Structure of the *Bacillus agaradherans* Family 5 Endoglucanase at 1.6 Å and Its Cellobiose Complex at 2.0 Å Resolution[†,‡]

Gideon J. Davies,*[,§] Miroslawa Dauter,[§] A. Marek Brzozowski,[§] Mads Eskelund Bjørnvad,[||] Kim V. Andersen,[||] and Martin Schülein[||]

Department of Chemistry, University of York, Heslington, York YO1 5DD, England, and Novo-Nordisk a/s, Novo Allé, DK-2880 Bagsvaerd, Denmark

Received August 29, 1997; Revised Manuscript Received November 24, 1997

ABSTRACT: The enzymatic degradation of cellulose, by cellulases, is not only industrially important in the food, paper, and textile industries but also a potentially useful method for the environmentally friendly recycling of municipal waste. An understanding of the structural and mechanistic requirements for the hydrolysis of the β-1,4 glycosidic bonds of cellulose is an essential prerequisite for beneficial engineering of cellulases for these processes. Cellulases have been classified into 13 of the 62 glycoside hydrolase families [Henrissat, B., and Bairoch, A. (1996) *Biochem J. 316,* 695−696]. The structure of the catalytic core of the family 5 endoglucanase, Cel5A, from the alkalophilic *Bacillus agaradherans* has been solved by multiple isomorphous replacement at 1.6 Å resolution. Cel5A has the $(\alpha/\beta)_8$ barrel structure and signature structural features typical of the grouping of glycoside hydrolase families known as clan GH-A, with the catalytic acid/base Glu 139 and nucleophile Glu 228 on barrel strands β4 and β7 as expected. In addition to the native enzyme, the 2.0 Å resolution structure of the cellobiose-bound form of the enzyme has also been determined. Cellobiose binds preferentially in the −2 and −3 subsites of the enzyme. Kinetic studies on the isolated catalytic core domain of Cel5A, using a series of reduced cellodextrins as substrates, suggest approximately five to six binding sites, consistent with the shape and size of the cleft observed by crystallography.

Cellulases are the enzymes responsible for the complete hydrolysis of the β-1,4 glycosidic bonds of cellulose. Cellulose is widely considered to be the most abundant natural polymer in the biosphere. It is also present in textiles, paper, and the raw materials for processed foods. This gives an understanding of the structural and mechanistic features of the cellulolytic enzymes' particular importance. Since the First Oil Crisis, the use of cellulases in the environmentally-friendly processing of municipal waste (the majority of which is cellulose-based) into fermentable sugars has frequently been proposed. Current applications of cellulases, however, are limited by a lack of detailed understanding of the structure, stability, and activity of these polysaccharide hydrolases.

Glycoside hydrolases have been classified into over 62 families on the basis of amino acid sequence similarities (*1*−*3*), with cellulases being found in 13 of these (families 5−10, 12, 26, 44, 45, 48, 60, and 61). Three-dimensional cellulase structures are known for representatives of families 5−10, 12, 45, and 48 (for review see refs *4* and *5*). Cellulases from most organisms are modular, with a catalytic core domain linked to one or more nonhydrolytic domains (*6*). These additional domains are responsible for polymeric substrate binding and thus modulate the action of the enzyme on insoluble substrates such as intact crystalline cellulose. In some anaerobic organisms the organization is more elaborate with several catalytic, cellulose-binding, and docking domains organized as a supermolecular complex known as the cellulosome (*7*). Cellulolytic microorganisms thus overcome the difficult task of cellulose breakdown by producing batteries of different catalytic and noncatalytic domains that act in synergistic harmony on the substrate and, thus, facilitate the complete breakdown of this otherwise refractory polymer (for review see ref *8*).

In this paper, we describe the structure of the catalytic core domain of the family 5 endoglucanase (hereafter "Cel5A"), from *Bacillus agaradherans*. Family 5 contains a very large number of members that possess a very low sequence identity (just eight residues are invariant (*9*, *10*)). For this reason, a classification into more closely related subfamilies has been proposed (*9*). The *B. agaradherans* Cel5A belongs to subfamily 5-2 for which no structures have yet been determined, but one of whose members, the EGZ from *Erwinia chrysanthemi*, has received extensive biochemical characterization (*11*). The native *B. agaradherans* Cel5A structure has been determined by multiple isomorphous replacement utilizing Hg and U derivatives and refined

[†] This work was funded in part, by the Biotechnology and Biological Sciences Research Council, Novo-Nordisk a/s and the European Union (contract no. BIO4-CT97-2303). G.J.D. is a Royal Society University Research Fellow.

[‡] Coordinates for the structure described in this paper have been deposited with the Brookhaven Protein Data Bank (accession references 1A3H and 2A3H).

* To whom correspondence should be addressed. Tel.: 44-1904-432596. Fax: 44-1904-410519. E-mail: davies@yorvic.york.ac.uk.

[§] University of York.

[||] Novo Nordisk a/s.

[1] Abbreviations: Cel5A, family 5 endoglucanase; GH-A, glycoside hydrolase clan A; DP, degree of polymerization.

S0006-2960(97)02162-4 CCC: $15.00
Published on Web 01/30/1998

at a resolution of 1.6 Å. Kinetic determinations on a series of reduced cellodextrins suggest that Cel5A has approximately five to six binding subsites, consistent with the shape and size of the substrate-binding surface observed in the 3-D structure. Complex crystals have been obtained by soaking with the disaccharide cellobiose and the structure elucidated at 2.0 Å. A well-ordered cellobiose molecule binds in the −2 and −3 subsites of the enzyme and allows a complete mapping of the protein−saccharide interactions in this region.

MATERIALS AND METHODS

Purification and Characterization of Cel5A. The 44 kDa endoglucanase produced by *B. agaradherans* strain AC13 (NCIMB 40482) was purified and the amino acid sequence for the first 19 *N*-terminal residues determined by conventional means. A degenerate PCR primer corresponding to the DNA encoding the *N*-terminal amino acids present was synthesized and used to isolate the gene encoding the *B. agaradherans* Cel5A. The gene was then subcloned into the expression vector pMOL995 under transcriptional control of the Thermamyl-amylase promoter and signal peptide. The cellulase negative *Bacillus subtilus* strain (PL2306) was transformed with this plasmid and incubated in shaker flasks using TY as medium. After 18 h incubation the fermentation broth was purified by filtration and concentrated by ultrafiltration. The concentrate was adjusted to pH 7.0 and the Cel5A purified using Avicel affinity chromatography. The naturally-hydrolyzed catalytic core domain elutes in the runoff since it contains no cellulose-binding domain. The core enzyme was purified using anion-exchange chromatography on an HPQ column in 50 mM sodium phosphate buffer at pH 7.0. The pure catalytic core domain has a relative molecular mass of 38 000.

Kinetics on reduced cellodextrins from DP3 to DP6 were performed as outlined by Schülein and co-workers (*12*) based on a linked assay with cellobiose dehydrogenase and the concomitant reduction of a colored substrate. Determinations were made at both pH 7.5 and 9.0, with the reducing substrate 2,6-dichloroindophenol at pH 7.5 and cytochrome c at pH 9.0. Data for steady-state kinetic determinations were analyzed using GRAFIT (Leatherbarrow, R. J. Erihacus Software, Ltd.).

Crystallization, Data Collection, and Phasing. B. agaradherans Cel5A catalytic core domain was desalted and concentrated to 20 mg mL^{-1} in distilled water. The protein was crystallized by the hanging-drop method using 0.8−1.2 M ammonium sulfate as both precipitant and buffer (pH 4.5) in the presence of 15% (v/v) glycerol. Crystals were mounted in a rayon fiber loop and placed in a boiling nitrogen stream at 120 K. A cryoprotectant solution was made consisting of 1.2 M ammonium sulfate at pH 5.5, with the addition of glycerol to a final concentration of 25% (v/v). Data were collected using an MAR Research image plate system together with a Cu rotating anode and utilizing long, focusing, mirror optics (Yale/Molecular Structure Corporation). A total of 135° of native data was collected for both the native and derivative crystals to ensure complete data coverage and a high multiplicity of observation for the derivative anomalous measurements. Data were processed and reduced using the DENZO/SCALEPACK programs (*13*). All further calculations used the CCP4 suite of programs unless otherwise stated (*14*). Heavy-atom derivatives were prepared by presoaking the crystals for 12 h in solutions of either 1 mM methyl mercury chloride (MeHgCl) or 10 mM uranyl acetate (UO_2Ac_2). The MeHgCl derivative crystal was "back-soaked" for 20 min in native mother liquor, prior to data collection. Initial heavy-metal positions, for two uranyl sites, were found by manual inspection of the UO_2Ac_2 difference Patterson calculated between 10 and 2.5 Å resolution. Phases were calculated using the MLPHARE program. Further uranium sites and a single MeHgCl site were found by "cross-phase" difference Fourier analysis. The resultant MIR map was improved with cycles of density modification, utilizing histogram matching, solvent flattening, and Sayre's equation, using the DM program (*15*).

Model Building and Refinement. The model was built into the DM modified MIR map with the O program (*16*). Five percent of the observations were then set aside for cross validation analysis (*17*) and were used to monitor various refinement strategies such as geometric and temperature-factor restraint values and the insertion of solvent water and as the basis for the maximum likelihood refinement using the REFMAC program (*18*). As all observed data from 15 Å resolution were employed in the refinement, a low-resolution bulk solvent correction was applied. Since the scattering from the Cel5A crystals is highly anisotropic, particular use was made of the anisotropic F_{obs} vs F_{calc} scaling options in REFMAC (*18*). Cycles of maximum-likelihood-based least-squares refinement were interspersed with rebuilding using O. Water molecules were added in an automated manner using ARP (*19*) and inspected manually prior to coordinate deposition. Coordinates for the protein structures described in this paper have been deposited with the Brookhaven Protein Databank (*20*).

Cellobiose Complex. Native crystals of Cel5A were soaked in a stabilizing mother liquor with the addition of 30 mM β-D-cellobiose for 12 h. Synchrotron data were collected on an EMBL Hamburg beamline X-31 to 2.0 Å resolution, also at 120 K, and the data processed and reduced in the same manner as for the native enzyme. To ensure correct cross-validation, the same "free" subset of reflections was maintained for refinement of the cellobiose complex and refinement was performed as for the native enzyme structure. Electron density maps to identify the bound saccharide were calculated prior to the incorporation of such a species in the refinement. Stereochemical dictionaries for refinement of the cellobiose moiety were calculated on the basis of the structure of cellotetraose as a model (*21*).

Structure Comparison. Multiple structure alignments, and structure-based sequence alignment, were performed with the MODELER (INSIGHT II, 95.5, Molecular Simulations, Inc. USA) program using a dynamic programming algorithm. Cα positions were considered "equivalent" if closer than 3.5 Å. The other family 5 coordinate sets used were the endoglucanase CCA from *Clostridium cellulolyticum*, (*22*), endocellulase E1 from *Acidothermus cellulolyticus* (*10*), and the endoglucanase A from *Clostridium thermocellum* (*23, 24*).

RESULTS AND DISCUSSION

MIR Structure Determination. The native Cel5A data consist of 222 447 observations of 40791 unique reflections with 120 observations rejected during the data reduction

Table 1: Heavy Atom Phasing Statistics for the *B. agaradherans* Cel5A Structure Determination (Statistics Are Given for All Observed Data between 15 and 2.3 Å Resolution)

derivative	no. of sites	*X*	*Y*	*Z*	*B* (Å^2)	R_{deriv}[a]	R_{Cullis}[b]	phasing power[c]	occupancy
UO_2Ac_2		−0.111	−0.126	−0.031	15				0.60
		−0.320	−0.663	−0.050	18				0.50
	5	−0.260	−0.401	−0.165	21	0.20	0.66	1.58 (1.2)	0.30
		−0.084	−0.675	−0.144	33				0.26
		−0.251	−0.350	−0.137	24				0.25
CH_3HgCl	1	0.280	0.751	−0.005	12	0.13	0.72	1.1 (0.9)	0.48

[a] $R_{deriv} = \Sigma|F_{deriv} - F_{nat}|/\Sigma F_{nat}$. [b] $R_{Cullis} = \Sigma||FPH_{obs}| - |FPH_{calc}||/\Sigma||FPH_{obs}| - |FP_{obs}||$ for centric or acentric terms, and $\Sigma|\Delta ano_{obs} - \Delta ano_{calc}|/\Sigma|\Delta ano_{obs}|$ for anomalous differences. [c] The phasing power is the mean value of the heavy-atom structure amplitude divided by the lack of closure error. The values quoted are for acentric reflections with the centric value in parentheses.

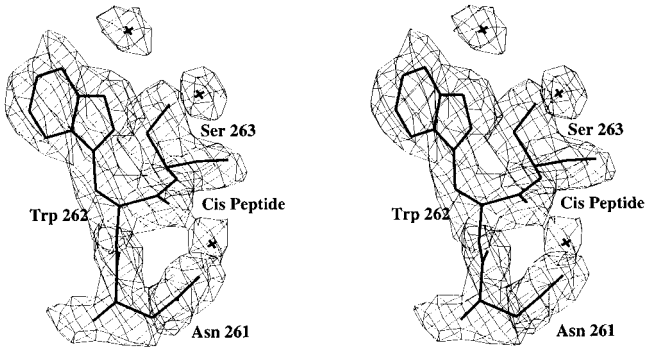

FIGURE 1: Divergent stereo representation of the DM-improved MIR map for the *B. agaradherans* Cel5A at 2.3 Å resolution contoured at 1.1σ. The region shown includes the nonproline *cis*-peptide between residues Trp 262 and Ser 263, a feature of family 5 glycoside hydrolases, which maintains Trp 262 in an orientation in which it can interact with substrate in the −1 subsite of the enzyme.

procedure. The data are 99.9% complete in the range 20−1.58 Å, with an overall R_{merge} of 0.048, a mean $I/\sigma I$ of 30.4 and a multiplicity of observation of 5.3. Data for the cellobiose complex extend to 2.0 Å resolution and have an R_{merge} of 0.034, with a mean $I/\sigma I$ of 39.2 and a multiplicity of 3.2. Heavy-atom data, to 2.3 Å resolution, were of a similar quality (data not shown). Details of the heavy-metal sites and phasing statistics are given in Table 1. The initial MIR map was of extremely high quality (figure of merit 0.58 at 2.3 Å resolution) allowing such features as nonprolyl cis peptides to be observed, Figure 1, and any poor regions of the map were healed after DM modification. Model building with O (*16*) proceeded smoothly. The Cel5A structure proved sufficiently dissimilar to the previously determined members of glycoside hydrolase family 5 that attempts at molecular replacement using the known family 5 structures (*10*, *22*, *23*) (both intact enzymes and the highly conserved β-sheet cores) were unsuccessful. This apparent failure of molecular replacement presumably reflecting the fact that some of the α-helices, whose repetitive motifs form dominant Patterson vectors, lie in structurally inequivalent positions.

The final native model featuring residues 4−303 has a crystallographic *R* value of 0.13, with a corresponding R_{free} of 0.17 for observed data between 20 and 1.57 Å resolution. This model has deviations from stereochemical target values of 0.017 and 0.032 Å (corresponding to approximately 1.1°), for 1−2 and 1−3 bonds, respectively. Final refinement statistics for the 1.6 Å native structure and the 2.0 Å cellobiose complex are given in Table 2. The low values for R_{cryst} and R_{free} for the native structure presumably reflect the high quality of the original native data (as discussed in ref *25*). All the nonglycine residues have conformational angles (φ,ψ) in permitted regions of the Ramachandran plot (*26*) with none of these in the "generously allowed" or "disallowed" regions as defined by PROCHECK (*27*).

Native Structure, Comparison with Other Family 5 Enzymes. The *B. agaradherans* Cel5A structure is of the $(\beta/\alpha)_8$ barrel type, the fold originally described for the triose-phosphate isomerase structure, Figure 2, and essentially as described for the previous family 5 structures. Three structures of endoglucanases from family 5 have previously been determined: endoglucanase EGCCA from *Clostridium cellulolyticum* (*22*), endocellulase E1 from *Acidothermus cellulolyticus* (*10*), and the endoglucanase CelC from *Clostridium thermocellum* (*23*, *24*). Structure-based sequence alignments reveal that these four family 5 endoglucanases display 15% sequence identity (between the two *Clostridial* enzymes) with just 10% sequence identity between the *B. agaradherans* Cel5A and either of the two *Clostridial* enzymes. These low values reveal how few residues, in this case only 29, are conserved in these structurally similar enzymes. Indeed, throughout the whole of family 5, just eight residues are invariant (*10*). At the 3-D structural level the *B. agaradherans* Cel5A has approximately 200 amino acids out of 299 in structurally equivalent positions to the other family 5 enzymes, with a root mean square (rms) Cα deviation of approximately 1.6 Å. Significant differences occur in the loop regions and in the position of the peripheral α-helices. There is a single nonprolyl *cis*-peptide between Trp 262 and Ser 263, Figure 1. Family 5 and most of the related clan GH-A enzymes (for review see ref *10*) have this *cis*-peptide, which allows the invariant tryptophan to both form the base of the −1 subsite and make a 2.9 Å hydrogen bond through its NE1 to the O-2 hydroxyl of the −2 subsite sugar, as described below.

The large number of distantly related sequences in glycoside hydrolase family 5 has led to some authors proposing a subfamily classification of more closely related sequences (*9*). This classification provides evidence for both divergent evolution from a common family 5 ancestral protein but also for convergent evolution of some structural features. A number of aspects of the Cel5A structure provide direct structural evidence for this subfamily classification. In particular, analysis of the regions contributing to the aglycon (+1/+2) binding sites provides direct evidence for convergent evolution toward a pyranoside binding motif. Tryptophan 178 lies in a position where it would form the

Table 2: Refinement and Structure Quality Statistics for Native and Complex *B. agaradherans* Cel5A Structures

	native	cellobiose complex
resolution of data (outer shell), Å	15−1.57 (1.64−1.57)	20−2.00 (2.10−2.00)
R_{merge}[a] (outer shell)	0.048 (0.196)	0.034 (0.065)
mean $I/\sigma I$ (outer shell)	30.4 (8.0)	39.2 (20.2)
completeness (outer shell), %	99.8 (98.4)	98.7 (99.8)
multiplicity (outer shell)	5.3 (4.8)	3.2 (3.0)
no. protein atoms (residues 4−303)	2398	2398
no. ligand atoms	N/A	23
no. solvent waters	504	497
resolution used in refinement	15−1.57 Å	15−2.0 Å
R_{cryst}	0.13	0.13
R_{free}	0.17	0.18
rms deviation 1−2 bonds (Å)	0.017	0.010
rms deviation 1−3 angles (Å)	0.032	0.028
rms deviation chiral volumes (Å^3)	0.126	0.119
avg main chain B (Å^2)	12.1	11.4
avg side chain B (Å^2)	15.6	14.2
avg solvent B (Å^2)	37.9	38.5
mean B ligand (Å^2)	N/A	−3 subsite (18.4), −2 subsite (24.4)
main chain ΔB, bonded atoms (Å^2)	1.1	1.7

[a] $R_{merge} = \sum_{hkl}\sum_i|I_{hkli} - \langle I_{hkl}\rangle|/\sum_{hkl}\sum_i\langle I_{hkl}\rangle$.

FIGURE 2: (A) Ribbon diagram and (B) stereo C_α-trace of the *B. agaradherands* Cel5A The catalytic acid/base and nucleophile, residues Glu 139 and Glu 228, are indicated in a "ball-and-stick" representation. These figures were prepared with the MOLSCRIPT program (*45*) and are in divergent stereo.

basis of the aglycon (+1) subsite, presumably through aromatic stacking with the hydrophobic faces of the pyranoside rings. This residue is invariant in all the subfamily 5-2 sequences, and in the Cel5A structure, it occurs on a small turn between strand β5 and the shortened helix α5. In the subfamily 5-4 enzyme, EGCCA from *C. cellulolyticum* (*22*), there is a comparably positioned, aglycon-binding residue, Trp 180. While the indole ring sits in an equivalent position, the residue itself is instead donated from a topologically inequivalent loop between strand β4 and helix α4. Again, this residue is invariant in that subfamily. In the family 5-3 structure, the *C. thermocellum* celC (*23, 24*), the residue is donated, as in the Cel5A case, on the loop between β5 and α5, but in this case the aromatic stacking potential is provided by a structurally equivalent tyrosine Tyr 176. In the subfamily 5-1 enzyme, the endocellulase E1 from *A. cellulolyticus* (*10*), the +1 binding site Trp 213, is located between β5 and α5, as in the 5-2 and 5-3 enzymes, but in this case it is preceded by a long loop region that extends the potential substrate binding surface considerably in the *A. cellulolyticus* enzyme compared to Cel5A. Together, these features are indicative of a convergent evolution to a functionally equivalent, but topologically unrelated, aglycon binding site.

The role of some of the conserved residues in family 5-2 has been examined on the enzyme EGZ from *E. chrysanthemi* in an elegant study combining conventional sequence alignments together with an Amber suppression system for site-specific mutagenesis (*11*). Two residues, conserved in family 5-2, warranted particular attention and were subjected to suppression analysis. In light of the 3-D structure of Cel5A presented here we observe that Arg 163 (equivalent to Arg 155 in EGZ) lies at the C-terminal end of helix α4. It hydrogen bonds to the main-chain carbonyls of residues Asn 169 and Pro 167 in the turn between α4 and strand β5 and forms a buried salt link with Asp 192. Asp 192 lies at the C-terminal end of helix α5. Conventional sequence alignments do not indicate that this residue is conserved, but all known family 5-2 sequences do possess an aspartate residue in this region of the sequence. Since the length of helix α5 is frequently variable and this is not taken into account in sequence alignments, we propose that Asp 192 may be conserved at the structural level. Thus, residue Arg 163 provides crucial interactions between helix α4 with helix α5, strand β5, and their interconnecting loops. These structural observations are entirely in agreement with the proposals made by Barras and co-workers mutagenesis that this residue must be extremely important in stabilizing the 3-D structure (*11*). Indeed, structural comparison with other family 5 members indicates that an equivalent Arg-Asp salt link is present in some, but not all, also as predicted by the EGZ study. His 206 (equivalent to His 198 in EGZ) displayed a semitolerant suppression pattern indicative of an important but not crucial role in catalysis. From the structure of Cel5A it is not apparent what role this residue could play.

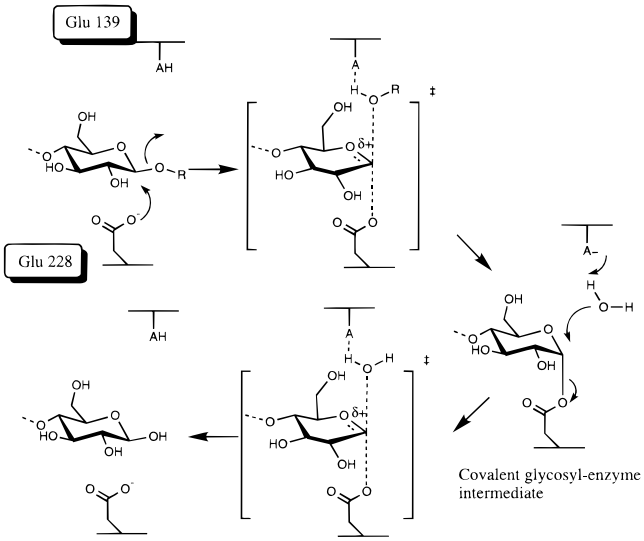

FIGURE 3: Double-displacement reaction mechanism for the family 5 *B. agaradherans* Cel5A.

His 206 is wholly surface exposed and over 20 Å from the active center. It is possible that given a conformational change it could play a role in substrate binding at the extreme reducing end of the substrate-binding cleft. As such it may play a role in the binding of long polymeric substrates such as (carboxymethyl)cellulose. A further possibility that cannot be ruled out is that it is somehow involved in polymeric substrate binding through an interaction with the cellulose-binding domain, which is absent from the Cel5A catalytic core structure presented here.

Catalytic Center and Cellobiose Binding. Cel5A is from glycoside hydrolase family 5, which, as discussed above, has structural, sequence, and mechanistic similarity with the many other glycoside hydrolases that form clan GH-A (*3*, *28*). It is particularly relevant that this clan includes the *Agrobacterium* β-glucosidase and the multifunctional *Cellulomonas fimi* cellulase/xylanase Cex, which are perhaps the most widely studied and understood of all the glycoside hydrolases (*29*−*33*). Cel5A, in common with these and other retaining cellulases, performs catalysis *via* a double-displacement mechanism (*34*) with a net retention of the anomeric configuration. A covalent glycosyl−enzyme intermediate is formed and subsequently rehydrolyzed via oxocarbenium-ion transition states, Figure 3. The covalent glycosyl−enzyme intermediate has been trapped sufficiently long for X-ray crystallographic analysis through the use of 2-fluoro sugars (*33*, *35*). In the case of Cel5A, sequence identity allows us to identify the two essential enzymatic functions. The Brønsted acid/base and enzymatic nucleophile are residues Glu 139 and Glu 228, respectively.

In order to determine the number of enzyme subsites that contribute to catalysis, the catalytic constants for a series of reduced cello-oligosaccharides were determined, both pH 7.5 and pH 9.0. The values obtained for the increase in $k_{cat.}/K_M$ for cellodextrins of increasing chain length, Table 3, suggest that the *B. agaradherans* Cel5A has approximately five to six binding subsites. The catalytic residues are situated at the bottom of a steep-sided gully across the surface of the enzyme, not a long continuous active site groove as has been observed for other endoglucanase structures. We estimate that there is sufficient room for just four of the binding

Table 3: Catalytic Activity on Reduced Cellodextrins at pH 7.5 and 9.0

		$k_{cat.}$ (s^{-1})	K_M (μM)	$k_{cat.}/K_M$ (s^{-1} μM^{-1})
rDP4	pH 7.5	0.05	208	0.002
	pH 9.0	0.05	333	0.002
rDP5	pH 7.5	18	77	0.23
	pH 9.0	13	100	0.13
rDP6	pH 7.5	36	58	0.62
	pH 9.0	38	47	0.81

[a] Kinetics were determined in a coupled reaction with cellobiose dehydrogenase (*12*, *46*). Standard errors are within 10%. The enzyme is not active against reduced oligosaccharide substrates smaller than rDP4.

subsites along the whole depression, with the −3 subsite (see below) external to this gully. Data collected for a crystal soaked in cellobiose revealed that cellobiose is bound to the −3 and −2 subsites of the enzyme (nomenclature according to ref *36*). There is no evidence for any ligand-induced conformational changes in the protein structure, as has been observed for the *C. thermocellum* enzyme (*24*), but none seems necessary since the catalytic glutamates are appropriately positioned even in the native enzyme structure. The electron density for the disaccharide is extremely well defined, Figure 4, and the individual pyranosides refine with mean temperature factors of 18 and 24 $Å^2$ for the −3 and −2 subsites respectively. At very low contour levels (< 0.08 e $Å^{-3}$) there is evidence of a second cellobiose molecule with an alternative occupancy spanning the −2 and −1 subsites. This species would appear to be highly disordered beyond the C3 and C5 atoms of the −1 subsite sugar and could not be modeled satisfactorily. This is entirely consistent with the view that the −1 subsite should favor transition-state and not substrate binding. Cel5A is catalytically inactive at the pH of the crystallization (unpublished results). We therefore interpret the low level density as a second discrete species and not as a transglycosylation event as others have reported on related family 5 enzymes (for example, see ref *10*). The interactions between cellobiose (−3/−2) and Cel5A are shown in Figure 5. The −3 subsite binds primarily through a hydrophobic stacking above Trp 39 on the surface of the enzyme. It makes few direct hydrogen bonds with the protein, with the O-2, O-3 and O-4 hydroxyls interacting only with solvent water. Only the O-6 hydroxyl interacts with the protein, making direct hydrogen bonds to both Tyr 40 and Lys 267. The −2 subsite sugar makes more direct interactions. The O-6 hydroxyl makes a direct H-bond to Tyr 66 and a water-mediated interaction with Ser 69. The O-3 hydroxyl makes H-bonds to Lys 267, His 35, and Glu 269. The interaction with Glu 269 would appear to be a low-barrier interaction at a distance of only 2.5 Å. The O-2 hydroxyl makes direct H-bonds with both Glu 269 and Trp 262. We would predict that, in addition to this hydrogen bonding role, Trp 262 also forms an aromatic stacking residue in the −1 subsite. The O-1 hydroxyl is present only in its β-configuration with not even weak density for the "wrong" α-anomer as might have been expected by mutarotation. The O-1 hydroxyl makes interactions only with solvent water.

CONCLUDING REMARKS

In addition to the families of related sequences, the increasing number of glycoside hydrolase structures reveals

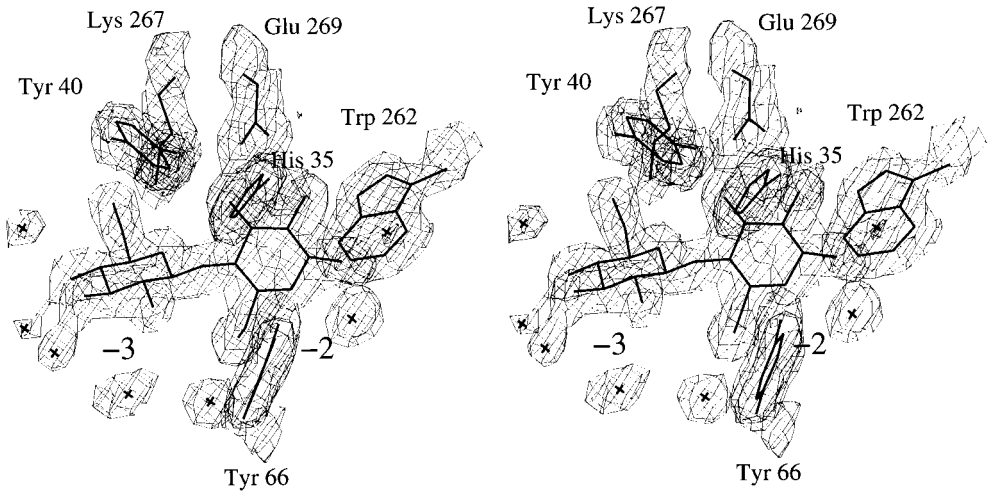

FIGURE 4: Divergent stereo representation of the electron density figure for cellobiose bound in the −3 and −2 subsites. The map is a maximum-likelihood weighted $2F_o - F_c$ map, based upon the native structure-factor phases together with the cellobiose complex structure-factor amplitudes. The contour level is approximately 0.41 e Å^{-3}.

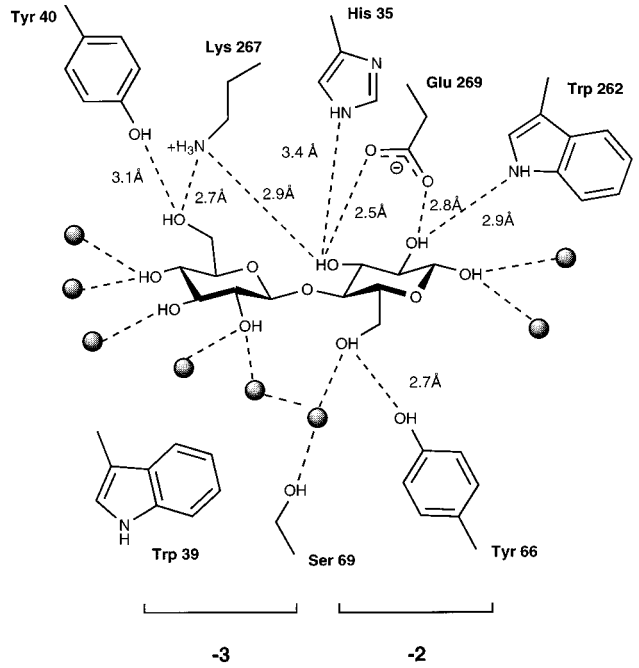

FIGURE 5: A schematic representation of the *B. agaradherans* cel5A−cellobiose (−3/−2) interactions.

relationships between 3-D structures that could not be detected by regular sequence analysis. We (*28*) and others (*37*) reported that many of the glycoside hydrolase families could be related by certain key signature motifs, a common 3-D fold, a topological similarity of the substrates, and a conserved catalytic machinery and stereochemistry. The confusion that surrounds the relationship between "families" and "superfamilies" led Henrissat to use the term "clan" to describe groupings of such families. Thus, the *B. agaradherans* Cel5A described here falls into clan GH-A, which, in addition to having a large number of cellulase structures, also contains enzymes with other distinct specificities such as the family 2 *Escherichia coli* β-galactosidase (*38*), the family 1 enzymes cyanogenic β-glucosidase (*39*), 6-phosphogalactosidase (*40*), and myrosinase (the only example of an *S*-glucosidase) (*35*), the family 10 xylanases (*41*−*43*), and the family 17 1,3- and mixed 1,4−1,3-glucanases (*44*). At present, 3-D representatives for oligosaccharide complexes for most of these families are not yet available, which prevents a comparison of their substrate specifities at the structural level. Eventually, however, this should allow an analysis of the exact molecular interactions that govern substrate specificity in these enzymes. The structure of the *B. agaradherans* Cel5A represents the first member of subfamily 2 of glycoside hydrolase family 5 to have been determined. Kinetic determinations suggest five to six binding subsites, consistent with the size and extent of the substrate-binding region elucidated by crystallography. Cellobiose is found bound in the −3/−2 subsites, allowing the dissection of all the protein−ligand interactions in those subsites. Studies of further oligosaccharide complexes of this and related enzymes are planned in order to dissect the nature of product specificity in this diverse clan of enzymes.

REFERENCES

1. Henrissat, B. (1991) *Biochem. J. 280*, 309−316.
2. Henrissat, B., and Bairoch, A. (1993) *Biochem. J. 293*, 781−788.
3. Henrissat, B., and Bairoch, A. (1996) *Biochem. J 316*, 695−696.
4. Henrissat, B., and Davies, G. J. (1997) *Curr. Op. Struct. Biol. 7*, 637−644.
5. Davies, G., and Henrissat, B. (1995) *Structure 3*, 853−859.
6. Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller, R. C., Jr., and Warren, R. A. J. (1991) *Microbiol. Rev. 55*, 303−315.
7. Bayer, E. A., Morag, E., and Lamet, R. (1994) *Trends Biotechnol. 12*, 379−386.
8. Gilbert, H. J., and Hazlewood, G. P. (1993) *J. Gen. Micro. 139*, 187−194.
9. Wang, Q., Tull, D., Meinke, A., Gilkes, N. R., Warren, R. A. J., Aebersold, R., and Withers, S. G. (1993) *J. Biol. Chem. 268*, 14096−14102.
10. Sakon, J., Adney, W. S., Himmel, M. E., Thomas, S. R., and Karplus, P. A. (1996) *Biochemistry 35*, 10648−10660.
11. Bortoli-German, I., Haiech, J., Chippaux, M., and Barras, F. (1995) *J. Mol. Biol. 246*, 82−94.
12. Schou, C., Rasmussen, G., Kaltoft, M.-B., Henrissat, B., and Schülein, M. (1993) *Eur. J. Biochem. 217*, 947−953.
13. Otwinowski, Z. (1993) in *Data Collection and Processing: proceedings of the CCP4 study weekend* (Sawyer, L., Issacs, N., and Bailey, S., Eds.) Science and Engineering Research Council, Daresbury, U.K.
14. Collaborative Computational Project Number 4. (1994) *Acta Crystallogr. D50*, 760−763.

15. Cowtan, K. D., and Main, P. (1996) *Acta Crystallogr. D49*, 148−157.
16. Jones, T. A., Zou, J.-Y., Cowan, S. W., and Kjeldgaard, M. (1991) *Acta Crystallogr. A47*, 110−119.
17. Brünger, A. T. (1992) *Nature 355*, 472−475.
18. Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) *Acta Crystallogr. D 53*, 240−255.
19. Lamzin, V. S., and Wilson, K. S. (1993) *Acta Crystallogr. D49*, 129−147.
20. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. T., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) *J. Mol. Biol. 112*, 535−542.
21. Raymond, S., Heyraud, A., Qui, D. T., Kvick, A., and Chanzy, H. (1995) *Macromolecules 28*, 2096−2100.
22. Ducros, V., Czjzek, M., Belaich, A., Gaudin, C., Fierobe, H.-P., Belaich, J.-P., Davies, G. J., and Haser, R. (1995) *Structure 3*, 939−949.
23. Dominguez, R., Souchon, H., Spinelli, S., Dauter, Z., Wilson, K. S., Chauvaux, S., Béguin, P., and Alzari, P. M. (1995) *Nat. Struct. Biol. 2*, 569−576.
24. Dominguez, R., Souchon, H., Lascombe, M.-B., and Alzari, P. M. (1996) *J. Mol. Biol 257*, 1042−1051.
25. Kleywegt, G. J., and Brünger, A. T. (1996) *Structure 4*, 897−904.
26. Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963) *J. Mol. Biol 7*, 95−99.
27. Laskowski, R. A., McArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) *J. Appl. Crystallogr. 26*, 282−291.
28. Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J.-P., and Davies, G. (1995) *Proc. Natl. Acad. Sci. U.S.A. 92*, 7090−7094.
29. Withers, S. G., Street, I. P., Bird, P., and Dolphin, D. H. (1987) *J. Am. Chem. Soc 109*, 7530−7531.
30. Street, I. P., Kempton, J. B., and Withers, S. G. (1992) *Biochemistry 31*, 9970−9978.
31. Tull, D., and Withers, S. G. (1994) *Biochemistry 33*, 6363−6370.
32. Namchuk, M. N., and Withers, S. G. (1995) *Biochemistry 34*, 16194−16202.
33. White, A., Tull, D., Johns, K., Withers, S. G., and Rose, D. R. (1996) *Nat. Struct. Biol. 3*, 149−154.
34. Koshland, D. E. (1953) *Biol. Rev. 28*, 416−436.
35. Burmeister, W. P., Cottaz, S., Driguez, H., Palmieri, S., and Henrissat, B. (1997) *Structure 5*, 663−675.
36. Davies, G. J., Wilson, K. S., and Henrissat, B. (1997) *Biochem. J. 321*, 557−559.
37. Jenkins, J., Leggio, L. L., Harris, G., and Pickersgill, R. (1995) *FEBS Lett. 362*, 281−285.
38. Jacobson, R. H., Zhang, X.-J., DuBose, R. F., and Matthews, B. W. (1994) *Nature 369*, 761−766.
39. Barrett, T., Suresh, C. G., Tolley, S. P., Dodson, E. J., and Hugues, M. A. (1995) *Structure 3*, 951−960.
40. Wiesmann, C., Beste, G., Hengstenberg, W., and Schulz, G. E. (1995) *Structure 3*, 961−968.
41. Derewenda, U., Swenson, L., Green, R., Wei, Y., Morosoli, R., Shareck, F., Kluepfel, D., and Derewenda, Z. S. (1994) *J. Biol. Chem. 269*, 20811−20814.
42. White, A., Withers, S. G., Gilkes, N. R., and Rose, D. R. (1994) *Biochemistry 33*, 12546−12552.
43. Harris, G. W., Jenkins, J. J., Connerton, I., Cummings, N., Leggio, L. L., Scott, M., Hazlewood, G. P., Laurie, J. I., Gilbert, H. J., and Pickersgill, R. W. (1994) *Structure 2*, 1107−1116.
44. Varghese, J. N., Garrett, T. P. J., Coleman, P. M., Chen, L., Høj, P. B., and Fincher, G. B. (1994) *Proc. Natl. Acad. Sci. U.S.A. 91*, 2785−2789.
45. Kraulis, P. J. (1991) *J. Appl. Crystallogr. 24*, 946−950.
46. Schülein, M. (1997) *J. Biotechnol. 57*, 71−81.

BI972162M