# Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose

José Tormo[1,2,3,4], Raphael Lamed[5], Arthur J.Chirino[1,2], Ely Morag[6], Edward A.Bayer[6], Yuval Shoham[7] and Thomas A.Steitz[1,2,8]

[1]Department of Molecular Biophysics and Biochemistry, [8]Department of Chemistry and [2]Howard Hughes Medical Institute, Yale University, New Haven, CT, USA, [5]Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Ramat Aviv, Israel, [6]Department of Membrane Research and Biophysics, The Weizmann Institute of Science, Rehovot, Israel and [7]Department of Food Engineering and Biotechnology, Technion-Israel Institute of Technology, Haifa, Israel

[3]Present address: Consejo Superior de Investigaciones Científicas, Centro de Investigación y Desarrollo, Jordi Girona 18–26, E-08034 Barcelona, Spain

[4]Corresponding author

**The crystal structure of a family-III cellulose-binding domain (CBD) from the cellulosomal scaffoldin subunit of *Clostridium thermocellum* has been determined at 1.75 Å resolution. The protein forms a nine-stranded β sandwich with a jelly roll topology and binds a calcium ion. Conserved, surface-exposed residues map into two defined surfaces located on opposite sides of the molecule. One of these faces is dominated by a planar linear strip of aromatic and polar residues which are proposed to interact with crystalline cellulose. The other conserved residues are contained in a shallow groove, the function of which is currently unknown, and which has not been observed previously in other families of CBDs. On the basis of modeling studies combined with comparisons of recently determined NMR structures for other CBDs, a general model for the binding of CBDs to cellulose is presented. Although the proposed binding of the CBD to cellulose is essentially a surface interaction, specific types and combinations of amino acids appear to interact selectively with glucose moieties positioned on three adjacent chains of the cellulose surface. The major interaction is characterized by the planar strip of aromatic residues, which align along one of the chains. In addition, polar amino acid residues are proposed to anchor the CBD molecule to two other adjacent chains of crystalline cellulose.**

*Keywords*: cellulose-binding domain (CBD)/cellulosome *Clostridium thermocellum*/crystal structure/scaffoldin

## Introduction

Cellulose, the major polysaccharide component of plant cell walls, is degraded in nature by the concerted action of a number of bacterial and fungal organisms (Béguin and Aubert, 1994). Cellulose degradation poses an interesting problem, not only from the biotechnological point of view (cellulose is the most abundant, renewable source of organic compound on the planet), but also from the point of view of basic research, since the insoluble nature and inherent stability of crystalline cellulose constitute a challenge for its enzymatic hydrolysis.

The initial event in the cellulose degradation process is the binding of the cellulolytic enzyme(s) or the entire microorganism to the cellulose substrate. This binding is mediated by a separate domain, named cellulose-binding domain or CBD. CBDs appear to play a multiple role in cellulolysis. They often comprise a distinct domain of a free enzyme, linked to one or more catalytic domains (not necessarily cellulases). In some cases, they occur in a discrete subunit, together with additional non-catalytic domains which serve to integrate the catalytic subunits into a multifunctional enzyme complex, the cellulosome (Lamed *et al.*, 1983; Lamed and Bayer, 1988). When the cellulases or cellulosomes are attached to the cell surface, their CBDs mediate the binding of the cell to cellulose. In addition to their more obvious role as a targeting vehicle, it has been proposed that CBDs may mediate the non-hydrolytic disruption of cellulose fibers, thereby facilitating subsequent enzymatic degradation by the catalytic domains (Din *et al.*, 1991, 1994a).

Over 100 different CBD sequences have already been identified, which range in size from only 33 to over 170 amino acid residues. These CBDs can be grouped into distinctive families on the basis of amino acid sequence similarities (Gilkes *et al.*, 1991; Tomme *et al.*, 1995). The smallest and simplest type of CBD, comprising family I, is found only in fungal cellulases and contains between 33 to 36 residues. The three-dimensional structure of one member of this family, CBH1-CBD, derived from the cellobiohydrolase I of *Trichoderma reesei*, has been determined by NMR spectroscopy (Kraulis *et al.*, 1989). The secondary structure of CBH1-CBD is organized into a wedge-shaped irregular β sheet. One face of the molecule, dominated by three conserved tyrosine side chains, forms a hydrophobic and planar surface that has been shown to be involved in cellulose binding (Linder *et al.*, 1995; Reinikainen *et al.*, 1995).

In contrast to the small CBDs observed in the fungal cellulases, the CBDs of bacteria are substantially larger. The structure of one of these, $C_{ex}$CBD, a 110-residue member of family-II CBDs derived from a β-1,4-glycanase of *Cellulomonas fimi*, has recently been solved using NMR spectroscopy (Xu *et al.*, 1995). $C_{ex}$CBD forms an elongated, nine-stranded β barrel, and the substrate-binding site appears to include three solvent-exposed tryptophans, together with other hydrophilic residues, located on one edge of the barrel.

Family-III CBDs comprise ~150 amino acid residues. They have been identified in many different bacterial enzymes, and in the non-hydrolytic proteins CbpA (Shoseyov *et al.*, 1992), CipA (Gerngross *et al.*, 1993), CipB (Poole *et al.*, 1992) and CipC (Pagès *et al.*, 1996) which are responsible for the structural organization of the cellulosomes present in *Clostridium cellulovorans* (CbpA), *Clostridium thermocellum* (CipA and CipB from strains ATCC 27405 and YS, respectively), and *Clostridium cellulolyticum* (CipC). These non-hydrolytic cellulosomal structural proteins, named scaffoldins (Bayer *et al.*, 1994), consist of a single large polypeptide chain of similar size (~1800 amino acid residues). Besides a discrete number of domains which interact with different catalytic subunits to form the cohesive cellulosome structure, the scaffoldins characterized so far contain a single CBD.

In this communication, we report the high-resolution crystal structure of the CBD from the cellulosomal scaffoldin subunit of *C.thermocellum*. The structure of this family-III CBD is compared and contrasted with other known CBD structures, highlighting important novel features. We propose its mode of binding and interaction with the cellulose substrate, based on available structural and mutagenesis data for this and related CBDs.

## Results

### Crystal structure determination

The CBD of the scaffoldin subunit Cip (Cip-CBD) of the cellulosome from *C.thermocellum* (comprising residues 361 to 527) was expressed in *Escherichia coli*, purified and crystallized as previously described (Lamed *et al.*, 1994; Morag *et al.*, 1995). The crystals were grown by vapor diffusion using PEG as a precipitant. They belong to the monoclinic space group C2, and contain two molecules in the asymmetric unit. The crystal structure of Cip-CBD was solved by conventional multiple isomorphous replacement including anomalous scattering (MIRAS) techniques, using two heavy atom derivatives, with data extending to 2.5 Å resolution (see Materials and methods). The MIRAS phases, although of good quality, were further improved by solvent flattening, non-crystallographic symmetry averaging, and histogram matching. The combination of these procedures provided a high-quality electron density map which was readily interpretable. The final atomic model has been refined to an *R* factor of 0.193 using 2σ data extending from 10.0 to 1.75 Å resolution (Table I). Deviations from ideal stereochemistry and the distribution of conformational angles about the expected values are within the ranges expected for well-refined X-ray structures determined at this resolution. The dihedral angles of the polypeptide backbone for 90% of the non-glycine residues fall within the 'most favored regions' of the Ramachandran plot, as defined by the program PROCHECK (Laskowski *et al.*, 1993); the rest being inside the 'additional allowed regions'. Seven residues at the N-terminus and five residues at the C-terminus do not show electron density in the final refined model and are probably disordered. The rest of the residues, apart from a few solvent-exposed side chains, present well-defined electron density for both molecules in the asymmetric unit. When the two independent molecules in the asymmetric unit are superimposed they have an r.m.s. fit of 0.23 Å and 0.49 Å for main chain and all atoms, respectively. The major differences between the two copies are localized in flexible loops connecting β strands (loops 5–6 and 8–9). These solvent-exposed regions form the rims of the concave groove on one β sheet, and are in different crystal environments.

**Table I.** Model refinement

| | |
|---|---|
| Resolution range (Å) | 10.0–1.75 |
| *R* factor (%) | 19.3 |
| Number of reflections with $F>2\sigma$ | 26 090 |
| Number of protein atoms | 2436 |
| Number of calcium ions | 2 |
| Number of water molecules | 280 |
| Average temperature factor (Å$^2$) | |
| main chain atoms | 27.0 |
| side chain atoms | 29.6 |
| calcium atoms | 22.8 |
| R.m.s. deviations B-factors between bonded atoms (Å$^2$) | 2.4 |
| R.m.s. deviations from ideal values | |
| bond lengths (Å) | 0.010 |
| bond angles (°) | 1.45 |
| dihedral angles (°) | 28.67 |
| improper torsion angles (°) | 1.40 |
| R.m.s. deviations between subunits | |
| main-chain atoms (Å) | 0.23 |
| all atoms (Å) | 0.49 |
| B-factors equivalent atoms (Å$^2$) | 1.83 |

Throughout the discussion, amino acids in Cip-CBD shall be numbered starting at the first residue which shows electron density as Asn1, which corresponds to residue 368 in CipA from *C.thermocellum* strain ATCC 27405 (Gerngross *et al.*, 1993). The last residue observed in the electron density maps is Pro155 (residue 522 in CipA). This residue is well conserved in family-III CBDs, and constitutes the ultimate or penultimate amino acid in those proteins where the CBD is located at the C-terminus.

### Description of the overall structure of Cip-CBD

The crystal structure shows that 155 amino acid residues of Cip-CBD fold into a single, compact domain that has an overall prismatic shape with approximate dimensions of 30 Å×30 Å×45 Å. Cip-CBD belongs to the all-β family of proteins and is arranged in two antiparallel β sheets that stack face-to-face to form a β sandwich with jelly roll topology (Figures 1 and 2). Except for the nine major β strands arranged in the two sheets, the protein consists mostly of loops connecting the secondary structure elements. However, the connections between strands 1–2, 7–8, on one side of the molecule, and 2–3, 6–7 on the other side, form very short antiparallel β strands. Since they do not form part of the β sheets, they have not been included in the sequential numbering of the β strands, and have been labeled 1′, 2′, 6′, 7′, respectively. The short β strand 4′ forms a β-hairpin with part of strand 4, but it has not been included either because it is not present in the majority of family-III CBDs.

In the following discussion the β sheet comprised of strands 1, 2, 7, and 4 will be referred to as the bottom sheet; whereas the top sheet contains strands 9, 8, 3, 6, 5. The two sheets show different curvatures. The bottom sheet is composed of long β strands (average length nine residues) and presents a smoothly bent surface. Strand 4, located on one edge of the sheet, is characterized by a β bulge that projects two consecutive residues, Asp56 and

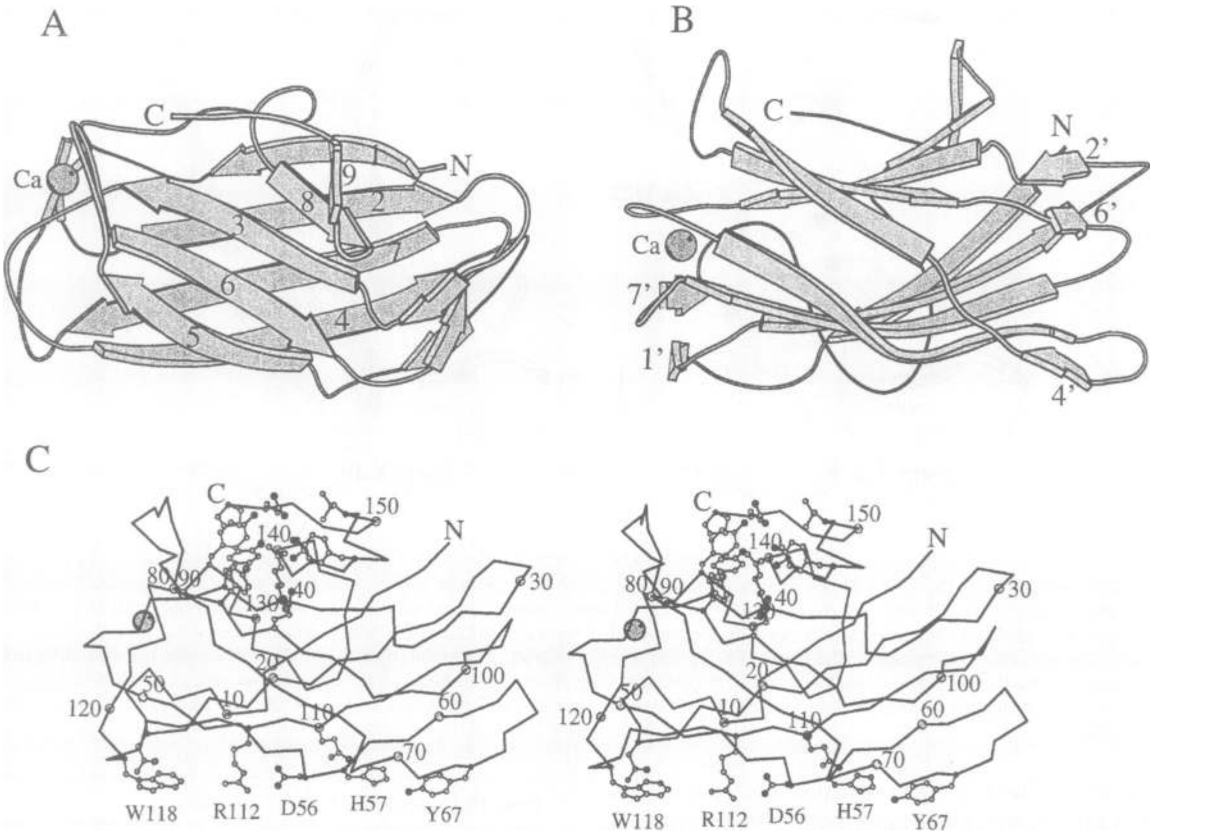
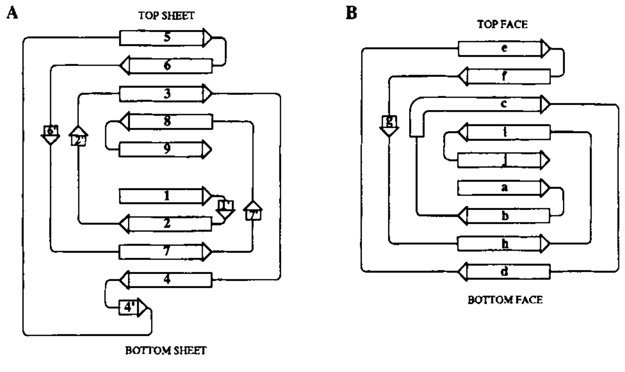


**Fig. 1.** Overall structural organization of the CBD. (**A**) Ribbon diagram of Cip-CBD. The C- and N-termini are labeled C and N, respectively, and the position of the calcium ion (Ca) is shown. β Strands are depicted as arrows and unstructured loops as tubes. Strands that form the two β sheets are labeled 1–9. Short strands, not included in the β sheets (1′, 2′, 6′, 7′), and strand 4′, which forms a short β hairpin with β strand 4, are labeled in (B), for clarity. (**B**) 90° rotation around the horizontal axis of the CBD molecule in (A). (**C**) Stereo diagram of the $C_\alpha$ trace of Cip-CBD with every tenth $C_\alpha$ position labeled. The molecule has been rotated 40° around the horizontal axis with respect to the view in (B). Side chains of conserved surface-exposed residues are shown, and those forming the aromatic strip proposed to interact with a single glucose chain of crystalline cellulose are labeled. The figure was generated using MOLSCRIPT (Kraulis, 1991).

His57, on the protein surface. The top sheet, made of shorter strands (average length six residues), is more curved and presents a concave surface with a shallow groove. The interface between the sheets is packed mostly with conserved hydrophobic side chains but also contains two hydrophilic patches formed by hydrogen-bonded side chains of residues from both sheets and connecting loops, as well as a few buried solvent molecules. The two patches cluster on the end of the sandwich which binds a $Ca^{2+}$ ion. One group involves side chains from residues Asn8, Tyr41, Gln123, Asp126, and three water molecules. Asp126 and a solvent molecule are $Ca^{2+}$-binding residues (Figure 3). The second patch contains residues Gln51, Tyr43 and Tyr121.

The structure-based alignment of the amino acid sequences of bacterial family-III CBDs shows that most of the secondary structure elements observed in the Cip-CBD structure are conserved throughout the family (Figures 4 and 5). Only the solvent-exposed β hairpin formed by strands 4 and 4′ is not conserved in the group and appears to be an insertion which is present exclusively in the well-characterized cellulosomal scaffoldin proteins (i.e. CipA and CipB from two different strains of *C.thermocellum*, CipC from *C.cellulolyticum*, and CbpA from *C.cellulovorans*). The $Ca^{2+}$-binding residues and the hydrophilic patch linked to them seem also to be conserved, whereas the second hydrophilic patch is absent in a few members of the family.



*Clostridium thermocellum* Cip-CBD

(Family III CBD)

*Cellulomonas fimi* Cex-CBD

(Family II CBD)

**Fig. 2.** Comparative strand topology in families-II and -III CBDs. Schematic diagrams of the topologies of the families-II and -III CBDs showing the β sheet architectures and labelings of secondary structure elements. The family-II CBD topology corresponds to that of $C_{ex}$-CBD (i.e. the CBD of the glycanase $C_{ex}$ from *C.fimi*). Both families exhibit very similar folds with the same jelly roll topology; however, in family-III CBDs the nine β strands are arranged in two β sheets, whereas in family II they are arranged in a half-closed β barrel. The top sheet of the *C.thermocellum* CBD comprises β strands 5, 6, 3, 8 and 9, and the bottom sheet includes β strands 1, 2, 7 and 4. The top face of the *Cellulomonas* CBD contains β strands e, f, c, i and j, and the bottom face contains strands a, b, h and d.

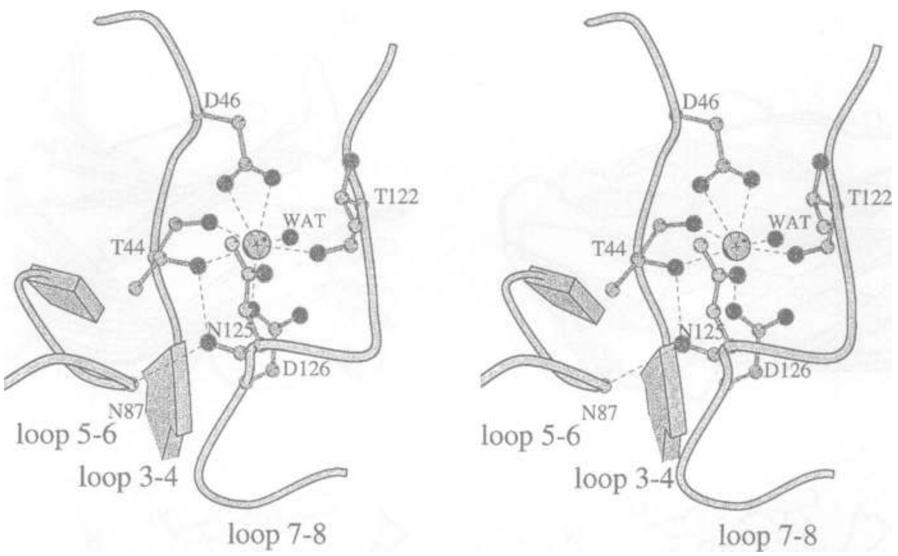


**Fig. 3.** Calcium-binding site of Cip-CBD. Stereo diagram showing the amino acid residues and the structured water molecule which participate in the octacoordination of the calcium ion. The oxygen atoms are located approximately at the vertices of a pentagonal bipyramid. The hydroxyl and carbonyl oxygens of Thr44, the amine oxygen of Asn125, the carbonyl oxygen of Thr122, and a water molecule lie in a distorted plane, approximately perpendicular to the plane of the figure. The carboxylates of Asp46, a bidentate ligand, and Asp126 are the trans-axial ligands and are located above and below the pentagonal equatorial plane, respectively. The close contact between the carbonyl oxygen of Asn125 and the $C_\alpha$ carbon of Asn87, located at the shallow groove on the top sheet ($C_\alpha$–O contact distances of 3.08 Å and 3.11 Å for the two molecules in the asymmetric unit), is also shown.

Comparison of the Cip-CBD structure with the known database of protein structures carried out with the program DALI (Holm and Sander, 1993) reveals a close relationship to the bacterial family-II CBD of the β-1,4-glycanase $C_{ex}$ ($C_{ex}$CBD) from *C.fimi*, and to the receptor-binding domain of diphtheria toxin (Choe *et al.*, 1992). Both bacterial CBDs share a common jelly roll topology (Figures 2 and 6), despite the different sizes of the two proteins (155 residues in Cip-CBD and 110 residues in $C_{ex}$CBD) and the lack of recognizable amino acid sequence similarities between the two families of CBDs. The same fold is adopted by the receptor-binding domain of diphtheria toxin, which is involved in the binding of the toxin to a cell-surface receptor. This binding domain has, however, an additional short N-terminal β strand, located at the edge of the bottom sheet. There is also a topological similarity between a subset of the secondary structural elements of these three domains and part of the P-domain of the capsid protein from tomato bushy stunt virus (TBSV) (Harrison *et al.*, 1978). β Strands 1–8 of Cip-CBD are topologically equivalent to β strands 3–10 of the P-domain in TBSV.

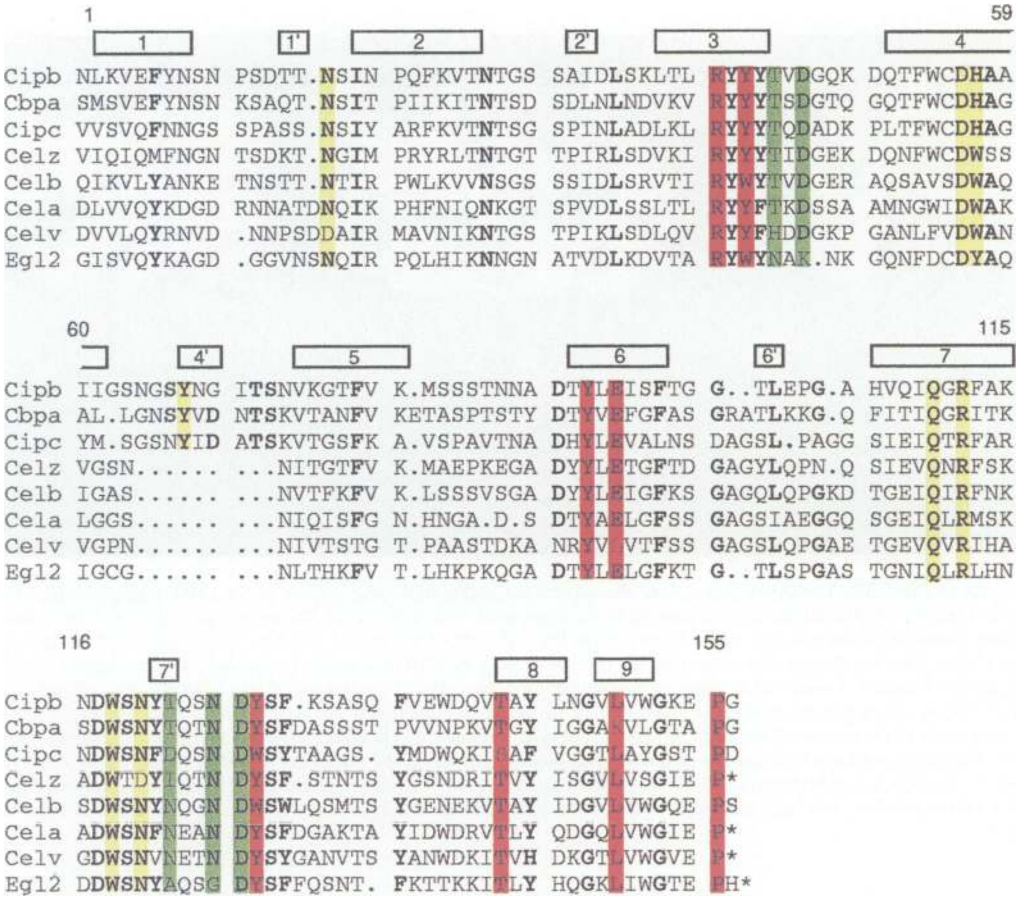
### *Cip-CBD contains a $Ca^{2+}$-binding site*

During the course of refinement, a strong peak was observed in each independent protein molecule while inspecting difference Fourier maps. Based on its coordination chemistry, it was modeled as $Ca^{2+}$, and the assignment was later corroborated by atomic absorption (data not shown). Although no $Ca^{2+}$ was added during the purification and crystallization processes, $Ca^{2+}$ seems to bind very tightly to Cip-CBD. The thermal parameters of the $Ca^{2+}$ ions, refined with full occupancy, are 22.2 $Å^2$ and 23.6 $Å^2$, for the two copies in the asymmetric unit (the average temperature factor for the protein main chain atoms is 27.0 $Å^2$).

**Table II.** Distances for $Ca^{2+}$ coordination (Å)

| | Subunit 1 | Subunit 2 |
|---|---|---|
| Thr44 O | 2.44 | 2.52 |
| Thr44 Oγ1 | 2.45 | 2.47 |
| Asp46 Oδ1 | 2.80 | 2.91 |
| Asp46 Oδ2 | 2.19 | 2.28 |
| Thr122 O | 2.28 | 2.30 |
| Asn125 Oδ1 | 2.57 | 2.57 |
| Asp126 Oδ1 | 2.40 | 2.42 |
| Water | 2.40 | 2.65 |

The $Ca^{2+}$ is not accessible to the solvent and binds in a cavity formed between loops 3–4 and 7–8, both of them cross-over connections between the two β sheets (Figure 1). The coordination sphere of $Ca^{2+}$ consists of oxygen atoms from residues Thr44 and Asp46 on loop 3–4, and residues Thr122, Asn125 and Asp126, on loop 7–8 (Figure 3 and Table II). Residues Asp46 and Asp126 bind $Ca^{2+}$ by means of their carboxylate groups. Asp126 uses only one of the carboxylate oxygens. On the other hand, the second carboxylate oxygen of Asp46 is located at 2.8 Å from $Ca^{2+}$ and, therefore, seems to constitute a bidentate ligand. Other ligands are the amide oxygen of Asn125, the hydroxyl and carbonyl oxygens of Thr44, and the carbonyl oxygen of Thr122. The eighth $Ca^{2+}$ ligand is provided by a water molecule, which is buried in the structure, and involved in one of the hydrophilic patches in the protein core. Cip-CBD thus presents an octacoordinate $Ca^{2+}$-binding site. The geometric arrangement of the ligands resembles a typical pentagonal bipyramid, as observed in other protein $Ca^{2+}$-binding sites (McPhalen *et al.*, 1991), with one of the vertices shared by the oxygen atoms of the carboxylate group of Asp46. The hydroxyl and carbonyl oxygens of Thr44, the amide oxygen of Asn125, the carbonyl oxygen of Thr122, and the water molecule lie in a distorted plane, whereas one oxygen of



```
Cipb NLKVEFYNSN PSDTT.NSIN PQFKVTNTGS SAIDLSKLTL RYYYTVDGQK DQTFWCDHAA
Cbpa SMSVEFYNSN KSAQT.NSIT PIIKITNTSD SDLNLNDVKV RYYYTSDGTQ GQTFWCDHAG
Cipc VVSVQFNNGS SPASS.NSIY ARFKVTNTSG SPINLADLKL RYYYTQDADK PLTFWCDHAG
Celz VIQIQMFNGN TSDKT.NGIM PRYRLTNTGT TPIRLSDVKI RYYYTIDGEK DQNFWCDWSS
Celb QIKVLYANKE TNSTT.NTIR PWLKVVNSGS SSIDLSRVTI RYWYTVDGER AQSAVSDWAQ
Cela DLVVQYKDGD RNNATDNQIK PHFNIQNKGT SPVDLSSLTL RYYFTKDSSA AMNGWIDWAK
Celv DVVLQYRNVD .NNPSDDAIR MAVNIKNTGS TPIKLSDLQV RYYFHDDGKP GANLFVDWAN
Egl2 GISVQYKAGD .GGVNSNQIR PQLHIKNNGN ATVDLKDVTA RYWYNAK.NK GQNFDCDYAQ

Cipb IIGSNGSYNG ITSNVKGTFV K.MSSSTNNA DTYLEISFTG G..TLEPG.A HVQIQGRFAK
Cbpa AL.LGNSYVD NTSKVTANFV KETASPTSTY DTYVEFGFAS GRATLKKG.Q FITIQGRITK
Cipc YM.SGSNYID ATSKVTGSFK A.VSPAVTNA DHYLEVALNS DAGSL.PAGG SIEIQTRFAR
Celz VGSN...... ...NITGTFV K.MAEPKEGA DYYLETGFTD GAGYLQPN.Q SIEVQNRFSK
Celb IGAS...... ...NVTFKFV K.LSSSVSGA DYYLEIGFKS GAGQLQPGKD TGEIQIRFNK
Cela LGGS...... ...NIQISFG N.HNGA.D.S DTYAELGFSS GAGSIAEGGQ SGEIQLRMSK
Celv VGPN...... ...NIVTSTG T.PAASTDKA NRYVEVTFSS GAGSLQPGAE TGEVQVRIHA
Egl2 IGCG...... ...NLTHKFV T.LHKPKQGA DTYLELGFKT G..TLSPGAS TGNIQLRLHN

Cipb NDWSNYTQSN DYSF.KSASQ FVEWDQVTAY LNGVLVWGKE PG
Cbpa SDWSNYTQTN DYSFDASSST PVVNPKVTGY IGGAKVLGTA PG
Cipc NDWSNFDQSN DWSYTAAGS. YMDWQKISAF VGGTLAYGST PD
Celz ADWTDYTQTN DYSF.STNTS YGSNDRITVY ISGVLVSGIE P*
Celb SDWSNYNQGN DWSWLQSMTS YGENEKVTAY IDGVLVWGQE PS
Cela ADWSNFNEAN DYSFDGAKTA YIDWDRVTLY QDGQLVWGIE P*
Celv GDWSNVNETN DYSYGANVTS YANWDKITVH DKGTLVWGVE P*
Egl2 DDWSNYAQSG DYSFFQSNT. FKTTKKITLY HQGKLIWGTE PH*
```

**Fig. 4.** Structure-based sequence alignment of selected CBDs from family III. Regions of secondary structure are marked and labeled. Conserved residues are shown in bold type. Residues at a given position are defined as conserved if seven out of the eight CBDs exhibit an identical or similar amino acid, where similarity is limited to aromatic residues (F, H, W, Y). Conserved surface-exposed residues, which are not involved in the stabilization of loops of secondary structure elements, cluster into different discrete regions, and have been color-coded as follows: yellow, conserved surface residues which form a planar strip at the bottom of the CBD molecule; red, conserved surface residues which form a shallow groove on the top of the CBD molecule; green, residues which interact with the calcium ion, where the non-conserved residue at position 122 interacts with the calcium via its backbone carbonyl group. The CBDs shown in the figure are from the following proteins: CipA = CipB, the scaffoldin subunit from *C.thermocellum* (Swiss-Prot accession number Q06851); CipC, the scaffoldin subunit of *C.cellulolyticum* (GenBank accession number U40345); CbpA, the scaffoldin subunit from *C.cellulovorans* (Swiss-Prot accession number P38058); CelZ, cellulase CelZ (Avicelase I) from *C.stercorarium* (Swiss-Prot accession number P23659); CelB, cellobiohydrolase/endocellulase from *Caldocellum saccharolyticum* (Swiss-Prot accession number P10474); CelA, endo-glucanase (cellulase A) from *Bacillus lautus* (Swiss-Prot accession number P29719); CelV, endo-glucanase V from *Erwinia carotovora* (Swiss-Prot accession number S39962); Egl2, endo-glucanase (cellulase Egl2) from *B.subtilis* (Swiss-Prot accession number A27198).

the carboxylate group of Asp126 constitutes the second trans-axial ligand.

### *Surface features in Cip-CBD*

Mapping of conserved residues on the surface of Cip-CBD shows that they tend to cluster into two discrete regions of the molecule (Figures 4 and 5). One set of residues, formed by four aromatic rings (Tyr42, Tyr91, Tyr127 and Tyr144), three charged or polar residues (Arg40, Glu93 and Thr142), Leu149 and Pro155 occupy the shallow groove on the top β sheet. The second conserved cluster is located on the opposite side of the molecule. This conserved surface is dominated by a planar linear array of aromatic and charged residues located on one edge of the bottom sheet (Asp56, His57, Tyr67, Arg112, Trp118), flanked on one side by additional conserved polar residues (Asn16, Gln110) positioned on the surface of the sheet.

The $Ca^{2+}$-binding site is located at a distance from the two conserved surfaces and is most probably involved in stabilizing the protein fold. It is also possible that $Ca^{2+}$ binding is required for the correct orientation of residues that form the shallow groove on the top sheet. Tyr127, one of the aromatic residues at the bottom of the groove, is located immediately adjacent to one of the $Ca^{2+}$-binding residues. Thr44 and Asn125, both of which are involved in $Ca^{2+}$ coordination, may also have a role in locking loop 5–6 in place through several contacts. The carbonyl of Asn125 is located a short distance (3.1 Å) from the $C_\alpha$ carbon of Asn87, within loop 5–6.

Interestingly, both conserved surfaces are formed by residues typically involved in protein–carbohydrate interactions which rely on hydrophobic van der Waals contacts, usually carried out by aromatic residues, and on hydrogen bonding networks by polar side-chains (Quiocho, 1993; Toone, 1994).

### *Structural comparison between family-II and family-III CBDs*

The NMR and crystal structures of $C_{ex}$CBD and Cip-CBD, which belong to CBDs of families II and III, respectively, show that both proteins share the same architecture, although no recognizable amino acid sequence similarity can be detected between the two

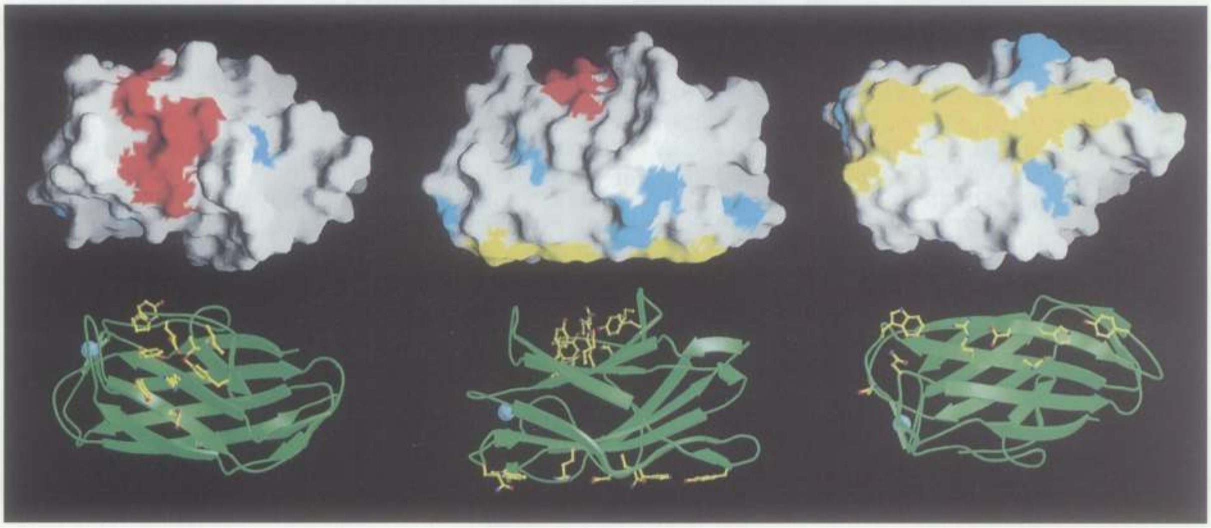


**Fig. 5.** Surface features in Cip-CBD. The upper views show the solvent-accessible molecular surface of the CBD, rendered with the GRASP program (Nicholls and Honig, 1993), and the lower views show the equivalent ribbon diagrams. The two central structures have the same orientation as the ribbon diagram shown in Figure 1B. The views shown at the left and right of the central one have been rotated 90° in opposite directions around the horizontal axis, showing the top and bottom views, respectively, of the CBD molecule. Conserved surface residues are color-coded, according to the legend to Figure 4. Conserved residues in yellow form a planar strip at the bottom of the molecule; red-colored residues form the shallow groove at the top of the molecule; residues in green participate in the coordination of the calcium ion. Other conserved residues are shown in light blue and correspond to the uncolored residues shown in bold type in Figure 4. The lower views show the two clusters of conserved solvent-exposed side chains, located on opposite sides of the molecule. The shallow groove on the top of the molecule is formed by four tyrosine residues (Tyr42, Tyr49, Tyr127, Tyr144), a salt bridge (Arg40, Glu93), Thr142 and Leu149. The flat conserved surface on the bottom of Cip-CBD, is proposed to interact with crystalline cellulose, and includes the amino acid residues which form the planar strip (from left to right: Trp118, Arg112, Asp56, His57, Tyr67).

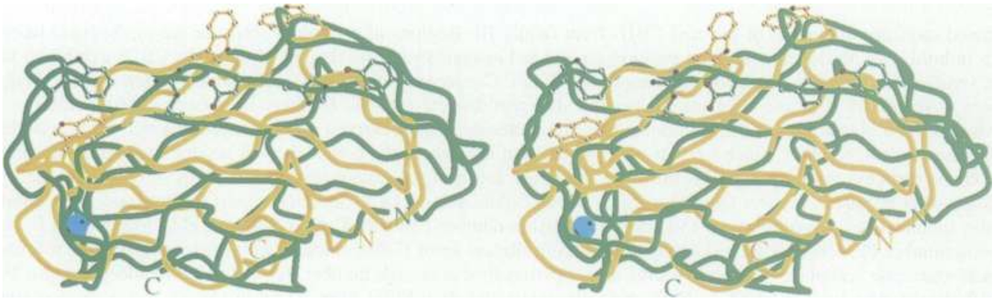


**Fig. 6.** Structural comparison between *C.thermocellum* Cip-CBD and *C.fimi* $C_{ex}$CBD. Stereoview of *C.thermocellum* Cip-CBD superimposed on *C.fimi* $C_{ex}$CBD. The $C_\alpha$ carbon traces of Cip-CBD and $C_{ex}$CBD are shown in green and orange lines, respectively. The blue sphere indicates the position of the $Ca^{2+}$ ion present only in Cip-CBD. The N- and C-termini of both proteins have also been labeled. The r.m.d.s. on 48 $C_\alpha$ carbon atoms selected from both β sheets and used to optimize the fit is 1.66 Å. Side chains of residues forming the aromatic strip in the cellulose-binding site of both proteins are also shown. For Cip-CBD, residues Trp118, Arg112, Asp56, His57 and Tyr67 are shown in green; whereas $C_{ex}$CBD residues Trp17, Trp54 and Trp72 are shown in orange. In this view, the aromatic strip of Cip-CBD is aligned along the horizontal direction, parallel to the plane of the figure. Although both planar strips are located at the same approximate location, on one edge of the bottom sheet, residues forming the strip are contributed by different secondary structure elements in the two proteins and the strips are oriented in different directions.

families (Figures 2 and 6). The structures were initially superimposed by an automatic structural alignment procedure (Stuart *et al.*, 1979), and the fit was maximized by a least-squares procedure using 48 $C_\alpha$ carbon atoms selected from the nine β strands. The r.m.s. deviation on test atoms is 1.66 Å.

Both CBDs are organized in two β sheets packed face-to-face in a β sandwich. The main differences between the two structures are in the length and conformation of the loops connecting the secondary structure elements. In addition, two long insertions in Cip-CBD are lacking in $C_{ex}$CBD, and the position of the C-terminus is also different in the two proteins. The differences in the connections between the β strands are most evident in the top sheet. In Cip-CBD, loop 5–6 is long and, together with the β hairpin produced by strands 8 and 9, they form the rims of the shallow groove on top of the molecule. On the other hand, in $C_{ex}$CBD, those connections are made by tight turns and no groove is evident at that location. An insertion of 10 amino acid residues at the beginning of strand 8 forms a loop that extends towards the bottom sheet, partially covering one edge of the β sandwich in Cip-CBD. In $C_{ex}$CBD, this loop is absent and the last β strand (g) is inserted between strands a and i on the top and bottom sheets respectively, closing the β sandwich on this side of the molecule. In Cip-CBD the C-terminus moves towards the groove; the last residue seen in the electron density maps, Pro155, is stacked against the phenyl ring of one of the conserved aromatic side chains (Tyr127) which form the bottom of the groove.

The last major difference in the overall architecture of both proteins is another long insertion in Cip-CBD with respect to $C_{ex}$CBD. This insertion of 20 amino acid residues is arranged in a β hairpin formed by a prolongation of strand 4 and a new strand, 4′, which may be a characteristic of the cellulosomal scaffoldin proteins.

Family-II and family-III CBDs are characterized by the conservation of separate sets of aromatic residues (Tomme *et al.*, 1995). As discussed above, in Cip-CBD these residues map into two defined surfaces. The conserved patch that forms the shallow groove on the top sheet in Cip-CBD does not, however, have its counterpart in $C_{ex}$CBD. The corresponding surface in this protein is rich in short-chain hydroxyl residues, but these are not generally conserved in family-II CBDs. On the other hand, the three conserved tryptophan residues in $C_{ex}$CBD, which have been shown by mutagenesis studies and NMR spectroscopy to be involved in cellulose binding (Poole *et al.*, 1993; Din *et al.*, 1994b; Xu *et al.*, 1995), appear to be located at a position analogous to that of the planar strip which contains aromatic and polar residues on the bottom sheet of Cip-CBD. Furthermore, residues Trp17, Trp57 and Trp72 in $C_{ex}$CBD, are also aligned along a planar surface exposed to the solvent. However, two out of three conserved aromatic residues in the respective planar strips are contributed by different structural elements in the two proteins (Figure 6). In Cip-CBD, His57, Tyr67 and Trp118 are located on β strand 4, loop 4–4′, and loop 7–8, respectively; whereas in $C_{ex}$CBD, Trp17, Trp54, and Trp72 sit on the loop connecting strands a and b (equivalent to 1–2 in Cip-CBD), strand d (equivalent to strand 4) and loop f–g (equivalent to 6–7), respectively. Therefore, only His57 in Cip-CBD and Trp54 in $C_{ex}$CBD are located in topologically equivalent positions.

### A mechanism for binding to cellulose, common to fungal and bacterial CBDs

Although there is no direct evidence linking any of the conserved surfaces in Cip-CBD to substrate binding, the striking similarity between the two linear strips of aromatic residues in both bacterial CBDs suggests that this surface contributes to the cellulose-binding site. Moreover, this linear strip of aromatic residues is not an exclusive characteristic of bacterial CBDs. The small fungal CBH1-CBD from *T.reesei*, belonging to family I, presents an analogous planar array of three tyrosine residues that has been demonstrated to be involved in cellulose binding (Linder *et al.*, 1995; Reinikainen *et al.*, 1995). Therefore, the fact that this surface feature is conserved among the three families of CBDs for which three-dimensional structures are known reinforces the assumption that such a structure constitutes a cellulose-binding site in all of them and also probably in CBDs from other families.

Cellulose, although chemically simple, presents a complex physical structure. The current models of native cellulose describe it as a composite of crystalline microfibrils embedded in a paracrystalline matrix. Furthermore, along the microfibrils the crystalline regions are interspersed with less-ordered or 'amorphous' sections. In the microcrystals of cellulose I—the naturally occurring form—the individual cellulose chains are organized in a parallel orientation in the direction of the long axis of the microfibril (Gardner and Blackwell, 1974). The cellulose chains are arranged in layered sheets where the chains are extensively hydrogen bonded with each other. On the other hand, there are no, or at most weak, hydrogen bonds across the sheets, which stack with each other mainly by van der Waals forces. The individual glycosyl residues are oriented with their ring planes nearly parallel to the surface of these sheets, which correspond to the (020) crystallographic plane, making them essentially hydrophobic. These sheets are arranged in such a way that the most exposed faces of the microfibrils correspond to the (110) and (1-10) crystallographic planes, whereas the (020) faces are located at the corners of the perfect crystals. We favor the (020) faces as the putative CBD binding site, as originally proposed by Reinikainen *et al.* (1995) for CBH1-CBD, because on these surfaces the glucopyranoside rings are fully exposed, making them available for hydrophobic interactions with the aromatic rings of the flat linear strips. Stacking of aromatic residues against the faces of sugar rings is a common structural feature of protein–carbohydrate interactions (Quiocho, 1993; Toone, 1994). Interestingly, the spacing of the aromatic rings is a multiple of that between the pyranoside rings along one cellulose chain, suggesting the orientation of the CBDs on the (020) faces (Figure 7). The possibility of a CBD binding along one cellulose chain was first suggested by Kraulis *et al.* (1989) for *T.reesei* CBH1-CBD, where three conserved tyrosine residues would stack onto every second glucose ring in the cellulose chain. In a similar manner, His57 and Tyr67 of Cip-CBD from *C.thermocellum* would lie on contiguous glucose units (the first and second rings), and Trp118 would stack onto the sixth ring, thus leaving a gap of three glucose units. The central moiety of this gap (the fourth ring) would then be aligned with the carboxylic group of Asp56 and the guanidinium group of Arg112, held together by a salt bridge. Consequently, four stacking interactions would be formed between residues of the Cip-CBD and four of six consecutive glucose units of the cellulose chain. $C_{ex}$CBD also spans six glucose units but involves only three aromatic rings, since the salt bridge is absent.

Our modeling of the three CBDs onto crystalline cellulose also shows that additional conserved polar residues are located at hydrogen bonding distances from polar groups on neighboring glucose chains, providing further protein–cellulose contacts (Figure 7 and Table III). For instance, in Cip-CBD these anchoring residues would establish hydrogen bonds with atoms on the second and third cellulose chains on one side of the aromatic strip. In $C_{ex}$CBD, polar residues are located at hydrogen bonding distances from atoms in chains on both sides of the aromatic strip, similar to the arrangement in the *T.reesei* CBD.

## Discussion

The CBDs of bacterial and fungal cellulolytic systems are believed to assist the hydrolysis of various forms of cellulose, since catalytic activities for insoluble crystalline forms of cellulose are drastically reduced if the CBDs are removed from the enzymes by proteolysis or genetic manipulation (Gilkes *et al.*, 1988; Tomme *et al.*, 1988). However, the actual role of CBDs in the hydrolysis of the substrate is not fully understood. They could simply

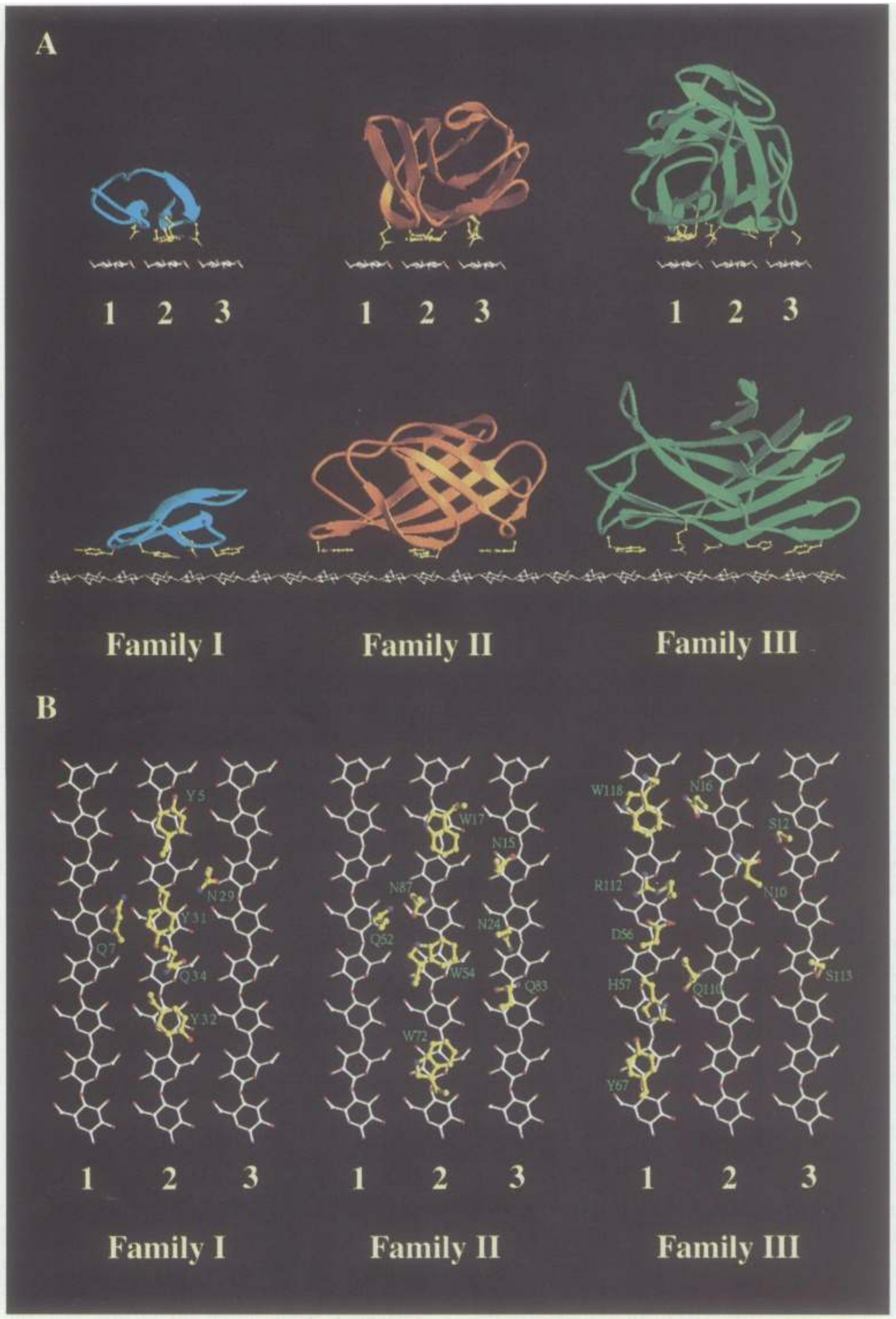


**Fig. 7.** Model for the interaction of families-I, -II and -III CBDs with cellulose. (**A**) The lower view shows the three CBDs aligned along a single cellulose chain. The spacing of the aromatic residues on the cellulose-binding sites is close to that of the sugar rings on the cellulose chain. In the upper panels the CBDs and the cellulose chains have been rotated 90°, separately, along the vertical axis, such that the CBDs are now aligned along three separate cellulose chains, designated 1, 2 and 3 for each structure. The family-I fungal CBD is shown in blue, the family-II bacterial CBD is shown in orange, and the family-III bacterial CBD is shown in green. Amino acid side chains, which appear to interact with the cellulose (see Table III), are shown in yellow. (**B**) Bird's-eye view of the residues proposed to interact with the cellulose chains. The residues, aligned along the cellulose chains, have been rotated 90° around the horizontal axis with respect to the orientations shown in the upper views of (A). The ribbon diagrams of the backbone traces have been omitted for clarity. The coordinates of CBH1-CBD (code 1cbh) and $C_{ex}$CBD (code 1exg) were taken from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977).

increase the effective concentration of the enzyme on the substrate, or they could have a more active role by facilitating the breakdown of the crystalline structure of cellulose (Din *et al.*, 1994a). The crystal structure of Cip-CBD, determined at 1.75 Å resolution, provides some suggestions for the biological role of this domain.

Cip-CBD adopts an uncommon jelly roll fold which is shared only by the bacterial family-II CBDs and the

**Table III.** Proposed interactions of CBD residues with successive chains of the crystalline cellulose lattice

| Chain I | Chain 2 | Chain 3 |
|---|---|---|
| Family-III bacterial CBD (from *Clostridium thermocellum* CipB) | | |
| (planar strip) | (anchor) | (anchor) |
| D56[b] | N10[a] | S12 |
| H57 | N16 | S133[a] |
| Y67 | Q110[a] | |
| R112[b] | | |
| W118 | | |
| Family-II bacterial CBD (from *Cellulomonas fimi* $C_{ex}$) | | |
| (anchor) | (planar strip) | (anchor) |
| Q52 | W17 | N15 |
| | W54[c] | N24 |
| | W72 | Q83 |
| | (N87)[d] | |
| Family-I fungal CBD (from *Trichoderma reesei* CBHI) | | |
| (anchor) | (planar strip) | (anchor) |
| Q7[a] | Y5 | N29 |
| | Y31 | |
| | Y32 | |
| | (Q34)[e] | |

[a] The side chains of the designated residues required reorientation, in order to form hydrogen bonds with the glucose residues in the indicated chains of cellulose.

[b] In the *C.thermocellum* CBD, D56 and R112 form a salt bridge within the planar strip, which is proposed to align across one of the glucose pyranose rings in the cellulose chain.

[c] In the *C.fimi* CBD, the orientation of W54 has been changed to the gauche+ rotamer in order to be aligned in a planar surface with the other aromatic residues.

[d] In the *C.fimi* CBD, N87 may stabilize the proposed orientation of W54 by forming a hydrogen bond with the indole nitrogen.

[e] In the *T.reesei* CBD, Q34 stabilizes the position of the Y32 side chain by bridging between the aromatic OH and the main chain carbonyl of Y5.

receptor-binding domain of diphtheria toxin. This fold is closer to the viral capsid β barrels than to other jelly roll sugar binding proteins, like lectins. In common with this group of proteins Cip-CBD contains a $Ca^{2+}$, which in this case seems to have a structural role. The presence of the metal ion in Cip-CBD was not known prior to the determination of its structure, and there are no direct data available on the influence of $Ca^{2+}$ binding on the stability and function of this CBD. However, in a mutation analysis of CbpA-CBD from the scaffoldin subunit of *C.cellulovorans* (Goldstein and Doi, 1994), the substitution of one of the aspartate residues that ligate the $Ca^{2+}$, equivalent to Asp126 in Cip-CBD, did not appear to have any effect on its binding to cellulose. Moreover, amino acid sequence alignments suggest that probably not all members of family-III CBDs contain a $Ca^{2+}$; e.g. in the CBD of the *Bacillus subtilis* cellulase Egl2, two of the $Ca^{2+}$-binding residues are not conserved.

Mapping of conserved residues among members of family-III CBDs revealed two surfaces that may have biological importance. One site is characterized by a linear and strikingly planar array of polar and aromatic residues that is probably involved in the binding of the CBD to cellulose, as will be discussed in more detail below. The orientation of these amino acid side chains is often stabilized by hydrogen bonding with other residues in the site. This may explain the fact that residues are conserved in pairs. For example, in the verified and well-characterized scaffoldin CBDs, a histidine is present at position 57 on β strand 4, which forms a hydrogen bond with the hydroxyl group of a tyrosine (Tyr67) located in a β hairpin. These two residues appear to be present only in these scaffoldin CBDs. In other members of family-III, where this tyrosine residue is absent, the histidine side chain is replaced by a tryptophan. Another example of pairwise conservation is the salt bridge formed by Asp56 and Arg112.

The detailed structure of the planar strip does not, however, seem to be conserved fully among all the CBDs that have been grouped under family III (Tomme *et al.*, 1995). On the basis of the differences in the fine structure of the planar strip, we can divide the CBDs presently grouped in this family into three different subgroups. One subgroup would include the CBDs of the scaffoldins, i.e. CipA, CipB, CipC and CbpA, which are distinguished from the other family-III CBDs by an extra β hairpin, formed by β strands 4 and 4′, which contributes an additional aromatic residue to the planar strip. The second subgroup would include most of the other family-III CBDs (see Figure 4) which do not exhibit the β strand 4′. Nevertheless, their three-dimensional structures would be expected to be very close to that of the scaffoldin CBDs. Finally, a third subset (not shown in Figure 4) would include CBDs of cellulases CelI and CelF from *C.thermocellum*, CenB from *C.fimi*, and CelCCG from *C.cellulolyticum*, which do not exhibit all of the conserved residues that characterize the planar strip in the other two subgroups. The residues of these CBDs which interact with cellulose are as yet unknown.

Interestingly, the residues that form the second conserved site, located in a groove on the other side of the molecule, seem to be strictly conserved among all the members of the family-III CBDs, thereby unifying the group. This surface constitutes a novel structural feature which has not yet been observed in other CBDs. The residues forming this shallow groove, mostly tyrosines and polar side chains, are characteristic of both the sugar-binding sites of lectins and glycosyl hydrolases, as well as the antigen-binding sites of antibodies and MHC molecules. Nothing is known about the actual function of this site, but several possibilities may be proposed. For example, the shallow groove could be a secondary binding site for the substrate, but the width of the cleft would presumably provide room for only a single cellulose chain. In this regard, it may indeed be a binding site for a single chain displaced from the crystalline cellulose and, as such, may serve to disrupt the crystalline arrangement of the substrate. However, such a catalytic function has previously been suggested for family-II CBDs which lack this particular feature (Din *et al.*, 1994a). Moreover, fluorescence measurements and attempts at co-crystallization using a variety of cellodextrins (data not shown) have thus far failed to support binding of Cip-CBD to short cellooligomers. Alternatively, the shallow groove may be involved in protein–protein or protein–carbohydrate interactions by binding either to other domains or structures on the same polypeptide chain, such as the glycosylated linkers that connect the different scaffoldin domains (Gerwig *et al.*, 1993), or to those of other (e.g. enzymatic) subunits of the cellulolytic system. It is worth noting that the location of this site is analogous to the carbohydrate binding site of legume lectins where the carbohydrate molecules bind to the concave face of the β sandwich with loops surrounding the ligand (Rini, 1995). In any event, our future efforts will be directed toward identifying the target structures of this unique groove, which may shed light on the specialized multifunctional role of the family-III CBDs.

The comparison of the Cip-CBD structure with members of family I and II has allowed us to identify the types of amino acid residues which may play a role in the putative cellulose-binding site. The analysis of the cellulose-binding structures from these different families of bacterial and fungal CBDs has revealed unexpected similarities which may reflect a common mechanism of substrate binding. Moreover, this mechanism may also be shared by other types of domain, which are involved in binding to other insoluble carbohydrates, like chitin. The variations in specificities, affinities and desorption conditions observed for different CBDs, even within representatives of a given family, could reflect, however, inherent differences in the detailed binding interactions. The three CBD structures currently available show a planar cellulose-binding surface composed of conserved residues which can be grouped in the following categories: (i) A planar linear strip, which includes a set of aromatic rings with one of their faces fully exposed to the solvent. In Cip-CBD and other family-III CBDs, an additional hydrophobic surface is provided by salt-bridged aspartate and arginine residues. (ii) A group of polar residues, interspersed among the aromatic rings in the planar strip, which appear to be responsible for stabilizing the appropriate orientation of the cellulose-binding residues, mainly through the formation of hydrogen bonds. In some cases, this architectural role appears to be contributed by the same side chains which form the planar strip, as in the hydrogen bonding interactions between His57 and Tyr67 residues in Cip-CBD. (iii) Finally, a group of polar anchoring residues, located on one or both sides of the aromatic planar strip, which appear to establish hydrogen bonding interactions with oxygen atoms and hydroxyl groups of glucose moieties on adjacent chains of the cellulose microcrystal. These residues not only contribute to the overall affinity of the CBDs for the substrate, but may also account for the cellulose-disrupting properties, which have been proposed for CBDs, by destabilizing the hydrogen-bonded structure of cellulose. Recent studies on the *T.reesei* CBH-I CBD, involving site-directed mutagenesis and molecular dynamics simulation, support the involvement of invariant polar residues (Gln7, Asn29 and Gln34) on the binding of this family-I CBD onto cellulose (Hoffrén *et al.*, 1995; Linder *et al.*, 1995).

The suggested mechanism of binding to cellulose, common to fungal and bacterial CBDs, proposes the attachment of these proteins to the (020) crystal faces which, probably due to their hydrophobic character, would only be exposed at the edges of perfect cellulose microcrystals. There is indeed experimental evidence, obtained by electron microscopy, for members of family-I (Chanzy *et al.*, 1984) and family-II (Gilkes *et al.*, 1993) CBDs, that cellulases or their CBDs bind preferentially to the edges of the cellulose microfibrils. Other observations consistent with the hypothesis that CBDs bind to the hydrophobic (020) faces are the following. CBDs often show higher binding capacities with amorphous cellulose, where the crystalline order of the cellulose chains have been mechanically or chemically disrupted, than with microcrystalline cellulose (Lamed *et al.*, 1985; Morag *et al.*, 1995). It may well be that in amorphous cellulose the (020) surfaces have become more accessible due to the disruption of the crystal packing, thereby increasing the number of binding sites. In addition, cellulases or their CBDs, but not their isolated catalytic domains, have been shown to prevent the flocculation of microcrystalline cellulose suspensions (Klyosov, 1990; Gilkes *et al.*, 1993) which appears to occur through interparticle hydrophobic interactions.

Finally, it is interesting to note that the distribution of the above-described set of amino acid residues on the putative cellulose-binding face of a given CBD appears to be contributed by different structural elements, depending on the family to which it belongs. This is especially significant regarding the two bacterial proteins, which exhibit a similar fold but show no sequence similarity. Thus, most of their amino acids that are purported to interact with cellulose as described above are contributed by different loops and β strands. Indeed, the character of the interaction with cellulose of the bacterial family-II CBD from *Cellulomonas fimi* more closely resembles that of the fungal family-I CBD. In both, the planar strip interacts with a central chain on the cellulose surface, and the presumed polar anchoring residues bind to glucose moieties on the two chains which occur on both sides of this central chain (Table III). This is in contrast to the interaction proposed for the family-III CBD from *C.thermocellum*, wherein the anchoring residues interact with two vicinal glucose chains which occur on one side of the chain which purportedly interacts with the planar strip. This would indicate that the two families of bacterial CBD did not evolve from a common molecular prototype; rather, their common fold and selection of functional elements appear to have been dictated by evolutionary convergence.

**Table IV.** Statistics for X-ray data collection and phase determination

| | Native | EMP[a] | GdCl$_3$ |
|---|---|---|---|
| Crystallographic data | | | |
| Resolution (Å) | 1.75 | 2.5 | 2.3 |
| Observed reflections | 113 902 | 65 749 | 75 585 |
| Unique reflections | 26 267 | 9849 | 11 094 |
| $R_{\mathrm{merge}}$ (%)[b] | 8.2 | 7.3 | 7.2 |
| Data coverage (%) | 87.4 | 93.6 | 82.6 |
| Outer shell (1.83–1.75 Å) (%) | 51.0 | | |
| MIR analysis | | | |
| Resolution (Å) | | 3.0 | 2.5 |
| Mean isomorphous difference[c] | | 15.8 | 17.7 |
| Phasing power[d] | | 1.43 | 1.98 |
| Mean figure of merit[e] | 0.53 to 2.5 Å | | |

[a] Ethyl mercuric phosphate.

[b] $R_{\mathrm{merge}}$, $\Sigma_h \Sigma_i |I_{ih} - <I_h>| / \Sigma_h \Sigma_i <I_h>$, where $<I_h>$ is the mean intensity of the i observations of reflection $h$.

[c] Mean isomorphous difference, $\Sigma ||F_{ph}| - |F_p|| / |F_p|$, where $|F_{ph}|$ and $|F_p|$ are the observed derivative and native structure factor amplitudes, respectively.

[d] Phasing power, r.m.s. ($|F_h|/E$), where $|F_h|$ is the calculated heavy atom structure factor amplitude, and $E$ is the residual lack of closure.

[e] Mean figure of merit before phase improvement with DM.

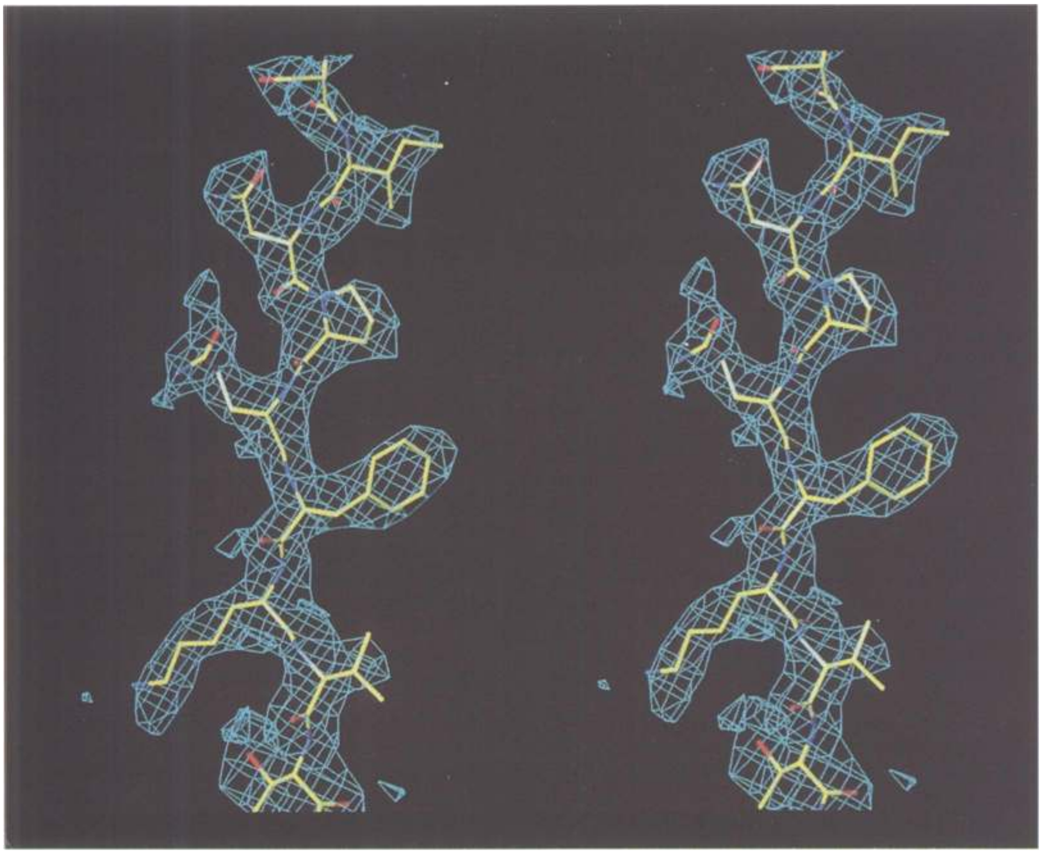


**Fig. 8.** Stereoview of the experimental electron density map. The final refined model and the electron density from the DM-modified MIRAS map at 2.5 Å resolution are shown for residues Ser17 to Thr25 in β strand 2. The electron density has been contoured at the r.m.s. value of the density for the unit cell.

## Materials and methods

### *Crystallization and preparation of heavy atom derivatives*

Cip-CBD was overexpressed, purified and crystallized as described (Lamed *et al.*, 1994; Morag *et al.*, 1995). The CBD portion of the CipB gene product (Poole *et al.*, 1992), derived from *C.thermocellum* strain YS, is identical to the sequence of the CipA-CBD from *C.thermocellum* strain ATCC 27405 (Gerngross *et al.*, 1993). Crystals of Cip-CBD were routinely grown by mixing the protein (8 mg/ml in water) with the same volume of mother liquor [50 mM 2-(morpholino)ethanesulfonic acid buffer, pH 6.5, 100 mM sodium acetate, 25% PEG 3350] and equilibrating the mixture against 1 ml of the latter solution at 12°C. Crystals grew as prisms which reached sizes of 0.4×0.4×0.05 mm$^3$ in 1–2 months. The crystals were harvested in solutions containing 35% PEG 3350 and 50 mM Tris at a variety of pHs, prior to data collection or heavy atom screening.

Two heavy atom derivatives were used for phase determination. The mercurial derivative was prepared by soaking in a solution containing 35% PEG 3350, 50 mM Tris–HCl, pH 8.0, 1 mM ethyl mercuric phosphate for 6 h at room temperature, whereas the lanthanide derivative was obtained by soaking the crystals in 35% PEG 3350, 50 mM Tris–HCl, pH 8.0, 40 mM $GdCl_3$ for 4 days.

### *Data collection and processing*

Cip-CBD crystals belong to the monoclinic space group C2 with unit cell dimensions $a$ = 63.57 Å, $b$ = 50.32 Å, $c$ = 95.96 Å, and β = 99.47°. There are two molecules in the asymmetric unit giving a $V_m$ value equal to 2.1 Å$^3$/Da, and an estimated solvent content of 42%.

X-ray intensity data were collected at room temperature with an $R_{axis}$-II image plate detector using Cu$K_\alpha$ radiation produced by a Rigaku RU200 rotating anode generator, equipped with mirrors. Data were collected with individual oscillations of 1.5° for the heavy atom derivatives, and 0.75° for the native data set. No special procedures were used to enhance the collection of Bijvoet pairs. Data collection was very often hindered by crystal multiplicity and twinning.

The images were processed and integrated with the DENZO package (Otwinowski, 1993), and reflections were subsequently scaled and merged using SCALEPACK (Otwinowski, 1993). Structure factors were derived from intensities by the program TRUNCATE from the CCP4 package (Collaborative Computing Project, Number 4, 1994). Statistics of the data sets used in the structure determination are given in Table IV.

### *Structure determination*

The derivative data sets were scaled to the native data with anisotropic temperature factors. Heavy atom positions were initially identified from difference and anomalous scattering Patterson maps and refined with the program MLPHARE (Otwinowski, 1991). Cross-phase difference Fourier maps computed using single isomorphous replacement with anomalous scattering (SIRAS) phases were used to confirm the heavy atom positions. Mercurial compounds bound to the sulfhydryl group of Cys55, whereas lanthanides substituted for the coordinated calcium ion. The initial 2.5 Å multiple isomorphous replacement with anomalous scattering (MIRAS) phases were improved by solvent leveling (Wang, 1985), and histogram matching using the program DM (Cowtan, 1994) in the CCP4 package. At this stage, a conservative solvent content of 30% was used for the automatic envelope determination. β strands could be clearly identified in the resulting electron density which, together with the location of the heavy atom sites, allowed an unambiguous determination of the position of the local 2-fold axis relating the two molecules in the asymmetric unit, and their solvent boundaries. Maps were computed for both heavy atom enantiomorphs and the twist of the β sheets was used to choose the correct set. The independent molecules are related by an ~2-fold rotation and translated 3.2 Å along an axis parallel to $y$ and located at the fractional coordinates ($x$ = 0.25, $z$ = 0.25). Masks covering the two monomers were created from skeletonized electron density using the program MAMA (Kleywegt and Jones, 1994), and were used for local averaging and solvent flattening, together with density histogram matching, in further cycles of density modification. The improved phases were used to further refine the heavy atom parameters (Rould *et al.*,

1992) which generated higher-quality electron density maps for a second density modification cycle. After this cycle, an electron density map at 2.5 Å was calculated for model building (Figure 8).

### Model building and refinement

All model building was performed on a Silicon Graphics workstation using the program O (Jones *et al.*, 1991). An α-carbon trace was built into the electron density using skeletonized maps obtained with the BONES option in MAPMAN program. This trace was used to produce an initial model from a database of refined structures (Jones *et al.*, 1991). The initial model comprised 155 amino acid residues in both independent monomers and had a crystallographic R factor of 0.41 for reflections in the resolution range 10.0–2.5 Å.

The atomic model was refined against native data by simulated annealing and least-squares minimization using the program X-PLOR (Brünger, 1992a). To monitor progress of the refinement and to guard against overfitting of the model, ~10% of the data were removed from refinement and used to calculate an $R_{\text{free}}$ value (Brünger, 1992b). Each monomer was treated independently during the refinement process. This was initiated with a round of least-squares minimization, followed by simulated annealing, using data in the range 10.0–2.5 Å which reduced the $R$ factor to 0.240 and the $R_{\text{free}}$ value to 0.329. Further cycles of standard least-squares refinement interspersed with manual model building resulted in an improved model which had $R$ and $R_{\text{free}}$ factors of 0.253 and 0.318, respectively, for reflections in the resolution range 10.0–1.75 Å. Ordered water molecules were then added to the model by examination of difference Fourier maps, using steric and hydrogen bonding criteria. The two highest peaks in the difference maps were modeled as calcium ions based in their coordination chemistry; their nature was later confirmed by atomic absorption (data not shown). When the crystallographic $R$ and $R_{\text{free}}$ factors fell below 0.209 and 0.250, respectively, the use of the last factor was discontinued and all data were incorporated into the working set. An overall anisotropic temperature factor was applied during the last refinement cycles. The refinement parameters for the current model are summarized in Table I.

Structural comparisons between proteins were made with the program SHP (Stuart *et al.*, 1979) and DALI (Holm and Sander, 1993).

The coordinates of the refined model are being deposited in the Brookhaven Protein Data Bank (code 1nbc) (Bernstein *et al.*, 1977).

## Acknowledgements

We thank S.Lee for his help with the preparation of figures. This work was supported in part by a grant from the Israel Science Foundation, administered by the Israel Academy of Sciences and Humanities, and by National Institutes of Health grant GM-22778. J.T. was the recipient of a Fulbright-Ministerio de Educación y Ciencia (Spain) fellowship.

## References

Bayer,E.A., Morag,E. and Lamed,R. (1994) The cellulosome – a treasure-trove for biotechnology. *Trends Biotechnol.*, **12**, 379–386.

Béguin,P. and Aubert,J.P. (1994) The biological degradation of cellulose. *FEMS Microbiol. Lett.*, **13**, 25–58.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.G.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Brünger,A.T. (1992a) *X-PLOR Version 3.1: A System for X-ray Crystallography and NMR.* Yale University Press, New Haven, CT.

Brünger,T.A. (1992b) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, **355**, 472–475.

Chanzy,H., Henrissat,B. and Vuong,R. (1984) Colloidal gold labelling of 1,4-β-D-glucan cellobiohydrolase adsorbed on cellulose substrates. *FEBS Lett.*, **172**, 193–197.

Choe,S., Bennett,M.J., Fujii,G., Curmi,P.M.G., Kantardjieff,K.A., Collier,R.J. and Eisenberg,D. (1992) The crystal structure of diphtheria toxin. *Nature*, **357**, 216–222.

Collaborative Computing Project, Number 4 (1994) The CCP4 Suite: programs for protein crystallography. *Acta Crystallogr.*, **D50**, 760–763.

Cowtan,K. (1994) DM, an automated procedure for phase improvement by density modification. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography*, **31**, 24–28.

Din,N. Gilkes,N.R., Tekant,B., Miller,R.C.,Jr, Warren,R.A.J. and Kilburn,D.G. (1991) Non-hydrolytic disruption of cellulose fibres by the binding domain of a bacterial cellulase. *BioTechnology*, **9**, 1096–1099.

Din,N., Damude,H.G., Gilkes,N.R., Miller,R.C.,Jr, Warren,R.A.J. and Kilburn,D.G. (1994a) $C_1$–$C_x$ revisited: intramolecular synergism in a cellulase. *Proc. Natl Acad. Sci. USA*, **91**, 11383–11387.

Din,N., Forsythe,I.J., Burtnick,L.D., Gilkes,N.R., Miller,R.C.,Jr, Warren,R.A.J. and Kilburn,D.G. (1994b) The cellulose-binding domain of endoglucanase A (CenA) from *Cellulomonas fimi*: evidence for the involvement of tryptophan residues in binding. *Mol. Microbiol.*, **11**, 747–755.

Gardner,K.H. and Blackwell,J. (1974) The structure of native cellulose. *Biopolymers*, **13**, 1975–2001.

Gerngross,U.T., Romaniec,M.P.M., Kobayashi,N.S. and Demain,A.L. (1993) Sequencing of a *Clostridium thermocellum* gene (cipA) encoding the cellulosomal $S_L$-protein reveals an unusual degree of internal homology. *Mol. Microbiol.*, **8**, 325–334.

Gerwig,G., Kamerling,J.P., Vliegenthart,J.F.G., Morag,E., Lamed,R. and 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.

Gilkes,N.R., Warren,R.A.J., Miller,R.C.,Jr and Kilburn,D.G. (1988) Precise excision of the cellulose binding domains from two *Cellulomonas fimi* cellulases by an homologous protease and the effect on catalysis. *J. Biol. Chem.*, **263**, 10401–10407.

Gilkes,N.R., Henrissat,B., Kilburn,D.G., Miller,R.C.,Jr and Warren, R.A.J. (1991) Domains in microbial β-1,4-glycanases: sequence conservation, function and enzyme families. *Microbiol. Rev.*, **55**, 305–315.

Gilkes,N.R., Kilburn,D.G., Miller,R.C.,Jr and Warren,R.A.J. (1993) Visualization of the adsorption of a bacterial endo-β-1,4-glucanase and its isolated cellulose-binding domain to crystalline cellulose. *Int. J. Biol. Macromol.*, **15**, 347–351.

Goldstein,M.A. and Doi,R.H. (1994) Mutation analysis of the cellulose-binding domain of the *Clostridium cellulovorans* cellulose-binding protein A. *J. Bacteriol.*, **176**, 7328–7334.

Harrison,S.C., Olson,A.J., Schutt,C.E., Winkler,F.K. and Bricogne,G. (1978) Tomato bushy stunt virus at 2.9 Å resolution. *Nature*, **276**, 368–373.

Hoffrén,A.-M., Teeri,T.T. and Teleman,O. (1995) Molecular dynamics simulation of fungal cellulose-binding domains: differences in molecular rigidity but a preserved cellulose binding surface. *Protein Engng*, **8**, 443–450.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Jones,T.A., Zou,J.-Y., Cowan,S.W. and Kjeldgaard,M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr.*, **A47**, 110–119.

Kleywegt,G.J. and Jones,T.A. (1994) Halloween... masks and bones. In Bailey,S., Hubbard,R. and Waller,D. (eds), *From First Map to Final Model*. SERC Daresbury Laboratory, UK, pp. 59–66.

Klyosov,A.A. (1990) Trends in biochemistry and enzymology of cellulose degradation. *Biochemistry*, **29**, 10577–10585.

Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.

Kraulis,P.J., Clore,G.M., Nilges,M., Jones,T.A., Pettersson,G., Knowles,J. and Gronenborn,A.M. (1989) Determination of the three-dimensional solution structure of the C-terminal domain of cellobiohydrolase I from *Trichoderma reesei*. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. *Biochemistry*, **28**, 7241–7257.

Lamed,R. and Bayer,E.A. (1988) The cellulosome of *Clostridium thermocellum*. *Adv. Appl. Microbiol.*, **33**, 1–46.

Lamed,R., Setter,E., Kenig,R. and Bayer,E.A. (1983) The cellulosome – a discrete cell surface organelle of *Clostridium thermocellum* which exhibits separate antigenic, cellulose-binding and various cellulolytic activities. *Biotech. Bioengng Symp.*, **13**, 163–181.

Lamed,R., Kenig,R., Setter,E. and Bayer,E.A. (1985) The major characteristics of the cellulolytic system of *Clostridium thermocellum* coincide with those of the purified cellulosome. *Enzyme Microb. Technol.*, **7**, 37–41.

Lamed,R., Tormo,J., Chirino,A.J., Morag,E. and Bayer,E.A. (1994) Crystallization and preliminary X-ray analysis of the major cellulose-binding domain of the cellulosome from *Clostridium thermocellum*. *J. Mol. Biol.*, **244**, 236–237.

Laskowski,R.A., McArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26**, 283–291.

Linder,M., Mattinen,M.L., Kontteli,M., Lindeberg,G., Ståhlberg,J., Drakenberg,T., Reinikainen,T., Pettersson,G. and Annila,A. (1995) Identification of functionally important amino acids in the cellulose-binding domain of *Trichoderma reesei* cellobiohydrolase I. *Protein Sci.*, **4**, 1056–1064.

McPhalen,C.A., Strynadka,N.C.J. and James,M.N.G. (1991) Calcium-binding sites in proteins: a structural perspective. In Anfinsen,C.B., Edsall,J.T., Richards,F.M. and Eisenberg,D.S. (eds), *Advances in Protein Chemistry*. Academic Press, New York, pp. 77–144.

Morag,E., Lapidot,A., Govorko,D., Lamed,R., Wilchek,M., Bayer,E.A. and Shoham,Y. (1995) Expression, purification and characterization of the cellulose-binding domain of the scaffoldin subunit from the cellulosome of *Clostridium thermocellum*. *Appl. Environ. Microbiol.*, **61**, 1980–1986.

Nicholls,A. and Honig,B. (1993) *GRASP*. Columbia University, New York, NY.

Otwinowski,Z. (1991) Maximum likelihood refinement of heavy atom parameters. In Wolf,W., Evans,P.R. and Leslie,A.G.W. (eds), *Isomorphous Replacement and Anomalous Scattering*. SERC Daresbury Laboratory, UK, pp. 80–86.

Otwinowski,Z. (1993) Oscillation data reduction programs. In Sawyer,L., Isaacs,N. and Bailey,S.W. (eds), *Data Collection and Processing*. SERC Daresbury Laboratory, UK, pp. 56–62.

Pagès,S., Belaich,A., Tardif,C., Reverbel-Leroy,C., Gaudin,C. and Belaich,J.P. (1996) Interaction between the endoglucanase CelA and the scaffolding protein CipC of the *Clostridium cellulolyticum* cellulosome. *J. Bacteriol.*, **178**, 2279–2286.

Poole,D.M., Morag,E., Lamed,R., Bayer,E.A., Hazlewood,G.P. and Gilbert,H.J. (1992) Identification of the cellulose binding domain of the cellulosome subunit S1 from *Clostridium thermocellum*. *FEMS Microbiol. Lett.*, **99**, 181–186.

Poole,D.M., Hazlewood,G.P., Huskisson,N.S., Virden,R. and Gilbert,H.J. (1993) The role of conserved tryptophan residues in the interaction of a bacterial cellulose binding domain with its ligand. *FEMS Microbiol. Lett.*, **106**, 77–84.

Quiocho,F.A. (1993) Probing the atomic interactions between proteins and carbohydrates. *Biochem. Soc. Trans.*, **21**, 442–448.

Reinikainen,T., Teleman,O. and Teeri,T.T. (1995) Effects of pH and high ionic strength on the adsorption and activity of native and mutated cellobiohydrolase I from *Trichoderma reesei*. *Proteins*, **22**, 392–403.

Rini,J.M. (1995) Lectin structure. *Annu. Rev. Biophys. Biomol. Struct.*, **24**, 551–577.

Rould,M.A., Perona,J.J. and Steitz,T.A. (1992) Improving multiple isomorphous replacement phasing by heavy-atom refinement using solvent-flattened phases. *Acta Crystallogr.*, **A48**, 751–756.

Shoseyov,O., Takagi,M., Goldstein,M.A. and Doi,R.H. (1992) Primary sequence analysis of *Clostridium cellulovorans* cellulose binding protein A. *Proc. Natl Acad. Sci. USA*, **89**, 3483–3487.

Stuart,D., Levine,M., Muirhead,M. and Stammers,D. (1979) The crystal structure of cat pyruvate kinase at a resolution of 2.6 Å. *J. Mol. Biol.*, **134**, 109–142.

Tomme,P., Van Tilbeurgh,H., Pettersson,G., Van Damme,J., Vandekerckhove,J., Knowles,J., Teeri,T.T. and Claeyssens,M. (1988) Studies of the cellulolytic system of *Trichoderma reesei* QM 9414: analysis of domain function in two cellobiohydrolases by limited proteolysis. *Eur. J. Biochem.*, **170**, 575–581.

Tomme,P., Warren,R.A.J. and Gilkes,N.R. (1995) Cellulose hydrolysis by bacteria and fungi. *Adv. Microb. Physiol.*, **37**, 1–81.

Toone,E.J. (1994) Structure and energetics of protein-carbohydrate complexes. *Curr. Opin. Struct. Biol.*, **4**, 719–728.

Wang,B.C. (1985) Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol.*, **115**, 90–112.

Xu,G.Y., Ong, E., Gilkes,N.R., Kilburn,D.G., Muhandiram,D.R., Harris-Brandts,M., Carver,J.P., Kay,L.E. and Harvey,T.S. (1995) Solution structure of a cellulose-binding domain from *Cellulomonas fimi* by nuclear magnetic resonance spectroscopy. *Biochemistry*, **34**, 6993–7009.

*Received on March 6, 1996; revised on July 17, 1996*