinski, T. J., Hettich, R. L. & Brown, S. D. (2012). *BMC Genomics*, **13**, 336.

Ye, Y. & Godzik, A. (2004). *Nucleic Acids Res.* **32**, W582–W585.

Yoder, M. D. & Jurnak, F. (1995). *Plant Physiol.* **107**, 349–364.

Zheng, Y., Huang, C.-H., Liu, W., Ko, T.-P., Xue, Y., Zhou, C., Guo, R.-T. & Ma, Y. (2012). *Biochem. Biophys. Res. Commun.* **420**, 269–274.