Acta Cryst. (1997). D53, 599–604

Crystallization and preliminary X-ray analysis of the major endoglucanase from *Thermoascus aurantiacus*

LEILA LO LEGGIO,[a] NEIL J. PARRY,[a] J. VAN BEEUMEN,[b] MARC CLAEYSSENS,[b] MAHALINGESHWARA K. BHAT[a] AND RICHARD W. PICKERSGILL[a]* at [a]*Institute of Food Research, Department of Food Macromolecular Science, Earley Gate, Reading RG6 6BZ, England, and* [b]*State University of Ghent, Faculty of Science, Department of Biochemistry, Physiology and Microbiology, K. L. Ledeganckstraat 35, B-9000, Ghent, Belgium. E-mail: richard.pickersgill@bbsrc.ac.uk*

(*Received 4 November 1996; accepted 8 April 1997*)

Abstract

The major endoglucanase (35 kDa) from the thermophilic fungus *Thermoascus aurantiacus* has been purified from culture filtrates using an affinity method and the sequence for 35 N-terminal amino acids determined. This has allowed assignment of the enzyme to subtype A6 of family 5 endoglucanases. The enzyme has been crystallized as thick plates by the hanging-drop method using ammonium sulfate as precipitant. The crystals belong to space group $P2_12_12_1$ with cell edges $a = 76.4$, $b = 85.7$ and $c = 89.5$ Å, with two molecules in the asymmetric unit, and diffract to 1.62 Å resolution using synchrotron radiation. The structure will be solved by isomorphous replacement.

1. Introduction

Microorganisms which degrade plant biomass have a full complement of enzymes with often subtly different specificities in order to efficiently hydrolyse cellulose and the associated hemicellulose. Cellulases (E.C. 3.2.1.4) are glycosyl hydrolases which are part of this catalytic machinery, and cleave internal β-1,4 linkages in cellulose.

The catalytic domains of cellulases, and in general endoglucanases, have been found to belong to several structural families, encompassing nine families in the glycosyl hydrolase classification based on sequence comparison by homology and hydrophobic cluster analysis (Henrissat, 1991; Henrissat & Bairoch, 1993). Previously cellulases and xylanases had been classified apart from other glycosyl hydrolases (Gilkes, Henrissat, Kilburn, Miller & Warren, 1991) to form families A to L (corresponding to glycosyl hydrolases families 5 to 12, 44, 45 and 48). Structures for enzymes belonging to several of these families have been solved, showing that the same function can be accommodated in folds as diverse as the twisted 12-helix barrel of *C. thermocellum* endoglucanase D (family 9) (Juy *et al.*, 1992) and the lectin-like all-β fold of *Trichoderma reesei* cellobiohydrolase I (family 7) (Divne *et al.*, 1994).

Here we report the classification and preliminary X-ray analysis of a 35 kDa thermostable endoglucanase produced by the moderately thermophilic fungus *T. aurantiacus* (optimally growing at 328 K). As will be described, this cellulase can be assigned to family 5 of the glycosyl hydrolases. This family of enzymes has been further subdivided in five subtypes, A1 to A5, where the A refers to the alternative classification of family 5 as family A (Béguin, 1990). From one subtype to the other there is considerable variation in sequence, such that similarities between them often cannot be easily identified by alignments based on sequence identity. The sequence identity is typically below 25%, so that more sophisticated methods, such as hydrophobic cluster analysis, have been used to ascertain similarities between subtypes.

The division in subtypes does not merely reflect the evolutionary relationships between different organisms, as endoglucanases from *C. thermocellum* belong to at least three of the subtypes (Béguin, 1990). It may however reflect subtly different substrate specificities, as enzymes belonging to subtype A3 often show lichenase activity, and enzymes belonging to subtype A4 often show significant xylanase activity.

Two structures of cellulases belonging to family 5 have been solved so far: one belonging to subtype A3, the *C. thermocellum* cellulase C (Dominguez *et al.*, 1995), and one belonging to subtype A4, the *C. cellulolyticum* endoglucanase A (Ducros *et al.*, 1995). In the *C. cellulolyticum* enzyme, the cellulose binding domain had to be removed to allow crystallization, while the *C. thermocellum* enzyme, like the major endoglucanase from *T. aurantiacus*, is devoid of extra domains. The catalytic domains of cellulase C and endoglucanase A have an eightfold β-barrel architecture, or TIM barrel architecture after triose phosphate isomerase, where the fold was first observed (Banner *et al.*, 1975). From comparison of the structures, family 5 glycosyl hydrolases have been shown to belong to a superfamily of glycosyl hydrolases (Jenkins, Lo Leggio, Harris & Pickersgill, 1995; Henrissat *et al.*, 1995), which also includes family 10 xylanases and family 17 barley lichenases. As well as sharing their overall architecture, enzymes in the three families all retain configuration at the anomeric C atom on hydrolysis of the substrate, have their active-site glutamates positioned at the C-terminal end of β-strands 4 and 7 (hence the superfamily is referred to as the 4/7 superfamily) and show conservation of several residues in the substrate binding groove.

T. aurantiacus cellulase belongs to a subtype of family 5 endoglucanases, for which no structure is yet available. By combining structure solution and detailed biochemical characterization of this enzyme, we aim to gain a better understanding of the determinants of substrate specificity within the family 5 endoglucanases and glycosyl hydrolases in the 4/7 superfamily.

2. Classification of *T. aurantiacus* endoglucanase

The major endoglucanase from *T. aurantiacus* was purified using a novel affinity column according to the method of Parry (1996). N-terminal sequence analysis using a model 477A protein sequencer working with pulsed-liquid phase allowed the identification of 35 N-terminal amino acids. Homologous sequences in the GenBank and SwissProt Databases were searched using the programs *BLAST* (Altschul, Gish, Miller, Myers & Lipman, 1990) and *FASTA* in the *GCG* package

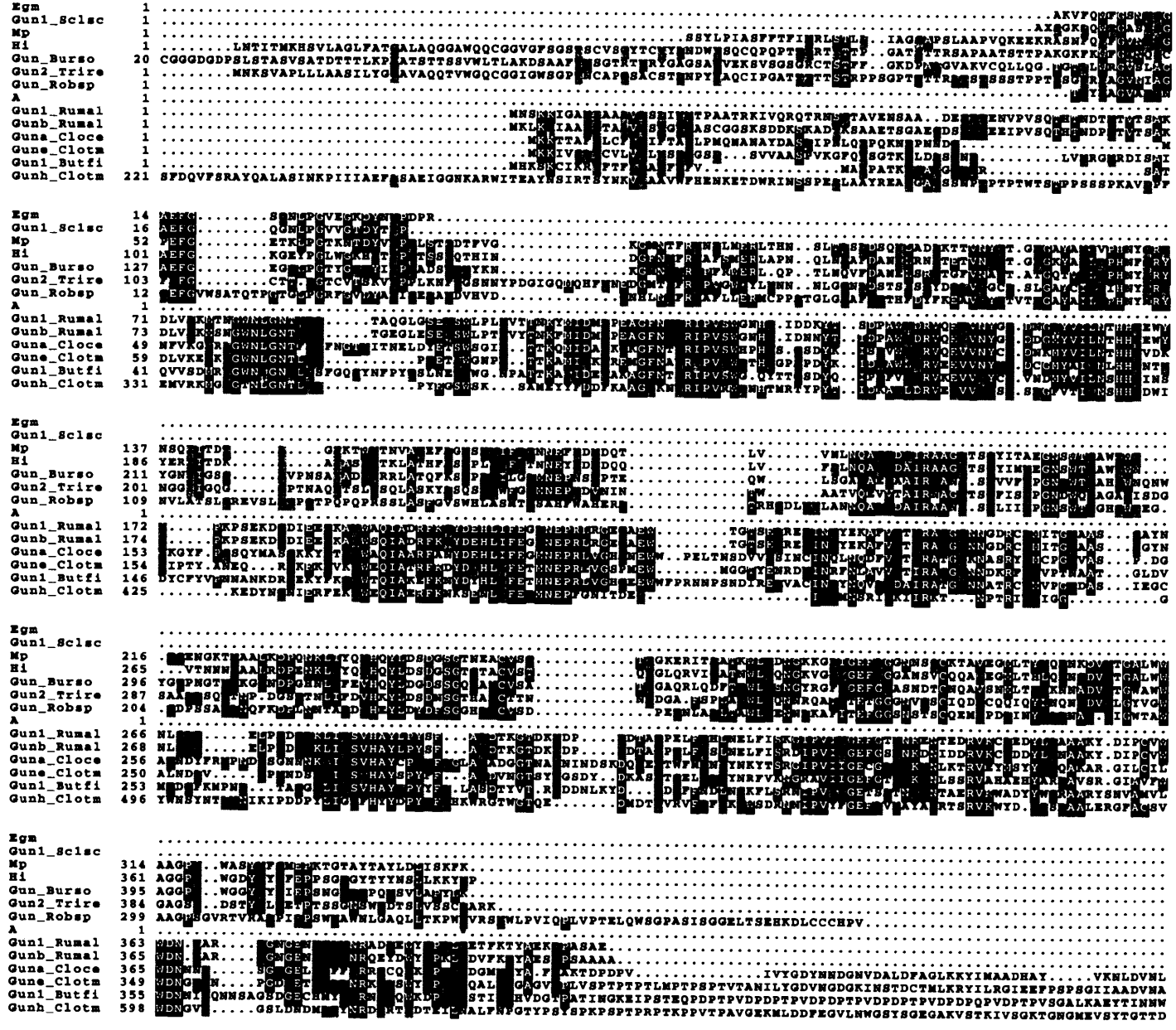

Fig. 1. Sequence alignment of family 5 cellulases belonging to subtype A6 (top 7) and subtype A4 (bottom 6). The N-terminal sequence of the major endoglucanases from *T. aurantiacus* is labelled EGM. All the other sequences are labelled as indicated in the text.

(Genetics Computer Group, 1991). Sequence comparison showed highest similarity with the N-terminal sequence (35 amino acids) of *Sclerotinia sclerotiorum* endoglucanase 1 (GUN1_SCLSC, 59.3% identity), *Macrophonima phaseolina* cellulase (MP, 57.5% identity), *Humicola insolens* cellulase (HI, 48.5% identity), and *P. solacenarum* cellulase (GUN_BURSO, 44.1% identity). Lower, but still significant similarity, was found with the *Trichoderma reesei* endoglucanase 2 (GUN2_TRIRE, 26.4% identity) and a *Robillarda sp.* endoglucanase (GUN_ROBSP, 34.3% identity). These sequences, although they can be confidently assigned to family 5 glycosyl hydrolases, are not assignable to either of the five originally recognised subtypes. Rather, they seem to form a distinct subtype A6, most closely related to the *C. thermocellum* endoglucanase E (GUNE_CLOTM) and *Ruminococcus albus* endoglucanase A and B (GUN1_RUMAL and GUNB_RUMAL), which belong to subtype A4 but share less then 25% sequence identity with the new subtype. All of the sequences above are fungal bar the *Pseudomonas solacenarum*, and this is perhaps one reason why they form a separate group in terms of sequence identity. A sequence alignment is shown in Fig 1, including also *Clostridium cellulolyticum* endoglucanase A (GUNA_CLOCE), *C. thermocellum* endoglucanase H (GUNH_CLOTM), and *Butyrivibrio fibrisolvens* endoglucanase 1 (GUN1_BUTFI).

3. Crystallization and preliminary X-ray analysis

Initial crystallization trials were set up using a Hampton Research Screen I and the vapour diffusion in hanging-drop technique. The hanging drops consisted of 2 µl of protein solution at 12.3 mg ml^{-1} (assuming an A_{280} of 2 for a protein concentration of 1 mg ml^{-1} on the basis of other protein estimation methods) and 2 µl of reservoir solution, and trials were stored at 291 K. A composite crystal grew from one of the conditions (number 4 in the screen) within two weeks. The reservoir consisted of 0.1 *M* Tris–HCl pH 8.5 and 2.0 *M* ammonium sulfate.

Exploration of similar conditions established that good X-ray quality crystals could be grown in hanging drops consisting of 2 µl of protein of concentration ranging between 12 and 32 mg ml^{-1} and 2 µl of reservoir (0.1 *M* Tris–HCl pH 7.5–9.0, 1.3 to 1.6 *M* ammonium sulfate). Crystals were plate like and started growing overnight, although they continued growing for some weeks to a maximum size of 0.7 × 0.5 × 0.05 mm (Fig. 2).

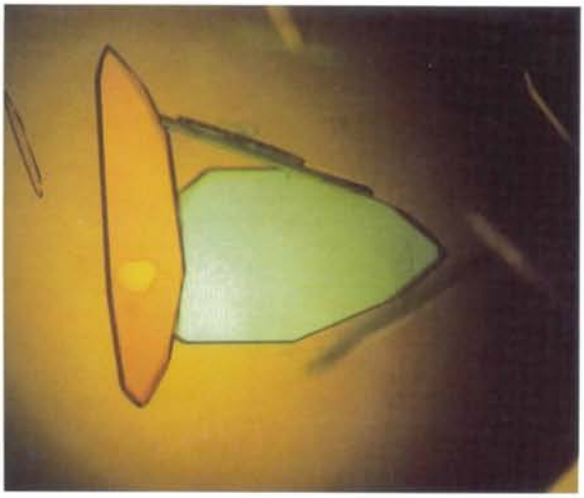

Fig. 2. Orthorhombic crystals of the major endoglucanase from *T. aurantiacus* grown in a hanging drop using ammonium sulfate as precipitant.

A native data set was collected using a Cu $K\alpha$ rotating-anode generator and a Xentronics area detector. The data were collected from two crystals grown using an endoglucanase concentration of 31.2 mg ml^{-1} and 0.1 *M* Tris–HCl pH 9.0, 1.6 *M* ammonium sulfate in the reservoir, and reduced using the *XENGEN* suite of programs (Howard *et al.*, 1987). The crystals belonged to the orthorhombic space group $P2_12_12_1$, as could be demonstrated by the pattern of systematic absences along the three axes a^*, b^* and c^*. The cell edges were $a = 76.4$, $b = 85.7$ and $c = 89.5$ Å. The statistics for the resulting data set are shown in Table 1(*a*). The cell edges are consistent with two molecules in the asymmetric unit, as this gives a calculated value of V_m (specific volume) = 2.09 Å^3 Da^{-1} which is within the typical range for proteins (Matthews, 1968).

So far, our attempts to solve the structure using molecular replacement with the *C. thermocellum* endoglucanase C model have been unsuccessful. Endoglucanase C belongs to subtype A3, and has only about 16% sequence identity with the *Macrophonima phaseolina* cellulase, the full sequence most similar to the N-terminal sequence of the *T. aurantiacus* endoglucanase. This level of sequence identity is at the current limit of successful structure solution by molecular replacement and reflects substantial structural differences between search model and target structure. The presence of two molecules in the asymmetric unit further hinders structure solution by molecular replacement. The non-crystallographic symmetry (NCS) relationship between the two molecules was initially unclear, as no significant peaks were found in a self-rotation map except for ones produced by crystallographic symmetry. This could indicate the presence a non-crystallographic twofold axis approximately parallel to one of the crystallographic axes, so that the peak corresponding to the non-crystallographic axis is obscured by one of the crystallographic axes' peaks.

While we are still pursuing the molecular replacement approach, we also have successfully prepared two heavy-atom derivatives. The first derivative was obtained by soaking an endoglucanase crystal in a solution containing 10 m*M* $SmCl_3$ for 2.5 h. The second derivative was obtained by soaking a crystal in a solution containing 75 m*M* $HgAc_2$ for 20 h. Both data sets were collected using Cu $K\alpha$ radiation and showed good internal statistics and isomorphous cell edges with the native (Table 2). Isomorphism was confirmed by the low R_{iso}* for the derivatives (Table 2). The scaling program *LOCAL*, which is distributed as unsupported with the *CCP*4 suite of programs (Collaborative Computational Project, Number 4, 1994), gave best results in terms of the clarity of peaks on the Harker sections of the isomorphous difference Patterson. The samarium derivative isomorphous difference Patterson was interpreted by visual examination aided by use of the *CCP*4 program *VECSUM* for checking consistency with the cross vector peaks. Although more noisy, the anomalous Patterson for this derivative was also interpretable. As an example, the Harker section $w = 1/2$ for the isomorphous difference and anomalous Patterson maps are shown in Fig. 3. The data are consistent with two heavy atoms bound per asymmetric unit, most likely one

* $R_{iso} = \sum |F_1 - F_2| / \sum F_1$.

Table 1. *Data-collection statistics for endoglucanase*

Resolution (Å)	No. of unique reflections	$R_{sym}(I)$* (%)	$I/\sigma(I) > 5$ (%)	Completeness (%)	Multiplicity
(*a*) Low-resolution data set					
35.35–4.27	4339	5.7	99.4	99.5	7.2
4.27–3.39	4211	7.0	90.3	100.0	6.0
3.39–2.96	4125	10.0	81.9	99.4	5.0
2.96–2.69	4042	12.5	70.3	98.0	3.6
2.69–2.50	3761	12.4	54.4	91.2	2.0
2.50–2.35	2405	14.1	44.1	58.6	1.5
35.35–2.35	22883	7.7	75.3	91.3	4.5
(*b*) High-resolution data set					
35.25–5.82	816	2.9	97.3	49.8	5.6
5.82–4.28	1404	2.8	98.9	55.2	5.5
4.28–3.54	2011	2.9	97.6	63.6	5.6
3.54–3.09	2774	3.2	97.3	75.5	5.7
3.09–2.77	3299	3.5	95.7	80.0	5.6
2.77–2.54	3709	4.0	94.0	81.8	5.5
3.54–2.36	4078	4.7	93.4	83.1	5.5
2.36–2.21	4390	5.3	92.3	83.7	5.4
2.21–2.08	4730	6.2	88.5	84.9	5.4
2.08–1.98	5034	7.3	86.0	85.8	5.4
1.98–1.89	5278	9.4	80.5	85.8	5.3
1.89–1.81	5573	12.3	73.3	86.6	5.2
1.81–1.74	5838	16.1	65.4	87.2	5.0
1.74–1.68	5965	21.9	55.0	85.9	5.0
1.68–1.62	6023	28.9	44.6	83.9	4.9
35.35–1.62	60922	5.6	78.7	81.6	5.3

* $R_{sym}(I) = \sum_{hkl} |I_i - \langle I \rangle| / \sum_{hkl} \langle I \rangle$, where I_i is the observed intensity and $\langle I \rangle$ the mean reflection intensity over all related observations.

Table 2. *Data collection and phasing for $SmCl_3$ and $HgAc_2$ endoclucanase derivatives*

(*a*) Data collection

	Resolution (Å)	No. of unique reflections	$R_{sym}(I)$* (%)	$I/\sigma(I) > 5$ (%)	Completeness (%)	Multiplicity
$SmCl_3$	35.35–2.67	12295	9.6	77.0	71.7	4.3
$HgAc_2$	33.35–3.32	8187	9.7	75.7	91.5	3.1

(*b*) Phasing statistics

	R_{iso} (%) †	Phasing power‡ centric	Phasing power‡ acentric	R_{Cullis} centric§	R_{Cullis} acentric§	R_{Cullis} anomalous¶
$SmCl_3$	13.7 (33.3–2.7)	1.20	1.66	0.75	0.77	0.90
$HgAc_2$	14.2 (33.3–3.4)	1.31	1.76	0.76	0.78	0.97

* $R_{sym}(I) = \sum_{hkl} |I_i - \langle I \rangle| / \sum_{hkl} \langle I \rangle$, where I_i is the observed intensity and $\langle I \rangle$ the mean reflection intensity over all related observations. † $R_{iso} = \sum |F_{PH} - F_P| / \sum F_P$, where F_P is the native and F_{PH} the derivative structure factor. ‡ Phasing power = $\langle F_H \rangle / \langle E \rangle$, where $\langle F_H \rangle$ is the root-mean-square calculated heavy-atom structure-factor amplitude and $\langle E \rangle$ is the root-mean-square lack-of-closure error. § $R_{Cullis} = \sum ||F_{PH} \pm F_P| - F_H| / \sum |F_{PH} \pm F_P|$ where F_H is the calculated heavy-atom structure factor, F_P is the native and F_{PH} the derivative structure factor. ¶ Anomalous $R_{Cullis} = [\sum(|\Delta F^{\pm}_{obs}| - |\Delta F^{\pm}_{calc}|)^2 / \sum(\Delta F^{\pm}_{obs})^2]^{1/2}$ where $\Delta F^{\pm}_{obs}$ and $\Delta F^{\pm}_{calc}$ are the observed and calculated anomalous differences, respectively.

per molecule as there are two molecules in the asymmetric unit. The Y coordinates for the two heavy atoms are similar, consistent with the presence of a twofold axis approximately parallel to the b cell axis and centred at the midpoint between the two heavy-atom positions. The midpoint has approximate coordinates $x = 0.25$ and $z = 0.25$, therefore indicating that the NCS twofold axis is at a distance of 0.25 in x from the crystallographic 2_1 axis parallel to b. A peak in the native Patterson map at $u = 0.5$, $v = 0.5$, $w = 0.0$, with a peak height of approximately a quarter of the origin peak, is also consistent with a twofold axis approximately parallel to b and at a distance of 0.25 in x and 0.0 in z from a crystallographic 2_1 axis parallel to b. After refinement in *VECREF*, the heavy-atom sites were input in *MLPHARE* for phasing using the isomorphous difference data only, and the newly calculated phases used for determination of the heavy-atom positions in the mercury derivative by difference Fourier techniques. The mercury sites appear to be in very similar positions to the samarium sites,

although one of the two mercury sites has much lower occupancy than the other. After refinement of all the sites in *VECREF*, anomalous data were also included for phasing in *MLPHARE*. Scaling and phasing statistics are summarized in Table 2. Both derivatives showed reasonably high phasing power and low R_{Cullis} for the isomorphous difference data, and usable anomalous data (anomalous $R_{\text{Cullis}} < 1$).

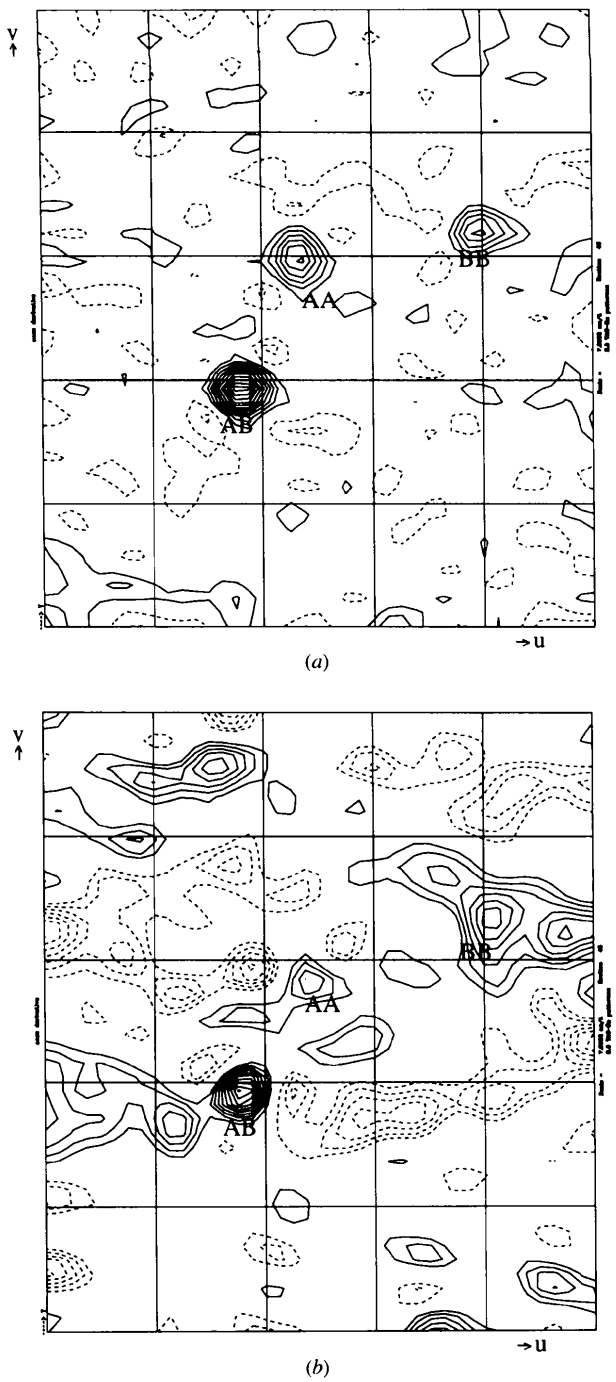

Fig. 3. Harker section $w = 1/2$ of (*a*) the isomorphous difference Patterson map and (*b*) anomalous Patterson map of the endoglucanase $SmCl_3$ derivative. AA and BB indicate the Harker peaks while AB is a cross peak between the two sites which lies on this Harker section by coincidence.

We have now collected another native data set using synchrotron radiation and the Weissenberg camera (Sakabe *et al.*, 1995) at the Photon Factory (BL6A). The data set was processed using *DENZO* (Otwinowski, 1991), merged with the *ROTAVATA/AGROVATA* programs in the *CCP*4 suite, and shows good statistics to 1.62 Å resolution and reasonable isomorphism with the data set collected in-house ($R_{\text{iso}} = 0.142$). The statistics for this data set are shown in Table 1(*b*). However, the completeness, especially at low resolution, is rather poor, mainly due to overlapping and overloaded reflections. This data set was therefore 'refilled' using the data set collected in house, so that merged reflections from the lower resolution data set which were not measured in the higher resolution data set were included in the latter. This 'refilled' data set had a completeness of 88.5% between infinity and 1.62 Å (99.2% completeness between infinity and 5.1 Å).

The major obstacle for structure solution is the lack of complete primary sequence, but we expect that multiple isomorphous replacement, in combination with density modification methods including averaging of the two molecules in the asymmetric unit, will allow us to build an initial model that can be at least partially refined. We are attempting to obtain the sequence by cloning the gene from *T. aurantiacus* DNA by the polymerase chain reaction technique, which will greatly benefit from an X-ray derived sequence (although this would clearly contain many ambiguities), in order to obtain sequence information to generate PCR primers corresponding to the C-terminal end of protein.

We would like to thank Professor N. Sakabe for provision of synchrotron time at the Photon Factory, Tsukuba, Japan, and Olga Mayans, John Jenkins and Gillian Harris for their help with data collection. This work was funded by the European Union under contract AIR2-CT93-1272 and by the Concerted Research Action (contract 12 052 293) of the Flemish Government to J. Van Beeumen.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403–410.

Banner, D. W., Bloomer, A. C., Petsko, G. A., Philips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E., Priddle, J. D. & Waley, S. G. (1975). *Nature (London)*, **255**, 609–614.

Béguin, P. (1990). *Ann. Rev. Microbiol.* **44**, 219–248.

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

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

Dominguez, R., Souchon, H., Spinelli, S., Dauter, Z., Wilson, K. S., Chauvaux, S., Béguin, P. & Alzari, P. M. (1995). *Nature 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.

Genetics Computer Group (1991). *Program Manual for the GCG Package*, Version 7, April 1991, 575 Science Drive, Madison, Wisconsin, USA.

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

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

Henrissat, B. & Bairoch, A. (1993). *Biochem. J.* **293**, 781–788.
Henrissat, B., Callebaud, I., Fabrega, S., Lehn, P., Mornon, J. P. & Davies G. (1995). *Proc. Natl Acad. Sci. USA*, **92**, 7090–7094.
Howard, A. J., Gilliland, G. L., Finzel, B. C., Poulos, T. L., Ohlendorf, D. H. & Salemme, F. R. (1987). *J. Appl. Cryst.* **20**, 383–387.
Jenkins, J., Lo Leggio, L., Harris, G. W. & Pickersgill, R. W. (1995). *FEBS Lett.* **362**, 281–285.
Juy, M., Amit, A. G., Alzari, P. M., Poljak, R. J., Claeyssens, M., Béguin, P. & Aubert, J.-P. (1992). *Nature (London)*, **357**, 89–91.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Otwinowski, Z. (1991). In *Isomorphous Replacement and Anomalous Scattering: Proceedings of the CCP4 Study Weekend, 25–26 January 1991*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie. Warrington: Daresbury Laboratory.
Parry, N. J. (1996). PhD thesis, Reading University, England.
Sakabe, N., Ikemizu, S., Sakabe, K., Higashi, T., Nakagawa, A., Watanabe, N., Adachi, S. & Sasaki, K. (1995). *Rev. Sci. Inst.* **66**, 1276–1281