Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Fine-structural variance of family 3 carbohydrate-binding modules as extracellular biomass-sensing components of *Clostridium thermocellum* anti-σ^I factors

Oren Yaniv,[a,b] Galit Fichman,[a] Ilya Borovok,[a] Yuval Shoham,[c] Edward A. Bayer,[d] Raphael Lamed,[a,b] Linda J. W. Shimon[e]* and Felix Frolow[a,b]*

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

Correspondence e-mail: linda.shimon@weizmann.ac.il, mbfrolow@post.tau.ac.il

Received 2 September 2013
Accepted 23 October 2013

PDB references:
0059CBM3b, 4b9c;
0267CBM3b, 4b9p;
0404CBM3b, 4b97

The anaerobic, thermophilic, cellulosome-producing bacterium *Clostridium thermocellum* relies on a variety of carbohydrate-active enzymes in order to efficiently break down complex carbohydrates into utilizable simple sugars. The regulation mechanism of the cellulosomal genes was unknown until recently, when genomic analysis revealed a set of putative operons in *C. thermocellum* that encode σ^I factors (*i.e.* alternative σ factors that control specialized regulon activation) and their cognate anti-σ^I factor (RsgI). These putative anti-σ^I-factor proteins have modules that are believed to be carbohydrate sensors. Three of these modules were crystallized and their three-dimensional structures were solved. The structures show a high overall degree of sequence and structural similarity to the cellulosomal family 3 carbohydrate-binding modules (CBM3s). The structures of the three carbohydrate sensors (RsgI-CBM3s) and a reference CBM3 are compared in the context of the structural determinants for the specificity of cellulose and complex carbohydrate binding. Fine structural variations among the RsgI-CBM3s appear to result in alternative substrate preferences for each of the sensors.

1. Introduction

Biomass is the term used to designate the biological material from a given ecosystem. It includes wood, herbaceous plants, algae and a high percentage of man-made disposable material. The carbohydrate components of biomass can be utilized as a renewable source of energy, but need to be converted to more utilizable simple sugars. This breakdown is obstructed by the structural and chemical complexity of the biomass environment. The cellulose found in biomass may be either crystalline, amorphous or fibrous and is tightly associated with other plant cell-wall components such as hemicelluloses, pectin and lignin. Xylan, the most abundant hemicellulose, is often branched and decorated with sugars with attached acetate, arabinofuranose and methyl glucuronic acid side groups. In addition, arabinofuranose on xylan may be esterified by aromatic acids such as ferulic acid and may participate in lignin–hemicellulose cross-linkages. The diverse nature of biomass carbohydrate composition requires the availability of a diverse and variable mixture of enzymes for efficient degradation. The collection of necessary enzymes required to break down the specific biomass carbohydrate composition must therefore be regulated. Until recently, the mechanism by which the production of these enzymes is regulated was unknown.

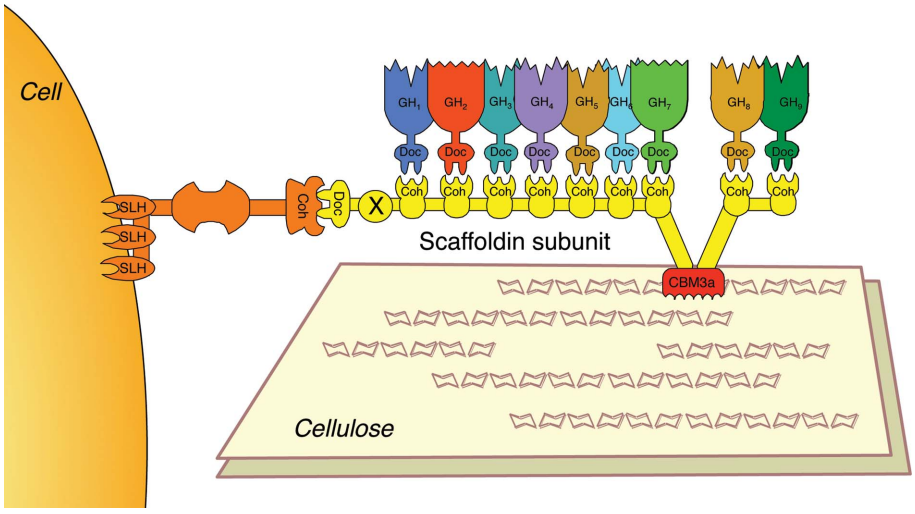


Figure 1
Schematic representation of the *C. thermocellum* cellulosome. The cellulosome is an extracellular multienzyme assembly attached to the cell wall *via* a surface layer homology (SLH)-like module of an anchoring scaffoldin. An assortment of the 70 possible glycoside hydrolases (GHs) can be incorporated into the cellulosome by binding to only nine available type I cohesin modules through the cohesin–dockerin interaction. The enzyme-borne dockerin (Doc) modules exhibit essentially the same specificity and can bind to any of the nine homologous cohesins (Coh). The cellulose-binding module (CBM3a) of the primary scaffoldin binds the cellulosome complex and the attached cell to the cellulosic substrate

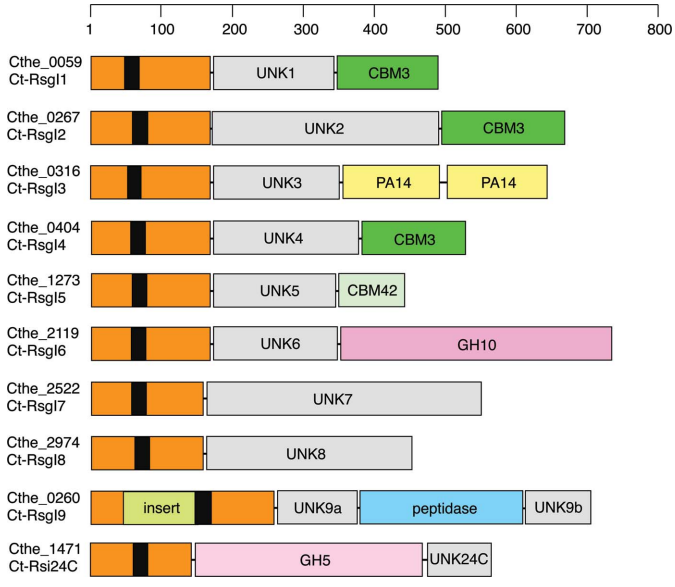


Figure 2
Schematic representation of the domain organization of RsgI-like proteins in *C. thermocellum*. Each of the multi-modular proteins has an RsgI-like domain (orange) with an N-terminal transmembrane region (black). Additional modules are marked in different colours as follows: GH10, family 10 glycoside hydrolase, pink; GH5, family 5 glycoside hydrolase, pink; CBM3, green; CBM42, mint; PA14, a conserved domain proposed to have a sugar-binding function, yellow; UNK, divergent domains of unknown function/s having no sequence identity to other proteins, grey; peptidase, a module sharing amino-acid sequence identity with proteins having domains homologous to family S1 (chymotrypsin)/family S6 (Hap) serine peptidases, cyan. The linkers between the modules are indicated as lines. The ruler above indicates the length of the proteins (number of amino acids). GH5, GH10, CBM3, CBM42 and PA14 have confirmed carbohydrate-active function.

The Gram-positive thermophilic anaerobic bacterium *Clostridium thermocellum* produces a highly efficient multienzyme cellulolytic system called a cellulosome. The cellulosome is capable of converting a wide variety of plant-derived polysaccharides directly into soluble sugars and fermentation products (Lamed, Setter, Kenig *et al.*, 1983; Lamed *et al.*, 1983; Bayer *et al.*, 2004; Fontes & Gilbert, 2010). The heart of the *C. thermocellum* cellulosome is the noncatalytic polypeptide scaffoldin CipA (Gerngross *et al.*, 1993), which includes nine repeated homologous modules called cohesins that integrate the various catalytic subunits into the cellulosome. Each cohesin interacts with high affinity with a complementary dockerin module located on the catalytic subunit (Salamitou *et al.*, 1992). The scaffoldin also includes a single carbohydrate-binding module of family 3 (CBM3) that mediates the binding of the entire complex to the cellulosic substrate (Fig. 1).

Even though there are only nine cohesins per *C. thermocellum* cellulosome and the scaffoldin can integrate only nine catalytic subunits at a time, the *C. thermocellum* genome encodes more than 70 different dockerin-containing enzymatic components (Schwarz *et al.*, 2004). Presumably, the binding of available dockerin-containing enzymes to the cellulosome is governed by their relative amounts and by their affinities for the cohesins (Shoham *et al.*, 1999). Studies have shown that the composition of the cellulosomal proteins is regulated by substrate availability (Johnson *et al.*, 1982; Nochur *et al.*, 1990; Nochur, Saraswathy *et al.*, 1992; Nochur, Jacobson *et al.*, 1992; Raman *et al.*, 2009). Moreover, the expression of many cellulose-related genes is influenced by the presence of extracellular polysaccharides (Dror *et al.*, 2003; Zhang & Lynd, 2005; Stevenson & Weimer, 2005; Riederer *et al.*, 2011; Raman *et al.*, 2011). The different carbohydrates of the diverse and variable forms of biomass must in some way regulate the cellulose-related genes.

The regulation mechanism of the cellulosomal genes was unknown until recently, when genomic analysis revealed a set of putative bicistronic operons in *C. thermocellum* that encode σ^{I} factors (*i.e.* alternative σ factors that control specialized regulon activation) and their cognate anti-σ^{I} factor RsgI. The second gene in these operons encodes a multi-modular membrane-associated protein containing a σ^{I}-interacting domain (Pfam PF12791; *RsgI_N*) with one predicted transmembrane helix and a putative C-terminal sensor. The deduced amino-acid sequences of the *C. thermocellum* anti-σ^{I} factors bear strong similarity to that of *Bacillus subtilis* RsgI. While many RsgI-like proteins have more than 300 residues, high homology between them has only been shown for the

N-terminal portions of approximately 165 amino-acid residues. The products of at least seven of these operons are assumed to be involved in the regulation of cellulosomal gene expression (Kahel-Raifer *et al.*, 2010), as the RsgI-like proteins have at their C-termini an extracellular module apparently possessing functions related to the recognition of polysaccharides (Fig. 2, Table 1; Nataf *et al.*, 2010; Bahari *et al.*, 2011).

Table 1
Extracellular modules and the corresponding target polysaccharides of the anti-σ^{I}-factor RsgI proteins.

	Module name†	Target polysaccharides‡
Cthe_0059 (RsgI1)	Family 3b carbohydrate-binding module (CBM3b)	Crystalline cellulose
Cthe_0267 (RsgI2)	Family 3b carbohydrate-binding module (CBM3b)	Crystalline cellulose
Cthe_0316 (RsgI3)	PA14 motifs (two in tandem)	Pectin and related polysaccharides
Cthe_0404 (RsgI4)	Family 3b carbohydrate-binding module (CBM3b)	Crystalline cellulose
Cthe_1273 (RsgI5)	Family 42 CBM (CBM42)	Arabinose
Cthe_2119 (RsgI6)	Family 10 glycoside hydrolase (GH10)	Mainly xylans; low enzymatic activity
Cthe_1471 (Rsi24C)	Unusual family 5 glycoside hydrolase (GH5-like)	Cellulose; no hydrolytic activity

† According to the CAZy website (http://www.cazy.org; Cantarel *et al.*, 2009). ‡ As demonstrated previously (Kahel-Raifer *et al.*, 2010; Bahari *et al.*, 2011).

According to the current model (Kahel-Raifer *et al.*, 2010; Fig. 3), the presence of a particular polysaccharide biomass component is detected extracellularly by a corresponding RsgI-borne CBM or glycoside hydrolase (GH)-like element. As a result, the cognate σ^{I} factor is released and interacts with RNA polymerase to promote transcription of the target genes (*e.g.* those encoding cellulosomal enzymes related to carbohydrate utilization). The results of isothermal titration calorimetry (ITC) binding measurements between several σ^{I} factors and their corresponding anti-σ^{I} domains support this model (Nataf *et al.*, 2010). These measurements demonstrated specific interaction between the σ^{I}–anti-σ^{I} couples and confirmed that they function together. Moreover, the expression of the σ^{I}-factor genes was found to be upregulated in the presence of extracellular polysaccharides (cellulose and xylan; Nataf *et al.*, 2010). The recently published genomes of

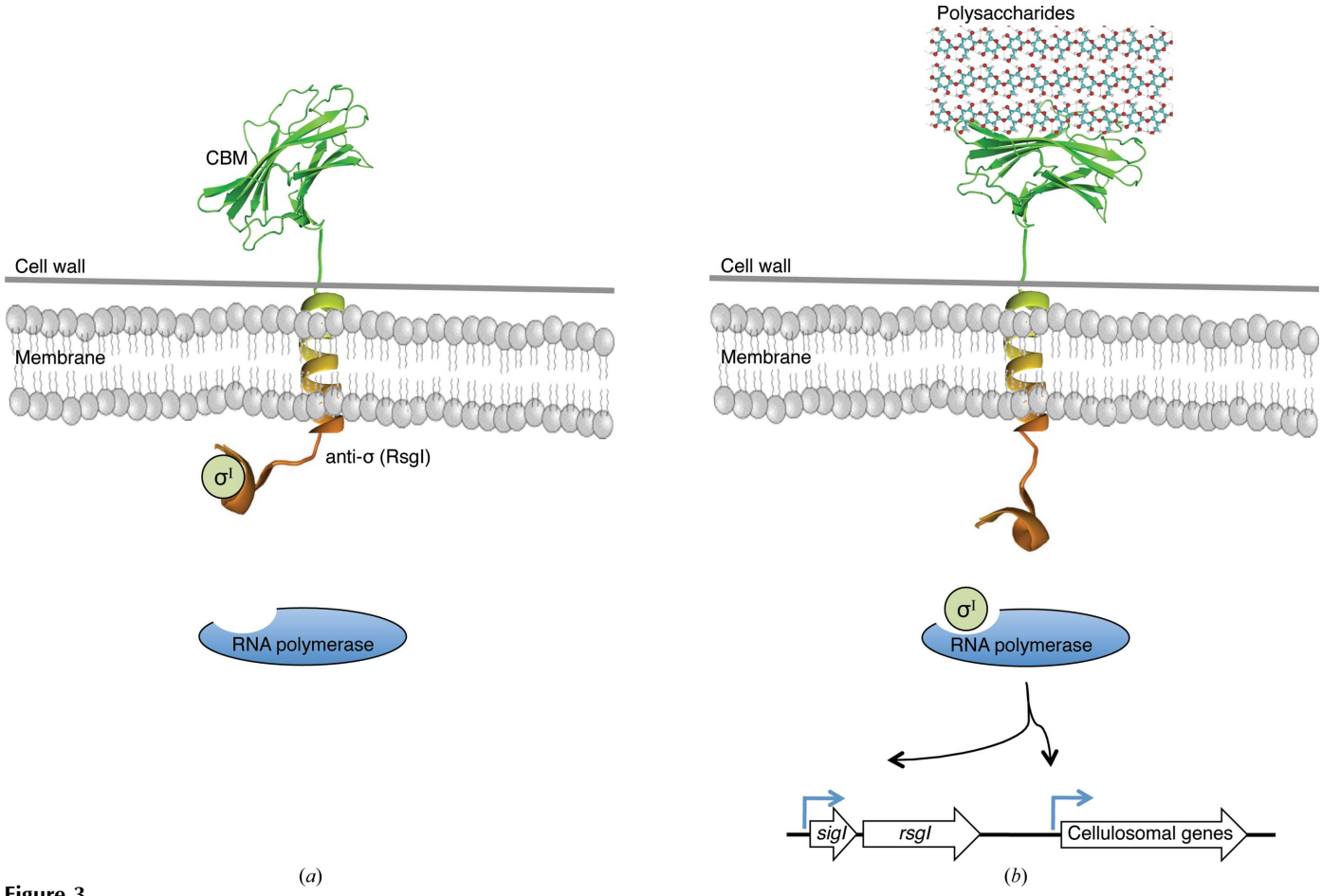


Figure 3
Proposed model of activation of σ^{I} factors. (*a*) Off state: in the absence of extracellular substrate, the cytosomal anti-σ^{I} domain interacts with its corresponding σ^{I} factor, thereby inactivating it. (*b*) On state: in the presence of target substrate, the extracellular C-terminal sensing domain (CBM) interacts with the target polysaccharide substrate, inducing a conformational change in the cytosomal anti-σ^{I} domain. As a result, the σ^{I} factor is released and interacts with RNA polymerase to promote transcription of the appropriate set of genes (encoding cellulosomal enzymes related to carbohydrate utilization).

Table 2
Oligonucleotide primers and restriction enzymes used in this study.

Construct	Nucleotide sequence	Restriction enzymes
0059CBM3b	F, 5′-GCACATATGGGGCTTAAAATACAA-TATTACAGC R, 5′-GCACTCGAGTCAGGGCTCAATTCC-CCATA	*Nde*I, *Xho*I
0267CBM3b	F, 5′-GGAATTCCATATGCCTACGGAAAC-TGACTTGTTAAC R, 5′-CGGGATCCTTAAGGTTCAATCCC-CAATATAG	*Nde*I, *Bam*HI
0404CBM3b	F, 5′-CGACCATGGGGAGTGTCAAAGTG-AGGTT R, 5′-CGAGAATTCTCATGGTTCTCTTCC-CCAT	*Nco*I, *Eco*RI

C. clariflavum (Izquierdo *et al.*, 2012) and *Acetivibrio cellulolyticus* (Dassa *et al.*, 2012) revealed similar putative biomass-sensing systems containing several family 3 CBMs.

Previous structural studies of the *C. thermocellum* CBM3s have focused on their function within cellulosomes (Jindou *et al.*, 2007; Yaniv, Petkun *et al.*, 2012) and scaffoldins (Tormo *et al.*, 1996), as well as on their role in some family 9 cellulase enzymes (Petkun *et al.*, 2010). The intriguing finding of three CBM3-like representatives that act as part of the proposed biomass-sensing mechanism initiated a comprehensive study with the purpose of discovering the structural basis of the substrate preferences of each of the RsgI-CBM3 modules RsgI1, RsgI2 and RsgI4 CBM3b.

2. Materials and methods

2.1. Cloning of CBM3b modules from anti-σ^{I} factors (RsgI) of *C. thermocellum*

DNA fragments encoding RsgI1 (Cthe_0059) CBM3b (GenBank accession No. YP_001036493), RsgI2 (Cthe_0267) CBM3b (GenBank accession No. YP_001036699) and RsgI4 (Cthe_0404) CBM3b (GenBank accession No. YP_001036835) CBM3b were amplified by PCR from *C. thermocellum* ATCC 27405 genomic DNA isolated as described by Murray & Thompson (1980) using specific primers (Table 2).

The PCR products were purified, cleaved with restriction enzymes (Table 2) and inserted into pET-28a(+) expression vector (Novagen, Madison, Wisconsin, USA). The obtained constructs pET-0059CBM3b and pET-0267CBM3b contained an N-terminal hexahistidine (His) tag and thrombin cleavage site. The pET-0404CBM3b construct was cloned without any additional His tag.

2.2. Protein expression and purification

Escherichia coli BL21 (DE3) cells harbouring the expression vector were aerated at 310 K in 4 l Terrific Broth supplemented with 25 µg ml^{-1} kanamycin for 16 h. The cells were harvested, resuspended in 50 m*M* phosphate buffer pH 8.0 containing 300 m*M* NaCl and sonicated in the presence of 10 µg ml^{-1} DNase.

The recombinant 0059CBM3b and 0267CBM3b used for the crystallization experiments were purified as described previously (Yaniv, Halfon *et al.*, 2012). The thrombin-cleaved proteins contained three additional residues (GSH) that originated from the cleavage site of the enzyme at their N-terminus.

For the recombinant 0404CBM3b, the sonicated material was heated for 30 min at 333 K and then centrifuged. The supernatant fluids were mixed with 40 m*M* Tris–HCl pH 7.5 in a 1:1 ratio and then loaded onto a 5 ml HiTrap Q Sepharose column (GE Healthcare) for anion-exchange chromatography. The 0404CBM3b was finally purified by fast protein liquid chromatography (FPLC) using a Superdex 75 16/60 column (GE Healthcare) and an ÄKTAprime system (GE Healthcare).

For the carbohydrate microarray assays (§2.6), another batch of proteins including 0404CBM3b now also containing an N-terminal hexahistidine tag were prepared. The purification steps were performed as described previously (Jindou *et al.*, 2007).

Protein purity was evaluated by SDS–PAGE (15%) and the gel was stained with Coomassie Brilliant Blue. The proteins were concentrated to 21, 10 and 23 mg ml^{-1} for 0059CBM3b, 0267CBM3b and 0404CBM3b, respectively, using Centriprep YM-3 centrifugal filter devices (Amicon Bioseparation, Millipore, Billerica, Massachusetts, USA). The protein concentration was determined by measuring the UV absorbance at 280 nm.

2.3. Crystallization and X-ray diffraction

The protein samples were screened for crystallization conditions using the hanging-drop vapour-diffusion method (McPherson, 1982) at 293 K. The initial crystallization conditions were optimized and the data and crystals are presented in Table 3 and Fig. 4. For 0267CBM3b crystals, a layer of 400 µl of oil (equal amounts of silicone and paraffin oils) was placed above the reservoir to slow the crystallization process (Chayen, 1997).

All crystals were harvested from the crystallization drops using a MiTeGen MicroMount (http://www.mitegen.com) made of polyimide and transferred for a short time into a cryostabilization solution mimicking the mother liquor supplemented with 18%(*w*/*v*) sucrose, 16%(*w*/*v*) glycerol, 16%(*w*/*v*) ethylene glycol and 4%(*w*/*v*) glucose. For data collection, crystals were mounted on the MiTeGen MicroMount, plunged into liquid nitrogen and placed in ESRF pucks for transfer and mounting at the synchrotron.

Diffraction data were measured on beamline ID29 at the European Synchrotron Radiation Facility (ESRF), Grenoble, France. A Pilatus 6M detector and X-ray radiation of 0.9762 Å wavelength were used. Diffraction data sets for 0059CBM3b, 0267CBM3b and 0404CBM3b consisting of 210, 165 and 200 images, respectively, were collected with 1° oscillation per image. The data were indexed and integrated with *DENZO* and scaled with *SCALEPACK* as implemented in *HKL*-2000 (v.0.98.703x; Otwinowski & Minor, 1997). During the scaling

Table 3
Crystallization-condition data.

	0059CBM3b (Fig. 4*a*)	0267CBM3b (Fig. 4*b*)	0404CBM3b (Fig. 4*c*)
Original condition	PEG/Ion 2 No. 2: 0.2 *M* sodium malonate pH 4.0, 20% PEG 3350	Crystal Screen 2 No. 27: 0.01 *M* $ZnSO_4.7H_2O$, 0.1 *M* MES pH 6.5 25% PEG MME 550 (protein at 5 mg ml^{-1})	Crystal Screen No. 9: 0.2 *M* ammonium acetate, 0.1 *M* trisodium citrate dihydrate, 30% PEG 400
No. of days	4 (291 K)	4 (298 K)	1 (291 K)
Optimized condition	0.2 *M* sodium malonate pH 4.0, 20% PEG 3350, 0.05 µg cellotriose (Sigma)	0.005 *M* $ZnSO_4.7H_2O$, 0.1 *M* MES pH 6.5, 23% PEG MME 550 (protein at 5 mg ml^{-1})	As original condition

of the diffraction data in *SCALEPACK*, Friedel pairs were kept separated in order to preserve the anomalous signal for further use. The 0059CBM3b, 0267CBM3b and 0404CBM3b crystals diffracted to 1.17, 1.18 and 1.28 Å resolution and belonged to space groups $P4_1$, $P2_12_12$ and $P2_12_12_1$, respectively. Matthews coefficients (Matthews, 1968) of 2.16, 2.10 and 1.81 Å^3 Da^{-1} calculated using the Matthews probability calculator (http://www.ruppweb.org/mattprob/; Kantardjieff & Rupp, 2003) suggested the presence of one molecule in the asymmetric unit, with solvent contents of 43.0, 41.4 and 33.1% for 0059CBM3b, 0267CBM3b and 0404CBM3b, respectively. The details of the X-ray data collection are presented in Table 4.

Table 4
Crystal and diffraction data.

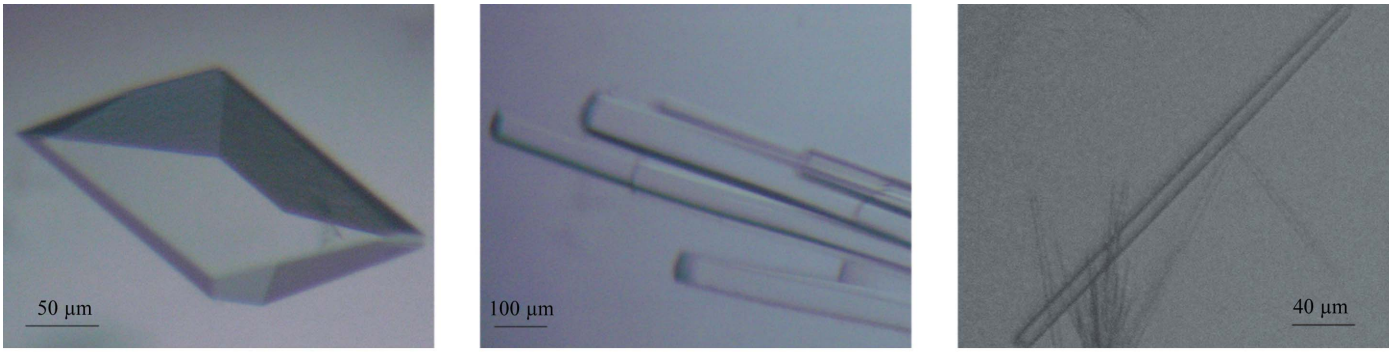
Values in parentheses are for the highest resolution shell.

	0059CBM3b	0267CBM3b	0404CBM3b
X-ray source	Beamline ID-29, ESRF	Beamline ID-29, ESRF	Beamline ID-29, ESRF
Space group	$P4_1$	$P2_12_12$	$P2_12_12_1$
No. of crystals	1	1	1
Total rotation angle (°)	210	165	200
Unit-cell parameters			
a (Å)	39.18	31.74	34.46
b (Å)	39.18	62.29	56.05
c (Å)	94.83	79.27	65.57
V (Å^3)	145610	156711	124710
No. of molecules in asymmetric unit	1	1	1
Resolution range (Å)	50–1.17 (1.19–1.17)	50–1.18 (1.20–1.18)	50–1.28 (1.30–1.28)
Total No. of reflections	171320	298808	227226
Unique reflections	48184	52558	33243
Mosaicity (°)	0.35–1.35	0.22–0.31	0.17–0.37
Multiplicity	3.57	5.68	6.83
Completeness (%)	97.5 (96.5)	99.9 (99.8)	99.8 (96.7)
Average $I/\sigma(I)$	13.7 (1.3)	16 (1.6)	30.8 (2.3)
R_{merge}†	0.084	0.093	0.056
B factor from Wilson plot (Å^2)	17.6	15.9	14.7

2.4. Structure determination and refinement

The structures of the CBM3b crystals were determined by molecular replacement (MR) employing *MOLREP* (v.11.0.05; Vagin & Teplyakov, 2010) or *Phaser* (v.2.1.4; McCoy *et al.*, 2007) from the *CCP*4 suite (v.6.3; Winn *et al.*, 2011). For determination of the structure of 0404CBM3b, the coordinates of Cel9V CBM3b (PDB entry 2wnx; Petkun *et al.*, 2010) from *C. thermocellum* were used as a search model in *MOLREP*. For determination of the structures of 0267CBM3b and 0059CBM3b, an ensemble composed of the coordinates of the newly determined structure of 0404CBM3b, ScaA CBM3b (PDB entry 3zqw; Yaniv, Petkun *et al.*, 2012) from *A. cellulolyticus* and CipA CBM3a (PDB entry 1nbc; Tormo *et al.*, 1996) from *C. thermocellum* was used as search model in *Phaser*.

The initial coordinates of 0404CBM3b and 0267CBM3b were subjected to several rounds of restrained refinement of positional and thermal parameters using *REFMAC*5 (Murshudov *et al.*, 2011), followed by manual building of the non-conserved side chains using *Coot* (Emsley *et al.*, 2010). The model building of 0059CBM3b was performed by *ARP*/*wARP* (v.7.2; Perrakis *et al.*, 1999). The structures were further refined with *PHENIX* (Adams *et al.*, 2010) and manually rebuilt using *Coot* until convergence. The corresponding R_{work} and R_{free} factors for the models were 0.169 and 0.185 for 0059CBM3b, 0.125 and 0.150 for 0267CBM3b and 0.139 and



(*a*) (*b*) (*c*)

Figure 4
Optimized cystals of (*a*) 0059CBM3b, (*b*) 0267CBM3b and (*c*) 0404CBM3b (see Table 3 for crystallization conditions).

0.156 for 0404CBM3b, respectively. The structures were validated using the *MolProbity* suite (Chen *et al.*, 2010) as implemented in *PHENIX*. The numerical details concerning the refinement of the three structures are summarized in Table 5. The refined structure and corresponding structure-factor amplitudes have been deposited in the PDB with accession codes 4b9c, 4b9p and 4b97 for 0059CBM3b, 0267CBM3b and 0404CBM3b, respectively.

2.5. Carbohydrate-binding assay based on SDS–PAGE

Qualitative carbohydrate-binding assays were performed as described previously (Xu *et al.*, 2004; Yaniv *et al.*, 2011). The protein samples were mixed with various carbohydrates (5 mg ml^{-1}): cellulose (Merck AG, Darmstadt, Germany), 1% amorphous cellulose (Lamed *et al.*, 1985), oat spelt xylan (arabinose $\leq$ 10% HPAE, glucose $\leq$ 15% HPAE, xylose $\geq$ 70% HPAE after hydrolosis; Sigma, St Louis, Missouri, USA), beech wood xylan (xylose $\geq$ 90% HPAE; Sigma), pectin from apple (70–75% esterification; Sigma), polygalacturonic acid (Sigma), neutral detergent fibre of alfalfa cell walls prepared according to Miron & Ben-Ghedalia (1993), banana fruit stem prepared according to VanSoest *et al.* (1991) and wheat straw prepared according to Miron *et al.* (2002). The alfalfa cell walls and wheat straw were a gift from Joshuah Miron, Institute of Animal Science, Bet Dagan, Israel. The banana fruit stem was a gift from Jose Francisco Calzada, ICAITI, Guatemala. Each assay was repeated at least three times.

Table 5
Summary of refinement statistics.

	0059CBM3b	0267CBM3b	0404CBM3b
No. of protein residues	150	166	151
No. of solvent atoms	167	214	154
No. of ions	1 (Ca^{2+})	4 (1 Ca^{2+}, 3 Zn^{2+})	1 (Ca^{2+})
R_{cryst}/R_{free} (%)	16.9/18.5	12.5/15.0	13.9/15.6
Geometry			
R.m.s.d., bonds (Å)	0.008	0.013	0.012
R.m.s.d., angles (°)	1.157	1.517	1.502
MolProbity validation			
Ramachandran favoured (%) (goal >98%)	99.3	98.3	99.4
Ramachandran outliers (%) (goal <0.2%)	0.0	0.0	0.0
Rotamer outliers (goal <1%)	1.5	0.0	0.0
C^{β} outliers (goal 0)	0	1	1
Clashscore†	4.43	4.70	5.32

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

2.6. Carbohydrate-binding assay based on microarrays

Carbohydrate microarrays containing 81 or 220 oligosaccharide samples were prepared by Professor W. G. T.

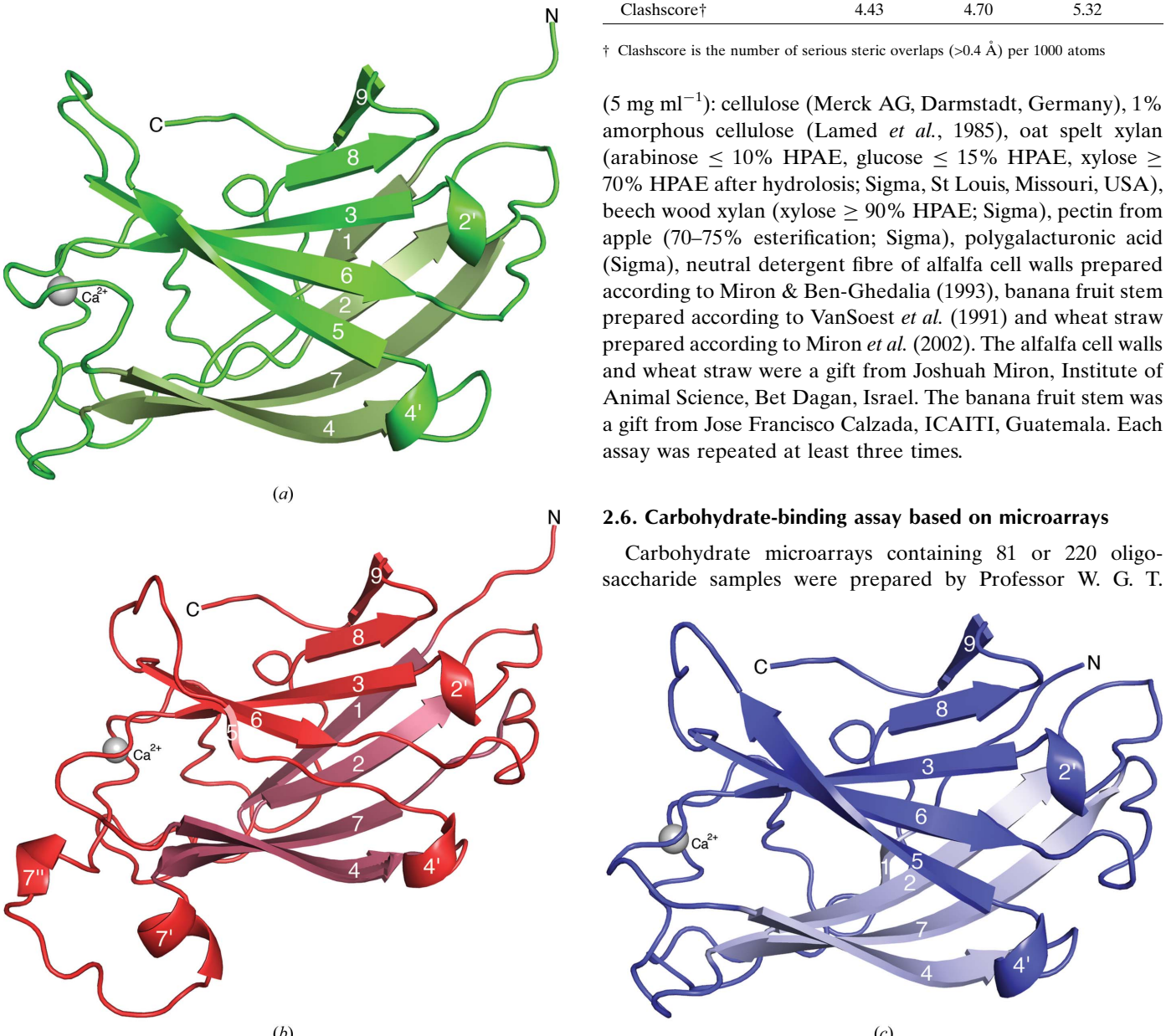


Figure 5
Overall structures of the RsgI-CBM3 modules. (*a*) 0059CBM3b, (*b*) 0267CBM3b, (*c*) 0404CBMb. A cartoon representation of the major secondary-structural elements (numbered according to sequence) is shown. The N- and C-termini are indicated. Calcium ions are shown as grey spheres. The top and bottom β-sheets are coloured in different shades.

Willats (Department of Plant Biology and Biotechnology University of Copenhagen, Denmark; Moller *et al.*, 2008). Binding assays for each of the RsgI-related CBM3s were performed as described previously (Cid *et al.*, 2010) with minor modifications as described below. Each assay was repeated between three and five times.

In brief, arrays were blocked in PBS (phosphate-buffered saline) containing 5% low-fat milk powder (5% MP-PBS) and were then incubated with a solution of one of the CBM3s (10 µg ml^{-1}) in 5% MP-PBS followed by three washes with PBS. The CBM3s were detected with anti-His secondary antibody conjugated to alkaline phosphatase (Sigma) diluted 1:10 000 in MP-PBS. The arrays were developed using SuperSignal West-Pico chemiluminescent substrate (Thermo Scientific Pierce Protein Research Products, Illinois, USA) according to the manufacturer's instructions.

3. Results and discussion

3.1. Overall structural analysis and comparison between the three structures

The RsgI-CBM3s exhibit structural characteristics typical of family 3 CBMs. They form a nine-stranded, antiparallel, immunoglobulin-like β-sandwich. The approximate overall dimensions of the reported molecules are 42 × 23 × 27 Å (as measured on the C$^{\alpha}$ skeleton), which is typical for family 3b CBMs. The two β-sheets of the sandwich are formed by strands 1, 2, 7 and 4 and by strands 5, 6, 3, 8 and 9 (Fig. 5). The first residue of the cloned 0404CBM3b protein and the first seven residues of the cloned 0267CBM3b protein were not visible in the electron-density maps.

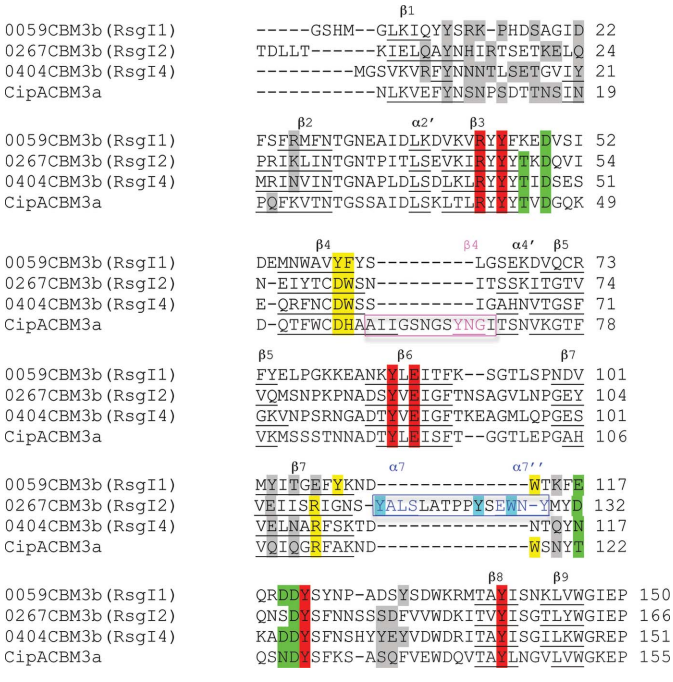


Figure 6
Structure-based sequence alignment of RsgI-CBM3 modules. Secondary-structural elements are underlined and labelled. The unique loop of 0267CBM3b is marked in a blue frame (residues 114–129). The unique helices are labelled in blue. The scaffoldin loop of *C. thermocellum* CipA CBM3a (PDB entry 1nbc) is framed in magenta (the unique β-strand is labelled in magenta). Proposed cellulose-binding residues are highlighted in yellow. Proposed cellulose-binding residues of the alternative site of 0267CBM3b are highlighted in light blue. Calcium-binding residues are shown in green. Anchoring residues are shown in grey. Residues that form the shallow groove are shown in red.

The amino-acid sequences of the three RsgI-CBM3s were aligned (Fig. 6) and analyzed. The most noticeable difference between the sequences is a 15-residue insertion of a unique loop in the sequence of 0267CBM3b (Fig. 6 and Supplementary Table S1[1]).

Superposition of the three RsgI-CBM3s with the cellulosomal CBM3a module of CipA scaffoldin from *C. thermocellum* (PDB entry 1nbc; Tormo *et al.*, 1996) resulted in relatively low r.m.s.d.s on C$^{\alpha}$ atoms of 0.988, 0.661 and 0.711 Å for 0059CBM3b, 0267CBM3b and 0404CBM3b (including 112, 113 and 118 out of 150, 151 and 166 atoms), respectively, as calculated on C$^{\alpha}$ positions using *PyMOL* (DeLano, 2002). The four compared CBMs are highly similar: one major difference between the RsgI-CBM3s and the classical cellulosomal CBM3a is the lack of the scaffoldin loop in the former (Fig. 6 and Supplementary Table S2).

In the structure of 0267CBM3b, the insertion and flanking residues form a protruding 15-residue loop (residues 114–129; Figs. 5 and 7) comprising two helical turns, $\alpha7'$ and $\alpha7''$ (indicated by a black ellipse in Fig. 7), located between β-strands 7 and 8, thereby creating a topological bulge ∼11 Å in height between the loop and the plane of the 'bottom' face.

The overall structures of 0059CBM3b and 0404CBM3b are similar (Figs. 5 and 7), with minor differences that can be specified as a shorter β-strand 1 and a longer β-strand 2 in 0404CBM3b in comparison to 0059CBM3b (pairwise r.m.s.d. of 0.738 Å for 106 of 150 residues as calculated using C$^{\alpha}$ atoms in *PyMOL*).

3.2. Ca^{2+}-binding site

The RsgI-CBM3s contain a solvent-accessible calcium-binding site. Calcium ions were found in the binding sites of all of the structures and were identified as calcium on the basis of coordination and anomalous difference Fourier maps (peak heights of 15.11σ, 26.89σ and 9.58σ for 0059CBM3b, 0267CBM3b and 0404CBM3b, respectively). Calcium is often found in carbohydrate-recognition domains. The sugar-binding calcium ions found in calcium-dependent (C-type) lectins such as mannose-binding protein (MBP) make direct hydrogen bonds to the sugar (Drickamer, 1999), whereas the calcium in concanavalin A arranges and stabilizes the geometry of the sugar-binding site [Kalb (Gilboa) & Helliwell, 2001]. However, the calciums in the CBMs reported here are located approximately 14 Å from the closest residue of the sugar-binding surface, which is too far for calcium-mediated coordination with the substrate. The calcium ions in these CBMs serve a purely structural role and tether the 'top' and 'bottom' β-sheets of the β-sandwich (Fig. 8).

[1] Supporting information has been deposited in the IUCr electronic archive (Reference: DZ5306).

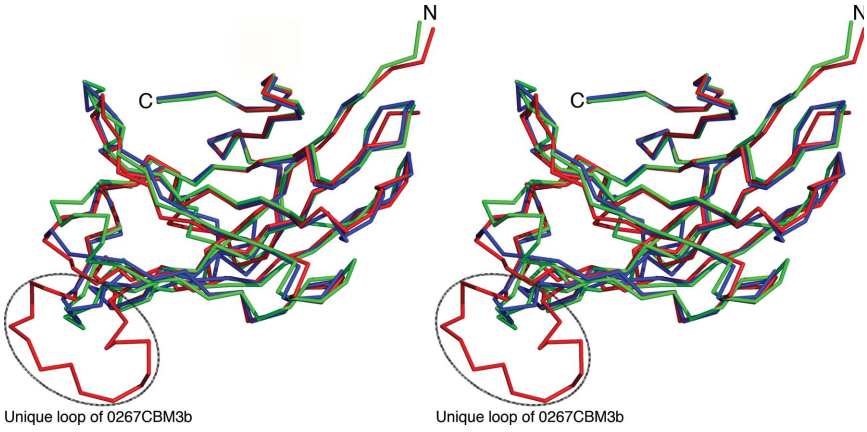


Figure 7
Stereo diagram of the superimposed C^{α} traces of the RsgI-CBM3 modules: 0059CBM3b, green; 0267CBM3b, red; 0404CBM3b, blue. The black ellipse denotes the unique loop in 0267CBM3b.

The calcium-binding sites of 0267CBM3b and 0404CBM3b are similar to each other, with an r.m.s.d. of 0.446 Å for the atoms of the coordination-sphere residues. The coordination of Ca^{2+} is formed by interactions with eight O atoms (five from side-chain groups, two from the main-chain carbonyl groups and one buried water; Figs. 8*b* and 8*c*). This Ca^{2+}-coordination sphere is similar to those previously observed in cellulosomal CBMs.

The calcium-binding site of 0059CBM3b differs from those of the other two RsgI-CBM3s (pairwise r.m.s.d.s of 1.56 and 1.62 Å to 0267CBM3b and 0404CBM3b, respectively) in the composition and geometry of the Ca^{2+}-coordination sphere (Fig. 8*a*). This site is also different from the calcium-binding sites of other family 3 CBMs with known structures. The calcium makes only four coordinative interactions with the protein (two side-chain O atoms and two main-chain carbonyl groups), and three water molecules complete the coordination sphere.

3.3. C-terminal interactions

The three RsgI-CBM3 modules are all located at the natural C-terminus of their respective multi-modular RsgI proteins and are positioned most distal from the cell wall.

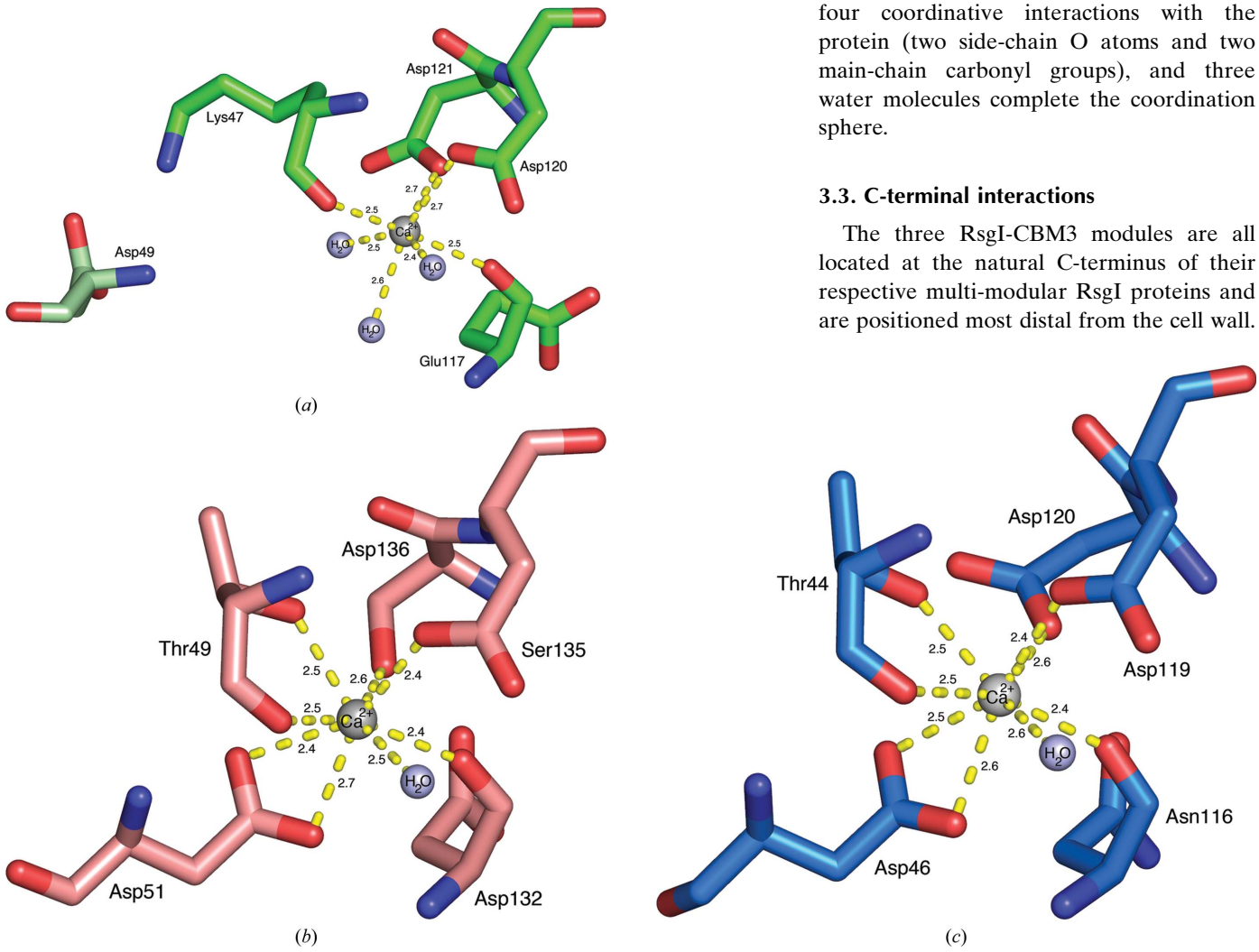


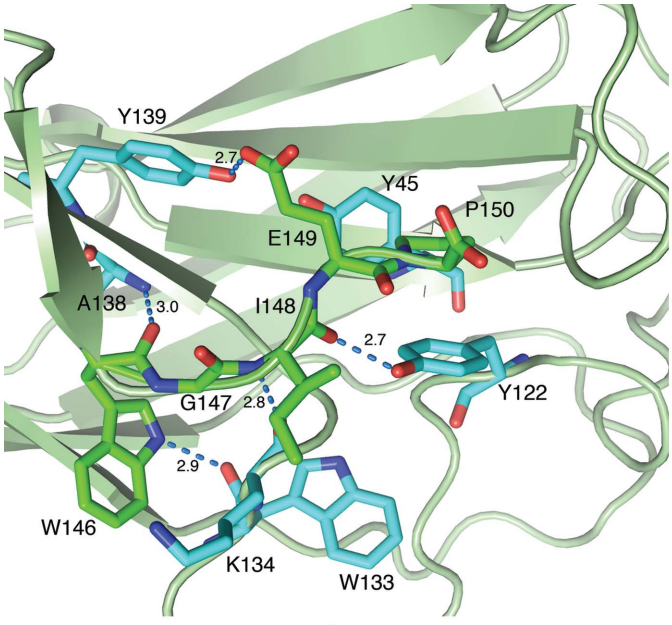
Figure 8
Ca^{2+}-binding sites in RsgI-CBM3 modules. (*a*) Coordination sphere of calcium in 0059CBM3b. (*b*) Coordination sphere of calcium in 0267CBM3b. (*c*) Coordination sphere of calcium in 0404CBM3b. Residues that coordinate Ca^{2+} are represented as sticks and coloured according to atom (O, red; N, blue). Calcium ions and water molecules are shown as spheres. Distances are in Å.

In this respect, they differ from the *C. thermocellum* CipA CBM3a, which is part of the multi-modular scaffoldin protein and is flanked on both sides by cohesin modules. The C-termini form a network of interactions that anchor the last residues of the proteins to the β-sandwich fold (Fig. 9).

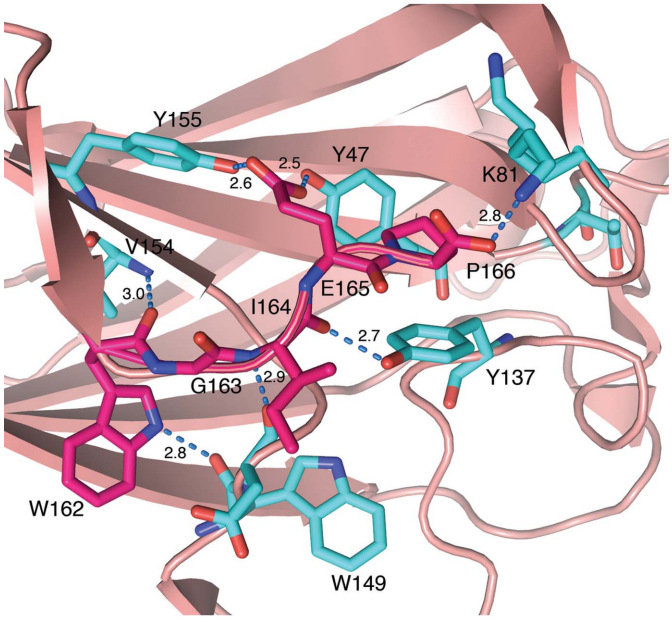
The last five residues in all three of the structures are nearly conserved. The last amino-acid residue in each of the proteins is a proline, which is sandwiched between two tyrosine rings. The tyrosine residues are held in position by hydrogen bonds to neighbouring residues in a network of interactions (Fig. 9). The C-terminal protein tail is clamped into position by overhanging residues into a complementary-shaped groove and is stabilized by van der Waals interactions and hydrogen bonds. The C-terminal proline amino-acid residues are tightly nestled against an amino-acid residue located on the loop between β-strand 5 and β-strand 6 (Fig. 9). Some of the residues participating in these anchoring interactions were previously identified as belonging to a shallow groove (Fig. 6, marked in red) that is highly conserved among cellulosomal CBMs. In the structure of *C. thermocellum* CBM3a-L (PDB entry 4jo5; Yaniv *et al.*, 2013), which included extended flanking linker residues, the secondary fold continues beyond the conserved proline residue (equivalent to the C-terminal proline residues in the above structures). However, the proline–threonine van der Waals interactions are maintained.

3.4. Carbohydrate binding

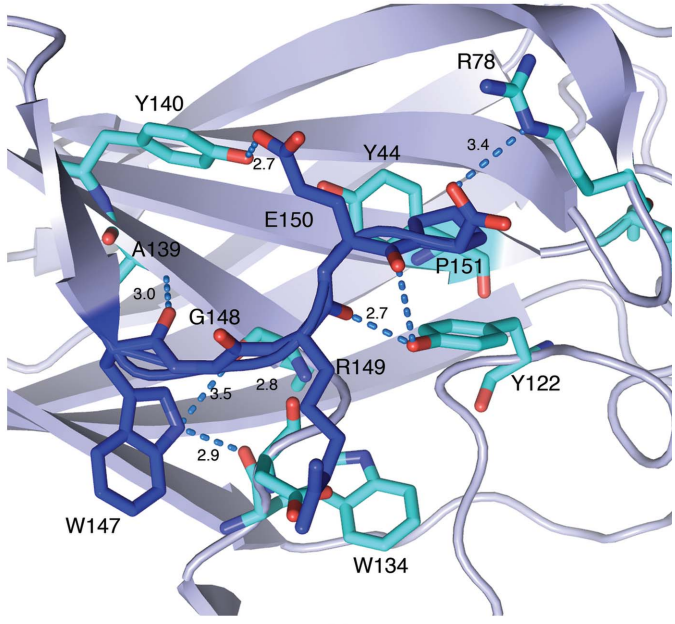
The RsgI-CBM3s show markedly different carbohydrate-binding preferences, most likely owing to differences in the lengths and amino-acid compositions of their carbohydrate-binding surfaces. The presented structures are family 3 CBMs, for which insoluble, highly crystalline cellulose is the primary binding target (Boraston *et al.*, 2004; Gilbert *et al.*, 2013). As such, the carbohydrate-binding properties of these molecules were analyzed for cellulose as the target substrate according to the accepted cellulose-binding requirements first proposed by Tormo (Tormo *et al.*, 1996; Shimon *et al.*, 2000; Petkun *et al.*, 2010). The formation of stable interactions between a CBM3 and cellulose requires (i) a planar face without topographic obstacles that enables extended proximity to the cellulose surface and (ii) a linear array of aromatic side chains on this planar face that is able to form stacking interactions with the cellulose.



(*a*)



(*b*)



(*c*)

Figure 9
The interactions in the C-terminal regions of the RsgI-CBM3 modules. (*a*) 0059CBM3b, (*b*) 0267CBM3b, (*c*) 0404CBMb. Cartoon representations presenting the interactions that maintain the C-terminal part of the module are shown.

In the carbohydrate-binding assays (see §§2.5 and 2.6), 0059CBM3b exhibits a broad specificity pattern and binds efficiently to crystalline and amorphous cellulose, xylan (from oat spelt) and three plant cell-wall extracts (Fig. 10). 0059CBM3b also binds to xyloglucan from two plant sources (tamarind and pea) and to galactan from lupin (data not shown). In 0059CBM3b, the bottom β-sheet comprising strands 1–2–7–4 is flat (Fig. 11*a*) and contains a long array

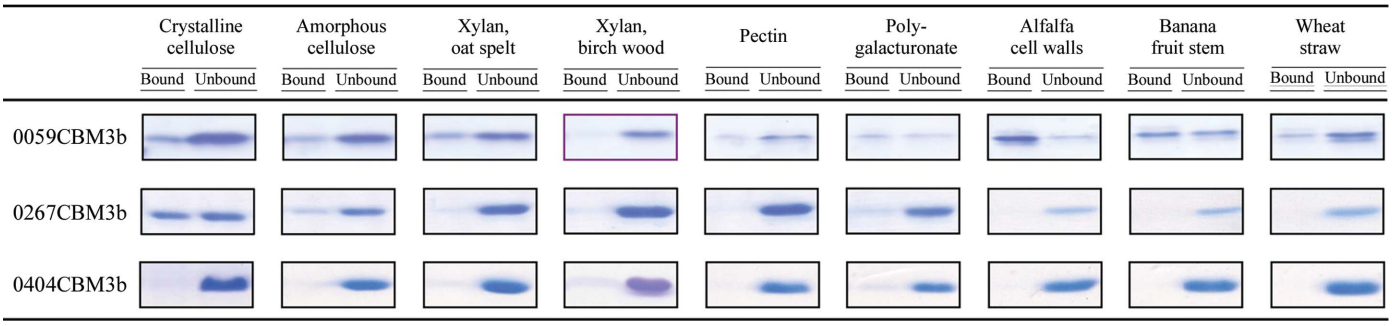


Figure 10
Carbohydrate-binding assay for RsgI-CBM3 modules. The partition of the CBM3 bands between the bound and unbound states is shown.

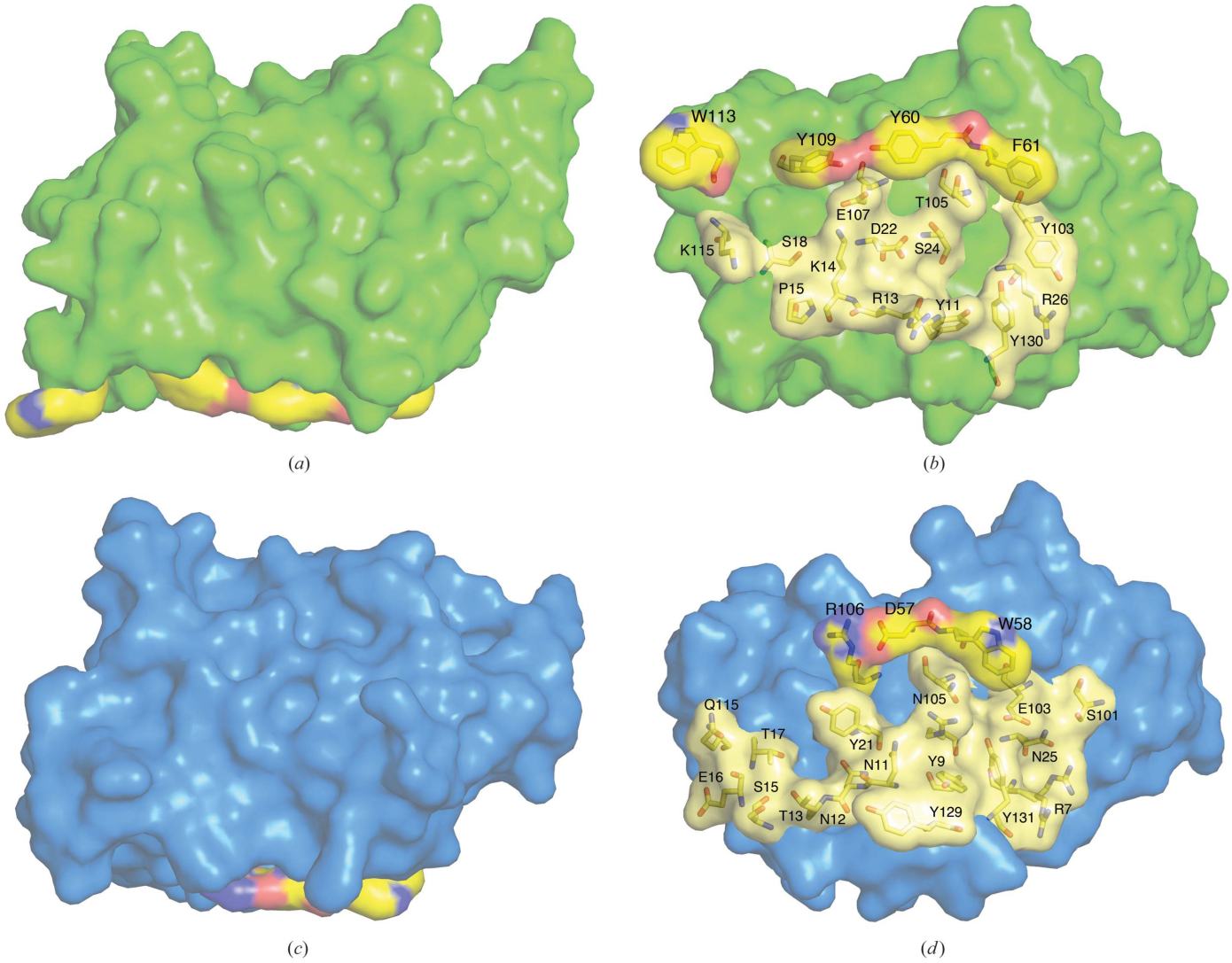


Figure 11
Cellulose-binding regions of the RsgI-CBM3 modules. (*a*, *b*) 0059CBM3b, (*c*, *d*) 0404CBMb. Each module is presented in two representations. (*a*, *c*) Lateral view of the cellulose-binding faces of the molecules. Putative cellulose-binding residues are shown in yellow and proposed anchoring residues are shown in pale yellow. (*b*, *d*) Surface representation of the proposed cellulose-binding region. Residues forming the cellulose-binding strip and anchoring residues are coloured by atom (C, yellow; O, red; N, blue) and numbered according to Fig. 6.

(32.5 Å in length) comprising four aromatic residues: Tyr60, Phe61, Tyr109 and Trp113 (Fig. 11*b*). This is the first known CBM3 that has four aromatic residues in the linear strip. Many of the other CBM3b modules have an Asp–Arg salt-bridge-forming pair that mimics an aromatic moiety in the linear strip (Tormo *et al.*, 1996; Shimon *et al.*, 2000).

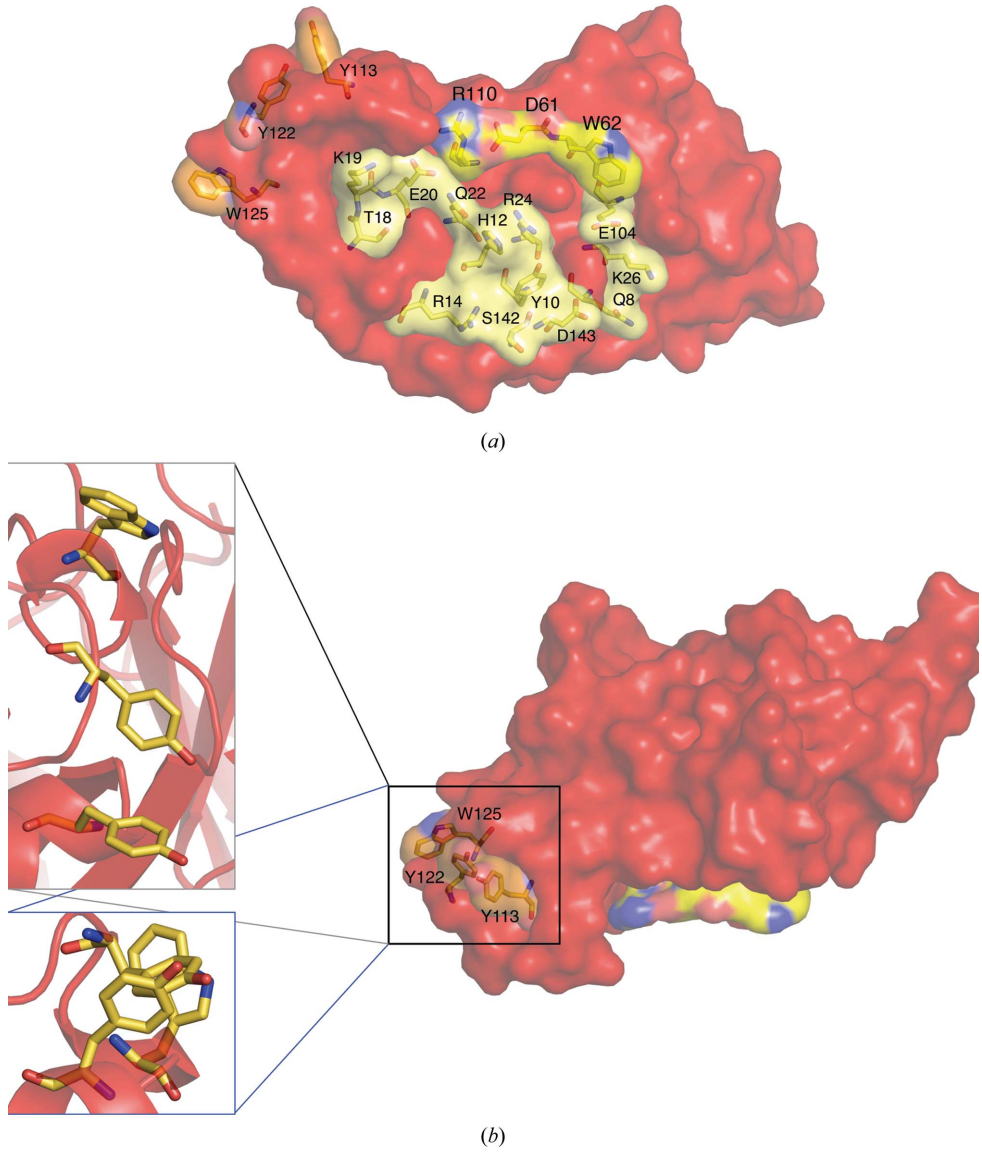


Figure 12
Proposed alternative cellulose-binding site in 0267CBM3b. Putative cellulose-binding residues are shown in shades of yellow: the cellulose-binding strip in yellow, the anchoring residues in pale yellow and the proposed alternate binding residues in amber. (*a*) Surface representation of the proposed cellulose-binding surface of 0267CBM3b. Putative cellulose-binding residues are shown in yellow and numbered according to sequence. (*b*) Magnification of the proposed alternative cellulose-binding site. Secondary-structure elements are represented as cartoons and proposed alternative cellulose-binding site residues are shown as sticks and numbered.

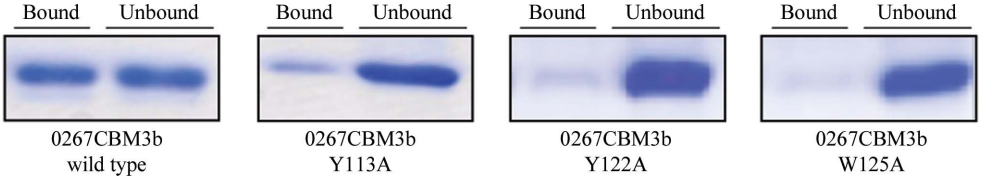


Figure 13
Crystalline cellulose-binding assay for 0267CBM3b variants. The partition of the CBM3 bands between the bound and unbound states is shown.

0404CBM3b only exhibited a very weak signal for xyloglucan (tamarind) binding (not shown) and did not bind any of the other approximately 250 carbohydrates and saccharides tested in our microarray assays. A possible explanation for the weak carbohydrate-binding signal that was measured is the shortness of the 0404CBM3b carbohydrate-binding linear strip (Fig. 11*c*). The 0404CBM3b structure reveals that a planar face alone is not sufficient for stable protein–cellulose interactions. It has been suggested that selectivity may also result from steric hindrance at the surface and the exclusion of some types of sugars from binding (Taylor & Drickamer, 2009). It is worth mentioning that site-specific mutagenesis (Asn113Trp) was performed in a separate study, thus providing an additional aromatic residue to the cellulose-binding linear strip. This mutant had increased cellulose-binding ability, as was shown previously for the *C. thermocellum* Cbh9A CBM3b, which also exhibited a short linear strip of aromatic residues (PDB entries 2ylk and 3zqx; Yaniv, Petkun *et al.*, 2012; data not shown).

0267CBM3b binds to both crystalline cellulose and amorphous cellulose (Lamed *et al.*, 1985) and also showed weak binding to xyloglucan from two plant sources (tamarind and pea) and to galactan from lupin. The linear strip of the 'cellulose-binding' 0267CBM3b is actually identical to that of the non-cellulose-binding 0404CBM3b (16.5 Å in length, composed of Trp62, Asp61 and Arg110; Fig. 12*a*). The differences in carbohydrate specificity may arise from the fact that the flat cellulose-binding face of 0267CBM3b is adjacent to the protruding loop

discussed above (residues 114–128; Figs. 5 and 7). This structure suggests a possible alternative mode of interaction between the protein and crystalline cellulose. In this context, the protruding loop is relatively rich in aromatic and proline rings and has two short α-helices. Its position at the edge of the binding plane would allow it to fit along the edge of cellulose crystals, not unlike a carpenter's edge guide (Fig. 12*b*). A binding site that resembles the proposed alternative site of 0267CBM3b was recently suggested by Cai *et al.* (2011) in the CBM3d from the Cel48A exoglucanase of *Cellulosilyticum ruminicola*. The Cel48A CBM3d has 24% homology to the CBM3b family and has a proposed cellulose-binding site located at the far end of the bottom face of the module. This site resembles the suggested binding site of 0267CBM3b in its special location and residue composition.

The detection of an RsgI-CBM3 that binds cellulose but fails to fulfil the two accepted requirements for cellulose binding (it does not contain a flat face or a sufficient linear strip) prompted a search for possible alternative cellulose-binding sites. In the case of 0267CBM3b the proposed cellulose-binding site of the unique loop, which comprises at least three aromatic residues (Tyr113, Tyr122 and Trp125; Fig. 12), may fit such a function. Alanine scanning of the three aromatic residues by site-specific mutagenesis resulted in a significant decrease in the cellulose-binding abilities of mutants modified at each of the three residues (Fig. 13). It can therefore be determined that the three aromatic residues on the unique loop of 0267CBM3b are functional residues for binding.

It has previously been noted (Tormo *et al.*, 1996) that the spacing between the aromatic rings of the linear strip is a multiple of that between the pyranoside rings along a single cellulose chain (approximately 5.2 Å from pyranoside to pyranoside; Simon *et al.*, 1988). In this context, the length of the alternative cellulose-binding site of 0267CBM3b is approximately 19 Å and falls between the 15.6 or 20.8 Å expected for protein–cellulose interactions formed between three or four consecutive pyranoside rings.

In addition to the aromatic stacking interactions *via* the planar strips, the CBMs present from their flat face a large number of both uncharged polar residues and charged residues. These can be considered to anchor the CBM to its sugar substrate through interactions with the sugar rings and sugar hydroxyl O atoms. These residues are arranged in two strips running parallel to the planar strip (Figs. 11 and 12) and are highly conserved both in the three CBMs presented here and in other family 3 CBMs (Tormo *et al.*, 1996; Shimon *et al.*, 2000).

4. Summary

The regulatory mechanism of the complicated cellulosomal system and its cellulose-related genes remained enigmatic for many years. Recently, a putative carbohydrate-biosensing system was discovered (Kahel-Raifer *et al.*, 2010; Nataf *et al.*, 2010). The premise of this system is that extracellular probe-like proteins detect polysaccharide components and subsequently transfer a signal into the cell, activating transcription of target genes. Three of those extracellular probes are classified as family 3 CBMs and have been shown to sense the presence of both crystalline and amorphous cellulose, xylan, xyloglucan and galactan from a variety of plant sources. Their high-resolution three-dimensional structures were determined in this work. The presence in *C. thermocellum* of numerous RsgI proteins with unique C-terminal CBMs with different carbohydrate-binding specificities suggests that each of them could be responsible for the regulation of a specific set of genes, which are thus activated as a consequence of the initial extracellular binding interaction with the target polysaccharide. The resultant complement of newly expressed enzymes and scaffoldins would then assemble into fresh cellulosomes and hydrolysis of the polysaccharide substrate would thus be modulated. Therefore, the biomass-sensing machinery of the bacterial cell would match the dynamic nature of the plant-wall cellulosic material during the deconstruction process.

We thank the ESRF, Grenoble, France for use of the macromolecular crystallographic data-collection facilities and the ID29 staff for their assistance. We thank Professor W. G. T. Willats from the Department of Plant Biology and Biotechnology University of Copenhagen, Denmark for his assistance with the carbohydrate arrays. This research was supported by the Israel Science Foundation (ISF; grant Nos. 715/12, 500/10 and 1349/13), by the Sidney E. Frank Foundation through the Israel Science Foundation (grant No. 24/11) and by grants from the US–Israel Binational Science Foundation (BSF), the Israeli Center of Research Excellence (I-CORE Center No. 152/11) Program of the Planning and Budgeting Committee, the Ministry of Environmental Protection and the Grand Technion Energy Program, comprising part of The Leona M. and Harry B. Helmsley Charitable Trust reporting on alternative energy series of the Technion – Israel Institute of Technology and the Weizmann Institute of Science. YS holds the Erwin and Rosl Pollak Chair in Biotechnology at the Technion – Israel Institute of Technology. EAB is the incumbent of The Maynard I. and Elaine Wishner Chair of Bio-Organic Chemistry at the Weizmann Institute of Science.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Bahari, L., Gilad, Y., Borovok, I., Kahel-Raifer, H., Dassa, B., Nataf, Y., Shoham, Y., Lamed, R. & Bayer, E. A. (2011). *J. Ind. Microbiol. Biotechnol.* **38**, 825–832.
Bayer, E. A., Belaich, J.-P., Shoham, Y. & Lamed, R. (2004). *Annu. Rev. Microbiol.* **58**, 521–554.
Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. (2004). *Biochem. J.* **382**, 769–781.
Cai, S.-C., Zheng, X. & Dong, X.-Z. (2011). *J. Bacteriol.* **193**, 5199–5206.
Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. (2009). *Nucleic Acids Res.* **37**, D233–D238.
Chayen, N. E. (1997). *J. Appl. Cryst.* **30**, 198–202.
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.

Cid, M., Pedersen, H. L., Kaneko, S., Coutinho, P. M., Henrissat, B., Willats, W. G. T. & Boraston, A. B. (2010). *J. Biol. Chem.* **285**, 35999–36009.
Dassa, B., Borovok, I., Lamed, R., Henrissat, B., Coutinho, P., Hemme, C. L., Huang, Y., Zhou, J. & Bayer, E. A. (2012). *BMC Genomics*, **13**, 210.
DeLano, W. L. (2002). *PyMOL*. http://www.pymol.org.
Drickamer, K. (1999). *Curr. Opin. Struct. Biol.* **9**, 585–590.
Dror, T. W., Morag, E., Rolider, A., Bayer, E. A., Lamed, R. & Shoham, Y. (2003). *J. Bacteriol.* **185**, 3042–3048.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Fontes, C. M. & Gilbert, H. J. (2010). *Annu. Rev. Biochem.* **79**, 655–681.
Gerngross, U. T., Romaniec, M. P. M., Kobayashi, T., Huskisson, N. S. & Demain, A. L. (1993). *Mol. Microbiol.* **10**, 1155.
Gilbert, H. J., Knox, J. P. & Boraston, A. B. (2013). *Curr. Opin. Sruct. Biol.* **23**, 669–677.
Izquierdo, J. A. *et al.* (2012). *Stand. Genomic Sci.* **6**, 104–115.
Jindou, S., Petkun, S., Shimon, L., Bayer, E. A., Lamed, R. & Frolow, F. (2007). *Acta Cryst.* F**63**, 1044–1047.
Johnson, E. A., Sakajoh, M., Halliwell, G., Madia, A. & Demain, A. L. (1982). *Appl. Environ. Microbiol.* **43**, 1125–1132.
Kahel-Raifer, H., Jindou, S., Bahari, L., Nataf, Y., Shoham, Y., Bayer, E. A., Borovok, I. & Lamed, R. (2010). *FEMS Microbiol. Lett.* **308**, 84–93.
Kalb (Gilboa), A. J. & Helliwell, J. R. (2001). *Handbook of Metalloproteins*, edited by A. Messerschmidt, R. Huber, T. Poulos & K. Weighardt. Chichester: John Wiley & Sons.
Kantardjieff, K. A. & Rupp, B. (2003). *Protein Sci.* **12**, 1865–1871.
Lamed, R., Kenig, R., Setter, E. & Bayer, E. A. (1985). *Enzyme Microb. Technol.* **7**, 37–41.
Lamed, R., Setter, E. & Bayer, E. A. (1983). *J. Bacteriol.* **156**, 828–836.
Lamed, R., Setter, E., Kenig, R. & Bayer, E. A. (1983). *Biotechnol. Bioeng. Symp.* **13**, 163–181.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
McPherson, A. (1982). *Preparation and Analysis of Protein Crystals*, 1st ed. New York: Wiley.
Miron, J. & Ben-Ghedalia, D. (1993). *Can. J. Microbiol.* **39**, 780–786.
Miron, J., Yosef, E., Ben-Ghedalia, D., Chase, L. E., Bauman, D. E. & Solomon, R. (2002). *J. Dairy Sci.* **85**, 89–94.
Moller, I., Marcus, S. E., Haeger, A., Verhertbruggen, Y., Verhoef, R., Schols, H., Ulvskov, P., Mikkelsen, J. D., Knox, J. P. & Willats, W. (2008). *Glycoconj. J.* **25**, 37–48.
Murray, M. G. & Thompson, W. F. (1980). *Nucleic Acids Res.* **8**, 4321–4325.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Nataf, Y., Bahari, L., Kahel-Raifer, H., Borovok, I., Lamed, R., Bayer, E. A., Sonenshein, A. L. & Shoham, Y. (2010). *Proc. Natl Acad. Sci. USA*, **107**, 18646–18651.
Nochur, S. V., Jacobson, G. R., Roberts, M. F. & Demain, A. L. (1992). *Appl. Biochem. Biotechnol.* **33**, 33–41.
Nochur, S. V., Roberts, M. F. & Demain, A. L. (1990). *FEMS Microbiol. Lett.* **71**, 199–204.
Nochur, S. V., Saraswathy, V., Demain, A. L. & Roberts, M. F. (1992). *Enzyme Microb. Technol.* **14**, 338–349.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Petkun, S., Jindou, S., Shimon, L. J. W., Rosenheck, S., Bayer, E. A., Lamed, R. & Frolow, F. (2010). *Acta Cryst.* D**66**, 33–43.
Raman, B., McKeown, C. K., Rodriguez, M. Jr, Brown, S. D. & Mielenz, J. R. (2011). *BMC Microbiol.* **11**, 134.
Raman, B., Pan, C., Hurst, G. B., Rodriguez, M. Jr, McKeown, C. K., Lankford, P. K., Samatova, N. F. & Mielenz, J. R. (2009). *PLoS One*, **4**, e5271.
Riederer, A., Takasuka, T. E., Makino, S., Stevenson, D. M., Bukhman, Y. V., Elsen, N. L. & Fox, B. G. (2011). *Appl. Environ. Microbiol.* **77**, 1243–1253.
Salamitou, S., Tokatlidis, K., Béguin, P. & Aubert, J. P. (1992). *FEBS Lett.* **304**, 89–92.
Schwarz, W. H., Zverlov, V. V. & Bahl, H. (2004). *Adv. Appl. Microbiol.* **56**, 215–261.
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.
Shoham, Y., Lamed, R. & Bayer, E. A. (1999). *Trends Microbiol.* **7**, 275–281.
Simon, I., Scheraga, H. A. & Manley, R. S. J. (1988). *Macromolecules*, **21**, 983–990.
Stevenson, D. M. & Weimer, P. J. (2005). *Appl. Environ. Microbiol.* **71**, 4672–4678.
Stout, G. H. & Jensen, L. H. (1968). *X-ray Structure Determination: A Practical Guide*. New York: McMillan.
Taylor, M. E. & Drickamer, K. (2009). *Glycobiology*, **19**, 1155–1162.
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. (2010). *Acta Cryst.* D**66**, 22–25.
Van Soest, P. J., Robertson, J. B. & Lewis, B. A. (1991). *J. Dairy Sci.* **74**, 3583–3597.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.
Xu, Q., Morrison, M., Nelson, K. E., Bayer, E. A., Atamna, N. & Lamed, R. (2004). *FEBS Lett.* **566**, 11–16.
Yaniv, O., Halfon, Y., Shimon, L. J. W., Bayer, E. A., Lamed, R. & Frolow, F. (2012). *Acta Cryst.* F**68**, 8–13.
Yaniv, O., Morag, E., Borovok, I., Bayer, E. A., Lamed, R., Frolow, F. & Shimon, L. J. W. (2013). *Acta Cryst.* F**69**, 733–737.
Yaniv, O., Petkun, S., Shimon, L. J. W., Bayer, E. A., Lamed, R. & Frolow, F. (2012). *Acta Cryst.* D**68**, 819–828.
Yaniv, O., Shimon, L. J. W., Bayer, E. A., Lamed, R. & Frolow, F. (2011). *Acta Cryst.* D**67**, 506–515.
Zhang, Y.-H. P. & Lynd, L. R. (2005). *J. Bacteriol.* **187**, 99–106.