Journal of Bacteriology, Apr. 1993, p. 1891–1899
0021-9193/93/071891-09$02.00/0
Copyright © 1993, American Society for Microbiology

Vol. 175, No. 7

Organization of a *Clostridium thermocellum* Gene Cluster Encoding the Cellulosomal Scaffolding Protein CipA and a Protein Possibly Involved in Attachment of the Cellulosome to the Cell Surface

TSUCHIYOSHI FUJINO,† PIERRE BÉGUIN, and JEAN-PAUL AUBERT*

Unité de Physiologie Cellulaire and URA 1300 Centre National de la Recherche Scientifique, Département des Biotechnologies, Institut Pasteur, 28, rue du Dr. Roux, 75724 Paris Cedex 15, France

Received 30 October 1992/Accepted 14 January 1993

The nucleotide sequence was determined for a 9.4-kb region of *Clostridium thermocellum* DNA extending from the 3′ end of the gene (now termed *cipA*), encoding the S1/S_L component of the cellulosome. Three open reading frames (ORFs) belonging to two operons were detected. They encoded polypeptides of 1,664, 688, and 447 residues, termed ORF1p, ORF2p, and ORF3p, respectively. The COOH-terminal regions of the three polypeptides were highly similar and contained three reiterated segments of 60 to 70 residues each. Similar segments have been found at the NH_2 terminus of the S-layer proteins of *Bacillus brevis* and *Acetogenium kivui*, suggesting that ORF1p, ORF2p, and ORF3p might also be located on the cell surface. Otherwise, the sequence of ORF1p and ORF2p gave little clue concerning their potential function. However, the NH_2-terminal region of ORF3p was similar to the reiterated domains previously identified in CipA as receptors involved in binding the duplicated segment of 22 amino acids present in catalytic subunits of the cellulosome. Indeed, it was found previously that ORF3p binds ^{125}I-labeled endoglucanase CelD containing the duplicated segment (T. Fujino, P. Béguin, and J.-P. Aubert, FEMS Microbiol. Lett. 94:165–170, 1992). These findings suggest that ORF3p might serve as an anchoring factor for the cellulosome on the cell surface by binding the duplicated segment that is present at the COOH end of CipA.

Clostridium thermocellum, a gram-positive anaerobic bacterium, produces a highly active, thermostable cellulase system in which the various cellulolytic components are associated into a high-molecular-weight complex termed the cellulosome (3, 12). The cellulosome is found both in the culture medium and at the surface of the bacteria, where it mediates adhesion of the cells to the substrate. The cellulosome components possess endoglucanase (1, 11), cellobiohydrolase (19), or hemicellulase (6, 9, 17) activity, with the exception of a 210- to 250-kDa glycoprotein previously termed S1 (11) or S_L (27, 28), that has recently been renamed CipA (for cellulosome-integrating protein [3a]). The complex is highly stable, and dissociation requires strongly denaturing conditions. Previous data suggested that CipA fulfills a dual function by promoting binding of the cellulosome to the substrate and by acting as a scaffolding protein around which the catalytic components are organized (10, 18, 22, 27). The

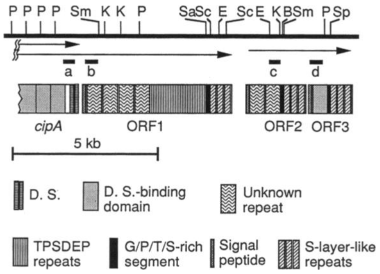


FIG. 1. Structural and transcriptional organization of the genes within the region adjacent to the 3′ end of *cipA*. The positions of probes a, b, c, and d used for mRNA hybridization are shown by horizontal bars. The positions and orientation of transcripts are indicated with arrows. The positions of the segments encoding the various regions identified within each polypeptide are shown by boxes of different patterns. E, *Eco*RI; K, *Kpn*I; P, *Pst*I; Sa, *Sal*I; Sc, *Sac*I; Sm, *Sma*I; D.S., 22-amino-acid duplicated segment.

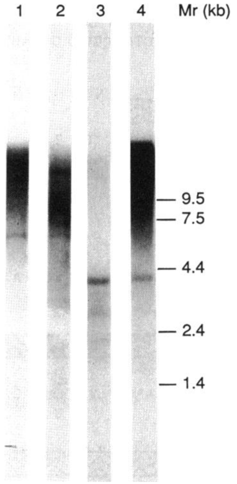
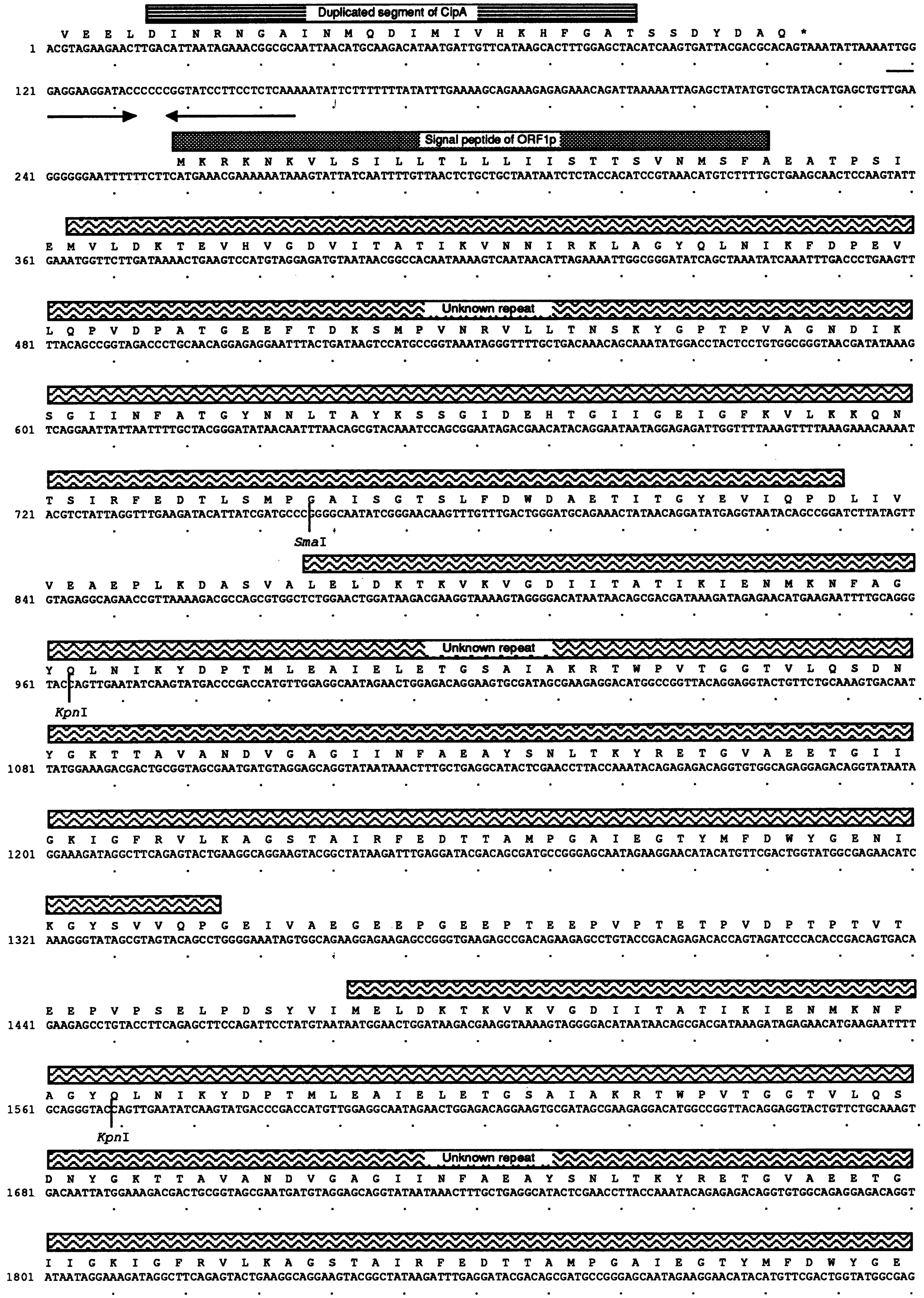


FIG. 2. Hybridization of mRNA from *C. thermocellum* grown in the presence of cellulose with probes derived from *cipA* (lane 1), ORF1 (lane 2), ORF2 (lane 3), and ORF3 (lane 4). Hybridizations to ORF1 and ORF2 were performed on the same blots as hybridizations to *cipA* and ORF3, respectively, after the previous probe had been stripped by boiling the nitrocellulose sheet in water for 5 min. Control autoradiograms showed that no detectable probe remained on the nitrocellulose after boiling. The positions of RNA size markers (GIBCO-BRL) are shown to the right.

* Corresponding author.

† Permanent address: Nagoya Seiraku Co. Ltd., Nagoya 468, Japan.



```
                          Duplicated segment of CipA
          V  E  E  L  D  I  N  R  N  G  A  I  N  M  Q  D  I  M  I  V  H  K  H  F  G  A  T  S  S  D  Y  D  A  Q  *
     1 ACGTAGAAGAACTTGACATTAATAGAAACGGCGCAATTAACATGCAAGACATAATGATTGTTCATAAGCACTTTGGAGCTACATCAAGTGATTACGACGCACAGTAAATATTAAAATTGG

   121 GAGGAAGGATACCCCCCGGTATCCTTCCTCTCAAAAATATTCTTTTTTTATATTTGAAAAGCAGAAAGAGAGAAACAGATTAAAAATTAGAGCTATATGTGCTATACATGAGCTGTTGAA

                          Signal peptide of ORF1p
                          M  K  R  K  N  K  V  L  S  I  L  L  T  L  L  L  I  I  S  T  T  S  V  N  M  S  F  A  E  A  T  P  S  I
   241 GGGGGGAATTTTTTCTTCATGAAACGAAAAAATAAAGTATTATCAATTTTGTTAACTCTGCTGCTAATAATCTCTACCACATCCGTAAACATGTCTTTTGCTGAAGCAACTCCAAGTATT

        E  M  V  L  D  K  T  E  V  H  V  G  D  V  I  T  A  T  I  K  V  N  N  I  R  K  L  A  G  Y  Q  L  N  I  K  F  D  P  E  V
   361 GAAATGGTTCTTGATAAAACTGAAGTCCATGTAGGAGATGTAATAACGGCCACAATAAAAGTCAATAACATTAGAAAATTGGCGGGATATCAGCTAAATATCAAATTTGACCCTGAAGTT

                          Unknown repeat
        L  Q  P  V  D  P  A  T  G  E  E  F  T  D  K  S  M  P  V  N  R  V  L  L  T  N  S  K  Y  G  P  T  P  V  A  G  N  D  I  K
   481 TTACAGCCGGTAGACCCTGCAACAGGAGAGGAATTTACTGATAAGTCCATGCCGGTAAATAGGGTTTTGCTGACAAACAGCAAATATGGACCTACTCCTGTGGCGGGTAACGATATAAAG

        S  G  I  I  N  F  A  T  G  Y  N  N  L  T  A  Y  K  S  S  G  I  D  E  H  T  G  I  I  G  E  I  G  F  K  V  L  K  K  Q  N
   601 TCAGGAATTATTAATTTTGCTACGGGATATAACAATTTAACAGCGTACAAATCCAGCGGAATAGACGAACATACAGGAATAATAGGAGAGATTGGTTTTAAAGTTTTAAAGAAACAAAAT

        T  S  I  R  F  E  D  T  L  S  M  P  G  A  I  S  G  T  S  L  F  D  W  D  A  E  T  I  T  G  Y  E  V  I  Q  P  D  L  I  V
   721 ACGTCTATTAGGTTTGAAGATACATTATCGATGCCCGGGGCAATATCGGGAACAAGTTTGTTTGACTGGGATGCAGAAACTATAACAGGATATGAGGTAATACAGCCGGATCTTATAGTT
                                           SmaI

        V  E  A  E  P  L  K  D  A  S  V  A  L  E  L  D  K  T  K  V  K  V  G  D  I  I  T  A  T  I  K  I  E  N  M  K  N  F  A  G
   841 GTAGAGGCAGAACCGTTAAAAGACGCCAGCGTGGCTCTGGAACTGGATAAGACGAAGGTAAAAGTAGGGGACATAATAACAGCGACGATAAAGATAGAGAACATGAAGAATTTTGCAGGG

                          Unknown repeat
        Y  Q  L  N  I  K  Y  D  P  T  M  L  E  A  I  E  L  E  T  G  S  A  I  A  K  R  T  W  P  V  T  G  G  T  V  L  Q  S  D  N
   961 TACCAGTTGAATATCAAGTATGACCCGACCATGTTGGAGGCAATAGAACTGGAGACAGGAAGTGCGATAGCGAAGAGGACATGGCCGGTTACAGGAGGTACTGTTCTGCAAAGTGACAAT
       KpnI

        Y  G  K  T  T  A  V  A  N  D  V  G  A  G  I  I  N  F  A  E  A  Y  S  N  L  T  K  Y  R  E  T  G  V  A  E  E  T  G  I  I
  1081 TATGGAAAGACGACTGCGGTAGCGAATGATGTAGGAGCAGGTATAATAAACTTTGCTGAGGCATACTCGAACCTTACCAAATACAGAGAGACAGGTGTGGCAGAGGAGACAGGTATAATA

        G  K  I  G  F  R  V  L  K  A  G  S  T  A  I  R  F  E  D  T  T  A  M  P  G  A  I  E  G  T  Y  M  F  D  W  Y  G  E  N  I
  1201 GGAAAGATAGGCTTCAGAGTACTGAAGGCAGGAAGTACGGCTATAAGATTTGAGGATACGACAGCGATGCCGGGAGCAATAGAAGGAACATACATGTTCGACTGGTATGGCGAGAACATC

        K  G  Y  S  V  V  Q  P  G  E  I  V  A  E  G  E  E  P  G  E  E  P  T  E  E  P  V  P  T  E  T  P  V  D  P  T  P  T  V  T
  1321 AAAGGGTATAGCGTAGTACAGCCTGGGGAAATAGTGGCAGAAGGAGAAGAGCCGGGTGAAGAGCCGACAGAAGAGCCTGTACCGACAGAGACACCAGTAGATCCCACACCGACAGTGACA

        E  E  P  V  P  S  E  L  P  D  S  Y  V  I  M  E  L  D  K  T  K  V  K  V  G  D  I  I  T  A  T  I  K  I  E  N  M  K  N  F
  1441 GAAGAGCCTGTACCTTCAGAGCTTCCAGATTCCTATGTAATAATGGAACTGGATAAGACGAAGGTAAAAGTAGGGGACATAATAACAGCGACGATAAAGATAGAGAACATGAAGAATTTT

        A  G  Y  Q  L  N  I  K  Y  D  P  T  M  L  E  A  I  E  L  E  T  G  S  A  I  A  K  R  T  W  P  V  T  G  G  T  V  L  Q  S
  1561 GCAGGGTACCAGTTGAATATCAAGTATGACCCGACCATGTTGGAGGCAATAGAACTGGAGACAGGAAGTGCGATAGCGAAGAGGACATGGCCGGTTACAGGAGGTACTGTTCTGCAAAGT
              KpnI

                          Unknown repeat
        D  N  Y  G  K  T  T  A  V  A  N  D  V  G  A  G  I  I  N  F  A  E  A  Y  S  N  L  T  K  Y  R  E  T  G  V  A  E  E  T  G
  1681 GACAATTATGGAAAGACGACTGCGGTAGCGAATGATGTAGGAGCAGGTATAATAAACTTTGCTGAGGCATACTCGAACCTTACCAAATACAGAGAGACAGGTGTGGCAGAGGAGACAGGT

        I  I  G  K  I  G  F  R  V  L  K  A  G  S  T  A  I  R  F  E  D  T  T  A  M  P  G  A  I  E  G  T  Y  M  F  D  W  Y  G  E
  1801 ATAATAGGAAAGATAGGCTTCAGAGTACTGAAGGCAGGAAGTACGGCTATAAGATTTGAGGATACGACAGCGATGCCGGGAGCAATAGAAGGAACATACATGTTCGACTGGTATGGCGAG
```

FIG. 3. Nucleotide sequence of the region extending downstream from *cipA*. Palindromic sequences are shown by arrows. The positions of the various regions identified within each polypeptide are indicated by boxes of the same patterns as in Fig. 1.

```
      N  I  K  G  Y  S  V  V  Q  P  G  E  I  V  A  E  G  E  E  P  T  E  E  P  V  P  T  E  T  P  V  D  P  T  P  T  V  T  E  E
1921 AACATCAAAGGGTATAGCGTAGTACAGCCTGGGGAAATAGTGGCGGAAGGAGAAGAGCCGACAGAAGAGCCTGTACCGACAGAGACACCAGTAGATCCCACACCGACAGTGACAGAAGAG

      P  V  P  S  E  L  P  D  S  Y  V  I  M  E  L  D  K  T  K  V  K  E  G  D  V  I  I  A  T  I  R  V  N  N  I  K  N  L  A  G
2041 CCTGTACCTTCAGAGCTTCCAGATTCCTATGTGATAATGGAATTGGATAAGACGAAGGTAAAAGAAGGCGACGTAATAATAGCAACAATAAGAGTAAATAACATAAAGAATCTTGCCGGA

      Y  Q  I  G  I  K  Y  D  P  K  V  L  E  A  F  N  I  E  T  G  D  P  I  D  E  G  T  W  P  A  V  G  G  T  I  L  K  N  R  D
2161 TATCAGATAGGCATCAAATATGACCCGAAAGTATTAGAGGCATTTAATATCGAGACAGGGGACCCAATAGATGAAGGAACATGGCCTGCAGTAGGGGGAACAATACTGAAGAATAGAGAT
                                                                                                    PstI

                                                 Unknown repeat
      Y  L  P  T  G  V  A  I  N  N  V  S  K  G  I  L  N  F  A  A  Y  Y  V  Y  F  D  D  Y  R  E  E  G  K  S  E  D  T  G  I  I
2281 TACCTGCCGACTGGGGTAGCAATAAACAATGTATCTAAAGGAATACTGAATTTTGCTGCTTATTACGTTTACTTCGATGACTATAGAGAGGAAGGAAAGTCAGAAGATACAGGAATTATA

      G  N  I  G  F  R  V  L  K  A  E  D  T  T  I  R  F  E  E  L  E  S  M  P  G  S  I  D  G  T  Y  M  L  D  W  Y  L  N  R  I
2401 GGAAATATAGGCTTTAGAGTACTGAAGGCGGAAGATACAACGATAAGATTTGAAGAGCTGGAGTCAATGCCGGGTTCAATAGACGGAACATATATGTTGGATTGGTATCTTAATAGAATC

      S  G  Y  V  V  I  Q  P  A  P  I  K  A  A  S  D  E  P  I  P  T  D  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P
2521 TCTGGCTATGTAGTAATACAACCGGCGCCTATAAAGGCGGCTAGTGACGAACCAATACCAACGGATACACCATCAGATGAACCGACACCGTCAGACGAGCCAACGCCATCTGACGAACCG

      T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  E  T  P  E  E  P  I  P  T  D  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S
2641 ACACCGTCTGATGAGCCAACACCGTCAGATGAACCGACTCCGTCAGAGACACCTGAGGAGCCGATACCGACGGATACACCATCAGATGAACCGACACCATCAGACGAGCCAACGCCATCT

      D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  E  T  P  E  E  P  I  P  T  D  T  P  S  D  E  P  T  P  S  D  E  P
2761 GATGAACCAACACCGTCTGATGAGCCAACACCATCTGATGAACCGACTCCGTCAGAGACACCTGAGGAGCCGATACCGACGGATACACCATCAGATGAACCGACACCGTCAGACGAGCCA

      T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  E  T  P  E  E  P  I  P  T  D  T  P  S  D  E  P  T  P  S
2801 ACGCCATCTGACGAACCAACACCGTCTGATGAGCCAACACCGTCAGATGAACCGACTCCGTCAGAGACACCTGAGGAGCCGATACCGACGGATACACCATCAGATGAACCGACACCGTCA

      D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  E  T  P  E  E  P  I  P  T  D  T  P  S  D  E  P
3001 GACGAGCCAACGCCATCTGACGAACCAACACCGTCTGATGAGCCAACACCGTCAGATGAACCGACTCCGTCAGAGACACCTGAGGAGCCGATACCGACGGATACACCATCAGATGAACCG

      T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  E
3121 ACACCGTCAGACGAGCCGACACCATCTGACGAACCAACACCGTCAGACGAGCCAACGCCATCTGACGAACCGACACCGTCTGATGAGCCAACACCATCTGATGAACCGACTCCGTCAGAG

      T  P  E  E  P  I  P  T  D  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T
3241 ACACCTGAGGAGCCGATACCGACGGATACACCATCAGATGAACCGACACCGTCAGACGAGCCGACACCATCTGACGAACCAACACCGTCAGACGAGCCAACGCCATCTGACGAACCGACA

      P  S  D  E  P  T  P  S  D  E  P  T  P  S  E  T  P  E  E  P  I  P  T  D  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D
3361 CCGTCTGATGAGCCAACACCATCTGATGAACCGACTCCGTCAGAGACACCTGAGGAGCCGATACCGACGGATACACCATCAGATGAACCGACACCGTCAGACGAGCCGACACCATCTGAC

                                                         TPSDEP repeats
      E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  E  T  P  E  E  P  I  P  T  D  T  P  S  D  E  P  T  P  S  D  E  P  T
3481 GAACCAACACCGTCTGATGAGCCAACACCGTCAGATGAACCGACTCCGTCAGAGACACCTGAGGAGCCGATACCGACGGATACACCATCAGATGAACCGACACCGTCAGACGAGCCAACG

      P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  E  T  P  E  E  P  I  P  T  D  T  P  S  D  E  P  T  P  S  D
3601 CCATCTGACGAACCGACACCGTCTGATGAGCCAACACCGTCAGATGAACCGACTCCGTCAGAGACACCTGAGGAGCCGATACCGACGGATACACCATCAGATGAACCGACACCGTCAGAC

      E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  E  T  P  E  E  P  I  P  T  D  T  P  S  D  E  P  T
3721 GAGCCAACGCCATCTGACGAACCGACACCGTCTGATGAGCCAACACCGTCAGATGAACCGACTCCGTCAGAGACACCTGAGGAGCCGATACCGACGGATACACCATCAGATGAACCGACA
```

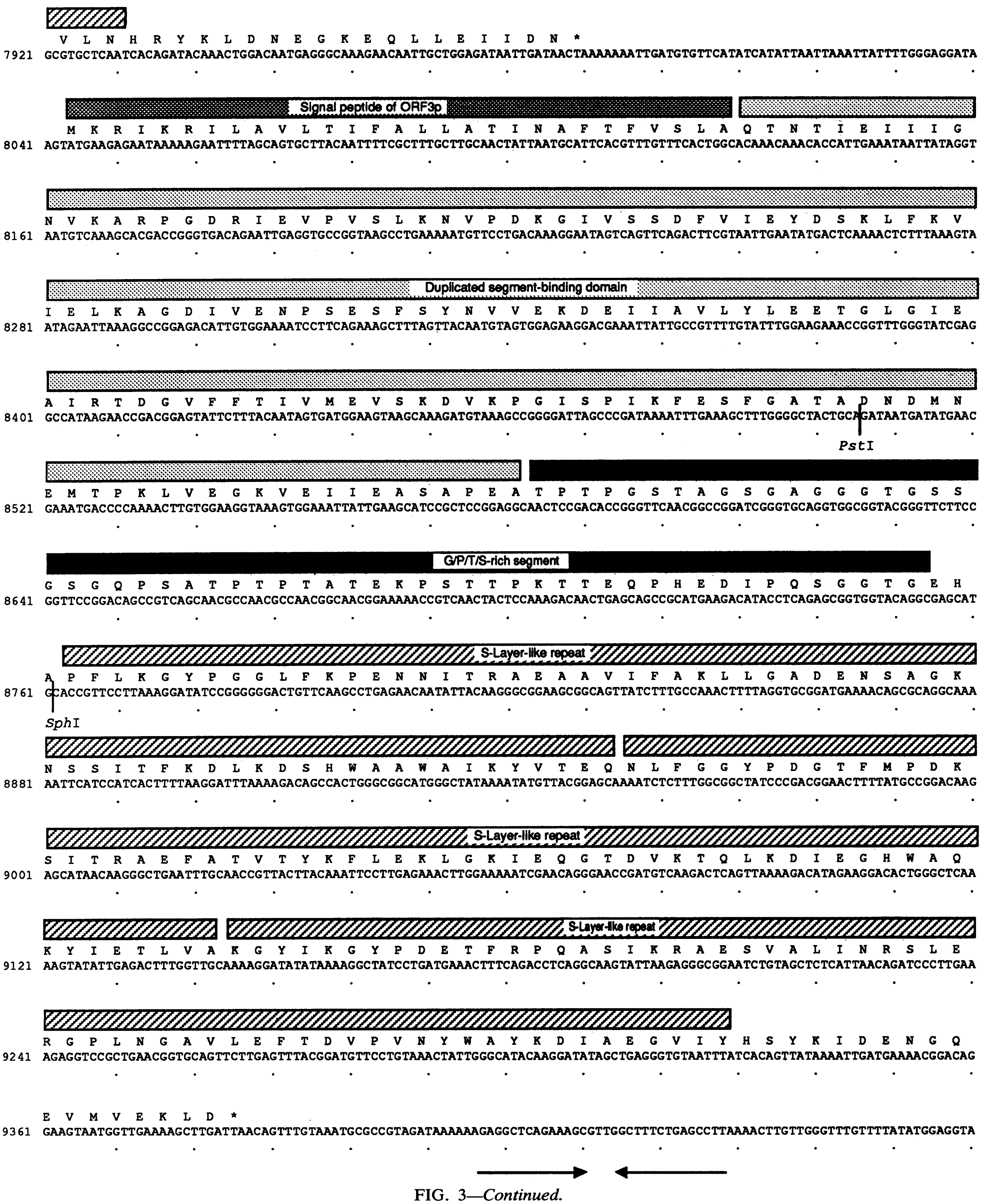
FIG. 3—*Continued.*

```
     P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  E  T  P  E  E  P  I  P  T  D  T  P  S  D
3841 CCATCAGACGAGCCAACGCCATCTGATGAACCAACACCGTCTGATGAGCCAACACCATCTGATGAACCGACTCCGTCAGAGACACCTGAGGAGCCGATACCGACGGATACACCATCAGAT

     E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  E  T  P  E  E  P  I  P  T  D  T
3961 GAACCGACACCGTCAGACGAGCCAACGCCATCTGACGAACCAACACCGTCTGATGAGCCAACACCGTCAGATGAACCGACTCCGTCAGAGACACCTGAGGAGCCGATACCGACGGATACA

     P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E
4081 CCATCAGATGAACCGACACCGTCAGACGAGCCGACACCATCTGACGAACCAACACCGTCAGACGAGCCAACGCCATCTGACGAACCGACACCGTCTGATGAGCCAACACCATCTGATGAA

     P  T  P  S  E  T  P  E  E  P  I  P  T  D  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P
4201 CCGACTCCGTCAGAGACACCTGAGGAGCCGATACCGACGGATACACCATCAGATGAACCGACACCGTCAGACGAGCCGACACCATCTGACGAACCAACACCGTCAGACGAGCCAACGCCA

     S  D  E  P  T  P  S  D  E  P  T  P  S  D  E  P  T  P  S  E  T  P  E  E  P  T  P  T  T  T  P  T  P  T  P  S  T  T  P  T
4321 TCTGACGAACCGACACCGTCTGATGAGCCAACACCATCTGATGAACCGACTCCGTCAGAGACACCTGAGGAGCCGACACCGACTACTACACCGACACCAACACCGTCGACAACGCCTACA
                                                                                                                  SalI

                                                     G/P/T/S-rich segment
     S  G  S  G  G  S  G  G  S  G  G  G  G  G  G  G  G  G  T  V  P  T  S  P  T  P  T  P  T  S  K  P  T  S  T  P  A  P  T  E
4441 AGTGGCAGCGGAGGCAGTGGTGGAAGCGGTGGTGGCGGCGGAGGTGGTGGAGGAACTGTACCTACATCTCCAACACCGACACCGACATCTAAACCGACGTCTACACCTGCACCGACAGAA

     I  E  E  P  T  P  S  D  V  P  G  A  I  G  G  E  H  R  A  Y  L  R  G  Y  P  D  G  S  F  R  P  E  R  N  I  T  R  A  E  A
4561 ATCGAAGAGCCTACACCATCTGATGTGCCTGGTGCAATCGGTGGAGAACATAGAGCATACTTAAGAGGATATCCGGATGGAAGCTTCAGGCCTGAAAGAAATATAACAAGAGCTGAAGCG

                                                     S-Layer-like repeat
     A  V  I  F  A  K  L  L  G  A  D  E  S  Y  G  A  Q  S  A  S  P  Y  S  D  L  A  D  T  H  W  A  A  W  A  I  K  F  A  T  S
4681 GCGGTAATCTTTGCTAAGTTGCTTGGAGCCGATGAAAGCTATGGAGCTCAGTCTGCAAGTCCATATAGTGATTTGGCTGATACTCACTGGGCTGCATGGGCAATCAAATTTGCAACAAGC
                                                 SacI

                                                     S-Layer-like repeat
     Q  G  L  F  K  G  Y  P  D  G  T  F  K  P  D  Q  N  I  T  R  A  E  F  A  T  V  V  L  H  F  L  T  K  V  K  G  Q  E  I  M
4801 CAGGGCTTGTTCAAAGGATATCCGGACGGTACGTTTAAACCTGATCAGAACATAACGAGAGCGGAATTCGCAACTGTGGTACTCCACTTCCTGACAAAAGTTAAGGGTCAGGAAATAATG
                                                                EcoRI

     S  K  L  A  T  I  D  I  S  N  P  K  F  D  D  C  V  G  H  W  A  Q  E  F  I  E  K  L  T  S  L  G  Y  I  S  G  Y  P  D  G
4921 AGCAAGCTTGCAACAATAGATATAAGTAATCCGAAGTTTGACGATTGTGTCGGACATTGGGCACAAGAGTTTATTGAGAAATTGACAAGCTTGGGTTATATTAGTGGCTATCCTGACGGA

                                                     S-Layer-like repeat
     T  F  K  P  Q  N  Y  I  K  R  S  E  S  V  A  L  I  N  R  A  L  E  R  G  P  L  N  G  A  P  K  L  F  P  D  V  N  E  S  Y
5041 ACGTTCAAGCCGCAAAACTATATTAAACGTTCCGAAAGTGTGGCACTGATTAACAGAGCTCTGGAGAGAGGTCCGCTTAATGGAGCGCCGAAGCTCTTCCCGGATGTTAACGAATCATAC
                                                             SacI

     W  A  F  G  D  I  M  D  G  A  L  D  H  S  Y  I  I  E  D  E  K  E  K  F  V  K  L  L  E  D  *
5161 TGGGCATTTGGCGACATTATGGACGGTGCTCTCGACCACAGTTACATTATCGAAGATGAGAAAGAAAAATTCGTTAAATTGCTCGAAGATTAATGTTGTGTAAGATAAATTGAAATGATA

5281 TGTCAAAAGCTGTTGCGTTTAGTTCTTTGACGCAGCAGCTTTTTATATTTGAAAGTAAAAAGCAAACAAGATTCAGATTTCCTGAAAAACAACAGATTATTCGCAAAATGACTTGAAAC
        ------------->          <-------------

5401 ATTGAAAAAACTAATTTATAATAAAAATATAACGGATATCCAATTCAAGTGCTTATAGAGACTGAAATATGTTCCAGTGGTTACTGTAATCGATTTATTTTTCTGGGGAGGAATTATTGT

5521 GGTGTGTCATTTTTAAAAATCATATCCTTTCTTGTTTAAAAAGATAACGCTTGAAAAATCATTCAGAGACGTTTGTAATGTACCAGCTCTTATCCCCGCTATTGCCACAGTCTATTCACT

5641 CCAATTGAAAAATTCAAAACGTTCGCTGTAAGTATAACGGTAGATTAATTAAAAGGTGTAATTGGGATAACATAACCCTAAGGTAGTTTTGCTGTTTTAAAGATTAAATACGACAATAGT

5761 GTTAGCAACTGTTTTTGAGCGTGATAACAATAAAGTTTGCTGTTTTTTATAAAAAACAATAAATAACATAAAAAATTAATATTGGGGGTGTTAAATGGATACATAAACATAAAAATAAAT

                                                                                Signal peptide of ORF2p
                                               M  K  K  N  N  V  L  T  I  A  A  M  I  A  L  L  L  T  S  L  L  T  S  I  T  F
5881 GGACAACAAATAAAACTATTATATGATATATTTGGGGGGTTTATGAAAAAAAACAATGTATTAACAATAGCAGCTATGATAGCGCTTCTTCTAACCAGCTTACTTACAAGTATAACTT
```

FIG. 3—*Continued.*

```
      G  E  T  S  S  I  P  S  R  I  S  M  E  L  D  K  T  K  A  N  I  G  D  I  I  I  A  T  I  R  I  D  N  I  N  N  F  S  G  Y
6001  TTGGGGAGACTTCGAGTATACCTTCAAGAATATCTATGGAGCTTGACAAGACAAAAGCAAACATAGGCGACATAATTATAGCCACAATAAGAATTGACAATATCAATAACTTTAGCGGAT

      Q  L  N  I  K  Y  D  P  S  Y  L  Q  A  V  N  P  L  T  G  E  P  I  K  K  R  T  M  P  A  V  N  G  T  V  L  L  K  G  D  Q
6121  ATCAATTAAATATAAAGTATGATCCGTCATACCTCCAGGCAGTTAATCCTTTGACAGGAGAACCGATAAAAAAGAGAACAATGCCGGCAGTGAACGGCACGGTGTTGTTAAAGGAGATC

                                            Unknown repeat
      Y  S  I  T  E  V  V  E  N  N  V  D  E  G  I  L  N  F  G  K  G  Y  A  N  L  T  E  Y  R  K  S  G  K  P  E  T  T  G  I  I
6241  AGTACAGTATTACTGAGGTTGTAGAAAATAACGTCGATGAAGGGATTTTAAATTTTGGCAAGGGATATGCAAATTTAACTGAATACAGGAAAAGCGGAAAACCTGAAACAACCGGAATTA

      G  K  I  G  F  K  A  L  K  L  G  K  T  E  I  K  F  E  N  T  P  V  M  P  G  A  K  E  G  T  L  L  F  D  W  D  A  E  T  I
6361  TTGGCAAGATAGGATTTAAAGCCTTAAAGCTTGGCAAGACGGAGATCAAATTTGAGAACACACCCGTCATGCCTGGGGCAAAAGAAGGAACACTGCTGTTTGACTGGGATGCAGAAACTA

      T  E  Y  N  V  I  Q  P  K  E  L  A  I  T  L  P  D  D  A  H  I  A  L  E  L  D  K  T  K  V  K  V  G  D  V  I  V  A  T  V
6481  TAACGGAATATAATGTAATTCAGCCTAAAGAACTTGCAATAACGTTACCGGACGATGCACACATTGCTTTGGAACTTGACAAGACAAAAGTGAAAGTGGGAGATGTAATTGTTGCGACAG

      K  A  K  N  M  T  S  M  A  G  I  Q  V  N  I  K  Y  D  P  E  V  L  Q  A  I  D  P  A  T  G  K  P  F  T  K  E  T  L  L  V
6601  TAAAAGCAAAGAATATGACTAGTATGGCGGGAATTCAGGTAAATATTAAATATGACCCTGAAGTATTGCAGGCGATTGATCCTGCGACGGGAAAACCGTTTACAAAAGAAACATTACTTG
                                     EcoRI
                                            Unknown repeat
      D  P  E  L  L  S  N  R  E  Y  N  P  L  L  T  A  V  N  D  I  N  S  G  I  I  N  Y  A  S  C  Y  V  Y  W  D  S  Y  R  E  S
6721  TGGACCCGGAACTGTTATCAAACAGAGAATATAATCCGTTGTTAACAGCAGTTAATGACATAAATTCCGGCATTATAAATTATGCATCTTGTTATGTATATTGGGATTCCTACAGAGAAT

      G  V  S  E  S  T  G  I  I  G  K  V  G  F  K  V  L  K  A  A  N  T  T  V  K  L  E  E  T  R  F  T  P  N  S  I  D  G  T  L
6841  CAGGAGTATCTGAAAGCACCGGAATAATTGGAAAGGTTGGCTTTAAAGTGCTGAAAGCTGCCAACACCACAGTAAAACTGGAAGAAACAAGATTTACACCAAATTCGATAGACGGTACTT

      V  I  D  W  Y  G  Q  Q  I  V  G  Y  K  V  I  Q  P  D  K  I  T  V  I  S  E  P  E  V  P  T  Q  T  P  T  Q  T  P  P  T  T
6961  TGGTAATTGATTGGTATGGCCAACAGATAGTTGGTTATAAAGTAATACAGCCCGACAAAATTACTGTGATTTCAGAGCCTGAGGTACCAACACAAACACCTACACAGACACCGCCAACAA
                                                                                                  KpnI
                                              G/P/T/S-rich segment
      T  A  P  S  Q  T  P  T  Q  T  P  P  T  T  T  A  P  S  Q  T  P  T  Q  T  P  A  V  T  P  T  Q  S  A  T  P  S  D  P  G  G
7081  CAACAGCACCATCGCAAACACCTACGCAGACACCGCCAACAACAACAGCACCATCACAGACACCTACACAGACACCGGCAGTAACGCCGACGCAAAGTGCAACTCCGTCGGATCCTGGCG
                                                                                                                        BamHI

      G  G  G  G  L  P  G  G  G  G  G  A  V  N  P  S  A  S  P  T  P  T  P  T  S  K  P  T  P  T  A  T  K  K  P  E  P  T  E  I
7201  GAGGTGGAGGAGGCCTCCCGGGTGGTGGAGGCGGCGCTGTTAATCCTTCAGCTTCACCGACACCAACACCGACATCCAAACCTACTCCTACTGCCACTAAAAAACCGGAGCCAACGGAAA
                     SmaI

      E  E  P  E  P  E  I  P  G  T  V  G  I  H  Y  S  Y  L  T  G  Y  P  D  K  M  F  R  P  E  K  S  I  T  R  A  E  A  A  V  I
7321  TAGAAGAACCCGAACCTGAAATACCGGGCACTGTTGGAATACATTATTCATACCTGACAGGTTATCCGGACAAAATGTTCAGACCTGAAAAGAGTATTACAAGAGCTGAAGCAGCCGTGA

                                            S-Layer-like repeat
      F  A  K  L  L  G  A  N  E  N  T  K  I  N  Y  N  V  S  Y  T  D  V  D  S  S  H  W  A  S  W  A  I  K  F  V  S  Y  K  K  L
7441  TTTTTGCAAAACTTTTGGGAGCAAACGAAAATACAAAGATAAACTATAATGTTTCATACACCGATGTTGACAGCTCCCATTGGGCAAGTTGGGCAATCAAATTTGTATCATACAAGAAAC

                                            S-Layer-like repeat
      F  T  G  Y  P  D  G  S  F  K  P  N  Q  N  I  T  R  A  E  F  S  T  V  V  F  K  L  L  V  S  E  K  G  L  K  E  E  K  I  E
7561  TGTTTACCGGATATCCTGATGGCTCGTTCAAGCCTAATCAGAATATAACGAGAGCCGAATTTTCAACGGTTGTGTTTAAGCTTCTTGTATCTGAGAAAGGTCTAAAAGAAGAAAAGATTG

      K  S  K  F  G  D  T  K  G  H  W  A  Q  Q  F  I  E  Q  L  S  D  L  G  Y  I  N  G  Y  P  D  G  T  F  K  P  N  N  N  I  K
7681  AAAAGTCCAAGTTTGGTGATACAAAGGGCCACTGGGCACAACAGTTTATTGAACAGCTGTCAGACCTTGGATACATCAACGGATATCCTGATGGTACATTCAAGCCCAACAACAATATCA

                                            S-Layer-like repeat
      R  S  E  S  V  A  L  I  N  R  A  M  G  R  G  P  L  H  G  A  P  Q  V  F  E  D  V  P  Q  T  H  W  A  F  K  D  I  A  E  G
7801  AACGATCAGAAAGTGTTGCCCTGATAAACAGAGCTATGGGAAGAGGGCCTTTGCATGGCGCACCGCAGGTATTCGAGGATGTTCCTCAGACACACTGGGCTTTCAAAGATATTGCAGAGG
```

FIG. 3—*Continued.*



```
        V  L  N  H  R  Y  K  L  D  N  E  G  K  E  Q  L  L  E  I  I  D  N  *
7921 GCGTGCTCAATCACAGATACAAACTGGACAATGAGGGCAAAGAACAATTGCTGGAGATAATTGATAACTAAAAAAATTGATGTGTTCATATCATATTAATTAAATTATTTGGGAGGATA
              .         .         .         .         .         .         .         .         .         .         .         .

         M  K  R  I  K  R  I  L  A  V  L  T  I  F  A  L  L  A  T  I  N  A  F  T  F  V  S  L  A  Q  T  N  T  I  E  I  I  I  G
8041 AGTATGAAGAGAATAAAAAGAATTTTAGCAGTGCTTACAATTTTCGCTTTGCTTGCAACTATTAATGCATTCACGTTTGTTTCACTGGCACAAACAAACACCATTGAAATAATTATAGGT
              .         .         .         .         .         .         .         .         .         .         .         .

      N  V  K  A  R  P  G  D  R  I  E  V  P  V  S  L  K  N  V  P  D  K  G  I  V  S  S  D  F  V  I  E  Y  D  S  K  L  F  K  V
8161 AATGTCAAAGCACGACCGGGTGACAGAATTGAGGTGCCGGTAAGCCTGAAAAATGTTCCTGACAAAGGAATAGTCAGTTCAGACTTCGTAATTGAATATGACTCAAAACTCTTTAAAGTA
              .         .         .         .         .         .         .         .         .         .         .         .

      I  E  L  K  A  G  D  I  V  E  N  P  S  E  S  F  S  Y  N  V  V  E  K  D  E  I  I  A  V  L  Y  L  E  E  T  G  L  G  I  E
8281 ATAGAATTAAAGGCCGGAGACATTGTGGAAAATCCTTCAGAAAGCTTTAGTTACAATGTAGTGGAGAAGGACGAAATTATTGCCGTTTTGTATTTGGAAGAAACCGGTTTGGGTATCGAG
              .         .         .         .         .         .         .         .         .         .         .         .

      A  I  R  T  D  G  V  F  F  T  I  V  M  E  V  S  K  D  V  K  P  G  I  S  P  I  K  F  E  S  F  G  A  T  A  D  N  D  M  N
8401 GCCATAAGAACCGACGGAGTATTCTTTACAATAGTGATGGAAGTAAGCAAAGATGTAAAGCCGGGGATTAGCCCGATAAAATTTGAAAGCTTTGGGGCTACTGCAGATAATGATATGAAC
              .         .         .         .         .         .         .         .         .         .         .         .
                                                                                                              PstI

      E  M  T  P  K  L  V  E  G  K  V  E  I  I  E  A  S  A  P  E  A  T  P  T  P  G  S  T  A  G  S  G  A  G  G  G  T  G  S  S
8521 GAAATGACCCCAAAACTTGTGGAAGGTAAAGTGGAAATTATTGAAGCATCCGCTCCGGAGGCAACTCCGACACCGGGTTCAACGGCCGGATCGGGTGCAGGTGGCGGTACGGGTTCTTCC
              .         .         .         .         .         .         .         .         .         .         .         .

      G  S  G  Q  P  S  A  T  P  T  P  T  A  T  E  K  P  S  T  T  P  K  T  T  E  Q  P  H  E  D  I  P  Q  S  G  G  T  G  E  H
8641 GGTTCCGGACAGCCGTCAGCAACGCCAACGCCAACGGCAACGGAAAAACCGTCAACTACTCCAAAGACAACTGAGCAGCCGCATGAAGACATACCTCAGAGCGGTGGTACAGGCGAGCAT
              .         .         .         .         .         .         .         .         .         .         .         .

      A  P  F  L  K  G  Y  P  G  G  L  F  K  P  E  N  N  I  T  R  A  E  A  A  V  I  F  A  K  L  L  G  A  D  E  N  S  A  G  K
8761 GCACCGTTCCTTAAAGGATATCCGGGGGGACTGTTCAAGCCTGAGAACAATATTACAAGGGCGGAAGCGGCAGTTATCTTTGCCAAACTTTTAGGTGCGGATGAAAACAGCGCAGGCAAA
              .         .         .         .         .         .         .         .         .         .         .         .
     SphI

      N  S  S  I  T  F  K  D  L  K  D  S  H  W  A  A  W  A  I  K  Y  V  T  E  Q  N  L  F  G  G  Y  P  D  G  T  F  M  P  D  K
8881 AATTCATCCATCACTTTTAAGGATTTAAAAGACAGCCACTGGGCGGCATGGGCTATAAAATATGTTACGGAGCAAAATCTCTTTGGCGGCTATCCCGACGGAACTTTTATGCCGGACAAG
              .         .         .         .         .         .         .         .         .         .         .         .

      S  I  T  R  A  E  F  A  T  V  T  Y  K  F  L  E  K  L  G  K  I  E  Q  G  T  D  V  K  T  Q  L  K  D  I  E  G  H  W  A  Q
9001 AGCATAACAAGGGCTGAATTTGCAACCGTTACTTACAAATTCCTTGAGAAACTTGGAAAAATCGAACAGGGAACCGATGTCAAGACTCAGTTAAAAGACATAGAAGGACACTGGGCTCAA
              .         .         .         .         .         .         .         .         .         .         .         .

      K  Y  I  E  T  L  V  A  K  G  Y  I  K  G  Y  P  D  E  T  F  R  P  Q  A  S  I  K  R  A  E  S  V  A  L  I  N  R  S  L  E
9121 AAGTATATTGAGACTTTGGTTGCAAAAGGATATATAAAAGGCTATCCTGATGAAACTTTCAGACCTCAGGCAAGTATTAAGAGGGCGGAATCTGTAGCTCTCATTAACAGATCCCTTGAA
              .         .         .         .         .         .         .         .         .         .         .         .

      R  G  P  L  N  G  A  V  L  E  F  T  D  V  P  V  N  Y  W  A  Y  K  D  I  A  E  G  V  I  Y  H  S  Y  K  I  D  E  N  G  Q
9241 AGAGGTCCGCTGAACGGTGCAGTTCTTGAGTTTACGGATGTTCCTGTAAACTATTGGGCATACAAGGATATAGCTGAGGGTGTAATTTATCACAGTTATAAAATTGATGAAAACGGACAG
              .         .         .         .         .         .         .         .         .         .         .         .

      E  V  M  V  E  K  L  D  *
9361 GAAGTAATGGTTGAAAAGCTTGATTAACAGTTTGTAAATGCGCCGTAGATAAAAAAGAGGCTCAGAAAGCGTTGGCTTTCTGAGCCTTAAAACTTGTTGGGTTTGTTTTATATGGAGGTA
              .         .         .         .         .         .         .         .         .         .         .         .
```

FIG. 3—*Continued.*

determinant involved in the attachment of the catalytic components to CipA was identified as a conserved, duplicated segment of 22 residues present in all of the cellulosome components identified so far (26).

In a previous article, we reported the screening of a gene bank of *C. thermocellum* DNA for clones producing proteins to which ^{125}I-labeled endoglucanase CelD bound by means of its duplicated segment (4). Two neighboring DNA regions were found to encode such proteins. One of the regions contains part of the *cipA* gene, which has been indepen-

```
  1' DLDAVRIKVDTVNAKPGYTVRIPVRFTGIPSKGIANCDFVYSYDPNVLEIIEIEPGE  CipA
166' DLDAVRIKVDTVNAKPGDTVRIPVRFSGIPSKGIANCDFVYSYDPNVLEIIEIEPGD  CipA
332'   NKLTLKIGRAEGRPGDTVEIPVNLYGVPQKGIASGDFVVSYDPNVLEIIEIEPGE  CipA
 30  QTNTIEIIIGNVKARPGDRIEVPVSLKNVPDKGIVSSDFVIEYDSKLFKVIELKAGD  ORF3p

 58' LIVDPNPTKSFDTAVYPDRKMIVFLFAEDSGTGAYAITEDGVFATIVRKVKSGAPNG  CipA
223' IIVDPNPDKSFDTAVYPDRKIIVFLFAEDSGTGAYAITKDGVFATIVAKVKEGAPNG  CipA
387' LIVDPNPTKSFDTAVYPDRKMIVFLFAEDSGTGAYAITEDGVFATIVAKVKEGAPEG  CipA
 87  IVE--NPSESFSYNVVEKDEIIAVLYLEETGLGIEAIRTDGVFFTIVMEVSKDVKPG  ORF3p

115' LSVIKFVEVGGFANNDLVEQKTQFFDGGVNVGDTTEPAT-PTTPVTTPTTTD       CipA
280' LSVIKFVEVGGFANNDLVEQKTQFFDGGVNVGDTTVPTTSPTTTPPEPTITP       CipA
444' FSAIEISEFGAFADNDLVEVETDLINGGVLVTN                          CipA
142  ISPIKFESFGATADNDMNEMTPKLVEGKVEIIEASAPEA                    ORF3p
```

FIG. 4. Alignment of the ORF3p receptor and the three COOH-terminal receptors of CipA responsible for binding of the duplicated segment. Residues that are identical or similar in the largest number of CipA receptors displayed are shown against a shaded background. Numbering of CipA residues is arbitrary and starts with the first residue of the partial sequence published previously (4); numbering of ORF3p residues starts with the putative initiation codon. Similar amino acids are: F, I, V, L, and M; R and K; S and T; D and E; N and Q; and F, Y, and W.

dently cloned and sequenced by another group, who used reactivity with anti-CipA antibodies as a screening test (4a, 21). The polypeptide sequence of CipA comprises several domains, each of about 146 residues, separated by Pro/Thr-rich segments of 17 to 19 residues. These domains appear to correspond to binding sites for the duplicated segment borne by the catalytic components. The protein also comprises a duplicated segment, which, although similar to that found in the catalytic components, is less well conserved (4) (Fig. 1). Another, previously unidentified region located some 8 to 9 kb downstream from *cipA* was found to encode a second polypeptide, which has affinity for ^{125}I-labeled CelD (4).

This article analyzes the nucleotide sequence and transcriptional organization of the corresponding gene, termed open reading frame 3 (ORF3), and of the genes lying between *cipA* and ORF3. From the similarity between the C-terminal region of the ORF3p polypeptide encoded by ORF3 and the N-terminal region of known S-layer proteins, we speculate that ORF3p might serve to anchor the cellulosome to the surface of *C. thermocellum* cells.

MATERIALS AND METHODS

DNA sequencing. Subclones were generated from the fragments of the genes cloned in λ-GEM-11 (4) by cloning appropriate restriction fragments in pTZ19U or pTZ19R (15). Nested deletions were created with exonuclease III (7) and mung bean nuclease. Single-stranded DNA templates were sequenced with the Taquence kit (United States Biochemicals, Cleveland, Ohio) as directed by the supplier. The entire sequence was determined at least once on each strand, and the sequence around each restriction site used for subcloning was checked by at least one overlapping gel reading. For regions containing highly repetitive sequences, the start point of each deletion was ascertained to within 0.1 kb by restriction mapping. The restriction map of cloned DNA fragments was consistent with the sizes of fragments of *C. thermocellum* DNA detected by Southern blotting (23), indicating the absence of artifacts in the cloning and sequencing processes.

Analysis of mRNAs. A *C. thermocellum* preculture was grown at 60°C to an optical density at 600 nm of 1.1 in CM3-3 medium containing 5 g of cellobiose per liter (24) and used to inoculate a 10-fold-larger volume of CM3-3 medium containing 10 g of cellulose per liter. The culture was incubated for 18 h at 60°C (i.e., until most of the cellulose had been hydrolyzed). Cells from 250 ml of culture were harvested, and RNA was extracted as described before (16). Northern (RNA) blotting onto nitrocellulose was performed as described before (25). DNA fragments a, b, c, and d, derived from *cipA*, ORF1, ORF2, and ORF3, respectively (Fig. 1), were obtained by electroelution after digesting appropriate subclones with *Sca*I and *Sna*BI (a), *Hpa*I and *Sma*I (b), *Eco*RI and *Kpn*I (c), or *Nsi*I and *Pst*I (d). They were labeled with ^{32}P with the Boehringer random primer labeling kit and used as hybridization probes. The GIBCO-BRL kit of RNA molecular weight standards was used to estimate the size of mRNAs.

RESULTS AND DISCUSSION

General organization and transcription of the genes. The general organization of the genes within the 14-kb region including the 3′ end of *cipA* is shown in Fig. 1. Three ORFs, encoding polypeptides of 1,664, 688, and 447 residues, were identified downstream from *cipA*. Previous deletion mapping

```
1453 AYLRGYPDGSFRPERNITRAEAAVIFAKLLGADESYGAQSAS--------PYSDLADTHWAAWAIKFATSQ ORF1p
1516 GLFKGYPDGTFKPDQNITRAEFATVVLHFLTKVKGQEIMSKLATIDISNPKFDDCV-GHWAQEFIEKLTSL ORF1p
1586 GYISGYPDGTFKPQNYIKRSESVALINRALERGPLNGAPK----------LFPDVNESYWAFGDIMDGALD ORF1p
 482 SYLTGYPDKMFRPEKSITRAEAAVIFAKLLGANENTKINYNV--------SYTDVDSSHWASWAIKFVSYK ORF2p
 545 KLFTGYPDGSFKPNQNITRAEFSTVVFKLLVSEKGLKEEKI------EKSKFGDTK-GHWAQQFIEQLSDL ORF2p
 609 GYINGYPDGTFKPNNNIKRSESVALINRAMGRGPLHGAPQ----------VFEDVPQTHWAFKDIAEGVLN ORF2p
 241 PFLKGYPGGLFKPENNITRAEAAVIFAKLLGADENSAGKNS-------SITFKDLKDSHWAAWAIKYVTEQ ORF3p
 305 NLFGGYPDGTFMPDKSITRAEFATVTYKFLEKLGKIEQGTD------VKTQLKDIE-GHWAQKYIETLVAK ORF3p
 368 GYIKGYPDETFRPQASIKRAESVALINRSLERGPLNGAVL----------EFTDVPVNYWAYKDIAEGVIY ORF3p
  50 GLVAGYGNGEYGVDKTITRAEFATLVVRARGLEQGAKLAQF------ SNTYTDVKSTDWFAGFVNVASGE MWP
 114 EIVKGFPDKSFKPQNQVTYAEAVTMIVRALGYE                                       MWP
  51 NITNGVGDPKFGVDQPVTRAQMITFVNRMLGYEDLAEMAKS------EKSAFKDVPQNHWAVGQINLAYKL A.k.
 116 GLAQGVGNGKFDPNSELRYAQALAFVLRALGFK                                       A.k.
```

FIG. 5. Alignment of the COOH-terminal repeats found in ORF1p, ORF2p, and ORF3p with the sequences of *B. brevis* 47 (29) and *A. kivui* (20) S-layer proteins. MWP, middle wall protein of *B. brevis*; A. k., S-layer protein of *A. kivui*. For each protein, numbering starts at the putative initiation codon. Residues that are similar or identical in at least five of the *C. thermocellum* segments and at least one of the *B. brevis* or *A. kivui* segments are shown against a shaded background. Similarity criteria are the same as for Fig. 4.

(4) had indicated that a portion of DNA lying to the right of the rightmost *Pst*I site was required to encode a polypeptide capable of binding ^{125}I-CelD. Accordingly, this polypeptide must be encoded by the rightmost ORF. Figure 2 shows that a 6.3-kb transcript was detected for *cipA*, which agrees with the size of the *cipA* gene (5.5 kb) (4a). A palindromic sequence, which may act as a transcription attenuator, was found downstream from *cipA*. The ORF1 transcript was about 12 to 14 kb, sufficiently large to include the sequence of *cipA* and ORF1. Possibly *cipA* and ORF1 mRNAs start from the same promoter, and ORF1 transcription results from readthrough past the transcriptional attenuator that accounts for the monocistronic, 6.3-kb *cipA* mRNA. Because of the high background observed in the high-M_r region for the *cipA* hybridization, a band of 12 to 14 kb would have escaped detection. For both ORF2 and ORF3, a 4-kb transcript was observed, compatible with cotranscription of the two genes.

Analysis of the polypeptide sequences encoded by ORF1, ORF2, and ORF3. The nucleotide sequence (EMBL accession number X67506) and the deduced polypeptide sequences of the *cipA*-ORF3 region are shown in Fig. 3. The three ORFs encode polypeptides starting with typical signal peptides. Furthermore, the COOH-terminal regions of the three polypeptides are highly conserved and consist of three similar segments of 60 to 70 residues each. These repeats, termed S-layer-like repeats in Fig. 1 and 3, display significant similarity to the NH$_2$-terminal regions of the S-layer proteins of *Bacillus brevis* 47 (29) and *Acetogenium kivui* (20), but in these organisms, there are one and a half copies of the basic motif rather than three copies (Fig. 5). In the three proteins, the COOH-terminal repeats are separated from the rest of the protein by stretches of 57 to 107 residues containing many Gly, Pro, Thr, and Ser residues (G/P/T/S-rich segment in Fig. 1 and 3).

Besides the S-layer-like repeats, the ORF1p polypeptide encoded by ORF1 is composed of various kinds of reiterated elements. The N-terminal region comprises four highly similar segments of 156 residues each. Two copies of this segment are also present in the ORF2p polypeptide encoded by ORF2, but no significant similarity with other recorded polypeptide sequences was found in the National Biomedical Research Foundation sequence data base with the FASTP algorithm (13). It is therefore designated unknown repeat in Fig. 1 and 3. The central region (TPSDEP repeats in Fig. 1 and 3) of ORF1p is extremely repetitive, so that the sequence of a stretch of 607 residues can be written as $(AB_5C)_4(AB_7C)_2(AB_5C)_5(AB_7C)_2$, where A is EPIPTD, B is TPSDEP, and C is TPSETPE. The TPSDEP repeats are separated by a G/P/T/S-rich segment from three COOH-terminal copies of the S-layer-like motif.

ORF2p is similar to ORF1p except that it contains only two copies of the unknown repeat and does not contain the TPSDEP repeats.

The NH$_2$-terminal region of ORF3p is highly similar to the domains identified in CipA as receptors responsible for binding the duplicated segment borne by CelD and other catalytic components of the cellulosome (4) (Fig. 4). However, the sequence of the ORF3p receptor appears to be more divergent from the consensus than any of the three CipA receptors known to date (4). As observed for ORF1p and ORF2p, the COOH-terminal region of ORF3p contains three S-layer-like repeats preceded by a G/P/T/S-rich segment.

Putative role of polypeptides ORF1p, ORF2p, and ORF3p. The presence of S-layer-like domains in ORF1p, ORF2p, and ORF3p suggests that these polypeptides might be located on the surface of *C. thermocellum* cells. This hypothesis is consistent with the fact that the sequences of the three polypeptides start with signal peptides. Furthermore, a G/P/T/S-rich region is also found in streptococcal M protein (8), which is also a bacterial surface protein.

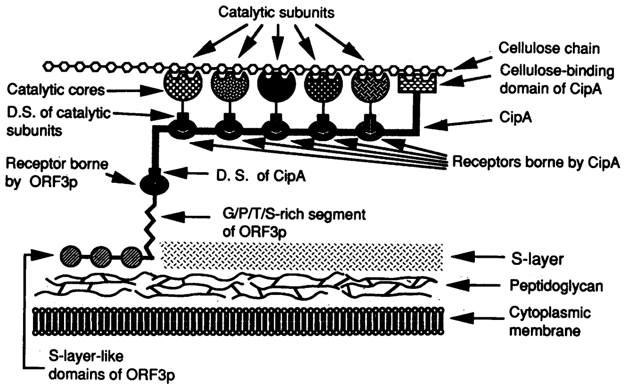


FIG. 6. Hypothetical model showing how ORF3p might mediate the attachment of the cellulosome to the cell surface of *C. thermocellum*. The scheme for the organization of the cellulosome itself is derived from evidence showing that CipA mediates binding to cellulose of the catalytic S_S subunit (27), that catalytic subunits bind to the repeated domains (receptors) of CipA by means of their duplicated segments (4, 26), and that CipA comprises a cellulose-binding domain (18, 22). The cores of the catalytic subunits are assumed to be poised along a cellulose chain, enabling quasi-simultaneous, multiple cutting events, as proposed by Mayer et al. (14). D.S., 22-amino-acid duplicated segment.

According to Lamed and Bayer (10), the surface of *C. thermocellum* is anionic, since it binds cationized ferritin. The presence of the TPSDEP repeats in ORF1p would certainly be consistent with an overall negative charge. Other examples of highly repetitive sequences are to be found among bacterial cell surface proteins, such as the streptococcal M protein (8) and the *Pseudomonas syringae* ice nucleation protein (5).

Previous results have shown that ORF3p can bind the duplicated, conserved segment that is responsible for anchoring catalytic components to the CipA subunit of the cellulosome (4). Not surprisingly, the sequence of ORF3p comprises a region that is similar to the CipA segments previously identified as receptors for the duplicated segment. However, the presence of a single receptor on ORF3p argues against its being a cellulosome scaffolding protein with a role similar to that of CipA. If, as suggested above, ORF3p is indeed a cell surface protein, it is tempting to speculate that its role may be to anchor the cellulosome to the cell surface. Indeed, although the duplicated segment present at the COOH terminus of CipA is similar to that found in the catalytic components, it is nonetheless more divergent from the consensus than any of the duplicated segments identified in the catalytic subunits. The ORF3p receptor is also the most divergent. A possible reason might be that it interacts preferentially with the duplicated segment of CipA rather than with those of the catalytic components. A diagram summarizing this hypothesis and extending previous models of the cellulosome (2, 14, 27) is presented in Fig. 6.

Although largely based on sequence analysis, the model provides a basis for experimental testing, e.g., by purifying the various proteins produced from cloned genes, quantifying their mutual affinities in vitro, and generating antibodies to check whether the corresponding antigens are indeed present on the cell surface.

ACKNOWLEDGMENTS

We thank U. T. Gerngross and A. L. Demain for communicating results prior to publication and M. P. Coughlan for helpful discussions. The expert technical help of O. Raynaud is acknowledged.

This work was supported by grant CPL C462 from the Commission of the European Communities and by research funds from the University of Paris 7.

REFERENCES

1. **Béguin, P.** 1983. Detection of cellulase activity in polyacrylamide gels using Congo red-stained agar replicas. Anal. Biochem. **131:**333–336.
2. **Béguin, P., J. Millet, and J.-P. Aubert.** 1992. Cellulose degradation by *Clostridium thermocellum*: from manure to molecular biology. FEMS Microbiol. Lett. **100:**523–528.
3. **Coughlan, M. P., K. Hon-Nami, H. Hon-Nami, L. G. Ljungdahl, J. J. Paulin, and W. E. Rigsby.** 1985. The cellulolytic enzyme complex of *Clostridium thermocellum* is very large. Biochem. Biophys. Res. Commun. **130:**904–909.

3a. **Demain, A. L.** Personal communication.

4. **Fujino, T., P. Béguin, and J.-P. Aubert.** 1992. Cloning of a *Clostridium thermocellum* DNA fragment encoding polypeptides that bind the catalytic components of the cellulosome. FEMS Microbiol. Lett. **94:**165–170.

4a. **Gerngross, U. T., and A. L. Demain.** Personal communication.

5. **Green, R. L., and G. J. Warren.** 1985. Physical and functional repetition in a bacterial ice nucleation gene. Nature (London) **317:**645–648.
6. **Grépinet, O., M.-C. Chebrou, and P. Béguin.** 1988. Purification of *Clostridium thermocellum* xylanase Z expressed in *Escherichia coli* and identification of the corresponding product in the culture medium of *C. thermocellum*. J. Bacteriol. **170:**4576–4581.
7. **Henikoff, S.** 1984. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene **28:**351–359.
8. **Hollingshead, S., V. A. Fischetti, and J. R. Scott.** 1986. Complete nucleotide sequence of type 6 M protein of the group A *Streptococcus*. J. Biol. Chem. **261:**1677–1686.
9. **Kohring, S., J. Wiegel, and F. Mayer.** 1990. Subunit composition and glycosidic activities of the cellulase complex from *Clostridium thermocellum* JW20. Appl. Environ. Microbiol. **56:**3798–3804.
10. **Lamed, R., and E. A. Bayer.** 1988. The cellulosome concept: exocellular/extracellular enzyme reactor centers for efficient binding and cellulolysis, p. 101–116. *In* J.-P. Aubert, P. Béguin, and J. Millet (ed.), FEMS Symposium No. 43: Biochemistry and genetics of cellulose degradation. Academic Press, Inc., New York.
11. **Lamed, R., E. Setter, and E. A. Bayer.** 1983. Characterization of a cellulose-binding, cellulase-containing complex in *Clostridium thermocellum*. J. Bacteriol. **156:**828–836.
12. **Lamed, R., E. Setter, R. Kenig, and E. A. Bayer.** 1983. The cellulosome: a discrete cell surface organelle of *Clostridium thermocellum* which exhibits separate antigenic, cellulose-binding and various cellulolytic activities. Biotechnol. Bioeng. Symp. **13:**163–181.
13. **Lipman, D. J., and W. R. Pearson.** 1985. Rapid and sensitive protein similarity searches. Science **227:**1435–1441.
14. **Mayer, F., M. P. Coughlan, Y. Mori, and L. G. Ljungdahl.** 1987. Macromolecular organization of the cellulolytic enzyme complex of *Clostridium thermocellum* as revealed by electron microscopy. Appl. Environ. Microbiol. **53:**2785–2792.
15. **Mead, D. A., E. Szczesna-Skorupa, and B. Kemper.** 1986. Single-stranded DNA "blue" T7 promoter plasmids: a versatile tandem promoter system for cloning and protein engineering. Protein Eng. **1:**67–74.
16. **Mishra, S., P. Béguin, and J.-P. Aubert.** 1991. Transcription of *Clostridium thermocellum* endoglucanase genes *celF* and *celD*. J. Bacteriol. **173:**80–85.
17. **Morag, E., E. A. Bayer, and R. Lamed.** 1990. Relationship of cellulosomal and noncellulosomal xylanases of *Clostridium thermocellum* to cellulose-degrading enzymes. J. Bacteriol. **172:**6098–6105.
18. **Morag, E., E. A. Bayer, and R. Lamed.** 1992. Unorthodox intra-subunit interactions in the cellulosome of *Clostridium thermocellum*: identification of structural transitions induced in the S1 subunit. Appl. Biochem. Biotechnol. **33:**205–217.
19. **Morag, E., I. Halevy, E. A. Bayer, and R. Lamed.** 1991. Isolation and properties of a major cellobiohydrolase from the cellulosome of *Clostridium thermocellum*. J. Bacteriol. **173:**4155–4162.
20. **Peters, J., M. Peters, F. Lottspeich, and W. Baumeister.** 1989. S-layer protein gene of *Acetogenium kivui*: cloning and expression in *Escherichia coli* and determination of the nucleotide sequence. J. Bacteriol. **171:**6307–6315.
21. **Romaniec, M. P. M., T. Kobayashi, U. Fauth, U. T. Gerngross, and A. L. Demain.** 1991. Cloning and expression of a *Clostridium thermocellum* DNA fragment that encodes a protein related to cellulosome component S_L. Appl. Biochem. Biotechnol. **31:**119–134.
22. **Salamitou, S., K. Tokatlidis, P. Béguin, and J.-P. Aubert.** 1992. Involvement of separate domains of the cellulosomal protein S1 of *Clostridium thermocellum* in binding to cellulose and in anchoring of catalytic subunits to the cellulosome. FEBS Lett. **304:**89–92.
23. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98:**503–517.
24. **Tailliez, P., H. Girard, J. Millet, and P. Béguin.** 1989. Enhanced cellulose fermentation by an asporogenous and ethanol-tolerant mutant of *Clostridium thermocellum*. Appl. Environ. Microbiol. **55:**207–211.
25. **Thomas, P.** 1983. Hybridization of denatured RNA transferred or dotted to nitrocellulose paper. Methods Enzymol. **100:**255–256.
26. **Tokatlidis, K., S. Salamitou, P. Béguin, P. Dhurjati, and J.-P. Aubert.** 1991. Interaction of the duplicated segment carried by *Clostridium thermocellum* cellulases with cellulosome components. FEBS Lett. **291:**185–188.
27. **Wu, J. H. D., and A. L. Demain.** 1988. Proteins of the *Clostridium thermocellum* cellulase complex responsible for degradation of crystalline cellulose, p. 117–131. *In* J.-P. Aubert, P. Béguin, and J. Millet (ed.), FEMS Symposium No. 43: Biochemistry and genetics of cellulose degradation. Academic Press, Inc., New York.
28. **Wu, J. H. D., W. H. Orme-Johnson, and A. L. Demain.** 1988. Two components of an extracellular protein aggregate of *Clostridium thermocellum* together degrade crystalline cellulose. Biochemistry **27:**1703–1709.
29. **Yamagata, H., T. Adachi, A. Tsuboi, M. Takao, T. Sasaki, N. Tsukagoshi, and S. Udaka.** 1987. Cloning and characterization of the 5′ region of the cell wall protein gene operon in *Bacillus brevis* 47. J. Bacteriol. **169:**1239–1245.