# Structural Features of the *Clostridium thermocellum* Cellulase $S_S$ Gene

WILLIAM K. WANG AND J. H. DAVID WU*

*University of Rochester, Department of Chemical Engineering, Rochester, NY 14627-0166*

## ABSTRACT

The *Clostridium thermocellum* cellulase $S_S$ is a subunit of the extracellular cellulase complex (cellulosome). It has previously been shown that $S_S$ hydrolyzes crystalline cellulose synergistically with another subunit, $S_L$. To study this synergism further, the authors cloned the gene coding for $S_S$ (*celS*) and compared its sequence to other known *cel* genes. The *celS,* although unique in its DNA sequence, has many structural features similar to those found in other *cel* genes. These features include a ribosome biding site, signal peptide sequence, the existence of a conserved reiterated amino acid sequence, and a palindromic structure downstream from its open reading frame.

**Index Entries:** *Clostridium thermocellum; celS* gene; cellulase; cellulose degradation; cellulase gene.

## INTRODUCTION

*Clostridium thermocellum* produces an extracellular cellulase system highly active on the crystalline cellulose. The Avicelase activity (activity against Avicel, a microcrystalline cellulose preparation) of this cellulase system resides mainly in an unusually large protein aggregate (cellulosome [1,2]), which has impeded the mechanistic study of this enzyme system. Purification of individual components from the cellulosome appears to be a prohibitive job. For this reason, studies of this enzyme system have been approached mostly through the molecular cloning of

*Author to whom all correspondence and reprint requests should be addressed.

cellulase genes. As a result, at least 15 endoglucanase genes, two xylanase genes, and two β-glucosidase genes have been cloned into *Escherichia coli* (*3–6*). Many of those genes (including *celA* [*7*], *celB* [*8*], *celC* [*9*], *celD* [*10*], *celE* [*11*], *celF* [*12*], *celH* [*13*], *xynZ* [*14*], *bglA* [*15*], and *bglB* [*16*]) have been sequenced.

Significant insights into the process of cellulose degradation by this enzyme system can be expected when the synergism between these cloned cellulase components is studied. However, at this time, it is not clear how many of these components are part of the cellulosome that accounts for most of the Avicelase activity. It would therefore be useful to target the cloning at the cellulosome subunits. In previous work, two cellulosome subunits essential for degrading crystalline cellulose, $S_S$ ($M_r$ = 82,000) and $S_L$ ($M_r$ = 250,000), have been identified (*17*). $S_S$ and $S_L$ act synergistically to degrade crystalline cellulose. The properties of their activity are consistent with those observed with the crude enzyme preparation (*18*). To study the synergism between $S_S$ and $S_L$ further, the gene coding for $S_S$ (*celS*) has been cloned using an oligonucleotide probe derived from the N-terminal amino acid sequence of $S_S$. The cloning of *celS* and its complete DNA sequence are published in a separate paper (*19*). In this article, the structural features of *celS* are compared to other known *cel* genes from *C. thermocellum*.

## MATERIALS AND METHODS

### Bacterial Strains and Vectors

*Clostridium thermocellum* ATCC 27405 was used as a source of the $S_S$ protein and for genomic DNA. The cloning vectors used were phage Lambda ZAPII (Stratagene, CA) and plasmid pBR322. *E. coli* XL-1 Blue {*rec*A1 *end*A1 *gyr*A96 *thi*-1 *hsd*R17 *supE*44 *rel*A1 *lac* [F′ *pro*AB *lac*I$^q$ZΔM15 Tn*10*(tet$^r$)]} was used as the cloning host for bacteriophage Lambda ZAPII. *E. coli* DH10B [F$^-$ *mcr*A Δ(*mrr*-*hsd*RMS-*mcr*BC) φ80Δ*lac*ZΔM15 Δ*lac*X74 *end*A1 *rec*A1 *deo*R Δ(*ara, leu*) 7697 *ara*D139 *gal*U *gal*K *nup*G *rps*L λ$^-$] was used as the host for pBR322. Growth and maintenance were carried out as described in a separate paper (*19*).

### Cloning and Sequencing of the $S_S$ Gene

The detailed procedures for cloning and sequencing the $S_S$ gene are described in a separate paper (*19*). In brief, the N-terminal amino acid sequence of the native $S_S$ protein was determined as described by Matsudaira and by LeGendre and Matsudaira (*20,21*). An oligonucleotide probe was constructed based on this amino acid sequence and used to screen *C. thermocellum* genomic libraries. Positive clones were sequenced using the dideoxy termination method (22). A DNA insert was found to

Table 1
Characteristics of Cloned C. *thermocellum* Endoglucanases[a]

| Cloned gene | Length, bp | Predicted mol wt, dalton | Actual mol wt, dalton[b] |
|---|---|---|---|
| *celA* | 1344 | 52,503 | 56,000 |
| *celB* | 1689 | 63,857 | 66,000 |
| *celC* | 1032 | 40,439 | 38,000 |
| *celD* | 1947 | 72,344 | 65,000 |
| *celE* | 2442 | 90,211 | - |
| *celF* | 2217 | 82,015 | - |
| *celH* | 2702 | 102,301 | - |
| *celS* | 2223 | 80,670 | 82,000 |

[a] Data taken from refs. (*7–13,19*).
[b] Molecular weight of the native protein produced by C. *thermocellum*.

contain a sequence coding for the N-terminal amino acid sequence of $S_S$. This clone contained a truncated structural gene of $S_S$. The complete open reading frame of the $S_S$ gene was obtained by a "chromosome walk" procedure and by subsequent reconstruction of the sequence from a total of four clones.

## DNA Sequence Analysis

The DNA sequence was analyzed using the computer software package developed by the University of Wisconsin Genetics Computer Group (*23,24*).

# RESULTS AND DISCUSSION

## The Open Reading Frame

The translated sequence of *celS* consists of 2223 bp, which encode a peptide of 741 amino acid residues, including a putative signal peptide of 27 amino acid residues. The 714 amino acid residue CelS protein has a predicted mol wt of 80,670 daltons (*19*). The comparison of its size to sizes of other C. *thermocellum cel* genes of known sequence is shown in Table 1. CelE and CelH are clearly larger than $S_S$. On the other hand, CelF has a very similar size to $S_S$, although they share no homologous sequences except the conserved reiterated sequence described below. The rest of the genes of known sequences are much smaller than *celS*.

## The Amino Acid Composition

The amino acid composition of the deduced *celS* product is shown in Table 2. The composition is not unique compared to other *cel* gene products, including CelA, CelB, CelC, CelD, and CelH. It is noteworthy that

Table 2
Amino Acid Composition of CelS
Deduced from the Nucleotide Sequence

| Amino Acid | Number | Percentage |
|---|---|---|
| *Ala* | 60 | 8.4 |
| *Cys* | 2 | 0.3 |
| *Asp* | 50 | 7.0 |
| *Glu* | 38 | 5.3 |
| *Phe* | 27 | 3.8 |
| *Gly* | 65 | 9.1 |
| *His* | 12 | 1.7 |
| *Ile* | 27 | 3.8 |
| *Lys* | 42 | 5.9 |
| *Leu* | 38 | 5.3 |
| *Met* | 18 | 2.5 |
| *Asn* | 33 | 4.6 |
| *Pro* | 37 | 5.2 |
| *Gln* | 21 | 2.9 |
| *Arg* | 25 | 3.5 |
| *Ser* | 47 | 6.6 |
| *Thr* | 50 | 7.0 |
| *Val* | 40 | 5.6 |
| *Trp* | 28 | 3.9 |
| *Tyr* | 54 | 7.5 |
|  | 714 | 99.9 |

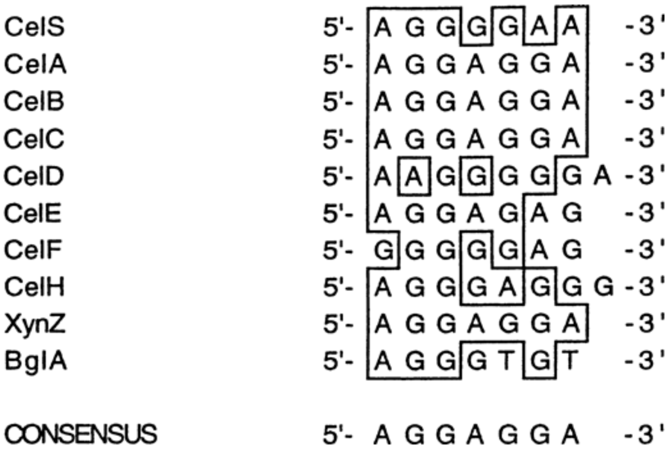
there are only two cysteine residues in the entire sequence of $S_S$. Other *C. thermocellum cel* genes also have low cysteine contents, except for *celF,* which contains 7% cysteine (*12*).

## The Ribosome Binding Site

A putative ribosome binding site is located in the 5′ end of *celS*. It is highly homologous to other Shine-Dalgarno sequences from *C. thermocellum* (Fig. 1). However, as in many other *cel* genes, no obvious promoter sequence can be identified in *celS*.

## The Signal Peptide Sequence

The deduced amino acid sequence of *celS* contains a sequence similar to the signal peptide sequence for prokaryotic secretory proteins. This signal peptide sequence is located upstream of the N-terminal amino acid sequence of the $S_S$ protein, confirming that $S_S$ is a secreted protein. As shown in Fig. 2, the signal peptides of various *C. themocellum* genes are of different length. However, they all share general characteristics, such as a



| | | | |
|---|---|---|---|
| CelS | 5'- | A G G G G A A | -3' |
| CelA | 5'- | A G G A G G A | -3' |
| CelB | 5'- | A G G A G G A | -3' |
| CelC | 5'- | A G G A G G A | -3' |
| CelD | 5'- | A A G G G G G A | -3' |
| CelE | 5'- | A G G A G A G | -3' |
| CelF | 5'- | G G G G G A G | -3' |
| CelH | 5'- | A G G G A G G G | -3' |
| XynZ | 5'- | A G G A G G A | -3' |
| BglA | 5'- | A G G G T G T | -3' |
| CONSENSUS | 5'- | A G G A G G A | -3' |

Fig. 1. A comparison of ribosome binding sites from the *celS, celA, celB, celC, celD, celE, celF, celH, xynZ,* and *bglA* genes from *C. thermocellum*. The consensus Shine-Dalgarno sequence is also shown.

short region rich in positively charged amino acid species, followed by a sequence of predominantly hydrophobic residues, a residue breaking the secondary structure (glycine or proline), and a cleavage site ending with alanine, glycine, or serine (*25*).

## The Conserved Reiterated Sequence

Comparison of the deduced CelS amino acid sequence with sequences of other cellulases from *C. thermocellum* and other organisms revealed no global homologies (*23,24*). However, the deduced *celS* peptide contains a highly conserved reiterated sequence of 24 amino acids, found also in CelA (*7*), CelB (*8*), CelD (*10*), CelE (*11*), CelF (*12*), CelH (*13*), CelX (*11*), and XynZ (*14*) of *C. thermocellum,* as well as in CelCCA (*26*) of *C. cellulolyticum* (Fig. 3). This sequence is located near the C-terminus of CelS and most of the other proteins. The reiterated sequences in *celS* are linked by eight amino acid residues similar to those in CelA and CelB (Fig. 3). It has been suggested that the reiterated sequences play a role in cellulose binding or in protein–protein interaction (*7,27*).

## The Palindromic Structure

Downstream from the open reading frame of *celS,* a palindromic sequence is located 16 bp from the stop codon. A possible stem and loop structure is shown in Fig. 4. Similar structures have been reported following the open reading frames of *celA* (*7*), *celB* (*8*), *celD* (*10*), *celH* (*13*), *xynZ* (*14*), and *bglA* (*15*) of *C. thermocellum*. These structures probably play a role in transcription termination (*28*).

| CelS | CelA | CelB | CelC | CelD | CelE | CelF | CelH | XynZ |
|---|---|---|---|---|---|---|---|---|
| | | | | fMET | | | | |
| | | | | S | | | | |
| | | | | R+ | | | | |
| | | | | M | | | | |
| | | | | T | | | | |
| | | | | L | | | | |
| | fMET | | | K+ | | | | |
| fMET | K+ | | | S | | | | |
| V | N | | fMET | S | | | | |
| K+ | V | | V | M | | | fMET | fMET |
| S | K+ | fMET | S | K+ | fMET | fMET | K+ | S |
| R+ | K+ | K+ | F | K+ | K+ | K+ | K+ | R+ |
| K+ | R+ | K+ | K+ | R+ | K+ | K+ | R+ | K+ |
| I | V | F | A | V | I | I | L | L |
| S | G | L | G | L | V | L | L | F |
| I | V | V | I | S | S | A | V | S |
| L | V | L | N | L | L | F | S | V |
| L | L | L | L | L | V | L | F | L |
| A | L | I | G | I | C | L | F | L |
| V | I | A | G | A | V | T | L | V |
| A | L | L | W | V | L | V | V | G |
| M | A | I | I | V | V | A | L | L |
| L | V | M | S | F | M | L | S | M |
| V | L | I | Q | L | L | V | I | L |
| S | G | A | Y | S | V | A | I | M |
| I | V | T | Q | L | S | V | V | T |
| M | Y | L | V | T | I | V | G | S |
| I | M | L | F | G | L | A | L | L |
| P | L | V | S | V | G | I | L | L |
| T | A | V | | F | S | P | S | V |
| T | M | P | | P | F | Q | F | T |
| A | P | G | | S | S | A | Q | I |
| F | A | V | | G | V | V | S | S |
| A | N | Q | | L | V | V | L | S |
| | T | T | | I | A | S | G | T |
| | V | S | | E | A | F | N | S |
| | S | A | | T | S | A | Y | A |
| | A | E | | K | P | | N | |
| | | G | | V | V | | S | |
| | | S | | S | K | | G | |
| | | Y | | A | G | | L | |
| | | A | | | F | | K | |
| | | | | | Q | | I | |
| | | | | | V | | G | |
| | | | | | | | A | |
| | | | | | | | W | |
| | | | | | | | V | |
| | | | | | | | G | |
| | | | | | | | T | |
| | | | | | | | Q | |
| | | | | | | | P | |
| | | | | | | | S | |
| | | | | | | | E | |
| | | | | | | | S | |

Fig. 2. A comparison of the signal peptides of the CelS, Cela, CelB, CelC, CelD, CelE, CelF, CelH, and XynZ proteins from *C. thermocellum*.

## The Hydrophilicity Plot

A hydrophilicity plot, generated by the Kyte-Doolittle algorithm (*29*), of the CelS protein is shown in Fig. 5. Except for the region of the signal peptide and a few other short hydrophobic regions, the protein is generally hydrophilic.

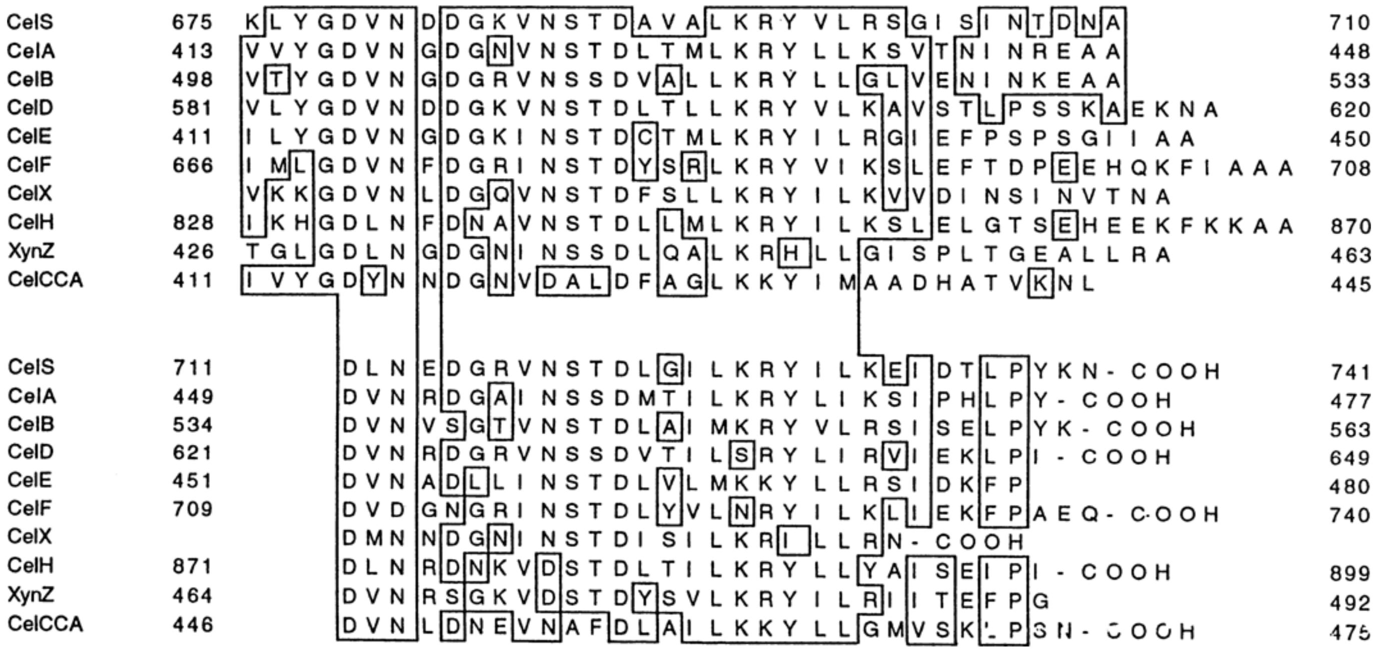

CelS 675 K L Y G D V N D D G K V N S T D A V A L K R Y V L R S G I S I N T D N A 710
CelA 413 V V Y G D V N G D G N V N S T D L T M L K R Y L L K S V T N I N R E A A 448
CelB 498 V T Y G D V N G D G R V N S S D V A L L K R Y L L G L V E N I N K E A A 533
CelD 581 V L Y G D V N D D G K V N S T D L T L L K R Y V L K A V S T L P S S K A E K N A 620
CelE 411 I L Y G D V N G D G K I N S T D C T M L K R Y I L R G I E F P S P S G I I A A 450
CelF 666 I M L G D V N F D G R I N S T D Y S R L K R Y V I K S L E F T D P E E H Q K F I A A A 708
CelX V K K G D V N L D G Q V N S T D F S L L K R Y I L K V V D I N S I N V T N A
CelH 828 I K H G D L N F D N A V N S T D L L M L K R Y I L K S L E L G T S E H E E K F K K A A 870
XynZ 426 T G L G D L N G D G N I N S S D L Q A L K R H L L G I S P L T G E A L L R A 463
CelCCA 411 I V Y G D Y N N D G N V D A L D F A G L K K Y I M A A D H A T V K N L 445
CelS 711 D L N E D G R V N S T D L G I L K R Y I L K E I D T L P Y K N - C O O H 741
CelA 449 D V N R D G A I N S S D M T I L K R Y L I K S I P H L P Y - C O O H 477
CelB 534 D V N V S G T V N S T D L A I M K R Y V L R S I S E L P Y K - C O O H 563
CelD 621 D V N R D G R V N S S D V T I L S R Y L I R V I E K L P I - C O O H 649
CelE 451 D V N A D L L I N S T D L V L M K K Y L L R S I D K F P 480
CelF 709 D V D G N G R I N S T D L Y V L N R Y I L K L I E K F P A E Q - C·O O H 740
CelX D M N N D G N I N S T D I S I L K R I L L R N - C O O H
CelH 871 D L N R D N K V D S T D L T I L K R Y L L Y A I S E I P I - C O O H 899
XynZ 464 D V N R S G K V D S T D Y S V L K R Y I L R I I T E F P G 492
CelCCA 446 D V N L D N E V N A F D L A I L K K Y L L G M V S K L P S N - C O O H 475

Fig. 3. The alignment of the reiterated, conserved region between the CelS, CelA, CelB, CelD, CelE, CelF, CelX, CelH, and XynZ of *C. thermocellum* and CelCCA of *C. cellulolyticum*. Boxed amino acid residues are identical or have similar chemical properties. Numbers indicate the position, within the sequence of each protein, of the first or the last amino acid residue shown on a line. Similar residues were: V,L,I,M,F;R,K;D,E;N,Q;Y,F,W;S,T.

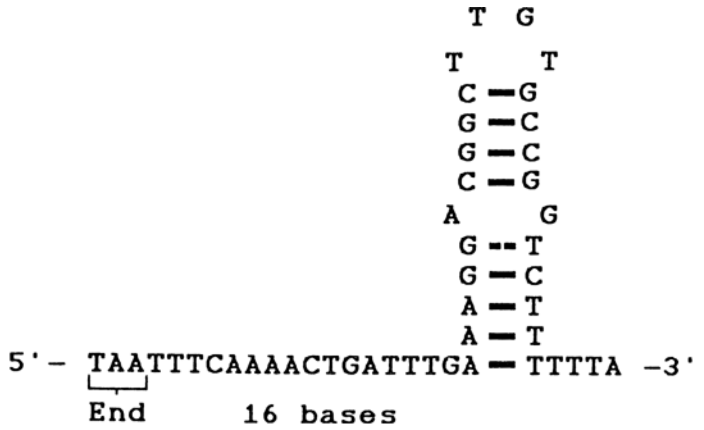


Fig. 4. A possible stem-and-loop structure located 16 bp downstream of the termination codon of the *celS* gene. This structure is composed of two palindromic sequences and may function as a transcription termination signal.

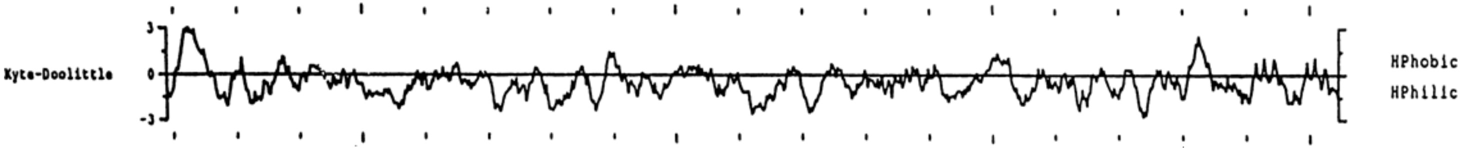


Fig. 5. A hydrophilicity plot of the CelS protein. The Kyte-Doolittle algorithm predicts that the CelS protein has a strong hydrophobic leader peptide and some weaker hydrophobic regions, but the CelS is generally hydrophilic.

## CONCLUSION

The *celS* is a new *cel* gene identified from *C. thermocellum*. It shares many structural features with other *cel* genes from the same bacterium. It is probably one of the first of the identified *cel* genes known to code for a cellulosome subunit. Since the *celS* product is one of the key subunits of the cellulosome for degrading crystalline cellulose, its availability will be useful for elucidating the mechanism of this enzyme system.

## ACKNOWLEDGMENTS

This project is suported in part by a Biomedical Research Support Grant from PHS. W. K. Wang appreciates the support of the Link Foundation Energy Fellowship.

## ADDENDUM IN PROOF

Since this article was submitted, the CelS deduced amino acid sequence has been found to have greater than 50% homology with two partial OFRs. One ORF is located upstream of the *celCCC* gene of *Clostridium cellulolyticum* (Bagnara-Tardif, C., Gaudin, C., Belaich, A., Hoest, P., Citard, T., and Belaich, J.-P. 1992. *Gene* **119:** 17–28) and the other precedes the *manA* gene of *Caldocellum saccharolyticum* (Luthi, E., Jasmat, N. B., Grayling, R. A., Love, D. R., and Bergquist, P. L. 1991. *Appl. Environ. Microbiol.* **57:** 694–700). The complete genes for these two polypeptides have not been reported.

## REFERENCES

1. Lamed, R., Setter, E., and Bayer, E. A. (1983), *J. Bacteriol.* **156,** 828–836.
2. Lamed, R., Kenig, R., Setter, E., and Bayer, E. A. (1985), *Enzym. Microb. Technol.* **7,** 37–41.
3. Cornet, P., Millet, J., Beguin, P., and Aubert, J.-P. (1983), *FEMS Microbiol. Lett.* **16,** 137–141.
4. Schwarz, W., Bronnenmeier, K., and Staudenbauer, W. L. (1985), *Biotechnol. Lett.* **7,** 859–864.
5. Grabnitz, F. and Staudenbauer, W. L. (1988), *Biotechnol. Lett.* **10,** 73–78.
6. Hazlewood, G. P., Romaniec, M. P. M., Davidson, K., Grepinet, O., Beguin, P., Millet, J., Raynaud, O., and Aubert, J.-P. (1988), *FEMS Microbiol. Lett.* **51,** 231–236.
7. Beguin, P., Cornet, P., and Aubert, J.-P. (1985), *J. Bacteriol.* **162,** 102–105.
8. Grepinet, O. and Beguin, P. (1986), *Nucleic Acids Res.* **14,** 1791–1799.

9. Schwarz, W. H., Schimming, S., Rucknagel, K. P., Burgschwaiger, S., Kriel, G., and Staudenbauer, W. (1988), *Gene* **63,** 23-30.
10. Joliff, G., Beguin, P., and Aubert, J.-P. (1986), *Nucleic Acids Res.* **14,** 8605-8613.
11. Hall, J., Hazlewood, G. P., Barker, P. J., and Gilbert, H. J. (1988), *Gene* **69,** 29-38.
12. Navarro, A., Chebrou, M.-C., Beguin, P., and Aubert, J.-P. (1991), *Res. Microbiol.* **142,** 927-936.
13. Yague, E., Beguin, P., and Aubert, J.-P. (1990), *Gene* **89,** 61-67.
14. Grepinet, O., Chebrou, M.-C., and Beguin, P. (1988), *J. Bacteriol.* **170,** 4582-4588.
15. Grabnitz, F., Seiss, M., Rucknagel, K. P., and Staudenbauer, W. L. (1991), *Eur. J. Biochem.* **200,** 301-309.
16. Grabnitz, F., Rucknagel, K. P., Seiss, M., and Staudenbauer, W. L. (1989), *Mol. Gen. Genet.* **217,** 10-76.
17. Wu, J. H. D., Orme-Johnson, W. H., and Demain, A. L. (1988), *Biochemistry* **27,** 1703-1709.
18. Wu, J. H. D. and Demain, A. L. (1988), in *Biochemistry and Genetics of Cellulose Degradation,* Aubert, J.-P., Beguin, P., and Millet, J., eds., Academic Press, New York, pp. 117-131.
19. Wang, W. K. and Wu, J. H. D. (1993), *J. Bacteriol.,* accepted.
20. Matsudaira, P. (1987), *J. Biol. Chem.* **262,** 10,035-10,038.
21. LeGendre, N. and Matsudaira, P. (1988), *BioTechniques* **6,** 154-159.
22. Sanger, F., Nicklen, S., and Coulson, A. R. (1977), *Proc. Natl. Acad. Scien.* **74,** 5463-5467.
23. Pearson, W. R. and Lipman, D. J. (1988), *Proc. Natl. Acad. Scien.* **85,** 2444-2448.
24. Devereux, J., Haeberli, P., and Smithies, O. (1984), *Nucleic Acids Res.* **12,** 387-395.
25. Watson, M. E. E. (1984), *Nucleic Acids Res.* **12,** 5145-5164.
26. Fature, E., Balaich, A., Bagnara, C., Gaudin, C., and Aubert, J.-P. (1989), *Gene* **84,** 39-46.
27. Tokatidis, K., Salamitou, S., Beguin, P., Dhurjati, P., and Aubert, J.-P. (1992), *Abstr. Am. Chem. Soc. Annu. Meet.,* San Francisco. BTD Paper 16.
28. Rosenberg, M., and Court, D. (1979), *Ann. Rev. Genet.* **13,** 319-353.
29. Kyte, J. and Doolittle, R. F. (1982), *J. Mol. Biol.* **157,** 105-132.