Transcription of Clostridium thermocellum Endoglucanase Genes celF and celD

SAROJ MISHRA,† PIERRE BÉGUIN,* AND JEAN-PAUL AUBERT

Unité de Physiologie Cellulaire et URA 1300 CNRS, Département des Biotechnologies, Institut Pasteur, 25, rue du Dr Roux, 75724 Paris Cedex 15, France

Received 1 June 1990/Accepted 8 October 1990

Transcripts of the Clostridium thermocellum endoglucanase genes celF and celD, encoding endoglucanases F and D, respectively, were characterized. The size of the mRNAs was about 2.35 kb for celF and 2.1 kb for celD, indicating monocistronic transcription of both genes. A unique 5' end, located 218 bp upstream from the initiation codon, was found for celF mRNA. No convincing homology could be identified between the sequence upstream from the celF 5' end and other procarboxyl promoters. Two 5' ends, located 124 and 294 bp upstream from the initiation codon, were mapped for celD mRNA. The −10 and the −35 sequences preceding the ATG-distal 5' end of celD mRNA were homologous to the consensus sequence of Bacillus subtilis σ28 promoters. The sequence upstream from the ATG-proximal 5' end had some similarity with the −10 sequence of B. subtilis σ28 promoters. During growth on cellobiose, the 5' end of celD transcripts was found predominantly at the −124 site during the late exponential phase but almost exclusively at the −294 site during the early stationary phase. The kinetics of appearance of celA, celC, celD, and celF mRNA was followed by dot blot analysis. Transcripts of celA, celD, and celF were detected during late exponential and early stationary phase. In contrast, the celC transcript was detected almost exclusively during early stationary phase. Since growth was limited by the availability of cellobiose, the results suggest that the genes are regulated by a mechanism analogous to catabolite repression.

The gram-positive thermophilic anaerobe Clostridium thermocellum produces a highly potent, thermostable cellulase complex organized in multimolecular structures termed cellulosomes, which contain at least 14 different polypeptides (23).

Cloning of chromosomal DNA fragments in Escherichia coli led to the isolation of 15 endoglucanase genes, two xylanase genes, and two β-glucosidase genes (11, 19, 26, 34). The genes celA (2), celB (15), celC (36), celD (21), celE (18), celF (unpublished results), and celH (41), encoding endoglucanases A, B, C, D, E, F, and H, respectively, as well as the genes xynZ (17) and bigB (12), encoding xylanase Z and β-glucosidase B, respectively, have been sequenced. Endoglucanases A (35), B (3), C (29), D (22), and E (18) and xylanase Z (16) have been purified and characterized from recombinant E. coli clones or, in the case of endoglucanase A, from C. thermocellum culture supernatant (30). However, little is known about the regulation of cellulase production in C. thermocellum or about the pattern of transcription of individual cellulase degradation genes. No catabolite repression by cellobiose or glucose was observed for the synthesis of endoglucanase activity (8, 37). By contrast, the synthesis of activity for crystalline cellulose (Avicelase) is repressed by readily metabolized carbon sources present in nonlimiting concentrations (20).

Regarding the transcription of individual cel genes, celA is the only C. thermocellum endoglucanase gene whose mRNA has been characterized (4). The celA gene is transcribed as a monocistronic unit. In C. thermocellum, transcription is initiated at two different sites. The major start site maps 57 bp upstream from the initiation codon and is preceded by −10 and −35 sequences homologous to Bacillus subtilis σ28 promoters. A minor start site is located 134 bases upstream from the initiation codon, with preceding promoter sequences analogous to B. subtilis σ28 and E. coli σ70 consensus sequences. The celA transcript isolated from recombinant E. coli starts exclusively at the −134 site.

The aim of the present work was to characterize the lengths and the start sites of the transcripts encoded by the celF and celD genes in C. thermocellum. The abundance of transcripts starting at either of the two identified celD 5' termini was studied as a function of the growth phase. The kinetics of appearance of the transcripts during growth on cellobiose was analyzed by dot blot hybridization for the four endoglucanase genes celA, celC, celD, and celF.

MATERIALS AND METHODS

Bacterial strain. The C. thermocellum strain used was NCIB 10682.

Media and growth conditions. For initial transcript mapping, cells were cultivated at 60°C in shaken 500-ml anaerobic bottles containing 250 ml of CM3-3 medium (30) with 0.5% (wt/vol) cellobiose as a carbon source. Culture (150 ml) was harvested anaerobically by centrifugation during the late exponential phase of growth. Cells were resuspended in 3 ml (1/50 of the original volume) of 100 mM Tris hydrochloride–1 mM EDTA (TE), pH 8.0, and RNA was extracted as described below. For the analysis of the time course of transcription, cells growing exponentially in CM3-3 medium were transferred into a fermentor (Biolaffite-LSL, St. Germain-en-Laye, France) containing 1.2 liters of CM3-3 medium with 0.5% cellobiose. The initial OD600 of the culture was 0.3. Cells were grown at 60°C with the pH regulated to 7.2 (38). Samples (150 ml) of the culture were collected at the time points indicated in Fig. 6. Cells were harvested, concentrated as described above, and stored in liquid nitrogen.
FIG. 1. Maps of celA, celC, celD, and celF and positions of the probes used for Northern and dot-blot hybridizations and for S1 nuclease mapping. The coding sequence of the genes is indicated by a rectangle. The orientation of transcription is shown by a straight arrow. Intragenic fragments used as probes for Northern and dot-blot hybridizations are indicated by a shaded bar. Single-stranded probes used for S1 nuclease mapping are shown by wavy arrows.

RNA extraction. The concentrated cell suspension in TE was made 0.9% sodium dodecyl sulfate and 0.7% Macaloid, and cells were disrupted with glass beads in the presence of phenol in a Mickie shaker. Isolation of RNA was performed by the method of Glatron and Rapoport (10). The final RNA pellet was dissolved in water and stored at -70°C.

Dot-blot and Northern (RNA blot) hybridization. Dot-blot and Northern hybridizations of mRNA were performed according to Thomas (39). For Northern blotting, RNA (up to 30 µg) was denatured in the presence of glyoxal and dimethyl sulfoxide, electrophoresed on 1.5% agarose gels, and blotted onto nitrocellulose. For celA, celC, celD, and celF probes, intragenic fragments ranging between 870 and 1,070 bp were derived from pCT105 (5), pCT301 (19), pCT600 (19), and pCT500 (19), respectively, using either the original clones or subclones that had been constructed for sequencing (Fig. 1). Labeling was performed according to Feinberg and Vogelstein (7), using a kit for random primer labeling (Boehringer, Mannheim, Federal Republic of Germany). The specific activity of the probes determined after filtration on Sephadex G-50 was 1 × 10⁶ to 1.5 × 10⁶ cpm/ng of DNA. Denaturated PstI fragments of λ DNA were used as molecular weight markers and detected by hybridization with labeled λ DNA.

Mapping of 5' ends by primer extension. Primer extension with reverse transcriptase was performed with synthetic oligonucleotides (26- and 28-mers) complementary to celF and celD (See Fig. 4 and 5). Preliminary experiments with primers hybridizing to the coding sequences of the genes demonstrated the absence of start sites located downstream from those detected with the primers shown in Fig. 4 and 5. RNA (50 µg) and 1 to 5 ng of the 5'-32P-labeled oligonucleotide probe (approximately 80,000 cpm) were heated for 10 min at 75°C and hybridized overnight at 42°C (celF) or 30°C (celD) in 30 µl of buffer (20 mM N-2-hydroxyethylpipera- zine-N'-2-ethanesulfonic acid [HEPES], pH 6.5, 80% deionized formamide, 0.4 M NaCl). The nucleic acids were precipitated and washed with 70% ethanol. Extension with reverse transcriptase (Boehringer) was performed as described before (1). The products of primer extension were analyzed on a 6% acrylamide–8 M urea sequencing gel along with sequencing ladders (33) derived from the same oligonucleotides.

S1 nuclease mapping. S1 nuclease mapping was performed (1). To prepare 5'-end-labeled single-stranded probes, the 5'-32P-labeled oligonucleotides (see Fig. 4 and 5) were hybridized to the single-stranded 5'-flanking sequence of celF, cloned in pTZ19R, and 5'-flanking sequence of celD, cloned in M13mp8. The oligonucleotides were extended with the Klenow fragment of DNA polymerase, and defined 3' ends were generated by digestion with AccI (for celF) and SspI (for celD) (1). The single-stranded probes (Fig. 1) were isolated either from a 0.7% low-melting-point agarose gel (celF) or from a 6% acrylamide–8 M urea sequencing gel (celD). Total RNA (50 µg) was mixed with 10 to 50 ng (approximately 50,000 cpm) of labeled probe, dissolved in 30 µl of the hybridization buffer described above, heated for 10 min at 75°C, and incubated overnight at 38°C (celF) or 36°C (celD). Nonhybridized DNA was digested with 300 to 500 U of S1 nuclease (Boehringer) for 1 h at 37°C. Nucleic acids were extracted once with phenol, ethanol precipitated, and analyzed on a 6% sequencing gel as described above.

RESULTS

Lengths of the celF and celD transcripts. Intragene probes derived from celF and celD hybridized with RNA species of about 2,350 and 2,100 nucleotides, respectively (Fig. 2, lanes 2 and 3). Thus, the sizes of the celF and celD transcripts closely matched the lengths of the corresponding structural genes, i.e., 2.2 kb for celF (unpublished data) and 1.95 kb for celD (21). Transcription of the genes is therefore monocis- tronic, as observed for other cellulase genes, such as celA of C. thermocellum (4), the endoglucanase genes from B. subtilis DLG (31), C. fimii (13, 14, 28), Erwinia chrysanthemi (17a), and Thermomonospora fusca (25), and the exoglucanase gene of C. fimii (13).

The 3' end of celD (21) is located in the vicinity of palindromic structures capable of forming hairpin loops analogous to E. coli rho-independent termination signals of transcription (32). No such structure was found at the 3' end of celF (unpublished data). The same feature has been described for E. coli rho-dependent terminators (27).

5'-end mapping of the celF and celD transcripts. Due to "breathing" of RNA-DNA hybrids, a set of closely spaced
bands was often observed with S1 nuclease mapping, suggesting that the nuclease was nibbling into the hybrid. In this respect, closely spaced multiple start sites reported for other genes (13, 14) on the basis of S1 nuclease mapping could be artifactual and not due to flexibility of RNA polymerase in selecting a start site, as suggested (14). Primer extension with reverse transcriptase was more accurate, but S1 nuclease was nonetheless useful in discriminating artifacts of mapping by primer extension with reverse transcriptase due to pausing or formation of hairpin loops.

The 5' end of the celF transcript was found by primer extension 218 nucleotides upstream from the putative initiation codon (TTG), with C as the first transcribed nucleotide (Fig. 3A, lane P, and Fig. 4). Minor bands corresponding to higher-molecular-weight primer extension products were also observed, but only the −218 start site was confirmed by S1 nuclease mapping (data not shown). The DNA sequences upstream of the celF start site did not contain convincing homologies with other described consensus promoters, such as those recognized by B. subtilis σ^73, σ^74, σ^75, or σ^78 by E. coli σ^70, σ^72, or σ^74 (mRNA) (6), or with other mapped cellulase promoter sequences, such as C. thermocellum celA (4), C. fimi ces, cenA, cenB, or cenC (13, 14, 28), and T. fusca celE (25).

For celD, primer extension revealed two start sites located 124 and 294 bp upstream from the ATG start codon (Fig. 3B and 5). Both sites were confirmed by S1 nuclease mapping (data not shown). Transcripts originating from the −124 site were most abundant during the late exponential growth phase (Fig. 3B, lane t₂), whereas most of the mRNA isolated during the early stationary phase started at the −294 site (Fig. 3B, lane t₄). No signal could be detected with RNA isolated either during the early exponential phase or during the late stationary phase (Fig. 3B, lanes t₁ and t₃).

The −10 and −35 regions preceding the −294 start site (pD2 promoter) were clearly similar to the consensus established for B. subtilis σ^75 and E. coli σ^70 promoters (6) (Fig. 5). A sequence showing homology to the −10 region of B. subtilis σ^78 promoters was tentatively identified upstream from the −124 site (pD1 promoter), but no corresponding −35 region could be found. A proximal promoter resembling B. subtilis σ^28 promoters and a distal promoter similar to B. subtilis σ^75 promoters were also described for the celA gene (4), and expression of the two genes may be subject to similar regulation.

Kinetics of appearance of celA, celC, celD, and celF transcripts. In order to check whether other endoglucanase genes follow the same time course of expression as celD, the levels of mRNA expressed from celA, celC, celD, and celF were assayed by dot-blot hybridization at different stages of growth. As shown in Fig. 6, the celA, celD, and celF transcripts were detected during the late exponential and the early stationary phase of growth. Transcripts of celA were more abundant in exponentially growing cells than in stationary-phase cells. In contrast, celC transcripts were detected mostly during the early stationary phase. For all genes studied, no mRNA was detected during the early exponential or the late stationary phase. Transfer of cellulose-grown cells producing high levels of the celA- and celF-specific transcripts into fresh cellulose medium resulted in suppression within 3 h of the celA and celF transcripts (unpublished data).

By comparing the intensities of the RNA spots with those of spotted DNA standards, it was estimated that the celA mRNA levels were 4- to 10-fold higher than those of celC and celD mRNAs, celF mRNA levels being intermediate (data not shown). These results are in agreement with the fact that various endoglucanases are present in differing quantities in the cellulosome (23). Indeed, endoglucanase A appears to be one of the major endoglucanase components found in the culture medium at the end of growth (5; unpublished data).

**FIG. 3.** Mapping of the 5' end of celF and celD mRNA. (A) 5'-3P-labeled celF-specific oligonucleotide (Fig. 4) was hybridized to mRNA from exponentially growing C. thermocellum cells and extended with reverse transcriptase, and the product was analyzed on a sequencing gel. P, Product of primer extension; T, C, G, and A, products of sequencing reactions obtained by using the same oligonucleotide as for lane P. (B) celD-specific oligonucleotide of Fig. 5 way hybridized to RNA isolated from C. thermocellum at the different time points indicated in Fig. 6. Reverse transcriptase-extended products are shown in lanes t₁, t₂, t₃, and t₄; T, C, G, and A, sequencing reactions derived from the same oligonucleotide as for lanes t₁ to t₄.

**DISCUSSION**

This article presents evidence that for the four individual C. thermocellum endoglucanase genes studied, mRNA levels are regulated. This was somewhat unexpected, since expression of endoglucanase activity in C. thermocellum is
FIG. 4. Position of the 5' end of celF mRNA within the sequence upstream from the celF structural gene. The first nucleotide of celF mRNA is indicated by a vertical arrow. The position of the oligonucleotide primer is shown by a horizontal arrow. Numbering starts at the first nucleotide of the coding sequence.

generally assumed to be constitutive. Garcia-Martinez et al. (8) and Shimmyo et al. (37) observed neither a delay nor a substantial inhibition of endoglucanase synthesis by *C. thermocellum* at initial cellulose concentrations of up to 30 g/liter. This apparent discrepancy may be due to the fact that the endoglucanase activity measured in these experiments (assayed by the release of trinitrophenol from trinitrophenylated carboxymethylcellulose [CMC]) (8, 37) could reflect the expression of still different genes, subject to a different type of control. There is indeed evidence that the expression of the various *C. thermocellum* cellulase components is differentially regulated. The synthesis of Avicelase but not CMCase activity is strongly repressed in the presence of readily metabolized carbon sources (8, 20). Furthermore, as pointed out above, the expression of celC is delayed relative to that of celA, celD, and celF. In this respect, is noteworthy that several features distinguish the celC-encoded endoglucanase C from other endoglucanases. Endoglucanase C is not part of the cellulosome (E. A. Bayer and R. Lamed, personal communication), and it is the only *C. thermocellum* endoglucanase identified so far that is able to hydrolyze cellobiose (36). Its function may be to act as a cellobextrinase. A shift in the ratios of enzymes with different specificities may be required for degradation of a dynamically changing substrate such as cellulose.

Regulation of transcription of cellulase-encoding genes has been reported for several other bacteria. In *C. fimii*, the transcription of *cenA*, *cenB*, *cenC*, and *cex* is CMC-inducible, while constitutive transcription in the presence of glycerol and glucose was only observed for *cenB* (13, 14, 28). Cellulase synthesis in *Thermomonospora fusca* and *Thermomonospora curvata* is subject to dual control by induction and growth rate-dependent repression by readily metabolizable carbon sources. The latter could be alleviated by addition of cyclic AMP (24, 40).

None of the *C. thermocellum* *cel* genes investigated was expressed during the early exponential phase, and the expression of *celC*, together with the expression of *celD* from the pD2 promoter, was turned on only after cessation of growth. Since the initial concentration of cellulose (0.5%) is the growth-limiting factor (data not shown), these findings are consistent with some kind of catabolite repression. However, the expression of *celA* and *celF* and the expression of *celD* from the pD1 promoter were turned on at a time when the cells were still growing exponentially and cellulose was therefore not limiting. Thus, the expression of *cel* genes might be controlled by concentrations of cellulose higher than those at which cellulose becomes limiting for growth.

The turn-off of the four genes during stationary phase may be correlated with a shift in the pattern of sigma factors as

---

**Fig. 5.** Positions of the celD mRNA start sites within the 5' sequence flanking the celD structural gene. The first nucleotide of the transcripts is indicated by a vertical arrow. The position of the oligonucleotide primer is shown by a horizontal arrow. Numbering starts at the first nucleotide of the coding sequence. The consensus *E. coli* (E.c.) σ70 and *B. subtilis* (B.s.) σ32 and σ80 promoter sequences are indicated under homologous sites of the *C. thermocellum* sequence.
the cells enter sporulation. The faster decline of transcripts originating at the \( \sigma^{28} \)-like pD1 promoter compared with the \( \sigma^{34} \)-like pD2 promoter matches the observations made in *B. subtilis* showing that \( \sigma^{28} \) transcripts decline very rapidly immediately after the beginning of the sporulation phase, whereas \( \sigma^{34} \) transcripts decline more slowly and to a lesser extent (9).

The sequence upstream from the 5' end of *celF* mRNA showed no significant similarity to previously described promoters and may thus represent a new type of promoter. Alternatively, the 5' end identified may result from the nuclease cleavage of an mRNA species whose true initiation site would be located further upstream. However, no such species could be identified up to 490 bp upstream from the coding sequence, and the 5' end detected by primer extension was quite sharp. Thus, if the 5' end of *celF* mRNA results from the processing of a larger species, such processing ought to be quite extensive and quite specific. The same considerations hold true for the start sites of *celD* mRNA.

The function of the long untranslated sequences found for both genes at the 5' ends of the transcripts is unknown. Computer searches extending from the coding sequence up to 90 bp upstream from the mRNA start sites failed to reveal obvious palindromes or reiterated sequences that would appear to be candidates for the regulation of gene expression.

ACKNOWLEDGMENTS

S.M. was a recipient of a fellowship from the Department of Biotechnology of the Government of India. This work was supported by grant EN3B-0082-F from the Commission of European Communities and by research funds from the University of Paris 7.

REFERENCES


