RESEARCH PAPERS

Acta Cryst. (1996). D**52**, 7–17

Structure Determination and Refinement of the *Humicola insolens* Endoglucanase V at 1.5 Å Resolution

GIDEON J. DAVIES,[a]* GUY DODSON,[a] MADELEINE H. MOORE,[a] SHIRLEY P. TOLLEY,[a] ZBIGNIEW DAUTER,[b] KEITH S. WILSON,[b] GRETHE RASMUSSEN[c] AND MARTIN SCHÜLEIN[c]

[a]*Department of Chemistry, University of York, Heslington, York YO1 5DD, England,* [b]*EMBL Hamburg Outstation, c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany, and* [c]*Novo-Nordisk a/s, Novo Alle, 2880-Bagsvaerd, Denmark*

(Received 4 March 1995; accepted 7 July 1995)

Abstract

The structure of the catalytic core of the endoglucanase V (EGV) from *Humicola insolens* has been determined by the method of multiple isomorphous replacement at 1.5 Å resolution. The final model, refined with *X-PLOR* and *PROLSQ*, has a crystallographic R factor of 0.163 ($R_{\rm free}$ = 0.240) with deviations from stereochemical target values of 0.012 Å and 0.037° for bonds and angles, respectively. The model was further refined with *SHELXL*, including anisotropic modelling of the protein-atom temperature factors, to give a final model with an R factor of 0.105 and an $R_{\rm free}$ of 0.154. The initial isomorphous replacement electron-density map was poor and uninterpretable but was improved by the use of synchrotron data collected at a wavelength chosen so as to optimize the f'' contribution of the anomalous scattering from the heavy atoms. The structure of *H. insolens* EGV consists of a six-stranded β-barrel domain, similar to that found in a family of plant defence proteins, linked by a number of disulfide-bonded loop regions. A long open groove runs across the surface of the enzyme either side of which lie the catalytic aspartate residues. The 9 Å separation of the catalytic carboxylate groups is consistent with the observation that EGV catalyzes the hydrolysis of the cellulose $\beta(1\rightarrow4)$ links with inversion of configuration at the anomeric C1 atom. This structure is the first representative from the glycosyl hydrolase family 45.

1. Introduction

Cellulose, a polymer of $\beta(1\rightarrow4)$ linked glucose residues, is one of the most abundant naturally produced compounds on the planet. Enzymatic hydrolysis of cellulose is catalyzed by a number of cellulases, both cellobiohydrolases and endoglucanases, which have been isolated from various bacterial and fungal sources (for reviews see Béguin & Aubert, 1994; Wood, 1992; Henrissat, 1994). Efficient and complete hydrolysis of crystalline cellulose is a complex procedure requiring many different cellulases acting in a synergistic manner. The majority of fungal cellulases are multidomain proteins, consisting of a catalytically competent core structure linked to a cellulose-binding domain (CBD) *via* a flexible, often highly glycosylated, linker region (Knowles, Lehtovaara & Teeri 1987; Gilkes, Henrissat, Kilburn, Miller & Warren, 1991). The structure of a fungal CBD has been determined by NMR (Kraulis *et al.*, 1989). It is a compact wedge-shaped moiety which locates the catalytic domains in an appropriate position for the breakdown of crystalline cellulose. Some CBD's may also serve to physically disperse the crystalline cellulose fibres thus rendering them more amenable to catalysis (Gilkes *et al.*, 1993). Cellulases are a subset of enzymes that are known to hydrolyse glycosidic bonds: the glycosyl hydrolases. Sequence comparisons and hydrophobic cluster analyses show that the catalytic domains of glycosyl hydrolases fall into 45 distinct families, 11 of which contain enzymes with cellulase/xylanase activities (Henrissat, 1991; Henrissat & Bairoch, 1993). These 11 families are sometimes referred to as cellulase families A–K, but throughout this paper only the numerical classification will be used.

Thus far, structures of the catalytic domains of cellulases and xylanases from seven of the families have been published: cellobiohydrolase II (CBH-II) from *Trichoderma reesei* (Rouvinen, Bergfors, Teeri, Knowles & Jones, 1990) and endocellulase E2 from *Thermomonospora fusca* (Spezio, Wilson & Karplus, 1993), both from family 6; CelA from *Clostridium thermocellum* (Juy *et al.*, 1992) a representative of family 9; the endoglucanase V from *H. insolens* (this paper and, Davies *et al.*, 1993) from family 45; cellobiohydrolase I (CBH-I) from *T. reesei* (Divne *et al.*, 1994), family 7; a number of family 11 xylanases (Campbell *et al.*, 1993; Wakarchuk, Campbell, Sung, Davoodi & Yaguchi, 1994; Törrönen, Harkki & Rouvinen, 1994); the family 10 xylanases (Derewenda *et al.*, 1994; Harris *et al.*, 1994; White, Withers, Gilkes & Rose, 1994) and most recently CelCCA, a cellulase from family 5 (Ducros *et al.*, 1996). Although all of these structures facilitate catalysis *via* a similar acid/base mechanism involving

© 1996 International Union of Crystallography
Printed in Great Britain – all rights reserved

ISSN 0907-4449 © 1996

two or more aspartate or glutamate residues (for reviews on these mechanisms of catalysis see Koshland, 1953; Sinnott, 1990, McCarter & Withers, 1994), the structures of the cellulases from each family may be quite distinct. CBH-II and E2 share a variant of the (α/β)-barrel topology first observed in triosephosphate isomerase. In CBH-II, an exocellulase, the active site is located within an enclosed tunnel, whereas endoglucanase E2 has a much more open active site as might be expected for an endoglucanase. CelA has an (α/α)-barrel of 12 helices linked to an 'antibody-like' domain, the active site being located in a long groove across the surface of the enzyme. CBH-I has a fold related to that found in concanavalin A and the family 16 hybrid *Bacillus* endo $(1\rightarrow3{:}1\rightarrow4)$ glucanase (Keitel, Simon, Borriss & Heinemann, 1993). The two xylanases have different folds, the family 10 enzyme being a standard $(\alpha/\beta)_8$-barrel whilst the family 11 xylanases exhibit a predominantly β-strand structure which has been likened to the plant legume lectin fold found in the family 7 and family 16 enzymes (Törrönen *et al.*, 1994).

We have determined the structure of the 210-residue catalytic core domain of the endoglucanase V (EGV) from *H. insolens* at 1.5 Å resolution (Davies *et al.*, 1993). EGV is a cellulase from family 45 which includes the 43 kDa endocellulase B from *Pseudomonas fluorescens* (Gilbert, Hall, Hazlewood & Ferreira, 1990), the EGV from *T. reesei* (Saloheimo, Henrissat, Hoffrén, Teleman, & Penttilä, 1994) and the EGV from *Fusarium oxysporum* (Sheppard *et al.*, 1994). Screening identified two independent heavy-metal derivatives. The initial multiple isomorphous replacement (MIR) electron-density map was, however, of poor quality and did not allow chain tracing to proceed. Data were recollected at a synchrotron source at wavelengths chosen so as to optimize the f'' component of the anomalous scattering from the derivative heavy atoms. The resulting map was of very high quality and permitted an unambiguous and rapid tracing of the polypeptide chain. In addition to conventional refinement strategies, the structure has also been refined using *SHELXL*, including anisotropic modelling of the protein atomic displacement parameters. The structure determination, in addition to revealing the high-resolution structure of the first member of a new family of cellulases, clearly indicates the value both of optimized-wavelength data-collection strategies and of anisotropic temperature-factor modelling in protein structure determination.

2. Crystallization, data collection and processing

Crystals of the catalytic core domain of EGV were prepared as described previously (Davies *et al.*, 1993). The crystals are in space group $P2_1$ with cell dimensions a = 42.1, b = 51.7, c = 45.1 Å and β = 109.4°. There is a single molecule of EGV in the asymmetric unit. Two derivatives were prepared by soaking the crystals in the appropriate mother liquor with the addition of either 2 m*M* methyl mercury chloride (5 d) or 10 m*M* lutetium chloride (10 h). Conventional source data (not shown) were collected on an MAR research image plate with a Cu $K\alpha$ source and were processed with the *MOSFLM* suite of programs (Leslie, Brick & Wonacott, 1986). Native and derivative synchrotron source data were collected at the Photon Factory synchrotron facility using the Weissenberg method. Screenless Weissenberg data were collected with a camera radius of 286.5 mm, 16° oscillation ranges per image and a coupling constant of 4° mm^{-1}. Images were recorded on Fuji storage phosphors and scanned on a Fuji BA 1000 flat-bed scanner. The wavelength chosen was 1.000 Å, just below the Hg L_{III} and Lu L_I edges. 180° of data were collected with the derivative crystals aligned so as to allow the recording of Friedel mates on the same image. All Weissenberg data were processed with the *WEISS* program (Higashi, 1989). Further high-resolution native data for refinement purposes were collected from a single crystal at the EMBL Hamburg outstation on the wiggler beamline BW7B, using an MAR research detector. These data were processed with the *MOSFLM* program suite.

3. Structure determination and refinement

Hg positions were determined by manual inspection of the relevant isomorphous and anomalous difference Patterson maps. Initial phasing on the Hg derivative alone allowed the relative position of the Lu atoms to be determined by difference Fourier syntheses. Phasing and refinement cycles were performed with the *CCP4* implementation of the *MLPHARE* program. For details of the heavy-metal sites and phasing statistics see Davies *et al.* (1993). Phases were further improved using the *SQUASH* program (Cowtan & Main, 1993). The electron-density maps were skeletonized and the structure of EGV built using *O* (Jones, Zou, Cowan & Kjeldgaard, 1991). All further computing utilized the *CCP4* suite of programs unless otherwise stated (Collaborative Computational Project, Number 4, 1994).

3.1. *Conventional refinement*

The model was refined against the Weissenberg native data using standard procedures with *X-PLOR* (Brünger, Kuriyan & Karplus, 1987) and *PROLSQ* (Hendrickson & Konnert, 1981). Refinement converged to give a model with an R factor of 0.20, but there was a very poor agreement between the observed and calculated structure-factor amplitudes for the strong data, presumably resulting from a systematic error during the data processing. It is possible that this was due to an incorrect assumption about the dynamic range of the image-plate scanner. To overcome this problem with the Weissenberg data, refinement was continued using the Hamburg native data set. In order to 'release' a

set of structure-factor amplitudes for cross-validation analysis (Brünger, 1992) the model that had previously been refined against the Weissenberg data was subjected to an *X-PLOR* slow-cooling protocol, with a starting temperature of 4000 K. We have little experience if this protocol is sufficient to remove memory effects from the calculated coefficients, but repetition of this protocol had no further discernible effect on the absolute or relative values of the conventional or free R factors. Refinement was completed with manual inspection of $3F_o - 2F_c$ and $F_o - F_c$ maps and several cycles of *PROLSQ*. The final water model was built using the ARP procedure (Lamzin & Wilson, 1993), with concomitant monitoring of the $R_{\rm free}$ to prevent overfitting of the diffraction data.

3.2. *SHELXL refinement*

The protein model coordinates from the *X-PLOR/PROLSQ* refinement were used as a starting model for refinement with *SHELXL* (George Sheldrick, University of Göttingen). Structure-factor amplitudes and not intensities were used as input to this refinement. Cross-validation data identical to those used in *X-PLOR/PROLSQ* were used to monitor the progress of refinement in *SHELXL* and to assess the validity of various refinement strategies such as incorporation of solvent atoms and anisotropic modelling of protein atom temperature factors. Initially, conjugate-gradient least-squares minimization with isotropic temperature-factor refinement was carried out on residues 1–111 and 118–210, a total of 1549 protein atoms each with unit occupancy. Automatic water divining (Sheldrick & Schneider, manuscript in preparation) from difference Fourier syntheses combined with graphical checking for well defined spherical $F_o - F_c$ density greater than 3 standard deviations in peak height was used to obtain solvent sites which were gradually included with unitary occupancy in subsequent cycles of refinement. Anti-bumping distance restraints were implemented to prevent unrealistic interactions with water molecules and a bulk solvent correction was applied according to Babinet's principle (Langridge *et al.*, 1960). The criteria for retaining solvent sites were that, after refinement, they had acceptable thermal parameters, they displayed well defined spherical $2F_o - F_c$ density at least a standard deviation in height and that they were between 2.2 and 3.3 Å of plausible hydrogen-bonding partners. The gradual addition of 242 water molecules reduced the maximum and the minimum of the $F_o - F_c$ electron-density synthesis from 0.99 and −0.54 to 0.33 and −0.31 e Å^{-3}, respectively, as well as both R and $R_{\rm free}$ values. At this stage the reduction in both conventional R and $R_{\rm free}$, 0.036 and 0.015, respectively, which resulted from restrained anisotropic temperature-factor refinement was considered significant. Both 'rigid bond' and 'similar' anisotropic displacement parameter restraints for spatially close protein atoms were employed. Anisotropic displacement parameters of water atoms were restrained to approximate isotropic behaviour. A further reduction of 0.011 in R value (0.005 in $R_{\rm free}$) was obtained following refinement when 'riding' H atoms were inserted in calculated positions. After 198 waters had been included during the isotropic refinement, difference Fourier maps indicated additional electron density for residue Ala117 and distinct double conformations for the side chains of residues Ser15, Ser55, Asp121, Ser181 and Asn202. Vastly superior quality difference maps, for building multiple conformations of these residues and Ala117 were obtained after the temperature factors of all other atoms had been refined anisotropically and H atoms included. The improvement in the model when disorder was correctly accounted for was not only observed in the electron-density maps but also in a final $R_{\rm free}$ value of 0.154. Neither hydrogen scattering nor anisotropic refinement were considered appropriate for the multiply configured side chains. Refinement was completed after 244 water molecules had been included and the largest peak and hole in the difference Fourier synthesis were 0.24 and −0.24 e Å^{-3}. Without any further changes to the model all data were included in a final conjugate-gradient refinement which converged with R = 0.106 for all 28 611 data and R = 0.104 for 25 700 $F_o > 4\sigma F_o$ data. Two subsequent cycles of overlapped blocks of full-matrix least squares were carried out using all the data, including the cross-validation set, both in order to obtain estimated standard deviations in atomic parameters and to achieve the best final model.

4. Discussion of the structure

4.1. *Analysis of the structure solution*

The MIR electron-density map calculated with data collected using a conventional source was of a poor quality. Features indicative of protein structure were visible, Fig. 1(*a*), but sensible interpretation proved impossible. Density-modification procedures with the *SQUASH* program could not improve the map to a sufficient extent. Data for the native protein and for the Hg and Lu derivatives were, therefore, recollected at the Photon Factory synchrotron at wavelengths chosen to optimize the signal for the f'' component of the anomalous scattering of the heavy atoms. Low-resolution diffractometer data collected on Sm, Eu, Gd, Dy, Ho and Lu derivatives indicated that salts of all these lanthanide series of elements bound in a similar fashion. Lu was chosen for higher resolution synchrotron data collection as it allowed f'' optimization at a shorter wavelength than would have been possible for other members of the lanthanide series. With a large crystal-to-detector distance and wide oscillations short wavelengths were clearly preferable. Using the data collected at the Photon Factory synchrotron source the MIR map (figure of merit

0.64 acentric, 0.78 centric) was excellent, Fig. 1(*b*), and permitted a tracing of the entire polypeptide chain with the exception of residues 112–117 which form a loop between two of the strands. The final map at 1.5 Å is of high quality, Fig. 1(*c*).

It is possible to analyze the phase improvement achieved through the optimization of the X-ray wavelength. A plot of the phase differences between the: (*a*) MIR phases (conventional source), (*b*) MIR (synchrotron source), (*c*) MIR (synchrotron source *SQUASH*-improved) and the final calculated phases from the *PROLSQ/X-PLOR*-refined model is shown in Fig. 2. Collecting the data at optimized wavelength has reduced the overall mean phase error (all data to 2.5 Å) from 68 to 57°. The use of *SQUASH* to further improve the phases results in a further 5° improvement in the mean phase error, resulting in a set of phases with a mean phase error of approximately 52° for all data to 2.5 Å resolution. Since the quality of the electron-density map is also dependent on the correctness of the weights used in the Fourier synthesis, analysis of the phase changes alone is not sufficient to examine the map improvement. An analysis of the map quality, as given by the residue-by-residue map correlation (main-chain atoms only) between the MIR maps and the F_{obs},α_{calc} map, calculated from the final refined structure, is given in Fig. 3. There is a substantial improvement in the map calculated with the synchrotron optimized-wavelength data, not only in terms of the average map correlation, but also in the absence of extremely low correlations corresponding to main-chain breaks which occur frequently in the MIR (conventional) map. For all atoms there is an improvement from a total correlation coefficient of 0.4 to 0.69 between the MIR (conventional source) and MIR (optimized λ) maps. Clearly, substantial improvement in the quality of heavy-atom derived phases can be achieved through the use of optimized-wavelength strategies at synchrotron sources (for review see, Dauter, 1991). It is, however, difficult to attribute all the phase improvement to the wavelength alone, as the synchrotron data may be superior from a number of other perspectives.

4.2. *Quality of the final model structure*

The (EMBL, Hamburg) native data merged from 142 071 observations to give 28 562 unique reflections. These data are 97% complete to 1.48 Å with an overall R_{merge} ($\sum_{hkl}|I-I_{mean}|/\sum_{hkl}I_{mean}$) of 0.059 and a mean multiplicity of 4.9 observations per reflection, Table 1.

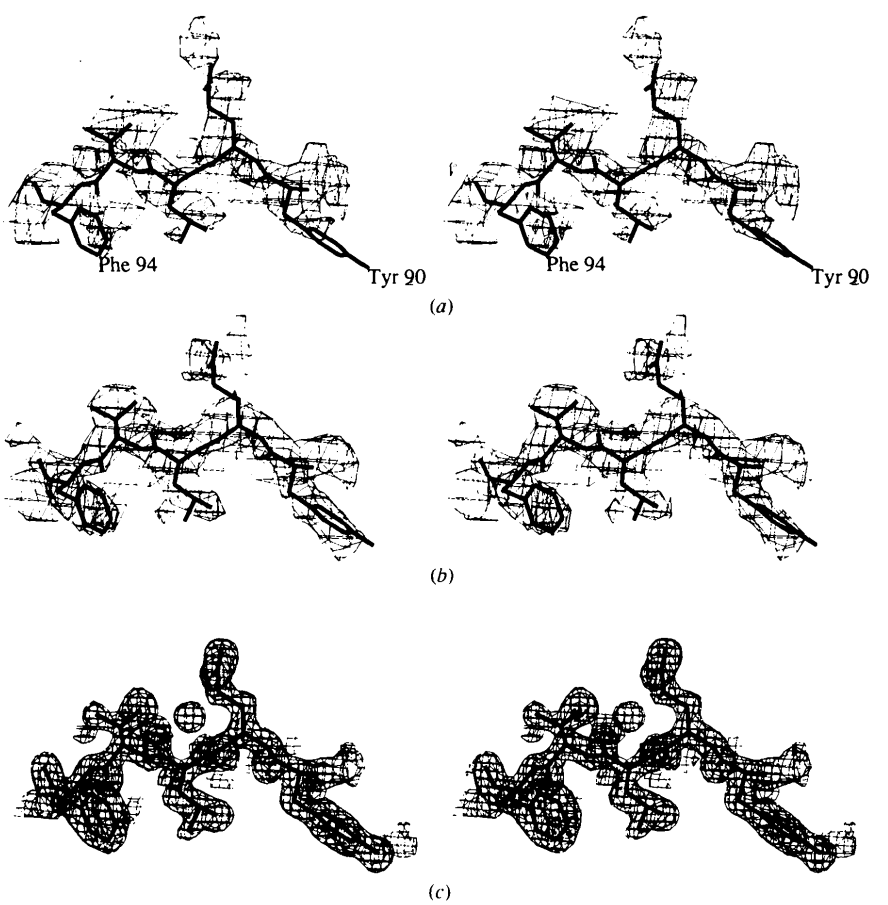

Fig. 1. Sections of the electron-density maps of EGV corresponding to residues 90–94. The three maps shown are (*a*) MIR map (conventional source), (*b*) MIR map (optimized λ) and (*c*) the 1.5 Å $2F_o-F_c$ map. The two MIR maps are contoured at a level of 1σ and the $2F_o-F_c$ map at a level of 0.4 e Å^{-3}.

The R_{merge} in the highest resolution shell (1.56–1.48 Å) is 0.189. The data have a temperature factor, calculated from the Wilson plot, of 23.4 Å^2, (Wilson, 1942). The *X-PLOR/PROLSQ* model structure, refined against these data, consists of 1546 protein atoms and 266 solvent water molecules. The crystallographic R factor is 0.163 with an R_{free} of 0.240 and deviations from stereochemical target values of 0.012, 0.037 and 0.010 Å for bonds, angles (1–3 bonding distance) and planes, respectively. One distinct double conformation was modelled for the side chain of one residue, Ser55. 89% of the non-glycine residues have conformational angles (φ,ψ) in the most favoured regions of the Ramachandran plot (Ramakrishnan & Ramachandran, 1965), Fig. 4. Seven of the non-glycine residues have conformational angles indicative of their involvement in left-handed turns: Cys11, Lys20, Phe35, Cys87, Lys175, Asn176 and Asn179. We believe that the involvement of two adjacent Cys residues (11/12 and 86/87) in the α-left region of the Ramachandran plot is extremely unusual in protein structures (Hutchinson & Thornton, personal communication). In addition to the unusual Cys11 conformation, Asp10, believed to be the catalytic base, has an unusual 'pep-flip' value (Jones *et al.*, 1991; Zou & Mowbray, 1994). Steric strain in residues involved in substrate binding and catalysis is not uncommon (for review see, Herzberg & Moult, 1991) and so it is possible that this conformation is required for its role in catalysis.

A plot of R factor against resolution (Luzzati, 1952) or the σ_A method of Read (1986) gives an upper estimate of coordinate error, assuming model errors to be solely responsible for the difference between F_{obs} and F_{calc}, of 0.18 Å for the *X-PLOR/PROLSQ*-refined model. The

Table 1. *Native data quality for the EMBL Hamburg native data used for refinement*

D_{min} (Å)	R factor	R_{cum}	Mean $I/\sigma I$	Completeness (%) (multiplicity)
4.66	0.045	0.046	60	99.9 (8.5)
3.31	0.053	0.051	53	100.0 (8.5)
2.70	0.059	0.053	42	100.0 (6.8)
2.34	0.071	0.054	31	99.8 (6.0)
2.10	0.089	0.057	27	99.6 (6.5)
1.91	0.094	0.058	20	99.2 (4.9)
1.77	0.072	0.058	14	98.1 (3.5)
1.66	0.094	0.058	10	96.4 (3.5)
1.56	0.134	0.058	7.0	96.6 (3.4)
1.48	0.189	0.059	4.6	85.1 (3.1)
Totals	—	0.059	21	97.0 (4.9)

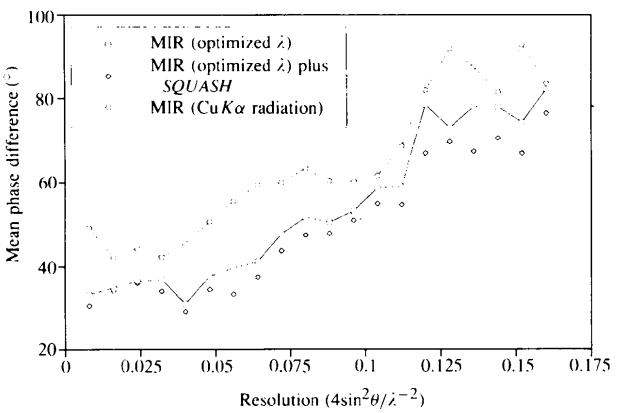

Fig. 2. Analysis of the phase difference, compared to the final calculated phases, for the MIR phases used in the EGV structure determination. The mean differences, to 2.5 Å resolution are 68, 57 and 52° for the conventional source MIR, optimized λ MIR and *SQUASH*ed MIR maps, respectively.

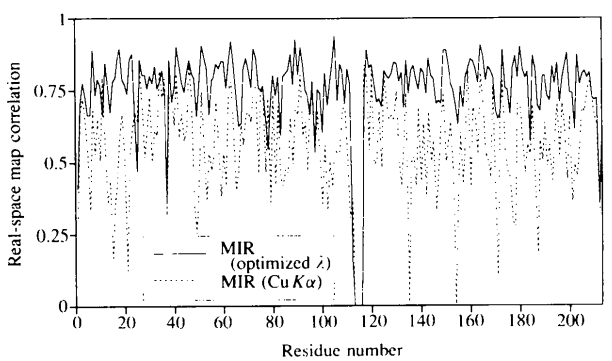

Fig. 3. A plot of the residue-by-residue map correlation (main-chain atoms only) for the two MIR maps compared to the final F_o,α_{calc} map at 2.5 Å resolution. The overall map correlation improved from 0.4 to 0.69 using the synchrotron optimized λ data.

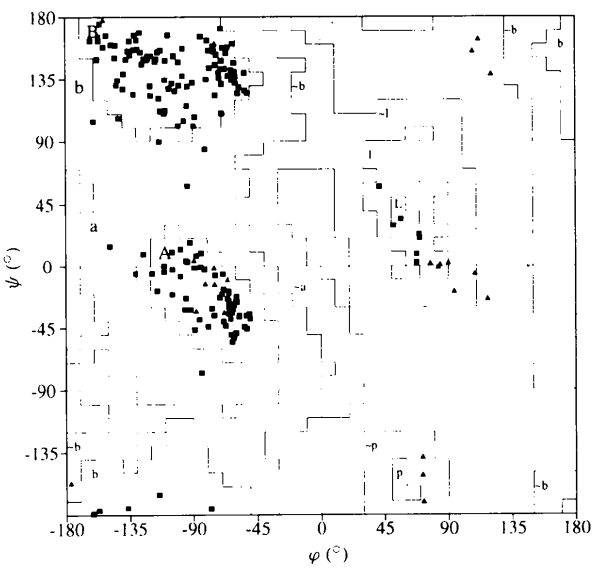

Fig. 4. Ramachandran plot (Ramakrishnan & Ramachandran, 1965) for the refined *H. insolens* EGV. Energetically favourable regions for the values of φ and ψ are indicated. Glycine residues are indicated by triangles and non-glycine residues by squares. This, and other, assessments of coordinate quality were assessed with the *PROCHECK* program (Laskowski, MacArthur, Moss & Thornton, 1993).

Table 2. *A summary of the stages of refinement with SHELXL*

Least squares		Hendrickson & Konnert conjugate gradient (CGLS)				Full matrix (LS)
Temperature-factor refinement		Isotropic		Anisotropic		Anisotropic
Positional refinement	All protein atoms included with unit occupancy			2 $conf^*_{sc}$ + H_{inc} †		2 $conf_{sc}$ + H_{inc}
No. of waters	0	0	242	242	244	244
Data/parameter ratio	4.2	4.2	3.6	1.6	1.6	1.7
(No. of parameters)	(6198)	(6198)	(7162)	(16112)	(16270)	(16270)
(Final no. of restraints)	(6134)	(6134)	(6487)	(19618)	(19805)	(19795)
(No. of data)	(25766)	(25766)	(25766)	(25766)	(25766)	(28611)
Overall R factor	0.276	0.215	0.156	0.120	0.105	0.106
(No. of all data)	(25766)	(25766)	(25766)	(25766)	(25766)	(28611)
Overall R factor	0.273	0.211	0.152	0.118	0.103	0.104
(No. of $F_o > 4\sigma F_o$ data)	(23125)	(23125)	(23125)	(23125)	(23125)	(25700)
Free R value (for 2845 data)		0.324	0.254	0.206	0.191	0.154

* 2 $conf_{SC}$, double conformations modeled for side chains of five residues and Ala117 included. † H atoms included in calculated positions.

average main-chain and side-chain B values are 24.1 and 32.4 Å^2, respectively. An analysis of the residue-by-residue main-chain and side-chain B values (with the exception of disordered residues 112–116) is given in Fig. 5.

In *SHELXL* it was considered appropriate to carry out anisotropic temperature-factor refinement despite the reduction in data/parameter ratio. A number of restraints in the nature of the anisotropic refinement were employed and the drop in both the conventional R and R_{free} values suggested that the refinement was meaningful. A summary of the *SHELXL* refinement is given in Table 2. An *ORTEP* plot of the atoms involved in, and close to, one of the disulfide bonds is shown in Fig. 6. For the five residues where double conformations were modelled the occupancies of any side-chain atoms in two different positions were tied to a free variable such that the sum of the two occupancies was unity. This refined the occupancies of the side-chain atoms to 0.6/0.4 for Ser15, 0.6/0.4 for Ser55, 0.7/0.3 for Asp121, 0.7/0.3 for Ser181 and 0.5/0.5 for Asn202. The second occupancy of Asp121 makes a steric clash (1.91 Å) with the single modelled position of His119. This side chain is also partially disordered but no discrete second conformation was visible in the electron-density maps. The single positions modelled for the side-chain atoms of His119 were, therefore, refined with an occupancy of 0.7. Two examples of residues exhibiting distinct static disorder are shown in Fig. 7. During refinement extra similar distance restraints were needed for the disordered Asp and Asn residues. The Cγ chiral volumes of leucines and the Cβ chiral volumes of isoleucines, threonines and valines decreased during refinement. Therefore, these residues also needed additional chiral volume restraints to prevent loss of tetrahedral geometry at the Cβ and Cγ atoms.

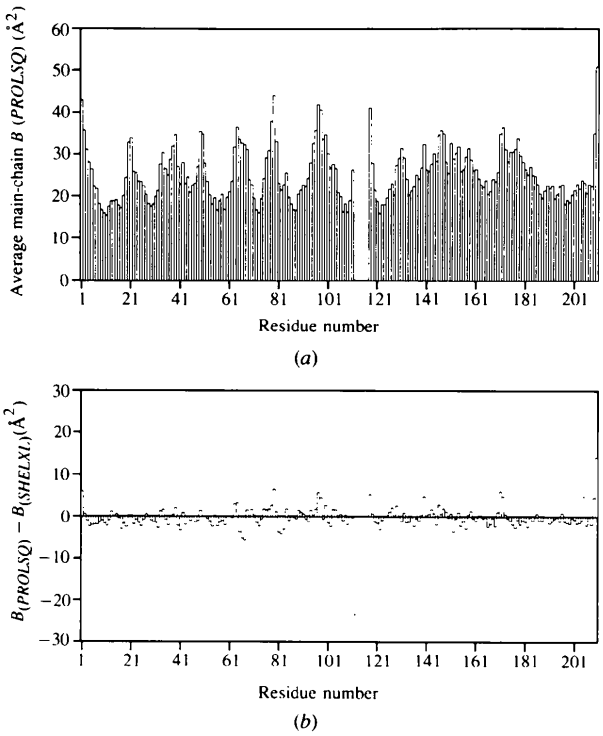

Fig. 5. (*a*) Average main-chain B values (Å^2) for the *X-PLOR/PROLSQ*-refined *H. insolens* EGV plotted as a function of residue number. Electron density for the loop from residues 112 to 116 is absent and these residues are, therefore, presumed to be disordered in the native EGV structure. (*b*) A plot of the difference in B factor between the *PROLSQ*- and *SHELXL*-refined models.

The main- and side-chain isotropic B values for the *SHELXL*-refined model are 24.9 and 35.7 Å^2 with r.m.s. differences betwen 1–2 bonded atoms of 1.4 and 3.1 Å^2, respectively. These are calculated by conversion from the anisotropic atomic displacement parameters obtained during the refinement. For comparative purposes, an analysis of the residue-by-residue main-chain *PROLSQ*-refined B values is given in Fig. 5 together with a plot of the difference between the *SHELXL* and *PROLSQ* B values. The two sets of B values are extremely similar with the biggest differences occurring, as expected, in regions of high mobility.

It is useful to compare the model obtained with this *SHELXL* refinement with that obtained with *X-PLOR/PROLSQ*. A summary of the protein model geometry is given in Table 3. The geometry is extremely similar between the two models, especially given the different weighting schemes and restraints used by the two programs. Two differences in geometry are clear. The first comes from an analysis of the standard deviation of the bond distances. If the standard deviation reflects the true error in the bond distance then the distribution of the standard deviations should have a mean of 0.0 and a σ of 1.0. The *SHELXL*-refined model reflects this theoretical distribution better than the *PROLSQ*-refined model. The positive value found with the *PROLSQ*-refined model probably reflects inadequate modelling of the scattering from the H atoms, causing a slight lengthening of the bond distances. If no modelling of the H-atom contribution is made the mean for this distribution increases to even greater values. We have observed this error in *PROLSQ* with many structures at moderately high (< 2 Å) resolutions (G. Davies, unpublished observations). The original *SHELXL*-refined model exhibited a significant error in the chiral volumes for many chiral centres (mostly Cα), with an r.m.s. deviation from the *PROLSQ* dictionary values of 0.223 Å^3 compared to the 0.037 Å^3 of the *PROLSQ*-refined model. This was corrected for by increasing the chiral volume restraints to twice the default value. Although the overall geometry improved with additional chiral volume restraints the final observed r.m.s. Δ on chiral volumes was still significantly higher than the target value in *PROLSQ*. This results entirely from the Cα chiral volumes of the proline residues which have a different target value, compared to other residues, in *SHELXL* but are treated equivalently to all other amino

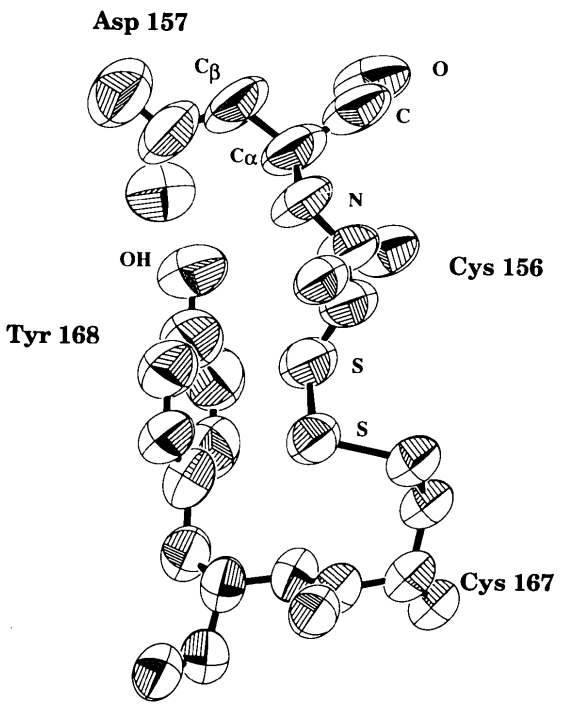

Fig. 6. An *ORTEP* plot of the 40% thermal ellipsoids involved in and adjacent to the disulfide bridge between residues Cys156 and Cys167. The S atoms in the disulfide link show relatively isotropic displacements whereas atoms of the adjacent, and more solvent exposed, residue Asp157 exhibit more significant anisotropic displacements. the main-chain atoms of Asp157 show displacements perpendicular to the peptide-bond plane.

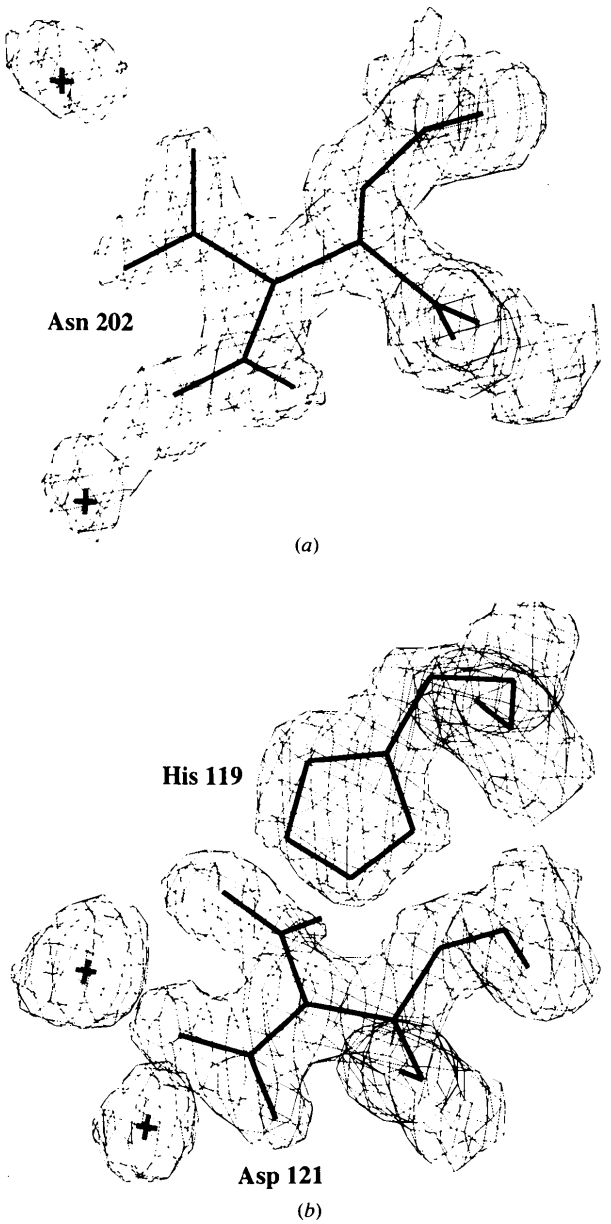

Fig. 7. $2F_o - F_c$ density for two disordered residues in EGV (*a*) Asn202 and (*b*) the catalytic residue Asp121. Both maps are contoured at a level of approximately 0.3 e Å^{-3}.

Table 3. *A comparison of the model geometry for the final PROLSQ- and SHELXL-refined models with PROLSQ target values given for reference*

Restraint	*PROLSQ* target value	Observed (*PROLSQ*)	Observed (*SHELXL*)
Bond and angles (Å)			
1–2 Bonds	0.020	0.012	0.012
1–3 Bonds	0.040	0.037	0.031
1–4 Bonds	0.050	0.038	0.053
Standard deviation of bond distances* (Å)			
Mean		0.297	0.10
σ		0.532	0.60
Planes and chiral volumes (Å, Å^3)			
Planes	0.020	0.010	0.025
Chirals	0.060	0.037	0.10
Torsion contacts (Å)			
Single torsion	0.3	0.160	0.156
Multiple torsion	0.3	0.260	0.268
Hydrogen bond	0.3	0.261	0.109
Angles (°)			
Planar	3	2.00	5.00
Staggered	15	15.83	15.22
Orthonormal	20	26.67	21.95

* If the standard deviation is a true estimate of the error then the distribution of the standard deviations should have a mean of 0.0 and a σ of 1.0.

acids in *PROLSQ*. Coordinate error for the *SHELXL*-refined model, calculated with the σ_A program (Read, 1986) is estimated to be 0.042 Å.

4.3. *Description of the structure*

With the exception of a disordered loop (residues 112–116) the entire polypeptide chain of EGV can be traced from residues 1 to 210. The overall shape of the molecule is that of a flattened sphere with dimensions 42 × 42 × 22 Å. The predominant structural feature is a six-stranded β-barrel domain, with an unusual topology containing both parallel and antiparallel strands. A seventh strand hydrogen bonds to one of the barrel strands, but may not formally be considered part of the barrel itself. The β-barrel falls into the '(6,10)' class (for reviews see Murzin, 1994; Murzin, Lesk & Chothia, 1994*a*,*b*), which so far includes four other topologies as exemplified by the acid proteases (Blundell, Cooper, Sali & Zhu, 1991), phthalate dioxygenase reductase (Correll, Batie, Ballou, & Ludvig, 1992), glutaminyl-tRNA synthetase (Rould, Perona & Steitz, 1991) and elongation factor Tu (Kjelgaard & Nyborg, 1992). The remainder of the structure consists of extensive disulfide-bonded loop regions (the seven disulfides are located between residues 11 and 135, 12 and 47, 16 and 86, 31 and 56, 87 and 199, 89 and 189, and 156 and 167) and three helices (Fig. 8).

4.4. *Catalytic centre and substrate-binding cleft*

EGV is an endocellulase (E.C. 3.2.1.4) that cleaves the $\beta(1\rightarrow4)$ glycosidic bonds of the substrate cellulose with inversion of the anomeric configuration. Kinetic studies have indicated that there are at least six, energetically significant, subsites for carbohydrate binding (Schou, Rasmussen, Kaltoft, Henrissat & Schülein, 1993). In previous papers we have described these sites as *A–F* with enzymatic cleavage taking place between subsites *D* and *E*. In order to obtain a more consistent and comparable nomenclature for the labeling of sugar binding sites in glycosyl hydrolases we now adopt the nomenclature in which enzymatic cleavage takes place between subsites –1 and +1. The leaving-group sites are, therefore, labelled +1, +2 to +*n* and the subsites before the point of cleavage are –1, –2 to –*n*. EGV, therefore, has four subsites before the point of cleavage, labelled –4 to –1 and two after the point of cleavage labelled +1 and +2 (Fig. 9).

A large groove, containing the active-site residues 10 and 121, runs across the surface of the enzyme. This substrate-binding groove is approximately 40 Å long.

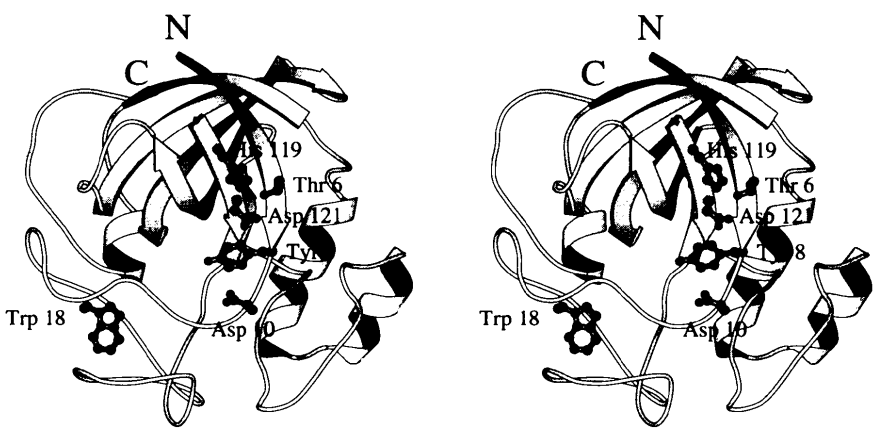

Fig. 8. Stereo diagram of the structure of the catalytic core of endoglucanase V from *H. insolens*. Residues thought to be involved in catalysis or binding as mentioned in the text are shown in ball-and-stick representation. This figure was drawn with the *MOLSCRIPT* program (Kraulis, 1990).

The catalytic residues sit either side of the groove with a tyrosine residue, Tyr8, making up the floor of the catalytic centre at this point. The spatial disposition of these residues, approximately 24 Å from one end of the active site and 14 Å from the other, is consistent with the preferred catalysis of cellohexaose to liberate cellobiose as the leaving group. The active site of EGV bares a superficial structural similarity to hen egg-white lysozyme (HEWL) in that both enzymes have one side of their active sites built up from three strands of a β-sheet, with one of the catalytic residues located in the middle strand. The other catalytic residue, in both cases, is situated on the opposite side of the active-site groove. In HEWL, the catalytic carboxylate groups are separated by approximately 5.5 Å, typical for enzymes that catalyze hydrolysis with a net retention of the anomeric configuration (for review see McCarter & Withers, 1994). In EGV, however, the carboxylates of Asp10 and Asp121 are separated by approximately 9 Å, in order to accommodate the water molecule required for catalysis with inversion of the anomeric configuration. This similarity between EGV and HEWL does not extend beyond a local structural level and perhaps only indicates a convergent evolution towards a similar $\beta(1 \rightarrow 4)$-linked oligosaccharide-binding motif.

Asp121, which we believe to be the catalytic proton donor (Davies *et al.*, 1993) sits in a predominantly hydrophobic environment. It is flanked on one side by the hydrophobic side chains of alanines 73 and 74. It is part of a hydrogen-bonding network interacting with the hydroxyl of Thr6 which in turn hydrogen bonds to His119 (Fig. 10). Asp121 exhibits distinct static disorder with two modelled conformations with relative occupancies of 0.7 and 0.3 (Fig. 7*b*). These are not dissimilar to the two conformations observed for the catalytic proton donor in the pH 5.5 and pH 6.5 xylanase structures (Törrönen *et al.*, 1994), which they ascribe to a difference in the protonation state of the glutamate. At the pH of crystallization (pH 8.0) of EGV, the enzyme has maximal activity and we would expect Asp121 to

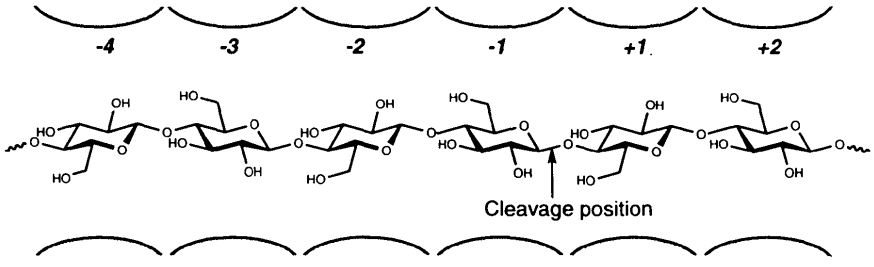

Fig. 9. Diagram showing the cleavage of cellohexaose by EGV. Cellulose is a polymer of $\beta(1 \rightarrow 4)$-linked glucose units. Kinetic measurements show the EGV has at least six significant subsites for binding labelled −4 to +2. and thus cleaves cellohexaose preferentially releasing cellobiose as the leaving group.

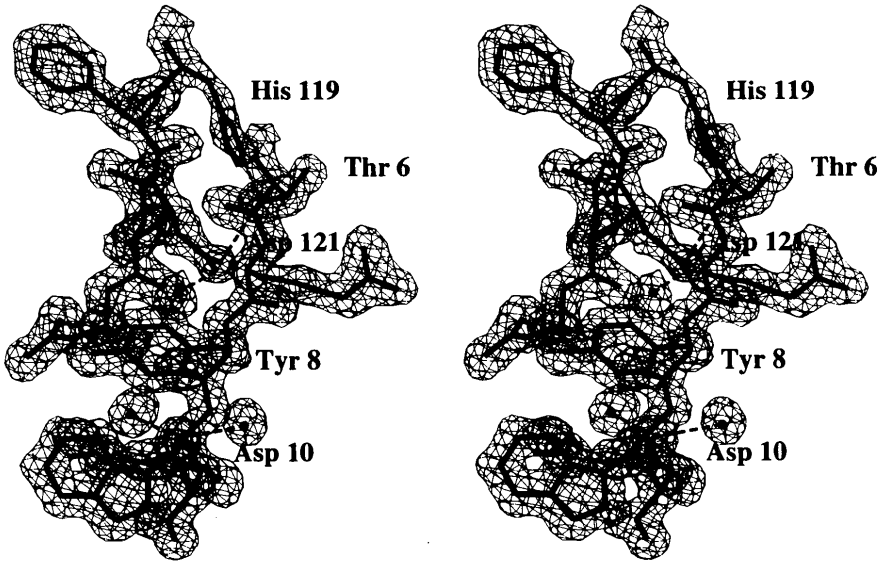

Fig. 10. $2F_o - F_c$ electron density, contoured at 0.44 e Å^{-3}, for the catalytic centre of EGV. The proposed proton donor, Asp121, and the catalytic base, Asp10, are labelled. Asp121 is involved in a hydrogen-bonding network involving residues Thr6 and His119, as indicated.

be fully protonated. It is quite possible, however, that in the native conformation the environment around Asp121 is not sufficient to maintain a suitably elevated pK_a for this residue and that the two conformation states seen do indeed represent the two protonation states. Substrate binding may, therefore, be required to elevate the pK_a of Asp121 to an appropriate level. A deep solvent channel penetrates into the structure adjacent to the tyrosine residue and contains at least eight well ordered water molecules which may play a role in catalysis. A similar feature is observed in the CBH-II structure (Rouvinen *et al.*, 1990) which has a narrow tube extending from the catalytic site into the solvent. The solvent cavity in EGV is different in that it is an enclosed pocket which extends from the active site into the core of the protein so it is not clear that water molecules in the EGV cavity could be replaced should they play a role in catalysis. Asp10, the only other completely conserved acidic group in the sequences of enzymes from family 45, is more solvent exposed. The carboxylate group primarily interacts with solvent water molecules (Fig. 10).

The substrate-binding cleft of EGV is slightly unusual in that it is not lined with aromatic residues as is typical for sugar-binding proteins (for review see, Vyas, 1991). In the other published cellulase structures, tryptophan residues are found lining the substrate-binding clefts, where they interact with the hydrophobic faces of the glucopyranosyl rings. Only two aromatic residues may be found across the substrate-binding cleft of EGV. Trp18 at one end of the cleft, presumably the –4 subsite, and Tyr8, at the point of cleavage. Instead, the substrate-binding cleft is lined with possible hydrogen-bonding groups, particularly a number of main-chain carboxyl and amide functions and many amino-acid side chains.

EGV shares its topology with barwin, a protein from barley seed related to wound-induced plant defence proteins (Svensson *et al.*, 1992; Ludvigsen & Poulsen, 1992*a*,*b*). Barwin is known to bind oligosaccharides and probably functions in response to stress, such as pathological invasion. Although barwin consists of only 125 residues compared to the 210 residues of the EGV catalytic core, the structures are quite similar with 38 of the $C\alpha$ atoms of the β-barrel overlapping with an r.m.s. distance of 1.56 Å. We have proposed (Davies *et al.*, 1993) that residue Asp121 may function as the catalytic acid in the glycosyl group hydrolysis of EGV and that Asp10 may function as the base, enhancing the nucleophilicity of the catalytic water. Barwin has a conserved aspartate residue (Asp94) in an analogous position to the EGV proton donor Asp121, but has no suitable residue directly equivalent to the catalytic base Asp10. This indicates that if barwin functions as a glycosyl hydrolase its mode of action is likely to be somewhat different to EGV. It is possible that either another acidic residue acts as the catalytic base (the barwin family of sequences have a number of other totally conserved acidic residues) or perhaps that catalysis by barwin involves an intramolecular base as is found in the *Escherichia coli* soluble lytic transglycosylase structure (Thunnissen *et al.*, 1994). It is interesting to note that the loop between strands V and VI, disordered in the native EGV structure, corresponds, in part, to the region of disorder in the solution structure of barwin (residues 89–95). This may suggest a common, and dynamic, functional role for this particular region of the structure.

5. Concluding remarks

The EGV structure presented here is the first representative of a new family of cellulose hydrolyzing enzymes. It represents a new protein fold for a cellulase and the structural similarity with barwin, another oligosaccharide binding protein suggests that this topology may represent a new oligosaccharide-binding motif. The success of optimized-wavelength data collection reiterates the usefulness of this application of synchrotron radiation to the elucidation of protein structure. Our knowledge of the structure and function relationships within EGV will best be enhanced by the study of mutant and substrate complexes of the enzyme and this work is presently in progress.

The refined coordinates and observed structure-factor amplitudes for the *H. insolens* endoglucanase catalytic core have been deposited with the Protein Data Bank (Bernstein *et al.*, 1977).* $C\alpha$ positions are already available as entry 1ENG.

The authors would like to thank Professor Nori Sakabe for assistance with data collection at the Photon Factory and Bernard Henrissat, Thomas Schneider, Alexei Murzin, Gail Hutchinson and Janet Thornton for many valuable discussions.

* Atomic coordinates and structure factors have been deposited with the Protein Data Bank, Brookhaven National Laboratory (Reference: 2ENG, R2ENGSF). Free copies may be obtained through The Managing Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England (Reference: JN0013). At the request of the authors, the atomic coordinates and structure factors will remain privileged until 29 June 1996.

References

Béguin, P. & Aubert, J.-P. (1994). *FEMS Micro. Rev.* **13**, 25–58.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Blundell, T. L., Cooper, J. B., Sali, A. & Zhu, Z.-Y. (1991). *Structure and Function of the Aspartic Proteinases: Genetics, Structures and Mechanisms*, edited by B. M. Dunn, pp. 443–453. New York: Plenum Press.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.

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

Campbell, R., Rose, D. R., Wakarchuk, W. W., To, R., Sung, W. & Yaguchi, M. (1993). *Trichoderma reesei Cellulases and Other Hydrolases. Enzyme Structures, Biochemistry, Genetics and Applications*, edited by P. Suominen & T. Rainikainen, pp. 63–72. Helsinki, Finland: Foundation for Biotechnical and Industrial Fermentation Research.

Collaborative Computational Project, Number 4. (1994). *Acta Cryst.* D**50**, 760–763.

Correll, C. C., Batie, C. J., Ballou, D. P. & Ludvig, M. L. (1992). *Science*, **258**, 1604–1610.

Cowtan, K. D. & Main, P. (1993). *Acta Cryst.* D**49**, 148–157.

Dauter, Z. (1991). *Isomorphous Replacement and Anomalous Scattering: Proceedings of the 1991 Daresbury CCP4 Study Weekend*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 163–172. Warrington, England: Daresbury Laboratory.

Davies, G. J., Dodson, G. G., Hubbard, R. E., Tolley, S. P., Dauter, Z., Wilson, K. S., Hjort, C., Mikkelsen, J. M., Rasmussen, G. & Schülein, M. (1993). *Nature (London)*, **365**, 362–364.

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

Divne, C., Ståhlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J. K. C., Teeri, T. T. & Jones, T. A. (1994). *Science*, **265**, 524–528.

Ducros, V., Czjzek, M., Belaich, A., Gaudin, C., Fierobe, H.-P., Belaich, J.-P., Davies, G. J. & Haser, R. (1996). *Structure*. In the press.

Gilbert, H. J., Hall, J., Hazlewood, G. P. & Ferreira, L. M. A. (1990). *Mol. Microbiol.* **4**, 759–767.

Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller, R. C. & Warren, R. A. J. (1991). *Microbiol. Rev.* **55**, 303–315.

Gilkes, N. R., Kilburn, D. G., Miller, R. C., Warren, R. A. J., Sugiyama, J., Chanzy, H. & Henrissat, B. (1993). *Int. J. Biol. Macromol.* **15**, 347–351.

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

Hendrickson, W. A. & Konnert, J. H. (1981). *Structure, Conformation and Evolution*, Vol. 1, edited by R. Srinivasan, pp. 43–57. Oxford: Pergamon.

Henrissat, B. (1991). *Biochem. J.* **280**, 309–316.

Henrissat, B. (1994). *Cellulose*, **1**, 169–196.

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

Herzberg, O. & Moult, J. (1991). *Proteins Struct. Funct. Genet.* **11**, 223–229.

Higashi, T. (1989). *J. Appl. Cryst.* **22**, 9–18.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Juy, M., Amit, A. G., Alzari, P. M., Poljak, R. J., Claeyssens, M., Béguin, P. & Aubert, J.-P. (1992). *Nature (London)*, **357**, 89–91.

Keitel, T., Simon, O., Borriss, R. & Heinemann, U. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 5287–5291.

Kjelgaard, M. & Nyborg, J. (1992). *J. Mol. Biol.* **223**, 721–742.

Knowles, J. K. C., Lehtovaara, P. & Teeri, T. (1987). *Trends Biotechnol.* **5**, 255–261.

Koshland, D. E. (1953). *Biol. Rev.* **28**, 416–436.

Kraulis, P. J. (1990). *J. Appl. Cryst.* **21**, 274–278.

Kraulis, P. J., Clore, G. M., Nilges, M., Jones, T. A., Pettersson, G., Knowles, J. & Gronenborn, A. M. (1989). *Biochemistry*, **28**, 7241–7257.

Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.

Langridge, R., Marvin, D. A., Seeds, W. E., Wilson, H. R., Hooper, C. W., Wilkins, M. H. F. & Hamilton, L. D. (1960). *J. Mol. Biol.* **2**, 38–64.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Leslie, A. G. W., Brick, P. & Wonacott, A. J. (1986). *CCP4 Newslett.* **18**, 33–39.

Ludvigsen, S. & Poulsen, F. M. (1992*a*). *Biochemistry*, **31**, 8771–8782.

Ludvigsen, S. & Poulsen, F. M. (1992*b*). *Biochemistry*, **31**, 8783–8789.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

McCarter, J. D. & Withers, S. G. (1994). *Curr. Opin. Struct. Biol.* **4**, 885–892.

Murzin, A. G. (1994). *Curr. Opin. Struct. Biol.* **4**, 441–449.

Murzin, A. G., Lesk, A. M. & Chothia, C. (1994*a*). *J. Mol. Biol.* **236**, 1368–1381.

Murzin, A. G., Lesk, A. M. & Chothia, C. (1994*b*). *J. Mol. Biol.* **236**, 1382–1400.

Ramakrishnan, C. & Ramachandran, G. N. (1965). *Biophys. J.* **5**, 909–933.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

Rould, M. A., Perona, J. J. & Steitz, T. A. (1991). *Nature (London)*, **352**, 213–218.

Rouvinen, J., Bergfors, T., Teeri, T., Knowles, J. K. C. & Jones, T. A. (1990). *Science*, **249**, 380–386.

Saloheimo, A., Henrissat, B., Hoffrén, A.-M., Teleman, O. & Penttilä, M. (1994). *Mol. Microbiol.* **13**, 219–228.

Schou, C., Rasmussen, G., Kaltoft, M.-B., Henrissat, B. & Schülein, M. (1993). *Eur. J. Biochem.* **217**, 947–953.

Sheppard, P. O., Grant, F. J., Oort, P. J., Sprecher, C. A., Foster, D. C., Hagen, F. S., Upshall, A., McKnight, G. L. & O'Hara, P. J. (1994). *Gene*, **150**, 163–167.

Sinnott, M. L. (1990). *Chem Rev.* **90**, 1171–1202.

Spezio, M., Wilson, D. B. & Karplus, P. A. (1993). *Biochemistry*, **32**, 9906–9916.

Svensson, B., Svendsen, I., Højrup, P., Roepstorff, P., Ludvigsen, S. & Pouslen, F. M. (1992). *Biochemistry*, **31**, 8767–8770.

Thunnissen, A.-M. W. H., Dijkstra, A. J., Kalk, K. H., Rozeboom, H. J., Engel, H., Keck, W. & Dijkstra, B. W. (1994). *Nature (London)*, **367**, 750–753.

Törrönen, A., Harkki, A. & Rouvinen, J. (1994). *EMBO J.* **13**, 2493–2501.

Vyas, N. K. (1991). *Curr. Opin. Struct. Biol.* **1**, 732–740.

Wakarchuk, W. W., Campbell, R. L., Sung, W. L., Davoodi, J. & Yaguchi, M. (1994). *Protein Sci.* **3**, 467–475.

White, A., Withers, S. G., Gilkes, N. R. & Rose, D. R. (1994). *Biochemistry*, **33**, 12546–12552.

Wilson, A. J. C. (1942). *Nature (London)*, **150**, 151–152.

Wood, T. M. (1992). *Biochem. Soc. Trans.* **20**, 46–53.

Zou, J.-Y. & Mowbray, S. L. (1994). *Acta Cryst.* D**50**, 237–249.