Crystal Structure of Thermostable Family 5 Endocellulase E1 from *Acidothermus cellulolyticus* in Complex with Cellotetraose[†,‡]

Joshua Sakon,[*,§] William S. Adney,[||] Michael E. Himmel,[||] Steven R. Thomas,[||] and P. Andrew Karplus[*,§]

Section of Biochemistry, Molecular and Cell Biology, Cornell University, Ithaca, New York 14853, and National Renewable Energy Laboratory, Golden, Colorado 80401

Received February 23, 1996; Revised Manuscript Received May 30, 1996[⊗]

ABSTRACT: The crystal structure of the catalytic domain of the thermostable endocellulase E1 from *Acidothermus cellulolyticus* in complex with cellotetraose has been solved by multiple isomorphous replacement and refined at 2.4 Å resolution to an *R*-factor of 0.18 (R_{free} = 0.24). E1cd is a member of the 4/7 superfamily of hydrolases, and as expected, its structure is an $(\alpha/\beta)_8$ barrel, which constitutes a prototype for family 5−subfamily 1 cellulases. The cellotetraose molecule binds in a manner consistent with the expected Michaelis complex for the glycosylation half-reaction and reveals that all eight residues conserved in family 5 enzymes are involved in recognition of the glycosyl group attacked during cleavage. Whereas only three residues are conserved in the whole 4/7 superfamily (the Asn/Glu duo and the Glu from which the name is derived), structural comparisons show that all eight residues conserved in family 5 have functional equivalents in the other 4/7 superfamily members, strengthening the case that mechanistic details are conserved throughout the superfamily. On the basis of the structure, a detailed sequence of physical steps of the cleavage mechanism is proposed. A close approach of two key glutamate residues provides an elegant mechanism for the shift in the pK_a of the acid/base for the glycosylation and deglycosylation half-reactions. Finally, purely structural based comparisons are used to show that significant differences exist in structural similarity scores resulting from different methods and suggest that caution should be exercised in interpreting such results in terms of implied evolutional relationships.

Cellulases catalyze the hydrolysis of cellulose, an unbranched β-1,4-linked homopolymer of glucose that is the major structural polysaccharide component of plant biomass. Complete hydrolysis of cellulose yields a single, easily fermentable product, glucose, and much cellulase research is aimed at improving the enzymatic hydrolysis of cellulose to a point that would enable biomass conversion to be economically competitive (Wyman et al., 1993). In 1985, *Acidothermus cellulolyticus* (new genus and species), a thermotolerant, acidophilic bacterium, was isolated from samples collected in Yellowstone National Park (Mohagheghi et al., 1986; Tucker et al., 1989). The *A. cellulolyticus* cellulase system is noncellulosomal, making it distinct from most cellulase systems produced by thermophiles. *A. cellulolyticus* produces at least three endoglucanases (Adney et al., 1994; Tucker et al., 1992), and one of them, E1 endoglucanase, is particularly interesting because it is highly thermostable (T_{opt} = 81 °C; Himmel et al., 1994) and has very high specific activity on carboxymethylcellulose (Thomas et al., 1995). These properties make E1[1] an attractive target for protein engineering to improve cellulase activity.

The E1 gene sequence (GenBank Accession No. U33212) shows that the mature E1 protein has 521 residues and consists of a catalytic domain (E1cd), a proline/serine/threonine-rich linker region, and a cellulose-binding domain. Among the 45 known families of glycosyl hydrolases (Henrissat & Bairoch, 1993), sequence comparisons clearly place the catalytic domain of E1 (E1cd) in family 5 (Wang et al., 1993). Family 5 (also known as cellulase family A) is the largest of the β-glycohydrolase families classified to date, including over 60 bacterial and fungal enzymes which all cleave with retention of configuration. The sequences of family 5 members are rather diverse and were reported to share only seven conserved residues (Wang et al., 1993), equivalent to Arg-62, His-116, Asn-161, Glu-162, His-238, Tyr-240, and Glu-282 in E1cd. Because of this diversity, family 5 has been further subdivided into five subfamilies within which amino acid sequence similarities are above 25% (Wang et al., 1993) and homology modeling methods can be used [e.g., Sali and Blundell (1993)].

Among the strictly conserved residues, the Glu-282 equivalent has been identified as the nucleophile in the displacement reaction (Tull et al., 1991), and the equivalent of Glu-162 (Bortoli-German et al., 1995; Wang et al., 1993) has been implicated as the proton donor. An extensive mutagenesis study has shown that cellulase activity is highly sensitive to the replacement of any of the seven conserved residues (Bortoli-German et al., 1995). Crystal structures have recently been published for a subfamily 3 member (CelC from *Clostridium thermocellum*; Dominguez et al., 1995) and a subfamily 4 member (endoglucanase A from *Clostridium cellulolyticum*; Ducros et al., 1995). These structures show that the seven conserved residues are all near

[†] This work was funded in part by the Ethanol from Biomass Program of the Biofuels System Division of the U.S. Department of Energy (RAH-5-15113).

[‡] The full coordinates (1ECE) and structure factor amplitudes (R1ECESF) have been deposited in the Protein Data Bank for immediate release.

* To whom correspondence should be addressed.

[§] Cornell University.

[||] National Renewable Energy Laboratory.

[⊗] Abstract published in *Advance ACS Abstracts,* July 15, 1996.

[1] Abbreviations: E1, mature *Acidothermus cellulolyticus* endocellulase E1; E1cd, catalytic domain of E1. All amino acids are abbreviated as recommended by the IUPAC−IUB Joint Commission on Biochemical Nomenclature (1985).

S0006-2960(96)00443-6 CCC: $12.00

each other in the active site, but neither of the structures provide detailed information on ligand binding or the catalytic mechanism.

As a member of glycohydrolase family 5, E1cd is also part of a large superfamily, encompassing families 1 (β-glucosidase, lactase phlorizin hydrolase, 6-phospho-β-glucosidase, 6-phospho-β-galactosidase, β-galactosidase, cyanogenic β-glucosidase), 2 (β-galactosidase, β-glucuronidase), 5 (cellulase, β-mannanase), 10 (xylanase), 17 (β-1−3,1−4-glucanase), 30 (glucocerebrosidase), 35 (β-galactosidase), 39 (α-L-iduronidase), and 42 (β-galactosidase). These families are proposed to share a common $(\alpha/\beta)_8$ barrel fold, three conserved residues, and a retaining cleavage mechanism (Henrissat et al., 1995; Jenkins et al., 1995). The superfamily has been termed the 4/7 superfamily because the three key residues are an adjacent Asn-Glu pair at the end of β-strand 4 and a Glu at the end of β-strand 7.

The generally accepted catalytic mechanism of these superfamily members is a double-displacement mechanism put forth originally by Koshland (1953). It involves an initial binding of the substrate to the enzyme, followed by a general acid-catalyzed attack of an enzymatic nucleophile upon the anomeric center to form a glycosyl-enzyme intermediate (McCarter & Withers, 1994; Sinnott, 1990). This intermediate is then hydrolyzed by a general base-catalyzed attack of water upon the anomeric center, forming the product and returning the enzyme to its original protonation state (McCarter & Withers, 1994; Sinnott, 1990). Mutagenesis and kinetic studies have allowed putative assignment of the first conserved Glu as the acid/base and the second conserved Glu as the nucleophile (McCarter & Withers, 1994). The use of substrates with a 2-fluoro group has allowed stabilization of covalent intermediates (Withers & Street, 1988) and very recently allowed elucidation of the structure of a covalent intermediate bound to Cex, a family 10 enzyme (White et al., 1996). This structure has strengthened the case for the double displacement mechanism and has given insight into the role of the conserved Asn.

To guide protein engineering of *A. cellulolyticus* endoglucanase E1, we have solved the structure of E1cd using crystals grown in the presence of cellobiose. The structure shows that the crystallized enzyme has cellotetraose bound, presumably resulting from an E1cd-catalyzed condensation of two cellobiose molecules. Our results thus provide a structural prototype for family 5−subfamily 1 cellulases including details of substrate binding and insights into the catalytic mechanism for the class of enzymes. We also analyze global and active site similarities with other structurally known glycosyl hydrolases and propose mechanistic details that structural and sequence comparisons suggest is conserved in the whole 4/7 superfamily.

MATERIALS AND METHODS

Substrates and Reagents. All reagents used were of the highest purity available and were purchased from Sigma Chemical Co. (St. Louis, MO) unless otherwise noted.

Enzyme Assays. Enzymatic activity was spectrophotometrically measured at 410 nm through changes in the absorbance due to hydrolysis of *p*-nitrophenyl β-D-cellobioside in the presence of 2.0 M Na_2CO_3. One unit of activity was defined as that amount of enzyme that cleaves 1.0 μmol of substrate/min at 45 °C in 50 mM acetate, pH 5.0.

Heterologous E1 Endoglucanase Expression System. Streptomyces lividans TK24 was transformed with pIJ702 (Hopwood et al., 1985) into which a 3.7 kb *Pvu*I fragment of *A. cellulolyticus* genomic DNA carrying native E1 gene had been subcloned. Cultures were kept frozen at −70 °C after addition of 77 μL of dimethyl sulfoxide/mL of culture suspension (Himmel et al., 1994).

S. lividans were grown in tryptone soya broth (TSB; Difco Laboratories, Detroit, MI) according to Hopwood et al. (1985). Plasmid pIJ702 and its derivatives were selected in liquid medium with 5 μg/mL thiostrepton. An aliquot of frozen *S. lividans* culture was thawed and transferred to 500 mL of TSB medium. After incubation for 2 days at 30 °C with rotary agitation (250 rpm), this culture was used to inoculate 8 L of TSB medium in NBS Microferm fermenters (New Brunswick Scientific, Brunswick, NJ). Fermented cultures were kept at pH 6.8 by addition of 1.0 N HCl while the temperature was maintained at 30 °C. The concentration of dissolved oxygen was maintained at 20% of saturation by addition of pure oxygen accompanied by impeller agitation at 300 rpm.

Purification of Recombinant E1 Endoglucanase from S. lividans Culture Broth. Culture broth from four 8 L recombinant *S. lividans* fermentations was clarified by ultracentrifugation using a CEPA (New Brunswick Scientific; Brunswick, NJ) Type LE continuous flow centrifuge operated at 25 000 rpm and 4 °C. Following concentration to a final volume of 300 mL using an Amicon (Danvers, MA) CH2 concentrator equipped with three 10 kDa cutoff (Type H1P10-43) hollow fiber membranes, ammonium sulfate was added to a final concentration of 0.5 M. The entire concentrate was loaded onto an Amicon G44 × 250 glass column packed with a 250 mL of Pharmacia (Piscataway, NJ) Fast Flow Phenyl Sepharose and attached to a Pharmacia FPLC system. The column was then washed with 5 volumes of 20 mM Tris, pH 8.0, containing 0.5 M $(NH_4)_2SO_4$. E1 was eluted at a flow rate of 2 mL/min by a descending linear gradient ending in 20 mM Tris, pH 8.0. Fractions eluting at 0% $(NH_4)_2SO_4$ and containing the recombinant enzyme were combined, concentrated, and diafiltered against 20 mM Tris, pH 8.0.

The diafiltered enzyme was next subjected to anion-exchange chromatography by loading to a Pharmacia Q-Sepharose HiLoad 16/10 high-performance column. The enzyme was eluted by a shallow linear gradient ending in 20 mM Tris, pH 8.0, containing 1.0 M NaCl. Active fractions eluted near 150 mM NaCl were combined, concentrated, and subjected to high-performance size exclusion chromatography (HPSEC) using a Pharmacia Superdex 200, HiLoad preparative grade column, with a flow rate of 0.5 mL/min in 50 mM ammonium acetate, pH 6.2. The elution profile showed a single symmetrical peak. Enzyme purity was supported by the migration of the protein as a single silver-stained band corresponding to 72 kDa on SDS−PAGE using Novex (San Diego, CA) precast 8−15% gradient gels.

Preparation of E1cd. Natural E1 was cleaved for 24 h at 28 °C using papain (Boehringer Mannheim, Indianapolis, IN) in 50 mM ammonium acetate, pH 6.2, with a molar ratio of E1 (72 kDa) to papain (23 kDa) of 6:1. E1cd was then separated from undigested E1, E1 cellulose-binding domain, and papain by HPSEC as described above. Purified E1cd migrates as a single band at $M_r = 42\,000$ when subjected to SDS−PAGE.

Table 1: X-ray Diffraction Data Statistics

	native	PIP	$PdCl_6$	$PdCl_6$	UO_2	HgI_4	KI_3
soak							
concn (mM)		satd	12	12	58	17	5
time (days)		3	1	1	0.5	3	1
no. of sites		4	9	9	8	11	2
R_{sym} (%)[a]	9.2	13.0	10.0	6.7	13.9	11.2	9.1
R_{iso} (%)[b]		23.0	24.0	24.0	24.0	18.0	20.0
unique reflections	47386	30430	22030	24847	18783	18675	21035
redundancy	2.2	4.8	2.5	2.0	2.2	1.9	2.3
completeness (%)	90.0	91.0	92.0	84.9	95.8	83.9	86.8
resolution range (Å)	∞−2.4	∞−2.9	∞−3.3	10−3.0	∞−3.4	∞−3.3	∞−3.2
phasing power		1.41	1.37	1.28	0.88	0.84	1.06
anomalous phasing power		0.50			0.51		

[a] $\sum|I - \langle I\rangle|/\sum I$, where I is the intensity measurement of reflections and $\langle I\rangle$ is the mean of the symmetry related reflections. [b] $\sum|F_N - F_H|/\sum F_H$, where F_N is the native structure factor amplitude and F_H is the derivative structure factor amplitude. Abbreviations were used for the following heavy-atom reagents: di-μ-iodobis(ethylenediamine)diplatinum nitrate (PIP), K_2PdCl_6 ($PdCl_6$), uranyl acetate (UO_2), K_2HgI_4 (HgI_4) with 0.1 M KI.

Crystallization. Condition 59 (2.0 M NaCl and 50 mM sodium acetate at pH 4.6) from a factorial screening method (Cudney et al., 1994; Jancarik & Kim, 1991) gave single crystals. Optimization of this condition yielded trigonal bipyramids measuring up to 1500 μm along the unique axis and 400 μm in width. The crystals grew in one month from 20 μL hanging drops containing 11.4 mg/mL E1cd, 2.3 M NaCl, 0.3 M cellobiose, and 0.1 M citrate buffer (pH 4) at room temperature. The precession photographs of *hh*0 and *h*0*l* zones suggested that the crystals belong to a trigonal space group, either $P3_121$ or $P3_221$, with unit cell dimensions $a = b = 97.13$ Å and $c = 258.71$ Å. The ambiguity in the space group was resolved later by using anomalous signals of a di-μ-iodobis(ethylenediamine)diplatinum nitrate (PIP) derivative, resulting in $P3_221$ as the true space group. There are two molecules per asymmetric unit with 68% (v/v) solvent. Unfrozen crystals have been seen to diffract to 1.6 Å resolution at the Cornell High Energy Synchrotron Source, but thus far freezing always has deteriorated the diffraction quality.

Crystallographic Data Collection and Isomorphous Replacement. X-ray diffraction data collected for 65 heavy-atom soaks yielded six isomorphous derivatives. The data were measured using two San Diego Multiwire Systems Mark II detectors (Hamlin, 1985) with graphite-monochromated Cu Kα X-rays from Rigaku RU200 operating at 50 kV and 150 mA with a 0.5 × 5 mm focus. This structural analysis was limited to 2.4 Å resolution because of the relatively long *c*-axis and the finite brilliance of the in-house source. Data reduction was carried out using the program SCALEPACK (Otwinowski, 1985). The major heavy-atom binding sites were obtained with isomorphous difference Pattersons and direct methods using SHELXS86 (Sheldrick, 1990). Both identification of minor sites and confirmation of individual sites were made by difference Fourier techniques. Individual heavy-atom parameters were refined (Tables 1 and 2) using PHASES (Furey & Swaminathan, 1996) to yield phases with an average figure of merit of 0.54 to 3.2 Å resolution.

Phase Improvement and Model Building. The multiple isomorphous replacement (MIR) map at 3.2 Å resolution was solvent-flattened and iteratively averaged using noncrystallographic symmetry. These operations were performed by PHASES suite programs. The resulting map revealed an $(\alpha/\beta)_8$ barrel motif, although complete chain tracing was not possible. Fragments of poly(Ala) were built into the averaged density using CHAIN (Sack, 1988), and the second molecule in the asymmetric unit was generated by NCS operators. The phases obtained from the partial model were combined with the MIR phases using sigmaA weights (Read, 1986). The electron density map calculated using the combined phases allowed improvement of the protein model which was then refined against data from infinity to 3.0 Å resolution using TNT (Tronrud et al., 1987). Fifteen cycles of such model building and phase combination yielded four chains of the poly(Ala/Ser) model which were further refined

Table 2: Individual Heavy-Atom Sites

heavy atom	site	*X*	*Y*	*Z*	Occ	*B*	neighbors[a]
PIP	1	0.175	0.207	0.957	0.63	2.0	M1075
PIP	2	0.160	0.251	0.917	0.66	48.6	M75
PIP	3	−0.069	0.131	0.936	0.66	80.0	M84
PIP	4	0.297	0.444	0.970	0.64	80.0	M1084
$PdCl_6$	1	0.175	0.207	0.957	0.30	2.0	M1075
$PdCl_6$	2	0.160	0.251	0.917	0.49	2.0	M75
$PdCl_6$	3	−0.069	0.131	0.936	0.24	2.0	M84
$PdCl_6$	4	0.297	0.444	0.970	0.40	2.0	M1084
$PdCl_6$	5	0.420	0.125	0.997	0.21	2.0	G1004
$PdCl_6$	6	0.986	0.480	0.980	0.42	2.0	H1164
$PdCl_6$	7	0.733	0.312	0.817	0.26	20.8	G4
$PdCl_6$	8	0.306	0.978	0.943	0.30	2.0	H164
$PdCl_6$	9	0.407	0.950	0.991	0.21	8.0	Q1142
UO_2	1	0.369	0.038	0.981	0.13	2.0	E282
UO_2	2	0.996	0.437	0.946	0.11	2.0	E1282
UO_2	3	0.786	0.471	0.925	0.45	80.0	D209
UO_2	4	0.759	0.426	0.925	0.49	80.0	D252
UO_2	5	0.549	0.108	0.914	0.62	80.0	D1209
UO_2	6	0.523	0.074	0.906	0.57	80.0	D1252
UO_2	7	0.242	0.089	0.863	0.39	80.0	D312
UO_2	8	0.241	0.062	0.874	0.42	72.1	E12
HgI_4	1	0.229	0.241	1.005	0.41	80.0	E134
HgI_4	2	0.176	0.741	0.977	0.45	80.0	R184
HgI_4	3	0.221	0.716	0.968	0.56	80.0	K149, N195, Y5
HgI_4	4	0.998	0.851	0.952	0.48	80.0	Q337, R44, Q52
HgI_4	5	0.191	0.109	0.913	0.31	80.0	A1, S352
HgI_4	6	0.454	0.184	0.897	0.27	80.0	T298, N257,Q301
HgI_4	7	0.316	0.678	0.882	0.47	80.0	K1149, N1195, Y1005
HgI_4	8	0.813	0.395	0.873	0.27	80.0	T1298, N1257, Q1301
HgI_4	9	0.039	0.017	1.005	0.16	80.0	D1334, K1333
HgI_4	10	0.953	0.171	0.924	0.15	80.0	D334, K333
HgI_4	11	0.286	0.408	0.979	0.12	80.0	Q1273
KI_3	1	0.169	0.201	0.917	0.79	2.0	Y82
KI_3	2	0.216	0.214	0.936	0.45	17.3	Y1082

[a] Neighbors are those residues having polar atoms within 3.5 Å of the heavy-atom site. The residue number for one of the molecules begins from 1 and the other begins from 1001.

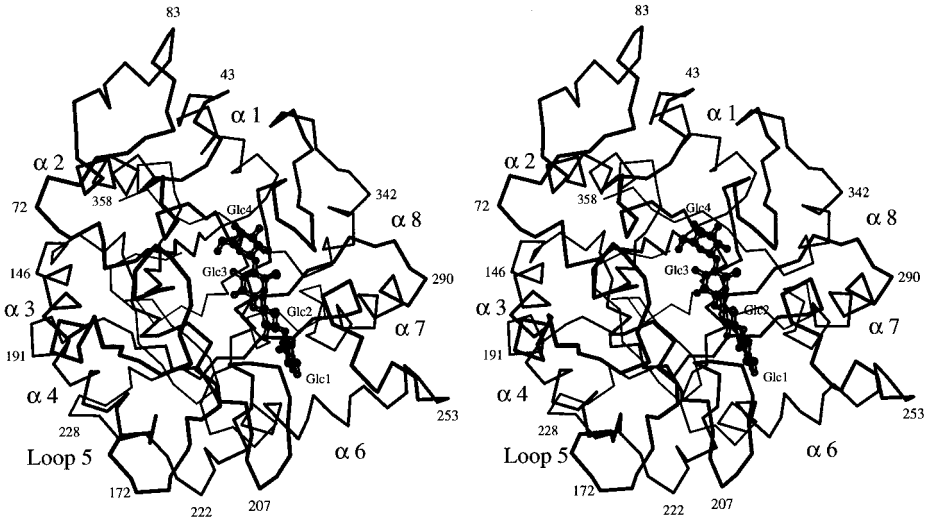

FIGURE 1: Stereo C_α plot of E1cd. The C_α backbone is traced with a thin line. A cellotetraose molecule bound in the active site of the enzyme is represented by thick lines. The enzyme exhibits an $(\alpha/\beta)_8$ barrel fold lacking helix 5. A very short or entirely lacking helix 5 is a feature that is conserved in known members of family 5 glycosyl hydrolases. This molecular orientation is maintained for Figures 4 and 5. The diagram and Figures 5 and 8 were drawn with MOLSCRIPT (Kraulis, 1991).

by the simulated annealing routines in X-PLOR (Brunger et al., 1987). Six cycles of manual model adjustment and simulated annealing yielded two continuous chains of the poly(Ala/Ser) model with a crystallographic *R*-factor of 38.8% and an R_{free} of 45.8% for all X-ray diffraction data from 10 to 3 Å resolution. R_{free} was calculated on the basis of 5% of the data which were not used in refinement.

Crystallographic Refinement. Twenty-five additional cycles of model building, refinement by simulated annealing, and phase combination yielded two complete models of E1 comprising of residues 1−358. Seven further cycles of manual adjustment and least-squares refinement of positional parameters and restrained individual temperature factors were carried out at 2.4 Å resolution. Water molecules were introduced where all three of the following criteria were met: electron density peaks $\geq 3\rho_{rms}$ in an $F_o - F_c$ difference Fourier map and $\geq 0.75\rho_{rms}$ in a $2F_o - F_c$ difference Fourier map and good hydrogen bond geometry. In the majority of cases, a solvent molecule assigned for one of the monomers was also found at an equivalent position in the other. As refinement progressed, difference Fourier maps showed a bound cellotetraose which was subsequently built to account for the observed density. The cellotetraose molecules bound in both E1cd molecules have refined temperature factors of 50−65 Å^2, which is about double the average protein temperature factor of 28 Å^2. The final model consisting of 716 amino acid residues (5702 atoms), two cellotetraose molecules (90 atoms), and 375 water molecules in the asymmetric unit has a crystallographic *R*-factor of 17.9% and an R_{free} of 23.6% for all X-ray diffraction data from 10 to 2.4 Å. The overall root-mean-square (rms) deviations from ideality for the refined model are 0.007 Å for bond lengths, 1.7° for bond angles, and 1.2° for fixed improper dihedrals, while the rms deviation of individual temperature factors between bonded atoms is 1.4 Å^2. Among the non-glycine and non-proline residues, 513 (85%) have ϕ, ψ values in the "most favored regions" of the Ramachandran plot (Laskowski et al., 1993). No residues are found in the "disallowed" regions. A tight noncrystallographic symmetry restraint was applied for all the main chain atoms and for cellobiose molecules. The rms difference between the C_α atoms of the two E1cd molecules in the asymmetric unit is 0.18 Å, and maximal deviations near 0.35 Å are seen in a loop (residues 82−85) involved in the intermolecular interaction.

RESULTS AND DISCUSSION

Structure Description. The structures of two noncrystallographic symmetry-related copies of E1cd have been determined by X-ray crystallography at 2.4 Å resolution using multiple isomorphous replacement. Both chains are visible from Ala-1 through Val-358. As E1cd was generated by papain cleavage of intact E1, the true carboxyl terminus is not known. Some residual electron density extending beyond Val-358 suggests that one or more additional disordered residues are present. Also, two loops (207−209 and 252−257) have high temperature factors (>50 Å^2), but the chain path is still clear in these regions.

The E1cd structure consists of a single domain with an $(\alpha/\beta)_8$ barrel fold in which a short irregular chain segment replaces helix 5 (Figure 1). The exact residues found in the β-strands and α-helices are shown in Figure 2 along with an alignment with other known family 5−subfamily 1 sequences. Aside from the structural elements of the barrel, there are only three short α-helices and five short β-strands. Three of the short β-strands occur in residues 1−21, and together with the extended C-terminal segment (residues 355−358), these strands seal the amino-terminal end of the barrel from solvent. At the carboxy-terminal end of the barrel, loops containing 16−26 residues each form the walls of a crevice running along the length of the molecule (9 Å wide, 30 Å long, and 10 Å deep). The cleft runs roughly from β-strand 1 to β-strand 6 and is the site of substrate binding. The two disulfide bonds of E1cd occur in the loops surrounding the active site cleft: one joining Cys-34 (loop 1) and Cys-120 (loop 3) and the other linking Cys-168 and Cys-171 (loop 4) in a tight turn with sequence Cys-Gly-Trp-Cys. *cis*-Peptide bonds occur preceding Pro-166, Pro-256, and Ser-320. The electron density for the rare non-prolyl *cis*-peptide at Ser-320 is unambiguous (Figure 3). Ser-320 is at the active site and is well conserved in subfamily 1, but Pro-166 and Pro-256 are not in the active site cleft

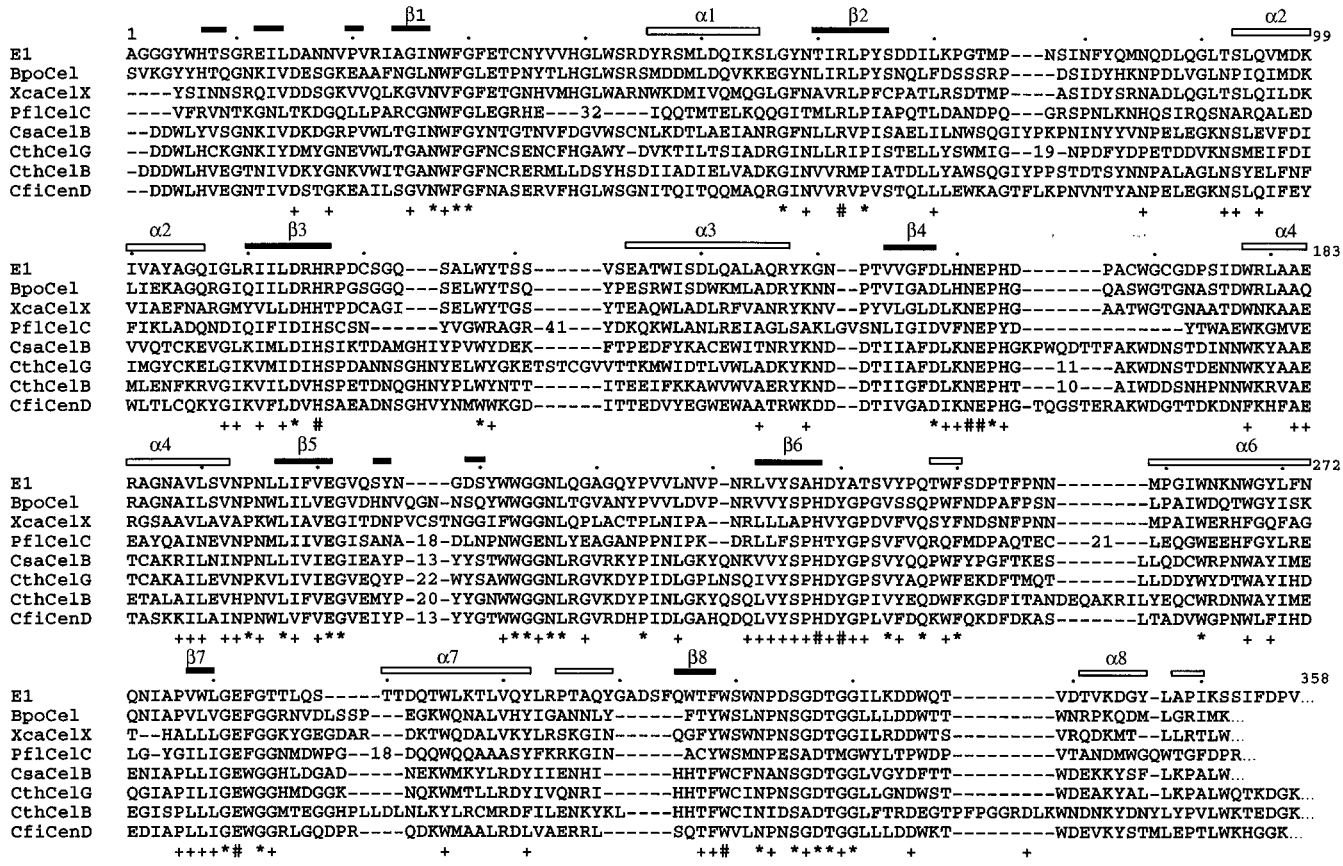

FIGURE 2: Amino acid sequence alignment of E1cd with family 5−subfamily 1 cellulases. The E1cd structure, which is the first for family 5−subfamily 1 enzymes, enables us to update the amino acid sequence alignment for this family (Wang et al., 1993). The updated alignment shows additional amino acid conservation in the subfamily such as Asn-27, Phe-29, and Glu-201 (see text). The sequences shown are as follows: BpoCel, *Bacillus polymyxa* endoglucanase (Baird et al., 1990); XcaCelX, *Xanthomonas campestris* endoglucanase XCA (Gough et al., 1990); PflCelC, *Pseudomonas fluorescens* endoglucanase C (Ferreira et al., 1991); CsaCelB, *Caldocellum saccharolyticum* endoglucanase B (Saul et al., 1989); CthCelG, *C. thermocellum* endoglucanase G (Lemaire & Beguin, 1993); CthCelB, *C. thermocellum* endoglucanase B (Grepinet & Beguin, 1986); CfiCenD, *Cellulomonas fimi* endoglucanase D (Meinke et al., 1993). β-Strands and α-helices as assigned by DSSP (Kabsh & Sander, 1983) are denoted by solid and open bars, respectively. Residue positions that are identical in all eight sequences ($*$) or in 6 or 7 sequences ($+$) are indicated. The eight residues conserved throughout family 5 are indicated by (#).

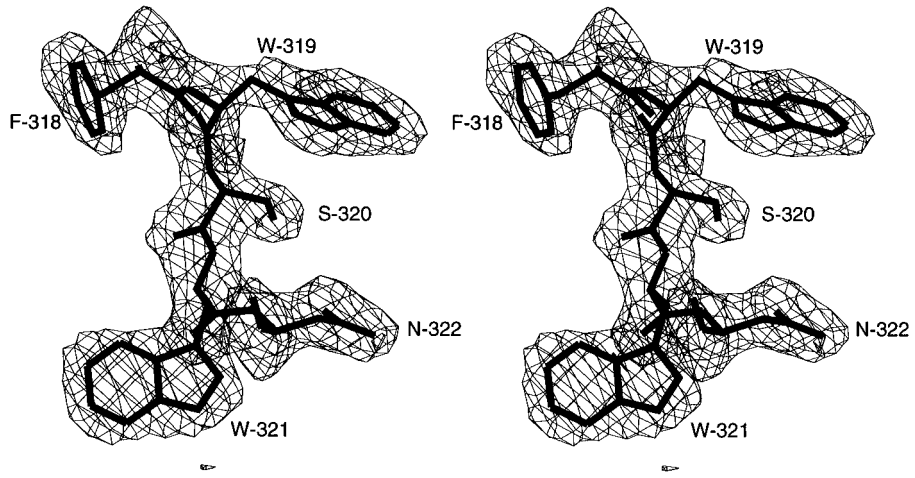

FIGURE 3: Stereo diagram of electron density corresponding to residues Phe-318-Trp-Ser-Trp-Asn, showing the *cis*-peptide bond between Trp-319 and Ser-320. The electron density map is from an annealed omit map, which is a $|F_o - F_c|$ difference Fourier map calculated after the final model without the relevant residues has been subjected to a round of simulated annealing. The contour level is 2.5 times the rms electron density of the map. Figures 3 and 4 were drawn with CHAIN (Sack, 1988).

and are not conserved (Figure 2), suggesting alternate conformations are adopted at those positions in the other subfamily members. Comparisons with the other two family 5 endoglucanase structures reveal a high similarity of the domain structure but rather extensive variation in the loops surrounding the active site cleft (see later).

The two molecules of E1cd in the asymmetric unit are related by approximately 2-fold symmetry ($\kappa = 182°$), so that they appear as a dimer with their active sites near each other. Predominantly polar residues from carboxy-terminal loops 1, 2, 3, and 8 are involved in the intermolecular interactions that bury about 960 Å^2 (7%) of the E1cd surface.

This polar interface is probably not of biological relevance, because the migration of E1cd during gel filtration indicates that it exists as a monomer in solution. Since the two molecules are very similar (see Materials and Methods), the discussion will refer only to chain 1.

Thermostability. Although the resolution of this structure is not sufficient to analyze details of the nonbonded interactions which may be important for thermostability, we can look at some of the larger scale features that others have noted may play a role. These include fraction of buried atoms, relative surface area, and lengths of loops (Chan et al., 1995; Russell et al., 1994). For its size, E1cd has a relatively large fraction of buried atoms (0.56) and relatively low surface area (0.87 of that expected). These values are both very similar to those of the hyperthermophilic enzyme aldehyde ferredoxin oxidoreductase and much more extreme than those of 30 nonthermophilic enzymes (Chan et al., 1995). With regard to loop size, the alignment of E1cd with its close homologs (Figure 2) shows that E1cd has the shortest C-terminal loops for all but one loop, and for that one E1cd is only one residue longer than the shortest ones. The only significant insertion in E1cd relative to the other sequences occurs just before β-strand 8. Whereas the true importance of these correlations is unclear, E1cd adds another example for where a low surface to volume ratio and short loops are correlated with thermostability.

Cellotetraose Binding. *(A) The Bound Ligand*. In the crystals of E1cd grown in the presence of 0.3 M cellobiose, instead of cellobiose molecules, the electron density suggests that a condensation product, cellotetraose, is bound to each E1cd molecule (Figure 4). Each glucosyl residue, numbered Glc1 to Glc4 from the reducing to the nonreducing end, has been fit to the density in a standard chair conformation ($^4C_1$). Although each residue's fit could be slightly improved if deviations from the standard chair conformation were allowed (for instance, a possible sofa conformation at the cleavage site), we have resisted this temptation because the resolution is not sufficient to justify such deviation. We speculate that the bound cellotetraose results from an enzyme-catalyzed condensation of two cellobiose units. Such reactions have been seen for other enzymes in this superfamily (Sinnott, 1990), and experiments with soluble E1 confirm that it also carries out this "reverse" reaction effectively. An incubation of 10 mg/mL E1cd with 300 mM cellobiose and 50 mM citrate (pH 5) showed significant amounts of cellotetraose, cellotriose, and glucose after 24 h. The cellotriose and glucose could result from cleavage of the newly formed cellotetraose.

The density for the ligand is not as strong as the surrounding active site atoms, and Glc3 has notably weaker density than the other three glucosyl units (Figure 4), signaling the presence of possible incomplete occupancy, disorder, and/or heterogeneity. With an active crystalline enzyme, the crystal should be composed of a mixture of states reflecting the on-enzyme equilibrium under the conditions of crystallization: some molecules having bound substrate (cellotetraose), some having bound product (one or two cellobiose molecules), and some having a covalent intermediate (one cellobiose covalently bound to Glu-282 and possibly a second noncovalently bound cellobiose). The observed density corresponding to cellotetraose suggests that this is the dominant form present. By analogy with lysozyme

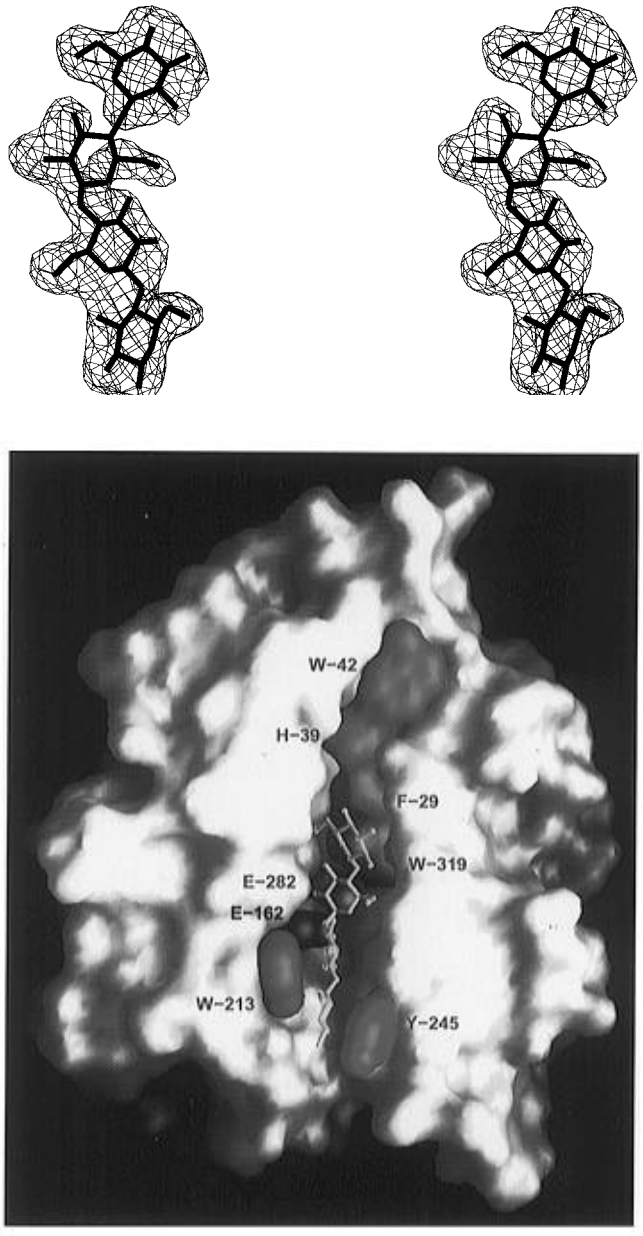

FIGURE 4: (a, top) Bound conformation of cellotetraose and its electron density. The electron density is from an annealed omit map. The contour level is at 2.5 times the rms electron density of the map. The nonreducing end of cellotetraose is at the top so the four glucosyl units are Glc4-Glc3-Glc2-Glc1 from top to bottom. According to a nomenclature for glucosyl unit binding sites relative to the cleavage site (Davies et al., 1995), Glc4, Glc3, Glc2, and Glc1 bind to sites −2, −1, +1, and +2, respectively. The weak electron density seen for Glc3 may be due to heterogeneity in the ligand (see text). The possible functional relevance of this heterogeneity is suggested by similarly weak density which is seen for the equivalent glucosyl unit in a covalent intermediate of a related glycosidase (White et al., 1996). (b, bottom) Molecular surface of the binding cleft of E1. The surfaces contributed by the observed and putative aromatic platforms (violet) and two key Glu residues (blue for the acid/base and red for the nucleophile) are highlighted. The width of the binding cleft above and below the catalytic center differs significantly. The narrowed binding cleft for Glc1 and Glc2 forces a twist in the polysaccharide chain, helping make the scissile glycosidic bond vulnerable (see text). The figure was prepared by GRASP (Nicholls, 1992).

studies which have shown that the glucosyl unit preceding the cleavage site binds poorly (Sacemski & Lienhard, 1974), we interpret the weak density for Glc3 compared to Glc1, Glc2, and Glc4 to suggest that the product complex is the

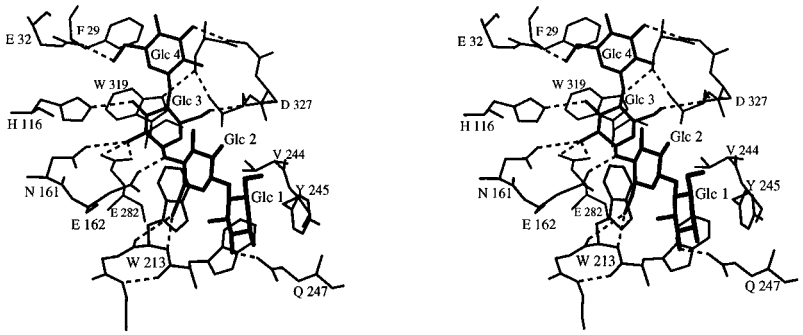

FIGURE 5: Cellotetraose recognition by E1cd. Side chains of the amino acid residues interacting with a cellotetraose molecule are shown. Dashed lines show probable hydrogen-bonding interactions of less than 3.2 Å. All of the labeled residues, except for Glu-32, are conserved in family 5−subfamily 1, suggesting a common substrate binding mode throughout the subfamily.

second most populated state and that it involves similar positions of Glc1, Glc2, and Glc4 and a less well ordered Glc3 site. The lack of significant difference density associated with Glu-282 leads us to believe that the covalent intermediate does not make up a significant fraction of the molecules. In crystals allowed to soak for 1 day in mother liquor without cellobiose the occupancy of the bound cellotetraose remained high, indicating that the ligand is locked into the active site by the crystal lattice. This lack of diffusion precludes simple measurement of the activity of the crystalline enzyme.

The conformation of cellotetraose approximates two standard cellobiosyl units (Gessler et al., 1994), Glc1-Glc2 and Glc3-Glc4, connected at an angle. The torsion angles (ϕ, ψ) at the three glycosidic bonds are (−99.9°, 95.9°), (−25.5°, 106.7°), and (−87.4°, 91.8°). Torsion angles ϕ and ψ are defined as ϕ ($O5_i-C1_i-O4_{i-1}-C4_{i-1}$) and ψ ($C1_i-O4_{i-1}-C4_{i-1}-C3_{i-1}$) (IUPAC−IUB, 1983). Although the torsion angles at the joining glycosidic linkage are not near those of crystalline cellotetraose (Gessler et al., 1994), they represent a low-energy conformation (Tayler et al., 1995). Conspicuously, Glu-162 and Glu-282, the expected acid/base and nucleophile of E1, and the five other known conserved residues of family 5 are involved in interactions focused on the recognition of Glc3 and the glycosidic bond that is attacked during cleavage, suggesting that these are the most crucial interactions for efficient catalysis. Descriptions of the interactions of these special residues and insights into the catalytic mechanism are presented below (see Catalytic Center and Enzyme Mechanism).

(B) Interactions Involved in the Extended Binding Site of Cellotetraose. The enzyme makes both hydrophobic and hydrogen-bonding contacts with all four residues of cellotetraose, and it appears that there may be additional binding sites for one to two more glucose residues beyond Glc4. As is commonly seen for polysaccharide binding enzymes [e.g., Rouvinen et al. (1990) and Spezio et al. (1993)], the hydrophobic face of each glucose unit interacts with an aromatic side chain that lines the active site cleft. Glc1, Glc2, Glc3, and Glc4 are interacting with Tyr-245, Trp-213, Trp-319, and Phe-29, respectively (Figures 4b and 5). Beyond Phe-29 sits Trp-42 which could play a similar role for a fifth or sixth glucosyl unit. The interactions with these aromatic platforms do not all seem to have optimal geometry. For Tyr-245 and Trp-213, the broad nonpolar surfaces are roughly coplanar with and appear ideal for packing against the hydrophobic, but weakly polar surface of the glucosyl units. However, the faces of Trp-319 and Phe-29 are parallel to each other and interact with the glucosyl units at about a 45° angle. Although it was not recognized due to inaccurate sequence alignments (Wang et al., 1993), Trp-319 is an eighth absolutely conserved residue in family 5. Other residues making close nonpolar interactions with the cellotetraose are Val-244, Trp-212, and the main chain of Ser-325.

Aside from the hydrogen-bonding interactions made by absolutely conserved residues to Glc3, an array of polar amino acids lining the active site cleft interact with glucosyl units. These interactions include Gln-247···O2-Glc1, Trp-213-N···O6-Glc2, Trp-213-O···O6-Glc2, Asp-327···O6-Glc3, Gly-326-N···O2-Glc4, and Glu-32···O6-Glc4 (Figure 5). Apparently no protein residues interact with the O5 atoms of the glucosyl units, although there are intermolecular hydrogen bonds from O3 of neighboring glucose units as are seen for crystalline cellotetraose (Gessler et al., 1994).

(C) Comparisons with Family 5 Enzymes. All five of the aromatic platforms (Tyr-245, Trp-213, Trp-319, Phe-29, and Trp-42) and each of the hydrogen-bonding residues (Gln-247, Asp-327, and Ser-325) are absolutely conserved in family 5−subfamily 1 (Figure 2), so that it is likely that most features of substrate binding will be well conserved within this subfamily. Also, although none of the residues making up the extended binding site are found throughout family 5, comparisons with the known structures for subfamilies show that the platforms for Glc1, Glc2, and Glc4 are well mimicked by nonequivalent residues: Tyr-245 that forms the platform for Glc1 is replaced by Phe-203 of CelC and Trp-259 of CelCCA; Trp-213 that interacts with Glc2 is replaced by the similarly oriented Tyr-176 of CelC and Trp-180 of CelCCA; and Trp-57 of CelCCA may replace Phe-29 as the platform for Glc4. Also, Trp-181 and Trp-287 of CelCCA may act as the additional platforms. These amino acid residues forming potential platforms are well conserved within their own subfamily members.

Catalytic Center and Enzyme Mechanism. (A) The Catalytic Center in E1cd. All eight absolutely conserved residues of family 5 (including Trp-319) are involved in interactions surrounding Glc3 and the glycosidic bond to be attacked during cleavage (Figure 6). Four of the residues hydrogen bond directly to the ligand (His-116···Glc3-O3, Asn-161···Glc3-O2, Glu-282···Glc3-O2, and Glu-162···Glc2-O4), Arg-62 and Tyr-240 hydrogen bond to the nucleophile Glu-282, and His-238 hydrogen bonds to the acid/base Glu-162. The Oϵ1 atom of the carboxylate of Glu-282 is within 3.5 Å of the anomeric carbon (C1) of Glc3 and is appropriately positioned for a nucleophilic attack. The

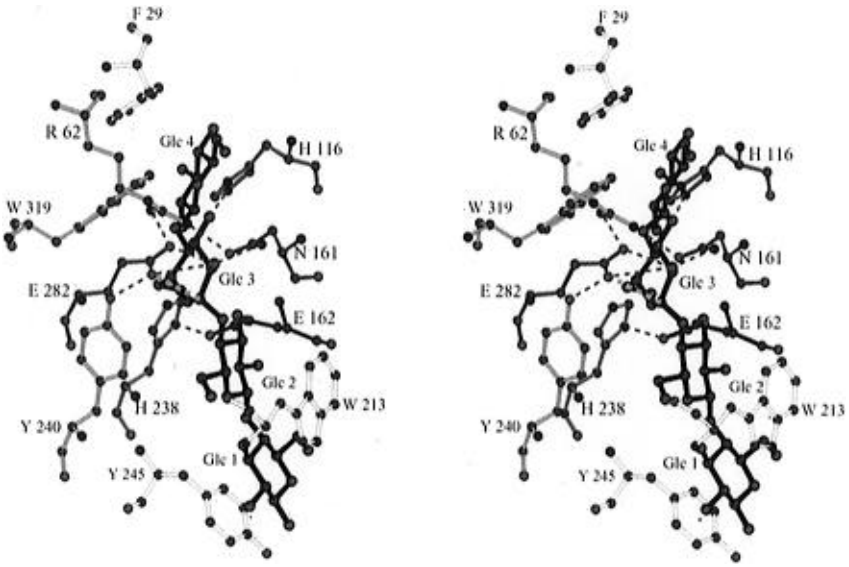

FIGURE 6: Interactions of the eight key active site residues. Stereo diagram of the E1cd−cellotetraose complex focusing on the geometry of the eight key residues. The coloring is by function: the putative nucleophile Glu-282 (red), residues aligning the nucleophile (pink), the acid/base (blue), residue aligning the acid/base (cyan), H-bond donor to Glc2-O2 (green), H-bond donor to Glc2-O3 (purple), and aromatic platform (yellow).

	62	116	161,2	238 240	282	319
5: E1cd	TIRLPY...	IILDR-H...	DLHNEP...	YSAHDY...	WLGEF-G...	WTFWS...
	91	137	182,3	324 326	397	446
1: 1CBG	AYRFSI...	PYVTLFH...	ITLNEP...	LGLNYY...	YITEN-G...	YFAWS...
	388 391	412	460,1	502,3	537	568
2: 1BGL	AVRCSH...	VVDEA.....	SLGNES...	ICPMY....	ILCEYAH...	GFVWD...
	41	80	126,7	203 205	233	273
10:2EXO	LVVA.....	LYGHT.....	DVVNEA...	VGFQSH...	RITEL-D...	VTVWG...
	31 33	53	92,3	168 170	232	275
17:1GHR	SMRLYA...	VVVGA.....	CVGNEV...	LMANIY...	VVSES-G...	TYIFA...

FIGURE 7: Putative functional counterparts in families 1, 2, 10, and 17 for the eight key residues in family 5. A structural alignment is presented for the six segments (from β-strands 2, 3, 4, 6, 7, and 8) containing the key residues. In this sequence alignment, all aligned residues have C_α atoms which overlay within 3.5 Å of their equivalent in E1cd and can be considered structurally equivalent residues. We use the term functional counterpart to designate residues which may or may not be structurally equivalent in terms of their main chain atoms but for which the side chains occupy a similar space and could carry out the same function. Putative functional counterparts are indicated by the color coding which matches the colors shown in Figure 6. Structures aligned are indicated by their PDB code: 1CBG (cyanogenic β-glucosidase from white clover), 1BGL (β-galactosidase from *E. coli*), 2EXO (β-1,4-glycanase Cex from *C. fimi*), and 1GHR (1−3,1−4-β-glucanase isoenzyme EII from barley). The assignment of functional equivalence for His-116 with His-80 of family 10 enzymes is confirmed by the complex recently reported by White et al. (1996) showing that His-80 of EXO hydrogen bonds to the O3 hydroxyl. In EXO, the replacement of Tyr-240 (E1cd) with His-205 (EXO) may be related to the absence of an Arg-62 (E1cd) counterpart as it maintains a positively charged group hydrogen bonding to the nucleophile. In GHR, the replacement of the His-116 (E1cd) interacting with the O3 atom with Tyr-33 may relate to its activity with β(1→3)-linked polymers which do not have an equivalent free hydroxyl.

eighth conserved residue, Trp-319, contacts the C3, C4, and C5 atoms of Glc3. These eight residues are present at the carboxy-terminal ends of β-strands 2, 3, 4, 6, 7, and 8, making the whole fold fairly directly involved in catalysis.

Our results clearly define important roles for Asn-161 in recognizing the C2 hydroxyl, His-116 in recognizing the C3 hydroxyl, Arg-62 and Tyr-240 in orienting and activating Glu-282, and His-238 in orienting Glu-162. In the E1cd structure, the hydrogen-bonding environment of His-238 is not sufficient to define whether it is protonated or not, but the replacement of His-238 with Asn/Gln residues in most superfamily members (see Figure 7) suggests that it is not protonated and this neutral hydrogen bond helps position Glu-162 without lowering its pK_a.

Additional residues conserved in subfamily 1 include four buried polar residues which help to form the catalytic center: Asn-27, Asp-158, and Glu-201 interact with Arg-62, and Asp-114 interacts with His-116. In a few cases there is also high conservation of short stretches of residues immediately surrounding the catalytic center. These include regions around Phe-29, Trp-213, and Asp-327 (Figure 2), each of which involves a special turn structure with conserved Asn and Gly residues. We describe the interesting stretch Trp-212-Trp-Gly-Gly-Asn-Leu here. Both Trp residues are interacting with cellotetraose. C_α-Gly-214 is 4 Å away from C_α-Glu-162 and has ϕ, ψ angles of (80°, −7°). A side chain at position 214 would bump and displace Glu-162. C_α-Gly-215 is 3.9 Å from O-His-248 and 3.8 Å from N-His-248. Although Gly-215 does not have unusual ϕ, ψ angles, a C_β atom would bump those atoms and thus displace His-248. Asn-216 makes hydrogen bonds to the main chain atoms O-204 and N-218, allowing the threading of this irregular loop 5 back to the surface.

(B) Other Family 5 Structures. In the other two family 5 structures which were determined without bound ligands, the catalytic center is structurally well conserved. In CelCCA (Ducros et al., 1995) all eight key residues appear to be in similar positions, and in CelC (Dominguez et al., 1995) all but the Asn161-Glu162 equivalents are similar. As discussed in the CelC publication (Dominguez et al., 1995), these residues point away from the active site but are thought to adopt an E1cd-like conformation upon substrate binding. The putative functional importance of some of these conserved residues is provided by mutagenesis experiments on subfam-

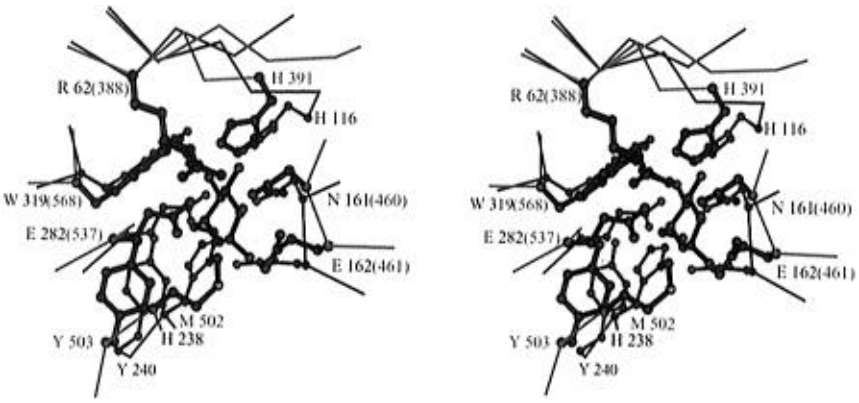

FIGURE 8: Illustration of the level of similarity leading to the assignment of putative functional counterparts. Shown are the overlaid active sites of E1cd (black) and BGL (violet). The positions of the side chains of His-116 and Tyr-240 of E1cd are remarkably well mimicked by nonequivalent residues His-391 and Tyr-503 in BGL, respectively. The substitution of His-238 (E1cd) for Met-502 (BGL) is surprising given the role of this residue as a neutral hydrogen bond donor to help position the acid/base catalyst [Glu-282 (461)]. Although the Sδ atom of Met is not a hydrogen bond donor, it is only very weakly negatively charged (ca 0.1 e^-) and thus could be considered to provide a largely nonpolar environment. We note that while this Met is conserved and may have a unique role in β-galactosidase sequences, β-glucuronidases, also in family 2, show a conserved Asn at this position.

ily 2 enzyme (Bortoli-German et al., 1995). In this work, mutations of the Arg-62, His-116, Glu-162, and His-238 equivalents to 13 other residues showed that all but one of the mutations of these key residues led to at least a 20-fold reduction in activity. The exception was the mutation of the His-116 equivalent to Phe which was reported to have 40% of wild-type activity.

(C) Functional Counterparts of Eight Key Residues in Other 4/7 Superfamily Members. Structural comparisons suggest that eight absolutely conserved residues and their interactions are of primary importance not only for family 5 but for all members of the superfamily. Except for the lack of an Arg-62 counterpart in family 10, and possibly the substitution of His-238 by Met in family 2, the eight conserved residues identified in E1cd have counterparts in the known structures from families 1, 2, 10, and 17 (Figure 7). Figure 8 illustrates two of the cases for which the side chains of nonequivalent residues are positioned so that they could carry out the same role. We suggest that these counterparts fulfill the same functional role even though they are not always identical residues and are not always even the structurally equivalent residues. In fact, only the three signature residues of the superfamily, the Asn-Glu duo and the nucleophilic Glu, are absolutely conserved in terms of position and sequence. The Trp-319 site is the next best conserved, being invariant in position and being either a Trp or a Phe. The relatively rare non-prolyl *cis*-peptide bond formed between Trp-319 and Ser-320 in E1cd (Figure 3) is also conserved in structures from families 1, 2, and 17. In family 10, a glycine (Gly-274 in EXO) has a *trans*-peptide bond and assumes ϕ, ψ angles near 130°, −180° to maintain the chain path.

Glycosyl hydrolase families 30, 35, 39, and 42 also have been classified in this superfamily, but structures for those families have not yet been determined. Those families' sequences also show appropriate residues well conserved within each family which can be lined up with each of the eight key residues, but because one must allow for possible changes in the identity and position of the key residues, the alignments are generally ambiguous. The identification of the five additional key residues does, however, place additional useful constraints on the sequence alignments. For instance, for family 35 we would revise the assignment of the nucleophile made by Henrissat et al. (1995). Using the sequence of human β-galactosidase as the representative for family 35, the assignment of Glu-339 (in the sequence PLSEAG) as the nucleophile allows the well-conserved sequence ASVNLY containing Asn-294 and Tyr-296 to correspond to His-238 and Tyr-240 of E1cd. The assignment of Glu-268 as the nucleophile (Henrissat et al., 1995) does not allow for reasonable equivalents to be assigned to His-238 and Tyr-240.

In addition to the nucleophile and the acid/base, whose importance for catalysis has been well documented (McCarter & Withers, 1994), the additional six key residues we identify here are virtually perfectly conserved within each of the families and correspond well with those residues which have been implicated as important by site-directed mutagenesis and chemical modifications in family 2 (Naider et al., 1972) and family 5 (Bortoli-German et al., 1995). On the basis of our amino acid sequence alignments, there are exceptions to the conserved residues. They are Ser-299 of *Escherichia coli* 6-phospho-β-glucosidase B replacing the otherwise conserved Asn-134 of CBG and similarly Glu-184 of XynA of *Filobasidium floriforme* replacing the conserved Gln-203 of 2EXO.

(D) Proposed Details of the Double Displacement Mechanism for the 4/7 Superfamily. The placement of the eight key residues in the E1cd−cellotetraose complex, combined with the results of the mutagenesis studies and the presence of putative functionally equivalent residues in all structurally known superfamily members, makes a compelling case that the complex we see is a catalytically relevant enzyme−substrate complex that can provide reliable insight into a catalytic mechanism common to this superfamily. Not only are the putative nucleophile (Glu-282) and the acid/base (Glu-162) well oriented to carry out their functions but, with only a slight rotation in the carboxylate of Glu-282, there is room in the complex for the Glc3 residue to tilt and move close enough to Glu-282 to form a covalent complex similar to that seen by White et al. (1996) for Cex, a family 10 enzyme.

The proposed mechanism includes the following steps (Figure 9): (1) The ES complex. This corresponds with the structure determined here. (2) The first transition state. Since Glu-282 is in an extended conformation already and is not likely to approach the anomeric carbon, the transition

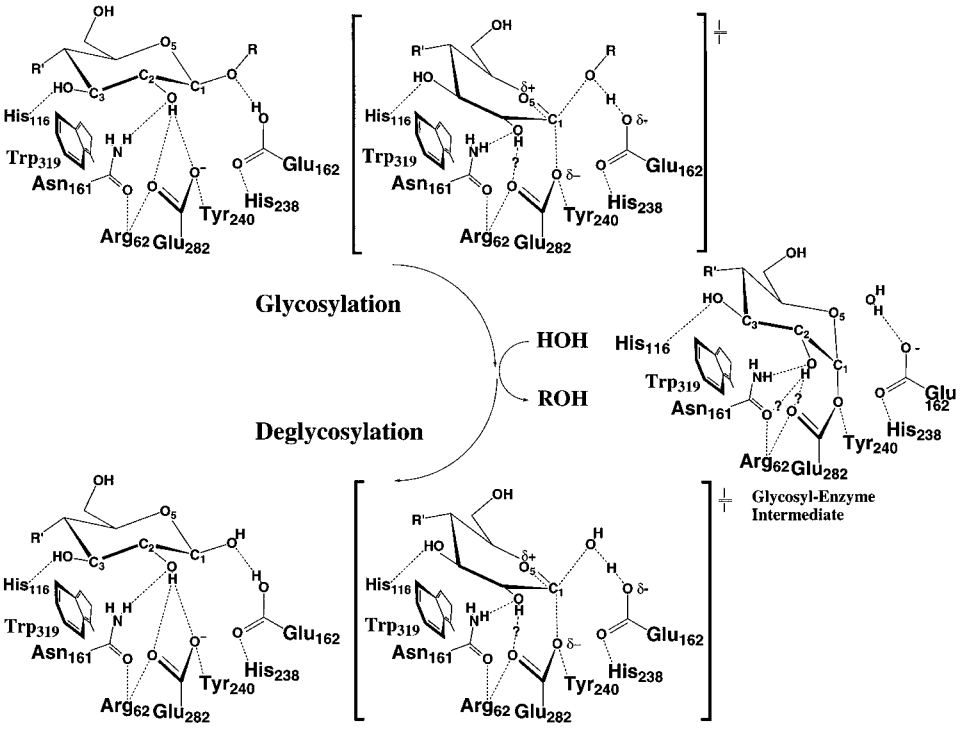

FIGURE 9: Proposed structurally detailed catalytic mechanism. A plausible covalent complex can be generated from the crystal structure by simply tilting Glc3 down to Glu-282. Although it is likely that the protein atoms move during catalysis, we do not have information about this movement and have drawn the mechanism assuming minimal motion of the substrate and no motion by the protein atoms. We have tried to maximize the structural accuracy of this scheme, and the view shown is simply a 90° clockwise rotation from that shown in Figure 6.

state for glycosylation can be generated by lowering Glc3 and converting it to a sofa conformation while maintaining the position of the glycosidic oxygen near Glu-162. Given the 4.4 Å distance from Glu-282-Oϵ1 to Glc2-O4, at the midway point the two partial bond distances would be about 2.2 Å. This geometry corresponds well to that expected for such a dissociative-type reaction with a high degree of oxocarbonium ion in the transition state (Tanaka et al., 1994), and the observed position of Glu-162 *trans* to the pyranose oxygen is stereoelectronically favorable for such an elimination reaction. At the transition state the shielding of the charge of Glu-282 would increase the acidity of Glu-162, enhancing the protonation of the leaving group. It has been noted that the C2-hydroxyl is the atom which undergoes the largest change in position during the chair−sofa transition and thus may play a key role in transition state binding (McCarter et al., 1992). An interesting possibility is that, in the transition state, the C2-hydroxyl might shift to interact with both the Oδ and Nδ atoms of Asn-161, making the role of this residue all the more important. (3) The glycosylated enzyme intermediate. In this structure, Glc3 is again in a chair conformation forming a α-glycosidic ester with Glu-282. The structure of the bound 2-fluoro analog (White et al., 1996) is the best model available for the true covalent intermediate, but the intimate interactions made by the 2-fluoro group in that structure mean that when a hydroxyl group is present, structural details may differ. The covalently bound Glc3 makes more extensive contact with Trp-319, and Glu-162 is unprotonated. (4) The transition state for deglycosylation. This is equivalent to the first transition state with the exception that water has replaced the glycosyl leaving group.

In cellulase E1, the apparent pK_a values for the nucleophile, Glu-282, and the acid/base, Glu-162, are near 4.5 and 8.0 (Thomas et al., unpublished). These pK_a values are also similar to those measured for other 4/7 superfamily members [e.g., Kempton and Withers (1992) and Tull and Withers (1994)]. The structural feature most likely related to the high pK_a of the acid, Glu-162, is its close approach to the nucleophile, Glu-282. The nearest oxygen atoms are only 3.4 Å apart, so a charged Glu-282 would greatly favor the protonated form of Glu-162. This close association is distinct from the textbook paradigm which shows these residues on opposite sides of the substrate. Also, since formation of the covalent intermediate would extinguish the charge of the Glu-282, it would cause the pK_a of Glu-162 to be much lower in the glycosylated enzyme, allowing Glu-162 to effectively function as a base in the deglycosylation half-reaction. This provides a plausible explanation for the puzzling observation that for enzymes in this family the pK_a of the acid/base seems to be 3 pK_a units different for the glycosylation and deglycosylation half-reactions (Kempton & Withers, 1992).

Structurally Similar Proteins. In addition to the structures of other proteins of known 4/7 superfamily members, the E1cd structure can be compared to all proteins of known structure to pick up more distant homologs. In such comparisons of proteins without significant sequence similarity, the assignment of divergent or convergent evolution is uncertain, and generally considerations of the level of structural, active site, and functional similarities are used to put forth a hypothesis of divergence or convergence [e.g., Matthews et al. (1981) and Orengo et al. (1994)]. E1cd is an especially interesting case, since the $(\alpha/\beta)_8$ barrel fold is very common and evolutionary relationships have been difficult to assign (Farber & Petsko, 1990). The $(\alpha/\beta)_8$ barrel has been described as a superfold because it occurs in 10 groups of enzymes having neither sequence nor functional similarities (Orengo et al., 1994).

Table 3: Structurally Similar Proteins

protein	PDB code	DALI[a] (rank)	SUPNEU[b] (rank)	ref
cellulase CelC*	1cec		8.2 (−)	Dominguez et al., 1995
cyanogenic β-glycanase*	1cbg		6.4 (−)	Barrett et al., 1995
β-galactosidase*	1bgl		5.8 (−)	Jacobson et al., 1994
xylanase*	1xys		4.9 (−)	Harris et al., 1994
β-amylase	1byb	15.5 (1)	4.8 (1)	Mikami et al., 1994
xylan xylanohydrolase*	1xas		4.5 (2)	Derewenda et al., 1994
1−3,1−4-β-glucanase*	1ghr	14.2 (2)	4.3 (3)	Varghese et al., 1994
exo-1,4-β-D-glycanase*	2exo	11.9 (4)	4.2 (4)	White et al., 1994
pyruvate kinase	1pkn	11.7 (6)	3.5 (5)	Larsen et al., 1994
cyclodextrin glycosyltransferase	1cdg	6.6 (25)	3.4 (6)	Lawson et al., 1994
α-amylase	6taa		3.3 (7)	Swift et al., 1991
narbonin	1nar	12.1 (3)	3.1 (8)	Hennig et al., 1992
tryptophan synthase	1wsy	8.0 (14)	3.1 (9)	Hyde & Miles 1990
1,4-α-D-glucan maltotetrahydrolase	1amg	6.9 (23)	3.0 (10)	Morishita et al., 1995
trimethylamine dehydrogenase	2tmd	7.8 (17)	3.0 (11)	Barber et al., 1992
fructose-1,6-bisphosphate aldolase	1fba	7.9 (16)	2.9 (12)	Hester et al., 1991
α-amylase	1ppi	9.1 (9)	2.9 (13)	Qian et al., 1993
RUBISCO	5rub	6.4 (26)	2.8 (14)	Schneider et al., 1990
chitinase A	1ctn	11.8 (5)	2.7 (21)	Perrakis et al., 1994
cellobiohydrolase II	3cbh	8.2 (11)	2.1 (37)	Rouvinen et al., 1990
endocellulase E2	1tml	6.3 (27)	2.1 (38)	Spezio et al., 1993

[a] The score for DALI is the strength (*Z*-score) of structural similarity in standard deviations above expected as reported by DALI. [b] The score for SUPNEU is calculated by (TSS score − average TSS score)/standard deviations. TSS represents topological similarity score defined as TSS = global sequentiality × local sequentiality as reported by SUPNEU. Asterisks indicate 4/7 superfamily members.

Table 4: Structural and Sequence Similarity between E1cd and Other Glycosyl Hydrolases with an $(\alpha/\beta)_8$ Barrel Fold

			β only			α/β core		
family	protein	PDB	rmsd[a]	% ide[b]	no. res[c]	rmsd	% ide	no. res
5	cellulase CelC	1cec	1.14	28.7	108	1.88	21.6	199
2	β-galactosidase	1bgl	1.60	22.6	93	1.91	16.7	168
1	cyanogenic β-glycanase	1cbg	1.57	17.3	81	2.78	14.3	154
10	exo-1,4-β-D-glycanase	2exo	2.18	12.3	81	3.02	15.9	164
17	1−3,1−4-β-glucanase	1ghr	1.56	14.7	73	3.10	15.6	154
14	β-amylase	1byb	1.55	8.8	68	2.55	8.4	178
18	chitinase A	1ctn	1.97	6.0	84	3.06	8.0	150
13	α-amylase	6taa	2.21	6.9	58	2.84	9.2	141
6	cellulase E2	1tml	2.47	4.0	70	3.25	5.6	125

[a] rmsd is the positional root mean square deviation of superimposed C_α atoms in angstroms. [b] % ide is the percentage of sequence identity over [c]the equivalenced positions (no. res). Overlays were initiated either by using amino acid residues on the eight β-strands (β only) or by using the eight β-strands and eight α-helices (α/β core) and then expanding these superpositions as far as possible to adjacent chain segments. The glycosyl hydrolases may be categorized as follows: (i) known 4/7 family is families 1, 2, 5, 10, and 17; (ii) a putative homologue based on structural similarity is family 14; and (iii) uncertain because of marginal functional, active site, and structural similarities are families 6, 13, and 18.

Structural similarity searches against the Protein Data Bank using both the DALI (Sander & Schneider, 1991) and SUPNEU (Diederichs, 1995) algorithms revealed that all top scorers were $(\alpha/\beta)_8$ barrel proteins (Table 3). As expected, among the most similar structures to E1 are the structurally known 4/7 superfamily members (Table 3). Using more traditional comparison statistics (Table 4), these enzymes are seen to have 14−22% amino acid sequence identities with E1cd over equivalent positions and conserve the key catalytic residues Asn-161, Glu-162, and Glu-282 (Dominguez et al., 1995; Jenkins et al., 1995).

β-Amylase. The very high structural similarity of E1cd with the family 14 enzyme β-amylase (Table 3) agrees with the observation of Wiesmann et al. (1995), who have suggested its homology with 4/7 superfamily members despite the fact that it has an inverting catalytic mechanism. A structural comparison of the β-amylase−maltotetraose complex (Mikami et al., 1994) and E1cd−cellotetraose molecules reveals that their corresponding substrates bind to the enzymes with the same directionality and at similar locations, and the key glutamic acids are at equivalent or adjacent positions. Inspection of the β-amylase structure reveals that the lack of a β-bulge on strand 7 places the key carboxylic acids ∼4 Å further apart, making room for the nucleophilic water molecule (McCarter & Withers, 1994). Other functional aspects are not similar between these two enzymes: the key asparagine residue (Asn-161 for E1cd) is missing in family 14 enzymes, and the α-linkages of amylose cause different faces of glucosyl units to interact with the cavity surface. As the very high structural/functional similarity seems sufficient to indicate homology, this appears to be an example where active site and mechanism have evolved more rapidly than global structure. We note that the recent conversion by mutagenesis of T4 lysozyme from an inverting to a retaining enzyme emphasizes the ease with which a mechanism can change (Kuroki et al., 1993; Wang et al., 1994). Consistent with the distinct mechanisms, the level of sequence identity between E1cd and β-amylase is considerably lower than among the 4/7 superfamily members (Table 4). What remains unusual about this case is that the extent of the structural changes is less than would be expected, given the extent of the sequence changes (Chothia & Lesk, 1986).

Discerning More Distant Relationships. In the SUPNEU results, pyruvate kinase is the highest scoring, clearly functionally unrelated enzyme and thus defines a natural cutoff for what might be considered as notably similar. However, in the DALI results two other proteins, narbonin, a seed storage protein, and chitinase A, a glycosyl hydrolase from family 18, score higher than pyruvate kinase, and narbonin even scores above the superfamily member exo-1,4-β-D-glycanase (Table 3). Although the scores are not much higher than that of pyruvate kinase, the scores seen for these two proteins in DALI, especially for the chitinase which is functionally related, could easily be used to support arguments for a homologous relationship between these proteins and E1cd. The scoring differences between SUPNEU and DALI may arise from DALI's employment of an elastic similarity scoring system and SUPNEU's use of a rigid body scoring system. But regardless of the origin of the differences, they emphasize that results are sensitive to the definition of structural similarity used and that discretion should be exercised in making conclusions based on one kind of comparison.

Just below pyruvate kinase in the SUPNEU results are the family 13 glycosyl hydrolases, α-amylase and cyclodextrin glycosyltransferase. Consistent with this result, structural/functional similarity arguments have been used to suggest that these enzymes are also homologs of the 4/7 superfamily (Wiesmann et al., 1995). However, for these enzymes, especially cyclodextrin glycosyltransferase, there are again discrepancies in the scoring of DALI and SUPNEU which affect the apparent strength of the argument. Given the marginal global structural similarity and the lack of active site similarity with E1cd (data not shown), we withhold judgment on the relationship of family 13 to the 4/7 superfamily.

The catalytic domains of the glycosyl hydrolases endocellulase E2 and CBHII (family 6) exhibit irregular $(\alpha/\beta)_8$ barrel motifs which are even less structurally similar to E1cd. In fact, the optimal overlay of these family 6 enzymes pairs their β-strand 1 with strand 2 of the 4/7 superfamily members. The previous overlay of the family 6 enzymes and family 5 member CelCCA resulted in the substrate-binding domain being rotated by 90° around the barrel axis (Ducros et al., 1995), but the overlay generated here shows that the active site clefts overlay well and the location of the putative proton donor, but not the base, is similar to the 4/7 superfamily members. As with the family 13 enzymes, the marginal structural similarities and the apparent differences in the active sites of family 6 enzymes confound assignment of their evolutionary relationship to E1cd. Interestingly, we note that although the sequence similarity between family 5 and families 1, 2, 10, and 17 is too weak to be discovered in the absence of structure (Henrissat et al., 1995), the level of sequence identity appears to be the main parameter that clearly distinguishes the 4/7 superfamily members which conserve the active site and mechanistic details from the possible other homologs which do not (Table 4).

ACKNOWLEDGMENT

We thank B. Ganem, D. Wilson, S. Savvides, C. Faerman, M. Pearson, M. Wisz, and Z. Deng for reading our manuscript. We also thank S. Ealick for allowing us to collect two sets of X-ray diffraction data for critical heavy-atom derivatives using his instrument and C. Ehrman for the confirmation of E1's transglycosylation reaction.

REFERENCES

Adney, W. S., Thomas, S. R., Nieves, R. A., & Himmel, M. E. (1994) U.S. Patent 5,366,884, Nov 22.

Baird, S. D., Hefford, M. A., Johnson, D. A., Sung, W. L., Yaguchi, M., & Seligy, V. L. (1990) *Biochem. Biophys. Res. Commun. 169*, 1035−1039.

Barber, M. J., Neame, P. J., Lim, L. W., White, S., & Mathews, F. S. (1992) *J. Biol. Chem. 267*, 6611−6619.

Barrett, T., Suresh, C. G., Tolley, S. P., Dodson, E. J., & Hughes, M. A. (1995) *Structure 3*, 951−960.

Bortoli-German, I., Haiech, J., Chippaux, M., & Barras, F. (1995) *J. Mol. Biol. 246*, 82−94.

Brunger, A. T., Kuriyan, J., & Karplus, M. (1987) *Science 235*, 458−460.

Chan, M. K., Mukund, S., Kletzin, A., Adams, M. W. W., & Rees, D. C. (1995) *Science 267*, 1463−1469.

Chothia, C., & Lesk, A. M. (1986) *EMBO J. 5*, 823−826.

Cudney, B., Patel, S., Weisgraber, K., Newhouse, Y., & McPherson, A. (1994) *Acta Crystallogr., Sect. D 50*, 414−423.

Davies, G. J., Tolley, S. P., Henrissat, B., Hjort, C., & Schulein, M. (1995) *Biochemistry 34*, 16210−16220.

Derewenda, U., Swenson, L., Green, R., Wei, Y., Morosoli, R., Shareck, F., Kluepfel, D., & Derewenda, Z. S. (1994) *J. Biol. Chem. 269*, 20811−20814.

Diederichs, K. (1995) *Proteins: Struct., Funct., Genet. 23*, 187−195.

Dominguez, R., Souchon, H., Spinelli, S., Dauter, Z., Wilson, K. S., Chauvaux, S., Beguin, P., & Alzari, P. M. (1995) *Nat. Struct. Biol. 2*, 569−576.

Ducros, V., Czjzek, M., Belaich, A., Gaudin, C., Fierobe, H.-P., Belaich, J.-P., Davies, G. J., & Haser, R. (1995) *Structure 3*, 939−949.

Farber, G. K., & Petsko, G. A. (1990) *Trends Biochem. Sci. 15*, 228−234.

Ferreira, L. M. A., Hazlewood, G. P., Barker, P. J., & Gilbert, H. J. (1991) *Biochem. J. 279*, 793−800.

Furey, W., & Swaminathan, S. (1996) in *Macromolecular Crystallography* (Cater, C., & Sweet, R., Eds.) Academic Press, Orlando, FL (in press).

Gessler, K., Krauss, N., Steiner, T., Betzel, C., Sandmann, C., & Saenger, W. (1994) *Science 266*, 1027−1029.

Gough, C. L., Dow, J. M., Keen, J., Henrissat, B., & Daniels, M. J. (1990) *Gene (Amsterdam) 89*, 53−60.

Grepinet, O., & Beguin, P. (1986) *Nucleic Acids Res. 14*, 1791−1799.

Hamlin, R. (1985) in *Methods in Enzymology* (Wyckoff, H. W., Hirs, C. H. W., & Timasheff, S. N., Eds.) Vol. 114, pp 416−451, Academic Press, Orlando, FL.

Harris, G. W., Jenkins, J. A., Connerton, I., Cummings, N., Leggio, L. L., Scott, M., Hazlewood, G. P., Laurie, J. I., Gilbert, H. J., & Pickersgill, R. W. (1994) *Structure 2*, 1107−1116.

Hennig, M., Schlesier, B., Dauter, Z., Pfeffer, S., Betzel, C., Hoehne, W. E., & Wilson, K. S. (1992) *FEBS Lett. 306*, 80−84.

Henrissat, B., & Bairoch, A. (1993) *Biochem. J. 293*, 781−788.

Henrissat, B., Callibaut, I., Fabrega, S., Lehn, P., Mornon, J.-P., & Davies, G. (1995) *Proc. Natl. Acad. Sci. U.S.A. 92*, 7090−7094.

Hester, G., Brenner, H. O., Rossi, F. A., Struck, D. M., Winterhalter, K. H., Smit, J. D. G., & Piontek, K. (1991) *FEBS Lett. 292*, 237−242.

Himmel, M. E., Adney, W. S., Grohmann, K., & Tucker, M. P. (1994) U.S. Patent 5,275,944, Jan 4.

Hopwood, D. A., Bibb, M. J., Chater, K. F., Kieser, T., Bruton, C. J., Kieser, H. M., Lydiate, D. J., Smith, C. P., Ward, J. M., & Schrempf, H. (1985) in *Genetic Manipulation of Streptomyces—A Laboratory Manual*, John Innes Foundation, Norwich, England.

Hyde, C. C., & Miles, E. W. (1990) *Bio/Technology 8*, 27−32.

IUPAC−IUB (1983) *Eur. J. Biochem. 131*, 5−7.

Jacobson, R. H., Zhang, X. J., DuBose, R. F., & Matthews, B. W. (1994) *Nature 369*, 761−766.

Jancarik, J., & Kim, S. H. (1991) *J. Appl. Crystallogr. 24*, 409−411.
Jenkins, J., Lo Leggio, L., Harris, G., & Pickersgill, R. (1995) *FEBS Lett. 362*, 281−285.
Kabsh, W., & Sander, C. (1983) *Biopolymers 22*, 2577−2637.
Kempton, J. B., & Withers, S. G. (1992) *Biochemistry 31*, 9961−9969.
Koshland, D. E., Jr. (1953) *Biol. Rev. 28*, 416−436.
Kraulis, P. (1991) *J. Appl. Crystallogr. 24*, 946−950.
Kuroki, R., Weaver, L. H., & Matthews, B. W. (1993) *Science 262*, 2030−2033.
Larsen, T. M., Laughlin, L. T., Holden, H. M., Rayment, I., & Reed, G. H. (1994) *Biochemistry 33*, 6301−6309.
Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993) *J. Appl. Crystallogr. 26*, 283−291.
Lawson, C. L., Van Montfort, R., Strokopytov, B., Rozeboom, H. J., Kalk, K. H., De Vries. G. E., Penninga, D., Dijkhuizen, L., & Dijkstra, B. W. (1994) *J. Mol. Biol. 236*, 590−600.
Lemaire, M., & Beguin, P. (1993) *J. Bacteriol. 175*, 3353−3360.
Matthews, B. W., Grütter, M. G., Anderson, W. F., & Remington, S. J. (1981) *Nature 290*, 334−335.
McCarter, J. D., & Withers, S. G. (1994) *Curr. Opin. Struct. Biol. 4*, 885−892.
McCarter, J. D., Adam, M. J., & Withers, S. G. (1992) *Biochem. J. 286*, 721−727.
Meinke, A., Gilkes, N. R., Kilburn, D. G., Miller, R. C., Jr., & Warren, R. A. J. (1993) *J. Bacteriol. 175*, 1910−1918.
Mikami, B., Degano, M., Hehre, E. J., & Sacchettini, J. C. (1994) *Biochemistry 33*, 7779-7787.
Mohagheghi, A., Grohmann, K., Himmel, M., Leighton, L., & Updegraff, D. (1986) *Int. J. Syst. Bacteriol. 36*, 435−443.
Morishita, Y., Matsuura, Y., Kubota, M., Sato, M., Sakai, S., & Katsube, Y. (1995) Protein Data Bank Entry 1amg.
Naider, F., Bohak, Z., & Yariv, J. (1972) *Biochemistry 11*, 3202−3207.
Nicholls, A. (1992) *GRASP: Graphical Representation and Analysis of Surface Properties*, Columbia University, New York.
Orengo, C. A., Jones, D. T., & Thornton, J. M. (1994) *Nature 372*, 631−634.
Otwinowski, Z. (1985) in *Data Collection and Processing* (Sawyer, L., Isaacs, N., & Bailey, S. S., Eds.) pp 56−62, SERC Daresbury Laboratory, Warrington, U.K.
Perrakis, A., Tews, I., Dauter, Z., Oppenheim, A. B., Chet, I., Wilson, K. S., & Vorgias, C. E. (1994) *Structure 2*, 1169−1180.
Qian, M., Haser, R., & Payan, F. (1993) *J. Mol. Biol. 231*, 785−799.
Read, R. J. (1986) *Acta Crystallogr., Sect. A 42*, 140−149.
Rouvinen, J., Bergfors, T., Teeri, T., Knowles, J. K. C., & Jones, T. A. (1990) *Science 249*, 380−386.
Russell, R. J. M., Hough, D. W., Danson, M. J., & Taylor, G. L. (1994) *Structure 2*, 1157−1167.
Sacemski, I. I., & Lienhard, G. E. (1974) *J. Biol. Chem. 249*, 2932.
Sack, J. S. (1988) *J. Mol. Graphics 6*, 224−225.
Sali, A., & Blundell, T. L. (1993) *J. Mol. Biol. 234*, 779−815.
Sander, C., & Schneider, R. (1991) *Proteins: Struct., Funct., Genet. 9*, 56−68.
Saul, D. J., Williams, L. C., Love, D. R., Chamley, L. W., & Bergquist, P. L. (1989) *Nucleic Acids Res. 17*, 439.
Schneider, G., Lindqvist, Y., & Lundqvist, T. (1990) *J. Mol. Biol. 211*, 989−1008.
Sheldrick, G. M. (1990) *Acta Crystallogr., Sect. A 46*, 467−473.
Sinnott, M. L. (1990) *Chem. Rev. 90*, 1171−1202.
Spezio, M., Wilson, D. B., & Karplus, P. A. (1993) *Biochemistry 32*, 9906−9916.
Swift, H. J., Brady, L., Derewenda, Z. S., Dodson, E. J., Dodson, G. G., Turkenburg, J. P., & Wilkinson, A. J. (1991) *Acta Crystallogr., Sect. B 47*, 535−544.
Tanaka, Y., Tao, W., Blanchard, J. S., & Hehre, E. J. (1994) *J. Biol. Chem. 269*, 32306−32312.
Tayler, J. S., Teo, B., Wilson, D. B., & Brady, J. W. (1995) *Protein Eng.* (in press).
Thomas, S. R., Laymon, R. A., Chou, Y. C., Tucker, M. P., Vinzant, T. B., Adney, W. S., Baker, J. O., Nieves, R. A., Mielenz, J. R., & Himmel, M. E. (1995) in *Enzymatic Degradation of Insoluble Polysaccharides* (Saddler, J. N., & Penner, M. H., Eds.) American Chemical Society, Washington, DC.
Tronrud, D. E., Ten Eyck, L. F., & Matthews, B. W. (1987) *Acta Crystallogr., Sect. A 43*, 489−501.
Tucker, M. P., Mohagheghi, A., Grohmann, K., & Himmel, M. E. (1989) *Bio/Technology 7*, 817−820.
Tucker, M. P., Grohmann, K., Mohagheghi, A., & Himmel, M. E. (1992) U.S. Patent 5,110,735, May 5.
Tull, D., & Withers, S. G. (1994) *Biochemistry 33*, 6363−6370.
Tull, D., Withers, S. G., Gilkes, N. R., Kilburn, D. G., Warren, R. A. J., & Aebersold, R. (1991) *J. Biol. Chem. 266*, 15621−15625.
Varghese, J. N., Garrett, T. P. J., Colman, P. M., Chen, L., Hoj, P. B., & Fincher, G. B. (1994) *Proc. Natl. Acad. Sci. U.S.A. 91*, 2785−2789.
Wang, Q., Tull, D., Meinke, A., Gilkes, N. R., Warren, R. A. J., Aebersold, R., & Withers, S. G. (1993) *J. Biol. Chem. 268*, 14096−14102.
Wang, Q., Graham, R. W., Trimbur, D., Warren, R. A. J., & Withers, S. G. (1994) *J. Am. Chem. Soc. 116*, 11594−11595.
White, A., Withers, S. G., Gilkes, N. R., & Rose, D. R. (1994) *Biochemistry 33*, 12546−12552.
White, A., Tull, D., Johns, K., Withers, S. G., & Rose, D. R. (1996) *Nat. Struct. Biol. 3*, 149−154.
Wiesmann, C., Beste, G., Hengstenberg, W., & Schulz, G. E. (1995) *Structure 3*, 961−968.
Withers, S. G., & Street, I. P. (1988) *J. Am. Chem. Soc. 110*, 8551−8553.
Wyman, C. E., Bain, R. L., Hinman, N. D., & Stevens, D. J. (1993) in *Renewable energy: sources for fuels and electricity* (Johansson, T. B., Kelly, H., Reddy, A. K. N., & Williams, R. H., Eds.) Island Press, Washington, DC.

BI9604439