The Solution Structure of the CBM4-2 Carbohydrate Binding Module from a Thermostable *Rhodothermus marinus* Xylanase[†]

Peter J. Simpson,[‡,§] Stuart J. Jamieson,[‡] Maher Abou-Hachem,[||] Eva Nordberg Karlsson,[||] Harry J. Gilbert,[⊥] Olle Holst,[||] and Michael P. Williamson*[,‡]

Krebs Institute, Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2TN, U.K., Department of Biotechnology, Center for Chemistry and Chemical Engineering, Lund University, P.O. Box 124, SE-221 00 Lund, Sweden, and Department of Biological and Nutritional Sciences, University of Newcastle, Newcastle upon Tyne NE1 7RU, U.K.

Received November 26, 2001; Revised Manuscript Received February 25, 2002

ABSTRACT: The solution structure is presented for the second family 4 carbohydrate binding module (CBM4-2) of xylanase 10A from the thermophilic bacterium *Rhodothermus marinus*. CBM4-2, which binds xylan tightly, has a β-sandwich structure formed by 11 strands, and contains a prominent cleft. From NMR titrations, it is shown that the cleft is the binding site for xylan, and that the main amino acids interacting with xylan are Asn31, Tyr69, Glu72, Phe110, Arg115, and His146. Key liganding residues are Tyr69 and Phe110, which form stacking interactions with the sugar. It is suggested that the loops on which the rings are displayed can alter their conformation on substrate binding, which may have functional importance. Comparison both with other family 4 cellulose binding modules and with the structurally similar family 22 xylan binding module shows that the key aromatic residues are in similar positions, and that the bottom of the cleft is much more hydrophobic in the cellulose binding modules than the xylan binding proteins. It is concluded that substrate specificity is determined by a combination of ring orientation and the nature of the residues lining the bottom of the binding cleft.

Plant and algal cell walls are composite insoluble structures that are highly recalcitrant to chemical and biological degradation. Consequently, plant cell wall degrading microbes have evolved a complex synergistic repertoire of enzymes to enable efficient digestion. This activity is central to the recycling of biomass, and therefore is essential for the turnover of cellulose and xylan, which are the most abundant organic molecules in the biosphere (*1*). In aerobic microbes, plant cell wall hydrolases are in general modular enzymes, which consist typically of a catalytic module joined by flexible linker sequences to one or more carbohydrate binding modules (CBMs)[1] (*2, 3*). The function of the CBMs is largely to attach the enzyme to its substrate, and therefore enhance the rate of catalysis by increasing the probability of enzyme/substrate interaction (*4*). The majority of CBMs bind specifically to cellulose (*5*), although CBMs have been characterized that bind to xylan, mannan, or other polysaccharides. CBMs have been classified into 28 families based on sequence similarities (*3*). Many of these families (for example 1, 2a, 3a, 5, and 10) bind only to crystalline cellulose (and often also to chitin). Family 4, however, contains members that bind to a range of single-stranded polysaccharides, including amorphous cellulose, β-1,3-glucan, and xylan (*6, 7*). The binding specificity of soluble cellulose binding CBM4s is conferred by the cleft structure of the ligand binding site, which cannot accommodate crystalline cellulose, and whose shape is complementary to the target polysaccharide (*8*). Family 4 is therefore a particularly interesting family for the study of protein carbohydrate recognition.

The objective of this paper is to study CBM4-2, the second family 4 CBM from *Rhodothermus marinus* xylanase 10A (Xyn10A). The modular enzyme consists of two *N*-terminal family 4 CBMs, followed by a domain of unknown function, a catalytic module, and a second domain of unknown function (*9*). *R. marinus* is an aerobic thermophilic bacterium first isolated from marine hot springs in Iceland, and is a useful source of thermophilic enzymes (*10*). CBM4-2 has previously been shown to bind strongly to xylan and barley β-glucan, and more weakly to laminarin and lichenan. The affinity for xylan and xylo-oligosaccharides is very high, with K_d values of 10 μM at 65 °C and 5 μM at 30 °C for xylohexaose (which contrasts with a more typical family 2b xylan binding domain, which has a K_d of 290 μM for xylohexaose at 30 °C: ref *11*). Interestingly, it also has affinity for cellooligosaccharides and phosphoric acid swollen

[†] The BBSRC provided support for P.J.S. and S.J.J., and the BBSRC and the Wellcome Trust gave equipment grants. E.N.K. was supported by the K&A Wallenberg Foundation, and M.A.-H. received support from the Helge Axson Johnsons Foundation.

* To whom correspondence should be addressed at the Krebs Institute, Department of Molecular Biology and Biotechnology, University of Sheffield, Firth Court, Western Bank, Sheffield S10 2TN, U.K. Email: m.williamson@sheffield.ac.uk. Tel: +44 114 222 4224. Fax: +44 114 272 8697.

[‡] University of Sheffield.

[§] Present address: Department of Biochemistry, Imperial College, Exhibition Rd., London SW7 2AY, U.K.

[||] Lund University.

[⊥] University of Newcastle.

[1] Abbreviations: CBM, carbohydrate binding module; CBM4-2, second family 4 CBM from *Rhodothermus marinus* xylanase 10A; CBD_{N1} and CBD_{N2}, first and second *N*-terminal CBMs from from *Cellulomonas fimi* endoglucanase C; HSQC, heteronuclear single quantum coherence; X_5, xylopentaose.

10.1021/bi012093i CCC: $22.00 © 2002 American Chemical Society
Published on Web 04/16/2002

cellulose, although as expected for a CBM4, it does not bind to crystalline cellulose (*12*). It is therefore of interest to study the structure of the CBM4, to understand how it binds so strongly to xylan, and yet also exhibits some affinity for amorphous cellulose.

MATERIALS AND METHODS

Protein Production and Purification. The region of the Xyn10A gene (*xyn10A*) encoding CBM4-2 (nucleotides 631−1119, corresponding to amino acid residues Leu211−Ile373 of the full-length protein, SwissProt P96988) was cloned into the expression vector pET-25b(+) (Novagen, Madison, WI) as described (*12*), and transformed into *E. coli* BL21(DE3). The cells were grown on defined medium (*13*), induced by 0.5 mM isopropyl β-D-thiogalactoside for 3 h at 37 °C, and purified, exploiting the C-terminal His tag, using an iminodiacetic acid−copper affinity column as described (*12*).

NMR Experiments. Samples for NMR were typically 1.5 mM CBM4-2 in 50 mM $CaCl_2$, 50 mM sodium acetate-d_3, pH 6.0, containing 10% D_2O, 10 mM sodium azide, and 0.1 mM sodium trimethylsilylpropionate (TSP). Most measurements were carried out at 313 K. NMR spectra were acquired at 500 and 600 MHz on Bruker DRX spectrometers using 5 mm probeheads with *z* gradients. Assignments and structure restraints were obtained using unlabeled, uniformly ^{15}N-labeled, and ^{13}C,^{15}N-double-labeled protein, using a standard range of NMR experiments (*13*). Stereospecific assignments for leucine and valine methyls were obtained using a uniformly 10% ^{13}C-labeled sample (*14*). All experiments used the States-TPPI scheme for quadrature detection in indirect dimensions. ^{1}H chemical shifts were referenced directly to TSP, and ^{15}N and ^{13}C shifts were referenced indirectly to the ^{1}H reference using gyromagnetic ratios (*15*). ^{15}N T_1 and T_2 relaxation times and NOEs were measured as described by (*16*), using uniformly ^{15}N-labeled protein.

Titrations of xylopentaose (X_5) into 500 μM protein were made at 313 K using a solution of 25 mM X_5 in the same buffer as described above, using 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.2, 1.6, 2.2, and 10.0 equiv of X_5. Chemical shift changes were ordered by the weighted total shift after addition of 10 equiv ($\delta_N + 2.0 \times \delta_H$: ref *11*). HSQC spectra were measured at each addition. All NMR spectra were processed and measured using FELIX (Accelrys Inc., San Diego, CA).

The assignment of ^{1}H, ^{13}C, and ^{15}N nuclei is almost complete (*13*). Stereospecific assignments were obtained for all resolvable leucine and valine methyls, and for a small number of aromatic Hβ protons. Distance restraints were obtained from 3D isotope-separated NOESY experiments with mixing times of 90 ms. Distances were obtained based on r^{-6} summation (*17*) from cross-peak intensities using five distance ranges (strongest, strong, medium, weak, and weakest, corresponding to upper bound distance ranges of 2.8, 3.3, 3.8, 4.2, and 5.5 Å, respectively). Additional restraints were obtained from HNHA experiments. In initial rounds of the calculation, the only restraints used were 2184 unambiguous NOEs (which in the initial calculations included symmetry-related and redundant NOEs), supplemented by 83 φ restraints from HNHA spectra, plus 162 backbone dihedral restraints based on backbone and ^{13}Cβ chemical shifts using the program TALOS (*18*). TALOS restraints were included with a range of ±2 standard deviations from the predicted dihedral angle. Three positive φ angles were identified for residues Asn7, Phe110, and Glu118 from consideration of their $J_{HNH\alpha}$ couplings and sequential NOEs (*19*). Of the 10 proline residues, only Pro77 is preceded by a cis amide bond, as evidenced by sequential NOEs and the chemical shift of the Pro77Cβ. At a later stage in the calculation, further side chain dihedral restraints, ambiguous restraints (*20*), and hydrogen bond restraints were added (the latter were included only when both the amide exchange rates with D_2O were low and the temperature coefficients were small: ref *21*). Useful restraints were only found for residues 1−168, which is the range of residues used in the structure calculation. Structures were calculated by hybrid distance geometry/simulated annealing in XPLOR, as described (*11*, *22*).

RESULTS

Structure of CBM4-2. (A) Structure Determination. The family of CBM4-2 structures was calculated using a hybrid distance geometry/simulated annealing protocol following methods similar to those used by Sorimachi et al. (*22*) and Simpson et al. (*11*). The final set of restraints contained 1654 nonredundant unambiguous NOEs and 17 ambiguous NOEs, 93 φ, 72 χ_1, and 1 χ_2 restraints, and 65 pairs of hydrogen bond restraints, plus 177 backbone dihedral restraints based on chemical shifts from TALOS. Almost all hydrogen bond restraints were limited to those expected from the regular secondary structure, except for Ala4−Ile164 and Asn79−Ala4 which contribute to tertiary structure, together with Trp28−Val25, Val36−Gly33, and Trp69−Asn67 which are involved in β-turns. The final restraint set had 12.8 restraints per residue. The distribution of NOE restraints is shown in Figure 1.

In the final calculation, 50 structures were calculated starting from random coordinates and were subsequently refined. The resulting structures were ranked by a combination of total XPLOR energy and NOE violations. The first 12 models had a very similar energy and showed no NOE violations greater than 0.2 Å and no angle violations greater than 5°. Hence, these models were selected to represent the solution structure of CBM4-2. The average structure was calculated from these 12 structures (by selecting the lowest energy structure and superimposing the remaining 11 on this one) and was subsequently subjected to restrained energy minimization to yield the minimized average structure. The 12 best structures and the average structure have been deposited with the Research Collaboratory for Structural Bioinformatics Protein Data Bank [*23*; PDB ID 1k42 (ensemble) and 1k45 (minimized average)].

Structural statistics for the ensemble are given in Table 1, and the ensemble together with a representation of the average structure are shown in Figure 2. When all residues between the *N*- and *C*-termini were used for superimposition, the atomic root-mean-square deviations for backbone and all heavy atoms are 0.42 ± 0.11 and 0.84 ± 0.14 Å, respectively. The atomic rmsd for backbone and all heavy atoms when superimposed by regions of secondary structure are 0.30 ± 0.11 and 0.82 ± 0.17 Å respectively.

The backbone and heavy-atom rmsd values from the mean structure together with the angular order parameters for the

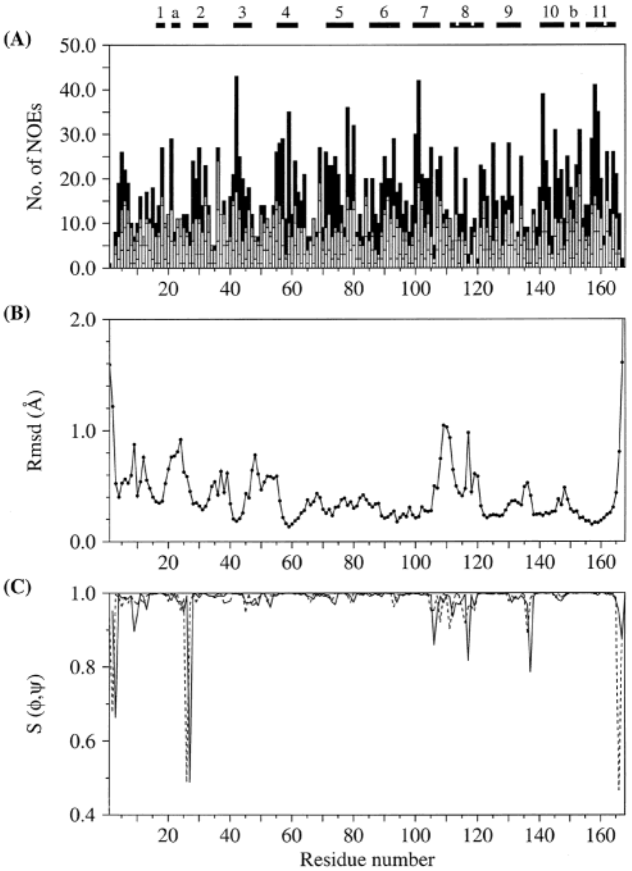

FIGURE 1: Structural parameters for CBM4-2. (A) Distribution of NOE restraints by residue. From bottom to top, intraresidue (white), sequential (light gray), medium-range (dark gray), and long-range (black) NOEs, respectively. (B) Root-mean-square differences from mean structure for backbone atoms. (C) Angular order parameters of backbone φ (solid) and ψ (dashed) angles. The bars along the top represent the regions of secondary structure. Bars 1−11 represent the β-strands while bars a and b are the positions of the helical twists. Thin regions in strands β8 and β11 indicate β-bulges.

φ and ψ dihedral angles are plotted in Figure 1B,C, respectively. Compared to the strands, the loop regions contain fewer restraints and are generally less well-defined. Interestingly, the aromatic residue Phe110, which is implicated in polysaccharide binding, is located within a less well-defined loop region (Gln108−Phe110). The first portion of the following strand (strand 8; Gln111−Ile121) is also generally less well-defined than other strands. This strand also contains two β-bulge regions (residues Tyr113 and Glu118). There is a third β-bulge in strand 11 at residue 161.

The quality of the structures generated has been evaluated using PROCHECK-NMR (*24*) and WHAT-CHECK (*25*). The structures display good stereochemical parameters with, for example, 99% of the residues having backbone dihedral angles in the most favored or additionally allowed regions of the Ramachandran plot (Table 1). One residue was found from WHAT-CHECK to have unusual dihedral distribution and packing parameters, namely, Phe110. This suggests that detailed analysis of the orientation of Phe110 is not warranted. However, Phe110 clearly does have an unusual conformation (e.g., positive φ angle), which is very likely to be related to its functional role in ligand binding, as described below.

(B) Description of the Structure. The structure of CBM4-2 consists of both a five-stranded β-sheet and a six-stranded β-sheet that form a twisted β-sandwich motif about an extensive hydrophobic core. The 11 β-strands are all arranged in an antiparallel fashion and comprise residues 16−18, 28−32, 41−46, 55−61, 71−79, 85−94, 99−107, 111−121, 126−133, 140−147, and 155−164. The structure also contains two short helical turns from residues 21−23 and 150−152. Most of the structure is well-defined (Figure 2A), except for the two termini.

Table 1: Structural Parameters for CBM4-2

	⟨CBM4-2⟩[a]	CBM4-2 avg min[b]
rmsd from experimental restraints		
distance restraints, 1801 (Å)[c]	0.036 ± 0.0007	0.028
dihedral restraints, 343 (deg)[d]	0.45 ± 0.048	0.38
rmsd from idealized covalent geometry		
bonds (Å)	0.003 ± 0	0
angles (deg)	0.67 ± 0.006	0.63
impropers (deg)	0.37 ± 0.008	0.34
XPLOR energies (kJ mol^{-1})		
E_{total}	509.0 ± 13.4	412.8
E_{repel}	26.3 ± 1.9	23.3
E_{NOE}	119.3 ± 5.0	70
E_{cdih}	4.2 ± 1.0	3
E_{bond}	24.6 ± 0.8	18.4
E_{angle}	306.5 ± 5.9	274.6
E_{improper}	28.0 ± 1.2	23.5
Ramachandran analysis		
most favored region (%)	76.7	75.2
additionally allowed regions (%)	22.1	24.8
generously allowed regions (%)	1.2	0
disallowed regions (%)	0.1	0

[a] Family of 12 best structures selected from 50 calculated. [b] Minimized average structure from the family of 12. [c] 510 intraresidue, 426 sequential, 115 medium range ($1 < i - j < 5$), and 603 long range, plus 17 ambiguous long-range and 57 pairs of hydrogen bond restraints. [d] 93 φ restraints from HNHA spectra, 179 restraints on φ or ψ from TALOS, 72 χ_1, and one χ_2 restraint.

The majority of the aromatic side chains are buried within the hydrophobic core of the protein. However, the side chains of Trp69 and Phe110 are situated on the exterior of CBM4-2. These two residues are located at the top of a groove on the surface of the protein, in positions that are unusually exposed for such large and hydrophobic side chains. This groove is formed by the β-sheet comprising strands β2−β5, β7, and β8, together with the loops between strands β2 and β3, β4 and β5, and β7 and β8 (Figure 2). Trp69 is located at the apex of the loop between strands β4 and β5, while Phe110 is located at the apex of the loop between strands β7 and β8.

The mobility of the backbone has been assessed by measuring ^{15}N T_1 and T_2 times for backbone nitrogens. The results are shown in Figure 3, which demonstrates that the structure has very little internal mobility on a subnanosecond time scale over essentially the whole length of the protein except the *C*-terminus and the loop between strands β3 and β4.

NMR titrations with Ca^{2+} indicated that CBM4-2 contained two bound calcium ions, which are bound at independent sites with very different affinities. The calcium ions were not used as distance restraints in the structure calculation, and their location is discussed in the accompanying paper (*26*).

Identification of the Polysaccharide Binding Site. To investigate the binding of xylo-oligosaccharides to CBM4-2, the

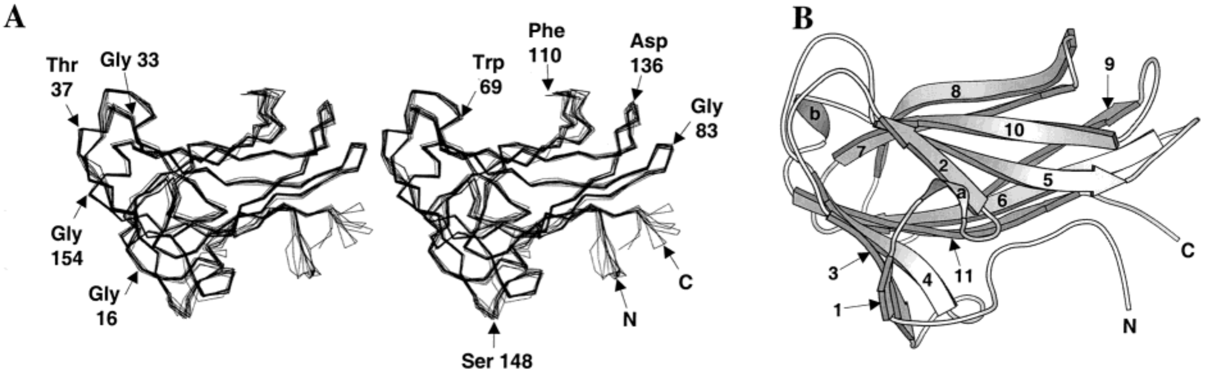

FIGURE 2: Backbone structure of CBM4-2. (A) Superposition of the best 12 CBM4-2 structures, selected from the 50 calculated, shown as a stereo diagram. The superposition used the backbone atoms (N, Cα, C′) of residues in regular secondary structure elements (residues 16−18, 21−23, 28−32, 41−46, 55−61, 71−79, 85−94, 99−107, 111−121, 126−133, 140−147, 150−152, 155−164). A number of residues are numbered, including Trp69 and Phe110, which line the groove at the top of the molecule and form the key residues interacting with xylan. (B) MOLSCRIPT (*43*) representation of CBM4-2, based on the minimized average structure, in a similar orientation to panel A.

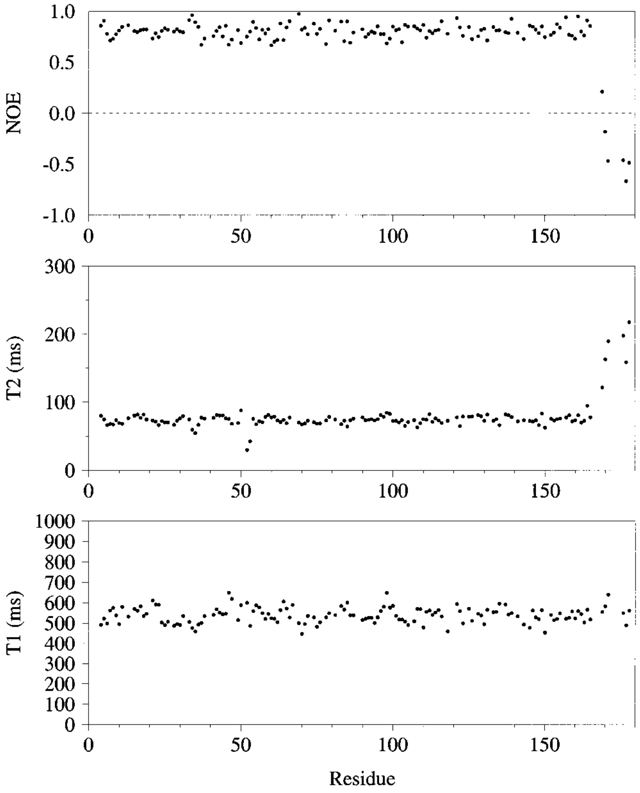

FIGURE 3: ^{15}N T_1 and T_2 relaxation times and NOEs for CBM4-2, by residue.

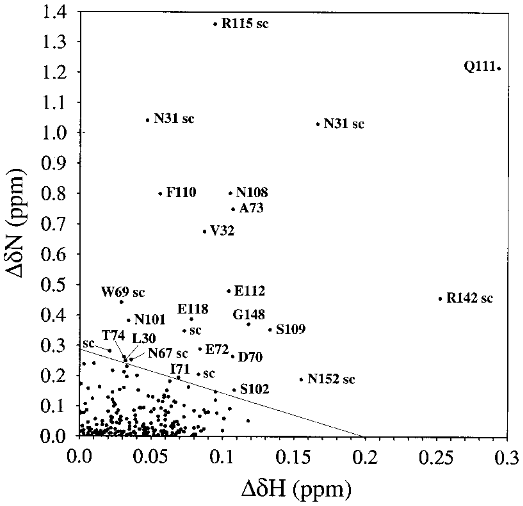

FIGURE 4: Chemical shift changes on addition of an excess of xylopentaose to CBM4-2. Resonances are labeled by the amino acid type and number, with sc indicating side chain resonances. The diagonal line shows the cutoff used to distinguish ‘large’ from ‘small’ shift changes.

protein was titrated with xylopentaose (X_5), and two-dimensional ^{1}H-^{15}N HSQC (heteronuclear single quantum coherence) NMR spectra were recorded following each addition. As noted previously (*12*), the spectral changes indicate a 1:1 binding stoichiometry. The majority (85%) of amide groups had ^{1}H and ^{15}N shifts that are not significantly affected by the binding of X_5. However, a number of backbone and side chain amide groups were observed to undergo large chemical shift changes upon the addition of the oligosaccharide, namely (in order, with the largest shift changes first), Gln111, Arg115 H^ϵ, Asn31 H^δ, Asn108, Ala73, Arg142 H^ϵ, Phe110, Val32, Glu112, Ser109, Gly148, Glu118, Trp69 H^ϵ, Asn152 H^δ, Asp70, Glu72, Asn101, Ser102, Ile71, Asn67 H^δ, Thr74, and Leu30 (Figure 4). These residues are shown shaded in Figure 5. All of the resonances that undergo large chemical shift changes on binding of X_5 are in and around the cleft in which the two solvent-exposed aromatic rings, Trp69 and Phe110, are located. This confirms this region as the xylan binding site. Exposed aromatic rings have previously been shown to be the main ligand binding residues in a range of CBMs (*27, 28*) including the only CBM4s previously determined (*8, 29*).

As in other known CBMs, the two aromatic rings in the binding site cleft are surrounded by a number of hydrophilic and charged residues, in particular Asn31, Glu72, Arg115, and His146. These residues are highlighted in Figure 6, and are all orientated such that their side chains are surface-exposed and able to form hydrogen bonds with a polysaccharide located within the binding cleft. Correspondingly, the side chain amide of Asn31, the guanidino group of Arg115, and the backbone amide of Glu72 display large chemical shift perturbations on ligand binding (Figure 4). Therefore, these residues are likely to be involved in xylan binding through hydrogen bonding. No titration data are available on the side chain of His146 as no signals from the

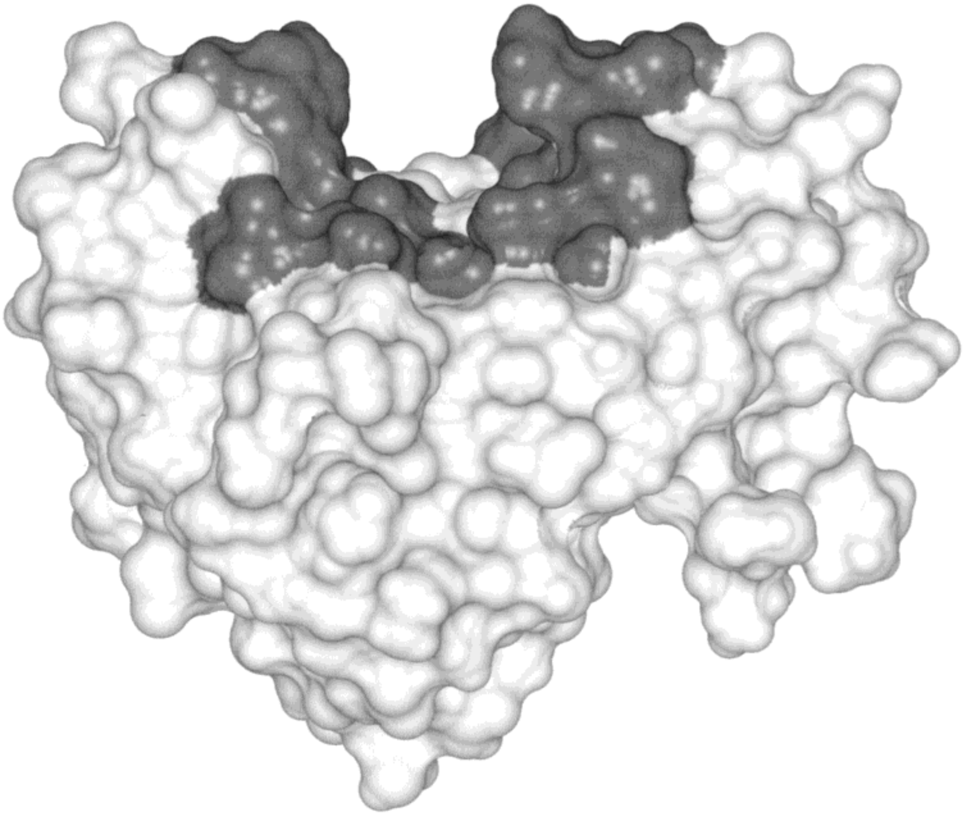

FIGURE 5: Surface of CBM4-2, shaded according to weighted ^{1}H and ^{15}N chemical shift changes upon addition of xylopentaose. Regions of large shift change (using the cutoff shown in Figure 4) are shaded in gray. White indicates small or no shift change.

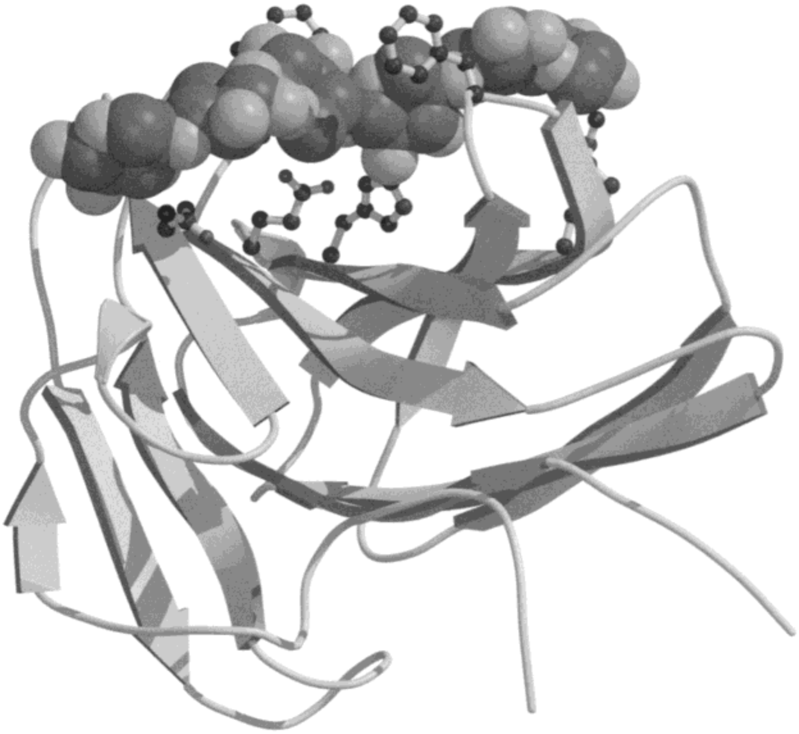

FIGURE 6: Structure of CBM4-2, showing the position of xylohexaose modeled into the binding site based on chemical shift changes on titration. Also shown are some of the side chains implicated in the binding. From left to right across the top of the figure, these are Asn31, Trp69, Glu72, His146, Phe110, and Arg115.

side chain have been identified in the ^{1}H-^{15}N HSQC NMR spectra. The backbone amide of His146 displayed a chemical shift change of 0.173 ppm in ^{15}N and 0.001 ppm in ^{1}H, a perturbation classified below the cutoff point for a shift significantly affected by X_5 binding.

Based on the structure presented, the reported 3-fold helical structure of xylan (*30*), and the X_5 titration data, we have constructed a model of protein−polysaccharide binding (Figure 6). The xylan sits at the bottom of the binding cleft, which permits all the residues showing very large chemical shift changes, and therefore implicated in hydrogen-bonding interactions with the substrate, to be close to the polysaccharide. It is noteworthy that only a limited area of the oligosaccharide is exposed to the solvent, which places important limitations on permissible side chain substitutions of the xylan. From the model, the xylose in the center of the site cannot be substituted at either O2 or O3, while two other xyloses can accommodate only limited substitution: the following xylose (toward the reducing end) can accommodate only very small substituents on O2 without steric contacts with Trp69, while the xylose another 2 units further can be substituted at O3 only following rotation around the glycosidic bond. There is therefore some steric restriction on xylose substitutions over a 4 unit stretch. This is compatible with the finding that the module can bind substituted xylans such as wheat arabinoxylan (*12*), provided that there is a sufficient density of unsubstituted sites, as has been shown, for example, in barley arabinoxylan (*31*). It should be noted that sterically there is little to distinguish the two possible orientations of the xylan chain, and it is possible to model the chain into the cleft in the other orientation. However, the alternative orientation carries similar steric restrictions on side chain substitution. Similar observations have previously been made for the CenC CBM4, which binds cello-oligosaccharides with approximately equal affinity in both orientations (*32*). It has been suggested that there is a functional advantage for CenC CBM4 to bind the substrate in both orientations (*32*).

The length of the cleft in the model is approximately 23 Å, similar to the length of xylopentaose, which is approximately 22 Å. The protein is therefore expected to bind almost equally strongly to the pentamer as it does to the hexamer (*12*). Analysis of line shapes and chemical shift changes on titration indicates that the affinities of xylopentaose and xylohexaose for CBM4-2 are approximately equal (10 and 5 μM, respectively), in agreement with this result.

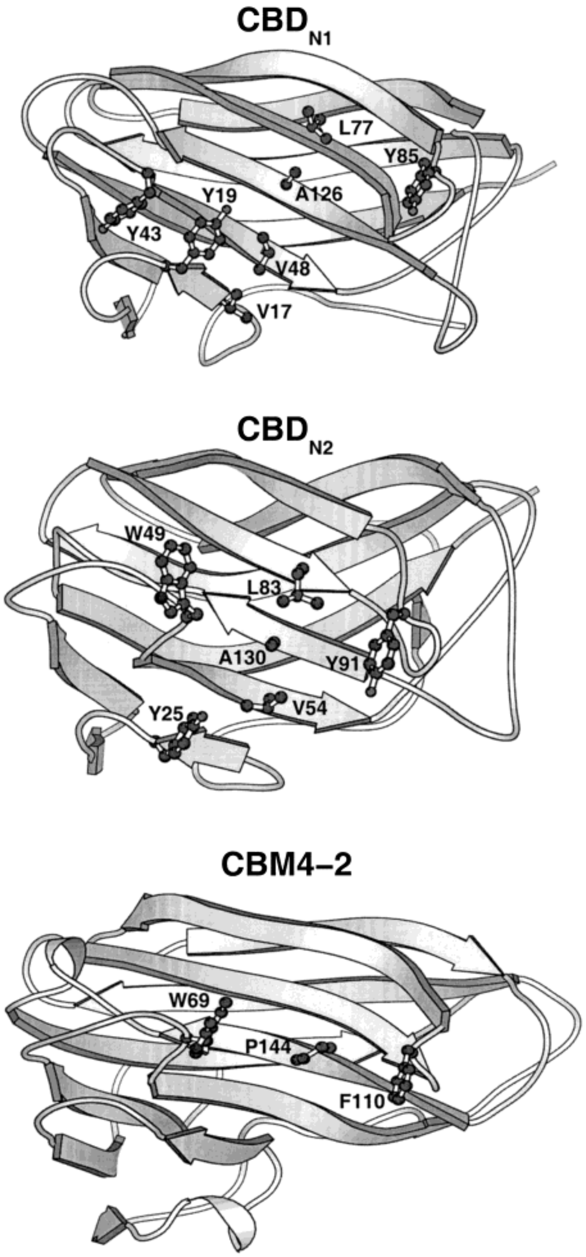

FIGURE 7: Binding site cleft of CBD$_{N1}$, CBD$_{N2}$, and CBM4-2, viewed from above. Aromatic and hydrophobic residues implicated in binding are drawn as ball-and-stick. Note the small number of hydrophobic residues in CBM4-2 compared to CBD$_{N1}$ and CBD$_{N2}$.

DISCUSSION

Comparison with Other CBMs. (A) Comparison with CBM4. On the basis of sequence similarities, the CBM studied here was originally suggested to form part of family 4 (*9, 33*). Functional characterization emphasized its relatedness to family 4 CBMs, but it also revealed some differences such as the affinity for xylans, which led us to suggest the division of family 4 into subfamilies (*12*). This suggestion has been confirmed by Sunna et al. (*6*) and by Zverlov et al. (*7*), who have classified CBM4-2 into subfamily CBM4C. The structure determined for CBM4-2 confirms it as belonging to family 4. There are two family 4 CBMs determined to date, namely, the two *N*-terminal domains from *Cellulomonas fimi* endoglucanase C, which have been abbreviated to CBD$_{N1}$ and CBD$_{N2}$ (*8, 29*). Both these domains bind to amorphous cellulose but not to xylan. The sequence similarity between CBM4-2 and CBD$_{N1}$ and CBD$_{N2}$ is 20% in both cases after alignment by secondary structure, and the backbone rmsd between corresponding regions of regular secondary structure is 2.09 and 2.73 Å, respectively. CBM4-2 is therefore reasonably similar to both domains. The β-sheet regions are essentially identical for all three proteins, although a short *N*-terminal β-strand is absent in CBD$_{N1}$. There are three β-bulges in CBM4-2, at positions 113, 118, and 161. The bulges at 113 and 161 match bulges at the equivalent locations in CBD$_{N1}$ (88 and 142, respectively), while the bulges at 113 and 118 match bulges at the equivalent locations in CBD$_{N2}$ (94 and 99, respectively). Interestingly, although many CBM4 domains contain a disulfide bridge that is reported to be essential for stability (*8*), CBM4-2 has no cysteine residues. CBD$_{N1}$ (but not CBD$_{N2}$) binds calcium, at a similar position to the high-affinity site of CBM4-2 (*26*).

On comparison of the ligand binding sites of CBD$_{N1}$ and CBD$_{N2}$ with CBM4-2, some interesting differences emerge. CBD$_{N1}$ has three aromatic rings close to the binding cleft (*8*), although only two are suggested to be involved in binding, namely, Tyr19 and Tyr85 (Figure 7) (*34, 35*). CBD$_{N2}$ also has three (*29*): Tyr25 and Tyr91 are in equivalent positions to the CBD$_{N1}$ residues, and there is a third aromatic residue, Trp49, which matches Tyr43 in

CBD_{N1}. There is no published evidence on whether Trp49 is involved in binding. The two residues identified in this work as the liganding residues, Trp69 and Phe110, are in equivalent positions to Tyr43 and Tyr85 in CBD_{N1} (Trp49 and Tyr91 in CBD_{N2}), although their orientations are different, which would be consistent with an altered substrate specificity. Thus, although the rings in CBM4-2 correspond to two of the rings in the CBDs, they are not the two identified as most important.

The cleft in CBD_{N1} and CBD_{N2} has a number of hydrophobic residues forming a strip along the bottom (Figure 7). In CBM4-2, the cleft is much less hydrophobic, with the only hydrophobic residue being Pro144. This fits with the structures of the ligands: cellulose has a flat hydrophobic surface, while xylan is helical and does not present the same hydrophobic surface. We conclude that the different ligand specificities of CBM4 (xylan) and CBD_{N1}/CBD_{N2} (cellulose) are due to a combination of the position and orientation of the aromatic ligands, and the nature of the residues lining the cleft, as discussed further below.

(B) Comparison with CBM22. It has been suggested that CBM4 forms part of a superfamily containing CBM families 4, 16, 17, 22, and 27 (*6*). A three-dimensional alignment of CBM4-2 with CBM22 (*36*) was therefore performed using the program DALI (*37*). The two structures have essentially identical topology, and their secondary structures overlay well, with insertions and deletions limited to loops. The β-sheet residues overlay with a backbone rmsd of 2.87 Å (i.e., residues 25−32, 41−46, 55−61, 71−77, 85−94, 99−107, 111−112, 114−121, 126−130, 131−133, 140−147, and 155−164 of CBM4-2 with residues 21−28, 31−36, 44−50, 56−62, 72−81, 88−96, 105−106, 107−114, 119−123, 125−127, 134−141, and 146−155 of CBM22). The binding groove is located in the same part of both proteins. Furthermore, the CBM4-2 high-affinity calcium site overlays with the calcium site of CBM22, both sites being located within a loop between strands β3 and β4 (*26*). There are a large number of hydrophobic side chains in the core of the proteins, which match well in sequence location. It is therefore clear that the two families CBM4 and CBM22 form part of one larger superfamily.

CBM22 and CBM4-2 both bind to xylan. On comparison of the suggested binding residues, the two aromatics identified here for CBM4-2 and by Charnock et al. (*36*) for CBM22 are in very similar positions. Moreover, the bottom of the cleft is much more hydrophilic than is seen for the CBDs, and looks much more like the CBM4-2 cleft. Following the argument above, this would be consistent with it binding to xylan rather than cellulose. Although aromatic residues are pivotal to ligand binding in CBM22, it is apparent that at least one of the hydrophilic amino acids, Glu138, also plays a critical role in xylan binding as its removal did not disrupt the structure of the protein but destroyed ligand binding (*38*). It is likely that hydrophilic residues in CBM4-2 may also play a critical role in ligand recognition.

Substrate Binding Specificity. The binding site for oligosaccharide on CBM4-2 is similar to that found in all CBMs to date, in that it consists of a rigid β-sheet framework, with exposed aromatic rings that form a crucial part of the site. Moreover, the structure is similar to that of the previously determined CBM4s, in that the binding site is a cleft rather than the exposed surface that is found in CBMs that interact with crystalline cellulose (*27, 39*). This explains how the CBM4 can bind specifically to amorphous cellulose but not to crystalline cellulose, which is sterically unable to enter the cleft. Thus, the binding specificity is determined to a great extent by the shape of the polysaccharide substrate.

The CBM4 family is relatively unusual, in that different members of the family have been found to display specificity for different polysaccharides. CBM4-2 is particularly interesting, in that it binds both to xylan and to amorphous cellulose, albeit with much lower affinity for the latter (*12*). This raises the question of how proteins with similar sequence and three-dimensional shape are able to distinguish between the polysaccharides xylan and amorphous cellulose that are also similar; and how CBM4-2 can bind to both ligands. A similar problem has previously been addressed in family 2 CBMs, in that family 2a binds amorphous and crystalline cellulose while family 2b binds xylan (*11, 40*). The answer in that case is that in family 2a, the key aromatic rings are coplanar (therefore matching the shape of cellulose), while in family 2b they are approximately perpendicular, therefore matching the twisted shape of xylan. Does CBM4-2 do something similar?

In the model (Figure 6), the aromatic rings of Trp69 and Phe110 are close to xylan pyranoside rings in the 3-fold helical structure of xylan, and only need to reorient by a small amount to make good stacking interactions with them. It is clear from Figure 6 that binding of an approximately planar cellulose would require much larger scale motion to make a good binding interaction, and is therefore likely to be less favorable. It therefore appears that binding specificity is achieved by steric complementarity of the binding site to the substrate, in a rather similar way to the CBM2 described above.

The capacity of CBM4-2 to bind, albeit weakly, to cellulose indicates that a conformational change must occur in either the ligand or the protein or possibly both. One could visualize that twisting cellulose out of its planar structure into a conformation that forms stacking interactions with the aromatic amino acids would require a considerable amount of energy, which could explain the low affinity for the ligand. Alternatively, it is possible that the planar cellulose and the perpendicularly orientated aromatic residues form weak hydrophobic interactions requiring no reorientation of the ligand or protein, rather than face-to-face stacking interactions, again providing an explanation for the low affinity. However, it is also plausible that the aromatic residues in the binding site undergo some conformational change to accommodate the binding of cellulose. Indeed, there are three main pieces of evidence which could suggest that the binding site is more flexible than that of most other CBMs studied to date. The first is the unusual positioning of the Trp69 and Phe110 rings, at the tips of highly exposed loops. The second is that all the NOEs involving the Trp69 and Phe110 rings are weak, potentially suggesting a loss in NOE intensity because of conformational averaging. The third is that residues in the whole length of the loops show chemical shift changes on titration with X_5. Thus, changes are seen for the backbone protons of Asn108, Ser109, Gln111, and Glu112 as well as Phe110; and residues Asp70, Ile71, Glu72, Ala73, and Thr74 as well as Trp69. Such extensive chemical shift changes would not be expected unless the loop alters the conformation significantly on binding. A change in conformation is not inconsistent with the ^{15}N relaxation data (Figure 3), which merely show that the protein is rigid on the subnanosecond time scale. The large-scale changes required for reorientation of the two loops would be expected to be slower time scale motions, and would not show up in the T_1/T_2 data. An unusual plasticity of the CBM4 binding site has also been suggested for the CenC CBM4s (*29*). One of the two binding sites of the *Aspergillus niger* glucoamylase starch binding domain has also been shown to have conformational flexibility, which has been proposed to aid substrate recognition (*22*). A recent analysis of the thermodynamics of polysaccharide binding to CBDs (*41*) has concluded that binding to an exposed site (as observed in all CBMs that bind crystalline cellulose) has very different requirements on ligand mobility than binding in a cleft, as found in CBM4s. We therefore propose that a limited plasticity of the polysaccharide binding site is functionally important, when the binding site is within a cleft, and may be expected in other CBMs with clefts. This generalization is not true for the CBM9 family, which have very specific requirements, in that they bind to the reducing end of polysaccharides (*42*).

Any movement of Trp69 and Phe110 to stack better against the xylan rings has the effect of narrowing the entrance to the cleft. It is therefore tempting to suggest that the repositioning of the rings, driven by an improved stacking interaction, also has the effect of hindering release of the saccharide chain and therefore increasing the strength of the binding interaction. This could be the reason for the unusually strong binding interaction of CBM4-2.

ACKNOWLEDGMENT

We thank Jeremy Craven for help with computing, and Andrea Hounslow for NMR support.

REFERENCES

1. Biely, P. (1985) *Trends Biotechnol. 3*, 286−290.
2. Tomme, P., Warren, R. A. J., and Gilkes, N. R. (1995) *Adv. Microb. Physiol. 37*, 1−81.
3. Coutinho, P. M., and Henrissat, B. (2001) Carbohydrate-Binding Module Family server (http://afmb.cnrs-mrs.fr/∼pedro/CAZY/cbm.html).
4. Gill, J., Rixon, J. E., Bolam, D. N., McQueen-Mason, S. J., Simpson, P. J., Williamson, M. P., Hazlewood, G. P., and Gilbert, H. J. (1999) *Biochem. J. 342*, 473−480.
5. Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller, R. C., Jr., and Warren, R. A. J. (1991) *Microb. Rev. 55*, 303−315.
6. Sunna, A., Gibbs, M. D., and Bergquist, P. L. (2001) *Biochem. J. 356*, 791−798.
7. Zverlov, V. V., Volkov, I. Y., Velikodvorskaya, G. A., and Schwarz, W. H. (2001) *Microbiology 147*, 621−629.
8. Johnson, P. E., Joshi, M. D., Tomme, P., Kilburn, D. G., and McIntosh, L. P. (1996) *Biochemistry 35*, 14381−14394.
9. Nordberg Karlsson, E., Bartonek-Roxå, E., and Holst, O. (1997) *Biochim. Biophys. Acta 1353*, 118−124.
10. Alfredsson, G. A., Kristjansson, G. O., Hjörleifsdottir, S., and Stetter, K. O. (1988) *J. Gen. Microbiol. 134*, 299−306.
11. Simpson, P. J., Bolam, D. N., Cooper, A., Ciruela, A., Hazlewood, G. P., Gilbert, H. J., and Williamson, M. P. (1999) *Structure 7*, 853−864.
12. Abou Hachem, M., Nordberg Karlsson, E., Bartonek-Roxå, E., Raghothama, S., Simpson, P. J., Gilbert, H. J., Williamson, M. P., and Holst, O. (2000) *Biochem. J. 345*, 53−60.
13. Jamieson, S. J., Williamson, M. P., Abou Hachem, A., and Simpson, P. J. (2002) *J. Biomol. NMR 22*, 187−188.
14. Neri, D., Szyperski, T., Otting, G., Senn, H., and Wüthrich, K. (1989) *Biochemistry 28*, 7510−7516.
15. Wishart, D. S., Bigam, C. H., Holm, A., Hodges, R. S., and Sykes, B. D. (1995) *J. Biomol. NMR 5*, 67−81.
16. Kördel, J., Skelton, N. J., Akke, M., Palmer, A. G., III, and Chazin, W. J. (1992) *Biochemistry 31*, 4856−4866.
17. Fletcher, C. M., Jones, D. N. M., Diamond, R., and Neuhaus, D. (1996) *J. Biomol. NMR 8*, 292−310.
18. Cornilescu, G., Delaglio, F., and Bax, A. (1999) *J. Biomol. NMR 13*, 289−302.
19. Ludvigsen, S., and Poulsen, F. M. (1992) *J. Biomol. NMR 2*, 227−233.
20. Nilges, M. (1995) *J. Mol. Biol. 245*, 645−660.
21. Baxter, N. J., and Williamson, M. P. (1997) *J. Biomol. NMR, 9*, 359−369.
22. Sorimachi, K., Le Gal-Coëffet, M.-F., Williamson, G., Archer, D. B., and Williamson, M. P. (1997) *Structure 5*, 647−661.
23. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) *Nucleic Acids Res. 28*, 235−242 (http://www.rcsb.org/pdb/).
24. Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) *J. Biomol. NMR 8*, 477−486.
25. Hooft, R. W. W., Vriend, G., Sander, C., and Abola, E. E. (1996) *Nature 381*, 272.
26. Abou Hachem, M., Nordberg Karlsson, E., Simpson, P. J., Linse, S., Sellers, P., Williamson, M. P., Jamieson, S. J., Gilbert, H. J., Bolam, D. N., and Holst, O. (2002) *Biochemistry 41*, 5720−5729.
27. Boraston, A. B., McLean, B. W., Kormos, J. M., Alam, M., Gilkes, N. R., Haynes, C. A., Tomme, P., Kilburn, D. G., and Warren, R. A. J. (1999) in *Recent Advances in Carbohydrate Engineering* (Gilbert, H. J., Davies, G. J., Svensson, B., and Henrissat, B., Eds.) pp 202−211, Royal Society of Chemistry, Cambridge, U.K.
28. Tormo, J., Lamed, R., Chirino, A. J., Morag, E., Bayer, E. A., Shoham, Y., and Steitz, T. A. (1996) *EMBO J. 15*, 5739−5751.
29. Brun, E., Johnson, P. E., Creagh, A. L., Tomme, P., Webster, P., Haynes, C. A., and McIntosh, L. P. (2000) *Biochemistry 39*, 2445−2458.
30. Nieduszynski, I. A., and Marchessault, R. H. (1972) *Biopolymers 11*, 1335−1344.
31. Izydorczyk, M. S., Macri, L. J., and MacGregor, A. W. (1998) *Carbohydr. Polym. 35*, 259−269.
32. Johnson, P. E., Brun, E., MacKenzie, L. F., Withers, S. G., and McIntosh, L. P. (1999) *J. Mol. Biol. 287*, 609−625.
33. Nordberg Karlsson, E., Bartonek-Roxå, E., and Holst, O. (1998) *FEMS Microbiol. Lett. 168*, 1−7.
34. Johnson, P. E., Tomme, P., Joshi, M. D., and McIntosh, L. P. (1996) *Biochemistry 35*, 13895−13906.
35. Kormos, J., Johnson, P. E., Brun, E., Tomme, P., McIntosh, L. P., Haynes, C. A., and Kilburn, D. G. (2000) *Biochemistry 39*, 8844−8852.
36. Charnock, S. J., Bolam, D. N., Turkenburg, J. P., Gilbert, H. J., Ferreira, L. M. A., Davies, G. J., and Fontes, C. M. G. A. (2000) *Biochemistry 39*, 5013−5021.
37. Holm, L., and Sander, C. (1993) *J. Mol. Biol. 233*, 123−138.
38. Xie, H., Gilbert, H. J., Charnock, S. J., Davies, G. J., Williamson, M. P., Simpson, P. J., Raghothama, S., Fontes, C. M. G. A., Dias, F. M. V., Ferreira, L. M. A., and Bolam, D. N. (2001) *Biochemistry 40*, 9167−9176.
39. 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.
40. Simpson, P. J., Xie, H., Bolam, D. N., Gilbert, H. J., and Williamson, M. P. (2000) *J. Biol. Chem. 275*, 41137−41152.
41. Xie, H., Bolam, D. N., Nagy, T., Szabó, L., Cooper, A., Simpson, P. J., Lakey, J. H., Williamson, M. P., and Gilbert, H. J. (2001) *Biochemistry 40*, 5700−5707.
42. Notenboom, V., Boraston, A. B., Kilburn, D. G., and Rose, D. R. (2001) *Biochemistry 40*, 6248−6256.
43. Kraulis, P. J. (1991) *J. Appl. Crystallogr. 24*, 946−950.

BI012093I