Solution Structure of the Cellulose-Binding Domain of the Endoglucanase Z Secreted by *Erwinia chrysanthemi*[†,‡]

Emmanuel Brun,[§,||,⊥] Fabrice Moriaud,[§] Pierre Gans,*[,§] Martin J. Blackledge,[§] Frederic Barras,[||] and Dominique Marion[§]

Institut de Biologie Structurale "Jean-Pierre Ebel" (CEA-CNRS), 41 avenue des Martyrs, 38027 Grenoble Cedex, France, and LCB-CNRS, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

Received July 29, 1997; Revised Manuscript Received October 2, 1997[⊗]

ABSTRACT: Two-dimensional proton nuclear magnetic resonance spectroscopy has been used to determine the three-dimensional structure of the 62 amino acid C-terminal cellulose-binding domain (CBD) of the endoglucanase Z (CBD_{EGZ}), secreted by *Erwinia chrysanthemi*. An experimental data set comprising 958 interproton nOe-derived restraints was used to calculate 23 structures. The calculated structures have an average root-mean-square deviation between Cys4 and Cys61 of 0.91 ± 0.11 Å for backbone atoms and 1.18 ± 0.12 Å for the heavy atoms. The CBD_{EGZ} exhibits a skiboot shape based mainly on a triple antiparallel β-sheet perpendicular to a less-ordered summital loop. Three aromatic rings (Trp18, Trp43, and Tyr44) are localized on one face of the protein and are exposed to the solvent in a conformation compatible with a cellulose-binding site. Based on its original folding, we have been able to relate the CBD sequence to those of several domains of unknown function occurring in several bacterial chitinases as well as other proteins. This study also provides a structural basis for analyzing the secretion-related information specific to the CBD_{EGZ}.

Plant cell wall is mainly composed of pectin and cellulosic substrates. Hence, plant pathogens make use of a battery of depolymerizing enzymes, including pectinases and cellulases. *Erwinia chrysanthemi* is a Gram negative bacterium plant pathogen that secretes a wide array of such depolymerizing enzymes, the action of which eventually lead to the so-called "soft-rot" disease *(1)*. Remarkably, this set of enzymes makes use of a single secretion machinery, hereafter referred to as Out, to cross the bacterial cell envelope and reach the plant cell wall polysaccharides. Hence, deciphering the structural determinants of both their enzymatic activity and their ability to use the Out machinery is a prerequisite to controlling *Erwinia chrysanthemi* mediated virulence.

In our studies, we have used the cellulase EGZ as a model to elucidate the molecular basis of the secretion and to identify the structural information, allowing the Out secretion machinery to recognize and specifically target proteins to the cell exterior *(2)*. EGZ has a modular architecture consisting of an N-proximal catalytic domain, which permits the hydrolysis of the β-1,4-glycosidic bond present either in small soluble substrates or in substituted cellulose, and is linked, via a typical Ser/Thr-rich linker region, to a C-proximal cellulose-binding domain (CBD),[1] which allows the binding to microcrystalline cellulose *(3)*. Studies of modified EGZ derivatives suggested that part of the secretion-related information is contained in the CBD *(2)*. Moreover, a series of mutagenesis studies established that EGZ adopts high-ordered structure prior to its secretion *(4)*. In particular, disulfide bond formation in the CBD was found to be a step required for EGZ secretion *(4)*. These observations prompted us to initiate a detailed study of the CBD_{EGZ} structure, in order to elucidate the secretion mechanism *(5)*.

Over 100 different CBD sequences have already been identified, ranging in size from only 33 to over 170 amino acid residues. Their role with respect to cellulase activity is not well understood, and at least two hypotheses have been put forward: (i) they may enhance enzyme activity simply by concentrating the enzyme on the substrate surface; (ii) they could be involved in the disruption of noncovalent interaction between adjacent substrate molecules *(6, 7)*.

CBDs can be grouped into distinctive families on the basis of amino acid sequence similarities *(8, 9)*. Family I contains 33−36 residue CBDs found only in fungal cellulases. The three-dimensional structure of one member of this family, CBD_{CBHI}, derived from the cellobiohydrolase I of *Trichoderma reesei*, has been determined by NMR spectroscopy *(10)*. The secondary structure of CBD_{CBHI} is organized into a wedge-shaped irregular β-sheet. One face of the molecule,

[†] This work was supported by grants from the Centre National de la Recherche scientifique (CNRS), the Commissariat à l'Energie Atomique (CEA), the University Aix-Marseille II, and the Bruker Co. (Wissembourg, France). E.B. was a recipient of a IMABIO-MRT fellowship. This is Publication No. 447 of the Institut de Biologie Structurale "Jean Pierre Ebel".

[‡] The coordinates have been deposited in the Brookhaven Protein Data Bank under Accession Number 1aiw.

[§] Institut de Biologie Structurale "Jean-Pierre Ebel".

[||] LCB-CNRS.

[⊥] Present address: Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, BC, Canada V6T 1Z3.

[⊗] Abstract published in *Advance ACS Abstracts*, December 1, 1997.

[1] Abbreviations: CBD, cellulose-binding domain; CBD_{EGZ}, cellulose-binding domain from *Erwinia chrysanthemi* endoglucanase Z; GSP, general secretory pathway; CBD_{BHI}, cellulose-binding domain from *Trichoderma reesei* cellobiohydrolase 1; CBD_{CEX}, cellulose-binding domain from *Cellulomonas fimi* xylanase−glucanase Cex; CBD_{Cip}, cellulose-binding domain from the cellulosomal scaffolding subunit of *Clostridium thermocellum*; CBD_{N1}, cellulose-binding domain from *Cellulomonas fimi* β-1,4-glucanase CenC; COSY-DQF, double quantum filtered J correlated spectroscopy; TOCSY, total correlated spectroscopy; NOESY, nuclear Overhauser spectroscopy; TPPI, time-proportional phase incrementation; SA, simulated annealing; rMD, restrained molecular dynamics; rmsd, root-mean-square deviation.

S0006-2960(97)01849-7 CCC: $14.00

dominated by three conserved tyrosine side chains, forms a hydrophobic and planar surface that has been shown to be involved in cellulose binding *(11, 12)*. CBDs of bacterial enzymes are substantially larger. The structures of three of these have been reported so far. CBD_{Cex}, a 110-residue member of family II CBDs derived from the β-1,4-glycanase Cex from *Cellulomonas fimi*, forms an elongated, 9-stranded β-barrel *(13)*. The substrate-binding site includes three solvent-exposed tryptophans, together with other hydrophilic residues, located on one edge of the barrel. CBD_{Cip}, a 155-residue member of family III CBDS derived from the cellulosomal scaffolding subunit of *Clostridium thermocellum*, forms a 9-stranded β-sandwich with a jelly roll topology *(14)*. One of the faces of the β-sandwich is dominated by a planar linear strip of conserved and surface-exposed aromatic and polar side chain residues which are proposed to interact with crystalline cellulose. CBD_{N1}, a 152-residue member of family IV CBDs derived from the CenC cellulase of *Cellulomonas fimi*, is made of 10 β-strands folded, like the CBD_{Cip}, into 2 antiparallel β-sheets with the topology of a jelly-roll β-sandwich *(15)*. In contrast to other CBD structures, instead of having a "flat" binding surface, the binding site of CBD_{N1} is a cleft containing a central strip of hydrophobic residues that is flanked on both sides by polar hydrogen-bonding groups. The presence of this cleft provides a structural explanation for the unique selectivity of CBD_{N1} for amorphous cellulose and other soluble oligosaccharides and the lack of binding to crystalline cellulose.

CBD_{EGZ} exhibits original features, and it was classified as an isolated and separated unit making family V on its own. Indeed, it is intermediate in size between the type I CBD_{CBHI} (33 amino acids) and the type II CBD_{Cex} (106 amino acids). Moreover, it shows no sequence similarity with known CBDs. This study provides an NMR-derived structure for CBD_{EGZ} which was found to adopt a different fold compared with other CBDs. Stuctural analysis allows us to point out residues potentially involved in cellulose binding as well as in protein secretion. Finally, unexpected relatedness with a series of modules thought to be chitin-binding domains was established.

MATERIALS AND METHODS

Bacterial Strains and Media. *E. coli* strains used in this study were TG1 [*supE hsd D5 thi* Δ*(lac-proAB)* Δ*(srl-recA)*-306::Tn*10*(Tet^r)] and BL21(DE3) [F-*ompT hsdSB* (rB^-mB^-) *dcm gal* (DE3)].

The culture medium used was Luria Bertani (LB) (which contains, per mL, 5 mg of bactotryptone/2.5 mg of yeast extract/2.5 mg of NaCl, adjusted pH 7.2). The culture medium used for the production of CBD was M9Cas (which contains 0.5% glycerol and, per mL, 6 mg of Na_2HPO_4/1.254 mg of KH_2PO_4/1 mg of NH_4Cl/0.5 mg of NaCl/0.24 mg of $MgSO_4$/0.01 mg of $CaCl_2$/0.001 mg of thiamine/2 mg of casamino acids). Ampicillin was added to 100 μg/mL.

Construction of the pMIA2 Plasmid. The CBD-encoding sequence was isolated as the 0.8 kb *Sma*I−*Sma*I fragment present in a plasmid previously used as an intermediate in the construction of plasmid pBSD8 *(3)*. This intermediate plasmid contains a *Sma*I restriction site, introduced by site-directed mutagenesis within the EGZ-coding gene *cel*Z, at the junction between the linker region and the CBD in the EGZ protein *(3)*. The other *Sma*I site lies within the untranslated region downstream of *cel*Z. The pET22b (Novagen), used as recipient, was linearized by *Nco*I restriction and made blunt-ended with T4 DNA polymerase (Amersham kit). The resulting pET22b vector was mixed with the 0.8 kb *Sma*I−*Sma*I restriction fragment (vector/insert ratio equal to 1/6 and a final DNA concentration of 10 μg/mL) in the presence of T4 DNA ligase for 16 h at 16 °C. The religation of the vector creates a new *Nsi*I restriction site while the ligation with the insert creates two new *Nco*I restriction sites. The DNA contained in the ligation mixture was first subjected to restriction by *Nsi*I. Plasmid DNA from 24 TG1 transformed clones was analyzed by *Nco*I restriction, permitting the identification of 8 clones, the plasmid of which contained the insert. Plasmid DNA of the eight clones was purified and used to transform competent *E. coli* BL21(DE3) cells. IPTG-induced cultures of 20 BL21(DE3) transformed clones were analyzed for their protein contents by Western blot using an anti-EGZ18 antibody *(3, 5)*. One clone, which expresses a protein that was recognized by anti-EGZ18 antibody and had a apparent molecular mass of about 6500 Da on SDS−PAGE, was selected and stored at −80 °C in 20% glycerol and referred to as BL21(DE3)/pMIA2.

Production and Purification of CBD_{EGZ}. A single colony of BL21(DE3)/pMIA2 was inoculated into M9Cas medium (20 mL), supplemented with ampicillin (100 μg/mL), and incubated overnight at 37 °C under agitation. A 20 mL sample was then used to inoculate 700 mL of the same medium in a 2 L toxin flask and incubated at 37 °C with agitation. Bacterial growth was monitored by recording the OD_{600}. At an OD_{600} of 2 (roughly 3 h after inoculation), IPTG was added to a final concentration of 0.2 mM. Cultures were grown until the end of exponential phase, which corresponds to an OD_{600} of 3−3.5 and 4−5 h of growth in the presence of IPTG. CBD_{EGZ} was purified from the culture supernatant in a three-step protocol as described *(5)*.

NMR Sample Preparation. Electrophoretically pure lyophilized sample of CBD_{EGZ} was dissolved in 0.05 M potassium phosphate buffer at pH 4.3 in 90% H_2O/10% ^2H_2O. The sample concentration was 0.8 mM, the maximum achieved due to the low solubility of CBD_{EGZ}. CBD concentration was determined by UV spectroscopy, measuring the Trp absorption at 285 nm.

NMR Measurements. NMR spectra were recorded at 27 and 37 °C on a Bruker AMX 600 spectrometer. Chemical shifts were referenced relative to the water resonance fixed at 4.75 and 4.65 ppm, for the two temperatures, respectively. ^1H 2D spectra, DQF-COSY *(16)*, TOCSY/HOHAHA *(17, 18)*, and NOESY *(19, 20)* Spectra were recorded in the States-TPPI mode *(21)*. The water resonance was attenuated by means of a coherent low-power ($\gamma B_2/2\pi = 50$ Hz) presaturation during the relaxation delay. For NOESY and TOCSY experiments, this presaturation was further combined with a "jump and return" read pulse *(22)*. The isotropic mixing time was set to 70 ms and the nOe mixing time to 60 and 150 ms. Two-dimensional spectra were collected as 512 (t_1) and 2048 (t_2) complex point time-domain matrices with spectral widths of 7201 Hz (t_1) and 8474 Hz (t_2). To obtain a sufficient signal/noise ratio, 96 scans were used per t_1 increment. They were transformed after zero-filling in the F_1 dimension, into 1024 and 1024 real points in the F_1 and F_2 dimension frequency-domain spectra using FELIX software version 2.3 (Biosym Technologies/Molecular Simulation Inc., San Diego, CA). Amide proton exchange rates in ^2H_2O were monitored over a half-day period by recording a series of one-dimensional spectra every 20 min at 37 °C immediately after dissolving CBD_{EGZ} in ^2H_2O/phosphate buffer.

Interatomic Distance Restraints. Interproton distance restraints were derived from NOESY experiments acquired at 37 °C with 60 and 150 ms mixing times, in 90% H_2O/10% ^2H_2O. A final set of 958 interproton distance restraints was used for the structure determination. This data set was comprised of 360 meaningful intraresidue, 247 sequential, 74 short-range ($1 < |i - j| \leq 4$), and 277 long-range ($|i - j| > 4$) restraints. The nOe-derived constraints were estimated as strong, medium, or weak (with corresponding upper limits of 2.8, 3.8, or 5.0 Å). Unresolved protons were replaced by pseudoatoms, and an appropriate correction was applied to the measured distance. Nonstereospecifically assigned but spectroscopically resolved protons were allowed to float between the two prochiral positions during the simulated-annealing protocols (see below).

Structure Calculations. Structure calculations were performed using DISCOVER (version 2.9.7, 1993) interfaced to the INSIGHT II program (version 2.3.0, 1993) for visualization and result analysis (Biosym Technologies Inc.). The AMBER4 force field *(23)* was used for all calculations except for the simulated annealing (SA) protocols, in which a simple quartic nonbond term was employed. All ω dihedrals were forced to *trans* except the one preceding Pro11, as there was strong evidence from the nOe data that this peptide bond was in the *cis* conformation (see results). The $S\gamma - S\gamma$ distance between Cys4 and Cys61 side chains was constrained in the structure calculation. A three-stage structure determination protocol was employed *(24)*.

(1) Determination of the Global Protein Fold. The global fold of the protein was determined using a 60 ps SA protocol starting from randomized initial coordinates containing 10 ps of slow cooling to 100 K. The exploratory period was calculated at a nominal temperature of 1000 K. Experimental data were scaled to a maximum of 50 kcal mol^{-1} Å^{-2} during the first 30 ps. The molecule was minimized in the complete AMBER force field. This protocol was employed twice, in the very beginning with an initial set of distance constraints, and once the final list of experimental constraints had been obtained after the iterative structure refinement.

(2) Structure Refinement. A two-stage SA/rMD (SI−SII) refinement protocol was used *(24)*. During the SA calculation (SI), the local conformational space available from the experimental data was investigated. This calculation still allowed a large sampling of the conformational space, but was less time-consuming than the global fold protocol as it started from already folded structures. In this step, the nominal temperature was held at 2000 K, and the nOe force constants were scaled up progressively to 50 kcal mol^{-1} Å^{-2}. Short-range distances which define the local conformation were introduced before long-range distances. After the scaling period, the molecule was allowed to sample the conformational space for 2 ps before slow-cooling to 100 K over a period of 11 ps. After 2 ps of dynamics, the molecule was subjected to energy minimization as described above. In the following restrained molecular dynamics (rMD) calculation (SII), the use of a more complete force field leads to a physically more viable structure. The effects of solvent were simulated implicitly by the use of a distance-dependent dielectric constant *(25, 26)* and reduced charges on the charged side chains *(27, 24)*. During this stage, the weight of the experimental distance restraints was reduced to 25 kcal mol^{-1} Å^{-2}. After equilibration, the temperature was regulated using weak coupling to a thermal bath *(28)*. The molecule was allowed to evolve at 750 K for 10 ps, and then slowly cooled to 300 K over a period of 5 ps, where dynamics were continued for a further 12 ps. Finally the structure was minimized using a conjugate gradient algorithm.

Structural Analysis. The function $F_{nOe} = k_{nOe}(r_{ij} - r_{ij}^{u})^2$ ($r_{ij} > r_{ij}^{u}$) has been used to create an experimental violation function (V_i) representing the sum of the energetic contributions from the violated experimental restraints. The root-mean-square deviation of the 23 best structures (with the lowest V_i values) was calculated over the specified atoms using the program INSIGHTII. After superposition of the backbone (N, C^α, C) atoms from residues 4 to 61 of each final model for a minimum rmsd to the structure with the lowest energy in the 23 final models, a geometric average was computed to describe a "mean-structure" for which no energy minimization was performed and which was used as a reference to describe the atomic rmsd. The solvent-accessible surfaces were calculated in the absence of protons using the united atom radii proposed by Conolly *(29)* and a water molecule radius of 1.4 Å.

Electrostatic Calculations. The program Delphi, as implemented in the Biosym package, was used to calculate the surface charge distribution of CBD_{EGZ}. The electrostatic potential is calculated by solving the Poisson−Boltzmann equations using a finite difference method *(30)*. Dielectric constants of 80 and 4 were used for the solvent and the protein interior, respectively, and an ionic strength of 0.145 $M{\cdot}L^{-1}$ was assumed for the solvent. A 2 Å Stern layer was used to define the exclusion region between the protein and the solvent. The full charge set was used as in the AMBER4 force field *(23)*. The protein was mapped onto a 65 × 65 × 65 point matrix, and each point was assigned a charge and dielectric value. These grids were used as bases for subsequent focusing calculations, giving a final resolution of approximately 0.6−0.8 Å. The charged exterior of the molecules was visualized by projecting the electrostatic values onto a Connolly solvent accessibility surface. In order to take into account the motional averaging of side chain orientations, the electrostatic surface has been calculated for a series of the 12 first NMR structures of CBD and the average grid value used to represent the effective encounter surface of the molecule.

RESULTS

Production, Purification, and Characterization of CBD_{EGZ}. Our previous studies made use of a vector, referred to as pMIA1, that exhibited several limitations among which were (i) poor yield of pure CBD (i.e., 0.25 mg/L culture) and (ii) aberrant location; i.e., a large proportion of the CBD remained cell-bound, presumably as a nonprocessed form *(5)*. Therefore, we decided to construct a new expression vector, referred to as pMIA2, as follows. First, the CBD_{EGZ}-encoding nucleotide sequence was inserted in the pET22b plasmid, downstream to that encoding the *Erwinia chrysanthemi* PelB leader sequence under the control of the T7 phage RNA polymerase. Second, an additional Met (Met1) was

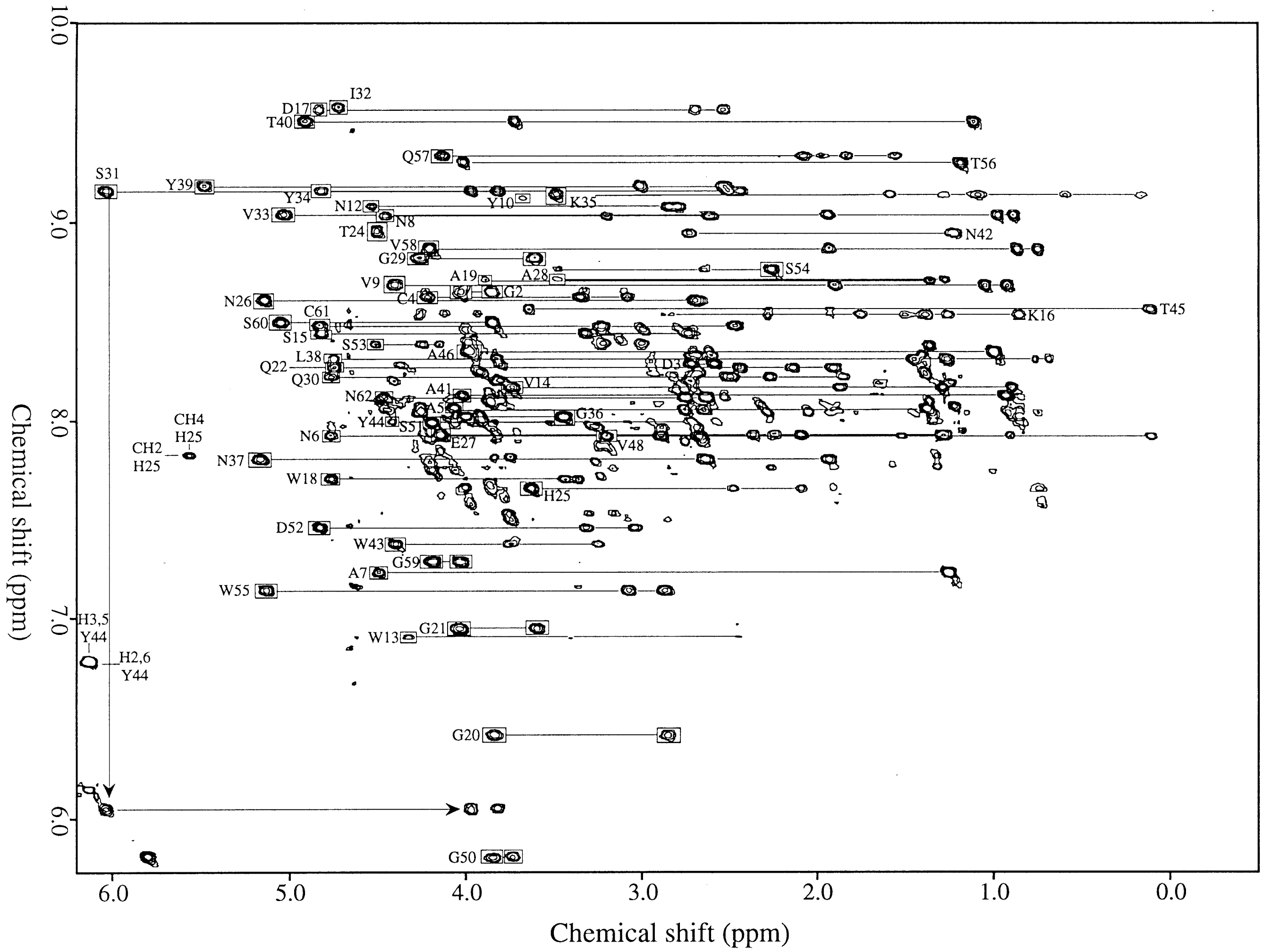

FIGURE 1: Amide/aliphatic region of a TOCSY spectrum of CBD_{EGZ} at 0.8 mM acquired with an isotropic mixing time of 70 ms, in H_2O at 37 °C and pH 4.8. The squared peaks designate the assigned $^1H^N-^1H^\alpha$ correlation peaks of each amino acid spin system also observed in the COSY spectra of CBD_{EGZ} at the same temperature. Assigned peaks that belong to the same spin system are related by a line. The $^1H^N-^1H^\alpha$ correlation peaks of Asp3, Lys16, Asp42, Thr45, and Thr56 are buried under the water resonance at this temperature; they have been detected in a COSY and TOCSY spectrum recorded at 27 °C. Unassigned peaks mostly observed in the central amide region are due to a contaminant molecule of unfolded CBD_{EGZ} (see Results).

added at the N-terminus; this was done to avoid the suspected steric hindrance between the enzyme involved in the disulfide bond formation between Cys4 and Cys61, namely DsbA and Lep, the signal sequence leader peptidase *(5)*. The pMIA2 expression vector allowed production of 10 mg of mature CBD_{EGZ} per liter of culture. Moreover, the CBD_{EGZ} produced was fully exported to the cell exterior, and no cell-bound CBD_{EGZ} precursor could be detected. By using the previously published purification protocol, we obtained 1 mg of pure CBD_{EGZ} per liter of culture. Albeit improved, this rather low yield could be attributed to the fact that only 10% of CBD_{EGZ} could be eluted under nondenaturant conditions from the cotton columns. Comparison of 1D NMR spectra of the CBD_{EGZ} obtained was very similar to that previously obtained (data not shown), indicating that the extra Met residue located at the N-terminus of CBD_{EGZ} did not significantly modify the structure of CBD_{EGZ}.

Proton Resonance Assignment. The assignment strategy is based on $^1H-^1H$ correlation experiments as proposed by Wüthrich *(31)*. In DQF-COSY and TOCSY (Figure 1) spectra at the two temperatures (27 and 37 °C), most of the spin systems expected were recognized with the exception of Ser47 which did not show any amide to aliphatic proton transfers. For Ser47, the sequential nOe transfers $^1H^\alpha46-^1H^N47$, $^1H^\alpha47-^1H^N48$, and $^1H^\beta47-^1H^N48$ observed on the NOESY spectrum were used to identify the amide frequency and the complete spin system. A few spin systems (such as Tyr10, Thr24, and Ile32) that deviate from standard patterns for lack of transfer from the amide proton to the side chain were identified in the aliphatic region of the TOCSY spectra, and their amide proton was assigned using nOe connectivity. Proton resonances of the three proline residues (Pro11, Pro23, and Pro49) were also easily assigned in the aliphatic region of the TOCSY spectra. The side-chain amide protons of Asn and Gln were stereospecifically assigned by using the relative intensity of their nOe correlation with either the H^β or the H^γ of the side chain. Only the side-chain amide protons of Gln22 were not found in the spectrum. None of the prochiral side-chain protons were stereospecifically assigned. Although the sample is made of electrophoretically pure CBD_{EGZ}, its NMR spectrum reveals the presence of contaminants (see Figure 1). This can be attributed to the presence of unfolded CBD_{EGZ}. Indeed, the contaminant peaks in the TOCSY spectra can be related to CBD_{EGZ} spin systems (Asp3, Cys4, Asn6, and Asn62 are the most obvious), their H^N chemical shifts correspond to random coil values (7.5−8.5 ppm), and they are not present in the NOESY experiments.

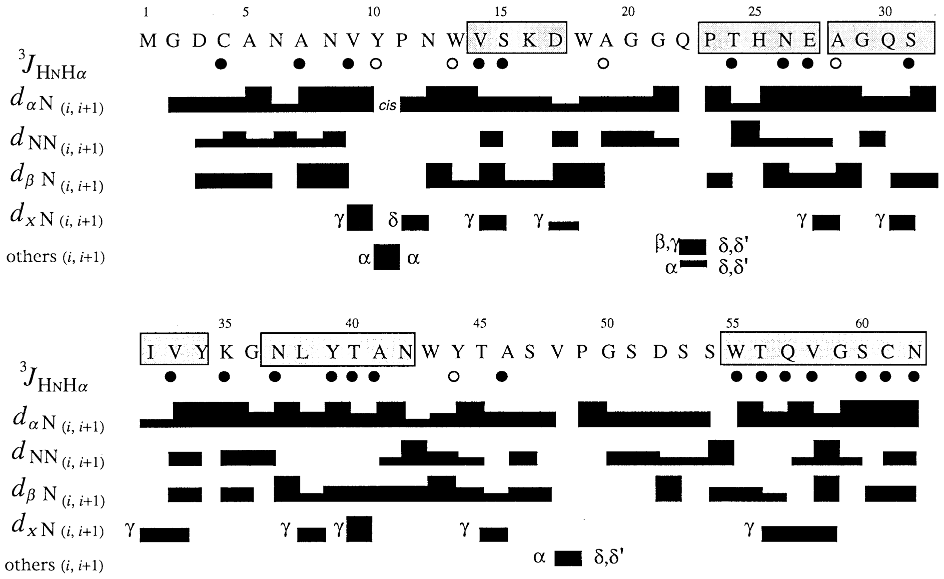

FIGURE 2: Secondary structure determination of the CBD$_{EGZ}$. Summary of the $^3J_{HN-H\alpha}$ values [estimated from the COSY: (○) J < 5 Hz; (●) J > 7 Hz] and of the sequential and short-range nOes (i to <i+4) observed in the $^1H-^1H$ NOESY spectra. The nOes have been categorized as strong, medium, and weak, which is reflected by bar height. Hatched bars indicate regions in extended conformations.

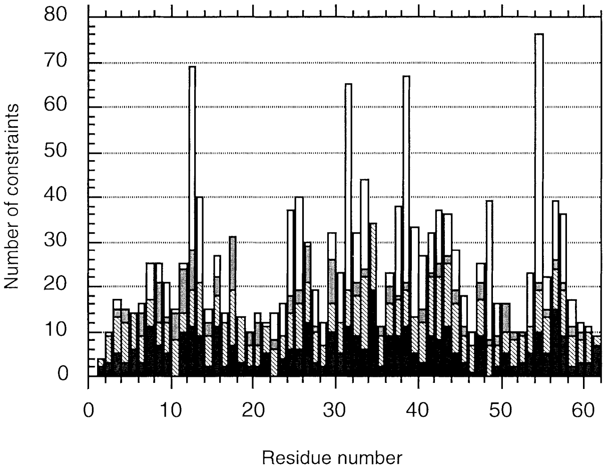

FIGURE 3: Sequential distribution and range of the nOe distance constraints used in the calculation of the CBD$_{EGZ}$. Black, intraresidue constraints; hatched, sequential nOe (i+1); light gray, medium-range nOe (i ≤ 4); white, long-range nOe.

Table 1: Statistical Analysis of the E_II Conformational Ensemble

(A) DISCOVER-AMBER Energetic Statistics	
F_{bond}	11.7 ± 0.95[a]
F_{angle}	85.7 ± 6.3[a]
$F_{torsion}$	83.8 ± 4.5[a]
$F_{out\ of\ plane}$	2.8 ± 0.5[a]
F_{H-bond}	−29.7 ± 2.0[a]
$F_{Lennard-Jones}$	−170.9 ± 7.0[a]
$F_{coulombic}$	−775.6 ± 18.6[a]
F_{total}	−789.9 ± 29.9[a]
(B) Experimental Statistics	
distance nOe violations	
>0.10 Å	9.00 ± 2.6
>0.20 Å	0.47 ± 0.8
>0.30 Å	0.1
>0.40 Å	0
maximal violation (Å)	0.325
violation energy[b]	
F_{nOe}	8.8 ± 1.5[a]
$F_{nOe(viol>0.10Å)}$	4.3 ± 1.7[a]
(C) Structural Statistics[c]	
atoms	
backbone[d] (1−62)	1.20 ± 0.16
all heavy (1−62)	1.45 ± 0.15
backbone[d] (4−61)	0.91 ± 0.11
all heavy (4−61)	1.18 ± 0.12
backbone β-sheet[e]	0.33 ± 0.10

[a] Values are in kcal·mol^{-1}. [b] F_{nOe} is calculated using a force constant of 25 kcal·mol^{-1}·Å^{-2}. [c] rmsd values are the average pairwise rmsd for the residues and atoms specified, relative to the geometric average of the corresponding ensemble; these values are given in Å. [d] N, C$^\alpha$, and C atoms were used for the superpositions. [e] β-Sheet is defined by residues 26−27, 31−34, 37−41, 44−45, and 55−60.

Sequential assignment was achieved using NOESY spectra at the two temperatures. Except for Ser54 and the three prolines, all H$^\alpha$ exhibit an nOe correlation with the next $^1H^N$ (Figure 2). Most of these assignments (as well as the link between Ser54 and Trp55) were confirmed by $^1H^N-^1H^N$ connectivities. Among the three prolines of the CBD$_{EGZ}$ sequence, only Pro11 exhibits a *cis* conformation as indicated by a strong $^1H^\alpha-^1H^\alpha$ nOe correlation between Tyr10 and Pro11 *(31)*.

Table S1 of the supporting information (see supporting information available) describes the overall assignment of CBD$_{EGZ}$ proton resonances. Some amide or H$^\alpha$ proton resonances were found outside their usual chemical shift range *(32, 33):* the amide protons of Gly20, Gly21, and Gly50 are below 7 ppm, whereas the $^1H^\alpha$ of Ser31 and Ser54 resonate respectively at 6.4 and 2.2 ppm.

Secondary Structure Pattern Recognition. Figure 2 summarizes the structural information contained in the sequential resonance assignment. A qualitative estimation of the $^3J_{HN-H\alpha}$ coupling constants was obtained from considering the $^1H^N-^1H^\alpha$ COSY cross-peak shapes and the TOCSY transfer efficiency in the region of Figure 1. Several segments showing sequential nOes and estimated $^3J_{HN-H\alpha}$ compatible with an extended β-strand conformation could be identified. No helical region can be detected. All amide protons are exchanged in less than 20 min (data not shown), so only long distance nOes involving backbone $^1H^N$ and $^1H^\alpha$ protons could formally define the limits of each different β-strand. At least three β-strands, Gln30 to Tyr34 (β_1), Asn37 to Ala41 (β_2), and Trp55 to Ser60 (β_3), can be identified from the NMR data summarized in Figure 2.

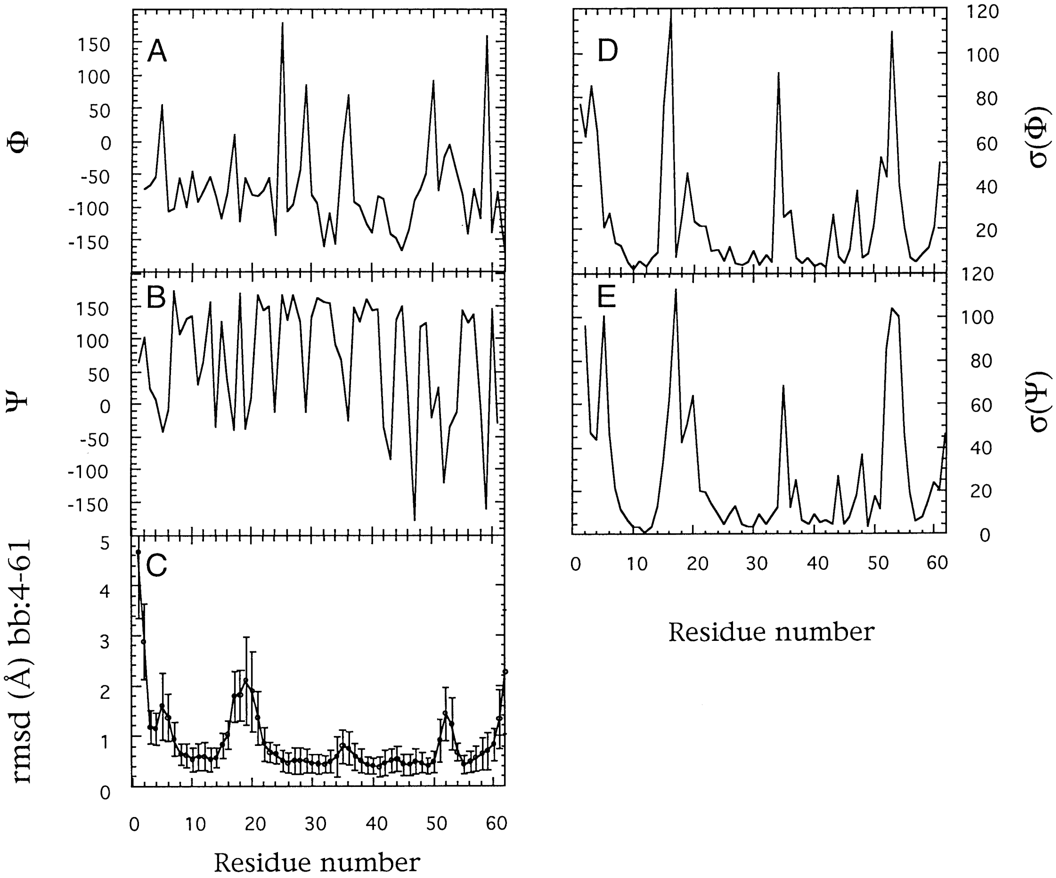

FIGURE 4: Structural analysis of the 23 best structures obtained after the rMD stage. (A and B) Mean backbone ϕ and ψ dihedral values of the E_II ensembles. (C) Positional rmsd values calculated with respect to the geometric mean structure of the corresponding E_II ensemble. (D and E) Standard deviation of ϕ and ψ angles compared to the mean value.

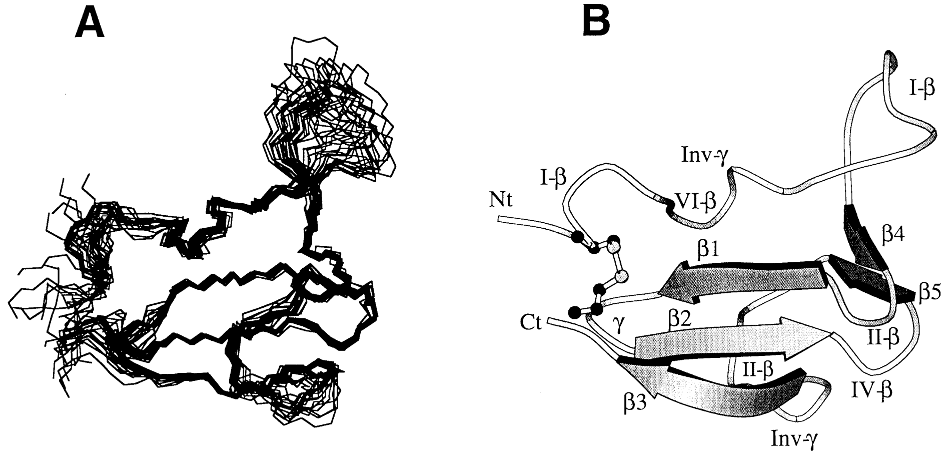

FIGURE 5: (A) Superimposition of the final ensemble of 23 structures calculated for CBD_{EGZ}. The coordinates were superimposed using the C, C^α, and N atoms from residues 26–27, 30–34, 37–41, 44–45, and 55–60. (B) MOLSCRIPT representation *(51)* of the closest CBD_{EGZ} structure to the geometric average of the E_II ensemble showing the secondary structure motifs.

These segments are connected by long-range $^1H^\alpha_i$–$^1H^\alpha_j$ and $^1H^\alpha_i$–$^1H^N_j$ nOes indicative of an antiparallel β-sheet.

Tertiary Structure Determination. An initial set of 770 distance constraints was used for the 62-residue protein structure determination. During the structure determination, additional constraints were assigned, giving a final set of 958 distance constraints. The sequential distribution of NOESY-derived distances according to their range is shown in Figure 3. The ensemble of 6 structures calculated with the final constraint set provided starting conformations for the final SA/rMD calculation giving 30 structures. Of this ensemble, seven were eliminated on the basis of experimental violation energy (>60 kcal mol^{-1}). The energetic and geometric statistics of the resulting ensemble, referred to as E_II, are given in Table 1 and Figure 4. Among these structures, the backbone was well-defined with an average rmsd of 0.8 Å for the backbone C, C^α, and N atoms when superimposed on the mean structure (1.2 Å for all heavy atoms). No violation greater than 0.35 Å was observed in E_II. Figure 5A shows the superposition of the 23 backbone-accepted structures of the E_II ensemble.

Overall Fold. The CBD_{EGZ} presents folding in "L" shape or "skiboot" shape. It consists of a well-defined part (the base of L) comprising several consecutive turns (residues

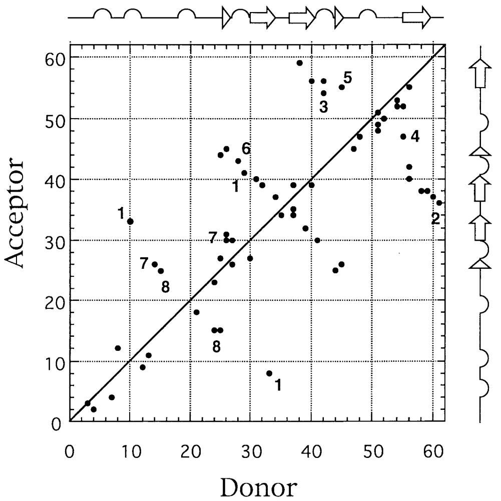

FIGURE 6: Matrix representation of the hydrogen-bonding networks present in the E_II ensemble. The numbers refer to the discussion of the hydrogen-bonding network identified in the text. Only the hydrogen bonds occurring in more than 25% of the E_II ensemble are shown. The donor residue appears along the *x*-axis and the acceptor residue along the *y*-axis. Analysis of the hydrogen bonds was done using the INSIGHTII routines using as criteria: (i) a distance less than 2.5 Å between the hydrogen of the donor group and the heavy acceptor atom; (ii) an angle formed by the heavy acceptor atom, the donor hydrogen, and the heavy donor atom linked to the hydrogen greater than 135°.

4−14) and an antiparallel β-sheet (residues 29−60). The upper part of L is formed by a less-ordered loop (residues 15−25) connected to the base by a small antiparallel β-sheet. The structure is maintained by a disulfide bridge between the two extremities of the peptide. The CBD_{EGZ} is mainly composed of β-strands and β- or γ-turns whose localizations are listed in Table 3; no helix structures are present in the molecule. The positions of the secondary structure elements of the protein are indicated on Figure 5B.

Secondary Structure Elements. The principal structural motif of the CBD_{EGZ} is the β-sheet composed of the three β-strands (β1 antiparallel to β2, β2 antiparallel to β3) (Figure 5B). It is observed in all the structures of the E_II ensemble, and the hydrogen bonds which stabilize it can be identified in Figure 6 by two lines perpendicular to the diagonal. The β_3 strand appears to be irregular at one end, forming a classic bulge from residues 58 to 59 *(34)*. A hydrogen bond between Leu38 H^N and Gly59 CO is observed in all E_II structures. In the same way, a G1T bulge can be localized at the beginning of the β_1 strand from residues Gly29 to Gln30 as we observe a hydrogen bond between Gly29 H^N and Ala41 CO in all E_II structures (Table 2).

A second short β-sheet can be identified between 26−27 (β4) and 44−45 (β5) with canonical ϕ, ψ values.

Side-Chain Conformation. The side-chain conformations are well-defined in most of the residues. Thirty residues adopt the same χ_1 rotamer in more than 80% of the structures, and 22 populate a single rotamer in all structures. These side chains correspond mainly to amino acids located in the β-sheets or in the turns. Only 10 residues out of 49 show a mixture of the 3 possible χ_1 rotamers; most of these residues correspond to the poorly-defined 51−54 region and the N- and C-terminus.

Of the eight aromatic side chains of the protein, five (Tyr10, Trp13, Tyr34, Tyr39, and Trp55) are localized in the same volume surrounding Pro49. In particular, Tyr34 and Trp55 are stacked over the Pro49 ring, with dihedral angles between the mean plane of Pro49 and the aromatic side-chain planes of Tyr34 and Trp55 of 30° ± 8 and 10° ± 3, respectively, with distances between the centroids of the side chains of 5.38 ± 0.20 Å and 4.0 ± 0.11 Å, respectively. These residues, with Tyr10 and Tyr39, form a bucket-like hydrophobic core closed at the bottom by Ile32 and containing Pro49 (Figure 7). The hydrophobic core also includes the Pro11 which is stacked over the Tyr10 ring (dihedral angle, 36.3° ± 8.8; mean centroid side-chain distance, 4.16 ± 0.17 Å) and the Trp13 side chain. All these residues are characterized by a low solvent accessibility and a very well-defined conformation. These residues populate only one χ_1 and χ_2 conformation, whose dihedral angles are defined at less than 15° standard deviation.

The three remaining aromatic residues Trp18, Trp43, and Tyr44 are located on the same side of the CBD_{EGZ} and are exposed to solvent (Figure 7). The side chain of the first residue (Trp18) which is at the apex of the large loop is loosely defined. On the other hand, the two other aromatic rings, Trp43 and Tyr44, appear to be well-defined, populating only one χ_1 conformer (47° ± 9.7 and −160° ± 5.7, respectively) and χ_2 conformer (−61° ± 4.7 and 106° ± 6.6, respectively). In the case of Tyr44, this is probably due to a stacking interaction between the aromatic rings of Tyr44 and His25 as we observe a dihedral angle of 18° ± 9 between the ring planes and a ring centroid distance of 3.8 ± 0.20 Å. The two Trp43 and Tyr44 rings are coplanar with a dihedral angle between the two aromatic planes of 19° ± 4 and a ring centroid distance of 9.4 ± 0.27 Å. The average distance between the aromatic ring of Tyr44 and the six-membered ring of Trp43 is 10.36 ± 0.22 Å.

Hydrogens Bonds: Turns. As the CBD_{EGZ} exhibits no helices and only 30% β-sheet, a large number of turns is observed (Table 3, Figure 5). The region 4−7 adopts two different major conformations. The first conformation (11 structures) can be assigned to a type I β-turn. A second irregular turn conformation is observed for 12 structures.

The region 8−13 consists of a double turn with residues Val9 to Asn12 forming a type VIa β-turn and residues Pro11 to Trp13 an inverse γ-turn. As usual in the type VI β-turn with a Pro residue at the i+2 position, the peptide bond between residues at positions i+1 (Tyr10) and i+2 (Pro11) is *cis (35, 36)*. Hydrogen bonds are observed between Asn12 H^N and Val9 CO as well as between Trp13 H^N and the Pro11 CO group in all 23 structures. The first turn is always further stabilized by a H-bond between the side chain NH_2 of Asn8 and the Asn12 CO group.

The large loop which extends from residues 15 to 26 is one of the most disordered parts of the CBD_{EGZ} structure with local rmsd values on the backbone of 6 Å. However, a type I turn can be defined between Trp18 and Gly21 for 17 structures.

A type II β-turn is localized from Glu27 to Gly30 in all 23 structures with a hydrogen bond between Gly30 H^N and Glu27 CO, and is further stabilized by a H-bond between the side-chain NH_2 of Asn26 and the Gln30 CO group in 22 of the 23 structures.

Residues 41−44, which end the β_2-strand, form in 19 structures a type IV β-turn with no CO_i−$H^N{}_{i+3}$ hydrogen bond. The Trp43 ϕ, ψ values (−139.8°; −72.5°) correspond however only to the "additional allowed" region of the Ramachandran plot.

Table 2: Bulges Present in the CBD_{EGZ}

residue[a]	ϕ_{i+1}[b]	ψ_{i+1}	ϕ_{i+2}	ψ_{i+2}	ϕ_x	ψ_x	type[a]
[24−25] [15]	−144 ± 9	−13 ± 11	178 ± 5	166 ± 5	−117 ± 33	126 ± 76	classic
[29−30] [41]	83 ± 4	−11 ± 5	−81 ± 4	133 ± 10	−84 ± 6	146 ± 4	G1T
[58−59] [38]	−116 ± 8	−9 ± 8	159 ± 16	−160 ± 11	−98 ± 7	126 ± 4	classic

[a] The nomenclature given refers to *(34)*. [b] Mean and standard deviation.

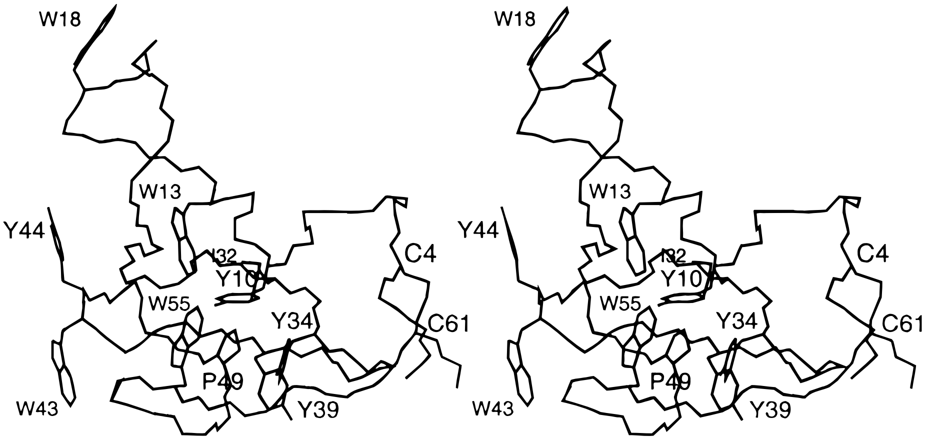

FIGURE 7: Stereoscopic view of the closest CBD_{EGZ} structure to the geometric average of the E_II ensemble with side-chain residues that are implicated either in the hydrophobic core of the protein or in the binding to cellulose. The disulfide bond between Cys4 and Cys61 is also presented.

Table 3: Turns Present in the CBD_{EGZ} and Their Occurrence in the E_II Ensemble of Structure

residue	ϕ_{i+1}[c]	ψ_{i+1}	ϕ_{i+2}	ψ_{i+2}	type[a]	H-bond[b]
Cys4−Ala7	−48 ± 11[d]	−31 ± 9	−60 ± 7	−28 ± 12	I(αα)	4−7(11)
Val9−Asn12	−46 ± 3[e]	134 ± 2	−94 ± 3	32 ± 5	VIa($\beta\alpha_R^{cis-Pro}$)	9−12(23)
Trp18−Gly21	−40 ± 5[f]	−60 ± 3	−87 ± 11	8 ± 9	I(αα)	18−21(14)
Glu27−Gln30	−44 ± 5[e]	126 ± 3	83 ± 4	−11 ± 5	II(βγ)	27−30(23)
Ala41−Tyr44	−89 ± 3[g]	−35 ± 2	−140 ± 5	−72 ± 4	IV(αγ)	−
Val48−Ser51	−49 ± 4[h]	125 ± 9	90 ± 18	−19 ± 22	II(βγ)	48−51(17)
Pro11−Trp13	−75 ± 1[e]	65 ± 3	−	−	inverse γ	11−13(23)
Lys35−Asn37	70 ± 3[i]	−42 ± 6	−	−	classic γ	35−37(15)
Gly50−Asp52	−77 ± 1[j]	65 ± 10	−	−	inverse γ	50−52(15)

[a] The nomenclature given refers to *(49)*, and a new nomenclature based on Ramachandran values *(50)* is given in parentheses. [b] Occurrence frequency of a hydrogen bond between CO_i and NH_{i-3} within the 23 model ensemble is given in parentheses. [c] Mean and standard deviation. [d] Calculation for a major conformation found in 11 structures. [e] Calculation over the 23 structures. [f] Calculation for a major conformation found in 17 structures. [g] Calculation for a major conformation found in 19 structures. [h] Calculation for a major conformation found in 17 structures. [i] Calculation for a major conformation found in 15 structures. [j] Calculation for a major conformation found in 15 structures.

The region from residue 48 to residue 54 is not well-ordered. A type II β-turn can, however, be recognized from Val48 to Ser51 with a H-bond between the Val48 CO and the H^N of Ser51 in 17 structures over the E_II ensemble. In 15 out of these 17 structures, this turn is followed by an inverse γ-turn from Gly50 to Asp52 stabilized by a hydrogen bond between the H^N of Asp52 and the CO of Gly50.

Hydrogens Bonds: Long-Range Stabilization between Secondary Structures. The two-dimensional plot in Figure 6 summarizes hydrogen bonds found in more than 7 out of 23 E_II structures. Apart from the hydrogen bonds occurring in the β-sheets and the turns, we observe several long-range hydrogen bond motifs probably important for the stabilization of the 3D structure of the protein. As noted above, the principal secondary motif of the structure is an antiparallel β-sheet, and most of the recognized long-range interactions connect the different parts of the molecule to this motif: the 14−26 loop which supports Trp18 is maintained above the large β-sheet by two interactions between this sheet and turns 9−12 and 27−30 (1); the C-terminal end is linked to the turn connecting the β1 and β2 strands (2); turn 41−44 interacts with the beginning of the β2 strand (3) and turn 48−51 with the beginning of the β3 strand (4). The large β-sheet and the small β-sheet are linked together by interactions between β3 and β5 (5) and turns 27−30 and 41−44 (6). Asn26, which is completely buried in the structure, links the first strand of the 14−26 loop to the β2 strand by two H-bonds with its NH_2 side chain (7). The two strands of the 14−26 loop are connected in a β-bulge-like conformation between 15 and 24−25 (8). The long-range hydrogen bonds are listed in Table 4.

Correlated Dihedral Angle Changes. Some local conformational subfamilies can be observed in the E_II ensemble from the cross-correlation of backbone dihedral variances shown in Figure 8. These correlations in no way imply the presence of dynamic processes, rather of two minima in the conformational space fulfilling equally well the NMR data. A correlated change of ϕ_{22} and the ψ_{14} corresponds to different possible orientations of the 14−22 loop above the large β-sheet (1). Similar correlations can be found between ϕ_{28}/ψ_{43} (2), corresponding to a second orientation of the Trp43 CO away from the Ala28 H^N, and ϕ_{43}/ϕ_{55} (3),

Table 4: Long-Range Hydrogen Bonds and Their Occurrence in the CBD_{EGZ}

donor	group	acceptor	group	*N*/23	function
10	HN	33	O	23	interaction of β-sheet 1/turn 9−12 (1)[a]
14	HN	26	OD1	12	stabilization of 14−26 loop (7)
15	HN	25	O	17	bulge [15] [24−25] (8)
24	HN	15	O	19	bulge [15] [24−25] (8)
25	HN	15	O	17	bulge [15] [24−25] (8)
26	HD21	31	O	11	interaction of β-sheet 1/loop 14−26 (7)
26	HN	45	O	23	stabilization of β-sheet 2
28	HN	43	O	19	interaction of turn 41−44/turn 27−30 (6)
29	HN	41	O	23	bulge [41] [29−30] (1)
31	HG	40	OG1	8	stabilization of β-sheet 1
32	HN	39	O	23	stabilization of β-sheet 1
33	HN	8	O	23	interaction of β-sheet 1/turn 9−12 (1)
34	HN	37	O	20	stabilization of β-sheet 1
38	HN	59	O	23	bulge [38] [58−59]
39	HN	32	O	23	stabilization of β-sheet 1
40	HN	56	O	23	stabilization of β-sheet 1
41	HN	30	O	21	bulge [41] [29−30]
42	HD21	54	O	15	interaction of β-sheet 1/turn 41−44 (3)
42	HD22	56	OG1	17	interaction of β-sheet 1/turn 41−44 (3)
42	HN	54	O	16	interaction of β-sheet 1/turn 41−44 (3)
44	HH	25	ND1	7	stabilization of β-sheet 2
44	OH	25	ND1	6	stabilization of β-sheet 2
45	HG1	55	NE1	9	interaction of β-sheet 1/β-sheet 2 (5)
45	HN	26	O	21	stabilization of β-sheet 2
55	HE1	47	OG	8	interaction of β-sheet 1/turn 48−51 (4)
55	HE1	47	O	17	interaction of β-sheet 1/turn 48−51 (4)
56	HN	40	O	23	stabilization of β-sheet 1
58	HN	38	O	23	bulge [38] [58−59]
59	HN	38	O	13	bulge [38] [58−59]
60	HG	37	OD1	8	interaction of β-sheet 1/C-terminus (2)
61	HN	36	O	19	interaction of β-sheet 1/C-terminus (2)

[a] Numbers in parentheses refer to the hydrogen-bonding network identified in the text.

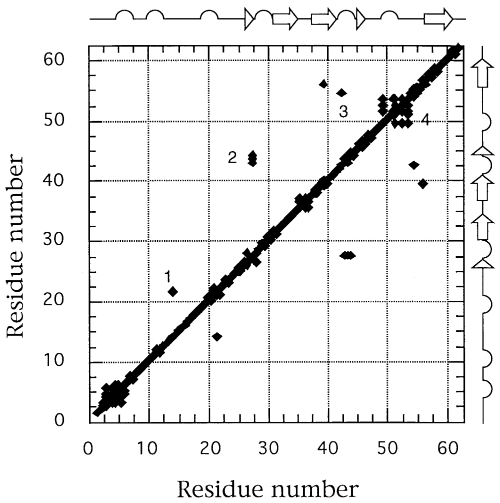

FIGURE 8: Covariances of ϕ and ψ angles throughout the E_II ensemble. Only values with $c_{i,j} > 0.75$ are shown. The numbers refer to the discussion of correlated dihedral changes identified in the text.

corresponding to a second orientation of the Ser54 CO group and the breaking of the two hydrogen bonds between the Ser54 CO and the Asn42 H^N and the side-chain NH_2. The ϕ_{50} variation, which reorients the Asp51 H^N toward the Gly49 CO, is accompanied by a large correlated difference of ϕ_{52}, ϕ_{53}, ϕ_{54}, and ψ_{51} (4). This locates the Ser52 H^N at less than 2 Å from the Val48 CO, and, in this case, segment 48−52 forms an irregular double turn stabilized by two hydrogen bonds (Val48 CO−Ser50 H^N and Gly49 CO−Ser51 H^N).

Electrostatic Surface. In Figure 9, the mean electrostatic surface calculated over the first 12 structures of the E_II family is shown. As expected from the cellulose-binding properties, the relative insolubility, and the low net charge (−**3**), the CBD_{EGZ} presents a rather neutral surface. Charged surfaces are apparent in only two regions of the protein: a high negative charge density due to the side chain of Glu27, located between Trp18 and Tyr44 in the putative cellulose-binding face, and a negative charge density near the disulfide bond, linked to the presence of Asp3. No high positive charge density region can be recognized in the mean electrostatic surface.

DISCUSSION

In this study, the structure of CBD_{EGZ} has been determined by 2D NMR and restrained molecular dynamics. The structure of CBD_{EGZ} presents an architecture in a “ski-boot” or “L” shape mainly composed of a well-defined β-sheet supporting a poorly ordered large loop. This folding appears to be original, and interrogation with available structure homology software revealed no significant similarity present in the Brookhaven protein database.

NMR Validation of CBD_{EGZ} Structure. Due to the relatively low quantity of pure protein that can be produced and due to the poor solubility of CBD_{EGZ} in solution, the set of NMR-derived restraints is fairly limited, and only nOe-derived distances restraints are used in the calculation process. Examination of our structures with the program PROCHECK-NMR *(37)* indicates a rather low percentage of residues in the allowed region (62%) and additional allowed region (31%) of the Ramachandran plot. Nevertheless, the validity of the calculated structures is corroborated by good agreement between qualitative NMR information and the calculated structure. Unusual chemical shifts observed for Gly20, Gly21, Pro49, Gly50, Ser31, and Ser54

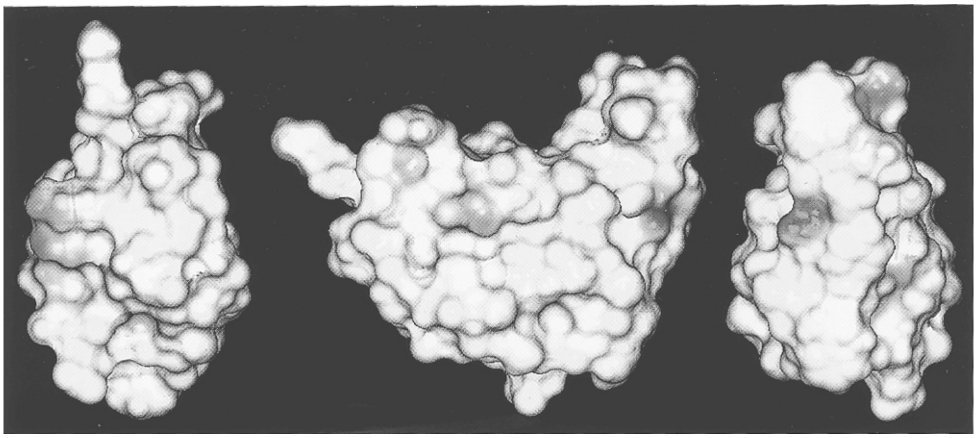

FIGURE 9: Electrostatic potential map of the CBD$_{EGZ}$. Middle: Orientation shown in Figure 5. The CBD has been rotated by −25° about the *z*-axis with respect to Figure 5. Right: Putative cellulose-binding side. The CBD has been rotated by +90° about the *y*-axis with respect to the middle view; the red area corresponds to the Glu27 side chain. Left: Disulfide bond side. The CBD has been rotated by −90° and 30° about the *y*-axis and the *x*-axis, respectively, with respect to the middle view; the orange area corresponds to the Asp3 side chain. Potential values between 0 and 3 kT and between 0 and −3 kT are scaled, between white, cyan, and blue and between white, yellow, and red, respectively.

can be explained by the aromatic ring current effect. For example, the upfield shift of Gly20 and Gly21 amide proton resonances may be due to their position above the Trp18 aromatic ring and Ser54 $^1H^\alpha$ is located above the aromatic ring of Trp43. Upfield shift of Gly50 $^1H^N$ and downfield shift of Ser31 $^1H^\alpha$ could be related to their position respectively above and coplanar to the Trp55 aromatic ring. In the same way, the dispersed chemical shift values of the Pro49 protons (see Table 1 of the supporting information) correlate with its central position in the middle of the hydrophobic core that is mainly composed of aromatic side chains. Several coupling constant $^3J_{\text{HN-H}\alpha}$ and $^3J_{\text{H}\alpha\text{H}\beta}$ values that have been estimated from NMR spectra (see Figure 2) are consistent with the dihedral angle values extracted from the structures although none were used in the calculation process.

Disordered Regions. Although the hydrophobic core embedded within the main β-sheet defines a fairly rigid region of the protein, the loop Ser15−Asn26 is less ordered, possibly due to the presence of two consecutive glycine residues (Gly20 and Gly21). On one extremity of the loop, the Asn8−Trp13 overall region, which adopts a double turn conformation, constitutes the first rigid anchor on the well-structured β-sheet region of the protein. Hence, the limited rotation of the Tyr10 aromatic ring, which is in the middle of this double turn structure, stacked in between Pro11 and Trp13, shows how much this region of the protein is constrained. At the other extremity, the well-defined type II β-turn localized from Glu27 to Gln30 also constitutes the second rigid anchoring of the loop. These two extremities are linked together via hydrogen bonding with the buried Asn26 residue. We could thus argue that the lack of structural information along the large loop (Ser15−Asn26) reflects an intrinsic mobility of this region of the protein in solution. This idea is also supported by the presented sequence alignment (see Figure 11) with several chitinase domains, where this region of CBD$_{EGZ}$ is not conserved at all and appears as an extension of the more conserved region that folds into the rigid part of the molecule. As discussed below, this region of the molecule is a part of the putative cellulose-binding site of CBD$_{EGZ}$, and its mobility may be important for its cellulose-binding ability.

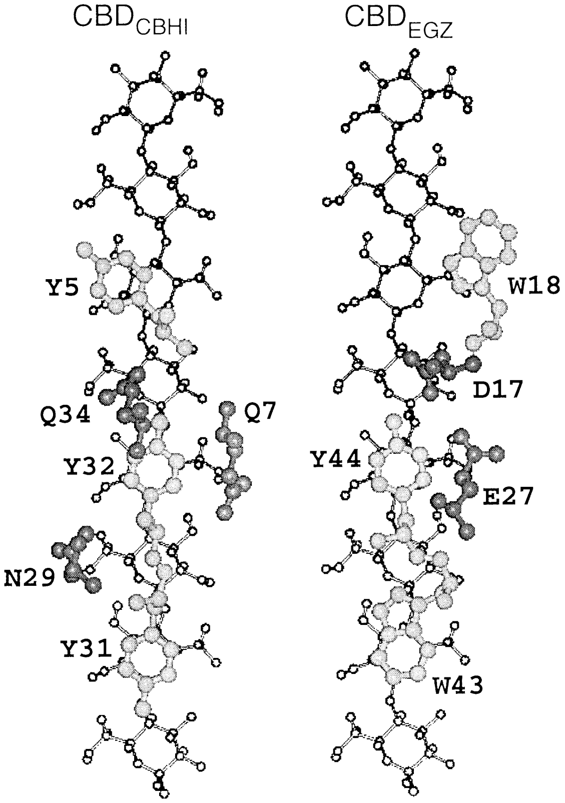

FIGURE 10: CBD$_{CBHI}$ and CBD$_{EGZ}$ placed above the cello-oligosaccharide with aromatic rings stacked on top of sugar rings. The distance between three glucose moieties of 10.3 Å has been proposed from X-ray crystallography data and corresponds to the b-unit cell of cellulose *(43)*. Only the heavy atoms (N, C$^\alpha$, O) of residues Tyr5, Gln7, Asn29, Tyr31, Tyr32, and Gln34 of the CBD$_{CBHI}$ cellulose-binding site and of residues Asp17, Trp18, Glu27, Trp43, and Tyr44 of the putative cellulose-binding site of CBD$_{EGZ}$ are shown for clarity. Aromatic side chains are shown in light gray, polar, or charged side chains are shown in dark gray and cello-oligosaccharides are shown in black.

CBD$_{EGZ}$ Stability. Previous studies suggested that the Cys4−Cys61 disulfide bond is not required for the CBD$_{EGZ}$ stability *(4)*, suggesting a major contribution of hydrophobic interactions. We indeed observe a hydrophobic core composed of seven well-defined aliphatic and aromatic side

```
GUN1A     353      DPGEYPAWDPN-------QIYTN-EIVYHNGQLWQAKW-WTQN-QEPGANQYG--------PWEPLG                                401
GUN1B     421      DPGEYPAWDPT-------QIYTN-EIVYHNGQLWQAKW-WTQN-QEP-GYPYG--------PWEPLN                                488
GUN2      364      DPGDYPAWDPN-------TIYTD-EIVYHNGQLWQAKW-WTQN-QEP-GYPYG--------PWEPLN                                409
ORF1      118      GSCTAPVWSSS-------TAYNGGWQVSYNGHTYTAKW-WTQG--VP-SSSTG-----DGSPWNDVGVCS                             172
ORF2      135      ASRTAAAWSAG-------TAYNGGTQVSYNGRNYTAKW-WTQG-NVP-SSSTG-----EGQPWASNGPCSG                            190
ORF3      117       GNCGGAWSSG-------TAYNGGAQVSYNGHLYTAKW-WTQG-NVP-SSSTG-----DGQPWTDSGICGG                            171
ORF4      136   GVRCAVLLPAWNTG-------TAYNGGVQVSYNGHTYTAKW-WTQG-NIP-SSSTG-----DGQPWTDNGVCG                             194
CHIII     390        TCAATWSSS-------TAYNGGATVAYNGHNYQAKW-WTQG-NVP-SSSTG-----DGQPWADLGVC                              443
JANLIV    149        TCALAWAAG-------TAYSAGATVSYAGTNYRANY-WTQG-DNP-STSSG---GATGKPWTSQEICST                            205
GRISEU     30       ATCATAWSSS-------SVYTNGGTVSYNGRNYTAKY-WTQN-ERPG----------TSDVWADKGACGT                            80
FURNIS     32     VDCSALAEWQSD-------TIYTGGDQVQYNGSAYQANY-WTQN-NDPE-QFSG-----DYQVWKLLDACT                             87
YHEB1      25         MEAWNNQ-------QGGNK-YQVIFDGKIYENAW-WVSSTNCPGKAKAN----DATNPWRLKRTATAAEISQFGNTLSCE              91
YHEB2     129         VVAWNKQ-------QGGQT-WYVVFNGAVYKNAW-WVASSNCPGDAKSN----DASNPWRYVRAATATEISETSNPQSCT              194
YHEB3     225      DNYAVVAWKGQ-------EGSST-WYVIYNGGIYKNAW-WVGAANCPGDAKEN----DASNPWRYVRAATATEISQYGNPGSCS              295
YHEB4     333      NDYSLQAWSGQ-------EGSEI-YHVIFNGNVYKNAW-WVGSKDCPRGTSAE----NSNNPWRLERTATAAELSQYGNPTTCE              403
YHEB5     454      KYEPAKAWSAS-------TVYVKGDRVVVDGQAYEALF-WTQS-DNPALVANQNATGSNSRPWKPLGKAQSYSNEELNNAPQFNPETLYASD 537
HARVE     507      GSNCAAAWDAN-------TVYVEGDQVSHDGATWVAGW-YTRG-EEPGTTG-------EWGVWKKASDSSCG                           562
LICHE     547       ACSYDEWKET-------NAYTGGERVAFNGKVYEAKW-WTKGDRRINPVN------GAYGGWSEAANNRKSNG                         604
AERCAV1   764    DPDAANHPAWSAG-------TVYNTNDKVSHKQLVWQAKY-WTQGN-EPS---------RTADQWKLVSQVQL                            815
AERCAV2   816        VQLGWDAG-------VVYNGGDVTSHNGRKWKAQY-WTKGD-EPG----------KAAVWVDQGAASCN                           865
ALTSO     778  VTIKDGAEYPTWDRS-------TVYVGGDRVIHNSNVFEAKW-WTQG-EEPG----------TADVWKAVTN                              820
SERMA     449     LPIMTAPAYVPG-------TTYAQGALVSYQGYVWQTKWGYITS--APG----------SDSAWLKVGRLA                             499
CBD         1  ADCANANVYPNWVSKDWAGGQPTHNEAGQSIVYKGNLYTANW-YTAS--VPG----------SDSSWTQVGSCN                            61

consensus      .......th.sWsst.......pshst...V.asGp.apstW Wsts.p.Ps..........tstsWt..t........................
```

FIGURE 11: Sequential alignment of the CBD_{EGZ} and several domains of chitinases, endoglucanases, and unknown proteins of bacterial origin. GUN1A and GUN1B, C-terminal domains a and b of *Bacillus sp.* strain N-4 endoglucanase A *(52)*; GUN2, C-terminal domain of *Bacillus sp.* strain N-4 endoglucanase B *(52)*; ORF1 to ORF4, *Aeromonas sp.* open reading frames coding for chitinase-like proteins *(53)*; CHIII, chitin-binding domain of *Aeromonas sp.* chitinase II *(54)*; JANLIV, chitin-binding domain of *Janthinobacterium lividum* chitinase *(55)*; GRISEU, chitin-binding domain of *Streptomyces griseus* chitinase *(48)*; FURNIS, N-terminal domain of *Vibrio furnissii* chitodextrinase *(56)*; YHEB1 to YHEB5, first to fifth repeat domains of *E. coli* hypothetical 97 kDa protein (SwissProt Accession No. P13656); HARVE, internal subdomain of *Vibrio harveyi* chitinase A (GenBank Accession No. U71214); LICHE, C-terminal domain of *Bacillus licheniformis* TP chitinase (GenBank Accession No. U81496); AERCAV1 and AERCAV2, first and second subdomains of *Aeromonas caviae* chitinase A *(57)*; ALTSO, C-terminal domain of *Alteromonas sp.* strain O-7 chitinase A *(58)*; SERMA, C-terminal domain of *Serratia marcescens* chitinase B *(59)*. The numbers at the left and right signify first and last residue positions in the protein sequence. The positions of the β-sheets of CBD_{EGZ} are squared. The residues of the CBD_{EGZ} hydrophobic core which are conserved in the alignment are in light gray. The two solvent-exposed aromatic residues in CBD_{EGZ} which are conserved in the alignment are in dark gray. The consensus sequence corresponds to a 80% consensus in the alignment. t, turn-like residue; h, hydrophobic; s, small; p, polar; a, aromatic.

chains buried in the center of the molecule (see Figure 7). It is likely that these residues participate in the stability of the protein as most of the ring centroid distances and aromatic side-chain dihedral angles are in the range described for aromatic−aromatic interactions *(38)*. Hydrophobicity is not only a property of the inside of the CBD_{EGZ} structure. The accessible surface is unusually hydrophobic; the surface area of the mean structure of the ensemble is comprised of 69% nonpolar residues, 26% polar, and 5% charged, as calculated with the program Vadar *(39, 40)*. This feature certainly accounts for the observed low solubility of the protein, but can also be related to its known low stability due to a fragile balance between buried and exposed hydrophobicity. We have observed unfolding of the CBD in either mild acidic conditions (pH < 4) or moderate temperature (t > 50 °C), and even completely buried amide protons probably involved in hydrogen bonds are exchanged within 30 min.

Evidence of a Cellulose-Binding Site. Face-to-face stacking of aromatic rings of amino acids with sugar rings has been observed in cases of protein/sugar interaction *(41)*. Aromatic tyrosine or tryptophan rings of the cellulose-binding site of CBDs are thought to stack directly against the pyranose rings of cellulose, providing a hydrophobic driving force for the binding *(12, 14, 42)*. CBD_{EGZ} structure presents three exposed and aligned aromatic side chains: Trp18, Trp43, and Tyr44 (Figures 7 and 10) on the same face. The distance between the aromatic centroids of Trp43 and Tyr44 (10.4 ± 0.27 Å) matches the distance between three glucose moieties in the cellulose structure (Figure 10) *(43)*, a situation which is reminiscent of those of the Tyr31 and Tyr32 cellulose-binding residues of CBD_{CBHI} *(12)*. The spatial position of the third aromatic ring (Trp18) at the apex of the large loop is poorly defined, and the distance between the two aromatic ring centroids of Tyr44 and Trp18 varies from 13.7 to 17.3 Å in the ensemble. In the structure where the three aromatic rings are almost coplanar, the distance between Tyr44 and Trp18 rings is about 15 Å. Thus, the total length of the CBD_{EGZ} putative cellulose-binding site formed by these three aromatic side chains would then cover five or six glucose moieties (Figure 10). It is important to stress that the relative orientation (parallel or antiparallel) of the aromatic planar strip of CBD versus the cellulose chain presented in Figure 8 simply reflects the complementarity between the periodicity of the aromatic rings of the binding face and the periodicity of the glucose moieties in the cellulose structure. Binding to natural cellulose is certainly much more complex considering the variety of cellulose super-structures (crystal irregularity and various type of cellulose crystal) in its natural plant cell-wall environment (for review, see *44)*. In this context, one could argue that the supplementary degree of freedom of the CBD_{EGZ} cellulose-binding site due to the intrinsic flexibility of one (Trp18) of its putative binding residues is an advantage that might allow binding to a wider range of cellulose substrates.

It is noticeable that among the four CBD structures currently available, the three that bind crystalline cellulose (CBD_{CBHI}, CBD_{Cex}, and CBD_{Cip}) show a planar cellulose-binding surface composed of a strip of three aromatic side-chain residues with one of their ring faces fully exposed to the solvent. Thus, beyond the apparent differences exist a series of properties that appear to be conserved throughout all CBDs. In this context, it might be interesting to note that recent studies on CBD_{CBHI}, involving site-directed mutagenesis and MD simulations, also support the involvement of invariant polar residues (Gln7, Asn29, and Gln34) in the binding of this family I CBD onto cellulose *(45, 46)*. As shown in Figure 10, the putative cellulose-binding site of CBD_{EGZ} also contains the exposed Asp17 side chain located in between Trp18 and Tyr44 aromatic rings and the Glu27 side chain located on one side of the aromatic planar strip. Two potential roles have been proposed for the polar residues present on the cellulose-binding face *(14)*: (i) to stabilize the appropriate orientation of the cellulose-binding residues, mainly through the formation of a hydrogen bond; (ii) to establish hydrogen-bonding interactions with oxygen atoms and/or hydroxyl groups of glucose moieties of the cellulose microcrystal. In the latter case, these residues not only contribute to the overall affinity of the CBDs for the substrate, but also may account for the cellulose-disrupting properties, which have been proposed for CBDs, by destabilizing the hydrogen-bonded structure of cellulose *(7, 8)*. Accordingly, in the CBD_{EGZ}, the position of Asp17 suggests a role of stabilization of the Tyr44 aromatic ring orientation, or a direct interaction with cellulose by forming a hydrogen bond with polar oxygen and hydroxyl groups of glucose moieties. The location of the Glu27 side chain, lateral to the aromatic ring of Tyr44, is consistent with its putative role in the formation of a hydrogen bond with cellulose. Site-directed mutagenesis studies are in progress in an attempt to define the role, if any, of Asp17, Trp18, Glu27, Trp43, and Tyr44.

Occurrence of CBD-like Domains in Chitinases. As mentioned in the introduction, the CBD_{EGZ} sequence exhibits no similarity with any known CBD, and, as a consequence, constitutes a new family. In fact, previous primary sequence comparisons had revealed local resemblance between residues 29−61 of the CBD_{EGZ} and bacterial chitinases which could hardly be taken as biologically meaningful in the absence of additional information about those matching regions. The availability of the 3D structure of the CBD_{EGZ} allowed us to reconsider those alignments. First, we realized that a region covering residues 29−61 corresponded to the β-sheet core of the CBD_{EGZ}. Using this region alone as a query, we have found additional chitinase regions as well as regions from endoglucanases and of an ORF of unknown function in *E. coli*. Analysis using the CLUSTAL W package *(47)* allowed us after editing the 15−24 loop to obtain a consensus sequence shared by 22 other sequences (Figure 11). It should be noted that a previous study revealed similarity between a subset of the sequences used here and chitin-binding domains of *Bacillus circulans* chitinases A and D *(48)*. Surprisingly, our derived consensus does not allow us to include *B. circulans* chitin-binding domain sequences in this family since *B. circulans* does not exhibit the highly conserved stWWst motif likely to be essential for CBD−cellulose interactions. All the residues forming the hydrophobic core in CBD_{EGZ} appeared to be either conserved or replaced by a similar residue in all other enzymes. The conserved region corresponds for the CBD to the two β-sheets. Most interestingly, in several chitinases, this consensus appears to be part of the regions proposed to bind chitin; they are about 50−60 residues long, include 2 Cys residues, 1 at each extremity, and are bracketed by long Pro/Thr-rich regions. Clearly, these features are reminiscent of the CBD_{EGZ}. It is hence highly tempting to propose that these regions form individual functional domains, structurally related to the CBD_{EGZ}. This supports the possibility that CBD_{EGZ} and some chitin-binding domains might form a new family of functional and structural related protein modules. Presumably, this family does not include the chitin-binding domains of *Bacillus circulans* chitinases. Biochemical and mutagenesis experiments are currently under way to test this prediction.

Secretion Motif. Our long-term goal is to identify structural information that specifically targets EGZ to the cell exterior through the Out machinery. Mutagenesis studies showed that each of the EGZ functional domains (CD, linker, CBD) is required for EGZ secretion *(4)*. This implies that the secretion motif is probably composed of several submotifs interacting either directly with one or several proteins of the Out machinery or together to form a unique secretion motif. In any case, a likely possibility is that the secretion motif(s) is (are) surface-exposed. The surface of CBD_{EGZ} is mainly composed of hydrophobic residues except in two regions: one in the putative cellulose-binding site, and the other around Asp3 (Figure 9). That this latter negative charge could be instrumental in establishing contacts with the Out proteins is an attractive hypothesis. Ongoing site-directed mutagenesis will allow us to test this proposition also.

ACKNOWLEDGMENT

Thanks are due to J. Haiech for his interest in this work.

SUPPORTING INFORMATION AVAILABLE

One table containing chemical shift assignments (4 pages). Ordering information is given on any current masthead page.

REFERENCES

1. Barras, F., Van Gijsegem, F., and Chatterjee, A. K. (1994) *Annu. Rev. Phytopathol. 32*, 201−234.
2. Py, B., Chippaux, M., and Barras, F. (1993) *Mol. Microbiol. 7*, 785−793.
3. Py, B., Bortoli-German, I., Haiech, J., Chippaux, M., and Barras, F. (1991) *Protein Eng. 4*, 325−333.
4. Bortoli-German, I., Brun, E., Py, I., Chippaux, M., and Barras, F. (1994) *Mol. Microbiol. 11*, 545−553.
5. Brun, E., Gans, P., Marion, D., and Barras, F. (1995) *Eur. J. Biochem. 231*, 142−148.
6. Din, N., Gilkes, N. R., Tekant, B., Miller, R. C., Warren, R. A. J., and Kilburn D. G. (1991) *Bio/Technology 9*, 1096−1099.
7. Din, N., Forsythe, I. J., Burtnick, L. D., Gilkes, N. R., Miller, R. C. J., Warren, R. A. J., and Kilburn, D. G. (1994) *Mol. Microbiol. 11*, 747−755.
8. Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller, R. C., and Warren, R. A. J. (1991) *Microbiol. Rev. 55*, 303−315.
9. Tomme, P., Warren, R. A. J., and Gilkes, N. R. (1995b) *Adv. Microb. Physiol. 37*, 1−81.
10. Kraulis, P. J., Clore, G. M., Nilges, M., Jones, T. A., Pettersson, G., Knowles, J., and Gronenborn, A. M. (1989) *Biochemistry 28*, 7241−7257.
11. Linder, M., Mattinen, M.-L., Kontteli, M., Lindberg, G., Stahlberg, J., Drakenberg, T., Reinikainen, T., Pettersson, G., and Annila, A. (1995) *Protein Sci. 4*, 1056−1064.
12. Reinikainen, T., Teleman, O., and Teeri, T. T. (1995) *Proteins: Struct., Funct., Gemet. 22*, 392−403.
13. 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) *Biochemistry 34*, 6993−7009.
14. Tormo, J., Lamed, R., Chirino, A. J., Morag, E., Bayer, E. A., Shoham, Y., and Steitz, T. A. (1996) *EMBO J. 15*, 5739−5751.
15. Johnson, P. E., Joshi, M. D., Tomme, P., Kilburn, D. C., and McIntosh, L. P. (1996) *Biochemistry 35*, 14381−14394.
16. Rance, M., Sørensen, O. W., Bodenhausen, G., Wagner, G., Ernst, R. R., and Wüthrich, K. (1983) *Biochem. Biophys. Res. Commun. 117*, 479−485.
17. Braunschweiler, L., and Ernst, R. R. (1983) *J. Magn. Reson. 53*, 521−528.

18. Davis, D. G., and Bax, A. (1985) *J. Am. Chem. Soc. 107*, 2820−2821.
19. Jeener, J., Meier, B. H., Bachmann, P., and Ernst, R. R. (1979) *J. Chem. Phys. 71*, 4546−4553.
20. Macura, S., Hyang, Y., Suter, D., and Ernst, R. R. (1981) *J. Magn. Reson. 43*, 259−281.
21. Marion, D., Ikura, M., Tschudin, R., and Bax, A. (1989) *J. Magn. Reson. 85*, 393−399.
22. Plateau, P., and Guéron, M. (1982) *J. Am. Chem. Soc. 104*, 7310−7311.
23. Pearlman, D. A., Case, D. A., Caldwell, J. C., Seibel, G. L., Singh, U. C., Weiner, P., and Kollman, P. A. (1991) *Amber 4.0*, University of California, San Fransisco.
24. Blackledge, M. J., Medvedeva, S., Poncin, M., Guerlesquin, F., Bruschi, M., and Marion, D. (1995) *J. Mol. Biol. 245*, 661−681.
25. Brooks, B. R., Broccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. J. (1983) *J. Comput. Chem. 4*, 187−217.
26. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984) *J. Am. Chem. Soc. 106*, 765−784.
27. Singh, U. C., and Kollman, P. A. (1983) *J. Comput. Chem. 5*, 129−145.
28. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) *J. Chem. Phys. 81*, 3684−3690.
29. Connolly, M. L. (1983) *Science 221*, 709−713.
30. Gilson, M. K., Sharp, K. A., and Honig, B. H. (1987) *J. Comput. Chem. 9*, 327−335.
31. Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley and Sons, New York.
32. Gross, K. H., and Kalbitzer, H. R. (1988) *J. Magn. Reson. 76*, 87−99.
33. Wishart, D. S., Sykes, B. D., and Richards, F. M. (1991) *J. Mol. Biol. 222*, 311−333.
34. Chan, A. W. E., Hutchinson, E. G., Harris, D., and Thorton, J. M. (1993) *Protein Sci. 2*, 1574−1590.
35. Lewis, P. N., Momany, F. A., and Scheraga, H. A. (1973) *Biochim. Biophys. Acta 303*, 211−229.
36. Yao, J., Dysson, H. J., and Wright, P. E. (1994) *J. Mol. Biol. 243*, 754−766.
37. Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R., and Thornton, J. (1996) *J. Biomol. NMR 8*, 477−486.
38. Burley, S. K., and Petsko, G. A. (1985) *Science 229*, 23−28.
39. Richards, F. M. (1974) *J. Mol. Biol. 82*, 1−14.
40. Richmond, T. J. (1984) *J. Mol. Biol. 178*, 63−89
41. Quiocho, F. A. (1996) *Annu. Rev. Biochem. 55*, 287−315.
42. Mattinen, M.-L., Kontteli, M., Kerovuo, J., Linder, M., Annila, A., Lindberg, G., Reinikainen, T., and Drakenberg, T. (1997) *Protein Sci. 6*, 294−303.
43. Preston, R. D. (1986) in *Cellulose: Structure, Modifications and Hydrolysis* (Young, R. A., and Rowell, R. M., Eds.) pp 3−26, John Wiley and Sons, New York.
44. Tomme, P., Warren, R. A. J., Miller, R. C., Jr., Kilburn, D. G., and Gilkes, N. R. (1995) in *Enzymatic Degradation of Insoluble Polysaccharides* (Salder, J. M., and Penner, M., Eds.) pp 142−161, American Chemical Society, Washington, DC.
45. Hoffren, A.-M., Teeri, T. T., and Telemann, O. (1995) *Protein Eng. 8*, 443−450.
46. Linder, M., Lindeberg, G., Reinikainen, T., Teeri, T., and Pettersson, G. (1995) *FEBS Lett. 372*, 96−98.
47. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res. 22*, 4673−4680.
48. Ohno, T., Armand, S., Hata, T., Nikaidou, N., Henrissat, B., Mitsumoti, M., and Watanabe, T. (1996) *J. Bacteriol. 178*, 5065−5070.
49. Richardson, J. S. (1981) *Adv. Protein Chem. 34*, 167−339.
50. Wilmot, C. M., and Thornton, J. M. (1990) *Protein Eng. 3*, 479−493.
51. Kraulis, P. J. (1991) *J. Appl. Crystallogr. 24*, 946−950.
52. Fukumori, F., Sashihara, N., Kudo, T., and Horikoshi, K. (1986) *J. Bacteriol. 168*, 479−485.
53. Shiro, M., Ueda, M., Kawaguchi, T., and Arai, M. (1996) *Biochim. Biophys. Acta 1305*, 44−48.
54. Ueda, M., Kawaguchi, T., and Arai, M. (1994) *J. Ferment. Bioeng. 78*, 205−211.
55. Gleave, A. P., Taylor, R. K., Morris, B. A., and Greenwood, D. R. (1995) *FEMS Microbiol. Lett. 131*, 279−288.
56. Keyhani, N. O., and Roseman, S. (1996) *J. Biol. Chem.* 271, 33414−33424.
57. Sitrit, Y., Vorgias, C. E., Chet, I., and Oppenheim, A. B. (1995) *J. Bacteriol. 177*, 4187−4189.
58. Tsujibo, H., Orikoshi, H., Tanno, H., Fujimoto, K., Miyamoto, K., Imada, C., Okami, Y., and Inamori, Y. (1993) *J. Bacteriol. 175*, 176−181.
59. Harpster, M. H., and Dunsmuir, P. (1989) *Nucleic Acids Res. 17*, 5395−5395.

BI9718494