Journal of Industrial Microbiology, 6 (1990) 285–290
Elsevier

SIM 00294

DNA sequence analysis of endoglucanase genes from *Pseudomonas fluorescens* subsp. *cellulosa* and *Pseudomonas* sp. NCIB 8634.

Bruce R. Wolff, Bernard R. Glick and J.J. Pasternak

Department of Biology, University of Waterloo, Waterloo, Ontario, Canada

(Received 11 September 1989; accepted 8 January 1990)

Key words: Cellulase; Endoglucanase; Gene sequence

SUMMARY

The DNA of two previously isolated recombinant clones, one from *Pseudomonas* sp. NCIB 8634 (= *Cellvibrio mixtus*) (pPC71) and another from *Pseudomonas fluorescens* subsp. *cellulosa* (pPFC4) that express endoglucanase activity in *E. coli* was sequenced. Plasmid pPC71 had three open reading frames, two of which include portions of plasmid pBR322. The third open reading frame occurs entirely within the *Pseudomonas* DNA insert and encodes a protein with a molecular mass of 5845 Da. The DNA insert in pPFC4 was found to contain an open reading frame (PFC-ORF) that encodes a protein of 32189 Da. The major endoglucanase produced in *E. coli* cells carrying pPFC4 is about 30000 Da [26]. It is concluded that PFC-ORF encodes this endoglucanase. Both ribosome and catabolite gene activator protein binding sites lie upstream from the initiating codon of PFC-ORF. An interesting feature of the PFC-ORF protein is the presence of amino acid motifs Val-Ser-Ser-Ser-Ser and Val-Val-Ser-Ser-Ser-Ser-Ser that occur within a 25 amino acid span.

INTRODUCTION

A number of microorganisms are capable of enzymatically degrading cellulose by the synergistic activities of endoglucanase and exoglucanase to produce cellobiose. In some microorganisms, cellulolysis proceeds further with β-glucosidase converting cellobiose to glucose. In many microbial systems, the cellulolytic process is mediated by a multienzyme complex [11,27]. In *Clostridium thermocellum*, for example, this complex (viz., cellulosome) is comprised of at least 15 different components [12,17]. How this and other similar complexes are assembled, the enzymic features of the constituent proteins and the derivation of these moieties have been the targets of recent research efforts [9,15]. Generally, it appears that families of genes encode the components that make up these multienzyme complexes although the multiplicity of some of the components may be due to post-translational modification (e.g., ref. 26).

Multiple forms of endoglucanase activity have been reported for both *Pseudomonas fluorescens* subsp. *cellulosa* and *Pseudomonas* sp. NCIB 8634 (a.k.a. *Cellvibrio mixtus*) [19,29,31]. As well, molecular cloning has resulted in the isolation of different endoglucanase genes from *P. fluorescens* subsp. *cellulosa* which appear on the basis of restriction endonuclease mapping to be distinct [5,6,14,25]. Previously, two cloned DNA fragments, one from *P. fluorescens* subsp. *cellulosa* and the other from *Pseudomonas* sp. NCIB 8634 were each shown to encode endoglucanase activity in *Escherichia coli* [25]. In the present study, the endoglucanase genes that are encoded by these two pseudomonad DNA fragments were sequenced. The gene from *P. fluorescens* subsp. *cellulosa* encodes a protein, which prior to cleavage of a putative leader sequence, is equivalent to 32189 Da; whereas, the open reading frame represented within the cloned DNA fragment from *Pseudomonas* sp. NCIB 8634 may code for an endoglucanase with a molecular mass of 5845.

MATERIALS AND METHODS

The methods for DNA isolation, bacterial transformation and DNA cloning followed established protocols [1,15]. Enzymes were purchased from Bethesda Research Laboratories, Pharmacia, Boehringer-Mannheim and New England Biolabs and used in accordance with the manufacturers' directions. For DNA sequencing reactions, [α-^{35}S]dATP was purchased from New England Nuclear. Endoglucanase activity in *E. coli* carrying plasmid-borne chromosomal fragments from either *Pseudomonas fluorescens* subsp. *cellulosa* or *Pseudomonas*

Correspondence: B.R. Glick, Department of Biology, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

sp. NCIB 8643 (= *Cellvibrio mixtus*) was scored by the congo red-carboxymethyl cellulose overlay method [23].

Restriction endonuclease fragments were subcloned into vectors M13mp18 and M13mp19 [30] and nucleotide sequences were determined using the dideoxynucleotide chain-termination method [20]. The DNA sequence of each subcloned fragment was usually determined from complementary strands using M13 sequencing primers. When overlapping segments of the same region were not available several independently isolated clones were sequenced. DNA sequences were determined across all restriction site junctions in the M13 subclones. DNA sequence data were entered into an Apple IIe computer and analyzed by means of the "DNA and Protein Sequence Analysis" program [16]. Hydropathy analysis was carried out using the algorithm of Kyte and Doolittle [10] and graphically represented using an 11-residue window. Searches of the EMBL (Release 15.0) and NBRF-PIR (Release 16.0) databases for sequence similarities were conducted through Bionet National Computer Resource for Molecular Biology (Mountain View, CA) using the FASTA program [18].

RESULTS

A *Pseudomonas* sp. NCIB 8643 genomic library that was constructed in pBR322 and maintained in *Escherichia coli* was previously screened for endoglucanase activity and two clones, pPC71 and pPC72, were isolated [25]. Further analysis revealed that both of these clones contained the same DNA fragment, i.e., about 450 bp and lacking *Pvu*II, *Pst*I, *Bam*HI, *Eco*RI, *Sal*I, *Hin*III, *Kpn*I, *Xho*I and *Xba*I restriction endonuclease sites [25].

Cells carrying pPC71 are unstable with respect to maintaining endoglucanase activity. After growth overnight in liquid culture, up to 50% of the cells give a negative response with the congo red-carboxymethylcellulose assay. These endoglucanase-negative cells carried plasmid DNA from which the *Pseudomonas* DNA had been deleted. To obtain a more stable clone, the insert of pPC71 with flanking regions from pBR322 (i.e., an *Eco*RI-*Sal*I fragment) was subcloned into pUC19 to yield pUC/71ES. *E. coli* cells that were transformed with pUC/7IES did not produce endoglucanase activity although the inserted DNA was stably maintained. The *Eco*RI-*Sal*I fragment from pUC19/71ES was cloned into the expression vector pHUB2 [2]. The *E. coli* cells that were transformed with pHUB2 and contained the *Eco*RI-*Sal*I fragment gave a positive endoglucanase response; but, as with pPC71 clones, stable maintenance of both endoglucanase activity and the insert DNA was not possible.

The nucleotide sequence of the *Eco*RI-*Sal*I fragment from pUC19/71ES was determined from both *Sau*3A fragments and the entire insert which had been cloned into M13mp vectors. The DNA segment that was derived from *Pseudomonas* sp. NCIB 8634 was found to be 430 bp in length and flanked on the 5′ and 3′ ends by 375 and 271 bp of DNA that originate from pBR322 (Fig. 1). The DNA sequences of the pBR322 segments agreed precisely with the published data [22].

Three open reading frames (ORFs) were identified within the *Eco*RI-*Sal*I insert from pUC19/71ES. The first ORF (ORF1) codes for a protein of 136 amino acids with a predicted molecular mass of 14522 (Fig. 1). However, only the final 39 amino acids of this protein are encoded by the DNA from *Pseudomonas* sp. NCIB 8634.

ORF2 comprises 166 amino acids and would yield a protein with a molecular mass of 18714 (Fig. 1). If ORF2 is expressed, the amino acid sequence would comprise

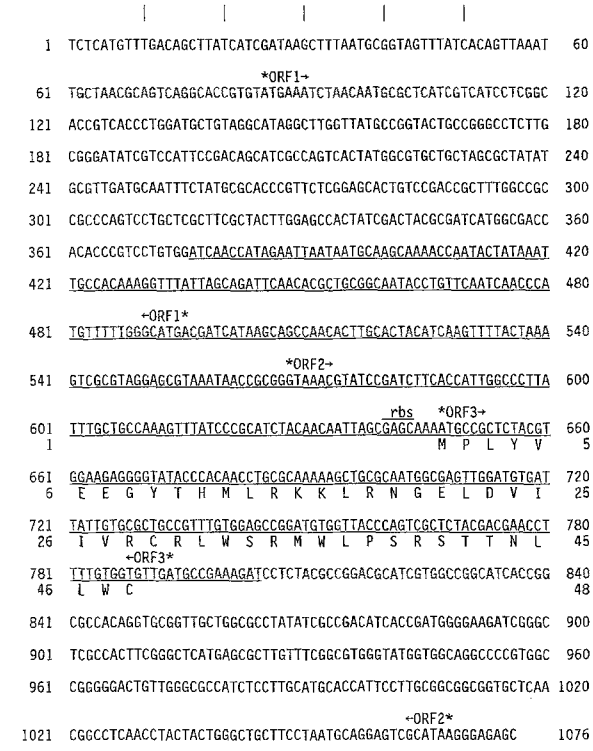

Fig. 1. Nucleotide sequence of the cloned *Eco*RI-*Sal*I fragment from pPC71. Nucleotides 1 to 375 and 805 to 1076 are derived from pBR322. The solid underlined segment denotes DNA from *Pseudomonas* sp. NCIB 8634. The predicted amino acid sequence, which is shown as a single-letter code, is for ORF3. A potential ribosome binding site for ORF3 is marked by a solid overline and rbs. The asterisks of ORF-labelled rightward arrows indicate the first nucleotide of ORF1, ORF2 and ORF3 and the asterisks of ORF-labelled leftward arrows indicate the sites of the last nucleotide of the stop codons of ORF1, ORF2 and ORF3.

79 residues from the N-terminus that would be encoded by DNA from *Pseudomonas* sp. NCIB 8634 while the remaining amino acid residues would be encoded by pBR322. The putative initiating codon for ORF2 is GUA which is not normally used for this role. The 5′ upstream region of ORF2 lacks sequences that resemble promoter regions from either *E. coli* or *Pseudomonas* and, as well, no potential ribosome binding site is evident. On this basis, it is unlikely that ORF2 would be translated in *E. coli*.

ORF3 is contained entirely within the cloned DNA from *Pseudomonas* sp. NCIB 8634 and encodes a protein of 48 amino acid residues (Fig. 1). The calculated molecular mass of this protein is 5845 Da. A potential ribosome binding site (GCGAG) lies four nucleotides from the initiating codon. No *E. coli*-like promoter regions are evident in the immediate upstream region although transcription could be initiated by the tetracycline resistance gene promoter from pBR322. Hydropathicity analysis revealed two hydrophilic regions that flank an internal hydrophobic domain (Fig. 2A). Searches of the entire EMBL and NBRF-PIR databases did not reveal any sequences with similarities to ORF3.

Digestion of plasmid pPFC4 with the exonuclease

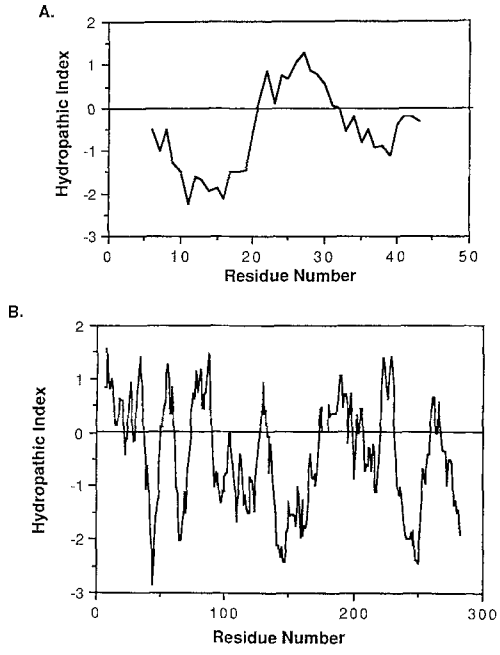

Fig. 2. Hydropathicity plots. (A) Hydropathicity index for ORF3 that occurs in cloned DNA from *Pseudomonas* sp. NCIB 8634. (B) Hydropathicity index for PFC-ORF that occurs in cloned DNA from *Pseudomonas fluorescens* subsp. *cellulosa*. The hydropathicity indices were determined by the method of Kyte and Doolittle [10] with an 11-residue window. Negative values denote regions of hydrophilicity on the Kyte-Doolittle aggregate scale. Hydrophobicity is represented by values greater than zero.

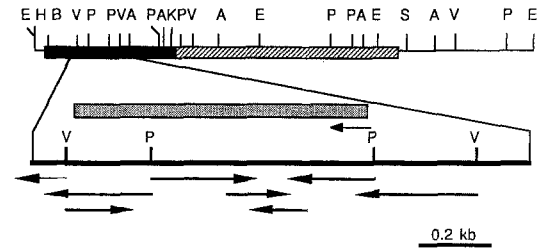

Fig. 3. Restriction endonuclease map and strategy for the sequencing of pPFC4. In the uppermost representation, the solid box denotes the DNA segment which after digestion with *Bal*31 retains endoglucanase activity, the hatched box represents that part of the original insert that can be removed by *Bal*31 digestion without loss of endoglucanase activity and the thin horizontal lines represent pBR322 DNA. A partial restriction endonuclease map is shown (A, *Ava*I; B, *Bam*HI; E, *Eco*RI; H, *Hind*III; K, *Kpn*I; P, *Pst*I; S, *Sal*I; V, *Pvu*II). The stippled box denotes the open reading frame, PFC-ORF, that occurs within the sequenced region. The thin arrow shows the direction of transcription. The thick arrows below the expanded segment of the original 10.6 kb insert represent the extent and direction of some of the sequence reactions. The kilobase scale measures only the expanded DNA segment that was sequenced and PFC-ORF (stippled box) and not the uppermost representation.

*Bal*31 after cleavage with restriction endonuclease *Sal*I yielded a fragment of about 4.6 kb from the 5′ end of the insert that expressed endoglucanase activity in *E. coli* (Fig. 3). DNA sequencing from the 5′ end of the 4.6 kb fragment revealed an open reading frame (PFC-ORF) which is 864 nucleotides in length (Fig. 4). Preceding the initiation codon, which is at nucleotide position 285, there is a potential ribosome binding site (GAGG) at nucleotide position 274. No promoter consensus sequences are evident in the immediate upstream vicinity (i.e., within 50 nucleotides) of the translation initiation site. The PFC-ORF terminates at nucleotide position 1146 with two successive amber codons.

The 288 amino acid residues encoded by PFC-ORF would yield a protein with a molecular mass of 32189. Using the weight-matrix method [24], a likely cleavage site for a leader peptidase is located at the alanine residue at amino acid position 38 of PFC-ORF (score = +4.95). Another possible cleavage site is at the alanine residue at position 12 (score = +2.25). Removal of a signal sequence at alanine-38 would result in a processed protein with a molecular mass of 28185 while if cleavage occurred at alanine-12 the product would have a molecular mass of 31100. Polyacrylamide gel electrophoresis of periplasmic extracts from *E. coli* carrying plasmid pPFC4, under non-denaturing conditions [8], yielded a value of 30000 ± 2000 Da (± S.D.; $n = 4$) for the major endoglucanase [26].

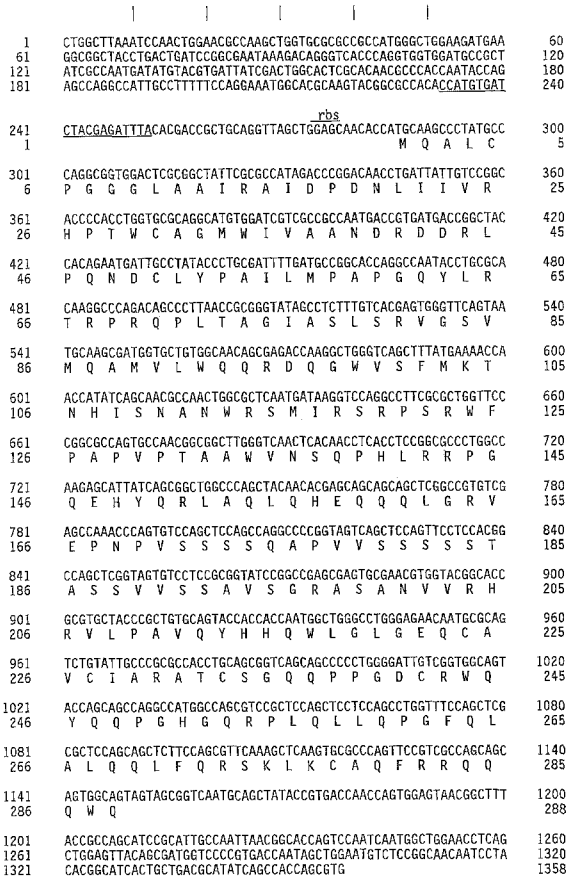

Fig. 4. Nucleotide sequence and predicted amino acid sequence (shown as a single-letter code) of PFC-ORF. A potential ribosome biding site (rbs) is marked at nucleotide positions 274–278. The underlined segment marks a potential catabolite gene activator protein binding site.

Of the 37 glutamine residues in PFC-ORF, 28 are confined to the region from amino acid sites 145 to 288. In the amino acid segment from sites 170 to 195, 15 of the 25 residues are serine and 6 are valine. Within this segment, two similar motifs, Val-Ser-Ser-Ser-Ser and Val-Val-Ser-Ser-Ser-Ser-Ser, occur. Hydropathy analysis predicts a protein with a moderately low hydrophobic nature (Fig. 2B). A computer search of the entire EMBL and NBRF-PIR databases did not uncover any sequences with significant similarity to PFC-ORF.

DISCUSSION

In the present study, a small DNA fragment from *Pseudomonas* sp. NCIB 8634 which encodes endoglucanase was sequenced. Of the three open reading frames (ORFs) that were found in pPC71, one (ORF3) is contained entirely within the *Pseudomonas* DNA fragment. The two other ORFs (ORF1 and ORF2) represent proteins that are derived from both pBR322 and *Pseudomonas* DNA sequences. As noted above, ORF2 is unlikely to form a protein in *E. coli*. If ORF1 produces endoglucanase activity, then the final 39 amino acids of the carboxy terminus which are encoded by *Pseudomonas* DNA are sufficient for this activity. This amino acid sequences has no distinctive features that would denote cellulolytic activity such as the egg-white lysozyme catalytic site [21,28]. The predicted molecular mass encoded by ORF3 is 5845 Da which, if this sequence expresses endoglucanase activity, would be, to date, the smallest cloned endoglucanase gene. Endoglucanases of this size have been reported to occur in *Cytophaga* [3] so that size per se does not necessarily preclude ORF3 from being an endoglucanase gene.

The originally cloned fragment form *P. fluorescens* subsp. *cellulosa* that expressed endoglucanase activity was 10.6 kb long. With *Bal*31 digestion, the size of the DNA fragment encoding endoglucanase activity was reduced to 4.6 kb. About 68% (i.e., 3.2 kb) of this segment was sequenced and found to contain a single open reading frame (PFC-ORF) that codes for a 32189 Da protein. This derived molecular mass corresponds closely to the molecular mass of the major endoglucanase that is produced in *E. coli* by pPFC4 as determined by polyacrylamide gel electrophoresis [26]. In addition, the major endoglucanase synthesized by *E. coli*/pPFC4 is found in the periplasm [26] which is consistent with the presence of a putative leader sequence in PFC-ORF.

There is additional evidence which is consistent with PFC-ORF being an endoglucanase gene. The motifs, $Val(Ser)_4$ and $(Val)_2(Ser)_5$, which are located within a span of 25 amino acids in PFC-ORF are very similar to domains that have been found in another carboxymethylcellulase (CMCase) gene from *P. fluorescens* subsp. *cellulosa* [6]. In this regard Knowles et al. [9] have noted that a Ser/Thr 'hinge' region may be a general feature common to a number of different cellulases.

Another similarity between PFC-ORF and the CMCase gene from *P. fluorescens* subsp. *cellulosa* is the presence of a putative catabolite gene activator protein (CAP) binding sequence that is upstream from the coding region of both of these DNA sequences. The CAP region for PFC-ORF lies at nucleotide position 231 (see Fig. 4). On the basis of the 14 bp consensus sequence from *E. coli* [4], the PFC-ORF CAP DNA site shows 9 matches out of 14. Discrepancies to this extent do occur among the CAP DNA binding sites of *E. coli* [4].

In summary, on the basis of commonality of unique amino acid sequence domains that are present in both PFC-ORF and an identified *Pseudomonas* CMCase gene [6] and the agreement between the calculated molecular mass of PFC-ORF and the observed value of the major

endoglucanase that is produced by *E. coli*/pPFC4 [25], it is inferred that PFC-ORF encodes an endoglucanase.

ACKNOWLEDGEMENT

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to B.R.G. and J.J.P.

REFERENCES

1 Berger, S.L., and A.R. Kimmel. 1987. Guide to Molecular Cloning. Methods Enzymol. 152: 1–812.
2 Bernard, H.-U., and D.R. Helinski. 1979. Use of the lambda phage promoter P_L to promote gene expression in hybrid plasmid cloning vectors. Methods Enzymol. 68: 483–492.
3 Chang, W.T.H., and D.W. Thayer. 1977. The cellulase system of a *Cytophaga* species. Can. J. Microbiol. 23: 1285–1292.
4 Ebright, R.H., P. Cossart, B. Gicquel-Sanzey, and J. Beckwith. 1984. Mutations that alter the DNA sequence specificity of the catabolite gene activator protein of *E. coli* Nature 311: 232–235.
5 Gilbert, H.J., G. Jenkins, D.A. Sullivan and J. Hall. 1987. Evidence for multiple carboxymethylcellulase genes in *Pseudomonas fluorescens* subsp. *cellulosa*. Mol. Gen. Genet. 210: 551–556.
6 Hall, J. and H.J. Gilbert. 1988. The nucleotide sequence of a carboxymethylcellulase gene from *Pseudomonas fluorescens* subsp. *cellulosa*. Mol. Gen. Genet. 213: 112–117.
7 Hawley, D.K., and W.R. McClure. 1983. Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nucleic Acids Res. 11: 2237–2255.
8 Hedrick, J.L., and A.J. Smith. 1968. Size and charge isomer separation and estimation of molecular weights of proteins by disk gel electrophoresis. Arch. Biochem. Biophys. 126: 155–164.
9 Knowles, J., P. Lehtovaara, and T. Teeri. 1987. Cellulase families and their genes. Trends Biotechnol. 5: 255–261.
10 Kyte, J., and R.F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157: 105–132.
11 Lamed, R., E. Setter and E.A. Bayer. 1983. Characterization of cellulose-binding, cellulase-containing complex in *Clostridium thermocellum*. J. Bacteriol. 156: 828–836.
12 Lamed, R., J. Naimark, E. Morgenstern and E.A. Bayer. 1987. Specialized cell surface structures in celluloytic bacteria. J. Bacteriol. 169: 3792–3800.
13 Lamed, R., and E.A. Bayer. 1988. The cellulosome of *Clostridium thermocellum*. In: Advances in Applied Microbiology, Vol. 33, (Laskin, A.I., ed.), pp. 1–46, Academic Press, San Diego.
14 Lejeune, A., C. Colson and D.E. Eveleigh. 1986. Cloning of an endoglucanase gene from *Pseudomonas fluorescens* var. *cellulosa* into *Escherichia coli* and *Pseudomonas fluorescens*. J. Indust. Microbiol. 1: 79–86.
15 Maniatis, T., E.F. Fritsch and J. Sambrook. 1982. Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratories, Cold Spring, New York.
16 Marck, C. 1986. Fast analysis of DNA and protein sequences on Apple IIe: restriction site search, alignment of short sequences and dot matrix analysis. Nucleic Acids Res. 14: 583–590.
17 Millet, J., D. Petre, P. Beguin, O. Raynaud and J.-P. Aubert. 1985. Cloning of ten distinct DNA fragments of *Clostridium thermocellum* coding for cellulases. FEMS Microbiol. Lett. 29: 145–149.
18 Pearson, W.R., and D.J. Lipman. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85: 2444–2448.
19 Ramasamy, K., and H. Verachtert. 1980. Localization of cellulase components in *Pseudomonas* sp. isolated from activated sludge. J. Gen. Microbiol. 117: 181–191.
20 Sanger, F., S. Nicklen and A.R. Coulson. 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. 74: 5463–5467.
21 Schwarz, W.H., K. Schimming, P. Rucknagel, S. Burgschwarger, G. Kreil and W.L. Staudenbauer. 1988. Nucleotide sequence of the *celC* gene encoding endoglucanase C of *Clostridium thermocellum*. Gen 63: 23–30.
22 Sutcliffe, J.G. 1979. Complete nucleotide sequence of the *Escherichia coli* plasmid pBR322. Cold Spring Harbor Symp. Quant. Biol. 43: 77–90.
23 Teather, R.M., and P.J. Wood. 1982. Use of congo red-polysaccharide interaction in the enumeration and characterization of cellulolytic bacteria of the bovine rumen. Appl. Environ. Microbiol. 43: 777–780.
24 von Heijne, G. 1986. A new method for predicting signal sequence cleavage sites. Nucleic Acids Res. 14: 4683–4690.
25 Wolff, B.R., T.A. Mudry, B.R. Glick and J.J. Pasternak. 1986. Isolation of endoglucanase genes from *Pseudomonas fluorescens* subsp. *cellulosa* and a *Pseudomonas* sp. Appl. Environ. Microbiol. 51: 1367–1369.
26 Wolff, B.R., D. Lewis, J.J. Pasternak and B.R. Glick, 1990. Partial characterization of *Pseudomonas fluorescens* subsp. *cellulosa* endoglucanase activity produced in *Escherichia coli*. J. Ind. Microbiol. 5: 59–64.
27 Wood, T.M., C.A. Wilson and C.S. Stewart. 1982. Preparation of cellulase from the celluloytic anaerobic rumen bacteria *Ruminococcus albus* and its release from the bacterial cell wall. Biochem. J. 205: 129–137.
28 Yaguchi, R.J., C. Roy, C.F. Rollin, M.G. Paice and L. Jurasek. 1983. A fungal cellulase shows sequence homology with the active site of hen egg-white lysozyme. Biochem. Biophys. Res. Commun. 116: 408–411.
29 Yamane, K., H. Suzuki and K. Nisizawa. 1970. Purification and properties of extracellular and cell-bound cellulase components of *Pseudomonas fluorescens* var. *cellulosa* J. Biochem. 67: 19–35.
30 Yannish-Perron, C., J. Vierra and J. Messing. 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13p18 and pUC19 vectors. Gene 33: 103–119.
31 Yoshikawa, T., H. Suzuki and F. Nisizawa. 1974. Biogenesis of multiple cellulase components of *Pseudomonas fluorescens* var. *cellulosa*. 1. Effects of culture conditions on the multiplicity of cellulase. J. Biochem. 75: 531–540.