Gene sequence and properties of Cell, a family E endoglucanase from *Clostridium thermocellum*

GEOFFREY P. HAZLEWOOD,^{1*} KEITH DAVIDSON,¹ JUDITH I. LAURIE,¹ NEVILLE S. HUSKISSON¹ and HARRY J. GILBERT²

¹Department of Biochemistry, AFRC Institute of Animal Physiology and Genetics Research, Babraham, Cambridge CB2 4AT, UK

²Department of Biological and Nutritional Sciences, University of Newcastle-upon-Tyne, Newcastle-upon-Tyne NE1 7RU, UK

(Received 17 September 1992; accepted 29 September 1992)

The *Clostridium thermocellum* *cell* gene, coding for endoglucanase I (Cell), consists of an open reading frame (ORF) of 2640 nucleotides and codes for a protein of M_r 98531. The ORF was confirmed as *cell* by comparing the N-terminal sequence of purified recombinant Cell with that deduced from the nucleotide sequence. Cell hydrolysed lichenan and carboxymethylcellulose, but was principally active against barley β -glucan. It exhibited significant sequence identity with subfamily E₂ endoglucanases, and by analogy with others in this group contains a catalytic domain of around 500 residues located in the N-terminal half of the protein. The C-terminal region of Cell was highly homologous with the cellulose-binding domain of the non-catalytic cellosome subunit, S1. A repeated segment, previously shown to be highly conserved in xylanase Z and in other endoglucanases from *C. thermocellum*, was absent from Cell. Antiserum raised against purified recombinant Cell cross-reacted with proteins contained in the cellosomes of two strains of *C. thermocellum*, suggesting that Cell is either a component of the cellosome or is homologous to other cellosome proteins. A second gene, located upstream of *cell*, consisted of an ORF of 1671 nucleotides, coding for a protein of M_r 61042. Based on its homology with the *Escherichia coli* *tar* gene product, the polypeptide encoded by the second gene is tentatively identified as a sensory transducer.

Introduction

Cellulase activity of the anaerobic thermophile *Clostridium thermocellum* is contained in a large extracellular aggregate (the cellosome) which binds tightly to cellulose and in the early stages of culture on cellulose mediates the attachment of cells to the substrate (Lamed & Bayer, 1988). Large cellosome clusters have been observed in certain strains (Mayer *et al.*, 1987), but the basic subunit in strain YS is a 2.1 MDa complex composed of at least 14 polypeptides ranging in size from 48 kDa to 210 kDa (Lamed & Bayer, 1988); cellosomes composed of as many as 50 polypeptides, many of which are glycosylated, have been described in other strains (Mayer *et al.*, 1987; Kohring *et al.*, 1990).

Multiple isomeric forms of endoglucanase (some with xylanase activity) and xylanases account for nearly all of the cellosome components of *C. thermocellum* (Lamed & Bayer, 1988; Kohring *et al.*, 1990; Morag *et al.*, 1990). Although several authors have presented evidence for exoglucanase-type activity, an enzyme analogous to the cellobiohydrolase of aerobic fungi has only recently been definitively identified as a component of the cellosome (Morag *et al.*, 1991).

Within the past decade biochemical analysis of the *C. thermocellum* cellulase system has been augmented by studies employing rDNA methodologies. A large number of *Escherichia coli* clones expressing *C. thermocellum* cellulase and xylanase genes have been described; these include at least 15 different endoglucanase genes, two xylanase genes and two β -glucosidase genes (Romaniec *et al.*, 1987; Hazlewood *et al.*, 1988). Complete nucleotide sequences have been determined for *celA* (Béguin *et al.*, 1985), *celB* (Grépinet & Béguin, 1986), *celC* (Schwarz *et al.*, 1988), *celD* (Joliff *et al.*, 1986), *celE* (Hall *et al.*, 1988), *celF* (Navarro *et al.*, 1991), *celH* (Yagüe *et al.*, 1990),

* Author for correspondence. Tel. 0223 832312; fax 0223 833511.

The nucleotide sequence data reported in this paper have been submitted to GenBank and have been assigned accession numbers L04735 (*cell*) and L04736 (ORF1).

bglB (Grabnitz *et al.*, 1989) and *xynZ* (Grépinet *et al.*, 1988).

Based on hydrophobic cluster analysis, the catalytic domains of the *C. thermocellum* endoglucanases and xylanase have been assigned to four of the six cellulase families initially defined by Henrissat *et al.* (1989). The majority of the endoglucanases (B, C, E and H) belong to family A, while CelD and CelF have been ascribed to family E and CelA and XynZ to families D and F respectively (Béguin, 1990).

Comparisons of the translated sequences of cloned *cel* and *xyn* genes revealed a conserved region containing a duplicated segment of approximately 24 amino acids separated by 8 to 15 residues. This region was present in XynZ (Grépinet *et al.*, 1988) and in all of the endoglucanases except CelC, and was located at the C-terminus, except in CelE and XynZ, where it was positioned towards the centre of the protein. The conserved domain was not required for the catalytic activity of CelE (Hall *et al.*, 1988), CelD (Chauvaux *et al.*, 1990) or XynZ (Grépinet *et al.*, 1988). Nor was it part of the cellulose-binding domain located at the C-terminus of CelE (Durrant *et al.*, 1991). The likelihood that the conserved domain interacts with a common determinant on a template protein during assembly of the cellulosome is supported by recent binding studies (Tokatlidis *et al.*, 1991) and by the observation that CelC, which lacks the domain, is not part of the cellulosome, while CelD, CelE and XynZ, which carry the duplicated segment, are known to be associated with the cellulosome (Béguin, 1990).

In this paper, we present the primary sequence of the *celI* gene and its encoded protein Cell, a family E endoglucanase from *C. thermocellum*. We have purified a truncated but catalytically active form of Cell from *E. coli* cultures and have determined some of its properties. In addition, we have shown that antiserum against purified recombinant Cell cross-reacts with proteins contained in the extracellular cellulase complex produced by two different strains of *C. thermocellum*.

Methods

Strains and plasmids. The bacteria used in this study were *Clostridium thermocellum* NCIB 10682, *Escherichia coli* JM83 [F^- *ara* Δ (*lac-proAB*) *rpsL* ϕ 80 *lacZ* Δ M15] and recombinants of *Escherichia coli* HB101 [F^- *hsdS20*(r_B^- m_B^-) *recA13* *leuB6* *ara-14* *proA2* *lacY1* *galK2* *rpsL20*(Str^r) *xyl-5* *mtl-1* *supE44* λ^-] containing *C. thermocellum* genomic DNA cloned in pBR322 (Romaniec *et al.*, 1987). The vectors used were pUC18 and pUC19 (Norlander *et al.*, 1983).

Culture conditions. *C. thermocellum* was cultured anaerobically at 60 °C in medium containing cellulose powder (MN300; Machery Nagel) as source of fermentable carbon, as described by Hazlewood *et al.* (1990a). *E. coli* was cultured at 37 °C in Luria broth (LB). Filter-sterilized ampicillin (100 μ g ml⁻¹) was added as required.

Preparation of cellulosome fraction. A cellulosome fraction was prepared from *C. thermocellum* culture supernatant using the affinity method described by Morag *et al.* (1992). High- M_r cellulase complex contained in the supernatant was selectively bound to acid-swollen cellulose at 4 °C and was released after digestion of the cellulose at 50 °C. Cellulosomes prepared by this method have the same composition as those purified by adsorption to crystalline cellulose and subsequent desorption, but are obtained in much higher yield. Total cellulase activity and a second cellulosome preparation were obtained as described previously (Hazlewood *et al.*, 1990a). Cellulosome from *C. thermocellum* strain YS was the kind gift of Dr E. Morag (Tel Aviv University, Israel).

Fractionation of *E. coli*. Total cell-free extract and periplasmic and cytoplasmic fractions were prepared as described by Hazlewood *et al.* (1990a).

Preparation of antiserum and immunological detection of endoglucanase I. Antiserum against purified recombinant Cell was raised in New Zealand White rabbits (Hazlewood *et al.*, 1990a). Proteins contained in *C. thermocellum* total cellulase or cellulosome preparations were fractionated by SDS-PAGE (Laemmli, 1970), and electrophoretically transferred to Immobilon membrane (Millipore) (Towbin *et al.*, 1979). Cell was detected on Western blots using enhanced chemiluminescence (ECL; Amersham).

Assays. Cellulase-related enzyme activities were assayed at 60 °C in PC buffer (50 mM-K₂HPO₄, 12 mM-citric acid, pH 6.5) as described by Hazlewood *et al.* (1990b). Protein was measured by dye-binding (Sedmak & Grossberg, 1977) or by the Lowry method, with bovine serum albumin as standard. For determining substrate specificity, the source of enzyme was cell-free extract from *E. coli* JM83 harbouring pKH9.

Purification of endoglucanase I. Recombinant Cell was purified from cell-free extract of *E. coli* JM83 harbouring pKH9. Cells cultured for 16 h at 37 °C in 3 litres of LB containing ampicillin (100 μ g ml⁻¹) were collected and resuspended in 50 mM-Tris/HCl, pH 8 (100 ml). Cell-free extract was prepared by sonicating the cell suspension on ice. Streptomycin sulphate (1 ml of a 10%, w/v, solution per g protein) was added and the precipitate which formed during 1 h of mixing at 4 °C was removed by centrifuging (12000 g; 4 °C; 15 min). The supernatant fraction was heated at 60 °C for 10 min and heat-denatured protein was removed by centrifuging (15000 g; 4 °C; 15 min) and by filtering (Whatman, GF-D). Endoglucanase activity was precipitated by the addition of ammonium sulphate (60% saturation) and was recovered by centrifuging before being dissolved in the minimum volume of 10 mM-phosphate buffer, pH 6.6, and dialysed against three changes of the same buffer. After mixing with an equal volume of 1 M-phosphate buffer, pH 6.6, the dialysed material was applied to a Phenyl-Sepharose column (2.6 cm $\times$ 12.0 cm) and was eluted in the final fractions of a linear gradient from 500 mM-phosphate, pH 6.6, to glass-distilled water (Chauvaux *et al.*, 1990). Active fractions were pooled and dialysed against 20 mM-Tris/HCl, pH 8, and Cell was further purified by anion-exchange chromatography on DEAE-Trisacryl M as described by Hazlewood *et al.* (1990b). Electrophoretically pure Cell was eluted between 150 mM- and 175 mM-NaCl and was dialysed against 10 mM-Tris/HCl, pH 8, and concentrated by vacuum evaporation.

Sequencing of Cell. The N-terminal amino acids of recombinant Cell were determined by automated Edman sequencing using a 470A gas-phase sequencer equipped with a 120A on-line PTH analyser (Applied Biosystems) (Hunkapillar *et al.*, 1983).

DNA sequencing. The major part of *celI* and a second upstream gene were contained in the *Sau3A* fragment cloned into pRV1.6 (Fig. 2; Romaniec *et al.*, 1987). The 3' extremity of *celI* was carried by a second overlapping *Sau3A* genomic fragment of 3.5 kb which was also cloned

Fig. 1. Southern hybridization of ^{32}P -labelled pRV1.6 with *Hind*III (H), *Sal*I (Sa), *Sph*I (Sp) and *Eco*RI (R) digests of *C. thermocellum* genomic DNA. *Hind*III fragments of λ phage DNA were run as standards and sizes are shown in kb.

in pBR322 (Fig. 2; pKG2). To construct full-length *cell*, the 5.3 kb *Sal*I/*Eco*RI fragment from pRV1.6 was ligated into pUC19 to generate pKG1. After digesting pKG1 with *Eco*RI and *Sac*I, the large fragment containing the vector, the N-terminal half of *cell* and the upstream gene was purified by electroelution from an agarose gel. The 1.5 kb *Eco*RI/*Sac*I fragment from pKG2, which contained the C-terminal half of *cell*, was similarly purified and ligated into the large fragment described above to produce full-length *cell* cloned into pUC19 (Fig. 2; pKG3). The nucleotide sequence of the cloned fragment was deter-

mined for both strands using the dideoxy chain-termination method of Sanger *et al.* (1977). Our method for sequencing double-stranded plasmid DNA was based on the protocol recommended for use with the Sequenase DNA sequencing kit (USB), with certain modifications. Plasmid DNA was prepared by the method of Birnboim & Doly (1979), and after denaturation by treatment with alkali was neutralized and further purified by spin-dialysis (Murphy & Kavanagh, 1988). Overlapping sequences were generated using oligonucleotide primers (17-mers) synthesized on an automated DNA synthesizer (Biosearch 8700 DNA synthesizer; New Brunswick Scientific). Sequences were compiled and ordered using software written by the Genetics Computer Group at the University of Wisconsin.

Bal31 deletions. Deletion of the *C. thermocellum* insert 5' of the *cell* gene was accomplished by linearizing pKG1 with *Sal*I and digesting with *Bal*31 (Hall & Gilbert, 1988). Blunt ends were produced by incubating with T4 DNA polymerase and the truncated insert, excised by digesting with *Eco*RI, was ligated into pUC18 which had been cleaved with *Sma*I and *Eco*RI. The 3' end of *cell* was deleted by linearizing pKG1 with *Eco*RI before digesting with *Bal*31 and filling in with T4 DNA polymerase. The deleted fragment was excised with *Sal*I and was ligated into pUC19 previously digested with *Sal*I and *Sma*I.

General recombinant DNA procedures. Plasmid DNA was prepared on a large scale from *E. coli* by alkaline lysis followed by caesium chloride density-gradient centrifugation. The method of Birnboim & Doly (1979) was used for small-scale extraction and rapid screening of plasmid-containing clones. Restriction enzymes and other DNA-modifying enzymes were used under the conditions specified by individual suppliers. DNA fragments were fractionated by electrophoresis in 0.8% (w/v) agarose gels.

Southern blot hybridization. Hybridization of the insert cloned in pRV1.6 with *C. thermocellum* genomic DNA was assessed as described by Romaniec *et al.* (1987). Probe DNA (25 to 50 ng) was denatured by boiling and labelled by incubating at 37 °C with [^{32}P]dCTP in the presence of Klenow fragment (20 U ml $^{-1}$) and random hexamer (Sambrook *et al.*, 1989).

Enzymes and reagents. Bacteriophage T4 ligase and *Bal*31 exonuclease were purchased from Boehringer. T7 DNA polymerase was from Cambridge Bioscience. Restriction enzymes were obtained from Northumbrian Biologicals. Deoxyadenosine 5'-[α - ^{35}S]thiotriphosphate

Fig. 2. Restriction map of the *C. thermocellum* DNA fragment carrying the *cell* gene. Hatched regions denote pBR322 DNA. Restriction enzyme sites are as follows: H, *Hind*III; P, *Pst*I; Pv, *Pvu*II; R, *Eco*RI; S, *Sal*I; Sa, *Sac*I; Sn, *Sna*BI; Sp, *Sph*I. The large arrow indicates the direction of transcription of ORF2. *Bal*31 digestion of the *Sal*I/*Eco*RI insert cloned in pKG1 resulted in fragments which were ligated into pUC18 or pUC19, as described in Methods, to produce pKH9, pKH10 and pKH11. The CMCase phenotype of *E. coli* JM83 harbouring the respective plasmids is shown. Recombinant *E. coli* HB101 harbouring pKG2 was isolated from a genomic library constructed in pBR322 by using a radiolabelled probe consisting of the *Sau*3A fragment between map positions 3662 and 4605 of pKG1.

GGTAACTACTAGGAAACGTAATAAATACGTAAGCAGAGTGAATTTTACATTTA
CTCTGCTTTTAAATACCAAAAGGCATGTCGGAATAATTTTGAATCTATACGGC
AATTTTCTGAAATTTGGTTTATCAAAATTTGACGAAGCATATATGATATAATAACTAA
TAAACGGGATAATATAGATTACTTAAATTAAGTTGTAATTTGATTTTATTAATTG
AGGAGGAGGATGCTATGAGGTGGTTACAGTTTGGGAAGAAGAAATTCMTTGTATA
M R L V N S L G R R K I L L I
CTTGCTGTTATTTAGCTTTTACGACGTGTTCTGTTGTTGCAAGCTATGGGGCGAAAG
L A V I V A F S T V L L F A K L W G R K
ACTTCAAGTACTTTGGATGAGGTGGTTCAAACTCATGTTGATTTGACGGCAGAAAT
T S S T L D E V G S K T H G D L T A E N
AAAAACGGCGGATATTACAGAGGAAGAGATTCCAGATCAGCCTCCGGCAACCGGGGCC
K N G G Y L P E E E I P D Q P P A T G A
TTCAACTACGGTGAAGCGTTGCAAAAGCAATTTTCTATGAGTGTCAGAAATCCGGA
F N Y G E A L Q K A I F F Y E C Q R S G
AAGTCGATCCCTCAACTCTTCCGCTAAATTTGGCGGGAGATTCCGGACTGGATGACGGA
K L D P S T L R L N W R G D S G L D D G
AAAGATGACGGAATTGATCTTACCGCGGATGGTATGATGCGGAGATCAGCTAAATTT
K D A G I D L T G G W Y D A G D H V K F
AATTTGCCATGCTTATTCCGGCGCTATGCTGGGGTGGCGGTGTATGAATGAAGAT
N L P M S Y S A A M L G W A V Y E Y E D
GCGTTTAAACAGAGTGGACAGTATAACCATATTTGAATAACATAAAATGGGCTTGTGAT
A F K Q S G Q Y N H I L N N I K W A C D
TATTTTATAAATGTCATCCGAAAAGGATGTGACTATTACAGGTGGGCGACGGTCAT
Y F I K C H P E K D V Y Y Y Q V G D G H
GCTGACCATGCGTGGTGGGTCTCGCAAGTATGCCTATGGAAGGCGCTGCTACAAA
A D H A W W G P A E W A M P M E R P S Y A K
GTTGACAGGTCATCGCCGGTTCACGGTTGTGGCAGAGACGTCGGCAGCTTTAGCTATA
V D R S S P G S T V V A E T S A A L A I
GCATCGATAATATTAAGAAAGTGGATGGTAATCTCGAAAGATGTTTAAAGCATGCA
A S I I F K K V D G E Y S K E C L K H A
AAAGAACTGTTGAATTTGCGGACACCAAAAGCGATGATGGGTACATGCAAGCCAAT
A L F E F A D C T K S D D G Y T A C N
GGTTTTATAATTCATGGAGCGGATTTTATGATGAGCTTCTCGGGCAGCTGTATGGCTT
G F Y N S W S G F Y D E L S W A A V W L
TATCTTGCTACCAATGATTCTTTCATATTTGGATAAAGCGGAAAGTATTCTGACAAATGG
Y L A T N D S S Y L D K A E S Y S D K W
GGTTATGAGCCACAGACGAACATACCGAAGTATAAGTGGGCTCAATGCTGGGATGATGTG
G Y E P Q T N I P K Y K W A C C W D D V
ACTTATGGCACTTCTTCTTTGGCCAGGATTAATAACGCAAGGAAATATAAGAA
T Y G T Y L L L A R I K N D N G K K Y K E
GCGATAGAAAGGCATCTTGATTTGGTGGACACCGGATACACGGTGAAGAAATACATAT
A I E R H L D W W T T G Y N G E R I T Y
ACTCGAAGGACTTGCATGGCTCGACAGTGGGGATCATTGAGGTATGCAACCAACAG
T P K G L A W L D Q W G S L R Y A T T T
GCATTTCTGGCATGTTTATTCCGATTTGGGAGAACGGTGATAAGGAAAAAGCAAAATCT
A F L A C V Y S D W E N G D K E K A K T
TATCTGGAGTTTGAAGAAGCCAGGCGGATTATGCTTTGGGAAGCACGGGAAGAGCTTT
Y L E F A R S Q A D Y A L G S T G R S F
GTTGTGGGTTTGGAGAAAATCCACGAAAAGGCCCATCAGAACTGCTCACGGTTCA
V V G F G E N P P K R P H H R T A H G S
TGGCGGACAGTCAGATGGAGCCTCCGCAACACAGGATGTTCTTTATGTTGCCCTTTGTG
W G A S Q M E P P E H R H V L Y G A L V
GGAGGACCTGACAGCAGGACAACTACACGACGACATCAGTAATTACACCTGCAATGAA
G G P D S T D N Y T D D I S N Y T C N E
GTTGCGTGTGACTATAATGTCAGGTTTGTGGGACTGCTTGC AAAATGTACAAGCTTTAT
V A C D Y N A G F V G L L A K M Y K L Y
GGCGGAAGTCCCGATCCCAAAATTTACCGGTATAGAAGAAGTTCCGGAGGATGAATATTC
G G S T D P K F P D E I E E V P D E I F
GTTGAAGCCGGTGTGAATGCATCGGGAACAATTTCAATTGAAATAAAAGCGATAGTTAAT
V E A G V N A S G N N F I E I K A I V N
AATAAATCGGGCTGCGCTGCAAGAGTATGTGAGAATTTATCTTTAGATATTTTATCAAC
N K S G W P A R V C E N L S F R Y F I N
ATTGAAGAGATTGTGAATGCGGGAAGGTGAAGCGACCTGCAAGTGAGCTCCAGCTAC
I E E I V N A G K S A S A D L Q V S S Y C
AATCAGGGGCAAACTGTCGATGTAAAGCACTACAAGCAATATTTTATTTATGTGGAA
N Q G A K L S D V K H Y K D N I Y Y V E

GTGGATTGTGCGGGACAAAAATATATCCCGGGGCAATCGGCATACAAGAAGGAAGTG
V D L S G T K I Y P G G Q S A Y K K E V
CAGTTTAGAATTTCCGCGCGGAGGCGACGGTGTATTAATCCGGAACGACTATTCCTAT
Q F R I S A P E G T V F N P E N D Y S Y
CAGGGACTTTCGGCAGGTACGGTTGTAAGTCTGAGTATATTCGGTATATGATGCGGG
Q G L S A G T V V K S E Y I P V Y D A E
GTGCTGTTATTTGAAGGGAACCGGGCTCAGCATCGAAAAGCAGCTCTAAGACAAATGGT
V L V F G R E P G S A S K S T S K D N G
TTGTCCAAGGCACTCCACGGTGAATACTGAATCTCAGCCGACAGCAAAACACACTCAA
L S K A T P T V K T E S Q P T A K H T Q
AATCTGCTCAGACTTTAAACTCCAGCAATCAGAACAGTGTAAAAAAGACCAAGGC
N P A S D F K T P A N Q N S V K K D Q G
ATAAAGGAGAAGTGTATTACAGTACGCAACGGGAATGCGAGTGTACGTCAACAGT
I K G E V V L Q Y A N G N A G A T S N S
ATTAATCCGAGGTTTAAATAATTAACAACGGTACAAAAGCCATAAATTTGTCGATGTC
I N P R F K I I N N G T K A I N L S D V
AAGATTAGATATTATTACAAAAAGAGGGGCGCATCTCAAACTTCTGGTGTGATTGG
K I R Y Y Y T K E G G A S Q N F W C D W
AGCAGTCCGGCAATTCAAATGTTACAGAACTTCTTTAATCTTTCTTCCCGAAAGAA
S S A G N S N V T G N F F N L S S P K E
GGAGCGGACCTGTCTTGAAGTTGGTTTTCGGAAGTGGGGCCGGAACCTTGATCTGGT
G A D T C L E V G F G S G A G T L D P G
GGAAGCGTTGAAGTACAGATAAGGTTTCAAAGGAAGACTGGTCAAACTATAACAGTCA
G S V E V Q I R F S K E D W S N Y N Q S
AACGATTATTTCTTCAAAACAGCTTGTTCACGGCAAGAACCTTGATTATCTGTACGCA
N D Y S F K Q A C L R Q R T L I Y L Y A
ACATGGCTAAGATAGGCTTTTGCAATACCGATGTAATATTCGGATTCTCTTACAATGTGA
T W L R *
TTAATTACACAGAGGCTGTCAAATGTTTGGCACTGCGGCACCTCTCTTTTGGCATGCA
TTAGGAAGTTTACGAAGGTTTGACTTTGACGTAGTGTATTTTCGAAT

Fig. 3. Nucleotide sequence of the *cell* gene of *C. thermocellum* and the deduced amino acid sequence of endoglucanase I (Cell). The proposed ribosome-binding site is boxed and the experimentally determined amino acid sequence corresponding to the N-terminus of mature Cell, produced by *E. coli* JM83, is underlined. The 3' ends of the truncated fragments cloned in pKH10 and pKH11 are denoted by black triangles.

(dATP^[35S]; 37 MBq mmol⁻¹) and deoxycytidine 5' [α -³²P]triphosphate (dCTP^[32P]; 37 MBq mmol⁻¹) were from Amersham. All other reagents were of a purity appropriate to their application.

Results

Nucleotide sequence of *C. thermocellum* DNA cloned in pRV1.6

Previous work (Romaniec *et al.*, 1987) showed that *C. thermocellum* DNA cloned in pRV1.6 encoded a functional endoglucanase, here designated Cell. In the course of the present study it became apparent that the *cell* gene was incomplete and a second overlapping *Sau*3A fragment containing the 3' extremity of the gene was isolated. Full-length *cell* was reconstituted by ligating the two genomic fragments (Fig. 2; pKG3). To confirm that the cloned copy of *cell* is the same as *cell* contained in the *C. thermocellum* genome, and does not result from spurious ligation, ³²P-labelled pRV1.6 was used to probe *C. thermocellum* genomic DNA. The pattern of hybridization (Fig. 1) was consistent with the proposed physical map of *cell*.

C. thermocellum DNA cloned in pKG3 and coding for CellI was sequenced in both strands. Translation of the nucleotide sequence revealed two open reading frames (ORFs). ORF1 (map positions 189 to 1859, Fig. 2) consisted of 1671 nucleotides (nt) coding for a polypeptide of 556 amino acids with a deduced M_r of 61042 and terminating in a TAG stop codon; the putative translational start was preceded (9 nt) by the sequence GAGG, which shows some similarity to the ribosome-binding sequence most frequently observed in *C. thermocellum* cloned *cel* and *xyn* genes (AGGAGGA). The 3' flanking sequence contained a 10 nt palindrome with the potential to form a stem-loop structure characteristic of factor-independent transcription termination. ORF2 (map positions 2106 to 4745, Fig. 2) comprised 2640 nt coding for a polypeptide of 879 amino acids with a deduced M_r of 98531. It was also terminated by a TAG stop codon and the proposed translational start was preceded (4 nt) by the sequence AGGAGG, which is complementary to the 3'-terminus of 16S rRNA from Gram-positive bacteria (McLaughlin *et al.*, 1981).

Within the 243 nt separating ORF2 from ORF1, the A + T content was 74.5 mol%; immediately upstream of ORF1, the A + T content was 61.4 mol%.

Identification of the *cell* gene

Plasmid pKG1, containing the *SalI*/*EcoRI* fragment from pRV1.6 ligated into pUC19, directed the synthesis of a functional endoglucanase in *E. coli* JM83. This suggested that the last 140 nt of ORF2 were not required for expression of CellI activity. To determine how much more of the insert could be deleted without affecting activity and to establish which of the two ORFs coded for CellI, pKG1 was linearized with *SalI* and digested with *Bal31* to remove progressively larger amounts of ORF1; the truncated fragment, ligated into pUC19, was then transformed into *E. coli* JM83. The most extensive deletion carried out removed almost 70% of ORF1 without resulting in loss of carboxymethylcellulase (CMCase) activity (Fig. 2; pKH9). Similarly, the removal of 1000 nt from the 3' end of ORF2 with *Bal31* (Fig. 2; pKH10) left CMCase activity unaffected, but the deletion of 1142 nt from the 3' end of ORF2 resulted in the complete loss of CMCase activity. These results suggest that CellI is encoded by ORF2 and furthermore that the catalytic region of the enzyme is located within the first 550 residues of this 879 residue polypeptide. Further evidence that ORF2 is *cellI* was obtained by purifying recombinant CellI from a cell-free extract of *E. coli* JM83 harbouring pKH9. Twelve of the first 14 experimentally determined N-terminal residues of recombinant CellI were identical to the deduced sequence between residues 56 and 69 of the polypeptide encoded by ORF2.

Characteristics of CellI

The deduced amino acid sequence of CellI is shown in Fig. 3. Codon usage was similar to that described for *xynZ* and for each of the sequenced *cel* genes of *C. thermocellum*, with a slight bias towards A or T in the wobble position reflecting the relatively low G + C content of the genomic DNA. With the exception of CelC (Schwarz *et al.*, 1988), the secreted endoglucanases of *C. thermocellum* contain putative signal peptides. At first glance, the same would appear to be true for CellI; the N-terminus of the deduced protein contains positively charged basic residues preceding a core of 18 residues of which 16 are hydrophobic. Beyond this point, however, the similarity with signal sequences of secreted prokaryotic proteins is less evident and there is no obvious signal peptide cleavage site which conforms with accepted criteria (Von Heijne, 1988). Since only Ala, Gly or Ser residues appear to be tolerated in the -1 position of bacterial signal peptides, the experimentally determined N-terminus of mature CellI synthesized by *E. coli*, beginning at residue 56 of the translated sequence, is more likely to indicate aberrant processing than a natural cleavage site. Certainly, a functional signal peptide was not recognized by *E. coli* harbouring *cellI* since endoglucanase activity remained in the cytoplasm exclusively and was not transported to the periplasm (data not shown). The highly conserved region consisting of duplicated segments of approximately 24 amino acids which has to date been found in XynZ and in each of the *C. thermocellum* endoglucanases except for CelC was absent from CellI. Furthermore, extended sequences rich in hydroxy amino acids and proline, which occur commonly in cellulase-related proteins, dividing the enzymes into discrete structural and functional domains, were not really apparent in CellI.

Homology with other cellulases

Comparison of the amino acid sequence of CellI with those of other cellulases revealed that CellI belongs to cellulase family E as defined by Henrissat *et al.* (1989). Extensive homology (Table 1) between CellI and CelF from *C. thermocellum* (Navarro *et al.*, 1991) and between CellI and two other cellulases [CenB from *Cellulomonas fimi* (Meinke *et al.*, 1991) and CelZ from *Clostridium stercoarum* (Jauris *et al.*, 1990)], which were recently shown to be very similar to CelF (Navarro *et al.*, 1991), indicates that CellI is a subfamily E₂ endoglucanase. Identity with the subfamily E₁ endoglucanases amounted to only 20 to 25% (Table 1). Comparison of the region of CellI between residues 71 and 574 with the sequences of CelF, CenB and CelZ revealed percentage identity values of 55.5%, 47.7% and 65.4% respectively (Fig. 4).

Cell	PATGAFTNYG	EAALQKA	IFFYEC	QRSGL	LD	PSTLRLN	WRGDS	GL	DDGKD	AGI	DLTG	GWYD	AGDH	VKFN	LP	MSYSA	143																																																								
CenB	AAAPTYYNYA	EAALQKSM	FFYQA	QRSGL	LD	PADFPVSW	RGDS	GL	TDGAD	VGK	DLTG	GWYD	AGDH	VKFG	FPM	AFSA	103																																																								
CelF	SFAADFNYG	EAALQKA	IMFFYEF	QRSGL	LD	PENKRNN	WRGDS	AL	NDGAD	NGL	DLTG	GWYD	AGDH	VKFN	LP	MAYAV	96																																																								
CelZ	AEAAGYNYG	EAALQKA	IMFFYEF	QRSGL	LD	PENKRDN	WRGDS	GL	NDGAD	VGL	DLTG	GWYD	AGDH	VKFN	LP	MAYSQ	96																																																								
Cell	AMLGW	AVY	EYEDAF	KQS	G	QYNHIL	NNIKWAC	DYFI	K	CH	PEKDV	Y	Y	QVGDG	HA	DHAW	WGPA	EVMP	M	ERP	S	Y	K	V	216																																																
CenB	TMLAW	GAI	ESPTGY	SKA	G	SLDELK	DNLR	FVSD	DYFV	K	AHTAP	NEL	Y	V	QVGDG	EAD	HKW	WGPA	EVMT	M	ARP	S	H	K	I	176																																															
CelF	TMLAW	SVY	ESRDAY	VQSG	QLPY	ILDNI	KWAT	DYFI	K	CH	PSPNV	Y	Y	Y	QVGDG	AL	DHS	W	WGPA	EVMT	M	PRP	S	F	K	V	169																																														
CelZ	TMLAW	AAAYE	AEEAL	ERS	G	QMGYLL	DAIK	WVSD	DYLI	K	CH	PSPNV	F	Y	Y	QVGDG	HL	DHS	W	WGPA	EVMT	M	DRP	A	Y	K	V	169																																													
Cell	DRSS	PGST	VVAET	S	AAL	L	A	IASI	IFKKV	D	GEY	SKECL	K	HAKEL	L	F	E	FADTT	KS	D	—	DGYT	A	ANG	F	Y	N	S	W	S	G	F	Y	D	E	287																																					
CenB	SASC	PGSD	VAAET	A	AAL	A	S	SAI	VLKGD	D	PAY	AATLV	SHAK	QLYT	F	ADTY	R	G	A	Y	S	D	C	V	T	A	A	S	A	Y	Y	K	S	W	S	G	Y	Q	D	E	249																																
CelF	DLTN	PGST	VVAET	A	AAM	A	S	S	I	VFKPT	D	PEY	AATLL	RHAK	ELFT	F	ADTT	R	S	D	—	AGYR	A	A	E	G	Y	Y	S	S	H	S	G	F	Y	D	E	240																																			
CelZ	DLAN	PGST	VVAE	A	AAA	A	L	A	S	A	V	F	ADR	D	P	A	Y	A	A	T	C	I	Q	H	A	K	E	L	Y	N	F	A	E	I	T	K	S	D	—	S	G	Y	T	A	A	S	G	F	Y	D	S	H	S	G	F	Y	D	E	240														
Cell	LSW	A	A	V	W	L	Y	L	A	T	N	D	S	S	Y	L	D	K	A	E	S	Y	S	D	K	W	G	Y	E	P	Q	T	N	I	P	K	Y	K	W	A	Q	C	W	D	V	T	Y	G	T	Y	L	L	A	R	I	K	N	D	N	G	K	Y	K	E	A	I	E	R	H	360			
CenB	LVW	G	A	Y	W	L	Y	K	A	T	G	D	A	T	Y	L	A	K	A	E	A	B	Y	D	K	L	G	T	E	N	Q	S	T	T	R	S	Y	K	W	T	I	A	W	D	N	K	Q	F	G	T	Y	A	L	L	A	—	M	E	T	G	K	Q	K	Y	V	D	D	A	N	R	W	321	
CelF	LTW	A	S	I	W	L	Y	L	A	T	G	D	Q	S	Y	L	D	K	A	E	S	Y	E	P	H	W	E	R	E	R	G	T	T	L	I	S	Y	S	W	A	H	C	W	D	N	K	L	Y	G	S	L	L	L	L	A	K	I	—	T	G	K	S	Y	Y	K	Q	C	I	E	N	H	312	
CelZ	LSW	A	G	V	W	L	Y	L	A	T	G	D	E	T	Y	L	N	K	A	E	Q	Y	V	A	Y	W	G	T	E	P	Q	T	N	I	S	Y	K	W	A	H	C	W	D	V	H	Y	G	A	C	L	L	L	A	K	I	—	T	G	K	W	I	Y	K	E	A	I	E	R	H	312			
Cell	LDW	W	T	T	G	Y	N	G	E	R	I	T	Y	T	P	K	G	L	A	W	L	D	Q	W	G	S	L	R	Y	A	T	T	T	A	F	L	A	C	V	Y	S	D	W	E	N	G	D	K	E	K	A	K	T	Y	L	E	F	A	R	S	Q	A	D	Y	A	L	G	S	T	G	R	433	
CenB	LDY	W	T	V	G	V	N	G	Q	K	V	P	Y	S	P	G	G	Q	A	V	L	D	S	W	G	A	L	R	Y	A	A	N	T	S	F	V	A	L	V	Y	S	D	W	M	T	D	A	T	R	K	A	R	—	Y	H	D	F	G	V	R	Q	I	N	Y	A	L	G	D	N	P	R	393	
CelF	LDY	W	T	V	G	F	N	G	S	R	V	Q	Y	T	P	K	G	L	A	Y	L	D	R	W	G	S	L	R	Y	A	T	T	Q	A	F	L	A	S	V	Y	A	D	W	S	G	C	D	P	A	K	A	A	V	Y	K	E	F	A	K	K	Q	V	D	Y	A	L	G	S	T	G	R	385	
CelZ	LDY	W	S	V	G	Y	N	G	E	R	V	H	Y	T	P	K	G	L	A	W	L	D	S	W	G	S	L	R	Y	A	T	T	T	A	F	L	A	S	V	Y	A	D	W	E	G	C	S	R	E	K	A	A	I	Y	N	D	F	A	K	Q	Q	I	D	Y	A	L	G	S	S	G	R	385	
Cell	—	S	F	V	V	G	F	G	E	N	P	P	K	R	P	H	H	R	T	A	H	G	S	W	A	D	S	Q	M	E	P	P	E	H	R	H	V	L	Y	G	A	L	V	G	G	P	D	S	—	T	D	N	Y	T	D	D	I	S	N	Y	T	C	N	E	V	A	C	D	Y	N	A	G	503
CenB	S	S	Y	V	V	G	F	G	A	N	P	P	T	A	P	H	H	R	T	A	H	G	S	W	L	D	S	I	T	T	P	A	Q	S	R	H	V	L	Y	G	A	L	V	G	G	P	G	S	P	N	D	A	Y	T	D	S	R	Q	D	Y	V	A	N	E	V	A	T	D	Y	N	A	G	468
CelF	—	S	F	V	V	G	F	G	K	N	P	P	R	N	P	H	H	R	T	A	H	S	S	W	S	A	L	M	T	E	P	A	E	C	R	H	I	L	V	G	A	L	V	G	G	P	D	G	—	S	D	S	Y	V	D	R	L	D	D	Y	Q	C	N	E	V	A	N	D	Y	N	A	G	455
CelZ	—	S	Y	V	V	G	F	G	V	N	P	P	K	R	P	H	H	R	T	A	H	S	S	W	A	D	S	M	S	V	P	D	Y	H	R	H	V	L	I	G	A	L	V	G	G	P	G	K	—	D	D	S	Y	T	D	D	I	N	N	Y	I	N	N	E	V	A	C	D	Y	N	A	G	455
Cell	FVGL	LAKMY	KLYGG	S	P	D	P	K	F	N	G	I	E	E	V	P	E	D	E	I	F	V	E	A	G	V	N	—	A	S	G	N	N	F	I	E	I	K	A	I	V	N	N	K	S	G	W	P	A	R	V	C	E	N	L	S	F	R	Y	F	I	574											
CenB	FTS	A	L	A	R	L	V	E	E	Y	G	G	T	P	L	A	S	F	P	T	P	E	Q	P	D	G	D	Q	L	F	V	E	A	M	L	N	—	A	A	G	P	G	F	T	E	V	K	A	M	I	R	N	K	S	G	S	F	P	A	R	S	L	K	N	A	K	V	R	Y	W	F	538	
CelF	FVGA	LAKMY	EKYGG	E	P	I	P	N	F	V	A	F	—	E	T	P	G	E	E	F	V	E	A	A	V	N	—	A	A	G	P	G	F	V	N	I	K	A	S	I	I	N	K	S	G	W	P	A	R	G	S	D	K	L	S	A	K	Y	F	V	525												
CelZ	FVGA	LAKMY	EDYGG	S	P	I	P	D	N	L	A	F	E	E	I	T	N	D	E	F	F	V	M	A	G	I	N	—	A	D	G	Q	N	F	I	E	I	K	A	L	L	H	N	Q	S	G	W	P	A	R	V	A	D	K	L	S	F	R	Y	F	V	526											

Fig. 4. Comparison of the amino acid sequence (between residues 71 and 574) of CellI from *Clostridium thermocellum* with the sequences of the catalytic domains of *Cellulomonas fimi* CenB (Meinke *et al.*, 1991), *Clostridium thermocellum* CelF (Navarro *et al.*, 1991) and *Clostridium stercorarium* CelZ (Jauris *et al.*, 1990). Residues identical in all four sequences are boxed. Numbers at the end of each sequence indicate the positions of the residues within their respective proteins.

Table 1. Percentage sequence identity between CellI from *C. thermocellum* and other cellulase subfamily E endoglucanases

Subfamily	Enzyme*	Organism	Percentage sequence identity	Reference
E ₁	CelD	<i>Clostridium thermocellum</i>	23.2	Joliff <i>et al.</i> (1986)
	CelA	<i>Pseudomonas fluorescens</i> subsp. <i>cellulosa</i>	21.6	Hall & Gilbert (1988)
	Ced1	<i>Butyrivibrio fibrisolvens</i>	25.2	Berger <i>et al.</i> (1990)
	CenC	<i>Cellulomonas fimi</i>	19.6	Coutinho <i>et al.</i> (1991)
E ₂	EG	<i>Persea americana</i>	39.2	Tucker <i>et al.</i> (1987)
	SGSP	<i>Dictyostelium discoideum</i>	35.0	Giorda <i>et al.</i> (1990)
	CelF	<i>Clostridium thermocellum</i>	55.5	Navarro <i>et al.</i> (1991)
	CenB	<i>Cellulomonas fimi</i>	47.7	Meinke <i>et al.</i> (1991)
	CelZ	<i>Clostridium stercorarium</i>	65.4	Jauris <i>et al.</i> (1990)

* Cel, Cen and EG denote endoglucanases; SGSP is a spore germination-specific protein and Ced is a cellodextrinase.

As one might predict, conservation of sequence between the four enzymes was 100% in the region of the histidine residue (His-448 in CellI) which has been shown to be present in the catalytic centre of CelD and, by inference, other family E endoglucanases (Tomme *et al.*, 1991). Homology observed between CellI and the catalytic domains of CenB, CelF and CelZ failed to extend beyond the first 574 residues of CellI, suggesting that the catalytic region of the enzyme is located within the N-terminal 65% of the polypeptide. Our observation that over 300 residues may be removed from the C-terminus of CellI without impairing catalytic function would appear to support this view. This raises the question as to the role of the apparently redundant C-terminal region. In CelE, from the same organism, a cellulose-binding domain, distinct from the catalytic domain, is located in the C-terminal half (Durrant *et al.*, 1991), but functions have yet to be assigned to the extended non-catalytic sequences in other enzymes from *C. thermocellum*, notably XynZ and CelH. However, it is worthy of note that the 150 residues proximal to the C-terminal of CellI exhibit 46% sequence identity (68% similarity) with the 167 residue recently shown by Poole *et al.* (1992) to mediate binding between cellulose and the cellulosome S1 subunit from *C. thermocellum* YS. It is therefore likely that the C-terminal region of CellI constitutes a cellulose-binding domain.

Substrate specificity

It was recently reported (Navarro *et al.*, 1991) that, in addition to sharing extensive sequence identity, CelF from *C. thermocellum* and CelZ from *C. stercorarium* exhibit similar biochemical characteristics, including little or no activity against chromogenic cellobiosides and high activity against barley β -glucan. In view of the substantial homology between CellI, CelF and CelZ, it

is not surprising that recombinant CellI exhibited considerable activity against barley β -glucan, being over 100 times more active against this substrate than against CMC (Table 2). Significant activity was recorded with lichenan as substrate, but activity against laminarin, xylan, crystalline or acid-swollen cellulose and the chromogenic substrates *p*-nitrophenyl glucoside and *p*-nitrophenyl cellobioside was very low.

Production of CellI by *E. coli* and *C. thermocellum*

CellI produced by *E. coli* JM83 harbouring full-length *cellI* was shown by Western analysis to vary in size from 102 kDa to 60 kDa, indicating that, in common with other cellulases expressed in *E. coli* from cloned genes, CellI is subject to proteolytic processing (Fig. 5). The size of the largest form produced (102 kDa) agrees well with the size of CellI deduced from the translated gene sequence.

To determine whether the *cellI* gene is normally expressed by *C. thermocellum*, total extracellular cellulase from a culture of strain NCIB 10682 and the cellulosome

Table 2. Substrate specificity of CellI

Substrate	Specific activity*
Barley β -glucan	108.7
Carboxymethylcellulose	0.8
Lichenan	2.4
Laminarin	0.01
Xylan	< 0.01
Acid-swollen cellulose	< 0.01
Avicel	< 0.01
Filter paper	< 0.01
<i>p</i> -Nitrophenyl glucoside	0.01
<i>p</i> -Nitrophenyl cellobioside	0.02

* nmol product min⁻¹ (mg protein)⁻¹.

Fig. 5. Production of Cell by *E. coli* and *C. thermocellum*. Cell-free extract from *E. coli* JM83 harbouring pKG3 (lanes 1 and 2), total extracellular proteins from *C. thermocellum* NCIB 10682 (lanes 3 and 6) and cellulosome fractions from *C. thermocellum* NCIB 10682 (lanes 4 and 7) and strain YS (lanes 5 and 8) were fractionated by SDS-PAGE (10% w/v, acrylamide) and were stained with Coomassie Blue (lanes 1, 3, 4 and 5) or transferred to Immobilon P (lanes 2, 6, 7 and 8). Proteins immunoreactive with anti-CellI antiserum were detected as described in Methods. The sizes of molecular mass markers are shown on the extreme left in kDa.

fractions from cultures of strains NCIB 10682 and YS were analysed by Western blotting, using antiserum prepared against purified recombinant CellI (Fig. 5). Major immunoreactive bands with M_r values of 86000 and 240000 were detected in the ethanol-precipitated proteins from cell-free culture supernatant of strain NCIB 10682. Proteins of similar size which cross-reacted with anti-CellI antiserum were also evident in the cellulosome fraction. The same results were obtained for cellulosomes prepared by two different procedures. A similar experiment conducted with cellulosomes purified from a culture of *C. thermocellum* strain YS revealed cross-reacting bands of 88 kDa and about 215 kDa (Fig. 5).

Identity of the protein encoded by ORF1

A search of protein sequence databases revealed that the polypeptide encoded by ORF1 exhibited 53% similarity and 30% identity overall with the *tar* gene product from *E. coli*; similarity was most apparent in the C-terminal halves of the respective sequences (Fig. 6). The *tar* gene belongs to a family of genes encoding structurally related transducer proteins which act as chemoreceptors and function in generating signals modulating flagellar rotation (Krikos *et al.*, 1983; Gardina *et al.*, 1992). Typically, members of this family of transmembrane proteins are composed of distinct functional domains,

```

.....MRYNTAKLLVYDAEQVSGIEKIQEHVENT
MINRIRVVTLTLLVMVLGVFALLQLISGSFFSSLHHSQKSFVVSQNLREQQ
EKYLSLYESTVINETDRIQLQELKALWEKYKSLVDKEVELVMSCKTEEAR
GELTSTWD.LMLQTRINLSRSASVRMMMDSSNQSNAKVELLSARKTLAQ
QL.....LLSDIDDITGDTLRDYFAFVEY.....NTITAA
AATHYKKFKSMAPLPEMVATSRNIDEKYKNVYATLTELIDYLDYGNICAY
KEKVDENKQMASTASTVMIVVIFVVGILITATLGVFISRTISKPIGQMVFA
FAQPTQGMQAMGERFAQ.....VALSSEKLYRDIIVTONADDYRFA
ADRLA.LGVVEVDVKAETRDITGK.LAESFKRMTENTREDAYVVERTAAG
QWQLAVIAIVVLLVAVWYGIIRMLLTPLAKITIAHIRE.....IAGC
DM..TVDVRVKSOKDILLGKKLKEMVDVTNNEVLSNINEVAAQVAGAKQVS
NLANTLTIDGRSEMGDLAQSVSHMORSRLDTVTTHVREGSDAIYAGITREIA
DSSMQLSQGATEQASSIEELTASLEQVANQTQLSAKNANQANPLAEVAKN
AGNTDLSRTEQASALEETAASMEQLTATVKQNADNARQASQLAQASAD
NAEQNKQMAEMLNMBEITNNSSNISRTIKVIDETAFQTNILALNAAVE
TAHSGSKVVDGVVKTMEIADSSKKIADIIISVIDGTAFTQTNILALNAAVE
AARAGOHGKGFVAVVAEEVRNLAARSANAAKETTELICTIKRTENGTKIA
AARAGEQGRGFVAVVAGEVRNLAARSAAAKEIKALIEDSVSRVDTGSVLV
RETAELINKIVEGSKAATLVNDIAMASNEQAAATITQINQIAQVSCVWQ
ESAGEITMNIIVNAVTRVTDIMGETASASDEQSRGIDQVALAVSEMQRWQ
TNSATSEESAAASEELSSQAEILKRSTAKFKLKNMGKMTSNRYKEVSPEI
QNASLVQESAAAAAALAEQASRLTQAVSAFRLAA.....
MRMLEDTYENKQPKSYKEENGEYS DGKETAEKDVGGGLKQKILLSDSEFG
.....SPLTNKHQTES.....RPASEQPPAQPRLRTAEQDPNWE
KY*
TF.

```

Fig. 6. Comparison of the amino acid sequence obtained by translating ORF1 (upper sequence, this work) with the amino acid sequence of the *E. coli tar* gene product (lower sequence, Krikos *et al.*, 1983). Identical residues are boxed.

and exhibit homology primarily in their C-terminal sequences, which contain the methylation sites involved in sensory adaptation. Based on this comparison of primary sequences, it is possible that the polypeptide encoded by ORF1 is a sensory transducer protein, similar in structure to the product of the *E. coli tar* gene. The likely functions of such a protein are to bind attractants or detect changes in the extracellular concentration of soluble sugars. Further work is required to establish if the putative polypeptide encoded by ORF1 is expressed by *C. thermocellum*, and whether there is any functional involvement with the cellulase system.

Discussion

Nucleotide sequencing has shown that a fragment of *C. thermocellum* genomic DNA, previously cloned in *E. coli* and shown to direct the synthesis of endoglucanase activity, contains two ORFs. *Bal31* deletion studies and comparison of the N-terminal sequence of the purified recombinant endoglucanase with the translated sequences derived for each of the ORFs confirmed that ORF2 encoded the endoglucanase, CellI.

Sequence alignments indicated that CellI is highly

homologous with CelF from *C. thermocellum* and with cellulases from *Cellulomonas fimi* (CenB) and *Clostridium stercoarium* (CelZ), and is a member of cellulase subfamily E₂; endoglucanases from *Persea americana* and *Dictyostelium discoideum* are also included in this family, suggesting that at some stage during cellulase evolution, gene transfer between prokaryotes and eukaryotes may have occurred. Based on deletion analysis and sequence comparisons, the catalytic domain of Cell is analogous in size to that of other family E endoglucanases, and is located within the first 550 N-terminal amino acids. The remainder of the polypeptide is not required for catalytic activity and, at its C-terminus, contains a putative cellulose-binding domain.

Cell, like CelC from *C. thermocellum*, lacks the 24 amino acid repeated sequence which has been found in xylanase Z and in each of the other *C. thermocellum* endoglucanases sequenced to date. Also, although the primary structure of Cell bears little resemblance to that of CelC, the substrate specificity of Cell is markedly similar to that of CelC, in that barley β -glucan was the preferred substrate and relatively little activity was seen against other cellulosic substrates.

Western analysis carried out using antiserum raised against purified recombinant Cell revealed that the *cell* gene is expressed by two different strains of *C. thermocellum* and that the product, Cell, is either homologous to cellulosome proteins or is itself a component of the high- M_r cellulase complex. Total extracellular proteins and cellulosomes from strain NCIB 10682 both contained proteins of 86 kDa and 240 kDa which were immunoreactive with anti-Cell antiserum. In cellulosomes from strain YS, cross-reacting proteins were 88 kDa and 215 kDa in size. These results suggest that Cell synthesized by *C. thermocellum* is subject to proteolytic processing. Alternatively, one might argue that, in view of the homology between Cell and CelF (an endoglucanase of 82 kDa), cross-reaction between anti-Cell antiserum and CelF could account for at least one of the immunoreactive bands visible in Fig. 5.

In conclusion, our results suggest that Cell is a component of the *C. thermocellum* cellulosome even though it lacks the repeated sequence which has been proposed to mediate the binding of xylanase Z and the majority of endoglucanases to the large non-catalytic subunit during cellulosome assembly. However, methods used for purifying cellulosomes depend upon their affinity for cellulose, and the presence of a putative cellulose-binding domain at the C-terminus of Cell may explain why the native enzyme copurifies with cellulosomes.

We wish to thank the AFRC (Grant no. LRG138) for supporting this work.

References

- BÉGUIN, P. (1990). Molecular biology of cellulose degradation. *Annual Review of Microbiology* **44**, 219–248.
- BÉGUIN, P., CORNET, P. & AUBERT, J.-P. (1985). Sequence of a cellulase gene of the thermophilic bacterium *Clostridium thermocellum*. *Journal of Bacteriology* **162**, 102–105.
- BERGER, E., JONES, W. A., JONES, D. T. & WOODS, D. R. (1990). Sequencing and expression of a cellobiohydrolase (*ced1*) gene from *Butyrivibrio fibrisolvens* H17c cloned in *Escherichia coli*. *Molecular and General Genetics* **223**, 310–318.
- BIRNBOIM, H. C. & DOLY, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Research* **7**, 1513–1523.
- CHAUVAUX, S., BÉGUIN, P., AUBERT, J.-P., BHAT, K. M., GOW, L. A., WOOD, T. M. & BAIRDOCH, A. (1990). Calcium-binding affinity and calcium-enhanced activity of *Clostridium thermocellum* endoglucanase D. *Biochemical Journal* **265**, 261–265.
- COUTINHO, S. B., MOSER, B., KILBURN, D. G., WARREN, R. A. J. & MILLER, R. C. (1991). Nucleotide sequence of the endoglucanase C gene (*cenC*) of *Cellulomonas fimi*, its high-level expression in *Escherichia coli*, and characterization of its products. *Molecular Microbiology* **5**, 1221–1233.
- DURRANT, A. J., HALL, J., HAZLEWOOD, G. P. & GILBERT, H. J. (1991). The non-catalytic C-terminal region of endoglucanase E from *Clostridium thermocellum* contains a cellulose-binding domain. *Biochemical Journal* **273**, 289–293.
- GARDINA, P., CONWAY, C., KOSSMAN, M. & MANSON, M. (1992). Aspartate and maltose-binding protein interact with adjacent sites in the *tar* chemotactic signal transducer of *Escherichia coli*. *Journal of Bacteriology* **172**, 1528–1536.
- GIORDA, R., OHMACHI, T., SHAW, D. R. & ENNIS, H. L. (1990). A shared internal threonine–glutamic acid–threonine–proline repeat defines a family of *Dictyostelium discoideum* spore germination proteins. *Biochemistry* **29**, 7264–7269.
- GRABNITZ, F., RÜCKNAGEL, K. P., SEIB, M. & STAUDENBAUER, W. L. (1989). Nucleotide sequence of the *Clostridium thermocellum* *bgIB* gene encoding thermostable β -glucosidase B: homology to fungal β -glucosidases. *Molecular and General Genetics* **217**, 70–76.
- GRÉPINET, O. & BÉGUIN, P. (1986). Sequence of the cellulase gene of *Clostridium thermocellum* coding for endoglucanase B. *Nucleic Acids Research* **14**, 1791–1799.
- GRÉPINET, O., CHEBROU, M.-C. & BÉGUIN, P. (1988). Nucleotide sequence and deletion analysis of the xylanase gene (*xynZ*) of *Clostridium thermocellum*. *Journal of Bacteriology* **179**, 4582–4588.
- HALL, J. & GILBERT, H. J. (1988). The nucleotide sequence of a carboxymethylcellulase from *Pseudomonas fluorescens* subsp. *cellulosa*. *Molecular and General Genetics* **213**, 112–117.
- HALL, J., HAZLEWOOD, G. P., BARKER, P. J. & GILBERT, H. J. (1988). Conserved reiterated domains in *Clostridium thermocellum* endoglucanases are not essential for catalytic activity. *Gene* **69**, 29–38.
- HAZLEWOOD, G. P., ROMANIEC, M. P. M., DAVIDSON, K., GRÉPINET, O., BÉGUIN, P., MILLET, J., RAYNAUD, O. & AUBERT, J.-P. (1988). A catalogue of *Clostridium thermocellum* endoglucanase, β -glucosidase and xylanase genes cloned in *Escherichia coli*. *FEMS Microbiology Letters* **51**, 231–236.
- HAZLEWOOD, G. P., DAVIDSON, K., CLARKE, J. H., DURRANT, A. J., HALL, J. & GILBERT, H. J. (1990a). Endoglucanase E produced at high level in *Escherichia coli* as a *lacZ'* fusion protein, is part of the *Clostridium thermocellum* cellulosome. *Enzyme and Microbial Technology* **12**, 656–662.
- HAZLEWOOD, G. P., DAVIDSON, K., LAURIE, J. I., ROMANIEC, M. P. M. & GILBERT, H. J. (1990b). Cloning and sequencing of the *celA* gene encoding endoglucanase A of *Butyrivibrio fibrisolvens* strain A46. *Journal of General Microbiology* **136**, 2089–2097.
- HENRISSAT, B., CLAEYSSENS, M., TOMME, P., LEMESLE, L. & MORNON, J.-P. (1989). Cellulase families revealed by hydrophobic cluster analysis. *Gene* **81**, 83–95.
- HUNKAPILLAR, M. W., HEWICK, R. E., DREYER, W. J. & HOOD, L. E. (1983). High sensitivity sequencing with gas-phase sequenator. *Methods in Enzymology* **91**, 399–413.

- Jauris, S., Rücknagel, K. P., Schwarz, W. H., Kratzsch, P., Bronnenmeier, K. & Staudenbauer, W. L. (1990). Sequence analysis of the *Clostridium stercorarium* celZ gene encoding a thermoactive cellulase (Avicelase I): identification of catalytic and cellulose-binding domains. *Molecular and General Genetics* **223**, 258–267.
- Joliff, G., Béguin, P. & Aubert, J.-P. (1986). Nucleotide sequence of the cellulase gene celD encoding endoglucanase D of *Clostridium thermocellum*. *Nucleic Acids Research* **14**, 8605–8613.
- Kohring, S., Wiegel, J. & Mayer, F. (1990). Subunit composition and glycosidic activities of the cellulase complex from *Clostridium thermocellum* JW20. *Applied and Environmental Microbiology* **56**, 3798–3804.
- Krikos, A., Mutoh, N., Boyd, A. & Simon, M. I. (1983). Sensory transducers of *E. coli* are composed of discrete structural and functional domains. *Cell* **33**, 615–622.
- Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature, London* **227**, 680–685.
- Lamed, R. & Bayer, E. A. (1988). The cellulosome of *Clostridium thermocellum*. *Advances in Applied Microbiology* **33**, 1–46.
- Mayer, F., Coughlan, M. P., Mori, Y. & Ljungdahl, L. G. (1987). Macromolecular organization of the cellulolytic enzyme complex of *Clostridium thermocellum* as revealed by electron microscopy. *Applied and Environmental Microbiology* **53**, 2785–2792.
- McLaughlin, J. R., Murray, C. L. & Rabinowitz, J. C. (1981). Unique features in the ribosome binding site sequence of the gram positive *Staphylococcus aureus* β -lactamase gene. *Journal of Biological Chemistry* **256**, 11283–11291.
- Meinke, A., Braun, C., Gilkes, N. R., Kilburn, D. G., Miller, R. C. & Warren, R. A. J. (1991). Unusual sequence organization in CenB, an inverting endoglucanase from *Cellulomonas fimi*. *Journal of Bacteriology* **173**, 308–314.
- Morag, E., Bayer, E. A. & Lamed, R. (1990). Relationship of cellulosomal and noncellulosomal xylanases of *Clostridium thermocellum* to cellulose-degrading enzymes. *Journal of Bacteriology* **172**, 6098–6105.
- Morag, E., Halevy, I., Bayer, E. A. & Lamed, R. (1991). Isolation and properties of a major cellobiohydrolase from the cellulosome of *Clostridium thermocellum*. *Journal of Bacteriology* **173**, 4155–4162.
- Morag, E., Bayer, E. A. & Lamed, R. (1992). Affinity digestion for the near-total recovery of purified cellulosome from *Clostridium thermocellum*. *Enzyme and Microbial Technology* **14**, 289–292.
- Murphy, G. & Kavanagh, T. (1988). Speeding-up the sequencing of double-stranded DNA. *Nucleic Acids Research* **16**, 5198.
- Navarro, A., Chebrou, M.-C., Béguin, P. & Aubert, J.-P. (1991). Nucleotide sequence of the cellulase gene celF of *Clostridium thermocellum*. *Research in Microbiology* **142**, 927–936.
- Norrander, J., Kemp, T. & Messing, J. (1983). Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. *Gene* **26**, 101–116.
- Poole, D. M., Morag, E., Lamed, R., Bayer, E. A., Hazlewood, G. P. & Gilbert, H. J. (1992). Identification of the cellulose binding domain of cellulosome subunit S1 from *Clostridium thermocellum* YS. *FEMS Microbiology Letters* (in the Press).
- Romaniec, M. P. M., Clarke, N. G. & Hazlewood, G. P. (1987). Molecular cloning of *Clostridium thermocellum* DNA and the expression of further novel endo- β -1,4-glucanase genes in *Escherichia coli*. *Journal of General Microbiology* **133**, 1297–1307.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463–5468.
- Schwarz, W. H., Schimming, S., Rücknagel, K. P., Burgesch-Waiger, S., Kreil, G. & Staudenbauer, W. L. (1988). Nucleotide sequence of the celC gene encoding endoglucanase C of *Clostridium thermocellum*. *Gene* **63**, 23–30.
- Sedmak, J. J. & Grossberg, S. E. (1977). A rapid, sensitive and versatile assay for protein using Coomassie Brilliant Blue G250. *Analytical Biochemistry* **69**, 544–552.
- Tokatlidis, K., Salamitou, S., Béguin, P., Dhurjati, P. & Aubert, J.-P. (1991). Interaction of the duplicated segment carried by *Clostridium thermocellum* cellulases with cellulosome components. *FEBS Letters* **291**, 185–188.
- Tomme, P., Chauvaux, S., Béguin, P., Millet, J., Aubert, J.-P. & Claeyssens, M. (1991). Identification of a histidyl residue in the active center of endoglucanase D from *Clostridium thermocellum*. *Journal of Biological Chemistry* **266**, 10313–10318.
- Towbin, H., Staehelin, T. & Gordon, J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 4350–4354.
- Tucker, M. L., Durbin, M. L., Clegg, M. T. & Lewis, L. N. (1987). Avocado cellulase: nucleotide sequence of a putative full-length cDNA clone and evidence for a small gene family. *Plant Molecular Biology* **9**, 197–203.
- Von Heijne, G. (1988). Transcending the impenetrable: how proteins come to terms with membranes. *Biochimica et Biophysica Acta* **947**, 307–333.
- Yagüe, E., Béguin, P. & Aubert, J.-P. (1990). Nucleotide sequence and deletion analysis of the cellulase-encoding gene celH of *Clostridium thermocellum*. *Gene* **89**, 61–67.