ACADEMIC PRESS

Biochemical and Biophysical Research Communications 296 (2002) 161–166

BBRC

www.academicpress.com

Atomic resolution structure of the major endoglucanase from *Thermoascus aurantiacus*

F. Van Petegem,[a] I. Vandenberghe,[a] M.K. Bhat,[b] and J. Van Beeumen[a,*]

[a] *Laboratorium voor Eiwitbiochemie en Eiwitengineering, Universiteit Gent, B-9000 Gent, Belgium*
[b] *Institute of Food Research, Food Materials Science Division, Norwich Research Park, Colney, Norwich NR4 7UA, UK*

Received 24 June 2002

Abstract

The crystal structure of the major endoglucanase from the thermophilic fungus *Thermoascus aurantiacus* was determined by single isomorphous replacement at 1.12 Å resolution. The full sequence supports the classification of the protein in a subgroup of glycoside hydrolase family 5 for which no structural data are available yet. The active site shows eight critical residues, strictly conserved within family 5. In addition, aromatic residues that line the substrate-binding cleft and that are possibly involved in substrate-binding are identified. A number of residues seem to be conserved among members of the subtype, including a disulphide bridge between Cys212 and Cys249. © 2002 Elsevier Science (USA). All rights reserved.

Keywords: Endoglucanase; Cellulose degradation; Glycoside hydrolase family 5; Crystal structure; Atomic resolution

Cellulases (E.C. 3.2.1.4) are glycoside hydrolases that cleave the internal β-1,4 linkages in cellulose, the most abundant natural polymer in the biosphere. The polymer is also present in paper, textile, and raw material for processed foods, indicating the industrial importance of cellulolytic enzymes [1,2].

Glycoside hydrolases have been classified into at least 87 different families [3]. Cellulases occur in at least 14 of these, a number of which have been classified into different clans. The largest clan, clan GH-A, contains at least 14 families and encompasses enzymes with many different substrate specificities. Initially, families from this clan were proposed to share a common $(\alpha/\beta)_8$ barrel fold, three conserved residues, and a retaining cleavage mechanism [4,5]. In the three-dimensional structures, it is observed that 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, and therefore the superfamily is also termed the 4/7 superfamily.

Within glycoside hydrolase family 5, which mainly contains endo-1,4-glucanases, five different subtypes (A1–A5) were initially proposed [6,7]. From one subtype to another, the sequence identity is typically below 25%. A number of enzymes which can be confidently assigned to family 5 do not share enough sequence identity to be classified with any of these 5 subtypes and were therefore proposed to form a distinct subtype A6 [8]. Furthermore, β-mannanases and exo-1,3-glucanases also occur in family 5 and were also proposed to form distinct subtypes [9,10]. X-ray structures have been determined for endoglucanases of subtypes A1 [11], A2 [12–14], A3 [15], A4 [16], two β-mannanases [9,17], and an exo-1,3-glucanase [10]. In this study, we describe the full sequence and tertiary structure of the 35 kDa thermostable *Thermoascus aurantiacus* endoglucanase, which can be grouped in a distinct subtype for which no crystal structure has been described yet.

Materials and methods

Data collection and phasing. The endoglucanase from *T. aurantiacus* was purified as described [18]. Crystallization and preparation of an $SmCl_3$ derivative of the *T. aurantiacus* endoglucanase were performed as described [8]. Native data up to 1.12 Å resolution were measured in a high-resolution and a low-resolution pass at beamline X11 of the EMBL, Hamburg Outstation. Together with the derivative data, collected up to 2.4 Å at the same beamline, phases were determined in

* Corresponding author. Fax: +32-0-9-264-53-38.
E-mail address: jozef.vanbeeumen@rug.ac.be (J. Van Beeumen).

PII: S0006-291X(02)00775-1

essentially the same way as described [8], except that no second $HgAc_2$ derivative was used. All data were processed using the programs DENZO and SCALEPACK [19]. The concomitant SIR phases, obtained with the CCP4 program MLPHARE [20], were density modified and extended using DM [20]. The phases were improved further with the program Arp-wArp [21], which, after choosing the correct hand of the heavy atom configuration, resulted in autobuilding of 299 residues in both molecules from the asymmetric unit. Further statistics are shown in Table 1.

Sequencing. The sequence of the protein was determined by interpretation of the electron density maps. Ambiguities, mainly in the identity of Asp/Asn and Glu/Gln residues and of residues with disordered side chains, were solved by different methods. The full amino acid sequence was determined by protein and peptide sequence analysis as well as by gene sequencing. In the former case, Edman degradation in combination with tandem mass spectrometry was applied on peptides obtained by enzymatic cleavages using trypsin, N-Asp endoproteinase, chymotrypsin, and pronase. Partial acid hydrolysis and chemical cleavages with CNBr in 50% formic acid were performed after reduction and aminopropylation of the cysteine residues. The sequence information for the protein fragment D108-F284 was completed using PCR techniques in combination with DNA sequencing.

Refinement. The structure was refined using SHELXL [22] and the graphics program TURBO-FRODO [23]. The occupancies of side chains with double conformations were refined. Side chains for which no sufficient electron density was observed were omitted from the model. After initial refinement rounds, anisotropic B-factors were introduced and in the final stages hydrogen atoms were generated in the riding positions (but not output in the resulting coordinates). Using a cut-off of $F_0 > 4\sigma(F_0)$, the structure was refined to a final R-factor of 10.93%, and an R_{free}-factor, based upon 5% of the reflections, of 14.48%. Further refinement characteristics are summarized in Table 1. The stereochemistry of the model was checked with PROCHECK [24]. A Ramachandran plot shows that Asn60 is in a disallowed region of the plot in both molecules from the asymmetric unit, although the residue fits very well in the electron density. All other residues are in the most favoured and additionally allowed regions. Coordinates with PDB entry code 1H1N will be available upon publication.

Results and discussion

Sequence

Fig. 1 shows the full sequence of the endoglucanase protein. This 305 residue enzyme displays a large degree of sequence identity with an endoglucanase of *Talaromyces emersonii* (70.5%), *Aspergillus kawachii* (69.6%), and *Aspergillus niger* (69.3%). A BLAST search showed that significant sequence identity is present for at least 20 different sequences, including the endoglucanase II from *Trichoderma reesei* (25.8%) at the lower end of the list. Although they can be confidently assigned to glycoside hydrolase family 5, some of these sequences were

Table 1
Data collection, phasing, and refinement statistics

Data collection statistics	Native	$SmCl_3$
Wavelength (Å)	0.81	0.81
Resolution (Å)	15–1.12 (1.14–1.12)	20–2.40 (2.44–2.40)
Space group	$P2_12_12_1$	$P2_12_12_1$
Unit cell parameters (Å)	$a=75.77$ $b=84.70$ $c=89.10$	$a=75.86$ $b=85.05$ 89.39
$I/\sigma(I)$	34.29 (2.57)	43.79 (31.0)
R_{merge} (%)	8.6 (43.8)	5.2 (9.3)
Completeness (%)	98.9 (97.6)	98.4 (97.8)
No. of reflections		
Total	2,988,510	292,244
Unique	232,102	23,281
Phasing		
$R_{cullis:}$ acentrics/centrics	0.71/0.61	
Phasing power: acentrics/centrics	1.57/1.30	
FOM: before/after density modification	0.3707/0.4978	
Refinement		
Resolution (Å)	15–1.12	
R/R_{free} without cut-off (%)	13.27/17.10	
R/R_{free} with $F_0 > 4\sigma(F_0)$ cut-off (%)	10.93/14.48	
rmsd Bond lengths (Å)	0.015	
rmsd Angles (Å)	0.030	
# Atoms		
Protein	4736	
Solvent	976	
Average B factor (Å^2)		
Main chain	12.76	
Side chain	14.63	
Solvent	33.22	
Ramachandran plot		
Most favoured regions (%)	90.2	
Additional allowed (%)	9.4	

Values in parentheses indicate the highest resolution shell in the data collection section.

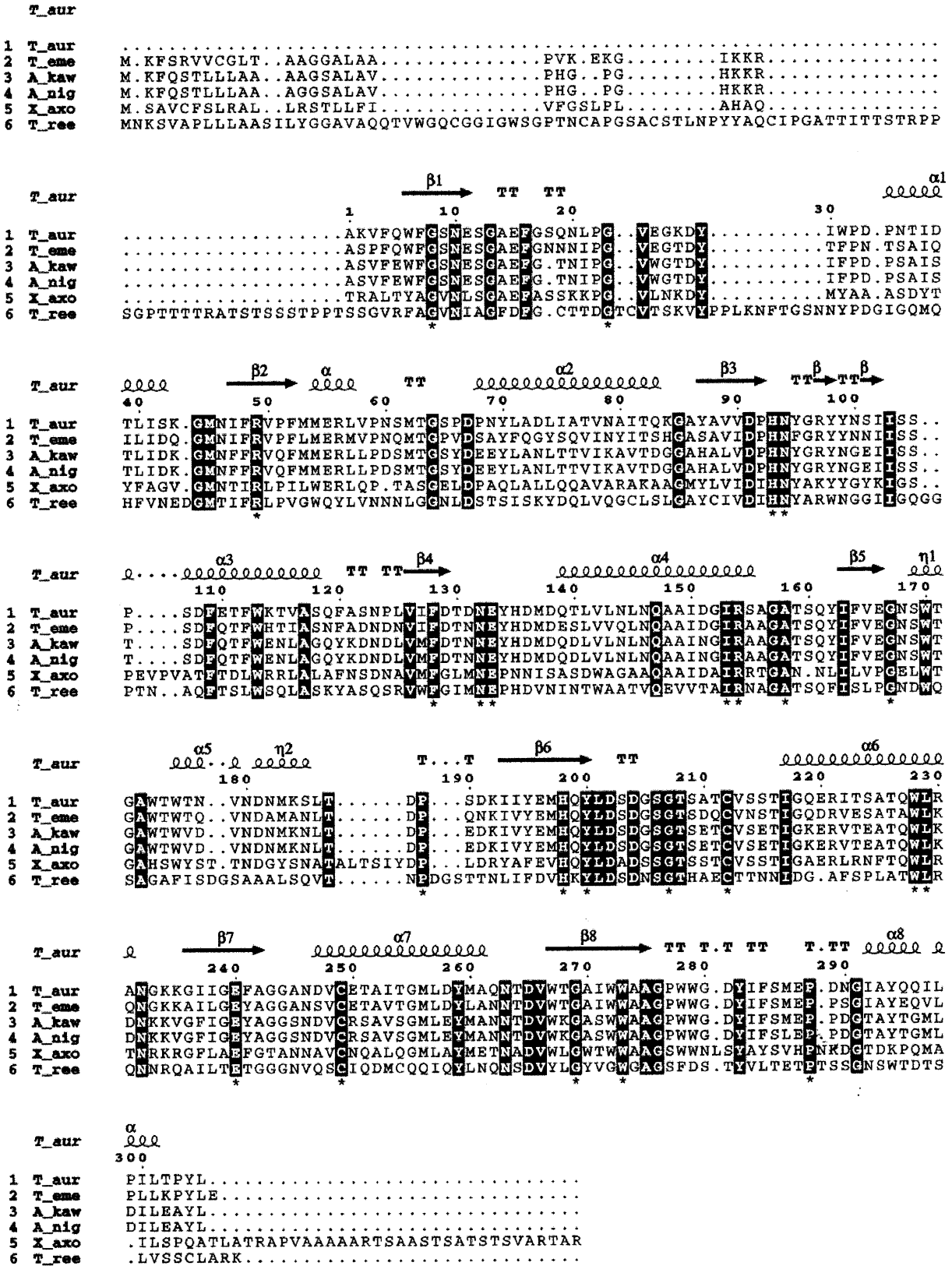

Fig. 1. Sequence alignment of the *T. aurantiacus* endoglucanase protein sequence and a selected subset of 5 endoglucanase sequences, showing significant sequence identity. T_aur: *T. aurantiacus* major endoglucanase; T_eme: *T. emersonii* endoglucanase; A_kaw: *A. kawachii* endoglucanase C; A_nig:*Aspergillus niger* endoglucanase B precursor; X_axo: *Xanthomonas axonopodis pv. citri str. 306* cellulase; T_ree: *T. reesei* endoglucanase II precursor. The secondary structure elements are for the *T. aurantiacus* crystal structure. Residues conserved between the six sequences are highlighted. Residues conserved between all 20 sequences analysed are indicated with an asterisk (*). The figure was generated with ESPript [26].

reported not to be assignable to any of the five originally recognized subtypes and proposed to form a distinct subtype [8].

General fold

Two molecules are present in the asymmetric unit of the endoglucanase. Both chains (A and B) can be superposed with an RMS deviation of 0.221 Å for the C_α atoms of residues 2–305. The N-terminal residue of molecule B was not visible in the electron density maps, probably due to disordering. All the following discussion is for molecule A from the asymmetric unit.

The general fold of the molecule can be described as a $(\beta/\alpha)_8$ barrel, typical for the 4/7 superfamily (Fig. 2A). An acidic cleft is visible at the C-terminal end of the β-barrel. Helix α_5 of the barrel has only one true α-helical twist. The N- and C-termini are close to each other, with 6.64 Å between the corresponding C_α atoms. In contrast to a lot of other family 5 structures, there are no distinct extra-barrel features, except for a small two-stranded β-sheet at the C-terminal end of β-strand 3 from the barrel. This differs from the situation in a subtype A3 endoglucanase from *Clostridium thermocellum* where an extra α-helical subdomain is present at the C-terminal end of the barrel [15]. In the structure of a subtype A-4 enzyme, an N-terminal α-helix is present [16], whereas in the three subtype A-2 structures, an extra two-stranded β-sheet contacts the N-terminal end of the $(\beta/\alpha)_8$ barrel [12–14].

The enzyme contains two cysteine residues (Cys212 and Cys249), implicated in a disulphide bridge. The bond connects the loop between strand β_6 and helix α_6 with helix α_7. These two cysteine residues appear to be conserved among the 20 sequences analysed (Fig. 1) and may therefore be a feature characteristic of this subtype.

Superposition with other family 5 endoglucanases

Structural similarity searches against the Protein Data Bank using the DALI server at http://www2.ebi.ac.uk/dali/ [25] revealed that the highest structural similarity is with the subtype A4 endoglucanase from *Clostridium cellulolyticum*. A corresponding structure-based superposition with members of subtypes A1–A4 is shown in Fig. 3. All numbering that follows is for the *T. aurantiacus* endoglucanase.

Eight different residues are completely conserved among the glycoside hydrolase family 5. These include the catalytic proton donor Glu133 and the neighbouring residue Asn132, lying at the C-terminal end of β_4, and the nucleophile Glu240 lying at the end of β_7. In the different structures, a cis peptide bond occurs after Trp273, a residue implicated in substrate-binding. Further conserved residues include His198 and Tyr200 which interact with the nucleophile Glu240, His93 which

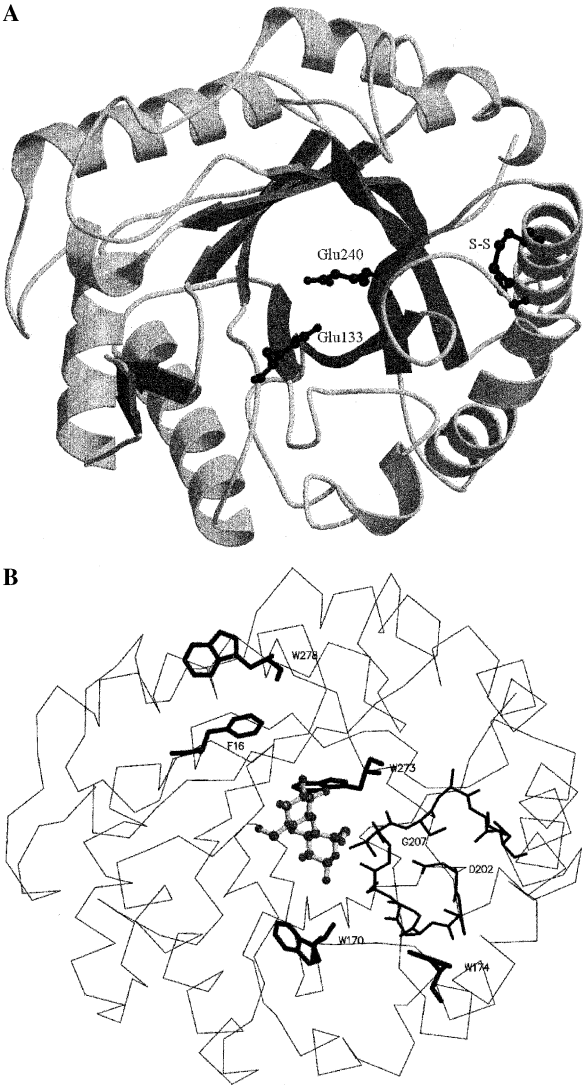

Fig. 2. (A) General fold of the *T. aurantiacus* major endoglucanase, showing secondary structure elements. The upper α-helix corresponds with α_1. The two catalytic residues, present at the C-terminal ends of β-strands 4 and 7, are shown in ball-and-stick representation. (B) Aromatic residues, possibly involved in the binding of substrate. Residues in the loop between β_6 and α_6 are also shown. Hydrogen bonds are not shown for clarity. The cellobiose shown in ball-and-stick is for the superposed subtype A3 endoglucanase from *C. thermocellum* occupying subsites −1 (bottom) and −2 (top). The figure was generated using MOLSCRIPT [27] and Raster3D [28].

contacts the sugar residue in subsite-1 [29], and Arg49 which forms hydrogen bonds with other strictly conserved residues (Asn132, His198, and Glu240). These eight strictly conserved residues are also spatially conserved in the *T. aurantiacus* endoglucanase. For the subtype A3 enzyme from *C. thermocellum*, it was observed that an induced fit occurs upon substrate-binding, including a repositioning of the proton donor Glu133 [29]. In this case, as well as in the other family 5 structures, the proton donor already has the same conformation as in the substrate-bound form.

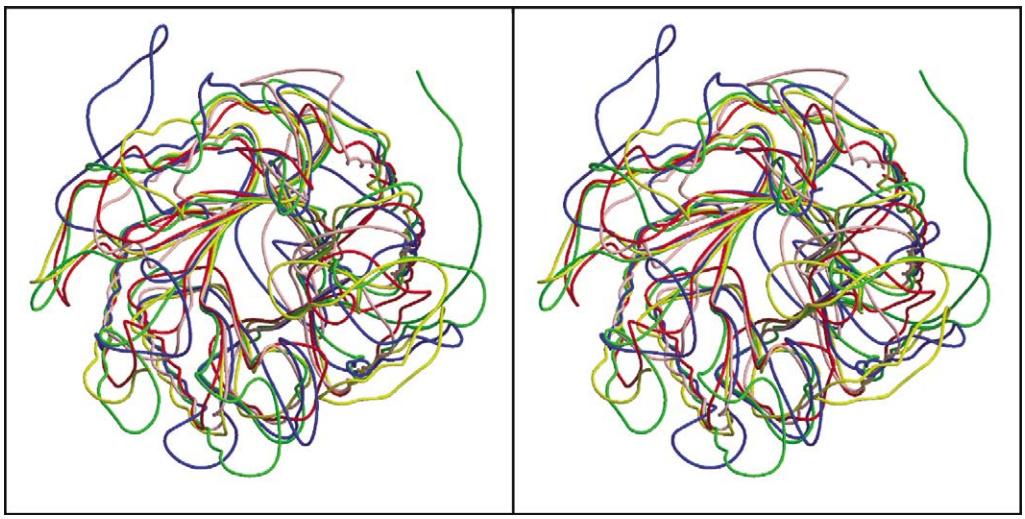

Fig. 3. Stereodrawing of a superposition of the *T. aurantiacus* endoglucanase (red) with family 5 endoglucanases belonging to different subtypes. Blue: *Acidothermus cellulolyticus*, subtype A1; pink: *Bacillus agaradhaerens*, subtype A2; yellow: *C. thermocellum*, subtype A3; green: *C. cellulolyticum*, subtype A4. The view is the same as in Fig. 2. The figure was generated with MOLSCRIPT [27] and Raster3D [28].

Other conserved residues and substrate-binding

Besides the eight strictly conserved residues, a number of other residues seem to be conserved among the 20 sequences that were analysed (Fig. 1). These include the two cysteine residues implicated in a disulphide bond. The exact role of all the other residues is not always clear, but some seem to be important for the general stabilization of the barrel structure. Phe128 and Ile153, for example, occur in a hydrophobic pocket near the N-terminal end of the barrel, while Arg154 forms a salt link with the highly conserved Asp188. In the other family 5 structures, these residues are also highly, but not strictly, conserved.

Of greater importance may be the loop containing Gly207, which is positioned near the substrate-binding site, between β_6 and α_6 (Fig. 2B). In a superposition with the substrate-bound form of the *C. thermocellum* endoglucanase C, it can be seen that this loop is very close to the substrate. An extensive hydrogen bonding pattern is present, including five hydrogen bonds with the side chain of the highly conserved Asp202. In only one of the 20 sequences analysed, this residue is an Asn. Moreover, the loop contains Cys212, implicated in the disulphide bridge. These interactions may be important for the proper positioning of the loop residues along the substrate-binding cleft.

Aromatic residues can stack against the hydrophobic faces of β-D-glucose. Analysis of the substrate-binding region by superposition with substrate-bound forms of family 5 endoglucanases suggests a number of residues that may be involved in this stacking (Fig. 2B). In subsite −1, Trp273 is strictly conserved among all family 5 members and was shown to interact with substrate. Phe16 is positioned well to stack against substrate in subsite −2, while Trp278 can stack in subsite −3. In subsites +1 and +3, Trp170 and Trp174 may fulfill these roles, while in subsite +2 no stacking residue can be proposed. In further support of this hypothesis, residues at positions 170 and 278 are always aromatic in all 20 sequences analysed, while the residue at position 16 is a Phe in 19 of the sequences. Complexes with substrate or inhibitor are, however, needed to unambiguously confirm this.

Acknowledgments

We gratefully acknowledge access to beamline X11 of the EMBL, Hamburg Outstation. F. Van Petegem is a research fellow of the Fund for Scientific Research-Flanders. J. Van Beeumen is indebted to the same institution for Grant 3G006896 and to the Bijzonder Onderzoeksfonds, Ghent University, for project 12050198. Also, we gratefully acknowledge the financial support from the BBSRC.

References

[1] M.K. Bhat, S. Bhat, Cellulose degrading enzymes and their potential industrial applications, Biotechnol. Adv. 15 (1997) 583–620.
[2] M.K. Bhat, Cellulases and related enzymes in biotechnology, Biotechnol. Adv. 18 (2000) 355–383.
[3] P.M. Coutinho, B. Henrissat, Carbohydrate-active enzymes server at URL http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html, 1999.
[4] B. Henrissat, I. Callibaut, S. Fabrega, P. Lehn, J.-P. Mornon, G. Davies, Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases, Proc. Natl. Acad. Sci. USA 92 (1995) 7090–7094.
[5] J. Jenkins, L. Lo Leggio, G. Harris, R. Pickersgill, Beta-glucosidase, beta-galactosidase, family A cellulases, family F xylanases and two barley glycanases form a superfamily of enzymes with 8-fold beta/alpha architecture and with two

conserved glutamates near the carboxy-terminal ends of beta-strands four and seven, FEBS Lett. 362 (1995) 281–285.
[6] P. Béguin, Molecular biology of cellulose degradation, Annu. Rev. Microbiol. 44 (1990) 219–248.
[7] Q. Wang, D. Tull, A. Meinke, N.R. Gilkes, R.A.J. Warren, R. Aebersold, S.G. Withers, Glu 280 is the nucleophile in the active site of *Clostridium thermocellum* CelC, a family A endo-Xβ-1,4-glucanase, J. Biol. Chem. 268 (1993) 14096–14102.
[8] L. Lo Leggio, N.J. Parry, J. Van Beeumen, M. Claeyssens, M.K. Bhat, R. Pickersgill, Crystallization and preliminary X-ray analysis of the major endoglucanase from *Thermoascus aurantiacus*, Acta. Cryst. D 53 (1997) 599–604.
[9] M. Hilge, S.M. Gloor, W. Rypniewski, O. Sauer, T.D. Heightman, W. Zimmermann, K. Winterhalter, K. Piontek, High-resolution native and complex structures of thermostable beta-mannanase from *Thermomonospora fusca*—substrate specificity in glycosyl hydrolase family 5, Structure 6 (1998) 1433–1444.
[10] S.M. Cutfield, G.J. Davies, G. Murshudov, B.F. Anderson, P.C. Moody, P.A. Sullivan, J.F. Cutfield, The structure of the exo-beta-(1,3)-glucanase from *Candida albicans* in native and bound forms: relationship between a pocket and groove in family 5 glycosyl hydrolases, J. Mol. Biol. 294 (1999) 771–783.
[11] J. Sakon, W.S. Adney, M.E. Himmel, S.R. Thomas, P.A. Karplus, Crystal structure of thermostable family 5 endocellulase E1 from *Acidothermus cellulolyticus* in complex with cellotetraose, Biochemistry 35 (1996) 10648–10660.
[12] G.J. Davies, M. Dauter, A.M. Brzozowski, M.E. Bjornvad, K.V. Andersen, M. Schulein, Structure of the *Bacillus agaradherans* family 5 endoglucanase at 1.6 AA and its cellobiose complex at 2.0 Å resolution, Biochemistry 37 (1998) 1926–1932.
[13] T. Shirai, H. Ishida, J. Noda, T. Yamane, K. Ozaki, Y. Hakamada, S. Ito, Crystal structure of alkaline cellulase K: insight into the alkaline adaptation of an industrial enzyme, J. Mol. Biol. 310 (2001) 1079–1087.
[14] V. Chapon, M. Czjzek, M. El Hassouni, B. Py, M. Juy, F. Barras, Type II protein secretion in gram-negative pathogenic bacteria: the study of the structure/secretion relationships of the cellulase Cel5 (formerly EGZ) from *Erwinia chrysanthemi*, J. Mol. Biol. 310 (2001) 1055–1066.
[15] R. Dominguez, H. Souchon, S. Spinelli, Z. Dauter, K.S. Wilson, S. Chauvaux, P. Béguin, P.M. Alzari, A common protein fold and similar active site in two distinct families of beta-glycanases, Nat. Struct. Biol. 2 (1995) 569–576.
[16] V. Ducros, M. Czjzek, A. Belaich, C. Gaudin, H.P. Fierobe, J.P. Belaich, G.J. Davies, R. Haser, Crystal structure of the catalytic domain of a bacterial cellulase belonging to family 5, Structure 3 (1995) 939–949.
[17] E. Sabini, H. Schubert, G. Murshudov, K.S. Wilson, M. Siika-Aho, M. Penttila, The three-dimensional structure of a *Trichoderma reesei* beta-mannanase from glycoside hydrolase family 5, Acta. Cryst. D 56 (2000) 3–13.
[18] M.K. Bhat, N.J. Parry, S. Kalogiannis, D.E. Beever, E. Owen, Thermostable cellulase and xylanase from *Thermoascus aurantiacus* and their potential applications, in: M.E. Himmel, J.O. Baker, J.N. Saddler (Eds.), Glycosyl hydrolases for biomass conversion, ACS Symposium series no. 769, American Chemical Society, Washington DC, 2001, pp. 204–219.
[19] Z. Otwinowski, W. Minor, Processing of X-ray diffraction data collected in oscillation mode, Meth. Enzymol. 276 (1996) 307–326.
[20] Collaborative Computational Project, Number 4, The CCP4 suite: programs for protein crystallography, Acta Cryst. D 50 (1994) 760–763.
[21] A. Perrakis, T.K. Sixma, K.S. Wilson, V.S. Lamzin, wARP: improvement and extension of crystallographic phases by weighted averaging of multiple refined dummy atomic models, Acta Cryst. D 53 (1997) 448–455.
[22] G. Sheldrick, T. Schneider, SHELXL: high-resolution refinement, Meth. Enzymol. 277 (1997) 319–343.
[23] A. Roussel, C. Cambillau, TURBO-FRODO, in: Silicon Graphics Geometry Directory, vol. 86, Silicon Graphics, Mountain View, CA, 1992.
[24] R.A. Laskowksi, M.W. McArthur, D.S. Moss, J. Thornton, PROCHECK: a program to check the quality of protein structures, J. Appl. Crystallogr. 26 (1993) 282–291.
[25] C. Sander, R. Schneider, Database of homology-derived protein structures and the structural meaning of sequence alignment, Proteins: Struct. Funct. Genet. 9 (1991) 56–68.
[26] P. Gouet, E. Courcelle, D.I. Stuart, F. Metoz, ESPript: multiple sequence alignments in PostScript, Bioinformatics 15 (1991) 305–308.
[27] P.J. Kraulis, MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures, J. Appl. Cryst. 24 (1991) 946–950.
[28] R.M. Esnouf, An extensively modified version of MOLSCRIPT that includes greatly enhanced coloring capabilities, J. Mol. Graphics 15 (1997) 132–134.
[29] R. Dominguez, H. Souchon, M. Lascombe, P.M. Alzari, The crystal structure of a family 5 endoglucanase mutant in complexed and uncomplexed forms reveals an induced fit activation mechanism, J. Mol. Biol. 257 (1996) 1042–1051.