CURRENT MICROBIOLOGY Vol. 40 (2000), pp. 333–340
DOI: 10.1007/s002849910066

Current Microbiology
An International Journal

Multidomain and Multifunctional Glycosyl Hydrolases from the Extreme Thermophile *Caldicellulosiruptor* Isolate Tok7B.1

Moreland D. Gibbs,[1] Rosalind A. Reeves,[1] G. King Farrington,[2] Paige Anderson,[2] Diane P. Williams,[2] Peter L. Bergquist[1,3]

[1]Department of Biological Sciences, Macquarie University, Sydney, New South Wales 2109, Australia
[2]Clariant Corporation, Biotech Research Division, Lexington, MA 02173, USA
[3]Department of Molecular Medicine, University of Auckland Medical School, Private Bag 92019, Auckland, New Zealand

Received: 12 November 1999 / Accepted: 30 November 1999

Abstract. DNA sequencing techniques have revealed widespread molecular diversity of the genomic organization of apparently closely related bacteria (as judged from SSU rDNA sequence similarity). We have previously described the extreme thermophile *Caldicellulosiruptor saccharolyticus*, which is unusual in possessing multi-catalytic, multidomain arrangements for the majority of its glycosyl hydrolases. We report here the sequencing of three gene clusters of glycosyl hydrolases from *Caldicellulosiruptor* sp. strain Tok7B.1. These clusters are not closely linked, and each is different in its organization from any described for *Cs. saccharolyticus*. The catalytic domains of the enzymes belong to glycosyl hydrolase families 5, 9, 10, 43, 44, and 48. The cellulose binding domains (CBDs) of these enzymes from *Caldicellulosiruptor* sp. Tok7B.1 are types IIIb, IIIc, or VI. A number of individual catalytic and binding domains have been expressed in *Escherichia coli*, and biochemical data are reported on the purified enzymes for cellulose degradation encoded by engineered derivatives of *celB* and *celE*.

The enzymes involved in the metabolism of plant carbohydrate polymers have been grouped into nearly 70 different families on the basis of sequence homologies [12–14]. Cellulases are found in glycosyl hydrolase families 5 to 9, 12, 44, 45, 48, and 61. Cellulases are often highly modular in structure and can be composed of either a single domain or a number of distinct domains broadly classified as catalytic or noncatalytic [1]. The usual situation is the covalent association of a single domain with enzyme activity plus one or more cellulose binding domains (CBDs). Linker peptides typically delineate the individual domains of multidomain enzymes into discrete and functionally independent units. Substrate binding domains can enhance the activity of glycosyl hydrolases by binding to cellulose and/or xylan substrates. Considerable homology exists between the CBDs of several xylanases and cellulases [11, 34, 37], of which the following features are highly conserved: (i) low contents of charged amino acids; (ii) two cysteines are present close to the N- and C-termini respectively; (iii) highly conserved tryptophan, glycine, and asparagine residues [35]; (iv) two surfaces with conserved residues, one a planar array of polar and aromatic residues that probably is involved in binding to cellulose, and the other a groove on the opposite side of the molecule that may be a secondary binding site [37].

Microorganisms from New Zealand hot springs are a recognized potential source of alkalophilic thermophilic enzymes [20]. We have examined a number of bacteria isolated from thermal pools for their cellulase activity under alkaline conditions. These samples were enriched on either amorphous cellulose or cotton fibers at pH 7.0 and pH 8.5 and 70°C. The enrichment strategy was based on the assumption that the presence of the cellulosic fibers would induce expression of the cellulase genes in the microorganisms, and from this collection of organisms the anaerobic cellulase producer, Tok7B.1, was isolated. This bacterium was determined to be a strain of the genus *Caldicellulosiruptor* [23] from SSU sequence analysis. We have reported on genes that encode cellulases, mannanases, and xylanases from *Cs. saccharolyticus* [1, 2, 16] and expected to find similar arrangements in *Caldicellulosiruptor* sp. Tok7B.1, but the extensive and

Correspondence to: P.L. Bergquist at Sydney

Table 1. N-terminal sequences of proteins containing cellulase activity isolated from the supernatant of *Caldicellulosiruptor* Tok7B.1 growth supernatants. Six protein bands were observed in the fractions containing cellulase activity. All six proteins were N-terminally sequenced, resulting in two N-termini. Four of the N-termini were translated from the *celE2* gene, and two from the *celB2* gene

Protein band detected	N-terminal amino acid sequence	Apparent molecular mass (kDa)	Genetic Domain (see Fig. 4)
1[a]	AAYNYGEALQKAIMFYEFXM	210	E2
3	AAYNYGEALQ	141	E2
5	GAYNYGEALQ	103	E2
6	GAYNY	84	E2
2	APDWSIPSLWESKYND	205	B2
4	APDWSIPSLW	150	B2

[a] Three separate sequences of B1 resulted in the identification of the first two amino acids as alanine alanine.

complex nature of the glycosyl hydrolase genes from this organism was novel and unusual. Three clusters of genes have been sequenced, and each of them codes for an array of multidomain enzymes. Some of these enzymes have been isolated, and the biochemical characteristics of the recombinant enzymes have been determined.

Materials and Methods

General procedures and assays. Agarose gel electrophoresis, plasmid isolation, M13 mp10 single-stranded DNA isolation, use of DNA modifying enzymes and *E. coli* transformation were performed as previously described [15, 17, 25]. High molecular mass genomic DNA suitable for gene library construction was prepared from cultures of *Caldicellulosiruptor* sp. Tok7B.1 grown anaerobically for 1–2 days with shaking at 70°C as described previously [8]. This DNA was used for the construction of a λZAPII expression library as described previously [8], and the two λ recombinants described below were isolated after staining for cellulase and xylanase activity at 70°C by the substrate overlay method [31]. The soluble cellulose derivative carboxymethyl cellulose and Oat Spelts xylan were used as substrates. Plaques were also screened for cellobiohydrolase activity with the chromogenic substrate methylumbelliferyl cellobioside as described by Saul et al. [26].

CMCase assays, pH rate profiles, and thermostability studies. CMCase activity was determined according to the method of Gilkes et al. [10]. Cellulases were assayed at 50°C for the derivation of pH rate profiles and thermostability as described by Farrington et al. [6].

Analysis of the Tok7B.1 cellulases secreted by the native host. The supernatant resulting from growth of *Caldicellulosiruptor* sp. Tok7B.1 was fractionated by Mono-S ion exchange chromatography. Virtually all cellulase activity from the supernatant was recovered in fractions from the mono-S column (data not shown). Fractions were assayed for CMCase activity, and appropriate samples were analyzed by SDS-PAGE. Proteins revealed by Coomassie Blue staining were blotted and N-terminally sequenced. Two different N-termini were identified of the six proteins sequenced (Table 1).

DNA sequencing and sequence analysis of the Tok7B.1 *xynA* and *celB* genes. Two positive λZAPII plaques, designated W2–4 and N17, were isolated which expressed thermophilic xylanase and/or cellulase activity. Each plasmid was mapped by use of a range of endonucleases after conversion to the plasmid form with the Exassist system (Stratagene, La Jolla, CA). Common digestion patterns indicated that they contained overlapping DNA from the same region of the *Cs.* Tok7B.1 genome. Recombinant DNAs from W2–4 and N17 were partially sequenced by creating simple plasmid deletions with known restriction sites within the Tok7B.1 inserts. Initial DNA sequence comparison data indicated the existence of at least two genes coding for enzymes, including xylanase and an α-arabinosidase domain and several internal cellulase binding domains (CBDs). The genomic DNA insert of W2–4 was sequenced in full, as well as portions of N17, by subcloning and sequencing internal restriction fragments and by using suitable synthetic DNA oligonucleotide primers (Table 2). The complete *xynA* sequence was obtained by genomic walking PCR (GWPCR) as described by Morris et al. [17].

Isolation of *celE* using consensus PCR. Two major cellulolytic enzymes are secreted by *Caldicellulosiruptor* sp. Tok7B.1, and one of them, the *celE* gene product (CelE), was identified by amino-terminal sequencing of purified peptides. CelE showed N-terminal homology to family 9 glycosyl hydrolases from other thermophilic clostridial microorganisms in the GenBank database (Fig. 1). Homology alignments indicated that it would be possible to design consensus oligonucleotide primers complementary to the coding sequence for highly conserved amino acids of these glycosyl hydrolases. Consensus primers could then be used in PCR to amplify any related gene or genes from Tok7B.1. Two primers were designed, the first, named TOKCELA, bound to DNA coding for the peptide sequence -QKAIMFYEF-, and the second, TOKCELR, which bound in the reverse orientation (with respect to the gene sequence) and corresponded to DNA coding for the peptide sequence -DYNAGFVGAL- (Table 2).

Long PCR to confirm postulated gene arrangements. Long PCR (Expand Long Template PCR System, Boehringer Mannheim) was used to determine the positional relationships of sequence information generated with GWPCR [17]. Two examples of PCR products generated by Long PCR are shown in Fig. 3a. Long PCR with the primer CELHGWF in combination with TOKCELERI (Table 2) generated a PCR product 11.2 kb in length. Similarly, the primer CELHGWF was used in combination with the primer AVICELR to generate an 8-kb product. A number of Long-PCR products were generated in this manner and mapped, confirming that the gene encoding ORF5 lies directly upstream of the ORF6 and *celE* genes as shown in Fig. 3a.

PCR cloning of individual genes into expression vectors in *E. coli*. The general method of Gibbs et al. [8] was used to transfer Tok7B.1 cellulase genes into controlled-expression plasmid vectors, initially into pJLA602 [27], but later into the pET9a vector (Novagen, CA, USA) [29]. Several truncated genes encoding either individual catalytic domains or catalytic domains connected to CBDs by linker sequences were constructed.

Phylogenetic analysis of Tok7B.1. The 16S (small subunit, SSU) rRNA gene was isolated by PCR with oligonucleotide primers designed to amplify the SSU rRNA gene from all known prokaryotic species. The approximately 1500-bp PCR fragment obtained was cloned into M13 mp10 in the forward and reverse orientation and sequenced. Close homologs of the Tok7B.1 SSU rRNA gene were aligned with the GCG multiple alignment software 'Pileup' and analyzed by parsimony methods [30]. Figure 2 shows the phylogenetic position of Tok7B.1 among cluster D of thermophilic clostridia [22]. It is closely related to two other *Caldicellulosiruptor* strains from which we have cloned glycosyl hydrolases previously, *Cs. saccharolyticus* and *Cs.* strain Rt8B.4 [5, 23].

Table 2. Oligonucleotide primers designed and synthesized for PCR amplification, GWPCR and sequencing of glycosyl hydrolase genes from *Caldicellulosiruptor* Tok7B.1. Engineered restriction sites are shown in reverse text

Primer name	Nucleotide sequence	Target gene	Engineered site
AVICELR	5′-TGTATCCCATGCCGTCTT	*orf4, orf6*	—
CELAF	5′-CAAAGCAGACGAATCTGT	*xynA*	—
CELAGWR	5′-TGGTGCTGGCAATGTTGAGTTGGC	*xynA*	—
CELAGWR2	5′-TCGGTAGTGCCACTTTCAAATCCA	*xynA*	—
CELHGWF	5′-GAGAAACATATCCTGCAA	*orf5*	—
CELHGWR	5′-CCCATTTTATACCCAGGC	*orf5*	—
CELHGWR2	5′-TCTTGAGCAGCCATTGGA	*orf5*	—
N17A	5′-GATGGCCAGTTCACGTTTATATGG	*celB, orf4*	—
TOKCBDF	5′-GAGGAACGGTCATATGAAGGTATGGTATGCGAATGGGAA	All	*Nde*I
TOKCBDXSR	5′-GTGCAGCTCGAGCTCCTCCCGGCTCCTGCCCCCA	All	*Xho*I, *Sac*I
TOKCELA	5′-CAAAAAGCAATTATGTTTTATGAATT	*celE*	—
TOKCELBR	5′-GAAGTATGGATCCATTTATTAATTCTTTGGG	*celB*	*Bam*HI
TOKCELEBAMR	5′-CCTGGATCCCTACGCTCCTCCCGGCTC	*celE*	*Bam*HI
TOKCELEFI	5′-ATGCAAGGCATGCAAGCAATTAAGAGGGTTG	*celE*	*Sph*I
TOKCELEFII	5′-GGGAATTCCATATGGCGGCGTATAATTACGGTG	*celE*	*Nde*I
TOKCELEGWF	5′-AGCACTGGTTGGTGGTCCTGGTAG	*celE*	—
TOKCELERI	5′-TCAACAAAGATCTAATCATTTGTGGGTGTTTC	*celE*	*Bgl*II
TOKCELERII	5′-GTGGATGAGATCTAACCCGGCTCTAAACCCCA	*celE*	*Bgl*II
TOKCELERIII	5′-GCAGCAGTGTCGACATTTTTATTCTTTAATCTAC	*celE, orf5*	*Sal*I
TOKCELERIV	5′-TATTATTATCATATGCGGC	*celE*	*Nde*I
TOKCELGGWF	5′-TTGAGGGATATGGTGACC	*orf6*	—
TOKCELGWFII	5′-GATTGACGGGTTACAATTGGGAGAAC	*celE, orf5F*	—
TOKCELR	5′-AGWGCACCNACAAATCCGGCATTGTARTC	*celE*	—
TOKGW	5′-CTCCAGAATGTCATTTGTAAGATACAT	*xynA*	—

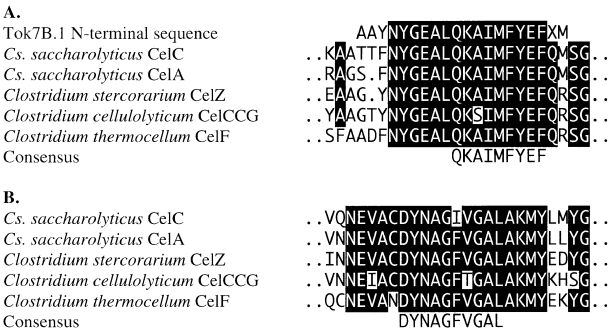


Fig. 1. Regions of highly conserved amino acid sequences of glycosyl hydrolase family 9 domains used to design the consensus primers (A) TOKCELA and (B) TOKCELR.

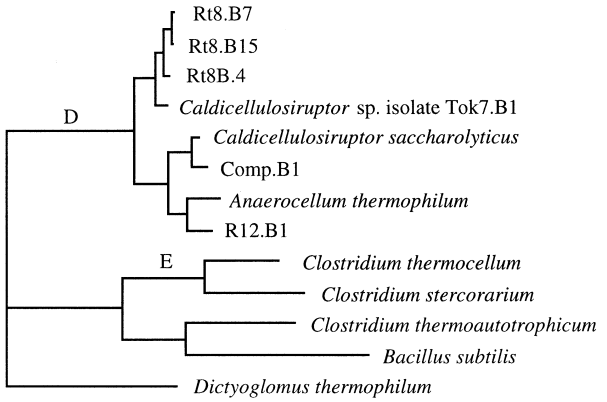


Fig. 2. Phylogenetic relationships of *Caldicellulosiruptor* strains in clostridial clusters D and E of Rainey et al. [24], as determined from SSU sequence data with PAUP version 3.1.1. [30]. The tree is rooted with *Dictyoglomus thermophilum* as an outgroup.

Results and Discussion

Glycosyl hydrolases expressed by the native organism. Analysis of cellulase proteins isolated from *Caldicellulosiruptor* sp. Tok7B.1 culture supernatant showed that CelE and CelB were expressed and secreted under the conditions used for the growth of the Tok7B.1 strain. Mono S-Sepharose fractionation and SDS-PAGE revealed two high-molecular-weight protein species that exhibited CMCase activity. These proteins were designated CelB and CelE. A number of lower molecular weight proteins were also detected, and the N-termini of each indicated that they arose from the N-terminal portion of the CelB or CelE proteins. In all cases, the proteins were exact matches for the Tok7B.1 gene sequences of *celE* and *celB* except for the N-terminus of CelE, which did not exactly match the expected sequence from the DNA. The N-terminus of the native protein would have been predicted to be GTYNY, but instead the sequence found was AAYNY. This sequence was incorporated into the N-termini of the *celE* gene products

expressed by *E. coli*. These results suggest that the *Caldicellulosiruptor* sp. Tok7B.1 may modify post-translationally the amino acids on the N-terminus of these proteins, or its translational machinery may deviate from the standard genetic code.

Architecture of glycosyl hydrolase genes on the *Caldicellulosiruptor* sp. Tok7B.1 genome. Many bacteria have been reported to carry a multiplicity of genes for cellulases and hemicellulases [9, 34, 35]. Figure 3a is a diagrammatic representation of the three clusters of glycosyl hydrolases from *Caldicellulosiruptor* sp. Tok7B.1, which form the subject of this report. Note that only *xynA*, *celB*, and *celE* have been sequenced on both strands and that some gene orders have been confirmed only by long PCR and Southern blots. They are different in their organization from the situation previously described for *Cs. saccharolyticus* [refs. 1, 16; see Fig. 3b). Several putative catalytic domains have been identified on the basis of sequence similarity to characterized enzymes, and it has not been possible to confirm the putative enzymatic activity on any substrate available to us. For example, a plasmid expressing the recombinant N-terminal domain of ORF6 demonstrated no enzymatic activity on any cellulosic substrate, any mannan or glucomannan, on lichenan, cellobiose, or β-glucoside (data not shown). While we cannot rule out the possibility that an inactive protein was expressed in *E. coli*, it should be noted that all other recombinant genes expressed were highly active, indicating that they have maintained their native confirmation. The catalytic domains belong to a limited number of glycosyl hydrolase families (5, 9, 10, 43, 44, and 48) [refs. 12, 13], and there are no genes coding for multidomain enzymes containing a family 5 β-mannanase as seen with *manA* and *celC* of *Cs. saccharolyticus* (see Fig. 3b). The CBDs are of either type III or VI [35]. The positions for the primers (listed in Table 2) used for the confirmation of gene order by Long-PCR are shown in Fig. 3a. The important features of the genes in the clusters are described below.

***xynA* and *celB*.** Representative GWPCR products spanning the region of the *xynA* gene and a representation of the complete DNA sequence containing the *xynA* and *celB* genes are displayed in Fig. 3a, with each gene shown according to its translated domain structure. The DNA can be seen to code for two genes, each of which contains a family 10 xylanase, in one case with putative thermostabilizing domains (TSDs) [ref. 40] and CBDs of type III [35]. There are different C-terminal catalytic domains. The gene *xynA* possesses an α-arabinosidase of family 43 and *celB* an endoglucanase from family 5 that is active on carboxymethyl cellulose (see Fig. 3a). At the 5′ and 3′ ends of the cluster are open reading frames whose sequences have no similarity to other sequences in the GenBank database. The translated product of the *celB* gene matches perfectly with two amino-terminal sequences obtained for native cellulolytic peptides secreted by Tok7B.1 (peptides B2 and B4, see Table 1), implying that the *celB* gene expresses one of the major cellulases secreted by the bacterium.

CelE and other putative glycosyl hydrolases. Homology alignments indicated that it would be possible to design consensus oligonucleotide primers that would bind to DNA coding for clusters of highly conserved amino acids found in all thermophilic clostridial family 9 glycosyl hydrolases (see Fig. 1). A PCR product amplified with the consensus primers TOKCELA and TOKCELR was ligated into M13 mp10, transformed into *E. coli* strain JM101, and plated to give individual recombinant plaques, the DNA of which were sequenced. After sequencing, GWPCR primers were designed and used to obtain the complete *celE* gene sequence (Fig. 3a and Table 2).

CelE consists of two catalytic domains, both endoglucanases, but from different families (9 and 44) and four type III binding domains with PT linkers. Analysis of the GWPCR products revealed that there were further sequence similarities with cellulases 5′- to *celE*. GWPCR was used to obtain DNA sequence from upstream of *celE* (Fig. 3a). Two further genes were identified in this way. Both of these genes, designated ORF5 and ORF6, are assumed to code for multidomain, multicatalytic proteins on the basis of sequence similarity comparisons, with the same general structure as XynA, CelB, and CelE. As the DNA sequence obtained was not contiguous, long-PCR was used to amplify DNA between the sequenced regions to confirm that they were linked (Fig. 3a). Approximately 13,500 bp of genomic DNA 5′- to the *celE* gene was partially sequenced.

Genes coding for ORF3 and ORF4. The primer N17A (Table 2) was used as a GWPCR primer during the isolation of the complete *xynA* gene sequence. A number of PCR products were obtained that did not match DNA sequence already obtained for the *celB* and *xynA* genes. It was clear from these results that the N17A primer was not uniquely specific for *celB* and was binding at other sites on the genome with high homology to the 5′- end of *celB*. A number of the GWPCR products were sequenced and shown to code for a protein with an amino-terminal domain identical to that of CelB. However, the gene sequence clearly was not derived from the *celB* gene as the intergenic sequence upstream of the start codon differed from the intergenic sequence upstream of *celB*. Also, the 3′-terminus of a second gene (coding for a family 48 glycosyl hydrolase) was identified upstream of

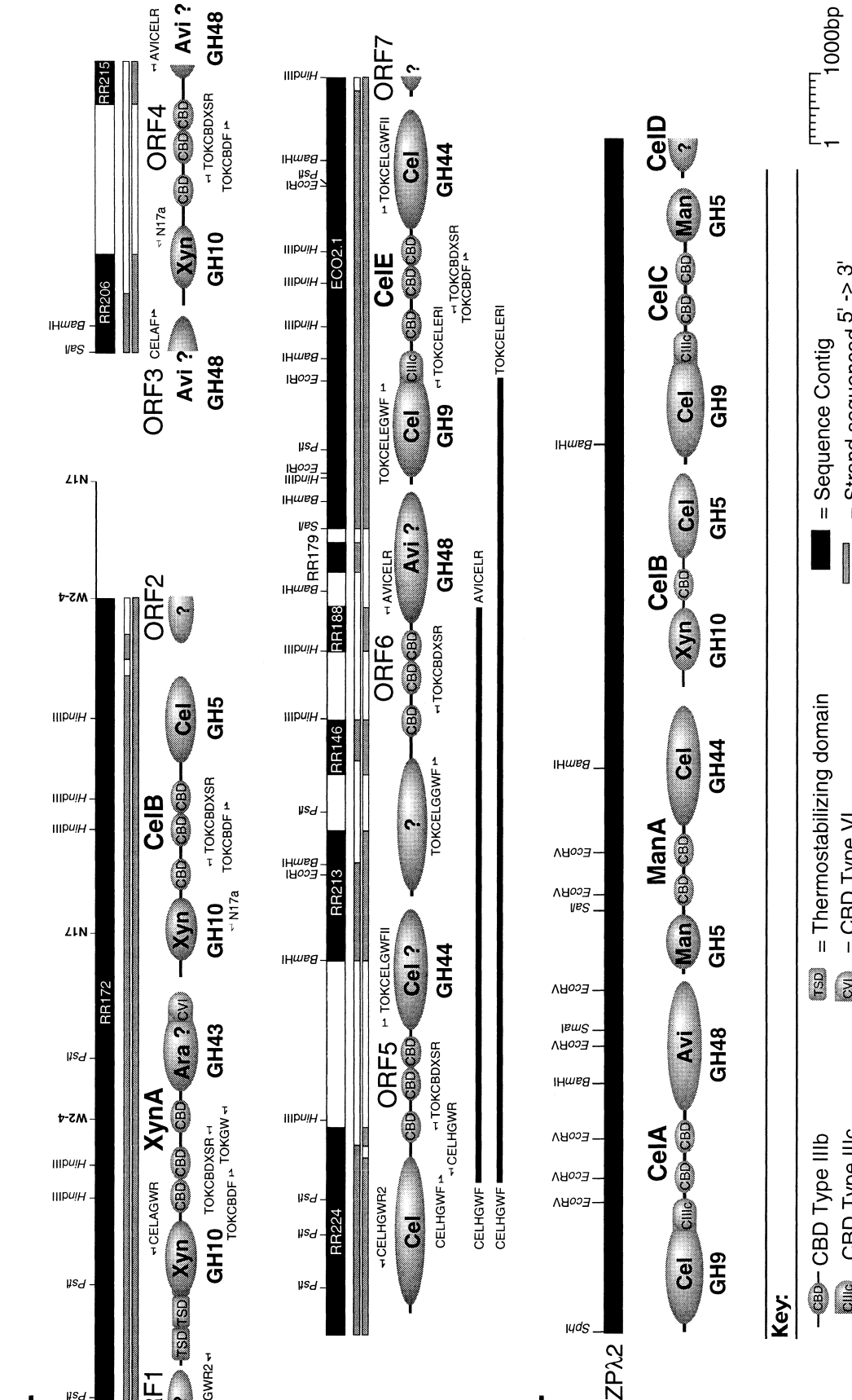
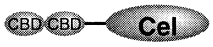


Fig. 3. Overall architecture of the three putative glycosyl hydrolase gene clusters sequenced from *Caldicellulosiruptor* Tok7B.1. Where complete sequence information is available, it is shown as a black box. Sequence information obtained for the forward and reverse strand is shown as thick grey lines. A stylized representation of the gene products is provided below the line. The glycosyl hydrolase family designation of each catalytic domain is indicated in brackets. Some restriction sites are named and λZAPII recombinant boundaries indicated (W2–4, N17). The positions of some PCR primers used for GWPCR and Long PCR are shown (see also Table 1). Representative examples of Long PCR products are shown as thin black lines. The key included in the figure shows the several types of bacterial CBDs encoded by the DNA. The class and presence of some CBDs are inferred by similarity comparisons and have not been sequenced in full. The DNA sequences shown in this figure can be obtained from GenBank under the accession numbers AF078038, AF078039, AF078040, AF078041, AF078042, AF078043, AF078044, AF078737. (A) Clusters of glycosyl hydrolase genes and their putative products from *Caldicellulosiruptor* strain Tok7B.1. (B) The cellulase gene cluster from *Cs. saccharolyticus* reported previously [3, 17, 32].

Name	Domain Structure	Thermostability [1]	pH [2]
CelE1/2/3		75°C	4-11
CelE1/2		75°C	4-11
CelE1		50°C	5-9
CelB5		70°C	4-10
CelB4/5		60°C	4-10
CelB3/4/5		60°C	5-10
CelE3/CelB5		60°C	4-10

[1] Maximum temperatures at which there was no loss of activity for 45 min. Assays were carried out at pH 7.0.

[2] pH at which at least 10% of the maximum activity is observed. Assays were carried out at 50°C.

Fig. 4. Diagrammatic representation of the recombinant proteins expressed from the various constructions involving CelB and CelE, showing the catalytic and binding domains and linker sequences. Each protein is named according to the domains included in the recombinant protein, where the catalytic and non-catalytic domains are numbered from the N-terminus of the native protein. Signal peptides and domain linkers are not numbered in this nomenclature. The pH range for activity and the temperature optimum of each recombinant protein under the conditions used for the assay are shown.

the *celB* homolog. This gene shares no similarity with the *xynA* gene lying upstream of *celB*. These genes were designated *orf3* and *orf4* (Fig. 3a). Oligonucleotide primers specific to the 3′-terminal end of the *orf3* gene and the 5′-terminal end of the *orf4* gene were synthesized and used in combination with primers that bound to DNA coding for the CBDs found in *xynA*, *celB*, *celE*, *orf5*, and *orf6*. The *orf3* and *orf4* genes were identified after the amplification and sequencing of PCR products. The results obtained indicated that they coded for proteins with the same basic domain structure of the other Tok7B.1 cellulolytic genes. The amino-terminal domain of ORF3 could not be identified, but the carboxy-terminal product of ORF4 is homologous to family 48 glycosyl hydrolases and has a high degree of similarity to the carboxy-terminal domains of ORF3 and ORF6. Clearly, there has been substantial domain duplication and recombination among the glycosyl hydrolases of Tok7B.1, and possibly, lateral gene transfer [reviewed in ref. 3].

Cloning of the genes for selected catalytic and binding domains. The structure of the catalytic and binding domains contained in the various pJLA602 and pET9a recombinant plasmids are shown in Fig. 4. References to related sequences are given, along with the predicted function from similarity comparisons. Strains of *E. coli* carrying each of these plasmids were grown up in enriched medium in a fermenter and induced with heat or IPTG for the production of the individual enzymes, which were then purified and characterized as described elsewhere [6].

Characterization of the activities of the *celB* and *celE* gene products. Under the assay conditions used, the highest temperature at which no loss of enzyme activity was detected after 45 min of incubation varied from 50° to 75°C depending on the number of CBDs associated with the purified recombinant enzyme (Fig. 4). There were no significant differences in temperature stability observed whether or not the catalytic domain was associated with CBDs of type IIIb and their accompanying PT linker sequences (for example, compare CelE1/2 with CelE1/2/3). However, the absence of a CBD that was not associated with a PT linker (for example, type IIIc CBD of CelE1/2) had a dramatic effect, as seen in the results for CelE1 compared with CelE1/2 (Fig. 4). The pH rate profiles of the CMC assays paralleled the thermostability results. The presence of catalytic domains plus CBDs with PT linkers had little effect, but the addition of a type IIIc CBD to a catalytic domain (for example, CelE1/2) significantly broadened the pH range for 50% activity. The pH profiles for CelB5 and CelB4/5 were identical, but the catalytic domain CelB5 was slightly more thermostable than the construction containing the PT linkers and the type IIIb CBD. A full description of the purification of the recombinant enzymes, discussion of the kinetic parameters of the recombinant cellulases, and their biochemical properties are included in the paper by Farrington et al. [6].

Overall architecture of the glycosyl hydrolases. A number of multifunctional enzymes involved in cellulose or hemicellulose degradation have been reported recently and their genes cloned and expressed in *Escherichia coli*. These include CenC from *Cellulomonas fimi* [36], which has both exo- and endocellulase activity, and the bifunctional xylanase from *Ruminococcus flavifaciens* [42]. Bacteria of the genus *Caldicellulosiruptor* seem to have a propensity for this multidomain structure, as shown by the characterization of CelA from the closely related '*Anerocellum thermophilum*' [41] and XynA, B, and C of *Caldicellulosiruptor* sp. isolate Rt69B.1 [18].

Multidomain/multifunctional cellulose/hemicellulose-degrading enzymes are a common arrangement in *Cs. saccharolyticus*. All the cellulase/hemicellulase enzymes from this organism (CelA, ManA, CelB, and CelC/ManB) contained at least one CBD with two to three proline-threonine-rich linkers, and two catalytic domains [16]. The situation in *Caldicellulosiruptor* sp. Tok7B.1 is related but more complex. Figure 3a shows a stylized line-up of three of the glycosyl hydrolase genes that we

have sequenced. The CBDs of XynA, CelB, and CelE that are separated from the catalytic domains by PT linkers are all classified in the Type IIIb subgroup of Tormo et al. [37], where the conserved amino acids of the planar array of polar and aromatic residues are believed to interact with crystalline cellulose. These residues have been shown to be involved in substrate binding in the example of the modular cellulase, CelZ, from *Clostridium stercorarium* [24]. Other open reading frames are inferred to contain similar CBDs, but the sequence data are incomplete. The Type IIIc CBD of CelE of Tok7B.1 has been demonstrated to have a thermostabilizing role [6], confirming similar results from *C. stercorarium* CelZ, where virtually the complete domain was required for thermostability and catalytic activity [24].

What is the explanation for the diversity in gene structure found in homologous genes in closely related bacteria? It has been generally assumed that the glycosylhydrolases have evolved by domain-shuffling [34, 35, 39], although exact mechanisms have not been described. Linkers are often found in xylanases and cellulases [7, 16], which are thought to function as flexible hinges between the catalytic and substrate-binding domains. The DNA encoding the repeated linkers may have a role analogous to that of introns, enabling sequences that encode discrete domains to be excised and fused to other genes, thus generating novel hybrid enzymes [9]. Another possibility is that, after duplication of glycosyl hydrolase genes by DNA replication, a recombinational event similar to that postulated for the origin of multiple tRNA genes ['unequal crossing-over', ref. 28] or intragenic recombination [4, 21] could give rise to the genes coding for multidomain enzymes seen on the genomes of most cellulolytic bacteria. Furthermore, there are superficial similarities in the organization of the CBDs. A simple example that could be attributed to intragenic recombination is shown by what appears to be an inverted orientation for the CBDs of XynA from *Caldicellulosiruptor* sp. Tok7B.1 (Fig. 3a) in comparison with the other related proteins (CelB, CelE). This may be explained by the occurrence of two intragenic cross-over events in the DNA coding for the PT-linker regions [3]. However, although this is a plausible model, alignment of the amino acid sequences of the CBDs and a dendrogram of their relationships suggests that none of the CBD arrangements observed was the immediate precursor of the inverted XynA structure (data not shown).

It may be significant that *Caldicellulosiruptor* is related phylogenetically to ancient and early-branching organisms, and the unique array of multifunctional enzymes with catalytic domains carrying out related activities in the hydrolysis of insoluble substrates may represent a persistent evolutionary experiment that developed before the organization of genes into operons. A cluster of genes in the same orientation is frequently part of an operon and is regulated by transcription from a single promoter. A possible primitive alternative regulatory mechanism may have had the genes encoding the hydrolytic enzymes fused, resulting in the production of a multifunctional protein on transcription. Multifunctional enzymes would guarantee equivalent transcription and translation of the related enzyme activities, and the substrate-binding domain(s) would ensure coordinate action at the same site on the substrate. Others have constructed gene fusions coding for multidomain enzymes by recombinant DNA techniques [19, 33, 38]. Riedel and Bronnenmeier [24] constructed a fusion between the *celY* and *celZ* of *C. stercorarium* and showed a significant increase in synergism of the fusion protein in the hydrolysis of crystalline cellulose as compared with the enzymes coded by the individual genes. Thus, a plausible alternative explanation for multidomain enzymes may be the selective growth advantage they offer cellulolytic bacteria.

ACKNOWLEDGMENT

This work was supported in part by a grant from the Macquarie University Research Committee.

Literature Cited

1. Bergquist PL, Gibbs MD, Saul DJ, Te'o VSJ, Dwivedi PP, and Morris DD (1993) Molecular genetics of thermophilic bacterial genes coding for enzymes involved in cellulose and hemicellulose degradation. In: Shimada K, Hoshino K, Ohiyma K, Sakka K, Kobayashi Y, Karita S (eds). Tokyo, Japan: Uni Publishers Co. pp 276–285
2. Bergquist PL, Gibbs MD, Saul DJ, Reeves RA, Morris DD, Te'o VSJ (1996) Families and functions of novel thermophilic hemicellulases in the facilitated bleaching of pulp. In: Jeffries TJ, Viikari L (eds) Enzymes for pulp and paper processing, vol. 655. Washington, D.C.: American Chemical Society Symposium Series, 85–100
3. Bergquist PL, Gibbs MD, Morris DD, Te'o VS, Saul DJ, Morgan HW (1999) Molecular diversity of thermophilic cellulolytic and hemicellulolytic bacteria. FEMS Microbiol Ecol 28:99–110
4. Cooper VJC, Salmond GPC (1993) Molecular analysis of the major cellulase (CelV) of *Erwinia carotovora*: evidence for an evolutionary ''mix-and-match'' of enzyme domains. Mol Gen Genet 241:341–350
5. Dwivedi PP, Gibbs MD, Saul DJ, Bergquist PL (1996) Cloning, sequencing and overexpression in *Escherichia coli* of a xylanase gene, *xynA* from the thermophilic bacterium Rt8B.4 genus *Caldicellulosiruptor*. Appl Microbiol Biotechnol 45:86–93
6. Farrington GK, Gibbs MD, Anderson P, Eldredge J, Bergquist PL, Williams DP (1999) Biochemical characterization of multidomain cellulase genes from the extreme thermophile *Caldicellulosiruptor* sp. isolate Tok7B.1. Biochem J, submitted
7. Ferreira LMA, Wood TM, Williamson G, Faulds C, Hazlewood GP, Black GW, Gilbert HJ (1993) A modular esterase from *Pseudomonas fluorescens* subsp. *cellulosa* contains a non-catalytic cellulose-binding domain. Biochem J 294:349–355
8. Gibbs MD, Reeves RA, Bergquist PL (1995) Cloning, sequencing, and expression of a xylanase gene from the extreme thermophile

Dictyoglomus thermophilum Rt46B.1 and activity of the enzyme on fiber-bound substrate. Appl Environ Microbiol 61:4403–4408
9. Gilbert HJ, Hazelwood GP (1993) Bacterial cellulases and xylanases. J Gen Microbiol 139:187–194
10. Gilkes NR, Langsford ML, Kilburn DG, Miller Jr RC, Warren RAJ (1984) Mode of action and substrate specificities of cellulases from cloned bacterial genes. J Biol Chem 259:10455–10459
11. Gilkes NR, Henrissat B, Kilburn DG, Miller MC, Warren RAJ (1991) Domains in microbial β-1,4-glycanases: sequence conservation, function, and enzyme families. Microbiol Rev 55:2303–2315
12. Henrissat B (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J 280:309–316
13. Henrissat B, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J 293:781–788
14. Henrissat B, Bairoch A (1996) Updating the sequence-based classification of glycosyl hydrolases. Biochem J 316:695–696
15. Messing J (1983) New M13 vectors for cloning. In: Wu R, Grossman L, Moldave K (eds) Recombinant DNA, vol. 101 (Part C). New York: Academic Press, pp 20–78
16. Morgan HW, Bergquist PL (1996) Cellulases and hemicellulases of extreme thermophiles. In: Asche V (ed) Melbourne, Australia: Australian Society for Microbiology Inc., vol. iv: pp 83–118
17. Morris DD, Reeves RA, Gibbs MD, Saul DJ, Bergquist PL (1995) Correction of the β-mannanase domain of the *celC* pseudogene from *Caldicellulosiruptor saccharolyticus* and activity of the gene product on kraft pulp. Appl Environ Microbiol 61:2262–2269
18. Morris DD, Gibbs MD, Ford M, Thomas J, Bergquist PL (1999) Sequence analysis of the multidomain family 10 and 11 xylanase genes from *Caldicellulosiruptor* sp. isolate Rt69B.1. Extremophiles 3:103–112
19. Olsen O, Thomsen KK, Weber J, Duus JØ, Svendsen I, Wegener C, von Wettstein D (1996) Transplanting two unique β-glucanase catalytic activities into one multienzyme which forms glucose. Bio/Technology 14:71–76
20. Peak K, Ruttersmith LD, Daniel RM, Morgan HW, Bergquist PL (1992) Thermophilic enzymes as industrial catalysts. Bioforum Eur 9:466–470
21. Quillet L, Barray S, Labedan B, Petit F, Guespinmichel J (1995) The gene encoding the β-1,4-endoglucanase (Cela) from *Myxococcus xanthus*—evidence for independent acquisition by horizontal transfer of binding and catalytic domains from Actinomycetes. Gene 158:23–29
22. Rainey F, Ward N, Morgan H, Toalster R, Stackebrandt E (1993) Phylogenetic analysis of anaerobic thermophilic bacteria: aid for their reclassification. J Bacteriol 175:4772–4779
23. Rainey FA, Donnison AM, Janssen PH, Saul D, Rodrigo A, Bergquist PL, Daniel RM, Stackebrandt E, Morgan HW (1994) Description of *Caldicellulosiruptor saccharolyticus* gen nov, sp nov—an obligately anaerobic, extremely thermophilic, cellulolytic bacterium. FEMS Microbiol Lett 120:263–266
24. Riedel K, Bronnenmeier K (1998) Intramolecular synergism in an engineered exo-endo-1,4-β-glucanase fusion protein. Mol Microbiol 28:767–775
25. Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd edu. New York, NY. Cold Spring Harbor Laboratory Press
26. 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
27. Schauder B, Blöcker H, Frank R, McCarthy JEG (1987) Inducible expression vectors incorporating the *Escherichia coli atpE* transcriptional initiation region. Gene 52:279–283
28. Smith JD, Barnett L, Brenner S, Russell RL (1970) More mutant tyrosine transfer ribonucleic acids. J Mol Biol 54:1–14
29. Studier FW, Rosenberg AH, Dunn JJ, Dubendorff JW (1990) Use of T7 RNA polymerase to direct expression of cloned genes. Methods Enzymol 185:60–89
30. Swofford DL (1993) Phylogenetic analysis using parsimony (PAUP). Washington D.C., USA: Smithsonian Institution
31. Teather RM, Wood PJ (1987) Use of congo red polysaccharide interaction in enumeration and characterisation of cellulolytic bacteria from bovine rumen. Appl Environ Microbiol 43:777–780
32. Teo VSJ, Saul DJ, Bergquist PL (1995) CelA, another gene coding for a multidomain cellulase from the extreme thermophile *Caldocellum saccharolyticum*. Appl Microbiol Biotechnol 43:291–296
33. Tomme P, Gilkes NR, Miller RC, Warren AJ, Kilburn DG (1994) An internal cellulose-binding domain mediates adsorption of an engineered bifunctional xylanase cellulase. Protein Eng 7:117–123
34. Tomme P, Warren RAJ, Miller RCJ, Kilburn DG, Gilkes NR (1995a) Cellulose-binding domains: classification and properties. In: Saddler JN, Penner MH (ed) In: The enzymatic degradation of insoluble polysaccharides. American Chemical Society Symposium 618:142–163
35. Tomme P, Warren RAJ, Gilkes NR (1995b) Cellulose hydrolysis by bacteria and fungi. Adv Microb Physiol 37:1–81
36. Tomme P, Kwan E, Gilkes NR, Kilburn DG, Warren RAJ (1996) Characterization of CenC, an enzyme from *Cellulomonas fimi* with both endo- and exoglucanase activities. J Bacteriol 178:4216–4223
37. Tormo J, Lamed R, Chirino AJ, Morag E, Bayer EA, Shoham Y, Steitz TA (1996) Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. EMBO J 15:5739–5751
38. Warren RAJ, Gerhard B, Gilkes NR, Owolabi JB, Kilburn DG, Miller RC (1987) A bifunctional exoglucanase-endoglucanase fusion protein. Gene 61:421–427
39. West CA, Elzanowski A, Yeh L-S, Barker WC (1989) Homologues of catalytic domains of *Cellulomonas* glucanases found in fungal and *Bacillus* glycosidases. FEMS Microbiol Lett 59:167–172
40. Winterhalter C, Heinrich P, Candussio A, Wich G, Liebl W (1995) Identification of a novel cellulose-binding domain within the multidomain 120 kDa xylanase XynA of the hyperthermophilic bacterium *Thermotoga maritima*. Mol Microbiol 15:431–444
41. Zerlov V, Mahr S, Reidel K, Bronnmeier K (1998) Properties and gene structure of a bifunctional cellulolytic enzyme (CelA) from the extreme thermophile '*Anaerocellum thermophilum*' with separate glycosyl hydrolase family 9 and 48 domains. Microbiology 144:457–465
42. Zhang J, Flint HJ (1992) A bifunctional xylanase encoded by the *xynA* gene of the rumen cellulolytic bacterium *Ruminococcus flavefaciens* 17 comprises two dissimilar domains linked by an asparagine/glutamine-rich sequence. Mol Microbiol 6:1013–1023