Appl Microbiol Biotechnol (1995) 43:291–296

ORIGINAL PAPER

V. S. J. Te'o · D. J. Saul · P. L. Bergquist

# *celA*, another gene coding for a multidomain cellulase from the extreme thermophile *Caldocellum saccharolyticum*

Received: 20 May 1994 / Received revision: 27 August 1994 / Accepted: 12 September 1994

**Abstract** *Caldocellum saccharolyticum* is an extremely thermophilic anaerobic bacterium capable of growth on cellulose and hemicellulose as sole carbon sources. Cellulase and hemicellulase genes have been found clustered together on its genome. The gene for one of the cellulases (*celA*) was isolated on a λ genomic library clone, sequenced and found to comprise a large open-reading frame of 5253 base pairs that could be translated into a peptide of 1751 amino acids. To date, it is the largest cellulase gene sequenced. The translated product is a multidomain structure composed of two catalytic domains and two cellulose-binding domains linked by proline-threonine-rich regions (PT linkers). The N-terminal domain of *celA* encodes for an endoglucanase activity on carboxymethylcellulose, consistent with its high homology to the sequences of several other endo-1,4-β-D-glucanases. The carboxylterminal domain shows sequence homology with a cellulase from *Clostridium thermocellum* (CelS), which is known to act synergistically with a second component to hydrolyze crystalline cellulose. In the absence of a *Caldocellum* homologue for this second protein, we can detect no activity from this domain.

## Introduction

The cellulase system comprises three general classes of enzymes: exoglucanases (β-1,4-D-glucan cellobiohydrolase), endoglucanases (β-1,4-D-glucan glucanhydrolase) and β-1,4-D-glucosidases. The first two enzymes depolymerize cellulose to cellobiose and cello-oligosaccharides, and β-glucosidase then hydrolyzes these sugars to form glucose (Aubert et al. 1988; Montencourt 1983). Most studies of cellulase production originally centred on the fungus *Trichoderma reesei* (Bhikhabhai et al. 1984; Coughlan 1990; Montencourt 1983). Several components of the cellulase system from *T. reesei* have been cloned (Knowles et al. 1987; Shoemaker et al. 1983) and the three-dimensional structure of cellobiohydrolase II determined (Rouvinen et al. 1990). Recently interest has shifted to cellulolytic bacteria. The thermophilic bacterium *Clostridium thermocellum* produces highly active and relatively thermostable enzymes and several groups have isolated cellulase genes from genomic libraries of this organism (reviewed by Béguin and Aubert 1994).

The bacterium used in this study, *Caldocellum saccharolyticum*, is an extremely thermophilic, obligate anaerobe that will grow and show cellulolytic activity at 80° C. This organism is a representative of a new genus of thermophilic bacteria, and recent sequence analysis of its SSU (16S) rRNA gene suggests that it is phylogenetically distinct from other known cellulolytic anaerobic bacteria (Rainey et al. 1993). We have described previously the isolation and expression of genes from this organism that encode cellulase and hemicellulase activities (Love et al. 1988; Lüthi et al. 1990; Saul et al.1990; Gibbs et al. 1992). One gene, *celB*, codes for an enzyme with two catalytic domains: the first has both exo-β-1,4-glucanase and endo-β-1, 4-xylanase activities, and the second has endo-β-1, 4-glucanase activity. The two domains are clearly delineated and are separated by a third domain lacking known enzymatic activity. By inference from the work of others, this domain is thought to be responsible for binding the enzyme to cellulose (Saul et al. 1989, 1990; Gibbs et al. 1992). The three domains are separated by PT linkers, regions of proline-threonine repeats which act as flexible hinges and are a common feature of multidomain cellulases. Another gene from *C. saccharolyticum*, *man*A, also encodes an enzyme with two catalytic domains. In this case, one domain is responsible for mannan degradation and the other shows endo-β-1,4-glucanase and endo-β-1,4-xylanase activities. This enzyme differs from CelB in that the two catalytic domains are separated by

V. S. J. Te'o · D. J. Saul · P. L. Bergquist (✉)
Centre for Gene Technology, Molecular Genetics and Microbiology, School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand,
Fax: 64-9-3737 414

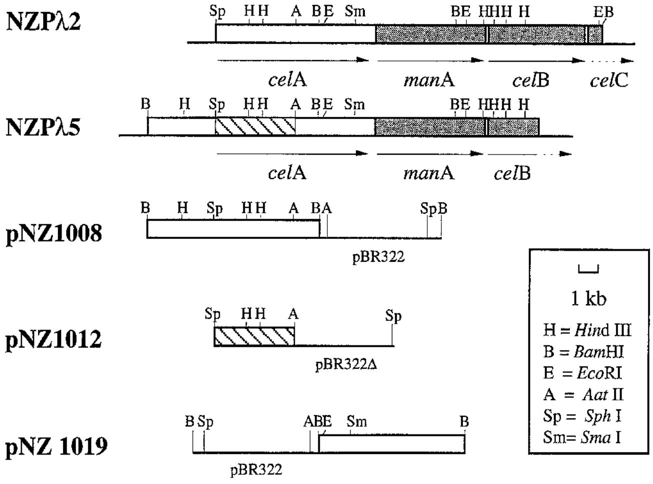

**Fig. 1** Restriction enzyme maps of the λ and plasmid recombinants used to isolate and sequence the *celA* gene. *Shaded areas* on the λ recombinants indicate the positions of previously published cellulase genes (Gibbs et al. 1992; Saul et al. 1989). *Cross-hatched shading* indicates the fragment to which carboxymethylcellulare activity was initially mapped

*two* binding domains, which are almost identical to the central domain of CelB (Gibbs et al. 1992). In this communication we describe the characterization and features of the sequence of a third multidomain enzyme from *C. saccharolyticum.*

## Materials and methods

### Bacterial Strains

*Escherichia coli* strain JM101 (F′[*traD*36 *pro*A, B$^+$, *lac*I$^q$, *lacZΔ*-MI5]*Δ*(*lac-pro*A,B), *thi*, supE44) (Yanisch-Perron et al. 1985) was used for all M13 recombinants. C600 (*thr*–1, *leu*B6, *lacY*1, *supE44, thi-1,hsd*R) (Appleyard 1954) was used for all other plasmid constructions.

### Manipulation of DNA

DNA preparation, manipulation and digestion with restriction enzymes were performed according to Sambrook et al. (1989). Preparation of a randomly sheared library of DNA fragments from plasmids pNZ1008 and pNZ1019 (see Fig. 1) was carried out according to Bankier et al. (1987) using a M13 mp10 vector cut with *Sma*I and dephosphorylated. *E. coli* strain JM101 was used for transformation. Polymerase chain reactions (PCR) were performed in a Perkin Cetus DNA thermal cycler. DNA sequencing was carried out on an Applied Biosystems 373A DNA sequencer utilizing either dye-labelled primers or dye-labelled dideoxy-DNA chain terminators and *Taq* DNA polymerase. Sequences were analyzed on a Silicon graphics 4D/30 using the GCG software package (Devereux et al. 1984).

### Enzyme activity assays

Activities on carboxymethylcellulose (CMC), xylan and mannan substrates were detected by the method of Teather and Wood (1982). Activity on *p*–nitrophenyl and *o*–nitrophenyl substrates was tested as follows: an overnight culture was diluted 1:10 in 25 ml fresh L broth with the appropriate antibiotic selection and grown at 28–32°C to an $A_{600}$ of 1.0. After induction at 42°C for 4 h, 1 ml cells was centrifuged and the cell pellet was resuspended in 100 µl buffer (0.1 M NaCl, 5% Triton X-100, 1 mM EDTA, 10 mM TRIS HCl, pH 8.0). A 1-µl aliquot of toluene was added and the tubes were placed at 37°C with lids open for 3 min. About 10 µl substrate at 20 mg/ml in dimethylformamide was added to the mixtures and the tubes were incubated at 70°C for 30 min. A yellow colour indicated activity.

## Results

### Sequence analysis of *celA*

Prior to this work, a CMCase-positive clone, NZPλ5, was isolated from a λ library of *C. saccharolyticum* genomic DNA (Saul et al.1990). Recombinant plasmids were constructed by ligating portions of the NZPλ5 insert into the *E.coli* vector pBR322. Activity on CMC was localised to a 2,8 × 10$^3$–base (2.8-kb) *Sph*I-*Aat*II fragment and the plasmid containing this fragment was designated pNZ1012 (Fig. 1). The putative gene encoding the activity was named *celA*. An M13 library of overlapping DNA fragments was generated from this fragment and the DNA sequence determined. On examination of the sequence, it became apparent that the open-reading frame of the *celA* gene continued beyond the *Aat*II restriction site on NZPλ5 (Fig. 1). The DNA close to this site codes for a PT linker, which indicated that *celA* was a multi-domain protein like its neighbours *manA* and *celB* (Gibbs et al. 1992) and that the CMCase activity shown by *E. coli* recombinants containing the 2.8-kb fragment was due to the expression of a single catalytic domain of a larger protein. Further M13 recombinants were generated from plasmids pNZ1008, pNZ1010 and the NZPλ2 derivative, pNZ1019, and the DNA sequence was determined as far as *manA* (Lüthi et al. 1991). The bulk of the sequence was obtained from random-shear and pseudorandom libraries and was completed by utilizing restriction fragments and primers designed to cover single-stranded gaps and butt joins. The sequence is available from GenBank under the accession number L32742.

The sequence contains a single open-reading frame of 5255 base pairs, which terminates close to the start of *manA*. Lüthi et al. (1991) noted the presence of an open-reading frame (ORF) upstream of *manA* but it was not then apparent that this was the 3′ end of the *celA* gene. The DNA sequence and the deduced peptide sequence are shown in Fig. 2. Translated, the *celA* gene is capable of producing a protein of 1751 amino acids (CelA), which would have a calculated molecular mass of 194550 Da.

**Fig. 2** Structure of the *celA* gene and deduced peptide sequence. ▶ Pro-Thr (PT) linkers are highlighted by *boxes*. A putative leader sequence is *underlined* with possible cleavage sites indicated by *shaded boxes*. The GenBank accession number for *celA* is L32742

```
                                                                                                       ┌─► celA
1     CCACGTATTTTGGGTATATAATTGAAATCGAAAATTATAATAAACTGAATGAGGGGGTTAGAGATTTTATGCAGCGTTACAGAAGAATTATTGCAATGGTTGTAACCTTTTTATTTATTT
                                                                                                      M  V  V  T  F  L  F  I  L
121   TAGGAGTGGTATATGGGGTTAAACCATGGCAGGAGGCTAGGGCTGGTTCGTTTAACTATGGGGAAGCATTACAAAAAGCTATCATGTTTTACGAATTTCAAATGTCTGGTAAACTTCCGA
         G  V  V  Y  G  V  K  P  W  Q  E  A  R  A  G  S  F  N  Y  G  E  A  L  Q  K  A  I  M  F  Y  E  F  Q  M  S  G  K  L  P  N
241   ATTGGGTACGCAACAACTGGCGTGGTGATTCAGCATTAAAGGATGGTCAAGACAATGGGCTTGATTTGACAGGTGGATGGTTCGATGCAGGTGATCATGTCAAATTTAACCTTCCAATGT
         W  V  R  N  N  W  R  G  D  S  A  L  K  D  G  Q  D  N  G  L  D  L  T  G  G  W  F  D  A  G  D  H  V  K  F  N  L  P  M  S
361   CGTATACTGGCACAATGTTGTCATGGGCAGCATATGAGTACAAAGATGCGTTTGTTAAGAGCGGCCAATTGGAACATATCTTAAACCAAATCGAGTGGGTTAATGACTATTTTGTAAAAT
         Y  T  G  T  M  L  S  W  A  A  Y  E  Y  K  D  A  F  V  K  S  G  Q  L  E  H  I  L  N  Q  I  E  W  V  N  D  Y  F  V  K  C
481   GTCATCCAAGCAAATATGTATACTATTACCAGGTTGGAGATGGCGGCAAGGACCATGCATGGTGGGGTCCTGCTGAGGTTATGCAAATGGAGAGACCTTCATTTAAGGTCACCCAGAGCA
         H  P  S  K  Y  V  Y  Y  Y  Q  V  G  D  G  G  K  D  H  A  W  W  G  P  A  E  V  M  Q  M  E  R  P  S  F  K  V  T  Q  S  S
601   GTCCAGGTTCTGCGGTAGTAGCAGAGACGGCAGCTTCCTTAGCAGCAGCTTCTATTGTTTTGAAAGACAGAAATCCCACTAAAGCAGCAACATATCTGCAACATGCAAAAGATTTATATG
         P  G  S  A  V  V  A  E  T  A  A  S  L  A  A  A  S  I  V  L  K  D  R  N  P  T  K  A  A  T  Y  L  Q  H  A  K  D  L  Y  E
721   AGTTTGCAGAAGTGACAAAAAGCGATTCTGGTTACACGGCAGCAAATGGATATTATAATTCATGGAGCGGTTTCTATGATGAGCTTTCTTGGGCAGCAGTTTGGTTGTATTTGGCAACAA
         F  A  E  V  T  K  S  D  S  G  Y  T  A  A  N  G  Y  Y  N  S  W  S  G  F  Y  D  E  L  S  W  A  A  V  W  L  Y  L  A  T  N
841   ATGATTCAACATATCTAACAAAAGCTGAGTCATATGTCCAAAATTGGCCAAAAATTTCAGGTAGTAACATAATTGACTACAAATGGGCTCATTGCTGGGATGATGTTCACAATGGAGCGG
         D  S  T  Y  L  T  K  A  E  S  Y  V  Q  N  W  P  K  I  S  G  S  N  I  I  D  Y  K  W  A  H  C  W  D  D  V  H  N  G  A  A
961   CATTATTGTTAGCAAAAATTACCGACAAGGATACTTATAAGCAAATTATTGAGAGTCACTTAGATTACTGGACTACAGGATACAACGGCGAAAGGATAAAGTATACTCCGAAAGGATTAG      CD1
         L  L  L  A  K  I  T  D  K  D  T  Y  K  Q  I  I  E  S  H  L  D  Y  W  T  T  G  Y  N  G  E  R  I  K  Y  T  P  K  G  L  A
1081  CATGGCTTGATCAATGGGGTTCTTTAAGATATGCGACAACTACAGCGTTTTTGGCATTTGTTTATAGCGATTGGTCTGGTTGCCCAACTGGTAAAAAAGAAACATATAGAAAATTTGGAG
         W  L  D  Q  W  G  S  L  R  Y  A  T  T  T  A  F  L  A  F  V  Y  S  D  W  S  G  C  P  T  G  K  K  E  T  Y  R  K  F  G  E
1201  AAAGCCAAATTGATTATGCATTAGGCTCAACTGGAAGAAGCTTTGTTGTTGGATTTGGCACAAATCCACCAAAGAGACCTCATCACAGAACTGCTCATAGCTCATGGGCAGACAGTCAGA
         S  Q  I  D  Y  A  L  G  S  T  G  R  S  F  V  V  G  F  G  T  N  P  P  K  R  P  H  H  R  T  A  H  S  S  W  A  D  S  Q  S
1321  GTATACCTTCATATCATAGACATACATTATATGGAGCGCTTGTTGGTGGTCCAGGCTCTGATGATAGCTATACAGATGACATTAGCAACTATGTGAATAATGAGGTAGCATGTGACTACA
         I  P  S  Y  H  R  H  T  L  Y  G  A  L  V  G  G  P  G  S  D  D  S  Y  T  D  D  I  S  N  Y  V  N  N  E  V  A  C  D  Y  N
1441  ACGCAGGATTTGTTGGTGCATTGGCAAAGATGTACTTATTGTATGGTGGAAATCCAATACCTGACTTCAAAGCCATTGAAACACCAACAAATGACGAGTTCTTTGTTGAAGCTGGTATAA
         A  G  F  V  G  A  L  A  K  M  Y  L  L  Y  G  G  N  P  I  P  D  F  K  A  I  E  T  P  T  N  D  E  F  F  V  E  A  G  I  N
1561  ATGCTTCTGGAACAAACTTTATTGAAATTAAGGCGATTGTTAATAATCAGAGTGGTTGGCCCGCAAGAGCAACAAATAAGCTTAAATTTAGATATTTTGTTGATCTGAGCGAATTAATTA
         A  S  G  T  N  F  I  E  I  K  A  I  V  N  N  Q  S  G  W  P  A  R  A  T  N  K  L  K  F  R  Y  F  V  D  L  S  E  L  I  K
1681  AAGCAGGATATTCACCAAATCAATTAACCTTAAGTACCAATTATAATCAAGGTGCAAAAGTAAGTGGACCTTATGTATGGGATTCAAGCAGGAATATATACTACATTTTAGTAGACTTTA
         A  G  Y  S  P  N  Q  L  T  L  S  T  N  Y  N  Q  G  A  K  V  S  G  P  Y  V  W  D  S  S  R  N  I  Y  Y  I  L  V  D  F  T
1801  CTGGCACATTGATTTATCCAGGTGGCCAAGACAAATATAAGAAAGAAGTTCAATTCAGAATTGCAGCGCCACAGAATGTACAGTGGGATAATTCCAACGACTATTCATTCCAAGATATAA
         G  T  L  I  Y  P  G  G  Q  D  K  Y  K  K  E  V  Q  F  R  I  A  A  P  Q  N  V  Q  W  D  N  S  N  D  Y  S  F  Q  D  I  K
1921  AGGGAGTTTCAAGTGGTTCAGTTGTTAAAACAAAATATATACCATTATATGATGAAGATATCAAGGTATGGGGTGAAGAGCCAGGTACATCTGGAGTATCACCAACACCAACTGCAAGTG
         G  V  S  S  G  S  V  V  K  T  K  Y  I  P  L  Y  D  E  D  I  K  V  W  G  E  E  P  G  T  S  G  V  S  P  T  P  T  A  S  V
2041  TAACCCCAACCCCAACACCTACACCGACTGCAACACCAACCCCGACACCTACACCAACAGTAACACCGACACCGACAGTAACAGCAACCCCGACACCAACACCAACACCGACCTCAACAC
         T  P  T  P  T  P  T  P  T  A  T  P  T  P  T  P  T  P  T  V  T  P  T  P  T  V  T  A  T  P  T  P  T  P  T  P  T  S  T  P
2161  CGACGGTAACACCTACACCTACACCTGTTAGCACACCTGCGACAAGTGGGCAGATAAAGGTACTGTATGCTAACAAGGAGACAAACAGCACGACAAACACGATAAGGCCATGGTTGAAGG
         T  V  T  P  T  P  T  P  V  S  T  P  A  T  S  G  Q  I  K  V  L  Y  A  N  K  E  T  N  S  T  T  N  T  I  R  P  W  L  K  V
2281  TAGTGAACAGTGGTAGCAGTAGCATAGATTTGAGCAGGGTAACGATAAGGTACTGGTACACGGTAGATGGTGAGAGGGCACAGAGTGCGATATCAGACTGGGCACAGATAGGAGCAAGCA
         V  N  S  G  S  S  S  I  D  L  S  R  V  T  I  R  Y  W  Y  T  V  D  G  E  R  A  Q  S  A  I  S  D  W  A  Q  I  G  A  S  N
2401  ATGTAACATTCAAGTTTGTGAAGCTGAGCAGTAGTGTGAGTGGAGCGGACTATTACTTGGAGATAGGATTTAAGAGTGGAGCAGGGCAGTTGCAGCCTGGGAAGGACACAGGAGAGATAC      BD1
         V  T  F  K  F  V  K  L  S  S  S  V  S  G  A  D  Y  Y  L  E  I  G  F  K  S  G  A  G  Q  L  Q  P  G  K  D  T  G  E  I  Q
2521  AGATAAGGTTTAACAAGGATGACTGGAGCAATTACAATCAGGGGAATGACTGGTCATGGATACAGAGCATGACGAGTTATGGAGAGAATGAGAAGGTAACGGCGTATATAGATGGTGTGC
         I  R  F  N  K  D  D  W  S  N  Y  N  Q  G  N  D  W  S  W  I  Q  S  M  T  S  Y  G  E  N  E  K  V  T  A  Y  I  D  G  V  L
2641  TGGTATGGGGACAGGAGCCGAGTGGAACAACACCTGCGCCGACGTCAACACCGACAGTAACGGTAACACCAACACCAACACCTACACCAACAGTAACACCGACACCGACAGTGACAGCAA
         V  W  G  Q  E  P  S  G  T  T  P  A  P  T  S  T  P  T  V  T  V  T  P  T  P  T  P  T  P  T  V  T  P  T  P  T  V  T  A  T
2761  CCCCGACTCCAACACCGACCCCGACGTCTACACCTGTTAGCACACCTGCGACAGGTGGGCAGATAAAGGTACTGTATGCTAACAAGGAGACAAACAGCACGACAAACACGATAAGGCCAT
         P  T  P  T  P  T  P  T  S  T  P  V  S  T  P  A  T  G  G  Q  I  K  V  L  Y  A  N  K  E  T  N  S  T  T  N  T  I  R  P  W
2881  GGTTGAAGGTAGTGAACAGTGGTAGCAGTAGCATAGATTTGAGCAGGGTAACGATAAGGTACTGGTACACGGTGGATGGTGAGAGGGCACAGAGTGCGATATCAGACTGGGCACAGATAG
         L  K  V  V  N  S  G  S  S  S  I  D  L  S  R  V  T  I  R  Y  W  Y  T  V  D  G  E  R  A  Q  S  A  I  S  D  W  A  Q  I  G
3001  GAGCAAGCAATGTAACATTCAAGTTTGTGAAGCTGAGCAGTAGCGTAAGTGGAGCGGATTATTACTTGGAGATAGGATTTAAGAGTGGAGCAGGGCAGTTGCAGCCTGGGAAGGACACAG      BD2
         A  S  N  V  T  F  K  F  V  K  L  S  S  S  V  S  G  A  D  Y  Y  L  E  I  G  F  K  S  G  A  G  Q  L  Q  P  G  K  D  T  G
3121  GAGAGATACAGATAAGATTTAACAAGGATGACTGGAGCAATTACAATCAGGGGAATGACTGGTCATGGATACAGAGCATGACGAGTTATGGAGAGAATGAGAAGGTAACGGCGTATATAG
         E  I  Q  I  R  F  N  K  D  D  W  S  N  Y  N  Q  G  N  D  W  S  W  I  Q  S  M  T  S  Y  G  E  N  E  K  V  T  A  Y  I  D
3241  ATGGTGTGCTGGTATGGGGACAGGAGCCGAGCGGAGCGACACCTGCGCCGACAGTGACGCCAACCCCGACAGTGACACCAACTCCGACCCCAGCACCGACCCCGACAGCGACTCCGACAC
         G  V  L  V  W  G  Q  E  P  S  G  A  T  P  A  P  T  V  T  P  T  P  T  V  T  P  T  P  T  P  A  P  T  P  T  A  T  P  T  P
3361  CAACTCCAACACCGACGGTGACACCGACGCCGACAGTGGCTCCGACACCTACACCGAGCAGCACACCGAGCGGTCTTGGAAAATATGGGCAGAGGTTTATGTGGTTGTGGAACAAGATAC
         T  P  T  P  T  V  T  P  T  P  T  V  A  P  T  P  T  P  S  S  T  P  S  G  L  G  K  Y  G  Q  R  F  M  W  L  W  N  K  I  H
3481  ATGATCCTGCGAGCGGGTATTTTAACCAGGATGGGATACCATATCATTCGGTAGAGACATTGATATGCGAAGCACCTGATTATGGTCATTTGACCACGAGTGAAGCATTTTCGTATTATG
         D  P  A  S  G  Y  F  N  Q  D  G  I  P  Y  H  S  V  E  T  L  I  C  E  A  P  D  Y  G  H  L  T  T  S  E  A  F  S  Y  Y  V
3601  TATGGCTTGAGGCAGTGTATGGGAAGTTGACGGGTGATTGGAGCAAGTTTAAGACGGCATGGGATACATTAGAGAAGTATATGATACCGTCGGCTGAAGATCAGCCGATGAGGTCATATG
         W  L  E  A  V  Y  G  K  L  T  G  D  W  S  K  F  K  T  A  W  D  T  L  E  K  Y  M  I  P  S  A  E  D  Q  P  M  R  S  Y  D
3721  ATCCTAACAAGCCAGCGACGTATGCAGGTGAGTGGGAGACACCGGACAAGTATCCATCGCCGCTTGAGTTTAATGTACCTGTTGGGAAGGATCCGTTGCATAATGAGCTTGTGAGCACAT
         P  N  K  P  A  T  Y  A  G  E  W  E  T  P  D  K  Y  P  S  P  L  E  F  N  V  P  V  G  K  D  P  L  H  N  E  L  V  S  T  Y
3841  ATGGTAGCACATTGATGTATGGTATGCACTGGTTGATGGACGTAGACAACTGGTATGGATATGGCAAGAGAGGGGACGGAGTGAGCCGGGCATCATTTATCAACACGTTCCAGAGAGGGC
         G  S  T  L  M  Y  G  M  H  W  L  M  D  V  D  N  W  Y  G  Y  G  K  R  G  D  G  V  S  R  A  S  F  I  N  T  F  Q  R  G  P
3961  CTGAGGAGTCTGTATGGGAGACTGTACCGCATCCGAGCTGGGAGGAATTCAAATGGGGCGGACCGAATGGATTTTTAGATCTGTTTATTAAGGATCAGAACTATTCGAAGCAGTGGAGGT
         E  E  S  V  W  E  T  V  P  H  P  S  W  E  E  F  K  W  G  G  P  N  G  F  L  D  L  F  I  K  D  Q  N  Y  S  K  Q  W  R  Y
4081  ATACGAACGCACCGGATGCGGATGCGAGAGCTATTCAGGCAACTTACTGGGCGAAGGTATGGGCGAAGGAGCAAGGTAAGTTTAATGAGATAAGCAGCTATGTAGGGAAGGCAGCGAAGA
         T  N  A  P  D  A  D  A  R  A  I  Q  A  T  Y  W  A  K  V  W  A  K  E  Q  G  K  F  N  E  I  S  S  Y  V  G  K  A  A  K  M
4201  TGGGAGACTATTTGAGGTATGCGATGTTTGACAAGTATTTCAAGCCATTAGGATGTCAGGACAAGAATGCAGCTGGAGGAACGGGATATGACAGTGCACATTATCTGTTATCATGGTATT
         G  D  Y  L  R  Y  A  M  F  D  K  Y  F  K  P  L  G  C  Q  D  K  N  A  A  G  G  T  G  Y  D  S  A  H  Y  L  L  S  W  Y  Y
4321  ATGCATGGGGCGGGGCATTGGATGGAGCATGGTCATGGAAGATAGGTTGCAGTCATGCGCACTTTGGATATCAGAATCCGATGGCTGCATGGGCATTAGCAAATGATAGTGATATGAAGC
         A  W  G  G  A  L  D  G  A  W  S  W  K  I  G  C  S  H  A  H  F  G  Y  Q  N  P  M  A  A  W  A  L  A  N  D  S  D  M  K  P
4441  CGAAGTCACCGAACGGAGCGAGTGACTGGGCGAAGAGTTTGAAGAGGCAGATAGAATTTTACAGGTGGTTGCAGTCAGCGGAGGGAGCGATAGCAGGAGGCGCGACAAATTCGTGGAATG      CD2
         K  S  P  N  G  A  S  D  W  A  K  S  L  K  R  Q  I  E  F  Y  R  W  L  Q  S  A  E  G  A  I  A  G  G  A  T  N  S  W  N  G
4561  GCAGGTACGAGAAGTATCCGGCAGGTACAGCGACATTTTATGGGATGGCATATGAACCGAACCCTGTGTATAGGGACCCGGGTAGCAATACATGGTTTGGATTCCAGGCATGGTCGATGC
         R  Y  E  K  Y  P  A  G  T  A  T  F  Y  G  M  A  Y  E  P  N  P  V  Y  R  D  P  G  S  N  T  W  F  G  F  Q  A  W  S  M  Q
4681  AGAGGGTAGCGGAGTATTACTATGTGACAGGAGATAAAGATGCAGGGACACTGCTTGAGAAGTGGGTAAGCTGGATAAAGAGTGTAGTGAAGTTGAATAGTGATGGTACATTTGCGATAC
         R  V  A  E  Y  Y  Y  V  T  G  D  K  D  A  G  T  L  L  E  K  W  V  S  W  I  K  S  V  V  K  L  N  S  D  G  T  F  A  I  P
4801  CATCGACGCTTGATTGGAGTGGGCAGCCAGACACATGGAATGGGACATATACAGGTAATCCGAACTTGCATGTGAAGGTAGTAGATTATGGGACGGATTTAGGAATAACGGCATCACTTG
         S  T  L  D  W  S  G  Q  P  D  T  W  N  G  T  Y  T  G  N  P  N  L  H  V  K  V  V  D  Y  G  T  D  L  G  I  T  A  S  L  A
4921  CGAATGCACTACTTTATTACAGTGCAGGGACGAAGAAGTATGGGGTATTTGATGAGGAAGCGAAGAATTTAGCGAAGGAATTGCTGGACAGGATGTGGAAGTTATACAGGGATGAGAAAG
         N  A  L  L  Y  Y  S  A  G  T  K  K  Y  G  V  F  D  E  E  A  K  N  L  A  K  E  L  L  D  R  M  W  K  L  Y  R  D  E  K  G
5041  GTTTATCGGCGCCAGAGAAGAGAGCGGACTACAAGAGGTTCTTTGAGCAAGAGGTATACATTCCGGCAGGATGGACAGGGAAGATGCCGAATGGAGATGTAATAAAGAGCGGAGTTAAGT
         L  S  A  P  E  K  R  A  D  Y  K  R  F  F  E  Q  E  V  Y  I  P  A  G  W  T  G  K  M  P  N  G  D  V  I  K  S  G  V  K  F
5161  TTATAGACATAAGGAGCAAGTACAAACAAGATCCTGATTGGCCGAAGTTAGAGGCGGCATACAAGTCAGGGCAGGTACCGGAGTTCAGATATCACAGGTTCTGGGCACAGTGTGACATAG
         I  D  I  R  S  K  Y  K  Q  D  P  D  W  P  K  L  E  A  A  Y  K  S  G  Q  V  P  E  F  R  Y  H  R  F  W  A  Q  C  D  I  A
5281  CAATTGTTAATGCAACATATGAAATTCTGTTCGGTAATCAATAATGAGTAGGTAAATGGAAATTTAGCGGGGTGGCACATCTATAAGTTTGGTGTGCTGCCTCGCTAAAATCCTGTATGG
         I  V  N  A  T  Y  E  I  L  F  G  N  Q  *  *
5401  AAGTGTTCGAAAAATAGTACAAAAAAATGGCGAGGTAAA  5439
```

CelA is clearly a multidomain cellulase consisting of two catalytic domains (CD1 and CD2) and two putative binding domains. These are connected by PT linkers similar to those seen in other *C. saccharolyticum* cellulases. This is the same structure as the *manA* gene product (Gibbs et al. 1992). The coding sequences from the two binding domains of CelA are highly homologous with each other and with those of ManA and CelB.

The deduced CelA peptide sequence was compared against the sequence databases. CD1 showed high homology to an avocado cellulase from *Persea americana* (Tucker et al. 1987), a bean abscission cellulase from *Phaseolus vulgaris* (Tucker and Milligan 1991), CenB from *Cellulomonas fimi*(Meinke et al. 1991), E4 cellulase from *Thermomonospora fusca* (Lao et al. 1991), CelF from *C. thermocellum* (Navarro et al. 1991), and 270–6 cellulase from *Dictyostelium discoideum* (Giorda et al. 1990). CD2 showed strong homology to translations of ORF1 in the cellulase gene cluster of *Clostridium cellulolyticum* (Bagnara-Tardif et al. 1992), ORF1 upstream of *celB* in *Clostridium josui* (Fujino et al. 1993) and a short open-reading frame downstream of the gene encoding binding protein A in *Clostridium cellulovorans* (Shoseyov et al. 1992). More informative is the homology (61% identity) with Ss (CelS) of *C. thermocellum,* which is known to degrade crystalline cellulose when associated with Sl, a second component of the *C. thermocellum* cellulosome (Wang et al. 1993). It is worth noting that the homology with *C. cellulovorans* ORF1 and *C. josui* ORF1 terminates immediately upstream of the re-iterated stretch of amino acids present in these peptides. Repeats of this kind are found in cellulases that associate to form a cellulosome and their absence in CelA is in line with other cellulases from *C. saccharolyticum* and the lack of a detectable cellulosome in this organism.

The presumed signal peptide of CelA is typical of those found in gram-positive bacteria (Fig. 2). Three basic amino acids (arginines) form a positively charged region, which is followed by a run of hydrophobic residues. Both Ala-21 and Ala-23 coincide with the rules summarized by Mackay et al. (1986) for possible signal peptidase cleavage sites.

## Expression studies with *celA*

Primers were designed and used to amplify the entire gene and individual catalytic domains (Fig. 3). A start codon has been included in primer P2 and a stop codon in P4. *Sph*I and *Sal*I sites have also been included in the primers to facilitate simple directional insertions into the expression vector pJLA602 (Schauder et al. 1987). The recombinant bacteriophage NZP2 was used as a PCR target for amplification of the CelA gene.

*celA* expression recombinants generated by PCR were tested on a variety of substrates to characterize the enzyme activities of individual domains and the entire CelA protein. Of the eight substrates, the full-length *celA* and CD1 were found to have activity on CMC, lichenan and konjac glucomannan (data not shown). All of the substrates testing positive contained $\beta$-1,4-linked glucose residues (lichenan consists of alternate $\beta$-1, $\beta$-1,3 and $\beta$-1,4 glucan residues and konjac glucomannan has a 1.8:1 ratio of mannose to glucose). The results are indicative of CD1 having endoglucanase activity. The full-length *celA* and CD1 showed no activity on laminarin, xylan and 4-methylumbelliferyl (meUmb) $\beta$-D-cellobioside. CD2 showed no detectable activity on any of these substrates, nor on any of a wide selection of *o*-nitrophenyl or *p*-nitrophenyl substrates. We were concerned that the failure to detect activity with CD2 on any substrates may be due to this domain misfolding in our truncated construction. As an alternative approach, an in-frame deletion of CD1 was made from the full-length construction but this too failed to show activity on any substrate.

**P1:** AGATTGCATGCAGCGTTACAGAAGAAT
**P2:** ACCGAGCATGCTTGGAAAATATGGGCAGAGG
**P3:** TAAATGTCGACTTACCTACTCATTATTGATT
**P4:** TCCAGGTCGACCTCACTCTTCACCCCATACCTTG

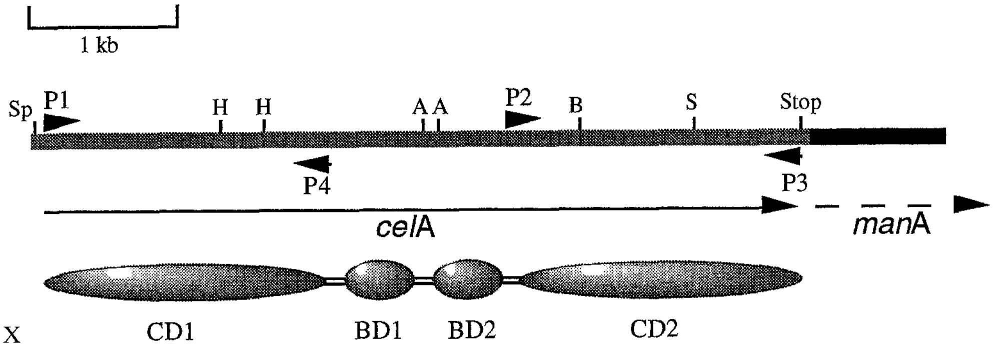

**Fig. 3** Schematic representation of *celA* gene and primers used in the polymerase chain reaction isolation of fragments used in the expression studies. *Underlined* on the primers are the introduced restriction sites

## Discussion

Like other cellulases, CelA is composed of catalytic domains, putative binding domains and linker regions (PT linkers). The linker regions are typically short sequences rich in proline or hydroxyamino acids, or both, joining discrete catalytic and cellulose-binding domains. These sequences vary considerably in length and in their proline and hydroxyamino acid contents. Many cellulases require binding domains to bind to the cellulosic substrates, but the mechanism and significance of this interaction are unclear. Catalytic domains and cellulose-binding domains can often retain their functions even when separated by proteolysis (Gilkes et al. 1988). *N*–Bromosuccinimide-inactivated cellobiohydrolase I of *Aspergillus ficum* still binds to cellulose, indicating that this enzyme also comprises discrete catalytic and binding domains (Hayashida et al. 1988). Din et al. (1991) have shown that the isolated cellulose-binding domain of endoglucanase A (CenA) from the bacterium *C. fimi* disrupts the structure of cellulose fibres and releases small particles, but has no detectable hydrolytic activity. In contrast, the isolated catalytic domain of this enzyme does not disrupt the fibril structure but 'polishes' the surface of the fibre concomitant with the release of reducing sugars.

Native cellulose is associated with other polysaccharides and, as a result, is a complex substrate for enzyme degradation. As a consequence, cellulolytic microorganisms produce several classes of enzymes. Thus, cooperation between the various cellulolytic enzymes in hydrolysing cellulose appears to be favoured. Several investigators have provided evidence for the synergistic effects of cellulolytic enzymes involved in the degradation of 'cellulosics'. Synergism between all components of the cellulase complex would undoubtedly increase the efficiency of cellulose hydrolysis, particularly with the more crystalline forms of cellulose. Evolution may have selected for efficiency of degradation, and therefore multigene cellulase families expressing 'complementary' activities have predominated, rather than perhaps less efficient, single enzymes with wider substrate specificities.

Henrissat et al. (1989) have grouped cellulases and xylanases into families on the basis of hydrophobic cluster analysis. It has been proposed that these enzymes have evolved by domain shuffling, with subsequent modification of the domains. Such a proposal is based on the observation that the catalytic domains from different cellulase and xylanase families are associated with the same type of cellulose-binding domains. *C. thermocellum* CelS and CD2 of *C. saccharolyticum* CelA do not conform to any of the glucanase families. Wang et al. (1993) have suggested that *C. thermocellum* CelS may be a member of a new family.

Synergism may apply for *C. saccharolyticum* cellulases. To date, a $\beta$-glucosidase, two MeUmb-$\beta$-*D*–cellobiosidases (and thus potential cellobiohydrolases), CelA, which has CMCase activity, and ManA, which has CMCase, mannanase and xylanase activities, have all been characterized from *C. saccharolyticum.* From these results, *C. saccharolyticum* cellulases appear to represent a family of multifunctional enzymes. Perhaps *C. saccharolyticum* has evolved cellulases with multifunctional enzyme activities that are positioned in a single protein with optimal three-dimensional configurations for synergistic hydrolysis because this organism does not have its enzymes organised into a cellulosome.

**Acknowledgements** This work was supported in part by grants from the University of Auckland Research Committee and from Pacific Enzymes Ltd.

## References

Appleyard RK (1954) Segregation of new lysogenic types during growth of a doubly lysogenic strain derived from *Escherichi coli* K12. Genetics 39:440–452

Aubert JP, Bèguin PJ Millet J (ed) (1988) Biochemistry and genetics of cellulose degradation. FEMS symposium no. 43. Academic Press, New York

Bagnara-Tardif C, Gaudin C, Belaich A, Hoest P, Citard T, Belaich J-P (1992) Sequence of a gene cluster encoding cellulases from *Clostridium cellulolyticum.* Gene 119:17–28

Bankier AT, Weston KM, Barrell BG (1987) Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. Methods Enzymol 155:51–93

Bèguin P, Aubert JP (1994) The biological degradation of cellulose. FEMS Microbiol Rev 13:25–58

Bhikhabhai R, Johansson G, Petterson LG (1984) Isolation of cellulolytic enzymes from *Trichoderma reesei* QM9414. J Appl Biochem 6:336–345

Coughlan MP (1990) Cellulose degradation by fungi. In: Fogarty WM, Kelly CT, (eds) *Microbial Enzymes and biotechnology*, 2nd edn. Elsevier Applied Science, London, pp 1–36

Devereux J, Haeberli P, Smithies DO (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12:387–395

Din N, Gilkes NR, Tekant B, Miller Jr. RC, Warren RAJ, Kilburn DG (1991) Non-hydrolytic disruption of cellulose fibres by the binding domain of a bacterial cellulase. Biotechnology 9:1096–1099

Fujino T, Karita S, Ohmiya K (1993) Nucleotide sequence of the *celB* gene encoding endo-1,4,-$\beta$-glucanase-2, ORF1 and ORF2 forming a putative cellulase gene cluster of *Clostridium josui.* J Ferment Bioeng 76:243–250

Gibbs MD, Saul DJ, Lüthi E, Bergquist PL (1992) The $\beta$-mannanase from *"Caldocellum saccharolyticum"* is part of a multidomain enzyme. Appl Environ Microbiol 58:3864–3867

Gilkes NR, Warren RAJ, Miller RC, Kilburn DG (1988) Precise excision of the cellulose binding domains from two *Cellulomonas fimi* cellulases by a homologous protease and the effect on catalysis. J Bio. Chem 120:97–120

Giorda R, Ohmachi T, Shaw DR, Ennis DHL (1990) A shared internal threonine glutamic acid-threonine-proline repeat defines a family of *Dictyostelium discoideum* spore germination-specific proteins. Biochemistry 29:7264–7269

Hayashida S, Mo K, Hosada A (1988) Production and characteristics of Avicel-digesting and non-Avicel-digesting cellobiohydrolases from *Aspergillus ficum.* Appl Environ Microbiol 54:1523–1529

Henrissat B, Driguez H, Viet C, Schulein M (1989) Cellulase families revealed by hydrophobic cluster analysis. Gene 81:83–95

Knowles J, Lehtovaara P, Teeri T (1987) Cellulase families and their genes. Trends Biotechnol 5:255–261

Lao G, Gurdev S, Ghangas ED, Wilson J, Wilson DB (1991) DNA sequences of three $\beta$-1,4-endoglucanase genes from *Thermononospora fusca.* J Bacteriol 173:3397–3407

Love DR, Fisher R, Bergquist PL (1988) Sequence structure and expression of a cloned $\beta$-glucosidase gene from an extreme thermophile. Mol Gen Genet 213:84–92

Lüthi E, Love DR, McAnulty J, Wallace C, Caughey PA, Bergquist PL (1990) Cloning, sequence analysis and expression of genes encoding xylan-degrading enzymes from the thermophile *"Caldocellum saccharolyticum".* Appl Environ Microbiol. 56:1017–1024

Lüthi E, Bhana-Jasmat N, Grayling RA, Love DR, Berguist PL (1991) Cloning, sequence analysis and expression in *Escherichia coli* of a gene coding for a b-mannanase from the extremely thermophilic bacterium "*Caldocellum saccharolyticum*". Appl Environ Microbiol 57:694–700

Mackay RM, Lo A, Willide G, Zuker M, Baird S, Dove M, Moranelli F, Seligy V (1986) Structure of a *Bacillus subtilis* endo-$\beta$-1,4-glucanase gene. Nucleic Acids Res. 14:9159–9170

Meinke A, Braun C, Gilkes NR, Kilburn DG, Miller Jr RC, Warren RAJ (1991) Unusual sequence organization in CenB, an inverting endoglucanase from *Cellulomonas fimi.* J Bacteriol 173:308–314

Montencourt BS (1983) *Trichoderma reesei* cellulases. Trends Biotechnol 1:156–161

Navarro A, Chebrou M-C, Béguin P, Aubert J-P (1991) Nucleotide sequence of the cellulase gene *celF* of *Clostridium thermocellum.* Res Microbiol 142:927–936

Rainey FA, Ward NL, Morgan HW, Stackebrandt E (1993) A phylogenetic analysis of anaerobic thermophilic bacteria, an aid for their reclassification. J Bacteriol 175:4772–4779

Rouvinen J, Bergfors T, Teeri T, Knowles JKC, Jones TA (1990) The three-dimensional structure of cellobiohydrolase II from *Trichoderma reesei.* Science 249:380–386

Sambrook J, Frisch EF, Maniatis T (1989) Molecular cloning: a laboratory manual. 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Saul DJ, Williams LC, Love DR, Chamley LW, Bergquist PL (1989) Nucleotide sequence of a gene from *Caldocellum saccharolyticum* encoding for exocellulase and endocellulase activity. Nucleic Acids Res 17:439

Saul DJ, Williams LC, Grayling RA, Chamley LW, Love DR, Bergquist PL (1990) *celB*, a gene coding for a bifunctional cellulase from the extreme thermophile *"Caldocellum saccharolyticum".* Appl Environ Microbiol 56:3117–3124

Schauder B, Blöcker H, Frank R, McCarthy JEG (1987) Inducible expression vectors incorporating the *Escherichia coli atpE* translational initiation region. Gene 52:279–283

Shoemaker S, Schweickart V, Ladner M, Gelfand D, Kwok S, Myambo K, Innis M (1983) Molecular cloning of exocellobiohydrolase I derived from *Trichoderma reesei* strain L27. Biotechnology 1:691–696

Shoseyov O, Takagi M, Goldstein MA, Doy RH (1992) Primary sequence analysis of *Clostridium cellulovorans* cellulose binding protein A Proc Natl Acad Sci USA 89:3483–3487

Teather RM, Wood PK (1982) Use of Congo Red-polysaccharide interactions in enumeration and characterization of cellulolytic bacteria from bovine rumen. Appl Environ Microbiol 43:777–780

Tucker ML, Milligan SB (1991) Sequence analysis and comparison of avocado fruit and bean abscission cellulases. Plant Physiol 95:928–933

Tucker ML, Durbin ML, Clegg MT, Lewis LN (1987) Avocado cellulase: nucleotide sequence of a putative full-length cDNA clone and evidence for a small gene family. Plant Mol Biol 9:197–203

Wang WK, Kruus K, Wu J HD (1993) Cloning and DNA sequence of the gene coding for *Clostridium thermocellum* cellulase $S_s$ (CelS), a major cellulosome component. J Bacteriol 175:1293–1302

Yanisch-Perron C, Vieira J, Messing J (1985) Improved M13 phage cloning vectors and host strains: Nucleotide sequences of M13mp18 and pUC19 vectors. Gene 33:103.119

**Author's note added in proof**

Further examination of the deduced peptide sequence of CelA revealed that the carboxylterminus of CD1 (amino acids 491–637) has a 44% identity with the C′ cellulose binding domain of *Clostridium stercorarium* (Jauris et al., 1990, Mol Gen Genet 223:258–267). This observation suggests that CelA has in fact two catalytic domains and three cellulose binding domains.