APPLIED AND ENVIRONMENTAL MICROBIOLOGY, May 1988, p. 1289–1292 Vol. 54, No. 5
0099-2240/88/051289-04$02.00/0

Structure of an Endo-β-1,4-Glucanase Gene from *Clostridium acetobutylicum* P262 Showing Homology with Endoglucanase Genes from *Bacillus* spp.

HAROLD ZAPPE, WINSOME A. JONES, DAVID T. JONES, AND DAVID R. WOODS*

Department of Microbiology, University of Cape Town, Rondebosch 7700, South Africa

Received 17 August 1987/Accepted 9 February 1988

The nucleotide sequence of an endo-β-1,4-glucanase gene of *Clostridium acetobutylicum* contained two putative extended promoter consensus sequences, a Shine-Dalgarno sequence and a TTG initiation codon. The nucleotide sequence of the gene coding for the C-terminal region of this enzyme was not required for activity. Extensive homology in the nucleotide and amino acid sequences of the endoglucanase genes from *C. acetobutylicum* and *Bacillus* spp. was demonstrated.

Clostridium acetobutylicum strains, which have been used for the production of acetone and butanol (9), produce endo-β-1,4-glucanases (1, 11) and xylanases (10, 12, 13). Previously, we described the cloning of an endoglucanase gene and a xylanase gene from *C. acetobutylicum* P262 and their expression in *Escherichia coli* (23, 24); here, we report the nucleotide sequence of the endoglucanase gene.

The DNA sequences of both strands of plasmids containing the cloned endoglucanase gene were determined by the dideoxynucleotide chain-termination method of Sanger et al. (19) with templates prepared from DNA subcloned in M13 or pUC vectors (22). Computer analysis of DNA and protein sequences was performed by using MicroGenie (Beckman). The DNA sequence contained an open reading frame encoding 448 amino acids with a calculated M_r of 49,354 (Fig. 1).

The nucleotide sequence of the *C. acetobutylicum* endoglucanase gene did not contain an in-frame ATG start codon. The putative ribosome-binding site resembled that reported for the β-lactamase gene from *Staphylococcus aureus* (15) and consisted of a TTG initiation codon and a strong Shine-Dalgarno complementarity containing 5 G-C base pairs situated 8 base pairs upstream of the initiation codon.

The G+C content of the nucleotides within the open reading frame (32.8%) was higher than that within the upstream regulatory region (19.9%). The average G+C ratio for the genome of *C. acetobutylicum* is approximately 28% (3). Codon usage was strongly biased towards codons in which A and U predominated.

The cloned endoglucanase gene from *C. acetobutylicum* was expressed from its own promoter in *E. coli* (23). The region upstream of the open reading frame, between nucleotides −221 and −175, contained a putative promoter sequence which consisted of a TTGTATT −35 region and a TACAAT −10 region separated by 16 nucleotides (Fig. 1). This promoter consensus sequence resembled the extended consensus sequence of the σ^{43} RNA polymerase recognition site of *Bacillus subtilis* and of other gram-positive bacteria (6). A second putative extended promoter consensus sequence was situated immediately upstream of the ribosome-binding site between nucleotides −66 and −19 (Fig. 1).

The nucleotide sequences of the genes encoding endoglucanases from a number of cellulolytic organisms, including *Clostridium thermocellum* (2, 7, 8), *Cellulomonas fimi* (21), *Trichoderma reesei* (17), *B. subtilis* (14, 16, 18), and three alkalophilic *Bacillus* spp. (4, 5), have been reported. The deduced amino acid sequence of the *C. acetobutylicum* endoglucanase gene showed only 11.9, 14.4, and 15.3% homology with those of the *celA*, *celB*, and *celD* endoglucanase genes of *Clostridium thermocellum*, respectively (2, 7, 8); 13.7% homology with the *cenA* endoglucanase gene of *Cellulomonas fimi* (21); and 14.5% homology with the endoglucanase I gene of *T. reesei* (17). There was also no discernable homology (13.8%) with the amino acid sequence of the endo-β-(1,3)-(1,4)-glucanase gene from *B. subtilis* C120 (16). However, comparison of the *C. acetobutylicum* endoglucanase gene amino acid sequence with the sequences of the genes coding for the two enzymes from the alkalophilic *B. subtilis* strain N-4 (5) and with those of the genes coding for the enzymes from *B. subtilis* strains PAP115 (14) and DLG (18) (which exhibit 93.4% homology) showed 43.9, 48.4, 43.4, and 44.6% homology, respectively. The gene coding for a larger endo-β-1,4-glucanase from another alkalophilic *Bacillus* sp. strain, 1139 (4), exhibited less overall homology (23.8%). The nucleotide sequence of the *C. acetobutylicum* endoglucanase gene showed 49% homology with the nucleotide sequences of the corresponding genes from *B. subtilis* PAP115 and DLG and 57% homology with the nucleotide sequence of the gene from the alkalophilic strain N-4.

Analysis of the aligned amino acid sequences of the endoglucanase genes from *C. acetobutylicum* and the *Bacillus* strains showed a region of homology which extended for approximately 350 residues from the N-terminal end (Fig. 2). In comparison with the four *Bacillus* endoglucanase genes, this region of the gene coding for the *C. acetobutylicum* enzyme showed approximately 61 to 65% nucleotide homology and 60% amino acid sequence homology. If amino acid replacement by conserved amino acids is taken into account, the overall homology of this region is increased to 80%. Although the amino acid homology of the entire endoglucanase gene from the alkalophilic *Bacillus* sp. strain, 1139, was only 23%, the amino acid homology of the N-terminal region was 40%.

The nonhomologous C-terminal regions of the endoglucanase genes from the different species varied in length from approximately 75 to 150 amino acids, and the junction

* Corresponding author.

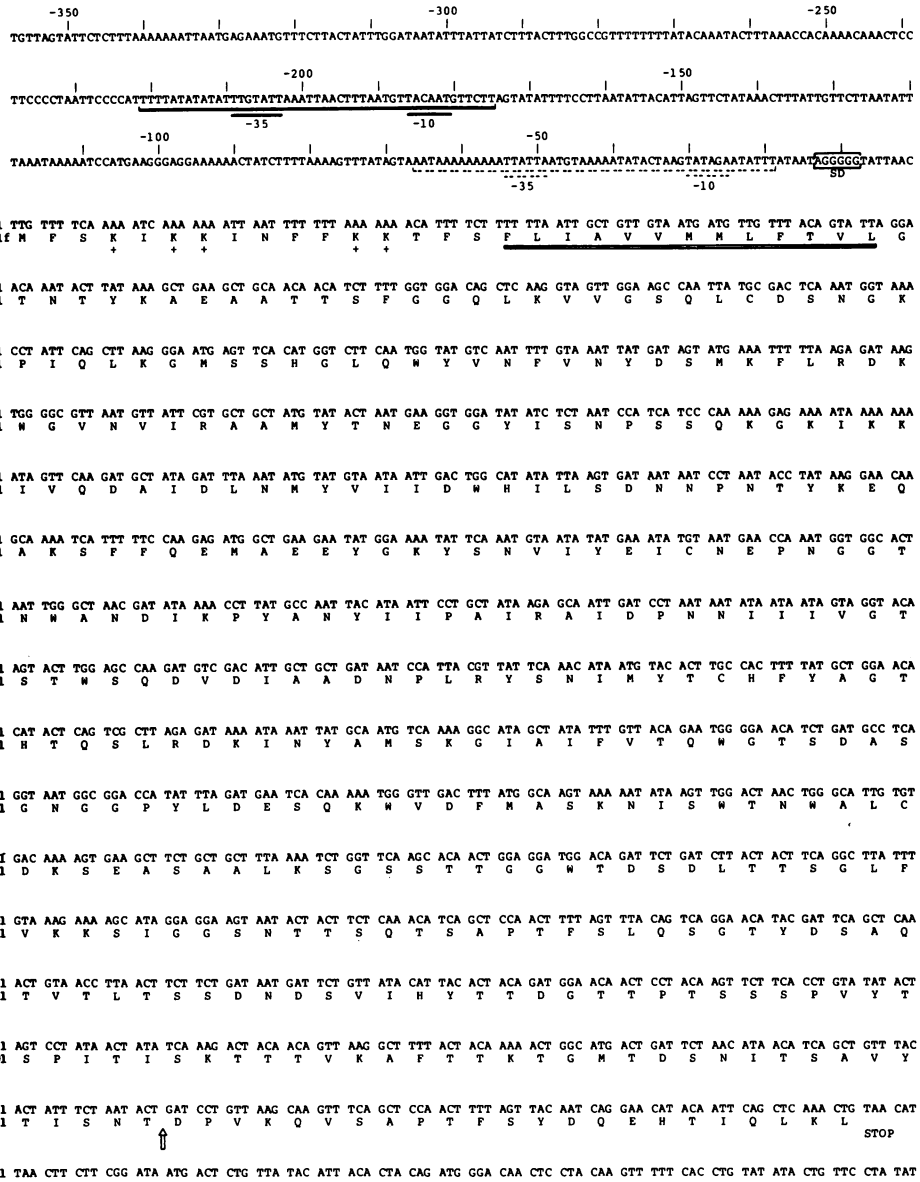

FIG. 1. Complete nucleotide sequence of the endo-β-1,4-glucanase gene and flanking regions from *C. acetobutylicum*. The derived amino acid sequence is given in the one-letter code from position 1 to 1,347 (448 residues). The putative promoter consensus sequence (from position −221 to −175) is underlined; the −35 and −10 regions are underlined twice. A second putative promoter sequence immediately upstream of the ribosome-binding site is underlined with a broken line. The Shine-Dalgarno (SD) sequence AGGGGG is boxed. The positively charged amino acids at the start of the putative signal sequence are indicated by plus signs, and the hydrophobic region is underlined twice. The extent of the truncated open reading frame (position 1,275) cloned in the recombinant plasmid pHZ100 is indicated with an arrow.

between the homologous and nonhomologous regions was characterized by an area rich in hydroxyl amino acids and proline (Fig. 2). Approximately 40% of the amino acids in the nonhomologous C-terminal region of the *C. acetobutylicum* endoglucanase gene were threonine or serine.

The C-terminal region was not essential for the function of the *B. subtilis* PAP115 endoglucanase gene, from which it was possible to delete approximately one-third of the C-terminal amino acids (position 343, Fig. 2) without loss of enzyme activity (14). The C-terminal region of the *C. acetobutylicum* gene is also not required for activity since an endoglucanase derivative lacking 24 C-terminal amino acids (position 426, Fig. 2) was functional.

Mackay et al. (14) identified a short region between amino acids 82 and 141 which showed some similarity to the amino acid sequences of the endo-β-1,4-glucanase *celB* gene of *Clostridium thermocellum* and suggested that this region contains some of the residues of the active site of the enzyme gene. Comparison with the aligned amino acid sequences of the endoglucanase genes from *C. acetobutylicum* and the *Bacillus* spp. confirmed the presence of two short regions situated between positions 85 to 97 and 115 to 143 which showed some homology with the amino acid sequence of the endoglucanase *celB* gene of *Clostridium thermocellum*.

A putative signal sequence consisting of a short stretch of five positively charged amino acids followed by a core of five

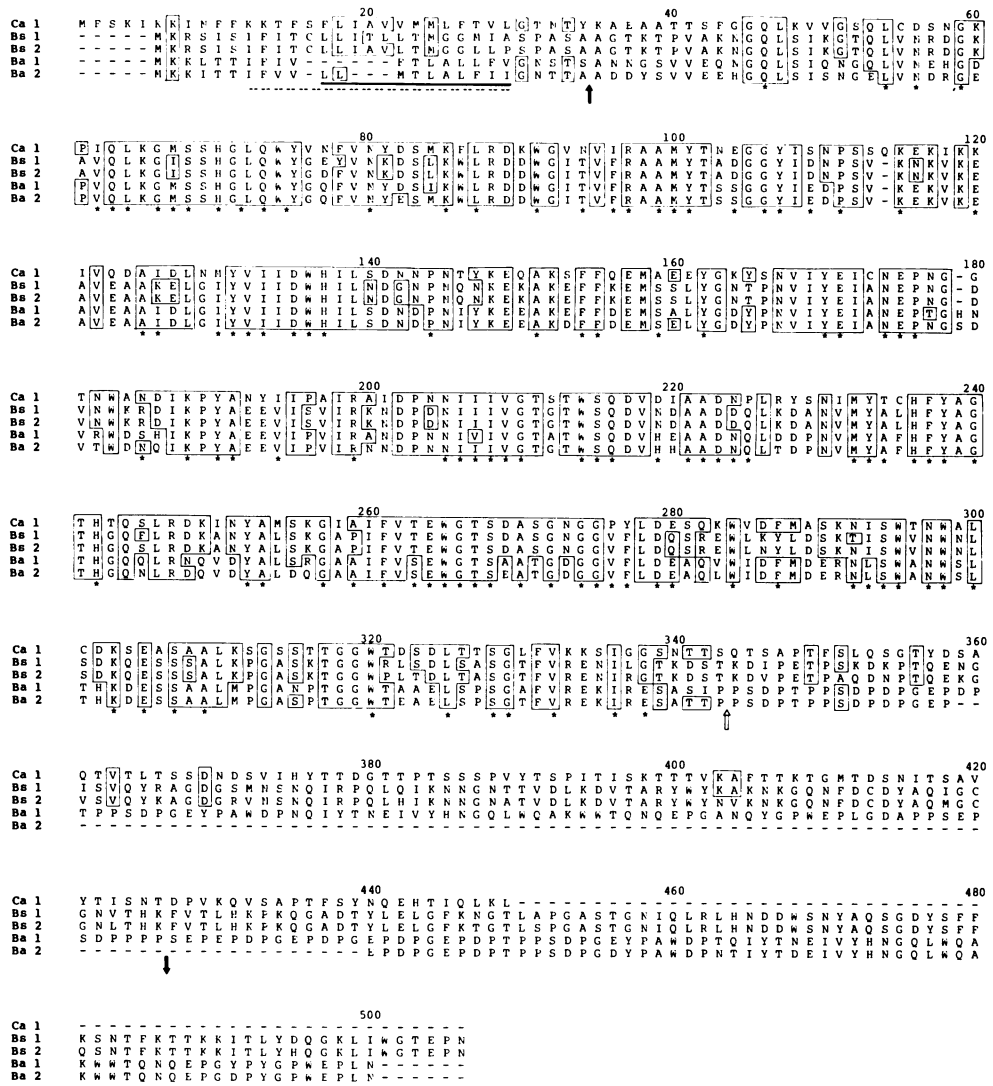

FIG. 2. Amino acid sequence alignments of the endo-β-1,4-glucanase genes from *C. acetobutylicum* P262 (Ca 1), *B. subtilis* PAP115 (Bs 1) (14), *B. subtilis* DLG (Bs 2) (18), alkalophilic *B. subtilis* N-4(pNK1) (Ba 1) and *B. subtilis* N-4(pNK2) (Ba 2) (5). The amino acids are identified by the single-letter code, and regions of identical homology are boxed. Amino acid residues which show homology with those of the endo-β-1,4-glucanase gene from the alkalophilic *Bacillus* sp. strain 1139 are indicated with asterisks. The hydrophilic region of the signal sequence of the endoglucanase gene from the *Bacillus* sp. is indicated by a broken line, and the region from *C. acetobutylicum* is indicated by an unbroken line. The signal sequence cleavage site identified in the endoglucanase genes from *B. subtilis* PAP115 and DLG is indicated by an arrow at position 34. The arrow at position 343 indicates the point to which BAL 31 digestion removed the C-terminal region of the endoglucanase gene from *B. subtilis* PAP115 such that the strain retained the ability to direct the synthesis of active extracellular endoglucanase. The arrow at position 426 indicates the extent of the truncated active endoglucanase gene from *C. acetobutylicum* cloned in the recombinant plasmid pHZ100.

hydrophobic amino acids, resembling signal sequences of gram-positive bacteria (14), was identified at the N-terminal region of the *C. acetobutylicum* endoglucanase gene (Fig. 2).

Previously, no close homology had been demonstrated between endoglucanase genes from different bacterial genera. Since there appeared to be little or no conservation of nucleotide or amino acid sequences between endoglucanase genes from different species and much greater conservation between endoglucanase genes from the same or closely related strains, it was postulated that the genes coding for the cellulolytic enzymes of the various microorganisms evolved independently (8). The occurrence of a conserved nucleotide and amino acid sequence in the endoglucanase genes from *C. acetobutylicum*, *B. subtilis*, and the alkalophilic bacilli indicated that, at least with the genes coding for these enzymes, this is not the case. The limited homology between the *C. acetobutylicum* endoglucanase gene and the genes coding for the three endoglucanases from *Clostridium thermocellum*, as well as the strong homology of the *C. acetobutylicum* endoglucanase gene with genes coding for enzymes from the *Bacillus* group, is interesting and tends to confirm the results of recent phytogenetic studies (20). These studies indicate a separation of the low-G+C-content gram-positive bacteria into a number of branches which delineate the thermophilic clostridia from the mesophilic clostridia and the bacilli (20). The endoglucanase genes showing strong

homology all originate from noncellulolytic species which produce only one or a few endoglucanases, whereas the endoglucanase genes which show little homology originate from cellulolytic species which produce a complex battery of cellulolytic enzymes.

This work was supported by grants to D.R.W. from Sentrachem Limited, Johannesburg, South Africa, and by the Foundation for Research and Development.

LITERATURE CITED

1. **Allcock, E. R., and D. R. Woods.** 1981. Carboxymethyl cellulase and cellobiase production by *Clostridium acetobutylicum* in an industrial fermentation medium. Appl. Environ. Microbiol. **41:** 539–541.
2. **Béguin, P., P. Cornet, and J.-P. Aubert.** 1985. Sequence of a cellulase gene of the thermophilic bacterium *Clostridium thermocellum*. J. Bacteriol. **162:**102–105.
3. **Cummins, C. S., and J. L. Johnson.** 1971. Taxonomy of the clostridia: wall composition and DNA homologies in *Clostridium butyricum* and other acid-producing clostridia. J. Gen. Microbiol. **67:**33–46.
4. **Fukumori, F., T. Kudo, Y. Narahashi, and K. Horikoshi.** 1986. Molecular cloning and nucleotide sequence of the alkaline cellulase gene from the alkalophilic *Bacillus* sp. strain 1139. J. Gen. Microbiol. **132:**2329–2335.
5. **Fukumori, F., N. Sashihara, T. Kudo, and K. Horikoshi.** 1986. Nucleotide sequences of two cellulase genes from alkalophilic *Bacillus* sp. strain N-4 and their strong homology. J. Bacteriol. **168:**479–485.
6. **Graves, M. C., and J. C. Rabinowitz.** 1986. *In vivo* and *in vitro* transcription of the *Clostridium pasteurianum* ferredoxin gene. J. Biol. Chem. **261:**11409–11415.
7. **Grepinet, O., and P. Beguin.** 1986. Sequence of the cellulase gene of *Clostridium thermocellum* coding for endoglucanase B. Nucleic Acids Res. **14:**1791–1799.
8. **Joliff, G. J., and J.-P. Aubert.** 1986. Nucleotide sequence of the cellulase gene *celD* encoding endoglucanase D of *Clostridium thermocellum*. Nucleic Acids Res. **14:**8605–8613.
9. **Jones, D. T., and D. R. Woods.** 1986. Acetone-butanol fermentation revisited. Microbiol. Rev. **50:**484–524.
10. **Lee, S. F., and C. W. Forsberg.** 1987. Isolation and some properties of a β-D-xylosidase from *Clostridium acetobutylicum* ATCC 824. Appl. Environ. Microbiol. **53:**651–654.
11. **Lee, S. F., C. W. Forsberg, and L. N. Gibbins.** 1985. Cellulolytic activity of *Clostridium acetobutylicum*. Appl. Environ. Microbiol. **50:**220–228.
12. **Lee, S. F., C. W. Forsberg, and L. N. Gibbins.** 1985. Xylanolytic activity of *Clostridium acetobutylicum*. Appl. Environ. Microbiol. **50:**1068–1076.
13. **Lee, S. F., C. W. Forsberg, and J. B. Rattray.** 1987. Purification and characterization of two endoxylanases from *Clostridium acetobutylicum* ATCC 824. Appl. Environ. Microbiol. **53:**644–650.
14. **Mackay, R. M., A. Lo, G. Willick, M. Zuker, S. Baird, M. Dove, F. Moranelli, and V. Seligy.** 1986. Structure of a *Bacillus subtilis* endo-β-1,4-glucanase gene. Nucleic Acids Res. **14:**9159–9170.
15. **McLaughlin, J. R., C. L . Murray, and J. C. Rabinowitz.** 1981. Unique features in the ribosome binding site sequence of the gram-positive *Staphylococcus aureus* β-lactamase gene. J. Biol. Chem. **256:**11283–11291.
16. **Murphy, N., D. J. McConnell, and B. A. Cantwell.** 1984. The DNA sequence of the gene and genetic control sites for the excreted *B. subtilis* enzyme β-glucanase. Nucleic Acids Res. **12:**5355–5367.
17. **Penttila, M., P. Lehtovaara, H. Nevalainen, R. Bhikhabhai, and J. Knowles.** 1986. Homology between cellulase genes of *Trichoderma reesei*: complete nucleotide sequence of the endoglucanase I gene. Gene **45:**253–263.
18. **Robson, L. M., and G. H. Chambliss.** 1987. Endo-β-1,4-glucanase gene of *Bacillus subtilis* DLG. J. Bacteriol. **169:**2017–2025.
19. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
20. **Woese, C. R.** 1987. Bacterial evolution. Microbiol. Rev. **51:** 221–271.
21. **Wong, W. K., B. Gerhard, Z. M. Guo, D. G. Kilburn, R. Anthony, J. Warren, and R. C. Miller, Jr.** 1986. Characterization and structure of an endoglucanase gene *cenA* of *Cellulomonas fimi*. Gene **44:**315–324.
22. **Yanisch-Perron, C., J. Vieira, and J. Messing.** 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene **33:**103–119.
23. **Zappe, H., D. T. Jones, and D. R. Woods.** 1986. Cloning and expression of *Clostridium acetobutylicum* endoglucanase, cellobiase and amino acid biosynthesis genes in *Escherichia coli*. J. Gen. Microbiol. **132:**1367–1372.
24. **Zappe, H., D. T. Jones, and D. R. Woods.** 1987. Cloning and expression of a xylanase gene from *Clostridium acetobutylicum* P262 in *Escherichia coli*. Appl. Microbiol. Biotechnol. **27:**57–63.