doi:10.1016/j.jmb.2010.07.028

J. Mol. Biol. (2010) **402**, 374–387

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

The Unique Binding Mode of Cellulosomal CBM4 from *Clostridium thermocellum* Cellobiohydrolase A

Markus Alahuhta, Qi Xu, Yannick J. Bomble, Roman Brunecky, William S. Adney, Shi-You Ding, Michael E. Himmel and Vladimir V. Lunin*

National Renewable Energy Laboratory, 1617 Cole Boulevard, Golden, CO 80401, USA

Received 26 April 2010; received in revised form 12 July 2010; accepted 14 July 2010
Available online
21 July 2010

Edited by G. Schulz

The crystal structure of the carbohydrate-binding module (CBM) 4 Ig fused domain from the cellulosomal cellulase cellobiohydrolase A (CbhA) of *Clostridium thermocellum* was solved in complex with cellobiose at 2.11 Å resolution. This is the first cellulosomal CBM4 crystal structure reported to date. It is similar to the previously solved noncellulosomal soluble oligosaccharide-binding CBM4 structures. However, this new structure possesses a significant feature—a binding site peptide loop with a tryptophan (Trp118) residing midway in the loop. Based on sequence alignment, this structural feature might be common to all cellulosomal clostridial CBM4 modules. Our results indicate that *C. thermocellum* CbhA CBM4 also has an extended binding pocket that can optimally bind to cellodextrins containing five or more sugar units. Molecular dynamics simulations and experimental binding studies with the Trp118Ala mutant suggest that Trp118 contributes to the binding and, possibly, the orientation of the module to soluble cellodextrins. Furthermore, the binding cleft aromatic residues Trp68 and Tyr110 play a crucial role in binding to bacterial microcrystalline cellulose (BMCC), amorphous cellulose, and soluble oligodextrins. Binding to BMCC is in disagreement with the structural features of the binding pocket, which does not support binding to the flat surface of crystalline cellulose, suggesting that CBM4 binds the amorphous part or the cellulose "whiskers" of BMCC. We propose that clostridial CBM4s have possibly evolved to bind the free-chain ends of crystalline cellulose in addition to their ability to bind soluble cellodextrins.

Keywords: carbohydrate-binding module; CBM4; cellulosome; CbhA; cellulose degradation

Introduction

In the United States, The Energy and Security Independence Act of 2007 mandated volumetric targets (2015 and 2022) for the integration of renewable transportation fuels to reduce reliance on foreign sources of oil. A recent technoeconomic analysis of the corn-stover-to-bioethanol process points to the high cost of pretreatment and enzymes as the key hurdle that needs to be overcome in the path to commercialization.[1] Although considerable effort has been expended towards understanding and improving the performance of free (nonaggregated) cellulases, the search for new and improved cellulose-degrading enzymes from nature is an obvious parallel strategy for meeting this cost goal.[2]

Some bacteria and a few anaerobic fungi use large macromolecular complexes of enzymes called cellulosomes, in addition to free enzymes in solution, to

*Corresponding author. E-mail address: vladimir.lunin@nrel.gov.

Abbreviations used: CBM, carbohydrate-binding module; CbhA, cellobiohydrolase A; BMCC, bacterial microcrystalline cellulose; GH, glycoside hydrolase; GH9, GH family 9; PDB, Protein Data Bank; MD, molecular dynamics; CMC, carboxymethyl cellulose.

hydrolyze plant cell walls.[3] Efficient cellulose conversion by these microorganisms is known to occur in complex populations, such as in the rumina of ungulates and in the guts of plant-matter-degrading insects, leading to the hypothesis that cellulosomes may have high specific activities on biomass. Cellulosomes contain a number of glycoside hydrolase (GH) enzymes with different functionalities that are bound to a single protein scaffold. These enzymes work together as a complex on the surface of the cell to degrade plant biomass. In contrast, fungal cellulase systems rely on a reducing-end-specific cellobiohydrolase (GH family 7) to achieve efficient cellulose hydrolysis. Cellulosomes use enzymes from GH families 5, 9, and 48, and do not contain a family 7 enzyme. In cellulosomal systems, it appears that members of GH family 9 (GH9) play an important role in cellulose conversion. One of the dominant family 9 catalytic domains in *Clostridium thermocellum* has been shown to be part of the cellobiohydrolase A (CbhA) multidomain complex, which is one of the largest cellulosomal enzymes known to date.[4] CbhA displays seven distinct domains starting from the N-terminus: carbohydrate-binding module (CBM) 4, an Ig-like domain, GH9, two X1 modules, CBM3b, and a type I dockerin.[5]

The protein domains thought to bind to cellulose and often linked to GHs were originally termed cellulose binding domains.[6] In more recent years, this unnecessarily specific term was replaced with the more general CBM.[7] Today, CBMs have been classified into 59 specific families based on amino acid similarity[8] (see CAZy[9] Web site). Although some ambiguity exists in the classification of some CBM families, most belong to type A, type B, or type C. Type A CBMs bind on the surfaces of insoluble polysaccharides (primarily cellulose) and display binding surfaces usually defined by a series of planar amino acid side chains (Tyr, Trp, and Phe) interacting with the hydrophobic surface of cellulose (i.e., the 1,0,0 face of cellulose 1β). Type B CBMs bind to the solution form of polysaccharides and usually display open cleft structures that are able to accommodate highly flexible polymers. Type B CBMs are probably unable to translate on the targeted polysaccharide chain in contrast to type A CBMs (i.e., CBM1 from *Trichoderma reesei* cellobiohydrolase I).[10] Type C CBMs interact with the ends of polysaccharide chains and contain binding sites that are pockets or short closed tunnels.[8] In general, CBMs are thought to play three possible roles: substrate proximity effect, substrate targeting, and substrate disruption.

It is our long-term strategy to study the structure–function relationships governing the action of this large cellulosomal enzyme complex. The crystal structures of the CbhA Ig-GH9 construct[11] and CBM3b[12] have been solved recently. CBM4, in general, is considered to be a type B CBM, binding putatively to single (and soluble) polysaccharide chains. CBM4 was predicted to play an important role in the degradation of plant cell wall.[5] In this study, we solved the structure of the CBM4-Ig construct from *C. thermocellum* CbhA and conducted a detailed comparison of this new structure with previously known CBM4 structures such as laminarinase 16A from *Thermotoga maritima*, cellulase 9B from *Cellulomonas fimi*, and xylanase 10A from *Rhodothermus marinus*. Finally, we analyzed computationally the ability of CbhA CBM4 to bind to different types of oligosaccharides and also tested the binding behavior of five site-directed CBM4 mutants.

Results and Discussion

Crystal structure of CbhA CBM4-Ig

The crystal structure of the *C. thermocellum* CbhA CBM4-Ig construct, in complex with cellobiose, was refined to a resolution of 2.11 Å with R and R_{free} of 0.161 and 0.224, respectively. The asymmetric unit contained one molecule, with a nine-residue-long completely ordered native linker region (between Pro178 and Pro186) connecting the CBM4 and Ig domains. The structure of the linker region was ordered due to intertwining with the neighboring symmetry-related molecule (Fig. 1). It should be noted that the threonine/proline-rich regions of some cellulosomal linker peptides are decorated by O-linked glycosylation with galactopyranose when expressed in *C. thermocellum*.[13] However, the *C. thermocellum* CbhA CBM4-Ig linker peptide has no threonines, but there is one water-exposed threonine side chain close to the linker peptide at position 176 and three prolines in the linker at positions 178, 184, and 186.

The contacts between symmetry-related molecules, as well as between the CBM4 domain and the Ig domain, appear to be nothing more than just regular crystal contacts (the interface area between CBM4 and Ig is 379 Å^2 out of 13,781 Å^2 in total) and most likely are not related to any biological function. The Ig structure was almost identical (rmsd of 0.69 Å for C$^\alpha$ atoms) to the previously described *C. thermocellum* CbhA Ig-GH9 structure (1RQ5[11]). The CBM4 domain has a typical β-sandwich fold with two β-sheets containing five antiparallel β-strands each. Found in the CBM4 domain are two metal ions, which have been identified as magnesium ions based on their coordination geometry, distances, and the presence of $MgCl_2$ in the crystallization solution (Fig. 2). One of the magnesium ions is located on the surface, interacting with five water molecules and one amino acid side chain (Glu19),

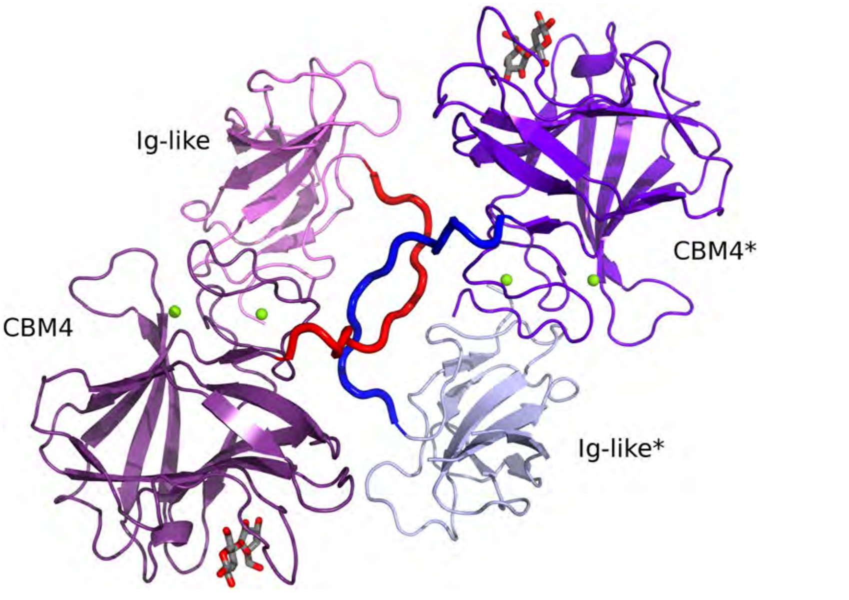

Fig. 1. CbhA CBM4-Ig shown in contact with its symmetry-related molecule. The asymmetric unit contains one molecule, with the completely ordered linker region (connecting the CBM4 and Ig domains) intertwined with the neighboring symmetry-related molecule. The symmetry molecule is shown on the right side (labeled with an asterisk). Magnesium ions are shown as green spheres, and cellobiose molecules are shown as sticks with gray carbons and red oxygens.

and another one is partially buried, interacting with one water molecule and five amino acid residues (Thr21, Asp23, Asn53, Lys54, and Asp166). The location of this partially buried magnesium ion corresponds with the position of calcium ions reported in other CBM structures.[8] We note that both magnesium ions are not in proximity to the binding site.

The binding pocket of CbhA CBM4 and the mode of binding to cellobiose

The binding pocket of CbhA CBM4 is located on the concave face of one of the β-sheets, with three loops (Gly104-Trp114, Glu34-Gly39, and Lys63-Ser69) forming the walls of the binding cleft. A cellobiose molecule was found in this pocket competing for this site with a glucose-1-phosphate molecule (Fig. 3). We assumed that the glucose-1-phosphate molecule was trapped in the binding cleft during protein expression (i.e., it was not replaced by cellobiose completely during cocrystallization). This cellobiose molecule was fixed in place by ring stacking with the Tyr110 and Trp68 side chains; by hydrogen bonding to Gln71, Arg73, and Lys102; and by several water-mediated H-bonds with Cys33, Asp35, Arg75, Gln105, Met106, Glu146, and His150 (Fig. 4).

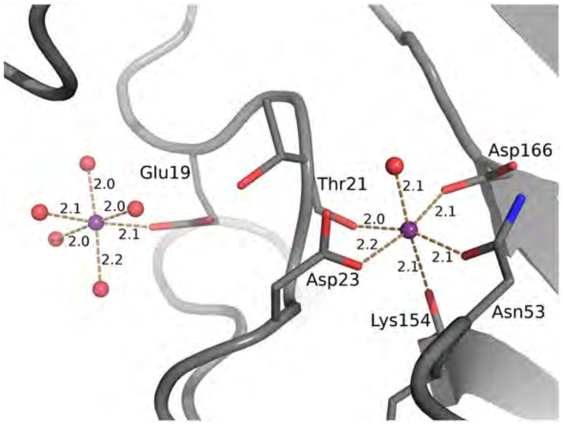

Fig. 2. CbhA CBM4 magnesium binding sites. Two magnesium ions were found in the CBM4 structure. The magnesium ions are shown in purple, and red balls indicate water molecules.

Structure comparison

In order to elucidate the structure–function relationship of CbhA CBM4, we conducted a comparison of related structures. Twenty proteins structurally similar to CBM4 and 372 proteins similar to Ig were found in the Protein Data Bank (PDB†) using the VAST server.[15] Three hundred twenty-nine structures with *Z*-scores of Œ5 for the CBM4 domain and 439 structures with *Z*-scores of Œ5 for the Ig domain were identified using the Dali[16] search tool. Manual comparison of the CBM4 domain with other CBM β-sandwich folds[17] yielded even more structurally similar PDB entries. The CBM4 structural homologues include many CBMs (families 2, 3, 4, 6, 11, 15, 16, 17, 22, 27, 28, 29, 30, 32, 35, 36, and 44), GHs (families 7, 16, 11, and 12), and various other proteins sharing the same β-sandwich fold.

† www.rscb.org

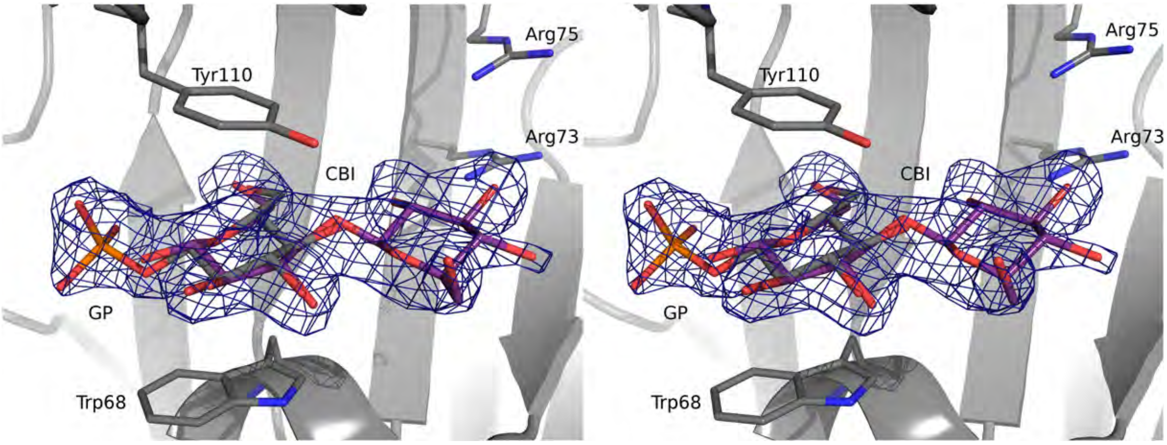

Fig. 3. The omit electron density map of cellobiose (CBI; purple carbons) and glucose-1-phosphate (GP; gray carbons). This $F_o - F_c$ map has been calculated at 3σ after 10 cycles of REFMAC5[14] refinement without ligands.

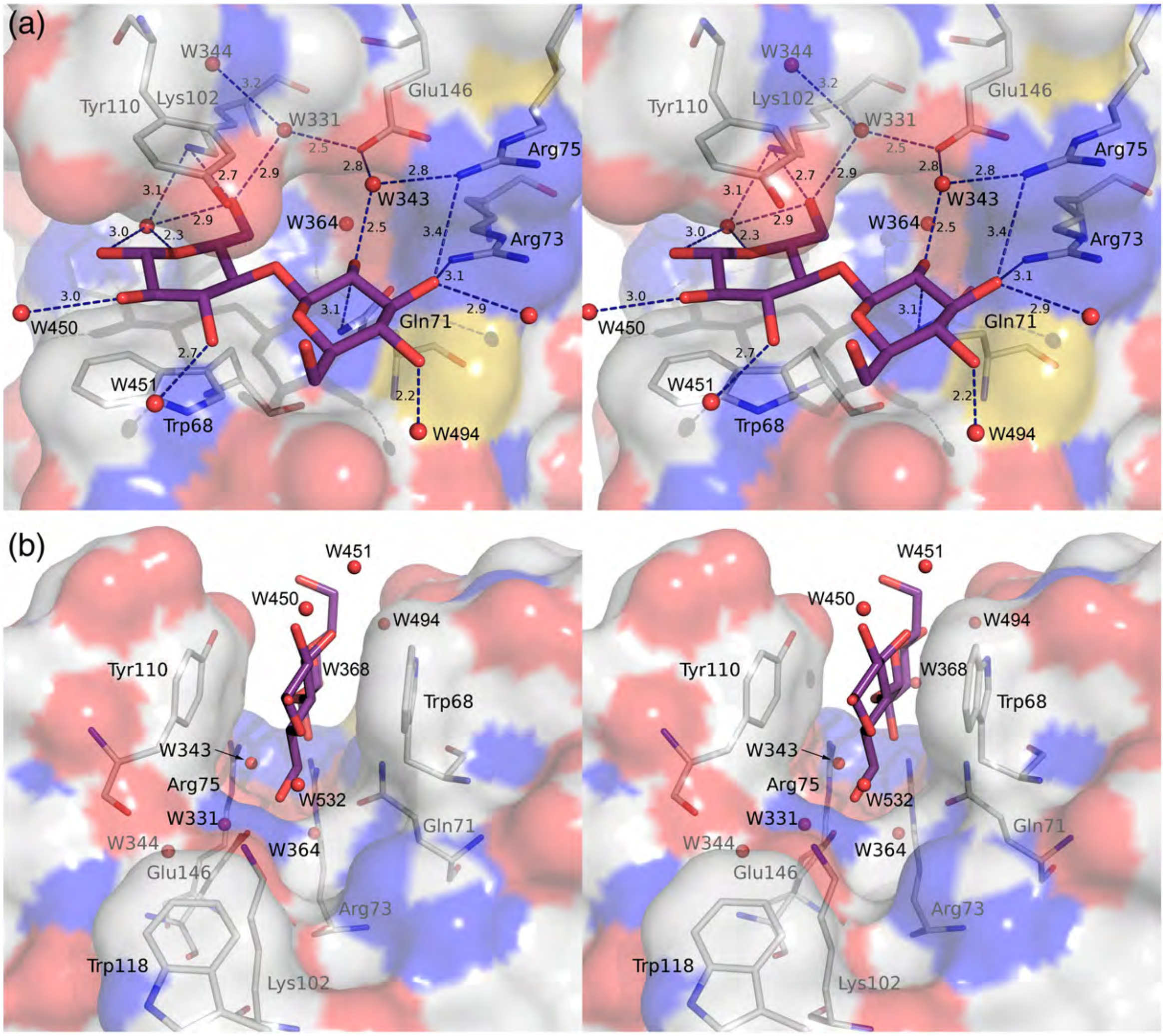

Fig. 4. Interactions of cellobiose in the binding pocket of CbhA CBM4. Top view (a) with important interactions, and side view (b) showing the shape of the binding cleft. Nearby residues and water molecules are labeled and shown as sticks. Note the position of water molecules 311, 343, 344, and 364 below the cellobiose molecule.

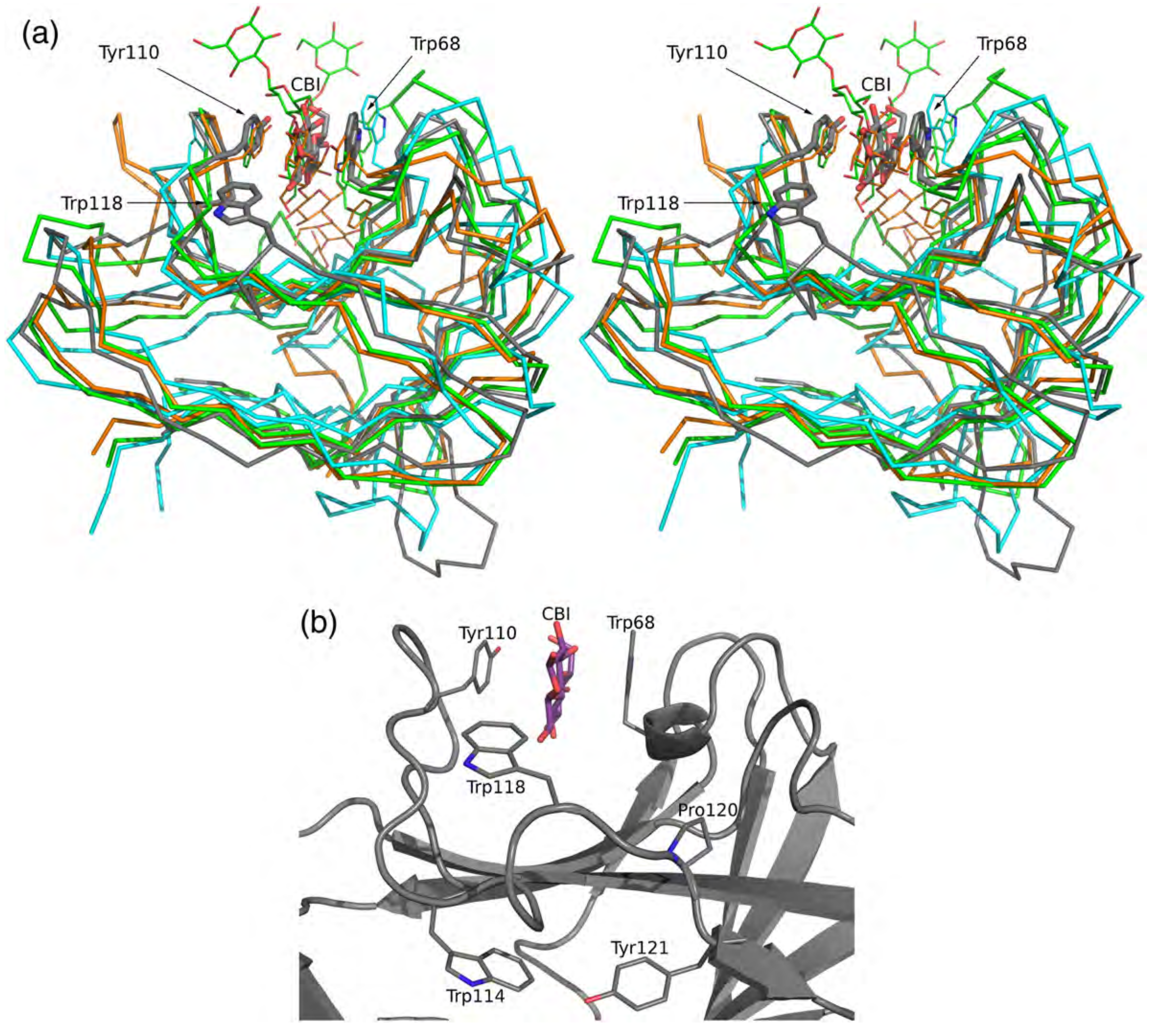

Fig. 5. Trp118 of CbhA CBM4. Comparison to other CBM4s (a) and details of the Trp118 environment (b). CbhA CBM4 ribbon and carbon atoms are shown in gray, 1K42 is shown in magenta, 1GUI is shown in green, and 1GU3 is shown in orange. The conserved binding site residues of 1K42, 1GUI, and 1GU3, as well as the cellohexaose ligands of 1GUI and 1GU3, are shown as sticks.

Naturally, the most closely related structures found were those from various CBM4 domains. For example, the core β-sandwich folds found in the structures of different CBMs were generally very similar, and even some proteins with unrelated functions, such as virus envelope proteins, have distinctly similar folds. Fourteen structures with ligands bound in the same pocket were available from other members of CBM4 (1GU3, 1GUI, and 1K42), CBM15 (1GNY and 1US2), CBM16 (e.g., 2ZEX), CBM27 (e.g., 1PMH), and CBM29 (e.g., 1GWM). From these 14 structures, nine structures belonging to families 16, 27, and 29 showed the ligands bound "flat" in the pocket (with approximately a 90° rotation of the sugar plane compared to the cellobiose found in the CBM4-Ig structure). The remaining five structures from CBM families 4 and 15 displayed ligands that are bound in approximately the same orientation that we observed. Cellulase 9B CBM4-1 in complex with cellopentaose from *Ce. fimi* (*Cf*CBM4, 1GU3; rmsd of 1.62 Å for all C^{α} atoms), laminarinase 16A CBM4-2 in complex with laminariheptaose from *T. maritima* (*Tm*CBM4, 1GUI; rmsd of 1.58 Å for all C^{α} atoms), and the NMR structure of xylanase 10A CBM4-2 from *R. marinus* (*Rm*CBM4, 1K42; modeled xylan, rmsd of 2.53 Å for all C^{α} atoms) showed binding modes very similar to that of the CbhA CBM4 reported here (i.e., with aromatic residues at almost identical positions). These aromatic residues are thought to facilitate ring stacking with the sugar units of their respective ligands (Fig. 5). Xylanase 10C CBM15 from *Pseudomonas cellulosa* (*Pc*CBM15, 1GNY) and xylanase from *Cellvibrio japonicus* (*Cj*CBM15, 1US2), both in complex with xylopentaose, also bind to their ligands similarly to CbhA CBM4. The results of the above comparisons immediately suggested that the binding pocket of CbhA CBM4 was cleft shaped (as seen in other type B[8] CBMs) with a similar mode of ligand binding (Fig. 4).

The unique binding site tryptophan of clostridial CBM4s and the extended binding pocket

A unique feature of the CbhA CBM4 structure is the loop between Trp114 and Tyr121, which includes a tryptophan at position 118 (Fig. 5). While comparing CBM4 sequences and structures, we could not find any aromatic residues equivalent to Trp118 in similar proteins with known three-dimensional structures. However, our sequence comparison revealed that this tryptophan or another aromatic residue can be found at the same or equivalent position in many CBMs (Table S1). All available sequences with an identity of over 31% relative to CbhA CBM4 are cellulosomal CBM4 modules from clostridia. A number of sequences from various species with identities below 31% might also have an aromatic side chain at this position, but this finding remains uncertain because there are gaps in the sequence alignments in this region. All the above sequences were annotated as CBMs or domains from family 9 or family 16 cellulases, laminarinases, or endoglucanases. It seems that Trp118 and this loop might be dominant features of cellulosomal CBM4 modules from clostridia.

Closer inspection of the *C. thermocellum* CbhA CBM4 ligand binding interactions reveals a series of side chains starting from Trp118, continuing through the binding cleft described above, and ending with Arg73 and Arg75 on the opposite side of the binding cleft (Fig. 4). Sequence conservation of this extended pocket between CBMs shows some similarity between these CBM4 domains. Specifically, aromatic residues at positions 68 and 110 are conserved, but Trp118 is unique to the *C. thermocellum* CbhA CBM4 structure (Table S2).

The Trp118 loop environment

In the structure reported here, Trp118 seems to be stabilized and protected from water by forming contacts with the equivalent area of the neighboring symmetry-related molecule. Flexibilities from normal-mode analysis calculations, using the generalized Born theory, indicated that this loop is fairly stable. The other two binding site residues in loops directly involved in substrate binding, Tyr110 and Trp68, are only somewhat flexible (Fig. 6). Trp118 is located in a fairly stiff loop, whereas Tyr110 is located in a flexible one. The flexibility of this loop is correlated, in all probability, to the recognition of various substrates. The first three normal modes of CBM4 show extreme motion of the catalytic cleft (Fig. S1) possibly related to the mechanism by which the CBM acquires the oligodextrin chain. It is acknowledged that the first 5–10 normal modes are, in most cases, enough to describe the overall dynamics of a given protein. Usually, these low-energy modes have the largest contributions to atomic fluctuations or the overall dynamics of a given protein.

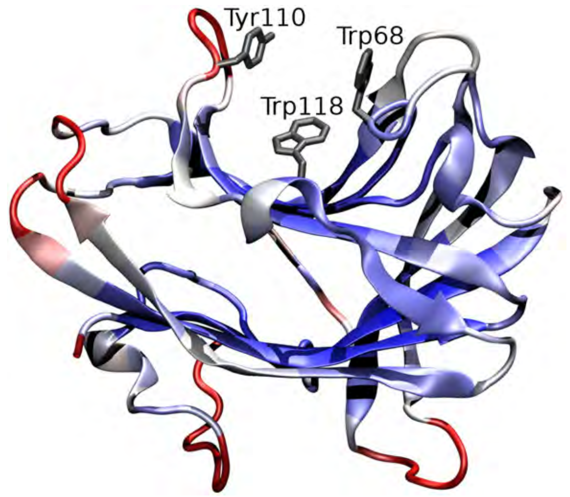

Fig. 6. Protein structure of CBM4 colored by *B*-factor from the most flexible (red) to the least flexible (blue). The residues predicted to be directly involved in substrate binding are shown in licorice representation (gray).

Trp118 is indeed very well stabilized even in the absence of crystal contacts. It is positioned on the tip of the loop between residues Trp114 and Tyr121 (Fig. 5). A proline at position 120 appears to lock the C-terminal "hinge," and a helical turn formed by residues 115–119 (with a hydrogen bond between OD1 of Asn115 and the main-chain nitrogen of Trp118) appears to further rigidify this peptide loop. The Trp118 side chain is firmly anchored in place by contacts with the side chains of Tyr100, Glu112, and Asn115, further increasing the stability (or inflexibility) of this loop.

Docking and molecular dynamics simulations

To confirm the existence of the extended binding pocket formed by Trp118, Tyr110, Trp68, Arg73, and Arg75, we conducted docking experiments and molecular dynamics (MD) simulations with different oligodextrins. The docking experiments with cellodextrin and oligoxylan ligands showed that this CBM4 can bind to both types of oligosaccharides. The best results were obtained when four water molecules (331, 343, 344, and 364) positioned between the ligand and protein molecules were included in the calculations. This result suggested an

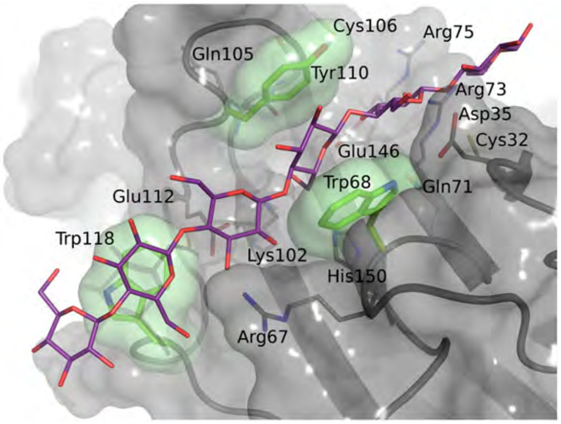

Fig. 7. MD simulation snapshot of CbhA CBM4 with cellohexaose. Docking experiments and MD simulation demonstrated that the binding pocket of CbhA CBM4 can bind to oligodextrins up to six sugar units long. The carbons of cellohexaose are shown in magenta.

important structural role for these water molecules. It is possible that the hydrogen bonds formed by these water molecules are important for ligand binding or specificity. However, Pell *et al.*, in their studies of CBM15 domains that are structurally similar to CBM4 modules, reported that individual water molecules are not necessary for binding.[18] The best ICM scores with a clear margin were obtained for cellopentaose (−27.55) and xylotetraose (−19.34) (Table S3). Together with manual inspection of the binding pocket area, this result indicated that the optimum number of interactions by CbhA CBM4 can be achieved with oligodextrins four to five sugar units in length. Another interesting observation is the predominance of the highest-scoring cellodextrins (larger than cellotriose) oriented with their nonreducing ends to the Trp118 side of the binding cleft.

Further investigation of two binding scenarios using MD simulations revealed stable stacking configurations (several nanoseconds during each simulation) between cellodextrin chains and Trp118 (Fig. 7; Figs. S2 and S3). Importantly, Trp118 displays a side chain in the same plane as the protein surface adjacent to the binding cleft (Fig. 5). This observation confirms the possible role of Trp118 and the existence of the extended binding pocket formed by Trp118, Tyr110, Trp68, Arg73, and Arg75 that can accommodate oligodextrins four to six sugar units long.

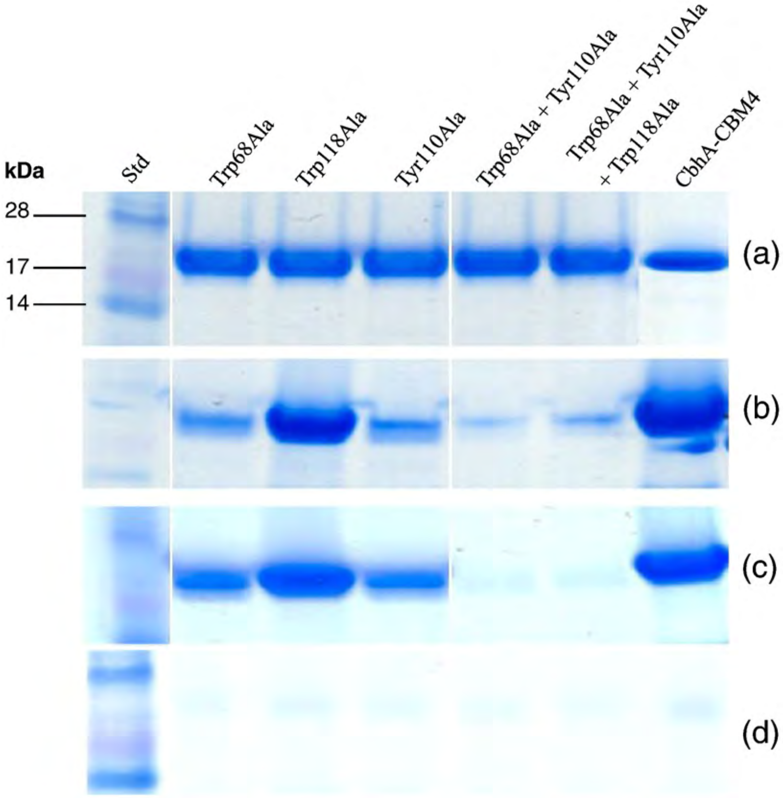

Fig. 8. Binding of various mutants to different insoluble polysaccharides revealed by SDS-PAGE. (a) Before binding. (b) Binding to BMCC. (c) Binding to amorphous cellulose. (d) Binding to insoluble xylan of birchwood. The gel figures have been cropped to remove unnecessary parts or results.

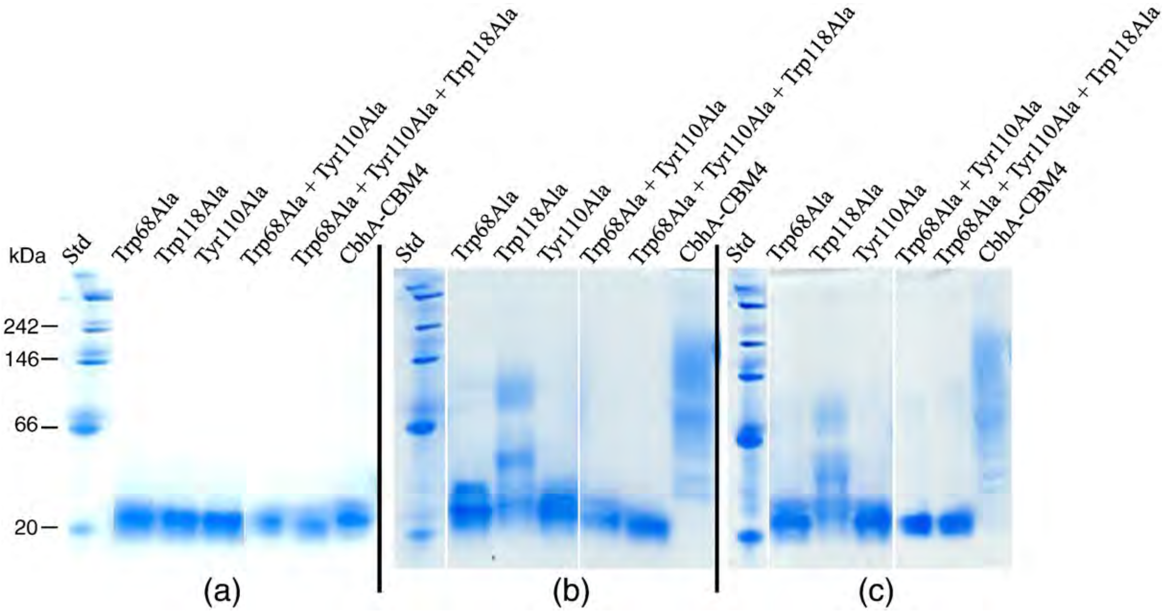

Fig. 9. CbhA CBM4 and its mutants binding to soluble xylan, as revealed by Native PAGE. (a) No soluble xylan added. (b) Soluble xylan of birchwood added. (c) Soluble xylan of switch grass added. Binding of soluble xylan causes multiple bands due to the heterogeneous nature of the xylan ligands. The gel figures have been cropped to remove unnecessary parts or results.

Site-directed mutagenesis and binding assays

In order to elucidate the mechanism of the binding of CBM4 to oligodextrins, we generated five mutants. These mutants included three single-residue mutants (Trp68Ala, Tyr110Ala, and Trp118Ala), one double-residue mutant (Trp68Ala + Tyr110Ala), and one triple-residue mutant (Trp68Ala + Tyr110Ala + Trp118Ala). With these mutations, we probed the importance of binding pocket aromatic residues. These mutants were designed to evaluate the possible binding ability of Trp118 compared to the binding cleft aromatic residues Trp68 and Tyr110.

The binding of the mentioned mutants to bacterial microcrystalline cellulose (BMCC; a cellulose standard material that contains a high crystalline cellulose content[19]) and amorphous cellulose (Avicel treated by phosphoric acid[20]) was determined by affinity electrophoresis. The results showed that *C. thermocellum* CbhA CBM4 can bind to BMCC and amorphous cellulose (Fig. 8b and c). In comparison to wild-type CBM4, only the Trp68Ala and Tyr110Ala mutants showed reduced binding to both BMCC and amorphous cellulose. Thus, both Trp68 and Tyr110 are crucial for CbhA CBM4 binding to these forms of cellulose. The binding strength of both the double-residue mutant (Trp68Ala + Tyr110Ala) and the triple-residue mutant (Trp68Ala + Tyr110Ala + Trp118Ala) was also considerably reduced compared to wild type (became essentially nonexistent). This result further suggests that Trp68 and Tyr110 are essential for CbhA CBM4 binding to cellulose, but that Trp118 is not strictly required.

To understand binding to xylan in lignocellulose, we tested wild-type CBM4 and the five CBM4 mutants for their binding to soluble and insoluble forms of xylan. There was no apparent binding of the wild-type and mutant modules to (insoluble) birchwood xylan (Fig. 8d). However, the electrophoretic mobility of CBM4 changed in Native PAGE when soluble xylan from birchwood and switch grass was added (Fig. 9). Therefore, CBM4 can bind to the soluble xylan fraction in birchwood and switch grass. The Trp118Ala mutant showed a similar performance in binding to the wild-type module, implying again that Trp118 is not imperative for binding to soluble xylan. However, the electrophoretic mobility of the Trp68Ala and Tyr110Ala mutants was not significantly retarded by electrophoresis in the presence of soluble xylan, suggesting that Trp68 and Tyr110 are crucial for CBM4 binding to soluble xylan. Indeed, the double-residue and triple-residue mutants with deletion of both Trp68 and Tyr110 showed no significant electrophoretic retardation, further demonstrating the critical contribution of Trp68 and Tyr110 to soluble xylan binding. Note that the smeared bands in these gel mobility experiments were caused by proteins bound to the soluble xylan (Fig. 9).

Binding to soluble oligodextrins

To assess the ability of CbhA CBM4 to bind to soluble oligodextrins, we performed experiments using fluorescence emission spectrum measurements. The effects of carboxymethyl cellulose

(CMC), xylobiose, xylohexaose, cellobiose, cellotetraose, cellopentaose, and cellohexaose were probed. Measurements with the native CBM4 showed binding only with CMC, cellopentaose, and cellohexaose (Fig. 10). Also, the ability of CBM4 to bind to these ligands was abolished if Trp68, Tyr110, or Trp118 was mutated to an alanine. This clearly demonstrates that Trp118 contributes to the binding of cellodextrins at least five sugar units in length.

The binding of ligands to Trp118 results in a strong signal, as shown by a comparison of the fluorescence emission spectra of native CBM4 and the Trp118Ala mutant. The fact that mutating any one of the binding cleft aromatic residues (Trp68 or Tyr110) also completely abolishes binding shows that Trp118 on its own cannot bind to cellodextrins; the strong "sandwich" binding interaction afforded by Trp68 and Tyr110 is required. This result is consistent with the outcome of affinity electrophoresis experiments, which showed that Trp118 is not needed for binding to cellodextrins.

The lack of binding signal with cellobiose, cellotetraose, xylobiose, and xylohexaose is inconsistent with the affinity electrophoresis experiments and our knowledge of the three-dimensional structure of the binding pocket. We believe that the reason for the apparent lack of signal in the case of these ligands is twofold. Firstly, from our structure, we can see that we have a tightly bound contaminant in the binding cleft with a high occupancy (~50%) that could not be fully removed during protein purification or even cocrystallization with 100 mM cellobiose. This could have easily removed at least half of the signal by already binding Trp68 when what was meant to be an unliganded emission spectrum was measured. Secondly, it is possible that Trp68 is in an environment where emission changes do not occur upon ligand binding.

The role of Trp118

We have suggested that Trp118 is a common feature in clostridial CBM4 modules and that its side chain interacts with the sugar rings of cellodextrins. The orientation of its side chain (normal to the planar side of the binding pocket) is indeed similar to the orientation of critical aromatic residues in type A CBMs that are known to bind to the crystalline cellulose surface. This implies two possible roles for Trp118 in substrate binding: (1) binding to the crystalline cellulose surface and/or (2) guiding or stabilizing a soluble cellodextrin in the binding cleft. At first glance, the first hypothesis appears to have been supported by experimental results showing that both *C. thermocellum* CbhA CBM4 and cellulase K CBM4 (with a 76% sequence similarity to CbhA CBM4 and has the predicted unique tryptophan-containing loop) can bind to BMCC.[21,22] We note, however, that BMCC is not purely crystalline cellulose.

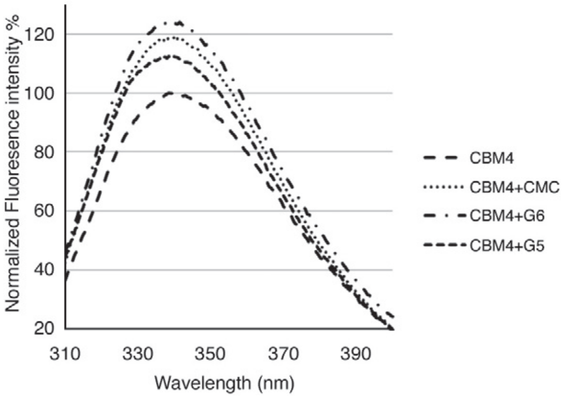

Fig. 10. Fluorescence emission spectrum measurements. G5 is cellopentaose, and G6 is cellohexaose. The results for mutants (Trp118; double and triple mutants), as well as the measurements with cellobiose, cellotetraose, xylobiose, and xylohexaose, are not shown because there was no signal (their curve equals CBM4).

With the use of targeted mutations, our results demonstrated that Trp118 is not required for the binding of CbhA CBM4 to BMCC, amorphous cellulose, or soluble xylan. Therefore, the binding of CbhA CBM4 to these substrates can be achieved by only two aromatic residues situated in the binding cleft (i.e., Trp68 and Tyr110). This appears to be in contradiction to what we have observed in the structural analysis of the binding pocket. The shape of the cleft, as well as the orientation of the aromatic side chains in it, does not support binding to the planar crystalline cellulose surface. Binding to crystalline cellulose requires a planar interactive protein modular face, as found in type A CBM modules, as well as a direct presentation of the interactive surface aromatic side chains to the cellulose surface.[8,23] Indeed, BMCC, as well as many other cellulose substrates, contains a significant amorphous content even though it is the most crystalline of the commonly used cellulose materials (crystalline content is 73.1–95.2%, depending on the measurement technique[19]). We suggest that the CBM4 interaction with BMCC occurs primarily through the amorphous regions of BMCC (which presumably expose cellodextrins), as well as through cellulose "whiskers" (i.e., free-chain ends extending from the crystalline core to the solution), rather than through the planar crystalline surfaces of the cellulose microfibril itself.

Fluorescence emission spectrum measurements clearly demonstrate that Trp118 contributes to the binding of CBM4 to soluble cellodextrins at least five sugar units in length. This experimental result is in good agreement with our computational results, where consistent and simultaneous ring stacking between the soluble cellodextrins and Trp68, Tyr110, and Trp118 can be observed (Fig. 7; Figs. S2 and S3). For all three contacts to be possible at the same time, only certain helical conformations with the right orientation of the cellodextrin chain are feasible. Thus, computer simulations and fluorescence emission spectrum measurements support our second hypothesis for the function of Trp118 (i.e., to direct soluble cellodextrins with specific helical conformations to the binding cleft or towards the catalytic domain). For example, Simpson *et al.* demonstrated with mutational studies of *Ce. fimi* xylanase 11A CBM2 that the orientation of the aromatic side chains at the binding site can be important for distinguishing between substrates.[23] By a single-point mutation, they succeeded in changing the orientation of one tryptophan side chain 90° from planar to a perpendicular orientation with respect to the CBM surface. This change was enough to make *Ce. fimi* xylanase 11A CBM2 specifically bind the helical conformation of xylan instead of its original substrate. This observation is consistent with part of our proposed function for Trp118—stabilizing a specific helical conformation of a cellodextrin substrate.

The function of clostridial CBM4s

Does Trp118 bind to soluble cellodextrins or can it also bind the flat surface of crystalline cellulose even though this amino acid is not necessary for binding to BMCC? Binding to cellulose "whiskers" still looks like the most probable binding mode for the CBM4 "cleft" aromatics Trp68 and Tyr110. Furthermore, binding of the module to the free ends of cellulose chains would presumably bring Trp118 close to the crystalline surface. Perhaps CbhA CBM4 has evolved to bind to these cellodextrin "whiskers" at a point near the crystalline cellulose surface, where Trp118 can bind to this planar surface. This scenario leads us to two possible binding modes for *C. thermocellum* CbhA CBM4: (1) binding of cleft residues to soluble cellodextrins both in solution and projecting from amorphous regions of cellulose, and (2) (weaker) binding of primarily Trp118 to the crystalline cellulose surface. Our experiments with CMC, cellohexaose, and BMCC confirm the first hypothesis, but do not disapprove the second hypothesis. Both are probably true. This result raises the possibility that clostridial CBM4 modules have evolved to bind to the free-chain ends of crystalline cellulose in addition to their ability to bind to soluble cellodextrins.

Conclusions

In this article, we have described the X-ray-diffraction-derived structure of the *C. thermocellum* CbhA CBM4-Ig domain in complex with cellobiose. The CBM4 module constitutes a β-sandwich fold and a cleft-like binding site typical of type B[8] CBM modules. CBM4 thus would seem to prefer soluble substrates to crystalline cellulose. A unique feature of this and other clostridial CBM4 modules not found in any other CBM4 structure reported to date is a peptide loop displaying a strategically placed tryptophan residue (Trp118) at the tip of the loop. The side chain of Trp118 is located close to the binding pocket and seems to be in position to facilitate binding. Our results show that *C. thermocellum* CbhA CBM4 has an extended binding pocket formed by Trp118, Tyr110, Trp68, Arg73, and Arg75 that can optimally bind cellodextrins containing five or more sugar units. Site-directed mutagenesis and binding studies of this module demonstrated that Trp68 and Tyr110 are critical for the binding of CbhA CBM4 to soluble sugars. This observation, together with the structural analysis of the binding cleft and the known amorphous contents of the commonly used cellulose substrates,[19] suggests that CBM4 binds the amorphous regions and/or the cellulose "whiskers" of BMCC. Our experiments indicate that clostridial CBM4 modules may have evolved to bind to the free-chain ends of crystalline cellulose in addition to soluble oligodextrins.

Materials and Methods

Expression, purification, and crystallization

CBM4-Ig (the dual domain of CBM4 and Ig) of *C. thermocellum* CbhA was amplified by PCR (the template is *C. thermocellum* genomic DNA) using the primers TCCGTG*CATATG*TTAGAAGATAATTCTTCGACT and CATCTG*CTCGAG*GATGTCTTTGCGAATGTCAA. The site-directed mutants of CBM4$_{\text{CbhA}}$ were synthesized by GenScript (Piscataway, NJ). The PCR fragment of CBM4-Ig or synthesized genes was inserted into the plasmid pET22b (Novagen, Madison, WI) via NdeI and XhoI to generate expression plasmids. CBM4-Ig was overexpressed in *Escherichia coli* (BL21) (Stratagene, La Jolla, CA) with induction of 0.3 mM IPTG. The recombinant CBM4-Ig containing a C-terminal His-tag (6× histidine) was purified with the QIAexpress Ni-NTA protein purification system (Qiagen, Valencia, CA), followed by size-exclusion chromatography with HiLoad Superdex 75 (26/60) (GE Healthcare, Piscataway, NJ) in buffer A [20 mM Tris–HCl (pH 7.0) containing 100 mM NaCl, 1 mM ethylenediaminetetraacetic acid, and 1 mM sodium azide], using the protocol recommended by the manufacturer. The purified fusion protein was concentrated with a Vivaspin 5K concentrator (Vivaproducts, Littleton, MA), and the protein concentration was measured by a NanoDrop UV spectrophotometer (NanoDrop, Wilmington, DE). Diffraction-quality crystals were obtained with sitting-drop vapor diffusion using a 96-well plate with Crystal Screen (Hampton Research, Aliso Viejo, CA). Fifty microliters of well solution was used with drops

Table 1. X-ray data collection and refinement statistics

Data collection	
Space group	$C222_1$
Unit cell parameters	
a, b, c (Å)	60.58, 94.32, 113.49
α, β, γ (°)	90.0, 90.0, 90.0
Wavelength (Å)	1.54178
Temperature (K)	100
Resolution (Å)	25.0–2.11 (2.15–2.11)
R_{int}[a]	0.100 (0.421)
Observed reflections	77,757 (2171)
Unique reflections	19,105 (807)
Redundancy	4.07 (2.69)
$I/\sigma(I)$	10.83 (2.33)
Completeness (%)	99.7 (98.1)
Refinement	
Resolution	25–2.11 (2.16–2.11)
Number of reflections	18,089 (1376)
R/R_{free}	0.161 (0.192)/0.224 (0.273)
Protein atoms	2292
Water molecules	275
Ligands and ions	48
rmsd bond length (Å)	0.021
rmsd bond angle (°)	1.790
Average B-factor (Å^2)	
Protein atoms	13.7 (with three TLS groups)
Ligands and ions	25.8
Water molecules	19.9
Ramachandran plot statistics (%)[b]	
Allowed	100
Favored	90
Outliers	0

Statistics for the highest-resolution bin are in parentheses, and the number of atoms includes alternative conformations.

[a] $R_{\text{int}} = \sum|I - \langle I \rangle| / \sum(I)$, where I is the intensity of an individual reflection, and $\langle I \rangle$ is the mean intensity of a group of equivalents.

[b] Ramachandran plot statistics were calculated using MolProbity.[34]

containing 1 μL of well solution and 1 μL of protein solution. The recombinant CBM4-Ig protein was crystallized at 25 °C in buffer B [0.1 M Tris–HCl (pH 8.5), 0.2 M $MgCl_2$, and 30% wt/vol polyethylene glycol 4000] containing 21 mg/mL protein kept in buffer A and supplemented with 100 mM cellobiose.

X-ray diffraction and structure determination

Before data collection, the CBM4-Ig crystal was soaked in 50%/50% (vol/vol) paraffin/paratone oil and flash frozen in a cold nitrogen gas stream at 100 K. Data collection was performed using a Bruker X8 MicroStar X-ray generator with Helios mirrors and Bruker Platinum 135 CCD detector. Data were indexed and processed with the Bruker Suite of programs (version 2008.1-0; Bruker AXS, Madison, WI). Intensities were converted into structure factors, and 5% of the reflections were flagged for R_{free} calculations using the programs F2MTZ, Truncate, CAD, and Unique from the CCP4 package of programs.[24] The automatic molecular replacement program MrBUMP[25] (version 0.4.4) was used to search, modify, and try different models with the programs FASTA,[26] SCOP,[27] and MOLREP[28] (version 10.2.23). The structure was built using combined models 1RQ5 (Ig part)[11] and 1GU3 (CBM4 part)[29] and an in-house homology model using 1GU3 as template, and then rebuilt in ARP/wARP[30] (version 7.0). Further refinement and manual correction were performed using REFMAC5[14] (version 5.5.01) and Coot[31] (version 0.6-pre-1). Three TLS groups generated using the TLS Motion Determination server[32,33] were used in the final cycle of refinement. The resulting structure has been deposited in the PDB with accession number 3K4Z. Data collection and refinement statistics are shown in Table 1.

Structure analysis

The programs Coot,[31] PyMOL‡, and ICM§ were used for comparing and analyzing structures. Ramachandran plot statistics were calculated using MolProbity.[34] The contact area between the CBM4 domain and the Ig domain was calculated using the program ICM after cutting the linker region between residues 177 and 185. Accessible surface was determined using a probe with a 0.7-Å radius to calculate the individual and combined surface areas. Sequence similarity searches were performed using NCBI BLAST.[35] Structural similarity searches were completed using the VAST server[15] and Dali.[16] Rmsds were calculated using ICM (version 3.6-1e; Molsoft LLC, La Jolla, CA). Docking studies were performed using the virtual ligand screening[36,37] module of the program ICM. After initial docking with oligodextrin (Gn; oligodextrin with n number of glucose units; G2–G10) and oligoxylan (Xn; oligoxylan with n number of xylose units; X2–X10) ligands, the binding pocket area was refined, with the most representative docking result obtained with cellopentaose (G5) and four water molecules (331, 343, 344, and 364) deemed structurally relevant to ligand binding. Final docking studies were repeated three times, using the refined receptor with the four water molecules and a ligand library with oligodextrin and oligoxylan ligands containing G2–G6 and X2–X6, as described above. This approach was used in consideration of the initial docking results, which indicated that longer oligosaccharides could not be accommodated in the binding pocket area. In all docking experiments, a thoroughness of 5 was used, and calculations were repeated three times. The sequence conservation of the binding pocket was analyzed by taking the docking result with xylohexaose (X6) and selecting all residues with atoms inside a 7-Å sphere from it. This residue selection was then combined with a sequence alignment of similar structures that were structurally superimposed onto CBM4-Ig.

MD simulations

The CBM with bound cellulose was modeled with the CHARMM27 force field using CMAP correction[38–40] for the protein. The new CHARMM force field for carbohydrates was applied to cellulose,[41,42] the TIP3P water model[43,44] was used for the solvent, and particle mesh Ewald summation was used for electrostatics[45] with a nonbonded interaction cutoff of 10 Å. The system was

‡ http://www.pymol.org

§ http://www.molsoft.com

initially equilibrated for 1 ns in an MD simulation with an NVT ensemble and a time step of 2 fs. We performed six 15-ns MD simulations with an NPT ensemble and a time step of 2 fs with different random seeds starting from two initial cellohexaose configurations.

Normal-mode analysis

Normal-mode analyses were carried out with the molecular mechanics program package NAB[46,47] (now part of Amber 10),[48,49] using the parameter set parm99SB,[50,51] and the pairwise approach of Hawkins *et al.* for the GB model.[52,53] The monomeric CBM4 structure without the Ig module was minimized using the Limited-Memory Broyden–Fletcher–Goldfarb–Shanno Truncated Newton Conjugate minimization technique to obtain an rms gradient below 1×10^{-8} kcal/mol Å. This level of convergence is necessary to avoid contamination from translational and rotational modes into true internal modes. Diagonalization of the Hessian matrix was performed using ARPACK[54] routines, in combination with Cholesky decomposition and inversion of the Hessian matrix, therefore providing better separation of eigenvalues to enhance convergence.

Fluorescence emission spectrum measurements

Steady-state fluorescence emission spectrum measurements were performed on a Jobin Yvon Fluoromax3 spectrofluorometer.[55] Excitation of the protein was selective for tryptophan at 295 nm, and spectra were collected from 310 to 400 nm. The CBM samples had a concentration of 0.1 mg/mL in 20 mM Tris buffer (pH 7.0). Various ligands were added at the following concentrations: CMC was added to a final concentration of 0.01%; and cellobiose, cellotetraose, cellopentaose, cellohexaose, xylobiose, and xylohexaose were added to a final concentration of 1 mM. The contribution of buffer and ligand addition was measured and subtracted from the final results.

Binding assays

The binding assays followed the procedure of Xu *et al.*,[56] where 0.3 mg of purified recombinant proteins was mixed with 1 mg of BMCC (a gift from the Cornell University, prepared in accordance with Jung *et al.*[57]), 0.8 mg of amorphous cellulose (treated with phosphoric acid[20]; Avicel PH101; FMC Corporation, Philadelphia, PA), and 1 mg of insoluble birchwood xylan (washed with water to remove partially soluble xylan; Sigma, St. Louis, MO). The suspension was brought to a final volume of 1.0 mL with 50 mM Tris (pH 8.0) buffer with 150 mM NaCl. These mixtures were kept at room temperature for 20 min with gentle rotation and centrifuged at 12,000*g* for 6 min to separate the polysaccharide containing the bound protein from the unbound protein in the supernatant. The cellulose particles with bound protein were washed three times with vortexing in 1 mL of the previously mentioned buffer. Finally, the polysaccharides were resuspended in 50 and 100 μL of loading buffer and heated in a boiling water bath for 10 min. After centrifugation, 20 μL of the supernatant was subjected to SDS-PAGE. Gel retardation assays with insoluble ligands were performed by running a normal 12% Native PAGE supplemented with 0.12% soluble xylan. The native sample buffer and protein standard were purchased from Invitrogen (Carlsbad, CA). Soluble xylan from birchwood and switch grass was prepared in accordance with Naran *et al.*[58]

PDB accession number

Atomic coordinates and structure factors have been deposited in the PDB with accession number 3K4Z.

Acknowledgements

This work was supported by the Department of Energy Office of Science, Office of the Biological and Environmental Research, through the BioEnergy Science Center, a Department of Energy Bioenergy Research Center. Computational time for this research was supported, in part, by the Golden Energy Computing Organization at the Colorado School of Mines using resources acquired with financial assistance from the National Science Foundation and the National Renewable Energy Laboratory. Simulations were also performed, in part, using the Texas Advanced Computing Center Ranger cluster under National Science Foundation Teragrid grant TG-MCB090159.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2010.07.028

References

1. Aden, A. & Foust, T. (2009). Technoeconomic analysis of the dilute sulfuric acid and enzymatic hydrolysis process for the conversion of corn stover to ethanol. *Cellulose*, **16**, 535–545.
2. Himmel, M. E., Ding, S. Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W. & Foust, T. D. (2007). Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science*, **315**, 804–807.
3. Bayer, E. A., Belaich, J. P., Shoham, Y. & Lamed, R. (2004). The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu. Rev. Microbiol.* **58**, 521–554.
4. Gold, N. D. & Martin, V. J. J. (2007). Global view of the *Clostridium thermocellum* cellulosome revealed by quantitative proteomic analysis. *J. Bacteriol.* **189**, 6787–6795.

5. Zverlov, V. V., Velikodvorskaya, G. V., Schwarz, W. H., Bronnenmeier, K., Kellermann, J. & Staudenbauer, W. L. (1998). Multidomain structure and cellulosomal localization of the *Clostridium thermocellum* cellobiohydrolase CbhA. *J. Bacteriol.* **180**, 3091–3099.
6. Vantilbeurgh, H., Tomme, P., Claeyssens, M., Bhikhabhai, R. & Pettersson, G. (1986). Limited proteolysis of the cellobiohydrolase I from *Trichoderma reesei*—separation of functional domains. *FEBS Lett.* **204**, 223–227.
7. Boraston, A. B., McLean, B. W., Kormos, J. M., Alam, M., Gilkes, N. R., Haynes, C. A. *et al.* (1999). Carbohydrate-binding modules: diversity of structure and function. In *Recent Advances in Carbohydrate Bioengineering*, pp. 202–211, Royal Society of Chemistry, Cambridge, UK.
8. Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. (2004). Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J.* **382**, 769–781.
9. Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. (2009). The Carbohydrate-Active enZYmes Database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* **37**, D233–D238.
10. Bu, L., Beckham, G. T., Crowley, M. F., Chang, C. H., Matthews, J. F., Bomble, Y. J. *et al.* (2009). The energy landscape for the interaction of the family 1 carbohydrate-binding module and the cellulose surface is altered by hydrolyzed glycosidic bonds. *J. Phys. Chem. B*, **113**, 10994–11002.
11. Schubot, F. D., Kataeva, I. A., Chang, J., Shah, A. K., Ljungdahl, L. G., Rose, J. P. & Wang, B. C. (2004). Structural basis for the exocellulase activity of the cellobiohydrolase CbhA from *Clostridium thermocellum*. *Biochemistry*, **43**, 1163–1170.
12. Jindou, S., Petkun, S., Shimon, L., Bayer, E. A., Lamed, R. & Frolow, F. (2007). Crystallization and preliminary diffraction studies of CBM3b of cellobiohydrolase 9A from *Clostridium thermocellum*. *Acta Crystallogr. Sect. F*, **63**, 1044–1047.
13. Gerwig, G. J., Kamerling, J. P., Vliegenthart, J. F. G., Morag, E., Lamed, R. & Bayer, E. A. (1993). The nature of the carbohydrate–peptide linkage region in glycoproteins from the cellulosomes of *Clostridium thermocellum* and *Bacteroides cellulosolvens*. *J. Biol. Chem.* **268**, 26956–26960.
14. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. Sect. D*, **53**, 240–255.
15. Gibrat, J. F., Madej, T. & Bryant, S. H. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377–385.
16. Holm, L., Kaariainen, S., Rosenstrom, P. & Schenkel, A. (2008). Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**, 2780–2781.
17. Hashimoto, H. (2006). Recent structural studies of carbohydrate-binding modules. *Cell. Mol. Life Sci.* **63**, 2954–2967.
18. Pell, G., Williamson, M. P., Walters, C., Du, H. M., Gilbert, H. J. & Bolam, D. N. (2003). Importance of hydrophobic and polar residues in ligand binding in the family 15 carbohydrate-binding module from *Cellvibrio japonicus* Xyn10C. *Biochemistry*, **42**, 9316–9323.
19. Park, S., Johnson, D. K., Ishizawa, C. I., Parilla, P. A. & Davis, M. F. (2009). Measuring the crystallinity index of cellulose by solid state C-13 nuclear magnetic resonance. *Cellulose*, **16**, 641–647.
20. Lamed, R., Kenig, R., Setter, E. & Bayer, E. A. (1985). Major characteristics of the cellulolytic system of *Clostridium thermocellum* coincide with those of the purified cellulosome. *Enzyme Microb. Technol.* **7**, 37–41.
21. Kataeva, I. A., Seidel, R. D., Li, X. L. & Ljungdahl, L. G. (2001). Properties and mutation analysis of the CelK cellulose-binding domain from the *Clostridium thermocellum* cellulosome. *J. Bacteriol.* **183**, 1552–1559.
22. Zverlov, V. V., Volkov, I. Y., Velikodvorskaya, G. A. & Schwarz, W. H. (2001). The binding pattern of two carbohydrate-binding modules of laminarinase Lam16A from *Thermotoga neapolitana*: differences in beta-glucan binding within family CBM4. *Microbiology*, **147**, 621–629.
23. Simpson, P. J., Xie, H., Bolam, D. N., Gilbert, H. J. & Williamson, M. P. (2000). The structural basis for the ligand specificity of family 2 carbohydrate-binding modules. *J. Biol. Chem.* **275**, 41137–41142.
24. Collaborative Computational Project, Number 4. (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. Sect. D*, **50**, 760–763.
25. Keegan, R. M. & Winn, M. D. (2008). MrBUMP: an automated pipeline for molecular replacement. *Acta Crystallogr. Sect. D*, **64**, 119–124.
26. Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
27. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a Structural Classification of Proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
28. Vagin, A. & Teplyakov, A. (1997). MOLREP: an automated program for molecular replacement. *J. Appl. Crystallogr.* **30**, 1022–1025.
29. Boraston, A. B., Nurizzo, D., Notenboom, V., Ducros, V., Rose, D. R., Kilburn, D. G. & Davies, G. J. (2002). Differential oligosaccharide recognition by evolutionarily-related beta-1,4 and beta-1,3 glucan-binding modules. *J. Mol. Biol.* **319**, 1143–1156.
30. Cohen, S. X., Ben, J. M., Long, F., Vagin, A., Knipscheer, P., Lebbink, J. *et al.* (2008). ARP/wARP and molecular replacement: the next generation. *Acta Crystallogr. Sect. D*, **64**, 49–60.
31. Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr. Sect. D*, **60**, 2126–2132.
32. Painter, J. & Merritt, E. A. (2006). Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr. Sect. D*, **62**, 439–450.
33. Painter, J. & Merritt, E. A. (2006). TLSMD Web server for the generation of multi-group TLS models. *J. Appl. Crystallogr.* **39**, 109–111.
34. Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X. *et al.* (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375–W383.
35. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. & Madden, T. L. (2008). NCBI

BLAST: a better Web interface. *Nucleic Acids Res.* **36**, W5–W9.

36. Abagyan, R., Totrov, M. & Kuznetsov, D. (1994). ICM—a new method for protein modeling and design—applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **15**, 488–506.
37. Abagyan, R. & Totrov, M. (1994). Biased probability Monte-Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235**, 983–1002.
38. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J. *et al.* (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, **102**, 3586–3616.
39. Mackerell, A. D., Jr, Feig, M. & Brooks, C. L., III (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **25**, 1400–1415.
40. Brooks, B. R., Brooks, C. L., III, Mackerell, A. D., Jr, Nilsson, L., Petrella, R. J., Roux, B. *et al.* (2009). CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614.
41. Guvench, O., Greene, S. N., Kamath, G., Brady, J. W., Venable, R. M., Pastor, R. W. & Mackerell, A. D., Jr (2008). Additive empirical force field for hexopyranose monosaccharides. *J. Comput. Chem.* **29**, 2543–2564.
42. Guvench, O., Hatcher, E., Venable, R. M., Pastor, R. W. & MacKerell, A. D. (2009). CHARMM additive all-atom force field for glycosidic linkages between hexopyranoses. *J. Chem. Theory Comput.* **5**, 2353–2370.
43. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935.
44. Durell, S. R., Brooks, B. R. & Ben-Naim, A. (2002). Solvent-induced forces between two hydrophilic groups. *J. Phys. Chem.* **98**, 2198–2202.
45. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H. & Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593.
46. Macke, T. & Case, D. A. (1998). Modeling unusual nucleic acid structures. In *Molecular Modeling of Nucleic Acids* (Leontes, N. B. & SantaLucia, J., eds), pp. 379–393, American Chemical Society, Washington, DC.
47. Macke T., Svrcek-Seiler W. A., Brown R. A., Kolossvary I., Bomble Y. J. & Case D. A. (2009). NAB version 6.
48. Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. A., Wang, J., Duke, R. E. *et al.* (2008). Amber 10. University of California, San Francisco, CA.
49. Case, D. A., Cheatham, T. E., III, Darden, T., Gohlke, H., Luo, R., Merz, K. M., Jr *et al.* (2005). The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688.
50. Wang, J., Cieplak, P. & Kollman, P. A. (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. *J. Comput. Chem.* **21**, 1049–1074.
51. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C. (2006). Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinf.* **65**, 712–725.
52. Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. (1995). Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **246**, 122–129.
53. Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. (1996). Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* **100**, 19824–19839.
54. Lehoucq, R., Sorensen, D. C. & Yang, C. (1998). *Arpack User's Guide: Solution of Large-Scale Eigenvalue Problems With Implicitly Restorted Arnoldi Methods.* SIAM, Philadelphia, PA.
55. Lakowicz, J. R. (1999). *Principles of Fluorescence Spectroscopy.* Springer, New York, NY.
56. Xu, Q., Morrison, M., Nelson, K. E., Bayer, E. A., Atamna, N. & Lamed, R. (2004). A novel family of carbohydrate-binding modules identified with *Ruminococcus albus* proteins. *FEBS Lett.* **566**, 11–16.
57. Jung, H., Wilson, D. B. & Walker, L. P. (2003). Binding and reversibility of *Thermobifida fusca* Cel5A, Cel6B, and Cel48A and their respective catalytic domains to bacterial microcrystalline cellulose. *Biotechnol. Bioeng.* **84**, 151–159.
58. Naran, R., Black, S., Decker, S. & Azadi, P. (2009). Extraction and characterization of native heteroxylans from delignified corn stover and aspen. *Cellulose*, **16**, 661–6675.