Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# Structure of a family 3b′ carbohydrate-binding module from the Cel9V glycoside hydrolase from *Clostridium thermocellum*: structural diversity and implications for carbohydrate binding

**Svetlana Petkun,[a,b] Sadanari Jindou,[a] Linda J. W. Shimon,[c] Sonia Rosenheck,[a] Edward A. Bayer,[d] Raphael Lamed[a,b] and Felix Frolow[a,b]***

[a]Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv 69978, Israel, [b]The Daniella Rich Institute for Structural Biology, Tel Aviv University, Tel Aviv 69978, Israel, [c]Department of Chemical Research Support, The Weizmann Institute of Science, Rehovot 76100, Israel, and [d]Department of Biological Chemistry, The Weizmann Institute of Science, Rehovot 76100, Israel

Correspondence e-mail: mbfrolow@post.tau.ac.il

Received 6 August 2009
Accepted 19 October 2009

**PDB References:** family 3b′ carbohydrate-binding module, 2wob; 2wo4; 2wnx.

Family 3 carbohydrate-binding modules (CBM3s) are associated with both cellulosomal scaffoldins and family 9 glycoside hydrolases (GH9s), which are multi-modular enzymes that act on cellulosic substrates. CBM3s bind cellulose. X-ray crystal structures of these modules have established an accepted cellulose-binding mechanism based on stacking interactions between the sugar rings of cellulose and a planar array of aromatic residues located on the CBM3 surface. These planar-strip residues are generally highly conserved, although some CBM3 sequences lack one or more of these residues. In particular, CBM3b′ from *Clostridium thermocellum* Cel9V exhibits such sequence changes and fails to bind cellulosic substrates. A crystallographic investigation of CBM3b′ has been initiated in order to understand the structural reason(s) for this inability. CBM3b′ crystallized in space group $C222_1$ (diffraction was obtained to 2.0 Å resolution in-house) with three independent molecules in the asymmetric unit and in space group $P4_12_12$ (diffraction was obtained to 1.79 Å resolution in-house and to 1.30 Å resolution at a synchrotron) with one molecule in the asymmetric unit. The molecular structure of Cel9V CBM3b′ revealed that in addition to the loss of several cellulose-binding residues in the planar strip, changes in the backbone create a surface 'hump' which could interfere with the formation of cellulose–protein surface interactions and thus prevent binding to crystalline cellulose.

## 1. Introduction

Cellulases and glycoside hydrolases are cellulose-degrading enzymes which usually contain multiple structurally and functionally diverse modules on the same polypeptide chain (Davies *et al.*, 1995; Gilkes *et al.*, 1991; Henrissat, 1997; Bayer *et al.*, 1998).

The cellulose-degradation process starts with the binding of the cellulolytic enzymes or the entire organism to the cellulosic substrate (Bayer *et al.*, 1998). A separate domain, the carbohydrate-binding module (CBM), mediates this step. CBMs can serve as targeting agents for the catalytic modules of free cellulases (Boraston *et al.*, 2004; Tomme *et al.*, 1995) or can act as a separate targeting module as part of the non-catalytic scaffoldin subunit of the cellulosome (Bayer *et al.*, 1998). The scaffoldin also contains numerous cohesin modules, which bind selectively to a complementary dockerin modular component of the individual cellulosomal enzymes. In addition to their role as a targeting vehicle, it has been proposed that CBMs may mediate the nonhydrolytic disruption of cellulose fibres, thereby facilitating subsequent enzymatic degradation by the catalytic domains (Din *et al.*, 1991; Tormo

*et al.*, 1996). Alternatively, CBMs may recognize the disengaged cellulose chains and assist hydrolysis by directing single cellulose chains into the active site of an adjacent hydrolyzing module (Gal *et al.*, 1997; Irwin *et al.*, 1998).

Family 3 CBMs comprise ~155 residues. They have been identified in many hydrolytic enzymes that contain the family 9 catalytic module (GH9). These include Cel9A from *Thermobifida fusca*, Cel9G from *Clostridium cellulolyticum*, Cel9I and CbhA from *C. thermocellum* and Cel9V from *Acetivibrio cellulolyticus*. CBM3s have also been identified in nonhydrolytic cellulosomal proteins (scaffoldins), *e.g.* CipA from *C. thermocellum*, CipC from *C. cellulolyticum* and CipV from *A. cellulolyticus* (Pagès *et al.*, 1996; Ding *et al.*, 1999). Schematic representations of some CBM3-containing proteins are shown in Fig. 1.

Three major subgroups have been identified in the CBM3 family on the basis of sequence similarity: 3a, 3b and 3c (Bayer *et al.*, 1998). Subfamilies 3a and 3b are more closely related to each other than to subfamily 3c. In biochemical studies, members of the 3a and 3b subgroups have been shown to bind strongly to the surface of microcrystalline cellulose and have been proposed to promote the cellulolytic reaction by concentrating the enzyme near the cellulose surface (Gilad *et al.*, 2003; Tormo *et al.*, 1996). In contrast to the characteristic cellulose-binding function observed for members of subfamilies 3a and 3b, CBM3c is believed to convey a single cellulose chain into the GH9 active site, thereby producing processive cleavage of crystalline cellulose by the enzyme. Moreover, CBM3s from subfamilies 3a and 3b contain well conserved cellulose-binding residues that differ from those of CBM3c (Jindou *et al.*, 2006). Cellulosomal cellulases generally lack CBM3a or CBM3b modules as they are targeted collectively to the cellulose substrate by means of the scaffoldin-borne CBM3a.

Recently, however, the commonly assumed hypothesis that all family 3a and 3b CBMs function primarily as cellulose-binding modules has been questioned by the discovery that certain CBM3b modules (designated CBM3b′) lack some of the normally conserved residues of the cellulose-binding planar strip previously identified in 'classic' CBM3s (Fig. 1*c*). For example, three intriguing cellulosomal enzymes (*C. thermocellum* Cel9V and Cel9U and *A. cellulolyticus* Cel9B) have been described that contain a CBM3c and a CBM3b′ module attached in tandem to the C-terminus of the GH9 catalytic module. These CBM3b′ molecules lack the ability to bind not only crystalline or amorphous cellulose but also the carbohydrates found in natural substrates such as banana stem, alfalfa and wheat (Jindou *et al.*, 2006). The ability to bind cellulose was not restored even after replacement of the nonconserved Ala residue in the cellulose-binding strip by the commonly conserved aromatic Trp (Jindou *et al.*, 2006). From this, it was inferred that there are additional changes in this molecule that must be responsible for its inability to bind cellulose. Furthermore, Cel9V appears to be a cellulosomal cellulase as it possesses a type I dockerin module (Fig. 1*c*).

Thus, X-ray structure determination was initiated in order to explore the structural reasons behind the inability of CBM3b′ to bind cellulosic substrates. Here, we report the molecular structure of *C. thermocellum* Cel9V CBM3b′, which is the first molecular structure of a subfamily 3b member of the family 3 carbohydrate-binding modules.

## 2. Materials and methods

### 2.1. Cloning of the CBM3b′ from Cel9V of *C. thermocellum*

A DNA fragment encoding Cel9V CBM3b′ was amplified by PCR from *C. thermocellum* genomic DNA using two specific primers, 5′-ATATCCATGGATCCCAGCCAAACTCCGGATGCT-3′ and 5′-CCCCGCGGCCGCCTCAACTCCATATTTGAGTTTATC-3′. The PCR products were purified and cleaved with the restriction enzymes *Nco*I and *Not*I and inserted into the pET-28a(+) expression vector (Novagen,

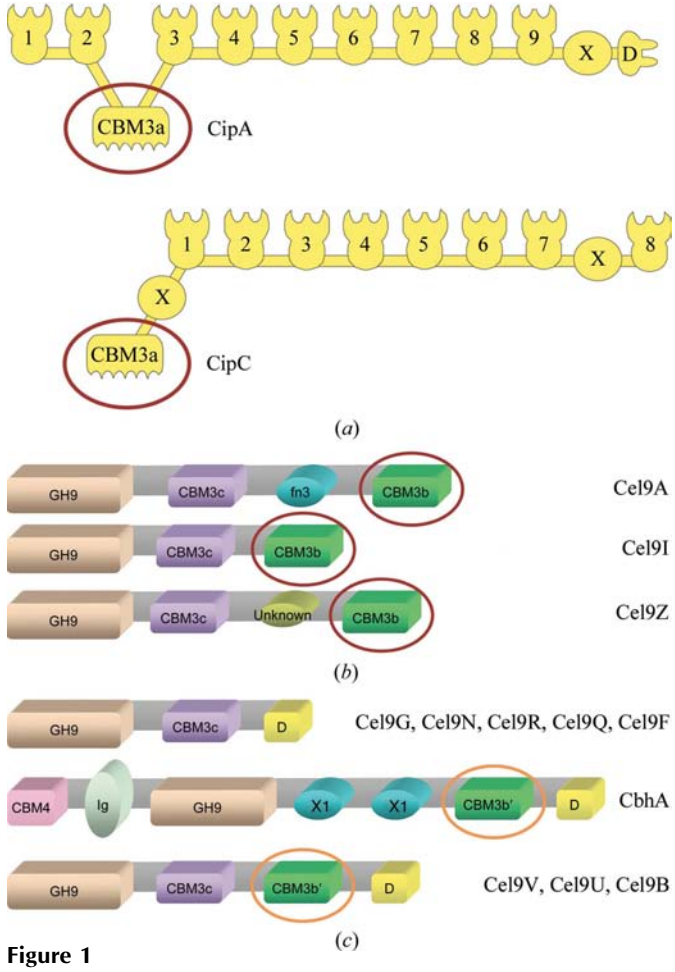


**Figure 1**
Schematic representation of CBM3-containing proteins. (*a*) Scaffoldin proteins: the 'classic' CBM3a modules of CipA from *C. thermocellum* and CipC from *C. cellulolyticum*. (*b*) Free cellulases: Cel9A from *T. fusca*, Cel9I from *C. thermocellum* and Cel9Z from *C. stercorarium*. (*c*) The cellulosomal cellulases Cel9G from *C. cellulolyticum* and Cel9N, Cel9R, Cel9Q and Cel9F from *C. thermocellum* and the cellulosomal cellulases containing the CBM3b′ module that fail to bind carbohydrates CbhA, Cel9V and Cel9U from *C. thermocellum* and Cel9B from *A. cellulolyticus*. Cellulose-binding CBM3b modules are emphasized by a dark red ellipse. Nonbinding CBM3b′ modules are emphasized by an orange ellipse. CBM, cellulose-binding module; GH9, family 9 glycoside hydrolase; D, dockerin; X1, X-domain; fn, fibronectin-like domain; Ig, immunoglobulin-like domain.

**Table 1**
Crystal and diffraction data.

Values in parentheses are for the highest resolution shell. TAU data were collected at Tel Aviv University. WIS data were collected at Weizmann Insitute of Science. ESRF data were collected at the European Synchrotron Radiation Facility.

| | TAU | WIS | ESRF |
|---|---|---|---|
| Space group | $C222_1$ | $P4_12_12$ | $P4_12_12$ |
| No. of crystals | 1 | 1 | 1 |
| Total rotation range (°) | 100 | 85 | 180 |
| Unit-cell parameters | | | |
| $a$ (Å) | 57.03 | 73.90 | 74.13 |
| $b$ (Å) | 86.61 | 73.90 | 74.13 |
| $c$ (Å) | 242.37 | 85.76 | 82.50 |
| $V$ (Å$^3$) | 1197147.5 | 467192.72 | 453351.96 |
| No. of molecules in ASU | 3 | 1 | 1 |
| Resolution range (Å) | 50–2.0 (2.03–2.00) | 30–1.79 (1.85–1.79) | 26–1.3 (1.35–1.30) |
| Total No. of reflections | 135627 | 150667 | 463125 |
| Unique reflections | 41368 | 23006 | 57111 |
| Mosaicity (°) | 0.344 | 0.25 | 0.40 |
| Redundancy | 3.5 | 6.6 | 8.6 |
| Completeness (%) | 95 (89.7) | 100 (100) | 94.6 (81.5) |
| Average $I/\sigma(I)$ | 16.9 (1.56) | 32.19 (1.66) | 48.15 (1.86) |
| $R_{\rm merge}$† | 0.106 (0.58) | 0.062 (0.54) | 0.08 (0.70) |
| Overall $B$ factor from Wilson plot (Å$^2$) | 31.6 | 23.6 | 17.1 |

† $R_{\rm merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$, where $\sum_{hkl}$ denotes the sum over all reflections and $\sum_i$ the sum over all equivalent and symmetry-related reflections (Stout & Jensen, 1968).

**Table 2**
Summary of refinement statistics.

| | TAU | WIS | ESRF |
|---|---|---|---|
| Space group | $C222_1$ | $P4_12_12$ | $P4_12_12$ |
| Resolution range (Å) | 50.0–2.0 | 30.0–1.79 | 26.0–1.3 |
| No. of protein atoms | 3545 | 1153 | 1207 |
| No. of solvent atoms | 273 | 143 | 353 |
| No. of ions | 3 | 1 | 1 |
| $R_{\rm cryst}$ | 0.169 | 0.178 | 0.123 |
| $R_{\rm free}$ | 0.220 | 0.216 | 0.156 |
| Geometry | | | |
| R.m.s.d. bonds (Å) | 0.019 | 0.013 | 0.012 |
| R.m.s.d. bond angles (°) | 1.765 | 1.452 | 1.431 |
| *MolProbity* validation | | | |
| Ramachandran favoured (%) (goal > 98%) | 96.47 | 98.67 | 98.69 |
| Ramachandran outliers (%) (goal < 0.2%) | 0.44 | 0.65 | 0.79 |
| $C^\beta$ deviations > 0.25 Å (goal 0) | 5 | 0 | 0 |
| Clash score† (all atoms) | 12.69 | 5.77 | 4.74 |
| Rotamer outliers (%) (goal < 1%) | 3.44 | 0.79 | 0.79 |
| Residues with bad bonds (%) (goal < 1%) | 0.22 | 0 | 0.00 |
| Residues with bad angles (%) (goal < 0.5%) | 0.44 | 0.66 | 0.00 |
| PDB code | 2wob | 2wo4 | 2wnx |

† Clash score is the number of serious steric overlaps (>0.4 Å) per 1000 atoms.

Madison, Wisconsin, USA) together with a C-terminal hexa-His tag, yielding pET-Cel9V-CBM3b′.

### 2.2. Protein expression and purification

*Escherichia coli* BL21 (DE3) harbouring pET-Cel9V-CBM3b′ was aerated at 310 K in 3 l Terrific broth supplemented with 25 µg ml$^{-1}$ kanamycin. After 3 h, when the culture had reached an $A_{600}$ of 0.6, 0.1 m*M* isopropyl $\beta$-D-1-thiogalactopyranoside (IPTG) was added to induce gene expression and cultivation was continued at 310 K for an additional 12 h. Cells were then harvested by centrifugation (5000*g* for 10 min) at 277 K and subsequently resuspended in 50 m*M* $NaH_2PO_4$ pH 8.0 containing 300 m*M* NaCl at a ratio of 1 g wet pellet per 4 ml buffer solution. A few micrograms of DNase powder were added prior to the sonication procedure. The suspension was kept on ice during sonication, after which it was centrifuged (20 000*g* at 277 K for 20 min) and the supernatant was collected. The expressed His-tagged protein was isolated by metal-chelate affinity chromatography using Ni–IDA resin (Rimon Biotech, Israel). The His tag was not removed. Fast protein liquid chromatography (FPLC) was performed using a Superdex 75 pg column and an ÄKTA Prime system (GE Healthcare). The protein was concentrated to 25 mg ml$^{-1}$ using Centriprep YM-3 centrifugal filter devices (Amicon Bioseparation, Millipore). The protein concentration was determined by measuring the UV absorbance at 280 nm.

### 2.3. Crystallization and X-ray diffraction

The protein sample (25 mg ml$^{-1}$ protein in 1.2 m*M* Tris–HCl pH 7.5, 1.5 m*M* NaCl and 0.025% sodium azide) was initially screened using 122 conditions from the Crystal Screen, Crystal Screen 2 and Grid Screen Ammonium Sulfate pre-formulated kits (Hampton Research, California, USA). Screening was performed using the vapour-diffusion method. Each drop contained a 2 µl:2 µl mixture of protein and precipitant solutions. The first crystals appeared in condition No. 9 of Crystal Screen [0.2 *M* ammonium acetate, 0.1 *M* trisodium citrate dihydrate pH 5.6, 30%(*w*/*v*) polyethylene glycol 4000] after 13 d. The crystals were harvested from the crystallization drop using a nylon loop (Teng, 1990; Hampton Research, California, USA) and transferred into a cryo-stabilization solution mimicking the mother liquor supplemented with 25% ethylene glycol. For data collection, crystals were mounted on a MiTeGen micro-mount (http://www.mitegen.com) made of polyimide and flash-cooled in a nitrogen stream produced by an Oxford Cryostream low-temperature generator (Cosier & Glazer, 1986) at a temperature of 100 K. Diffraction data from the CBM3b′ crystals were measured using an R-AXIS IV area detector mounted on a Rigaku ultraX 18 Cu rotating-anode X-ray generator (wavelength 1.5418 Å) equipped with Osmic confocal mirrors at Tel Aviv University (TAU). Diffraction data consisting of 200 images with 0.5° oscillation per frame were collected. Data were processed with *DENZO* and scaled with *SCALEPACK* as implemented in *HKL*-2000 (Otwinowski & Minor, 1997). The crystal diffracted to 2.0 Å resolution and belonged to the orthorhombic space group $C222_1$, with unit-cell parameters $a = 57.03$, $b = 86.61$, $c = 242.37$ Å. The presence of three monomers in the asymmetric unit gave an acceptable packing density $V_{\rm M}$ (Matthews, 1968) of 2.68 Å$^3$ Da$^{-1}$, corresponding to a solvent content of 54.2%. Diffraction data analysis statistics are given in Table 1.

Crystals with different morphology were obtained using condition C2 from the Grid Screen Ammonium Sulfate kit (2.4 *M* ammonium sulfate and 0.1 *M* citric acid pH 5.0). After one round of optimization, final crystallization conditions were defined consisting of 1.8 *M* ammonium sulfate, 0.1 *M*

citric acid pH 5.0. The crystals grew over a period of three weeks. Diffraction data from these crystals were measured using an R-AXIS IV$^{++}$ area detector mounted on a Rigaku RU-H3R Cu rotating-anode X-ray generator (wavelength 1.5418 Å) equipped with Osmic confocal mirrors at a temperature of 120 K at the Weizmann Institute of Science (WIS). The crystals diffracted to 1.79 Å resolution and belonged to the tetragonal space group $P4_12_12$, with unit-cell parameters $a = b = 73.90$, $c = 85.76$ Å. Diffraction data consisting of 170 images with 0.5° oscillation per frame were collected. The data were processed and scaled as described above for the orthorhombic crystal. The presence of one monomer in the asymmetric unit gave an acceptable packing density $V_M$ (Matthews, 1968) of 3.04 Å$^3$ Da$^{-1}$, corresponding to a solvent content of 59.6%. Diffraction data analysis statistics are given in Table 1. Diffraction data from these crystals were also measured on the ID14-4 beamline at ESRF, Grenoble, France. An ADSC Q315 detector and X-ray radiation of 0.976 Å wavelength were used. Diffraction data consisting of 360 images with 0.5° oscillation per frame were collected. The data were processed with *DENZO* and scaled with *SCALEPACK* as implemented in *HKL*-2000 (Otwinowski & Minor, 1997). The crystals diffracted to 1.3 Å resolution and belonged to the tetragonal space group $P4_12_12$, with unit-cell parameters $a = b = 74.13$, $c = 82.50$ Å. Diffraction data analysis statistics are given in Table 1.

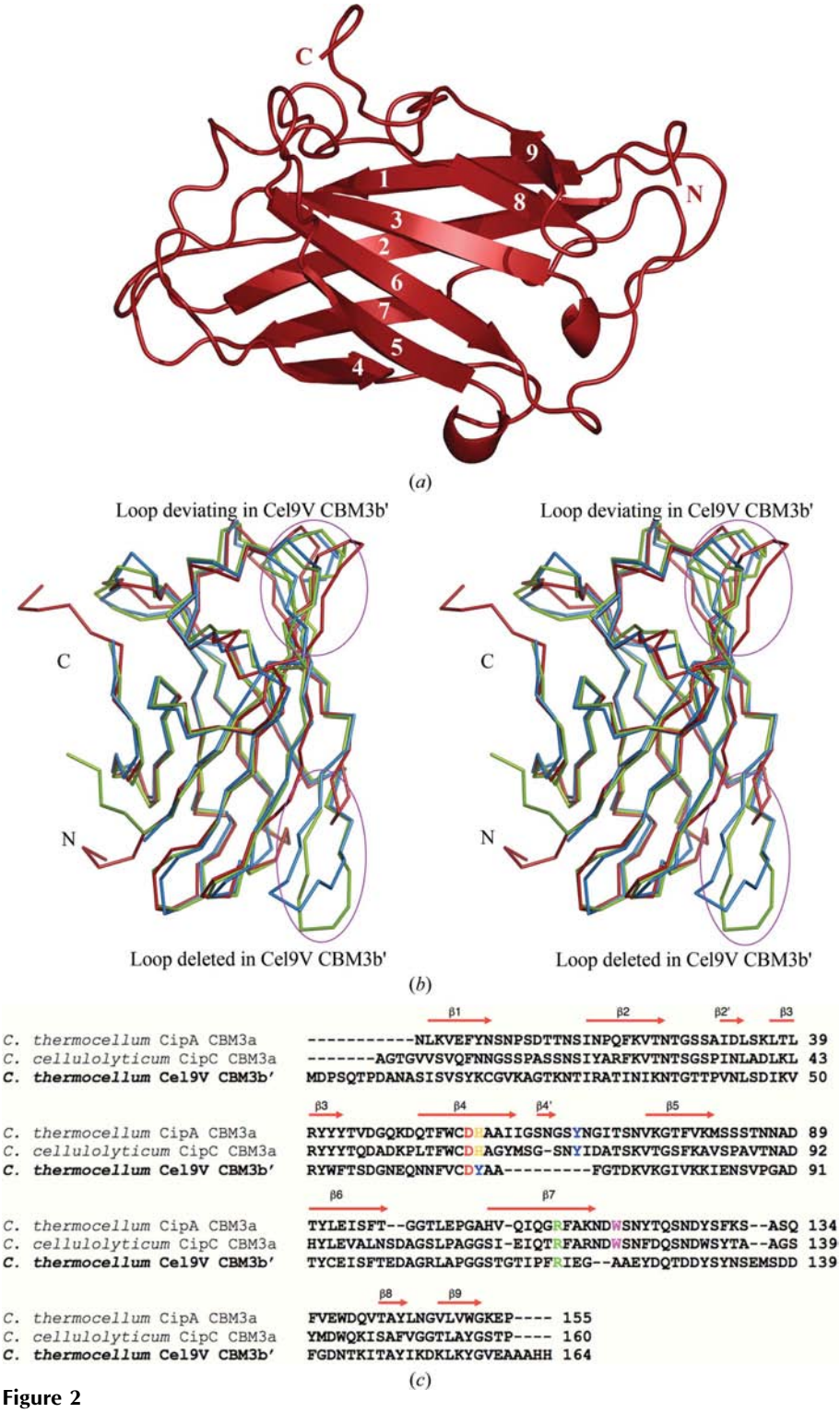


**Figure 2**
(*a*) Structural and sequence alignments of CBM3a and CBM3b′ modules. A cartoon representation of the Cel9V CBM3b′ structure is shown with the $\beta$-strands labelled. (*b*) Stereo diagram of the superimposed C$^\alpha$ traces of *C. thermocellum* Cel9V CBM3b′ (red), *C. thermocellum* CipA (green) and *C. cellulolyticum* CipC (blue). (*c*) Sequence alignment of CipA CBM3a, CipC CBM3a and Cel9V CBM3b′, with secondary-structure elements annotated.

### 2.4. Structure determination and refinement

The structure of the $P4_12_12$ crystal was determined by molecular replacement employing *Phaser* (Storoni *et al.*, 2004; McCoy *et al.*, 2005), using the atomic coordinates of *C. thermocellum* CipA CBM3a (PDB code 1nbc) as a search model. The sequence identity between the model and the target protein was 34%. A clear solution with a *Z* score of 17.74, a log-likelihood gain of 213.89 and two clashes was found after the maximal number of legal clashes (C$^\alpha$ backbone overlay) was increased from 0 to 6. After 20 cycles of rigid-body refinement using *REFMAC*5 (Murshudov *et al.*, 1997), the *R* and the $R_{free}$ (using a random 5% subset of the data set aside for cross-validation; Brünger, 1992) factors were 0.534 and 0.489, respectively.

The model was subjected to several rounds of restrained refinement of positional and thermal parameters using *REFMAC*5. The results of the refinement were assessed using *O* (Jones *et al.*, 1991) and *Coot* (Emsley & Cowtan, 2004). The $2F_o - F_c$ and $F_o - F_c$ electron-density maps revealed the positions of the nonconserved side chains of the Cel9V CBM3b′ sequence, thus allowing them to be manually built into the map using *O* and *Coot*. The lack of electron density corresponding to the extended loop between strands 4 and 4′ (residues 60–67), which is present in the search model CipA CBM3a but is deleted in the Cel9V CBM3b sequence, was obvious. The $R$ and $R_{free}$ factors converged to 0.347 and 0.358, respectively. The *ARP*/*wARP* procedure significantly improved the electron density (Perrakis *et al.*, 1999) and the $R$ and $R_{free}$ factors converged to 0.169 and 0.234, respectively. A $Ca^{2+}$ ion was located in the electron density and built into the model. After several additional rounds of refinement, *ARP*/*wARP* electron-density improvement, manual model correction and addition of solvent atoms, the final Cel9V CBM3b model for this data set at 1.79 Å resolution was obtained (see Table 2). The corresponding $R$ and $R_{free}$ factors for this model are 0.178 and 0.216, respectively. When the 1.3 Å resolution data measured at the synchrotron became available, the coordinates of the previously determined 1.79 Å resolution structure stripped of solvent and $Ca^{2+}$ ion were used as the starting model for refinement with a newly defined set of $R_{free}$ reflections. After 20 cycles of rigid-body fitting, the $R$ and the $R_{free}$ factors were 0.384 and 0.388, respectively. After numerous cycles of manual rebuilding, positional and thermal isotropic refinement and solvent-atom placement using *ARP*/*wARP* (Perrakis *et al.*, 1999), the $R$ and $R_{free}$ factors fell to 0.182 and 0.209, respectively. Introduction of H atoms at calculated positions resulted in a further improvement of the $R$ and $R_{free}$ factors, which at this stage were 0.165 and 0.195. One $Ca^{2+}$ and three sulfate ions were found during further electron-density analysis. Refinement of TLS (rigid-body translation/libration/screw motions) parameters (Winn *et al.*, 2001, 2003; Winn, 2003) improved the $R$ and $R_{free}$ factors, which finally converged to 0.130 and 0.165. The $R$ and $R_{free}$ factors further converged to 0.123 and 0.156, respectively, when anisotropic $B$ temperature factors were refined.

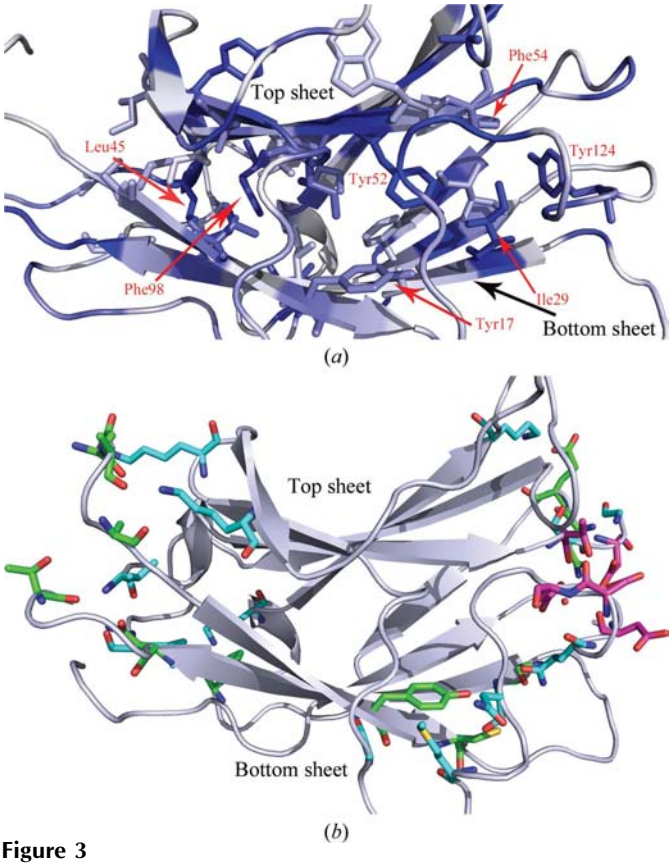


**Figure 3**
The aromatic and hydrophobic core of Cel9V CBM3b′. (*a*) The residues forming the hydrophobic core (shown in stick representation) were determined using the *Protein Interactions Calculator* server (http://crick.mbu.iisc.ernet.in/~PIC/). The blue colour gradient was prepared using *ProtSkin* (Deprez *et al.*, 2005) and *ClustalW* (Larkin *et al.*, 2007) alignment of CBM3a and CBM3b modules as in Fig. 5 (data not shown). Dark blue represents highly conserved and light blue less conserved. (*b*) Residues forming hydrogen bonds and salt bridges between the top and the bottom $\beta$-sheets are shown in stick representation and are coloured green. $Ca^{2+}$-coordinating residues are shown in pink. The contacting residues were found using *PISA* (Krissinel & Henrick, 2007).

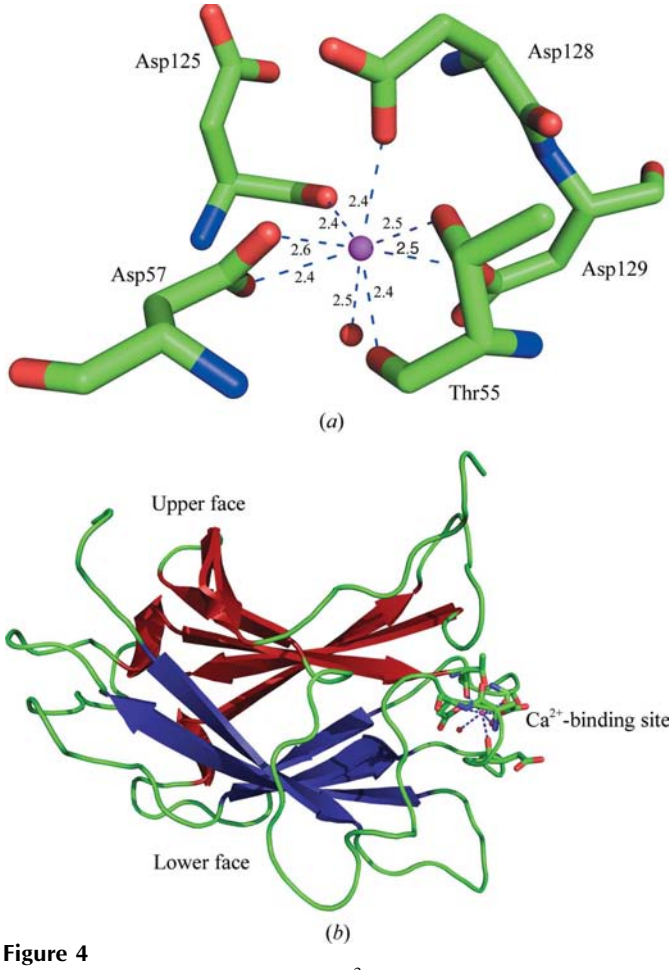


**Figure 4**
Coordination and location of the $Ca^{2+}$ ion in CBM3b′ of Cel9V from *C. thermocellum*. (*a*) $Ca^{2+}$-coordinating residues in the CBM3b′. (*b*) The $Ca^{2+}$-binding site plays a structural role, tethering the upper and the lower $\beta$-sheets together. The upper $\beta$-sheet is coloured red, the lower $\beta$-sheets is coloured blue and the loop regions are coloured green. $Ca^{2+}$-coordinating residues are shown as sticks.

Structure determination of the $C222_1$ crystal, which contained three independent molecules in the asymmetric unit, only became possible after the structure of the $P4_12_12$ crystals, which contained only one molecule in the asymmetric unit, was available for use as a search model. Using the coordinates of the tetragonal structure stripped of water molecules and $Ca^{2+}$, the molecular-replacement program *MOLREP* (Vagin & Teplyakov, 1997) found a clear solution with three molecules in the asymmetric unit. The *R* factor and correlation coefficient constantly improved when additional molecules were located, and were 0.541 and 0.341, respectively, for the first monomer, 0.455 and 0.532 for the second monomer, and 0.375 and 0.689 for the third monomer. It is important to note that previous attempts to solve the structure using the coordinates of closely related proteins and several molecular-replacement packages, including *MOLREP*, prior to determination of the tetragonal structure failed (data not shown). The molecular-replacement solution was then subjected to rigid-body refinement in *REFMAC*5 (Murshudov *et al.*, 1997) and converged to *R* and $R_{\rm free}$ factors of 0.242 and 0.268, respectively. Several cycles of restrained refinement, manual rebuilding and solvent-atom search improved the *R* and $R_{\rm free}$ factors to 0.195 and 0.24, respectively. Refinement of TLS parameters resulted in final *R* and $R_{\rm free}$ factors of 0.169 and 0.220, respectively. For confirmation of $Ca^{2+}$ ion binding sites, anomalous difference maps were calculated with the final phases and experimental anomalous differences.

The refinement statistics of the three structures are summarized in Table 2. The structures were validated using the *MolProbity* suite (Davis *et al.*, 2007). The validation results are shown in Table 2. The structures were deposited in the PDB and the PDB deposition codes are given in Table 2.

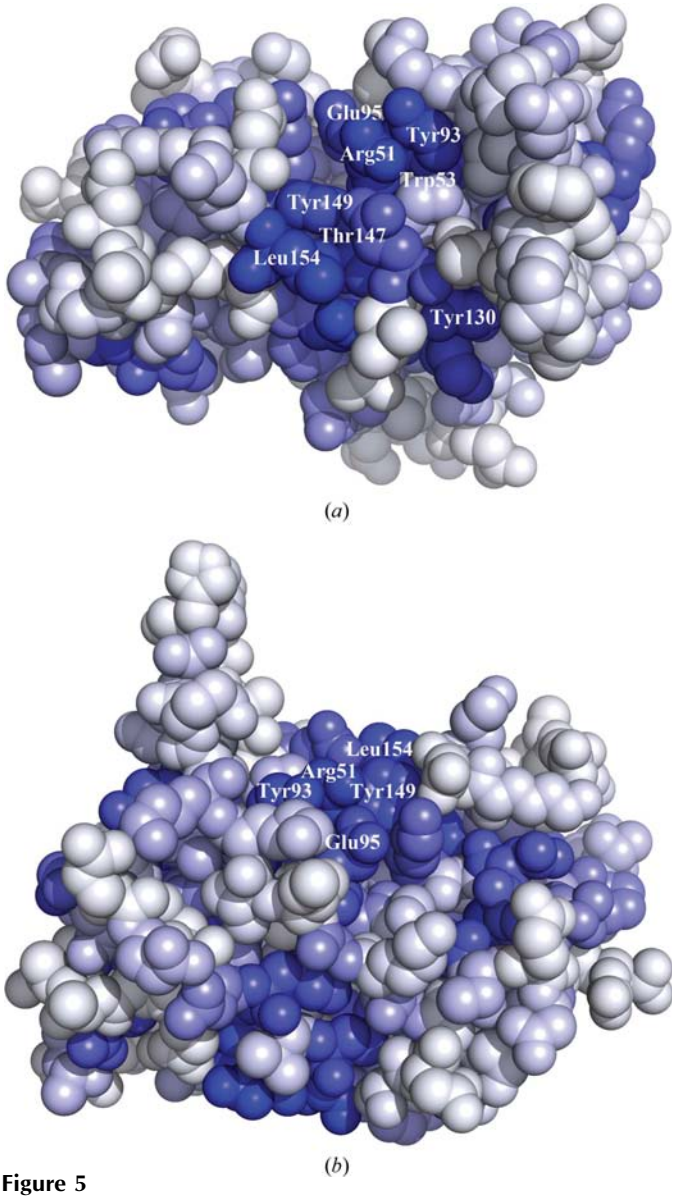


**Figure 5**
Space-filling representation of the shallow-groove region of CBM3b′ from *C. thermocellum* Cel9V. The degree of sequence conservation is shown as a blue gradient as in Fig. 3(*a*). (*a*) Upper view and (*b*) lateral view of the shallow groove.

## 3. Results and discussion

### 3.1. Overview

The 157-residue CBM3b′ module, spanning residues 730–887, of the cellulosomal cellulase Cel9V from *C. thermocellum* (UNP accession NO. A3DJ30; Fig. 1*c*) was cloned and expressed with an additional 11-residue C-terminal His tag, yielding a 168-residue protein molecule. The protein was crystallized by the hanging-drop vapour-diffusion method. Two crystal forms were obtained. Crystals of the first type diffracted to 2.0 Å resolution and belonged to the orthorhombic space group $C222_1$, with three molecules in the asymmetric unit. The second type of crystals diffracted to 1.79 Å resolution using an R-AXIS IV$^{++}$ area detector at Weizmann Institute of Science (WIS) and to 1.3 Å resolution at the European Synchrotron Radiation Source (ESRF), Grenoble, France. They belonged to the tetragonal space group $P4_12_12$, with one molecule in the asymmetric unit. The structure of the tetragonal crystals was determined by molecular replacement using the coordinates of CBM3a from CipA of *C. thermocellum* as a search model. The final atomic model was refined against the data to 1.3 Å resolution to crystallographic *R* and $R_{\rm free}$ factors of 0.123 and 0.156, respectively. The structure of the orthorhombic crystal was determined using the coordinates from the tetragonal crystal. The final *R* and $R_{\rm free}$ factors converged to 0.169 and 0.220, respectively. The final electron-density maps were of high quality in all three structures. In the 1.3 Å resolution structure the first five residues were not observed in the electron-density map. The first three residues in the 1.79 Å resolution tetragonal crystal and the first ten residues in the orthorhombic crystal were not observed in the electron-density map. However, the numbering of all the structures is the same as it starts according to the cloned sequence. The stereochemical quality of the structures is good, with over 96–98% of the residues in the most favourable regions of the Ramachandran plot (Table 2). Each structure contains one calcium ion and varying numbers of water molecules (Table 2).

### 3.2. Overall structure analysis of CBM3b′ of Cel9V and comparison with CBM3a of CipA and CBM3a of CipC

The 157 residues of *C. thermocellum* Cel9V CBM3b′ form a nine-stranded antiparallel $\beta$-sandwich (Fig. 2*a*). The 'bottom' $\beta$-sheet of the $\beta$-sandwich is formed by strands 1, 2, 7 and 4, while the 'top' $\beta$-sheet comprises relatively short antiparallel strands (5, 6, 3, 8 and 9; Figs. 2*a* and 2*c*). Both the N- and C-termini are located in the 'top' sheet region. The overall structure is similar to the related structures of *C. thermocellum* CipA CBM3a (Tormo *et al.*, 1996) and *C. cellulolyticum* CipC CBM3a (Shimon *et al.*, 2000; Figs. 2*b* and 2*c*) with a pairwise r.m.s.d. of about 1.5 Å between CBM3b′ of Cel9V, CBM3a of CipA and CBM3a of CipC.

The interface between the $\beta$-sheets is formed by 20 hydrophobic and eight aromatic side chains (Fig. 3*a*). The interacting hydrophobic residues were identified using the *Protein Interactions Calculator* (http://crick.mbu.iisc.ernet.in/~PIC/index.html) with a conventional contact criterion (Burley & Petsko, 1985). The residues establishing aromatic–aromatic interactions include the conserved Tyr52, which forms three separate edge-to-face interactions with the conserved Tyr17, with Phe54 and with the nonconserved Phe116. Phe54 forms an additional edge-to-face interaction with the conserved Tyr124. In addition, the aromatic core amino acids make multiple contacts with aliphatic residues, some of which are conserved between the CBM3a and CBM3b members. Residues Tyr52, Phe54 and Tyr124 also contact Ile29 and Phe98 contacts Leu45 (Fig. 3*a*). The conservation of the contacting hydrophobic core residues stresses the apparent importance of their precise arrangement for the overall structure.

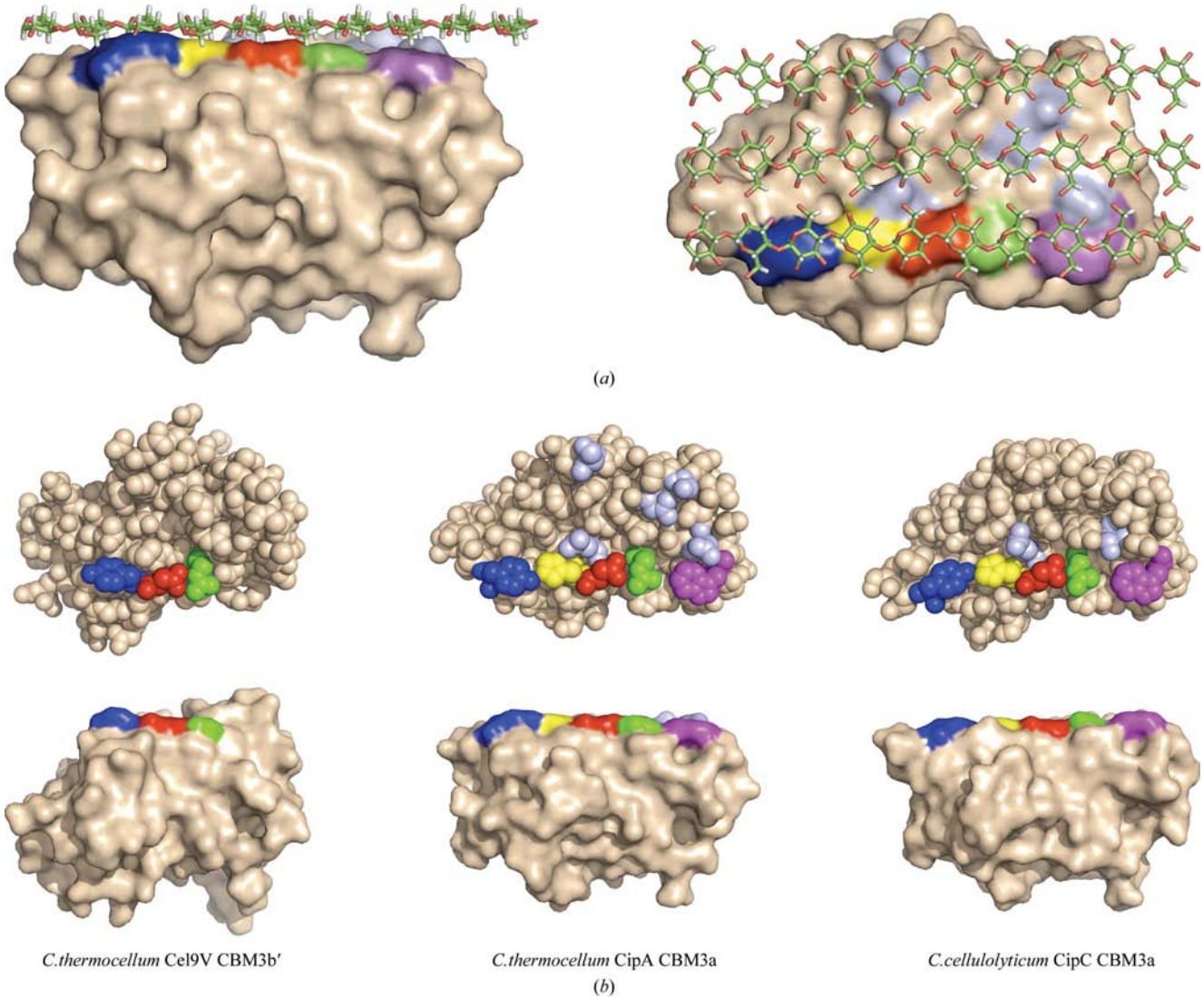


**Figure 6**
Cellulose-binding sites in CBM3s. (*a*) Putative binding of crystalline cellulose to the surface of *C. thermocellum* CipA CBM3a. Crystalline cellulose is shown as a stick model. Residues forming the cellulose-binding strip are highlighted by colouring: Arg, green; Asp, red; His, yellow; Trp, purple; Tyr, blue. The anchoring residues are shown in light blue. (*b*) Comparison of the putative cellulose-binding site of the three CBMs, CBM3b′ of Cel9V, CBM3a of CipA and CBM3c of CipC, in two mutually perpendicular views.

$\beta$-Strands are formed by 63 of 165 residues (about 37%) of CBM3b′. The remainder of the residues are in extended loops that contribute to the overall stability of the molecule by forming a network of interactions (Fig. 3*b*) and by binding a structural $Ca^{2+}$ ion.

### 3.3. $Ca^{2+}$ binding

In the CBM3b′ molecule one Ca atom was found in the region between strands 3 and 4 and was confirmed to be a $Ca^{2+}$ ion by its coordination and by its being the largest peak on the anomalous difference map in all three data sets. Its seven coordinating O atoms are the main-chain O and OG1 of Thr55, OD1 of Asp57, the main-chain O of Asp125, OD1 of Asp128, OD1 of Asp129 and a water molecule (Fig. 4*a*). Thr55 and Asp57 are located in a loop connecting $\beta$-strand 3 (from the top face) and $\beta$-strand 4 (from the bottom face). Asp25, Asp128 and Asp129 are located in the loop connecting $\beta$-strand 7 (from the bottom face) and $\beta$-strand 8 (from the top face). Thus, $Ca^{2+}$ binding appears to play a structural role, holding the two loops together and as a result tethering the upper and the lower $\beta$-sheets together (Fig. 4*b*). The residues forming this site are conserved among other known members of the CBM3 family, notably *C. thermocellum* CipA CBM3a (Tormo *et al.*, 1996), *T. fusca* Cel9A CBM3c (Sakon *et al.*, 1997) and CipC CBM3a from *C. cellulolyticum* (Shimon *et al.*, 2000) (Fig. 2*c*). The $Ca^{2+}$-binding sites are also structurally conserved: the r.m.s. differences between the CipA CBM3a, CipC CBM3a and Cel9V CBM3b′ $Ca^{2+}$-binding sites are very low at between 0.14 and 0.18 Å.

### 3.4. The 'shallow groove'

The 'upper' $\beta$-sheet layer of *C. thermocellum* Cel9V CBM3b′ has a structural feature called the 'shallow groove' that is highly conserved in all family 3 CBM-binding modules (Figs. 5*a* and 5*b*). The surface-facing residues which form the shallow groove are mostly tyrosines and polar side chains: Arg51, Trp53, Tyr93, Glu95, Tyr130, Thr147, Tyr149 and Leu154 from $\beta$-sheets 3, 6, 8 and 9. Nothing is known about the actual function of this site. One intriguing possibility is that the shallow groove may interact with peptide segments such as the highly conserved Pro/Thr-containing linkers of the scaffoldin and other cellulosomal components. It has been shown (Shimon *et al.*, 2000) that the groove region of the *C. cellulolyticum* CipC CBM3a is capable of binding proline residues from the linker of the adjacent symmetry-related molecule.

### 3.5. Cellulose-binding plane

It has been hypothesized (Tormo *et al.*, 1996) that the planar 'bottom' $\beta$-sheet surface of the *C. thermocellum* CipA CBM3a, which carries a linear array of aromatic residues and five non-aromatic 'anchoring' residues, interacts with the glucose rings of cellulose (Fig. 6*a*). The aromatic residues involved in *C. thermocellum* CipA CBM3a and *C. cellulolyticum* CipC CBM3a are Trp118, His57 and Tyr67. Additionally, Asp56 and Arg112 form a salt bridge, which presents a hydrogen-bonded ring that also can act in these surface interactions (Fig. 6*b*). In the Cel9V CBM3b′ protein there is a partial deletion in the planar strip (residues 60–69 of CBM3a; Figs. 2*b* and 2*c*), thus rendering it shorter compared with the cellulose-binding planar strips of the CipA and CipC CBM3a molecules (Fig. 6*b*). The putative cellulose-binding residue Tyr67 of CipA CBM3a is one of the deleted residues. Moreover, some of the designated cellulose-binding residues found on the surface are not conserved, *i.e.* Trp118 and His57 in CipA CBM are replaced by Ala121 and Tyr68 in Cel9V

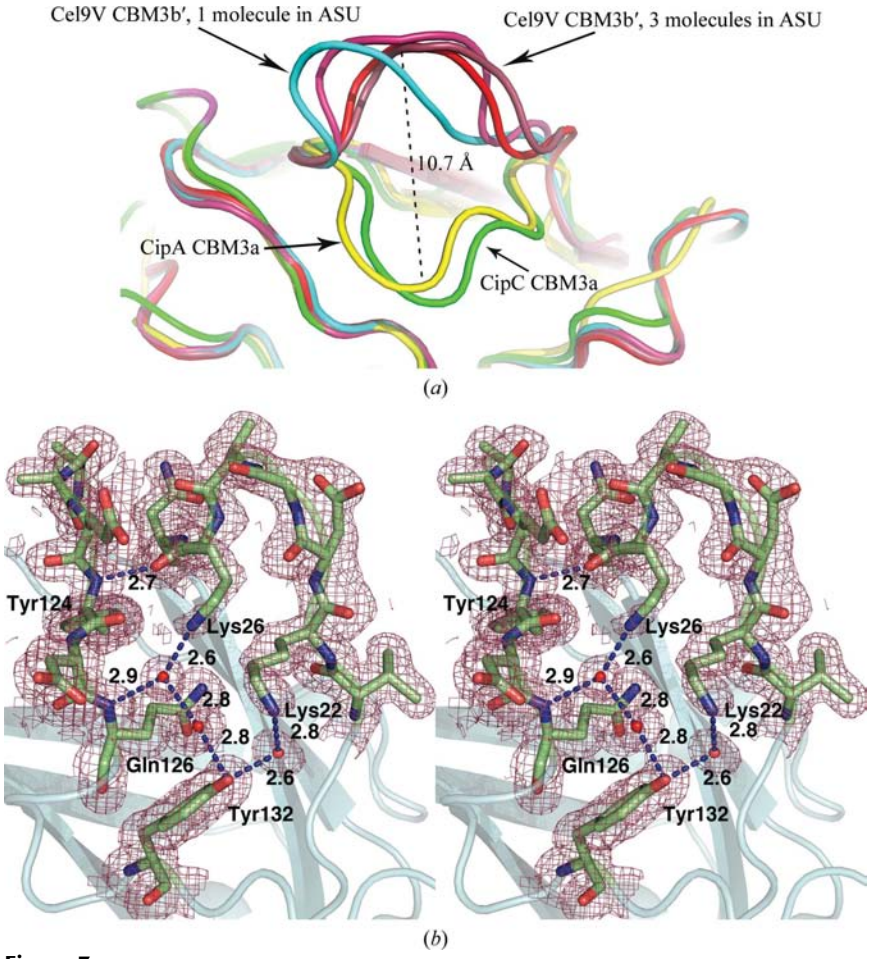


**Figure 7**
Loop shift forming a 'hump'. (*a*) Comparison of the $C^{\alpha}$ path of the 'hump'-forming loop of the three independent molecules of the $C222_1$ Cel9V CBM3b′ crystal (various reds), the $P4_12_12$ CBM3b′ crystal (cyan), CipC CBM3a (green) and CipA CBM3a (yellow). (*b*) Stereoview of the $2F_o - F_c$ electron density contoured at 0.7$\sigma$ overlaying the structure in the $\beta1$–$\beta2$ and $\beta7$–$\beta8$ loop regions of the high-resolution tetragonal structure (PDB code 2wnx).

CBM3b′. It has previously been shown (Jindou *et al.*, 2006) that site-directed mutation (A121W) of the nonconserved planar-strip residue in Cel9V CBM3b′ did not restore the cellulose-binding ability of Cel9V CBM3b′. This fact suggested that the loss of the conserved residues alone fails to explain the loss of cellulose-binding ability of this molecule.

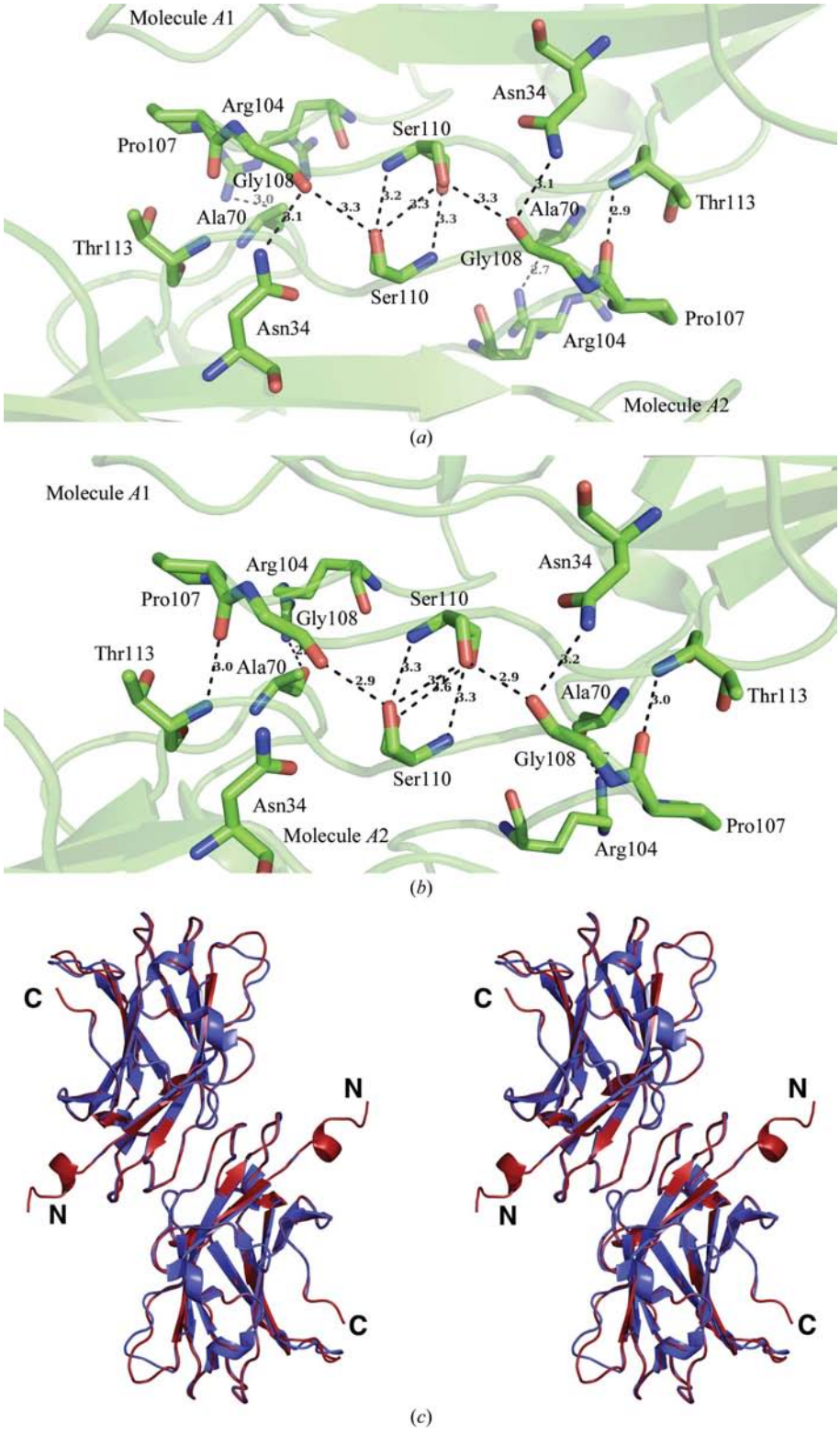


**Figure 8**
Contacts between the monomers related by the crystallographic twofold axis: (*a*) in the $P4_12_12$ crystal, (*b*) between the *A* monomers in the $C222_1$ crystal. (*c*) Superposition of cartoon representations of the $P4_12_12$ (brick red) and $C222_1$ (blue) crystallographic dimers.

Examination of the CBM3b′ structure revealed that the loop connecting the $\beta1$ and $\beta2$ strands is flipped by about 180°, producing a shift of approximately 10 Å in the $C^{\alpha}$ backbone relative to those of CipA and CipC CBM3a (Fig. 7*a*) and forming a 'hump' on the proposed cellulose-binding plane of this molecule. This loop in CBM3b′ is comprised of Val21, Lys22, Asp23, Gly24, Thr25 and Lys26. The 'hump' disrupts the planar topography of the cellulose-binding surface. This displacement would serve to prevent Cel9V CBM3b′ from forming stacking interactions with cellulose, thereby preventing binding. There are no insertions or deletions in this loop region, but the sequences differ (Fig. 2*c*). We can confidently rule out a crystal-packing or crystallization artifact as the source of this $C^{\alpha}$ shift as the 'hump' is observed in both of the crystal forms in which the CBM3b′ structure was determined and in four molecules (three noncrystallographically related protein molecules in one of the crystals and one in the other), regardless of the differences in their packing contacts (Fig. 7*a*). In all four CBM3b′ molecules two lysine residues (Lys22 and Lys26) protrude from this loop towards the loop connecting the seventh and eighth $\beta$-sheets (Fig. 7*b*). Lys22 forms a hydrogen bond to a water molecule, which in turn bonds to Tyr132. Lys26 forms a water-mediated hydrogen-bond connection to Tyr124 and to the main-chain N atom of Gln126. The extended hydrogen-bonded network stabilizes the $\beta1$–$\beta2$ and $\beta7$–$\beta8$ loops in an open conformation. This moves the $C^{\alpha}$ backbone of $\beta1$–$\beta2$ approximately 10 Å relative to its position in CBM3a from *C. thermocellum*, thereby forming a hump (Fig. 7*a*).

### 3.6. Crystal packing

Comparison of the crystal packing between crystallographically and non-crystallographically related molecules in different structures enables us to examine intermolecular interactions.

Although molecular interactions in a crystalline lattice do not necessarily have any immediate biological meaning, studying them contributes to a better understanding of protein–protein recognition, molecular recognition in crowded surroundings, crystallogenesis and protein surface chemistry. Analysis of the molecular-packing interfaces of the orthorhombic and tetragonal crystal forms using the *PISA* program (Krissinel & Henrick, 2007) showed that they have a similar interface. This interface is characterized by eight residues from molecule *A* (Figs. 8*a* and 8*b*) that form hydrogen bonds to eight twofold symmetry-related residues of the adjacent molecule; the dimers defined by these interfaces in both crystal forms are very similar (Fig. 8*c*).

Despite the similarity of these dimers, the general packing arrangements of the molecules in the orthorhombic $C222_1$ and tetragonal $P4_12_12$ crystals are different (Fig. 9). The layers formed in the orthorhombic structure by molecules *A* (red) and *B* (green) are interdigitized (Fig. 9*a*), while the *C* molecules (blue) form layers composed of only *C* molecules (Fig. 9*b*). Interestingly, the layers of *C* molecules do not form any hydrophobic and hydrophilic interactions with each other to form the sheets; rather, they are fixed only through their contacts with the *A* and *B* molecules. Molecules in the tetragonal structure form large rectangular pores (30 × 20 Å) that traverse the crystal along the fourfold axis.

The residues that participate in these packing interactions are not conserved in the family 3 CBMs, implying that they do not have a profound biological function in this group of proteins. The differences in the crystal packing apparently result from the influence of different precipitants (polyethylene glycol *versus* ammonium sulfate) on the crystallization.

## 4. Summary

The enzyme-borne CBM3b′ of Cel9V and the scaffoldin-borne CBM3as of CipA and CipC belong to different CBM3 subfamilies and are likely to have different functions in bacterial cellulose-degrading systems. Indeed, the discovery of cellulosomal cellulases such as *C. thermocellum* Cip9V that

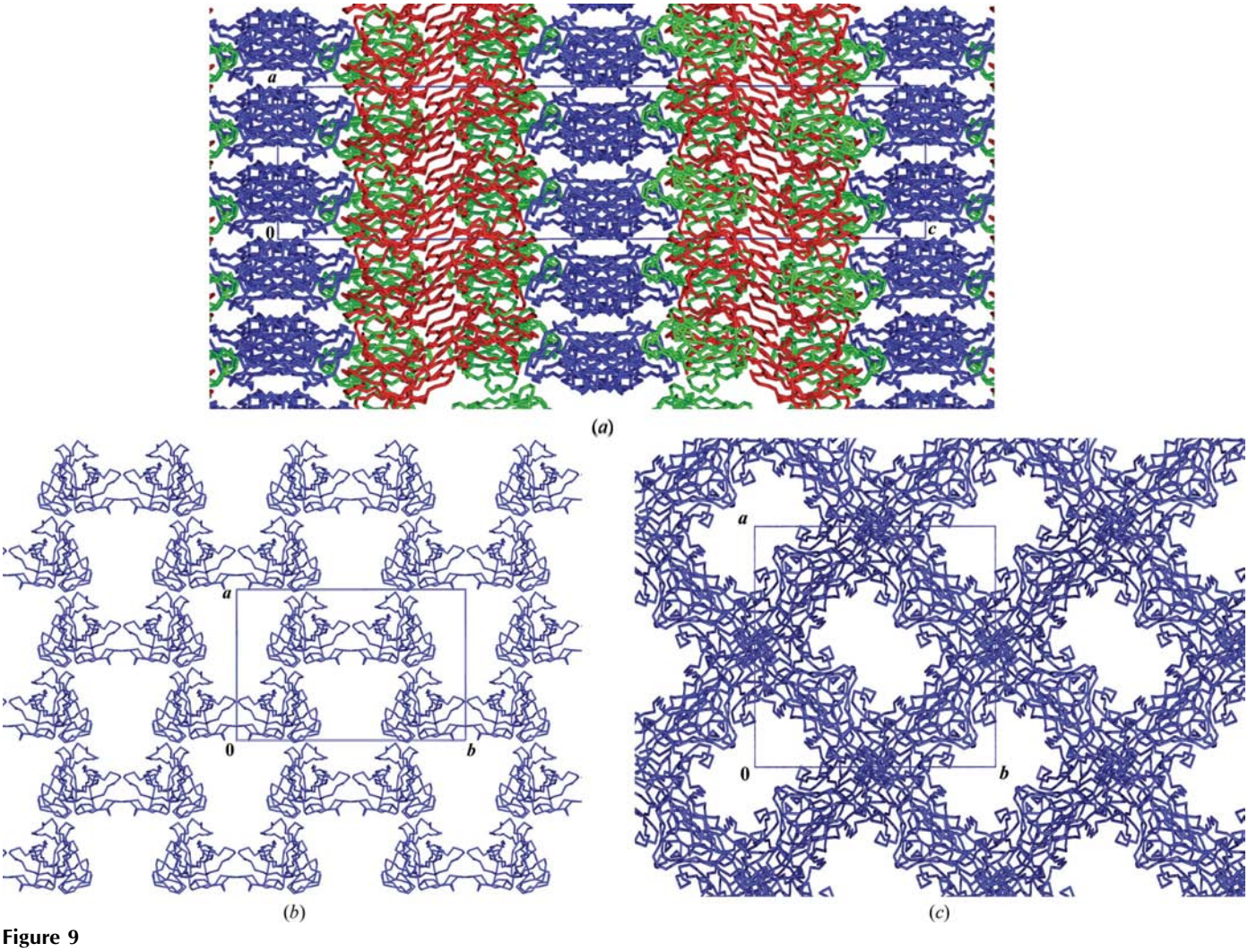


**Figure 9**
Packing diagrams of the $C222_1$ and $P4_12_12$ crystal structures. (*a*) Crystallographically independent molecules of the $C222_1$ structure are shown in different colours: *A* in red, *B* in green and *C* in blue. (*b*) Layer formed by the *C* molecules of the $C222_1$ structure. (*c*) Large pores formed in the $P4_12_12$ structure. The view is along the fourfold axis.

carry both a dockerin and a CBM3 was initially surprising. The lack of cellulose-binding capacity on the one hand and the high energy cost of producing additional modules on the other together imply that these CBMs are not actually cellulose-binding modules *per se*, but may possess some other as yet unknown function. For this reason it would be both interesting and informative to compare the topography of the cellulose-binding plane of the CBM3b′ of Cel9V with that of its more closely related 'classic' cellulose-binding CBM3 from subfamily 3b. Unfortunately, at present there is no molecular structure of such a cellulose-binding CBM3b available for comparison with CBM3b′. The proximity of the CBM3b′ to the CBM3c in *C. thermocellum* Cip9V and its closely related enzymes may suggest that this type of CBM serves to further modify the function of the parent GH9 enzyme, perhaps by combined action on a single glucan chain together with the adjacent CBM3c.

We thank the ESRF, Grenoble for use of the macromolecular crystallographic data-collection facilities and the ID14eh cluster staff for their assistance. This research was supported by the Israel Science Foundation (ISF; Grant Nos. 293/08, 966/09 and 159/07). EAB holds The Maynard I. and Elaine Wishner Chair of Bio-Organic Chemistry.

## References

Bayer, E. A., Shimon, L. J. W., Shoham, Y. & Lamed, R. (1998). *J. Struct. Biol.* **124**, 221–234.
Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. (2004). *Biochem. J.* **382**, 769–781.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Burley, S. K. & Petsko, G. A. (1985). *Science*, **229**, 23–28.
Cosier, J. & Glazer, A. M. (1986). *J. Appl. Cryst.* **19**, 105–107.
Davies, G. J., Tolley, S. P., Henrissat, B., Hjort, C. & Schulein, M. (1995). *Biochemistry*, **34**, 16210–16220.
Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B. III, Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). *Nucleic Acids Res.* **35**, W375–W383.
Deprez, C., Lloubes, R., Gavioli, M., Marion, D., Guerlesquin, F. & Blanchard, L. (2005). *J. Mol. Biol.* **346**, 1047–1057.
Din, N., Gilkes, N. R., Tekant, B., Miller, R. C., Warren, A. J. & Kilburn, D. G. (1991). *Biotechnology*, **9**, 1096–1099.
Ding, S. Y., Bayer, E. A., Steiner, D., Shoham, Y. & Lamed, R. (1999). *J. Bacteriol.* **181**, 6720–6729.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Gal, L., Gaudin, C., Belaich, A., Pages, S., Tardif, C. & Belaich, J. P. (1997). *J. Bacteriol.* **179**, 6595–6601.
Gilad, R., Rabinovich, L., Yaron, S., Bayer, E. A., Lamed, R., Gilbert, H. J. & Shoham, Y. (2003). *J. Bacteriol.* **185**, 391–398.
Gilkes, N. R., Claeyssens, M., Aebersold, R., Henrissat, B., Meinke, A., Morrison, H. D., Kilburn, D. G., Warren, R. A. & Miller, R. C. Jr (1991). *Eur. J. Biochem.* **202**, 367–377.
Henrissat, B. (1997). *Mol. Microbiol.* **23**, 848–849.
Irwin, D., Shin, D. H., Zhang, S., Barr, B. K., Sakon, J., Karplus, P. A. & Wilson, D. B. (1998). *J. Bacteriol.* **180**, 1709–1714.
Jindou, S., Xu, Q., Kenig, R., Shulman, M., Shoham, Y., Bayer, E. A. & Lamed, R. (2006). *FEMS Microbiol. Lett.* **254**, 308–316.
Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.
Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007). *Bioinformatics*, **23**, 2947–2948.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Acta Cryst.* D**61**, 458–464.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Pagès, S., Belaich, A., Tardif, C., Reverbel-Leroy, C., Gaudin, C. & Belaich, J. P. (1996). *J. Bacteriol.* **178**, 2279–2286.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Sakon, J., Irwin, D., Wilson, D. B. & Karplus, P. A. (1997). *Nature Struct. Biol.* **4**, 810–818.
Shimon, L. J. W., Pagès, S., Belaich, A., Belaich, J.-P., Bayer, E. A., Lamed, R., Shoham, Y. & Frolow, F. (2000). *Acta Cryst.* D**56**, 1560–1568.
Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Acta Cryst.* D**60**, 432–438.
Stout, G. H. & Jensen, L. H. (1968). *X-ray Structure Determination. A Practical Guide.* London: MacMillan.
Teng, T.-Y. (1990). *J. Appl. Cryst.* **23**, 387–391.
Tomme, P., Driver, D. P., Amandoron, E. A., Miller, R. C., Antony, R., Warren, J. & Kilburn, D. G. (1995). *J. Bacteriol.* **177**, 4356–4363.
Tormo, J., Lamed, R., Chirino, A. J., Morag, E., Bayer, E. A., Shoham, Y. & Steitz, T. A. (1996). *EMBO J.* **15**, 5739–5751.
Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.
Winn, M. D. (2003). *J. Synchrotron Rad.* **10**, 23–25.
Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). *Acta Cryst.* D**57**, 122–133.
Winn, M. D., Murshudov, G. N. & Papiz, M. Z. (2003). *Methods Enzymol.* **374**, 300–321.