INFECTION AND IMMUNITY, Feb. 1990, p. 480–488
0019-9567/90/020480-09$02.00/0

Vol. 58, No. 2

Molecular Characterization of the *Clostridium difficile* Toxin A Gene

C. H. DOVE, S.-Z. WANG, S. B. PRICE,† C. J. PHELPS, D. M. LYERLY, T. D. WILKINS, AND J. L. JOHNSON*

Department of Anaerobic Microbiology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

Received 10 July 1989/Accepted 21 October 1989

The gene encoding the toxin A protein of *Clostridium difficile* (strain VPI 10463) was cloned and sequenced. The coding region of 8,133 base pairs has a mol% G+C of 26.9 and encodes 2,710 amino acids. The deduced polypeptide has a molecular mass of ca. 308 kilodaltons. Nearly a third of the gene, at the 3′ end, consists of 38 repeating sequences. The repeating units were grouped into two classes, I and II, on the basis of length and the low levels of DNA sequence similarities between them. There were seven class I repeating units, each containing 90 nucleotides, and 31 class II units, which, with two exceptions, were either 60 or 63 nucleotides in length. On the basis of DNA sequence similarities, the class II repeating units were further segregated into subclasses: 7 class IIA, 13 class IIB, 5 class IIC, and 6 class IID. The dipeptide tyrosine-phenylalanine was found in all 38 repeating units, and other amino acid sequences were unique to a specific class or subclass. This region of the protein has epitopes for the monoclonal antibody PCG-4 and includes the binding region for the Galα1-3Galβ1-4GlcNAc carbohydrate receptor. Located 1,350 base pairs upstream from the toxin A translation start site is the 3′ end of the toxin B gene. Between the two toxin genes is a small open reading frame, which encodes a deduced polypeptide of ca. 16 or 19 kilodaltons. The role of this open reading frame is unknown.

Clostridium difficile is the major causative agent of pseudomembranous colitis in humans (2). The organism produces two toxins, designated toxin A and toxin B (1, 32, 33). They are both cytotoxic and lethal for animals, although toxin B is about 1,000-fold more cytotoxic than toxin A for most cell lines. Both toxins appear to be produced in all toxigenic strains; however, the toxicity of strains may vary by several orders of magnitude (20, 32). The actions of these toxins appear to be quite complex and at present are not understood. Although toxin A has a direct toxic effect on the intestinal mucosa, toxin B does not cause a significant response when given intragastrically to hamsters, unless it is initially mixed with a small amount of toxin A (19). Alternatively, toxin B is also toxic if it is given to hamsters with bruised (injured) ceca. The results are consistent with the initial binding and primary tissue damage being caused by toxin A or by mechanical injury, followed by the entry of toxin B.

Investigators in a number of laboratories have worked on the isolation and physical properties of the toxins (1, 2, 27, 30, 32). Both toxins have been purified to homogeneity (18, 32). An interesting and controversial property of the toxins has been their molecular weights. Initial molecular weight estimations obtained by using native proteins have ranged from 440,000 to 600,000 for toxin A and 360,000 to 500,000 for toxin B (1, 30, 32). However, in later studies there has been controversy as to whether the toxins dissociate into smaller subunits under denaturing conditions. Under these conditions, size estimations for the toxins range from 300,000 (18) to 50,000 (27, 30) and down to 42,500 and 16,000 (29). These discrepancies have been summarized by Lyerly et al. (16) and are difficult to explain (35). Perhaps in some of the isolation procedures a smaller contaminating protein copurified with the toxins and tended to mask the toxins in the polyacrylamide gels. The best approach to resolve this controversy is to clone and sequence the toxin genes.

Several investigators have begun cloning these genes. Muldrow and his collaborators (26) have reported the cloning of a 0.3-kilobase-pair (kb) fragment of the toxin A gene in the lambda bacteriophage expression vector gt11. The expressed peptides reacted with toxin A polyvalent antisera. When the cloned fragment from toxin A was used as a labeled probe, it reacted with a *Pst*I-generated fragment of *C. difficile* DNA, which they estimated as 4.5 kb. We have cloned a 4.7-kb *Pst*I fragment into a plasmid vector (28). This fragment has an internal *Pst*I site which is protected from digestion in the *C. difficile* DNA. When this fragment is expressed, the peptide reacts with both toxin A affinity-purified polyclonal antisera and with the monoclonal antibody PCG-4 (17). Preliminary results on the sequencing of this fragment have shown that there are many repeating sequence units within the fragment (14). Eichel-Streiber et al. (8) have recently cloned portions of the 4.7-kb *Pst*I fragment into a plasmid expression vector and obtained an expression product that also reacted with toxin A antisera. Wren et al. (36) have reported the cloning of toxin A in lambda phage. The clone expressed a protein that caused elongation of Chinese hamster ovary cells, and this protein had an estimated molecular weight of 235,000.

In this study, we have completed the cloning and sequencing of toxin A and its flanking regions. Nearly one-third of the gene (from the 3′ end) consists of a series of repeating units which appear to code for the receptor portion of the toxin.

MATERIALS AND METHODS

Bacteria, bacteriophages, and plasmids. DNA isolated from *C. difficile* VPI 10463 was used for cloning. Plasmids pBR322, pUC18, and pUC19 were used for the primary cloning of *C. difficile* DNA fragments. Subclones in the M13 phages mp18 and mp19 were used for DNA sequencing. The plasmids, phages, and *Escherichia coli* host strains JM109, DH5α, and DH5αF′ were all obtained from Bethesda Research Laboratories, Inc. *E. coli* strain Chi 1776 was purchased from the American Type Culture Collection.

Enzymes and radiolabeled compounds. Restriction endonu-

* Corresponding author.

† Present address: Department of Microbiology and Immunology, University of Kentucky Medical Center, Lexington, KY 40536.

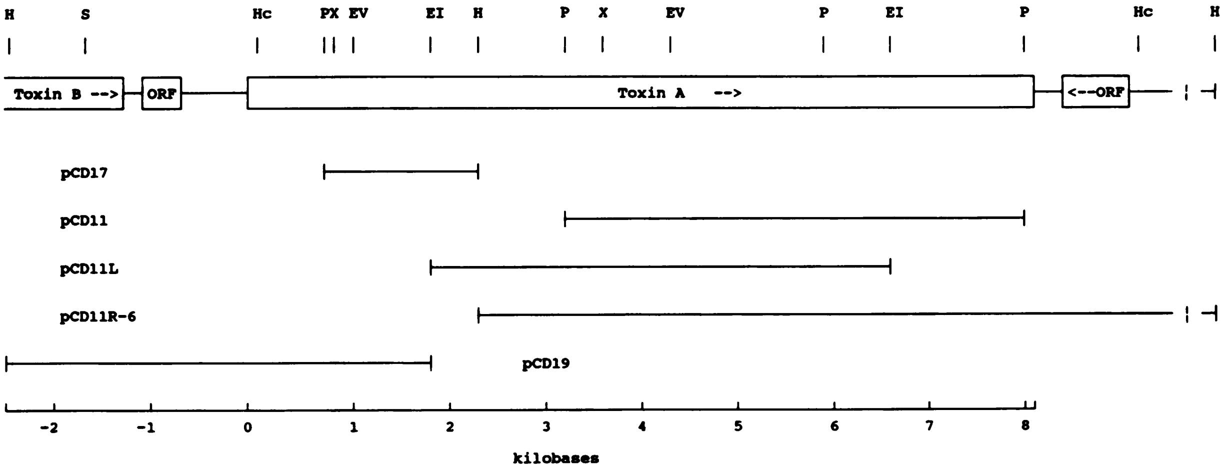

FIG. 1. Partial endonuclease restriction map of the cloned toxin A region from *C. difficile* strain 10463. Also shown are the sizes and locations of primary clones pCD11, pCD11L, pCD11R-6, pCD17, and pCD19.

clease enzymes were purchased from Bethesda Research Laboratories, Inc., International Biotechnologies, Inc., or Promega Biotec. Endonucleases III and VII, *E. coli* DNA polymerase I, and the polymerase I Klenow fragment were obtained from Bethesda Research Laboratories, Inc. DNA ligase and calf alkaline phosphatase were purchased from Boehringer Mannheim Biochemicals. The T7 DNA polymerase, Sequenase, was purchased from U.S. Biochemical. The labeled nucleotide triphosphate [α-^{35}S]dATP was obtained from New England Nuclear. Random primer labeling kits were obtained from International Biotechnologies, Inc. All of the enzymes were used according to the instructions provided by the manufacturers.

DNA isolations. High-molecular-weight *C. difficile* DNA was isolated by using a variation of the Marmur procedure (13, 22). The harvested cells (12) were suspended in 50 mM Tris–1 mM EDTA buffer (pH 8.0). After 50 μg of lysozyme per ml was added, the cell suspension was incubated at 37°C until the cells were susceptible to sodium dodecyl sulfate disruption. At the time of disruption, EDTA was added to a final concentration of 40 mM, proteinase K was added to a final concentration of 30 to 50 μg/ml, and β-mercaptoethanol was added to a concentration of 1% to inhibit endogenous nuclease activity in the lysate. Plasmid DNA and the replicating-form DNA of M13 phage were isolated by the Birnboim and Doly alkaline lysis procedure (4). DNA preparations used for probe fragment generation and nested deletions were further purified by CsCl centrifugation. Specific DNA restriction fragments were separated from others on low-melting-point agarose, and the individual bands were cut from the gel for use in the random priming labeling procedure. When restriction fragments of *C. difficile* DNA were needed in a particular size range (i.e., for cloning by chromosome walking), the gel was cut at the lower size range and this part was removed. A well was then cut into the gel at the upper size range, the polarity of the electrophoresis unit was reversed, and the fragments were electroeluted into the well. This tended to concentrate the fragments entering the well and resulted in a lower eluate volume. The fragments were then ethanol precipitated.

Primary cloning. Of the four primary clones used in this study (Fig. 1), three were cloned by the chromosome walking approach. The 2.6-kb *Pst*I fragment of pCD11 was used as a probe for cloning pCD11L, the 0.5-kb *Hin*dIII-*Eco*RI fragment of pCD11L was used as a probe for cloning pCD17, which was then used as a probe to clone pCD19, and pCD11 was used to detect clone pCD11R-6. The cloning was carried out under EK-2, BL-2 containment with *E. coli* Chi 1776 as host. Cells were made competent by the Hanahan procedure (11, 21).

Toxicity assays. Lysates from each primary clone were checked for animal and cell toxicity. Mouse lethality tests were performed by injecting five 8-week-old BALB/c mice (Dominion Laboratories, Dublin, Va.) intraperitoneally with 200 μl of lysate and observing them for illness or death. Cytotoxicity was checked in the Chinese hamster ovary (CHO) cell assay by following a procedure previously described (7). Lysates of *E. coli* Chi 1776 transformed with pUC18 were used as negative controls.

DNA sequencing and sequence analysis. Both strands of the DNA were sequenced by using the dideoxy-chain termination procedure developed by Sanger et al. (3, 31). DNA fragments were cloned into M13, and nested deletions were generated in replicating-form DNA by using the exonuclease III and exonuclease VII procedure (37). Restriction sites used for subcloning were sequenced across by using oligonucleotide primers and double-stranded sequencing. Synthetic oligonucleotide primers were also used for filling occasional sequence gaps not covered by the nested deletions.

Sequence analysis was done by using the Pustell programs from International Biotechnologies, Inc., and the Sequence Analysis Software Package from the Genetics Computer Group, University of Wisconsin. The data bases that were searched included the GenBank data base and the National Biomedical Research Foundation Protein Sequence Data Base. Unweighted Pair Group cluster analysis was done by using the NTSYS-pc programs (F. J. Rohlf, Exeter Publishing, Ltd.).

N-terminal sequencing. Toxin A was purified from culture filtrates of *C. difficile* VPI 10463 by sequential ammonium sulfate precipitation, ion-exchange chromatography, and precipitation at pH 5.6 as previously described (18). The highly purified protein was denatured with a final concentration of 2.5% sodium dodecyl sulfate–5% 2-mercaptoethanol at 100°C for 2 min and subjected to sodium dodecyl sulfate-polyacrylamide gel electrophoresis. After electrophoresis, the protein was transferred to polyvinylidene difluoride membranes by electroblotting and the N-terminal amino acid

```
                            ATAAAAAATC AATATTAATT TATTTTTAAA AAATAGAAAG GAGTGTATAA GATTTATTTT CAAAGTTTAA AAACAAGAAA ATCAATTTAA ATTTCAGAAG GAATAAATGT GGTTATAGAA      -41

        GTGGATTTAT TATCAAAAAT AATAATACTA GGAGGTTTTT ATG TCT TTA ATA TCT AAA GAA GAG TTA ATA AAA CTC GCA TAT AGC ATT AGA CCA AGA GAA AAT GAG TAT AAA ACT ATA CTA       81
                                                     M   S   L   I   S   K   E   E   L   I   K   L   A   Y   S   I   R   P   R   E   N   E   Y   K   T   I   L

ACT AAT TTA GAC GAA TAT AAT AAG TTA ACT ACA AAC AAT AAT GAA AAT AAA TAT TTA CAA TTA AAA AAA CTA AAT GAA TCA ATT GAT GTT TTT ATG AAT AAA TAT AAA ACT TCA AGC AGA      201
 T   N   L   D   E   Y   N   K   L   T   T   N   N   N   E   N   K   Y   L   Q   L   K   K   L   N   E   S   I   D   V   F   M   N   K   Y   K   T   S   S   R

AAT AGA GCA CTC TCT AAT CTA AAA AAA GAT ATA TTA AAA GAA GTA ATT CTT ATT AAA AAT TCC AAT ACA AGC CCT GTA GAA AAA AAT TTA CAT TTT GTA TGG ATA GGT GGA GAA GTC AGT      321
 N   R   A   L   S   N   L   K   K   D   I   L   K   E   V   I   L   I   K   N   S   N   T   S   P   V   E   K   N   L   H   F   V   W   I   G   G   E   V   S

GAT ATT GCT CTT GAA TAC ATA AAA CAA TGG GCT GAT ATT AAT GCA GAA TAT AAT ATT AAA CTG TGG TAT GAT AGT GAA GCA TTC TTA GTA AAT ACA CTA AAA AAG GCT ATA GTT GAA TCT      441
 D   I   A   L   E   Y   I   K   Q   W   A   D   I   N   A   E   Y   N   I   K   L   W   Y   D   S   E   A   F   L   V   N   T   L   K   K   A   I   V   E   S

TCT ACC ACT GAA GCA TTA CAG CTA CTA GAG GAA GAG ATT CAA AAT CCT CAA TTT GAT AAT ATG AAA TTT TAC AAA AAA AGG ATG GAA TTT ATA TAT GAT AGA CAA AAA AGG TTT ATA AAT      561
 S   T   T   E   A   L   Q   L   L   E   E   E   I   Q   N   P   Q   F   D   N   M   K   F   Y   K   K   R   M   E   F   I   Y   D   R   Q   K   R   F   I   N

TAT TAT AAA TCT CAA ATC AAT AAA CCT ACA GTA CCT ACA ATA GAT GAT ATT ATA AAG TCT CAT CTA GTA TCT GAA TAT AAT AGA GAT GAA ACT GTA TTA GAA TCA TAT AGA ACA AAT TCT      681
 Y   Y   K   S   Q   I   N   K   P   T   V   P   T   I   D   D   I   I   K   S   H   L   V   S   E   Y   N   R   D   E   T   V   L   E   S   Y   R   T   N   S

TTG AGA AAA ATA AAT AGT AAT CAT GGG ATA GAT ATC AGG GCT AAT AGT TTG TTT ACA GAA CAA GAG TTA TTA AAT ATT TAT AGT CAG GAG TTG TTA AAT CGT GGA AAT TTA GCT GCA GCA      801
 L   R   K   I   N   S   N   H   G   I   D   I   R   A   N   S   L   F   T   E   Q   E   L   L   N   I   Y   S   Q   E   L   L   N   R   G   N   L   A   A   A

TCT GAC ATA GTA AGA TTA TTA GCC CTA AAA AAT TTT GGC GGA GTA TAT TTA GAT GTT GAT ATG CTT CCA GGT ATT CAC TCT GAT TTA TTT AAA ACA ATA TCT AGA CCT AGC TCT ATT GGA      921
 S   D   I   V   R   L   L   A   L   K   N   F   G   G   V   Y   L   D   V   D   M   L   P   G   I   H   S   D   L   F   K   T   I   S   R   P   S   S   I   G

CTA GAC CGT TGG GAA ATG ATA AAA TTA GAG GCT ATT ATG AAG TAT AAA AAA TAT ATA AAT AAT TAT ACA TCA GAA AAC TTT GAT AAA CTT GAT CAA CAA TTA AAA GAT AAT TTT AAA CTC     1041
 L   D   R   W   E   M   I   K   L   E   A   I   M   K   Y   K   K   Y   I   N   N   Y   T   S   E   N   F   D   K   L   D   Q   Q   L   K   D   N   F   K   L

ATT ATA GAA AGT AAA AGT GAA AAA TCT GAG ATA TTT TCT AAA TTA GAA AAT TTA AAT GTA TCT GAT CTT GAA ATT AAA ATA GCT TTC GCT TTA GGC AGT GTT ATA AAT CAA GCC TTG ATA     1161
 I   I   E   S   K   S   E   K   S   E   I   F   S   K   L   E   N   L   N   V   S   D   L   E   I   K   I   A   F   A   L   G   S   V   I   N   Q   A   L   I

TCA AAA CAA GGT TCA TAT CTT ACT AAC CTA GTA ATA GAA CAA GTA AAA AAT AGA TAT CAA TTT TTA AAC CAA CAC CTT AAC CCA GCC ATA GAG TCT GAT AAT AAC TTC ACA GAT ACT ACT     1281
 S   K   Q   G   S   Y   L   T   N   L   V   I   E   Q   V   K   N   R   Y   Q   F   L   N   Q   H   L   N   P   A   I   E   S   D   N   N   F   T   D   T   T

AAA ATT TTT CAT GAT TCA TTA TTT AAT TCA GCT ACC GCA GAA AAC TCT ATG TTT TTA ACA AAA ATA GCA CCA TAC TTA CAA GTA GGT TTT ATG CCA GAA GCT CGC TCC ACA ATA AGT TTA     1401
 K   I   F   H   D   S   L   F   N   S   A   T   A   E   N   S   M   F   L   T   K   I   A   P   Y   L   Q   V   G   F   M   P   E   A   R   S   T   I   S   L

AGT GGT CCA GGA GCT TAT GCG TCA GCT TAC TAT GAT TTC ATA AAT TTA CAA GAA AAT ACT ATA GAA AAA ACT TTA AAA GCA TCA GAT TTA ATA GAA TTT AAA TTC CCA GAA AAT AAT CTA     1521
 S   G   P   G   A   Y   A   S   A   Y   Y   D   F   I   N   L   Q   E   N   T   I   E   K   T   L   K   A   S   D   L   I   E   F   K   F   P   E   N   N   L

TCT CAA TTG ACA GAA CAA GAA ATA AAT AGT CTA TGG AGC TTT GAT CAA GCA AGT GCA AAA TAT CAA TTT GAG AAA TAT GTA AGA GAT TAT ACT GGT GGA TCT CTT TCT GAA GAC AAT GGG     1641
 S   Q   L   T   E   Q   E   I   N   S   L   W   S   F   D   Q   A   S   A   K  /Y   Q   F   E   K   Y   V   R   D   Y   T   G   G   S   L   S   E   D   N   G

GTA GAC TTT AAT AAA AAT ACT GCC CTC GAC AAA AAC TAT TTA TTA AAT AAT AAA ATT CCA TCA AAC AAT GTA GAA GAA GCT GGA AGT AAA AAT TAT GTT CAT TAT ATC ATA CAG TTA CAA     1761
 V   D   F   N   K   N   T   A   L   D   K   N   Y   L   L   N   N   K   I   P   S   N   N   V   E   E   A   G   S   K   N   Y   V   H   Y   I   I   Q   L   Q

GGA GAT GAT ATA AGT TAT GAA GCA ACA TGC AAT TTA TTT TCT AAA AAT CCT AAA AAT AGT ATT ATT ATA CAA CGA AAT ATG AAT GAA AGT GCA AAA AGC TAC TTT TTA AGT GAT GAT GGA     1881
 G   D   D   I   S   Y   E   A   T   C   N   L   F   S   K   N   P   K   N   S   I   I   I   Q   R   N   M   N   E   S   A   K   S   Y   F   L   S   D   D   G

GAA TCT ATT TTA GAA TTA AAT AAA TAT AGG ATA CCT GAA AGA TTA AAA AAT AAG GAA AAA GTA AAA GTA ACC TTT ATT GGA CAT GGT AAA GAT GAA TTC AAC ACA AGC GAA TTT GCT AGA     2001
 E   S   I   L   E   L   N   K   Y   R   I   P   E   R   L   K   N   K   E   K   V   K   V   T   F   I   G   H   G   K   D   E   F   N   T   S   E   F   A   R

TTA AGT GTA GAT TCA CTT TCC AAT GAG ATA AGT TCA TTT TTA GAT ACC ATA AAA TTA GAT ATA TCA CCT AAA AAT GTA GAA GTA AAC TTA CTT GGA TGT AAT ATG TTT AGT TAT GAT TTT     2121
 L   S   V   D   S   L   S   N   E   I   S   S   F   L   D   T   I   K   L   D   I   S   P   K   N   V   E   V   N   L   L   G   C   N   M   F   S   Y   D   F

AAT GTT GAA GAA ACT TAT CCT GGG AAG TTG CTA TTA AGT ATT ATG GAC AAA ATT ACT TCC ACT TTA CCT GAT GTA AAT AAA AAT TCT ATT ACT ATA GGA GCA AAT CAA TAT GAA GTA AGA     2241
 N   V   E   E   T   Y   P   G   K   L   L   L   S   I   M   D   K   I   T   S   T   L   P   D   V   N   K   N   S   I   T   I   G   A   N   Q   Y   E   V   R

ATT AAT AGT GAG GGA AGA AAA GAA CTT CTG GCT CAC TCA GGT AAA TGG ATA AAT AAA GAA GAA GCT ATT ATG AGC GAT TTA TCT AGT AAA GAA TAC ATT TTT TTT GAT TCT ATA GAT AAT     2361
 I   N   S   E   G   R   K   E   L   L   A   H   S   G   K   W   I   N   K   E   E   A   I   M   S   D   L   S   S   K   E   Y   I   F   F   D   S   I   D   N

AAG CTA AAA GCA AAG TCC AAG AAT ATT CCA GGA TTA GCA TCA ATA TCA GAA GAT ATA AAA ACA TTA TTA CTT GAT GCA AGT GTT AGT CCT GAT ACA AAA TTT ATT TTA AAT AAT CTT AAG     2481
 K   L   K   A   K   S   K   N   I   P   G   L   A   S   I   S   E   D   I   K   T   L   L   L   D   A   S   V   S   P   D   T   K   F   I   L   N   N   L   K

CTT AAT ATT GAA TCT TCT ATT GGG GAT TAC ATT TAT TAT GAA AAA TTA GAG CCT GTT AAA AAT ATA ATT CAC AAT TCT ATA GAT GAT TTA ATA GAT GAG TTC AAT CTA CTT GAA AAT GTA     2601
 L   N   I   E   S   S   I   G   D   Y   I   Y   Y   E   K   L   E   P   V   K   N   I   I   H   N   S   I   D   D   L   I   D   E   F   N   L   L   E   N   V

TCT GAT GAA TTA TAT GAA TTA AAA AAA TTA AAT AAT CTA GAT GAG AAG TAT TTA ATA TCT TTT GAA GAT ATC TCA AAA AAT AAT TCA ACT TAC TCT GTA AGA TTT ATT AAC AAA AGT AAT     2721
 S   D   E   L   Y   E   L   K   K   L   N   N   L   D   E   K   Y   L   I   S   F   E   D   I   S   K   N   N   S   T   Y   S   V   R   F   I   N   K   S   N

GGT GAG TCA GTT TAT GTA GAA ACA GAA AAA GAA ATT TTT TCA AAA TAT AGC GAA CAT ATT ACA AAA GAA ATA AGT ACT ATA AAG AAT AGT ATA ATT ACA GAT GTT AAT GGT AAT TTA TTG     2841
 G   E   S   V   Y   V   E   T   E   K   E   I   F   S   K   Y   S   E   H   I   T   K   E   I   S   T   I   K   N   S   I   I   T   D   V   N   G   N   L   L

GAT AAT ATA CAG TTA GAT CAT ACT TCT CAA GTT AAT ACA TTA AAC GCA GCA TTC TTT ATT CAA TCA TTA ATA GAT TAT AGT AGC AAT AAA GAT GTA CTG AAT GAT TTA AGT ACC TCA GTT     2961
 D   N   I   Q   L   D   H   T   S   Q   V   N   T   L   N   A   A   F   F   I   Q   S   L   I   D   Y   S   S   N   K   D   V   L   N   D   L   S   T   S   V

AAG GTT CAA CTT TAT GCT CAA CTA TTT AGT ACA GGT TTA AAT ACT ATA TAT GAC TCT ATC CAA TTA GTA AAT TTA ATA TCA AAT GCA GTA AAT GAT ACT ATA AAT GTA CTA CCT ACA ATA     3081
 K   V   Q   L   Y   A   Q   L   F   S   T   G   L   N   T   I   Y   D   S   I   Q   L   V   N   L   I   S   N   A   V   N   D   T   I   N   V   L   P   T   I

ACA GAG GGG ATA CCT ATT GTA TCT ACT ATA TTA GAC GGA ATA AAC TTA GGT GCA GCA ATT AAG GAA TTA CTA GAC GAA CAT GAC CCA TTA CTA AAA AAA GAA TTA GAA GCT AAG GTG GGT     3201
 T   E   G   I   P   I   V   S   T   I   L   D   G   I   N   L   G   A   A   I   K   E   L   L   D   E   H   D   P   L   L   K   K   E   L   E   A   K   V   G

GTT TTA GCA ATA AAT ATG TCA TTA TCT ATA GCT GCA ACT GTA GCT TCA ATT GTT GGA ATA GGT GCT GAA GTT ACT ATT TTC TTA TTA CCT ATA GCT GGT ATA TCT GCA GGA ATA CCT TCA     3321
 V   L   A   I   N   M   S   L   S   I   A   A   T   V   A   S   I   V   G   I   G   A   E   V   T   I   F   L   L   P   I   A   G   I   S   A   G   I   P   S

TTA GTT AAT AAT GAA TTA ATA TTG CAT GAT AAG GCA ACT TCA GTG GTA AAC TAT TTT AAT CAT TTG TCT GAA TCT AAA AAA TAT GGC CCT CTT AAA ACA GAA GAT GAT AAA ATT TTA GTT     3441
 L   V   N   N   E   L   I   L   H   D   K   A   T   S   V   V   N   Y   F   N   H   L   S   E   S   K   K   Y   G   P   L   K   T   E   D   D   K   I   L   V

CCT ATT GAT GAT TTA GTA ATA TCA GAA ATA GAT TTT AAT AAT AAT TCG ATA AAA CTA GGA ACA TGT AAT ATA TTA GCA ATG GAG GGG GGA TCA GGA CAC ACA GTG ACT GGT AAT ATA GAT     3561
 P   I   D   D   L   V   I   S   E   I   D   F   N   N   N   S   I   K   L   G   T   C   N   I   L   A   M   E   G   G   S   G   H   T   V   T   G   N   I   D

CAC TTT TTC TCA TCT CCA TCT ATA AGT TCT CAT ATT CCT TCA TTA TCA ATT TAT TCT GCA ATA GGT ATA GAA ACA GAA AAT CTA GAT TTT TCA AAA AAA ATA ATG ATG TTA CCT AAT GCT     3681
 H   F   F   S   S   P   S   I   S   S   H   I   P   S   L   S   I   Y   S   A   I   G   I   E   T   E   N   L   D   F   S   K   K   I   M   M   L   P   N   A

CCT TCA AGA GTG TTT TGG TGG GAA ACT GGA GCA GTT CCA GGT TTA AGA TCA TTG GAA AAT GAC GGA ACT AGA TTA CTT GAT TCA ATA AGA GAT TTA TAC CCA GGT AAA TTT TAC TGG AGA     3801
 P   S   R   V   F   W   W   E   T   G   A   V   P   G   L   R   S   L   E   N   D   G   T   R   L   L   D   S   I   R   D   L   Y   P   G   K   F   Y   W   R

TTC TAT GCT TTT TTC GAT TAT GCA ATA ACT ACA TTA AAA CCA GTT TAT GAA GAC ACT AAT ATT AAA ATT AAA CTA GAT AAA GAT ACT AGA AAC TTC ATA ATG CCA ACT ATA ACT ACT AAC     3921
 F   Y   A   F   F   D   Y   A   I   T   T   L   K   P   V   Y   E   D   T   N   I   K   I   K   L   D   K   D   T   R   N   F   I   M   P   T   I   T   T   N

GAA ATT AGA AAC AAA TTA TCT TAT TCA TTT GAT GGA GCA GGA GGA ACT TAC TCT TTA TTA TTA TCT TCA TAT CCA ATA TCA ACG AAT ATA AAT TTA TCT AAA GAT GAT TTA TGG ATA TTT     4041
 E   I   R   N   K   L   S   Y   S   F   D   G   A   G   G   T   Y   S   L   L   L   S   S   Y   P   I   S   T   N   I   N   L   S   K   D   D   L   W   I   F
```

FIG. 2. Nucleotide and deduced amino acid sequences of *C. difficile* toxin A gene.

```
AAT ATT GAT AAT GAA GTA AGA GAA ATA TCT ATA GAA AAT GGT ACT ATT AAA AAA GGA AAG TTA ATA AAA GAT GTT TTA AGT AAA ATT GAT ATA AAT AAA AAT AAA CTT ATT ATA GGC AAT    4161
 N   I   D   N   E   V   R   E   I   S   I   E   N   G   T   I   K   K   G   K   L   I   K   D   V   L   S   K   I   D   I   N   K   N   K   L   I   I   G   N

CAA ACA ATA GAT TTT TCA GGC GAT ATA GAT AAT AAA GAT AGA TAT ATA TTC TTG ACT TGT GAG TTA GAT GAT AAA ATT AGT TTA ATA ATA GAA ATA AAT CTT GTT GCA AAA TCT TAT AGT    4281
 Q   T   I   D   F   S   G   D   I   D   N   K   D   R   Y   I   F   L   T   C   E   L   D   D   K   I   S   L   I   I   E   I   N   L   V   A   K   S   Y   S

TTG TTA TTG TCT GGG GAT AAA AAT TAT TTG ATA TCC AAT TTA TCT AAT ACT ATT GAG AAA ATC AAT ACT TTA GGC CTA GAT AGT AAA AAT ATA GCG TAC AAT TAC ACT GAT GAA TCT AAT    4401
 L   L   L   S   G   D   K   N   Y   L   I   S   N   L   S   N   T   I   E   K   I   N   T   L   G   L   D   S   K   N   I   A   Y   N   Y   T   D   E   S   N

AAT AAA TAT TTT GGA GCT ATA TCT AAA ACA AGT CAA AAA AGC ATA ATA CAT TAT AAA AAA GAC AGT AAA AAT ATA TTA GAA TTT TAT AAT GAC AGT ACA TTA GAA TTT AAC AGT AAA GAT    4521
 N   K   Y   F   G   A   I   S   K   T   S   Q   K   S   I   I   H   Y   K   K   D   S   K   N   I   L   E   F   Y   N   D   S   T   L   E   F   N   S   K   D

TTT ATT GCT GAA GAT ATA AAT GTA TTT ATG AAA GAT GAT ATT AAT ACT ATA ACA GGA AAA TAC TAT GTT GAT AAT AAT ACT GAT AAA AGT ATA GAT TTC TCT ATT TCT TTA GTT AGT AAA    4641
 F   I   A   E   D   I   N   V   F   M   K   D   D   I   N   T   I   T   G   K   Y   Y   V   D   N   N   T   D   K   S   I   D   F   S   I   S   L   V   S   K

AAT CAA GTA AAA GTA AAT GGA TTA TAT TTA AAT GAA TCC GTA TAC TCA TCT TAC CTT GAT TTT GTG AAA AAT TCA GAT GGA CAC CAT AAT ACT TCT AAT TTT ATG AAT TTA TTT TTG GAC    4761
 N   Q   V   K   V   N   G   L   Y   L   N   E   S   V   Y   S   S   Y   L   D   F   V   K   N   S   D   G   H   H   N   T   S   N   F   M   N   L   F   L   D

AAT ATA AGT TTC TGG AAA TTG TTT GGG TTT GAA AAT ATA AAT TTT GTA ATC GAT AAA TAC TTT ACC CTT GTT GGT AAA ACT AAT CTT GGA TAT GTA GAA TTT ATT TGT GAC AAT AAT AAA    4881
 N   I   S   F   W   K   L   F   G   F   E   N   I   N   F   V   I   D   K   Y   F   T   L   V   G   K   T   N   L   G   Y   V   E   F   I   C   D   N   N   K

AAT ATA GAT ATA TAT TTT GGT GAA TGG AAA ACA TCG TCA TCT AAA AGC ACT ATA TTT AGC GGA AAT GGT AGA AAT GTT GTA GTA GAG CCT ATA TAT AAT CCT GAT ACG GGT GAA GAT ATA    5001
 N   I   D   I   Y   F   G   E   W   K   T   S   S   S   K   S   T   I   F   S   G   N   G   R   N   V   V   V   E   P   I   Y   N   P   D   T   G   E   D   I

TCT ACT TCA CTA GAT TTT TCC TAT GAA CCT CTC TAT GGA ATA GAT AGA TAT ATA AAT AAA GTA TTG ATA GCA CCT GAT TTA TAT ACA AGT TTA ATA AAT ATT AAT ACC AAT TAT TAT TCA    5121
 S   T   S   L   D   F   S   Y   E   P   L   Y   G   I   D   R   Y   I   N   K   V   L   I   A   P   D   L   Y   T   S   L   I   N   I   N   T   N   Y   Y   S

AAT GAG TAC TAC CCT GAG ATT ATA GTT CTT AAC CCA AAT ACA TTC CAC AAA AAA GTA AAT ATA AAT TTA GAT AGT TCT TCT TTT GAG TAT AAA TGG TCT ACA GAA GGA AGT GAC TTT ATT    5241
 N   E   Y   Y   P   E   I   I   V   L   N   P   N   T   F   H   K   K   V   N   I   N   L   D   S   S   S   F   E   Y   K   W   S   T   E   G   S   D   F   I

TTA GTT AGA TAC TTA GAA GAA AGT AAT AAA AAA ATA TTA CAA AAA ATA AGA ATC AAA GGT ATC TTA TCT AAT ACT CAA TCA TTT AAT AAA ATG AGT ATA GAT TTT AAA GAT ATT AAA AAA    5361
 L   V   R   Y   L   E   E   S   N   K   K   I   L   Q   K   I   R   I   K   G   I   L   S   N   T   Q   S   F   N   K   M   S   I   D   F   K   D   I   K   K

CTA TCA TTA GGA TAT ATA ATG AGT AAT TTT AAA TCA TTT AAT TCT GAA AAT GAA TTA GAT AGA GAT CAT TTA GGA TTT AAA ATA ATA GAT AAT AAA ACT TAT TAC TAT GAT GAA GAT AGT    5481
 L   S   L   G   Y   I   M   S   N   F   K   S   F   N   S   E   N   E   L   D   R   D   H   L   G   F   K   I   I   D   N   K   T   Y   Y   Y   D   E   D   S

AAA TTA GTT AAA GGA TTA ATC AAT ATA AAT AAT TCA TTA TTC TAT TTT GAT CCT ATA GAA TTT AAC TTA GTA ACT GGA TGG CAA ACT ATC AAT GGT AAA AAA TAT TAT TTT GAT ATA AAT    5601
 K   L   V   K   G   L   I   N   I   N   N   S   L   F   Y   F   D   P   I   E   F   N   L   V   T   G   W   Q   T   I   N   G   K   K   Y   Y   F   D   I   N

ACT GGA GCA GCT TTA ACT AGT TAT AAA ATT ATT AAT GGT AAA CAC TTT TAT TTT AAT AAT GAT GGT GTG ATG CAG TTG GGA GTA TTT AAA GGA CCT GAT GGA TTT GAA TAT TTT GCA CCT    5721
 T   G   A   A   L   T   S   Y   K   I   I   N   G   K   H   F   Y   F   N   N   D   G   V   M   Q   L   G   V   F   K   G   P   D   G   F   E   Y   F   A   P

GCC AAT ACT CAA AAT AAT AAC ATA GAA GGT CAG GCT ATA GTT TAT CAA AGT AAA TTC TTA ACT TTG AAT GGC AAA AAA TAT TAT TTT GAT AAT AAC TCA AAA GCA GTC ACT GGA TGG AGA    5841
 A   N   T   Q   N   N   N   I   E   G   Q   A   I   V   Y   Q   S   K   F   L   T   L   N   G   K   K   Y   Y   F   D   N   N   S   K   A   V   T   G   W   R

ATT ATT AAC AAT GAG AAA TAT TAC TTT AAT CCT AAT AAT GCT ATT GCT GCA GTC GGA TTG CAA GTA ATT GAC AAT AAT AAG TAT TAT TTC AAT CCT GAC ACT GCT ATC ATC TCA AAA GGT    5961
 I   I   N   N   E   K   Y   Y   F   N   P   N   N   A   I   A   A   V   G   L   Q   V   I   D   N   N   K   Y   Y   F   N   P   D   T   A   I   I   S   K   G

TGG CAG ACT GTT AAT GGT AGT AGA TAC TAC TTT GAT ACT GAT ACC GCT ATT GCC TTT AAT GGT TAT AAA ACT ATT GAT GGT AAA CAC TTT TAT TTT GAT AGT GAT TGT GTA GTG AAA ATA    6081
 W   Q   T   V   N   G   S   R   Y   Y   F   D   T   D   T   A   I   A   F   N   G   Y   K   T   I   D   G   K   H   F   Y   F   D   S   D   C   V   V   K   I

GGT GTG TTT AGT ACC TCT AAT GGA TTT GAA TAT TTT GCA CCT GCT AAT ACT TAT AAT AAT AAC ATA GAA GGT CAG GCT ATA GTT TAT CAA AGT AAA TTC TTA ACT TTG AAT GGT AAA AAA    6201
 G   V   F   S   T   S   N   G   F   E   Y   F   A   P   A   N   T   Y   N   N   N   I   E   G   Q   A   I   V   Y   Q   S   K   F   L   T   L   N   G   K   K

TAT TAC TTT GAT AAT AAC TCA AAA GCA GTT ACC GGA TGG CAA ACT ATT GAT AGT AAA AAA TAT TAC TTT AAT ACT AAC ACT GCT GAA GCA GCT ACT GGA TGG CAA ACT ATT GAT GGT AAA    6321
 Y   Y   F   D   N   N   S   K   A   V   T   G   W   Q   T   I   D   S   K   K   Y   Y   F   N   T   N   T   A   E   A   A   T   G   W   Q   T   I   D   G   K

AAA TAT TAC TTT AAT ACT AAC ACT GCT GAA GCA GCT ACT GGA TGG CAA ACT ATT GAT GGT AAA AAA TAT TAC TTT AAT ACT AAC ACT GCT ATA GCT TCA ACT GGT TAT ACA ATT ATT AAT    6441
 K   Y   Y   F   N   T   N   T   A   E   A   A   T   G   W   Q   T   I   D   G   K   K   Y   Y   F   N   T   N   T   A   I   A   S   T   G   Y   T   I   I   N

GGT AAA CAT TTT TAT TTT AAT ACT GAT GGT ATT ATG CAG ATA GGA GTG TTT AAA GGA CCT AAT GGA TTT GAA TAT TTT GCA CCT GCT AAT ACG GAT GCT AAC AAC ATA GAA GGT CAA GCT    6561
 G   K   H   F   Y   F   N   T   D   G   I   M   Q   I   G   V   F   K   G   P   N   G   F   E   Y   F   A   P   A   N   T   D   A   N   N   I   E   G   Q   A

ATA CTT TAC CAA AAT GAA TTC TTA ACT TTG AAT GGT AAA AAA TAT TAC TTT GGT AGT GAC TCA AAA GCA GTT ACT GGA TGG AGA ATT ATT AAC AAT AAG AAA TAT TAC TTT AAT CCT AAT    6681
 I   L   Y   Q   N   E   F   L   T   L   N   G   K   K   Y   Y   F   G   S   D   S   K   A   V   T   G   W   R   I   I   N   N   K   K   Y   Y   F   N   P   N

AAT GCT ATT GCT GCA ATT CAT CTA TGC ACT ATA AAT AAT GAC AAG TAT TAC TTT AGT TAT GAT GGA ATT CTT CAA AAT GGA TAT ATT ACT ATT GAA AGA AAT AAT TTC TAT TTT GAT GCT    6801
 N   A   I   A   A   I   H   L   C   T   I   N   N   D   K   Y   Y   F   S   Y   D   G   I   L   Q   N   G   Y   I   T   I   E   R   N   N   F   Y   F   D   A

AAT AAT GAA TCT AAA ATG GTA ACA GGA GTA TTT AAA GGA CCT AAT GGA TTT GAG TAT TTT GCA CCT GCT AAT ACT CAC AAT AAT AAC ATA GAA GGT CAG GCT ATA GTT TAC CAG AAC AAA    6921
 N   N   E   S   K   M   V   T   G   V   F   K   G   P   N   G   F   E   Y   F   A   P   A   N   T   H   N   N   N   I   E   G   Q   A   I   V   Y   Q   N   K

TTC TTA ACT TTG AAT GGC AAA AAA TAT TAT TTT GAT AAT GAC TCA AAA GCA GTT ACT GGA TGG CAA ACC ATT GAT GGT AAA AAA TAT TAC TTT AAT CTT AAC ACT GCT GAA GCA GCT ACT    7041
 F   L   T   L   N   G   K   K   Y   Y   F   D   N   D   S   K   A   V   T   G   W   Q   T   I   D   G   K   K   Y   Y   F   N   L   N   T   A   E   A   A   T

GGA TGG CAA ACT ATT GAT GGT AAA AAA TAT TAC TTT AAT CTT AAC ACT GCT GAA GCA GCT ACT GGA TGG CAA ACT ATT GAT GGT AAA AAA TAT TAC TTT AAT ACT AAC ACT TTC ATA GCC    7161
 G   W   Q   T   I   D   G   K   K   Y   Y   F   N   L   N   T   A   E   A   A   T   G   W   Q   T   I   D   G   K   K   Y   Y   F   N   T   N   T   F   I   A

TCA ACT GGT TAT ACA AGT ATT AAT GGT AAA CAT TTT TAT TTT AAT ACT GAT GGT ATT ATG CAG ATA GGA GTG TTT AAA GGA CCT AAT GGA TTT GAA TAC TTT GCA CCT GCT AAT ACG GAT    7281
 S   T   G   Y   T   S   I   N   G   K   H   F   Y   F   N   T   D   G   I   M   Q   I   G   V   F   K   G   P   N   G   F   E   Y   F   A   P   A   N   T   D

GCT AAC AAC ATA GAA GGT CAA GCT ATA CTT TAC CAA AAT AAA TTC TTA ACT TTG AAT GGT AAA AAA TAT TAC TTT GGT AGT GAC TCA AAA GCA GTT ACC GGA CTG CGA ACT ATT GAT GGT    7401
 A   N   N   I   E   G   Q   A   I   L   Y   Q   N   K   F   L   T   L   N   G   K   K   Y   Y   F   G   S   D   S   K   A   V   T   G   L   R   T   I   D   G

AAA AAA TAT TAC TTT AAT ACT AAC ACT GCT GTT GCA GTT ACT GGA TGG CAA ACT ATT AAT GGT AAA AAA TAC TAC TTT AAT ACT AAC ACT TCT ATA GCT TCA ACT GGT TAT ACA ATT ATT    7521
 K   K   Y   Y   F   N   T   N   T   A   V   A   V   T   G   W   Q   T   I   N   G   K   K   Y   Y   F   N   T   N   T   S   I   A   S   T   G   Y   T   I   I

AGT GGT AAA CAT TTT TAT TTT AAT ACT GAT GGT ATT ATG CAG ATA GGA GTG TTT AAA GGA CCT GAT GGA TTT GAA TAC TTT GCA CCT GCT AAT ACA GAT GCT AAC AAT ATA GAA GGT CAA    7641
 S   G   K   H   F   Y   F   N   T   D   G   I   M   Q   I   G   V   F   K   G   P   D   G   F   E   Y   F   A   P   A   N   T   D   A   N   N   I   E   G   Q

GCT ATA CGT TAT CAA AAT AGA TTC CTA TAT TTA CAT GAC AAT ATA TAT TAT TTT GGT AAT AAT TCA AAA GCG GCT ACT GGT TGG GTA ACT ATT GAT GGT AAT AGA TAT TAC TTC GAG CCT    7761
 A   I   R   Y   Q   N   R   F   L   Y   L   H   D   N   I   Y   Y   F   G   N   N   S   K   A   A   T   G   W   V   T   I   D   G   N   R   Y   Y   F   E   P

AAT ACA GCT ATG GGT GCG AAT GGT TAT AAA ACT ATT GAT AAT AAA AAT TTT TAC TTT AGA AAT GGT TTA CCT CAG ATA GGA GTG TTT AAA GGG TCT AAT GGA TTT GAA TAC TTT GCA CCT    7881
 N   T   A   M   G   A   N   G   Y   K   T   I   D   N   K   N   F   Y   F   R   N   G   L   P   Q   I   G   V   F   K   G   S   N   G   F   E   Y   F   A   P

GCT AAT ACG GAT GCT AAC AAT ATA GAA GGT CAA GCT ATA CGT TAT CAA AAT AGA TTC CTA CAT TTA CTT GGA AAA ATA TAT TAC TTT GGT AAT AAT TCA AAA GCA GTT ACT GGA TGG CAA    8001
 A   N   T   D   A   N   N   I   E   G   Q   A   I   R   Y   Q   N   R   F   L   H   L   L   G   K   I   Y   Y   F   G   N   N   S   K   A   V   T   G   W   Q

ACT ATT AAT GGT AAA GTA TAT TAC TTT ATG CCT GAT ACT GCT ATG GCT GCA GCT GGT GGA CTT TTC GAG ATT GAT GGT GTT ATA TAT TTC TTT GGT GTT GAT GGA GTA AAA GCC CCT GGG    8121
 T   I   N   G   K   V   Y   Y   F   M   P   D   T   A   M   A   A   A   G   G   L   F   E   I   D   G   V   I   Y   F   F   G   V   D   G   V   K   A   P   G

ATA TAT GGC TAA AATATATGTT TGATAAAAAA TTATTCCTGT GCTACTAAGA AATTATTTTT ATATAATAAA TATTGAGATT TAATTAAAAT CATGTGTTAT TGTAATACAT GACTTTTAG TTAAA    8248
 I   Y   G   -
```

FIG. 2—*Continued*

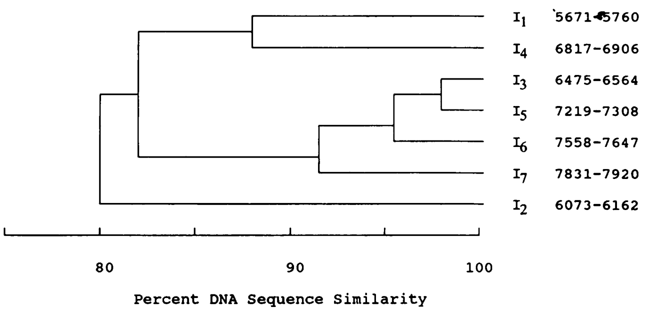

FIG. 3. Nucleotide sequence similarity cluster analysis of the class I repeating sequences.

sequence was determined by previously described methods (23).

RESULTS

Primary clones of toxin A. Relationships between the five primary clones, each containing a portion of the *C. difficile* toxin A gene, are shown in Fig. 1. Also included in the figure is a partial restriction map of this 15-kb region of the *C. difficile* genome. Clone pCD11 has been partially characterized and shown to contain a carbohydrate binding region and antigenic epitopes which react with the monoclonal antibody PCG-4 (28). Clone pCD11R-6, in addition to containing the entire pCD11 insert and most of the pCD11L insert, contains the last 80 bases of the toxin A gene and approximately 4.1 kb of additional sequences downstream from the toxin A gene. The downstream region contains two open reading frames (ORFs) and part of the third, one of which is shown in Fig. 1. All of these ORFs read in the direction opposite that of the toxin A gene (data not shown). Clone pCD11L contains an additional 1.5 kb of sequence upstream of the pCD11 insert. Clone pCD17 was used as a probe for cloning pCD19. Clone pCD19 codes for the 5′ end of toxin A, a small ORF that could code for a 16- or 19-kilodalton (kDa) protein and 1.2 kb of toxin B. These clones were not toxic for mice or CHO cells. The clone immediately upstream from the pCD19 insert was found to contain the remainder of the toxin B gene, and we have since been able to reconstruct the intact gene in a plasmid. The recombinant protein expressed by this plasmid is cytotoxic to tissue cells, is lethal to mice, and has immunological identity with toxin B (D. M. Lyerly and J. L. Johnson, unpublished data).

Nucleotide and amino acid sequences for toxin A. The nucleotide sequence and the deduced amino acids for the toxin A gene are shown in Fig. 2 (GenBank accession number, M30307). The open reading frame is 8,133 nucleotides long and codes for 2,710 amino acids. The gene contains 26.9 mol% G+C, and the deduced protein has a molecular mass of 308,103 Da. The amino acid sequence of the N-terminal end of toxin A was determined by microanalysis after electrophoresis under denaturing conditions (23), and the first 10 amino acids agree with the first 10 deduced amino acids of the toxin A open reading frame, indicating that there are no posttranslational modifications involving a signal peptide.

An interesting property of this gene is the repeating sequences at the 3′ end. A total of 2,551 nucleotides, or 31.5% of the gene, are in 38 contiguous repeating units. This region extends from nucleotides 5,545 to 8,106. The repeating units were grouped into two classes, I and II, on the basis of the low levels of DNA sequence similarities between them. There are 7 class I and 31 class II repeating units. Each of the class I repeats is 90 nucleotides long, and the class II repeats are either 60 or 63 nucleotides long, with the one exception being 66 nucleotides long. The class II repeats have been subdivided into 7 class IIA, 13 class IIB, 5 class IIC, and 6 class IID repeats.

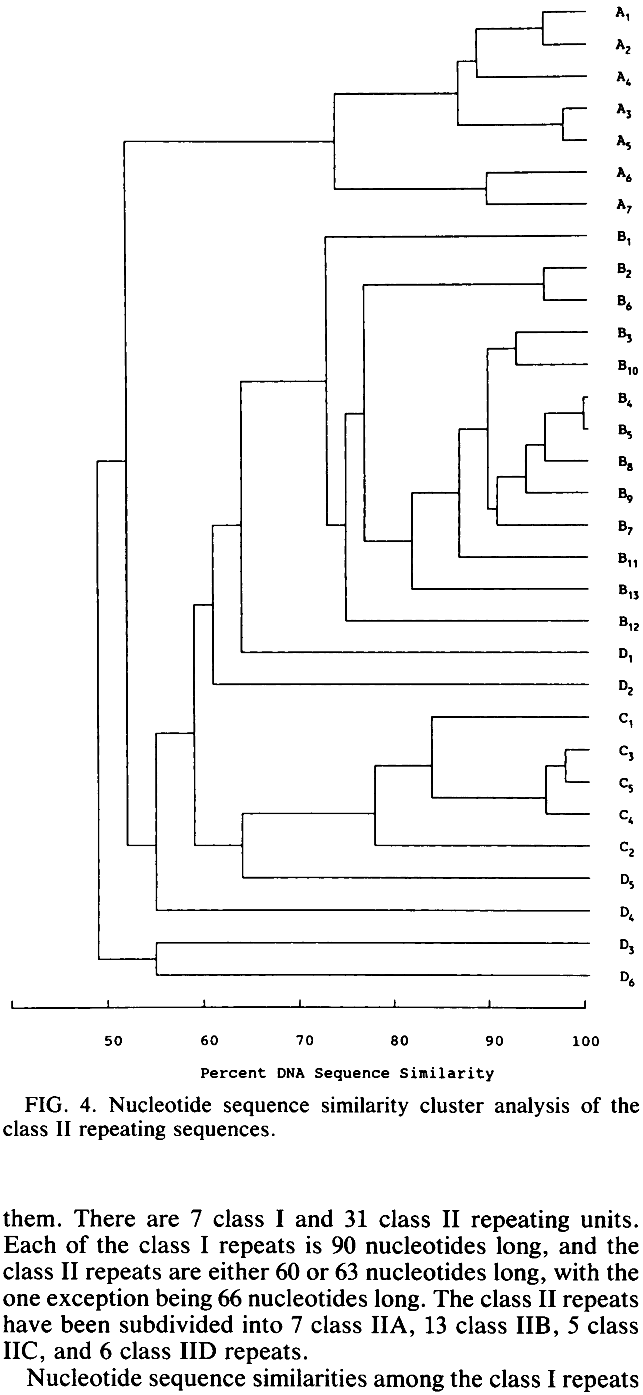

FIG. 4. Nucleotide sequence similarity cluster analysis of the class II repeating sequences.

Nucleotide sequence similarities among the class I repeats are shown in Fig. 3. Similarities ranged from 73 to 98%, with the average values in the cluster analysis being 80% or greater. Nucleotide sequence similarities among the class II repeats are shown in Fig. 4. With the exception of class IID, clustering within each subclass is high, being 70% or higher for class IIA, 65% or higher for class IIB, and 76% or higher for class IIC repeats. The class IID repeats are a diverse collection, in that all are very distinct. Two of them fit closer to the class IIB cluster and one fits closer to the class IIC group than to the others in class IID. This is also the only group in which there is any size variation; repeat unit class IID_4 has an extra AAA codon, while IID_5 has one fewer codon.

The deduced amino acid residues for the repeated se-

Class I peptides

I_1	1891-1920	M Q L G V F K G P D G F E Y F A P A N T Q N N N I E G Q A I
I_2	2025-2054	V K I G V F S T S N G F E Y F A P A N T Y N N N I E G Q A I
I_3	2159-2188	M Q I G V F K G P N G F E Y F A P A N T D A N N I E G Q A I
I_4	2273-2302	M V T G V F K G P N G F E Y F A P A N T H N N N I E G Q A I
I_5	2407-2436	M Q I G V F K G P N G F E Y F A P A N T D A N N I E G Q A I
I_6	2520-2549	M Q I G V F K G P D G F E Y F A P A N T D A N N I E G Q A I
I_7	2611-2640	P Q I G V F K G S N G F E Y F A P A N T D A N N I E G Q A I
	CONSENSUS	M Q I G V F K G P N G F E Y F A P A N T D A N N I E G Q A I

FIG. 5. Deduced amino acid sequences for the class I repeating units. Unit designations (I_1 to I_7) are listed in order from the N-terminal to C-terminal direction. The inclusive amino acid residue numbers are given for each unit, and the conserved amino acids are boxed.

quences are shown in Fig. 5 and 6. The inclusive amino acid residue numbers are given for each repeat unit, and the conserved amino acid residues within each class or class subgroup are boxed. Seventy percent of the amino acids in the class I peptides are conserved among the units, while less than 50% are conserved within each of the class II subgroups. The dipeptide tyrosine-phenylalanine (YF) is the most conserved and can be found in all 38 repeat units. It represents residues 14 and 15 in the class I units and residues 15 and 16 in the class II repeats, except for unit IID_4. Base differences in the regions of conserved amino acids involved the codon's third base as expected, whereas switching from one amino acid to another in a given position usually involved a total codon change or at least two of the bases.

A hydropathic index plot for the deduced toxin A protein and a map of the repeat units are shown in Fig. 7. There is no evidence for a signal peptide at the amino-terminal end; this finding is in agreement with the lack of posttranslational modification of the N-terminal end of the protein. The only strongly hydrophobic region in the deduced protein is from residues 1,050 to 1,100. There appears to be a periodicity in the hydropathic index within each repeat region. However, the repeat region is for the most part hydrophilic.

The 160 bases immediately upstream from the toxin A translation initiation site are shown in Fig. 2. There appears to be a ribosomal binding site (GGAGGT) starting six bases upstream of the initiation codon. Since we do not know where transcription initiates, it is difficult to predict promoter regions, although there are several TA-rich areas in the region of 160 bases upstream (Fig. 2). Other than these, there do not appear to be any other unique structures, such as inverted or tandem repeats.

Small protein. A small ORF (ca. 500 base pairs; Fig. 8; GenBank accession number, 30308) is located 122 bases downstream from the stop codon of the toxin B gene. Although the deduced amino acid sequence begins with the first start codon, there are two additional ATG codons at the amino acid residue positions 25 and 27. There appear to be ribosomal binding sites in the −10 regions of the first (GGTGGA) and third (GGAGGC) ATG codons. The deduced protein would have a molecular mass of 18,798 Da by using the longer sequence and a 15,878-Da molecular mass by using the shorter sequence. The pI values for the two peptides are 9.22 and 9.11, respectively. The hydropathic

CLASS IIA PEPTIDES

A_1	1921-1940	V Y Q S K F L T L N G K K Y Y F D N N S
A_2	2055-2074	V Y Q S K F L T L N G K K Y Y F D N N S
A_3	2189-2208	L Y Q N E F L T L N G K K Y Y F G S D S
A_4	2303-2322	V Y Q N K F L T L N G K K Y Y F D N D S
A_5	2437-2456	L Y Q N K F L T L N G K K Y Y F G S D S
A_6	2550-2569	R Y Q N R F L Y L H D N I Y Y F G N N S
A_7	2641-2660	R Y Q N R F L H L L G K I Y Y F G N N S
	CONSENSUS	V Y Q N K F L T L N G K K Y Y F G N N S

CLASS IIB PEPTIDES

B_1	1849-1869	N L V T G W Q T I N G K K Y Y F D I N T G
B_2	1941-1961	K A V T G W R I I N N E K Y Y F N P N N A
B_3	2075-2095	K A V T G W Q T I D S K K Y Y F N T N T A
B_4	2096-2116	E A A T G W Q T I D G K K Y Y F N T N T A
B_5	2117-2137	E A A T G W Q T I D G K K Y Y F N T N T A
B_6	2209-2229	K A V T G W R I I N N K K Y Y F N P N N A
B_7	2323-2343	K A V T G W Q T I D G K K Y Y F N L N T A
B_8	2344-2364	E A A T G W Q T I D G K K Y Y F N L N T A
B_9	2365-2385	E A A T G W Q T I D G K K Y Y F N T N T F
B_{10}	2457-2477	K A V T G L R T I D G K K Y Y F N T N T A
B_{11}	2478-2498	V A V T G W Q T I N G K K Y Y F N T N T S
B_{12}	2570-2590	K A A T G W V T I D G N R Y Y F E P N T A
B_{13}	2661-2681	K A V T G W Q T I N G K V Y Y F M P D T A
	CONSENSUS	K A V T G W Q T I D G K K Y Y F N T N T A

CLASS IIC

C_1	1870-1890	A A L T S Y K I I N G K H F Y F N N D G V
C_2	2004-2024	I A F N G Y K T I D G K H F Y F D S D C V
C_3	2138-2158	I A S T G Y T I I N G K H F Y F N T D G I
C_4	2386-2406	I A S T G Y T S I N G K H F Y F N T D G I
C_5	2499-2519	I A S T G Y T I I S G K H F Y F N T D G I
	CONSENSUS	I A S T G Y T I I N G K H F Y F N T D G I

CLASS IID

D_1	1962-1982	I A A V G L Q V I D N N K Y Y F N P D T A
D_2	1983-2003	I I S K G W Q T V N G S R Y Y F D T D T A
D_3	2230-2250	I A A I H L C T I N N D K Y Y F S Y D G I
D_4	2251-2271	L Q N G Y I T I E R N N F Y F D A N N E S K
D_5	2591-2611	M G A N G Y K T I D N K N F Y F R N G L
D_6	2682-2702	M A A A G G L F E I D G V I Y F F G V D G
	CONSENSUS	I A A - G - - T I - N - - Y Y F - - D - -

FIG. 6. Deduced amino acid sequences for the class II repeating units. Unit designations are made in the same manner as for the class I units.

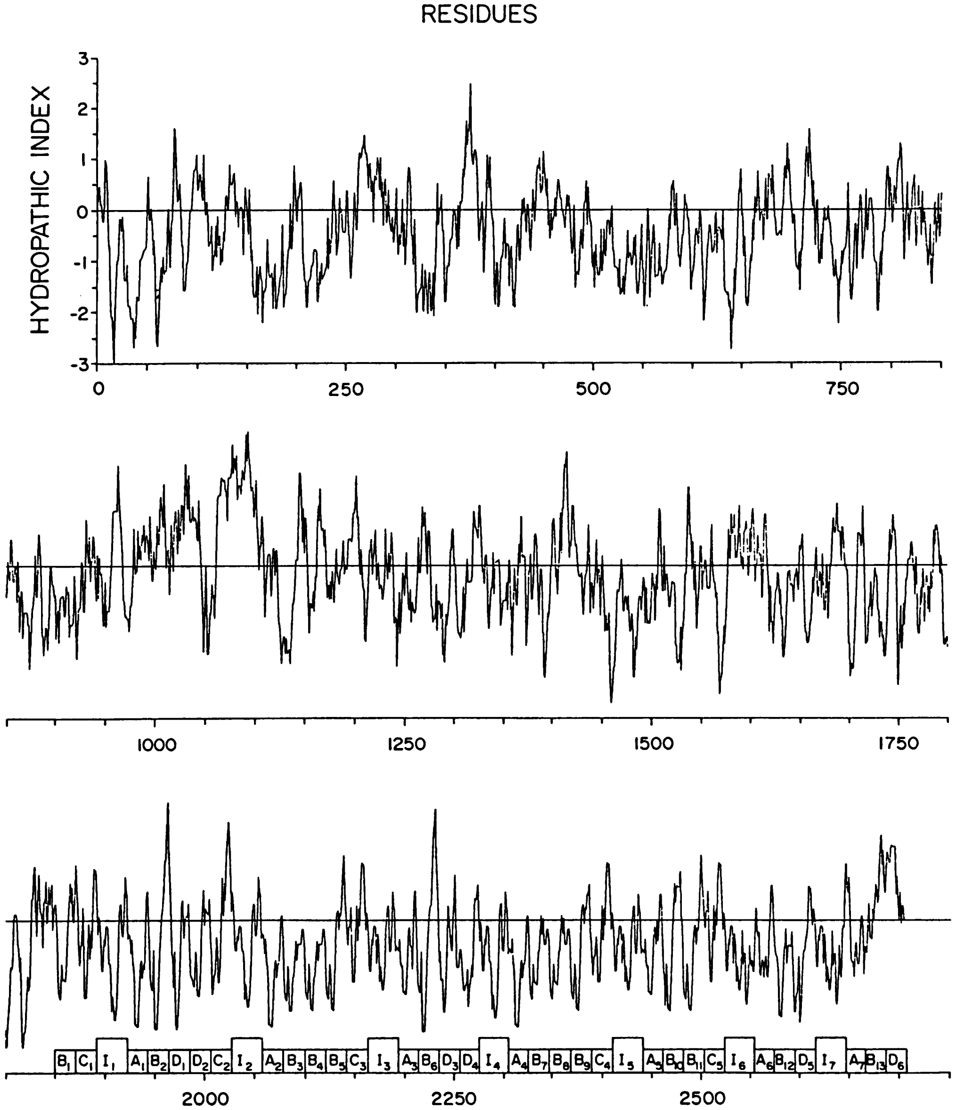

FIG. 7. Hydropathy plot and repeating unit map for *C. difficile* toxin A gene. Hydrophobic regions are indicated by positive values.

indexes were determined for both versions of the ORF (data not included). The deduced peptide is in general hydrophilic, and there does not appear to be a signal peptide in the first 25 amino acid residues; however, for a polypeptide starting at amino acid residue 27 (the third ATG codon), there is a short hydrophobic region that is characteristic of other signal sequences (24).

DISCUSSION

We report here the molecular mass of 308,103 Da for the deduced toxin A protein of *C. difficile*. This is in agreement with previous studies that reported a large size for this toxin (1, 18, 32, 33). Although we have not been able to express toxicity from the cloned fragments, the 2.1-kb *Pst*I fragment at the 3′ end of the gene has been used to express the major antigenic and carbohydrate binding sites of the toxin (15, 28). In fact, antiserum against this portion of the protein neutralizes the enterotoxicity of toxin A, and this is further evidence that the repeating units represent the binding portion (D. M. Lyerly and T. D. Wilkins, unpublished data).

The mechanism of action of toxin A is unknown. In the data base searches, we were unable to find any amino acid sequence similarities with other characterized toxins or enzymes. We cannot rule out a second peptide associating with this one, for example, the small ORF protein. The loss of such a protein would have very little effect on the electrophoretic migration and probably would not be detected if the protein existed in equimolar amounts with the large protein. Also, after electrophoresis under denaturing conditions, only antigenicity has been measured and not toxicity. It is not yet known whether the small protein is even expressed in *C. difficile*, so any presumed role for the small protein in the toxicity of the organism will have to await further study.

The most interesting feature of the toxin A gene is the repeating sequences in the carbohydrate binding region, which, as seen by the hydropathy plot, contains the most hydrophilic portion of the molecule. This is at the carboxyl end of the protein and includes over a third of the polypeptide. Proteins with repeating units have been reported from a wide range of organisms. Some of the highly antigenic

```
ATAAAAATAT GTTAAATATA TCCTCTTATA CTTAAATATA TAAAAATAAA CAAAATGATA        60

CACTACATAA AGTGTTCTAT CTAATATGAA GATTTACCAA TAAAAAGGTG GACTATGATG       120

A ATG CAC AGT AGT TCA CCT TTT TAT ATT TCT AAT GGT AAC AAA ATA TTT TTT   172
   M   H   S   S   S   P   F   Y   I   S   N   G   N   K   I   F   F

TAT ATA AAC CTA GGA GGC GTT ATG AAT ATG ACA ATA TCT TTT TTA TCA GAG     223
 Y   I   N   L   G   G   V   M   N   M   T   I   S   F   L   S   E

CAT ATA TTT ATA AAG TTA GTA ATT TTA ACT ATA TCA TTT GAT ACA TTA TTA     274
 H   I   F   I   K   L   V   I   L   T   I   S   F   D   T   L   L

GGA TGT TTA AGT GCA ATA AAA AGT CGT AAA TTT AAT TCT AGT TTT GGA ATA     325
 G   C   L   S   A   I   K   S   R   K   F   N   S   S   F   G   I

GAT GGA GGA ATC AGA AAA GTA GCA ATG ATA GCA TGT ATA TTT TTT TTA TCA     376
 D   G   G   I   R   K   V   A   M   I   A   C   I   F   F   L   S

GTA GTT GAC ATT CTT ACA AAG TTT AAC TTT TTA TTT ATG TTA CCA CAA GAT     427
 V   V   D   I   L   T   K   F   N   F   L   F   M   L   P   Q   D

TGT ATC AAT TTT TTA AGA CTA AAA CAT CTT GGA ATA TCT GAA TTT TTC TCT     478
 C   I   N   F   L   R   L   K   H   L   G   I   S   E   F   F   S

ATT TTA TTT ATT TTA TAT GAA AGT GTA AGT ATA TTA AAA AAT ATG TGC TTA     529
 I   L   F   I   L   Y   E   S   V   S   I   L   K   N   M   C   L

TGT GGA TTA CCA GTA CCT AAG AGA TTA AAG GAA AAA ATA GCA ATT TTA CTA     580
 C   G   L   P   V   P   K   R   L   K   E   K   I   A   I   L   L

GAT GCA ATG ACA GAT GAA ATG AAT GCT AAG GAT GAA AAG TAA GTAATGGT        630
 D   A   M   T   D   E   M   N   A   K   D   E   K   END

AGATATAATA AAGATATTAA CAAATAAAAA GTGTTATCCA AATAAGAATA GCTGAAAGTT       690

ATCATAATTC ATGAAACTAA TAATGAAAAC GAGGGAGCAG ATGCCAAGAG ACACACAAGT       750

ATTAAATACA TATAATTTCG AAGCAAGTGT TCATTACTAT ATAGATGACA AGGTAGTATA       810

TCAAACATTG GTTCACAAAG ATGGTGCATG GTCAGTTGGT AAAATCTATT AAGCTACATT       870

AGTTACAGAT ATCACAAACT ATAATAGTTA AACATAGAAA TATGTGTAAA TTGTGATGGA       930

AATTATTCAA AAACACAAAA ATACGTGATG AAGGACAAAA TGATATAGAA AATAAGTATC       990

AAACCTTAAT AAATGATTTA ATTGATAGTT TAAAAGTTAT AGGAAAAATA TATAAAGAAA      1050

TAAAAACATT AAAAAAATAT AAGATATGTT TACAAATTAC TATCAGACAA TCTCCTTATC      1110

TAATAGAAGA GTCAATTAAC TAATTGAGTA TCTTTAAATT GAAATGTTAG GAAGTGATTT      1170

AAATATGAAA ACTTAAATT     1189
```

FIG. 8. Nucleotide and deduced amino acid sequences of the small open reading frame located between the 3′ end of *C. difficile* toxin B and the 5′ end of toxin A. Also included are the sequences between toxin B and the open reading frame and between the open reading frame and the first nucleotide (−160) listed in the toxin A sequence (Fig. 1).

surface proteins of *Plasmodium* species have repeated sequences, several of which are believed to be target cell binding proteins (25). These repeating units range from 3 to 18 amino acids in length, are repeated from 5 to as many as 41 times, and may consist of nearly 40% of the protein (5, 6). Several toxin genes have been sequenced that contain repeating sequences at the C-terminal end of the proteins. The C-terminal region of the *E. coli* hemolysin polypeptide contains 13 8-amino-acid repeating units, which are required for hemolytic activity (9). The calmodulin-sensitive adenylate cyclase of *Bordetella pertussis* contains two regions that contain repeating units (10). Eleven repeating units of 15 amino acids have recently been reported for the insecticidal crystal proteins of *Bacillus thuringiensis* (34). Although the repeating sequences of *C. difficile* toxin A do not have any sequence similarities with any of these other proteins, location at the C-terminal end of the proteins is common, and some may have a common role for target cell binding. Because the repeating region constitutes about one-third of the entire toxin molecule and the repeats are highly hydrophilic, it would be interesting to determine the spatial distribution of these repeats in the native protein. It remains to be shown whether a periodicity on the surface of the toxin molecule confers certain unique biological properties to the protein. We are currently pursuing research in this area to gain more understanding of the structure and function of this toxin.

ACKNOWLEDGMENTS

This research was supported by Commonwealth of Virginia project 6127250 and by Public Health Service grant AI 15749 from the National Institutes of Health.

LITERATURE CITED

1. **Banno, Y., T. Kobayashi, K. Watanabe, K. Ueno, and Y. Nozawa.** 1981. Two toxins (D-1 and D-2) of *Clostridium difficile* causing antibiotic-associated colitis: purification and some characterization. Biochem. Int. 2:629–635.
2. **Bartlett, J. G., N. S. Taylor, T. Chang, and J. Dzink.** 1980. Clinical and laboratory observations in *Clostridium difficile* colitis. Am. J. Clin. Nutr. **33:**2521–2526.
3. **Biggin, M. D., T. J. Gibson, and G. F. Hong.** 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80:**3963–3965.
4. **Birnboim, H. C., and J. Doly.** 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. **7:**1513–1523.
5. **Coppel, R. L., A. F. Cowman, K. R. Lingelbach, G. V. Brown, R. B. Saint, D. J. Kemp, and R. F. Anders.** 1983. Isolate-specific S-antigen of *Plasmodium falciparum* contains a repeated sequence of eleven amino acids. Nature (London) **306:**751–756.
6. **Dame, J. B., J. L. Williams, T. F. McCutchan, J. L. Weber, R. A. Wirtz, W. T. Hockmeyer, W. L. Maloy, J. D. Haynes, I. Schneider, D. Roberts, G. S. Sanders, E. P. Reddy, C. L. Diggs, and L. H. Miller.** 1984. Structure of the gene encoding the immunodominant surface antigen of the sporozoite of the human malaria parasite *Plasmodium falciparum*. Science **225:** 593–599.
7. **Ehrich, M., R. L. Van Tassell, J. M. Libby, and T. D. Wilkins.** 1980. Production of *Clostridium difficile* antitoxin. Infect. Immun. **28:**1041–1043.
8. **Eichel-Streiber, C. V., D. Suckau, M. Wachter, and U. Hadding.** 1989. Cloning and characterization of overlapping DNA fragments of the toxin A gene of *Clostridium difficile*. J. Gen. Microbiol. **135:**55–64.
9. **Felmlee, T., and R. A. Welch.** 1988. Alterations of amino acid repeats in the *Escherichia coli* hemolysin affect cytolytic activity and secretion. Proc. Natl. Acad. Sci. USA **85:**5269–5273.
10. **Glaser, P., D. Ladant, O. Sezer, F. Pichot, A. Ullmann, and A. Danchin.** 1988. The calmodulin-sensitive adenylate cyclase of *Bordetella pertussis*: cloning and expression in *Escherichia coli*. Mol. Microbiol. **2:**19–30.
11. **Hanahan, D.** 1983. Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. **166:**557–580.
12. **Holdeman, L. V., E. P. Cato, and W. E. C. Moore (ed.).** 1977. Anaerobe laboratory manual, 4th ed. Virginia Polytechnic Institute and State University, Blacksburg.
13. **Johnson, J. L.** 1981. Genetic characterization, p. 450–472. *In* P. Gerhardt, R. G. E. Murray, R. N. Costilow, E. W. Nester, W. A. Wood, N. R. Krieg, and G. B. Phillips (ed.), Manual of methods for general bacteriology. American Society for Microbiology, Washington, D.C.
14. **Johnson, J. L., C. H. Dove, S. B. Price, T. W. Sickles, C. J. Phelps, and T. D. Wilkins.** 1988. The Toxin A gene of *Clostridium difficile*, p. 115–123. *In* J. M. Hardie and S. P. Borriello (ed.), Anaerobes today. John Wiley & Sons, Ltd., London.
15. **Krivan, H. C., G. F. Clark, D. F. Smith, and T. D. Wilkins.** 1986. Cell surface binding site for *Clostridium difficile* enterotoxin: evidence for a glycoconjugate containing the sequence Galα1-3Galβ1-4GlcNAc. Infect. Immun. **53:**573–581.
16. **Lyerly, D. M., H. C. Krivan, and T. D. Wilkins.** 1988. *Clostridium difficile*: its disease and toxins. Clin. Microbiol. Rev. **1:**1–18.
17. **Lyerly, D. M., C. J. Phelps, and T. D. Wilkins.** 1985. Monoclonal and specific polyclonal antibodies for immunoassay of *Clostridium difficile* toxin A. J. Clin. Microbiol. **21:**12–14.
18. **Lyerly, D. M., M. D. Roberts, C. J. Phelps, and T. D. Wilkins.** 1986. Purification and properties of toxins A and B of *Clostridium difficile*. FEMS Microbiol. Lett. **33:**31–35.
19. **Lyerly, D. M., K. S. Saum, D. K. McDonald, and T. D. Wilkins.** 1985. Effects of *Clostridium difficile* given intragastrically to

animals. Infect. Immun. **47**:349–352.

20. **Lyerly, D. M., N. M. Sullivan, and T. D. Wilkins.** 1983. Enzyme-linked immunosorbent assay for *Clostridium difficile* toxin A. J. Clin. Microbiol. **17**:72–78.
21. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
22. **Marmur, J.** 1961. A procedure for the isolation of deoxyribonucleic acid from microorganisms. J. Mol. Biol. **3**:208–218.
23. **Matsudaira, P.** 1987. Sequence from picomole quantities of proteins electroblotted onto polyvinylidene difluoride membranes. J. Biol. Chem. **262**:10035–10038.
24. **Michaelis, S., and J. Beckwith.** 1982. Mechanism of incorporation of cell envelope proteins in *Escherichia coli*. Annu. Rev. Microbiol. **36**:435–465.
25. **Miller, L. H., R. J. Howard, R. Carter, M. F. Good, V. Nussenzweig, and R. S. Nussenzweig.** 1986. Research toward malaria vaccines. Science **234**:1349–1356.
26. **Muldrow, L. L., G. C. Ibeanu, N. I. Lee, N. K. Bose, and J. Johnson.** 1987. Molecular cloning of *Clostridium difficile* toxin A gene fragment in lambda gt11. FEBS Lett. **213**:249–253.
27. **Pothoulakis, D., L. M. Barone, R. Ely, B. Faris, M. E. Clark, C. Franzblau, and J. T. LaMont.** 1986. Purification and properties of *Clostridium difficile* cytotoxin B. J. Biol. Chem. **261**:1316–1321.
28. **Price, S. B., C. J. Phelps, T. D. Wilkins, and J. L. Johnson.** 1987. Cloning of the carbohydrate-binding portion of the toxin A gene of *Clostridium difficile*. Curr. Microbiol. **16**:55–60.
29. **Rihn, B., J. M. Scheftel, R. Girardot, and H. Monteil.** 1984. A new purification procedure for *Clostridium difficile* enterotoxin. Biochem. Biophys. Res. Commun. **124**:690–695.
30. **Rolfe, R. D., and S. M. Finegold.** 1979. Purification and characterization of *Clostridium difficile* toxin. Infect. Immun. **25**:191–201.
31. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
32. **Sullivan, N. M., S. Pellett, and T. D. Wilkins.** 1982. Purification and characterization of toxins A and B of *Clostridium difficile*. Infect. Immun. **35**:1032–1040.
33. **Taylor, N. S., G. M. Thorne, and J. G. Bartlett.** 1981. Comparison of two toxins produced by *Clostridium difficile*. Infect. Immun. **34**:1036–1043.
34. **Widner, W. R., and H. R. Whiteley.** 1989. Two highly related insecticidal crystal proteins of *Bacillus thuringiensis* subsp. *kurstaki* possess different host range specificities. J. Bacteriol. **171**:965–974.
35. **Wilkins, T. D.** 1987. Role of *Clostridium difficile* toxins in disease. Gastroenterology **93**:389–391.
36. **Wren, B. W., C. L. Clayton, P. P. Mullany, and S. Tabaqchali.** 1987. Molecular cloning and expression of *Clostridium difficile* toxin A in *Escherichia coli* K12. FEBS Lett. **225**:82–86.
37. **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 **34**:103–119.