Gene, 69 (1988) 287–300
Elsevier

GEN 02569

Sequence and analysis of the DNA encoding protective antigen of *Bacillus anthracis*

(Recombinant DNA; *E. coli* host; anthrax; toxin gene; inverted repeat; signal peptide; promoter; start and stop codons; transcription terminator)

S.L. Welkos[a], J.R. Lowe[a], F. Eden-McCutchan[b], M. Vodkin[a*], S.H. Leppla[a] and J.J. Schmidt[a]

Bacteriology Division, U.S. Army Medical Research Institute of Infectious Diseases[a], Frederick, MD 21701-5011 (U.S.A.) and Molecular Biology and Recombinant DNA Laboratories[b], Bionetics Research, Inc., Rockville, MD 20850 (U.S.A.) Tel. (301)258-5200

Received 5 April 1988
Revised 4 May 1988
Accepted 5 May 1988
Received by publisher 7 June 1988

SUMMARY

The nucleotide sequence of the protective antigen (PA) gene from *Bacillus anthracis* and the 5′ and 3′ flanking sequences were determined. PA is one of three proteins comprising anthrax toxin; and its nucleotide sequence is the first to be reported from *B. anthracis*. The open reading frame (ORF) is 2319 bp long, of which 2205 bp encode the 735 amino acids of the secreted protein. This region is preceded by 29 codons, which appear to encode a signal peptide having characteristics in common with those of other secreted proteins. A consensus TATAAT sequence was located at the putative −10 promoter site. A Shine–Dalgarno site similar to that found in genes of other *Bacillus* sp. was located 7 bp upstream from the ATG start codon. The codon usage for the *PA* gene reflected its high A + T (69%) base composition and differed from those of genes for bacterial proteins from most other sequences examined. The TAA translation stop codon was followed by an inverted repeat forming a potential termination signal. In addition, a 192-codon ORF of unknown significance, theoretically encoding a 21.6-kDa protein, preceded the 5′ end of the *PA* gene.

INTRODUCTION

B. anthracis is an important pathogen of animals and of people exposed to infected animals or their products. It can cause cutaneous anthrax, gastrointestinal anthrax, and an often fatal systemic pulmonary form of the disease (Hambleton et al., 1984; Leppla et al., 1985; Lincoln and Fish, 1970). The

Correspondence to: Dr. S.L. Welkos, Bacteriology Division, U.S. Army Medical Research Institute of Infectious Diseases, Frederick, MD 21701-5011 (U.S.A.) Tel. (301)663-7341.
* Present address: Department of Veterinary Pathobiology, University of Illinois, Urbana, IL 61801 (U.S.A.) Tel. (217)333-2449.

Abbreviations: aa, amino acid(s); bp, base pair(s); CHO, Chinese hamster ovary; EF, edema factor; EtdBr, ethidium bromide; kb, 1000 bp; LF, lethal factor; nt, nucleotide(s); ORF, open reading frame; PA, protective antigen; *PA*, gene coding for PA; PolIk, Klenow (large) fragment of *E. coli* DNA polymerase I; PAGE, polyacrylamide-gel electrophoresis; SDS, sodium dodecyl sulfate; [ ], designates plasmid-carrier state.

two major virulence factors of *B. anthracis* are a poly-D-glutamic acid capsule and 'anthrax toxin'. DNA functions controlling toxin and capsule production are carried on *B. anthracis* plasmids pXO1 and pXO2, respectively (Green et al., 1985; Mikesell et al., 1983). The toxin is composed of three separate proteins, PA, EF and LF. The three proteins alone are nontoxic. However, PA in combination with LF causes death in rats (Beall et al., 1962), and PA combined with EF produces edema in the skin of guinea pigs and rabbits (Leppla et al., 1985; Lincoln and Fish, 1970). In addition to mediating the toxic effects of LF and EF, protective antigen induces immunity to infection and is the major component of the currently licensed human vaccine (Hambleton et al., 1984; Ivins and Welkos, 1986; Leppla et al., 1985; Little and Knudson, 1986; Turnbull et al., 1986).

To understand the role of PA in the pathogenesis of disease and the induction of protective immunity, the DNA encoding PA has been cloned and sequenced. All three of the toxin proteins are encoded by the 176-kb plasmid pXO1 (Mikesell et al., 1983; Robertson and Leppla, 1986; Vodkin and Leppla, 1983). Vodkin and Leppla (1983) first reported the cloning of the *PA* gene in *Escherichia coli*. The gene was contained in a 6-kb *Bam*HI fragment of pXO1 cloned into plasmid pBR322. Full-size, biologically active PA was produced. The *Bacillus* promoter was present but expression of the gene by the recombinant plasmid (pSE36) in *E. coli* was low.

In a recent study, we subcloned the 6-kb insert of pSE36 into the plasmid vector pUB110 and transformed *B. subtilis* with the recombinant DNA (Ivins and Welkos, 1986). Two recombinants were isolated which produced large amounts of full-size PA despite the presence of deletions in the 6-kb insert of approx. 2.7 kb and 3.4 kb, respectively. In vitro concentrations of PA produced by the recombinants were similar or greater than those observed with *B. anthracis* (Ivins and Welkos, 1986). Protective antigen, a protein of approx. 85 kDa by SDS–PAGE (Ivins and Welkos, 1986; Leppla et al., 1985; Vodkin and Leppla, 1983), requires a coding region of 2–2.5 kb.

The purpose of the present study was to map and sequence the coding region of PA. Partial digestion and religation of plasmid pSE36 (which has the 6-kb insert) yielded a smaller derivative plasmid, pPA26, which contains a 4.2-kb insert encoding full-size PA. In this report, the nucleotide sequence of this insert and analysis of the PA-coding region are presented.

MATERIALS AND METHODS

(a) Bacteria and plasmids

Isolates of *E. coli* K-12 strain HB101, transformed with pSE36 or pPA26, were the sources of plasmid DNA; and strain JM103 (Messing, 1983) was used to propagate M13 phage derivatives.

(b) Subcloning and detection of protective antigen-producing recombinants

The isolation of recombinant *E. coli* [pSE36] has been described (Vodkin and Leppla, 1983). Briefly, pSE36 consists of plasmid pBR322 with a 6-kb *Bam*HI fragment encoding the PA protein from plasmid pXO1 of *B. anthracis*. To obtain derivatives having smaller insert DNA, plasmid pSE36 DNA was partially digested with *Hin*dIII and religated. *E. coli* strain HB101 was transformed with the plasmid DNA, and recombinants were tested for the presence of the *PA* gene by immunological assay (Vodkin and Leppla, 1983). The size and biological activity of PA produced by the recombinants were tested by a Western-blot procedure and the CHO cell elongation assay, respectively (Ivins and Welkos, 1986; Leppla, 1984; Vodkin and Leppla, 1983).

(c) Isolation of DNA

Plasmid pPA26 DNA was prepared from cleared lysates by ultracentrifugation in CsCl-EtdBr gradients according to methods described by Maniatis et al. (1982). The DNA was digested simultaneously with *Hin*dIII + *Bam*HI, and the 4.2-kb insert encoding PA was isolated as a 2.2-kb *Hin*dIII and 2.0-kb *Hin*dIII-*Bam*HI fragment (see Fig. 1). The DNA fragments were purified by preparative gel electrophoresis, the bands excised, and the DNA extracted with phenol for cloning in M13.

(d) Nucleotide sequence analysis

The two fragments were each cloned into phages M13mp10 and M13mp11, and the dideoxy chain-termination method (Messing and Seeburg, 1981; Sanger et al., 1977) was used to sequence the DNA. Initially, data were collected by using the universal primer (Pharmacia P-L Biochemicals, Piscataway, NJ). Using these data, we synthesized oligodeoxynucleotide primers 18 nt long to collect each additional data segment (Sanchez-Pescador and Urdea, 1984). The oligodeoxynucleotides were prepared by the phosphoramidite method (Applied Biosystems, Foster City, CA). The products of the sequencing reactions were separated in 7% denaturing polyacrylamide gels, and data read from the autoradiograms were compiled and melded by using the GEL program in the IntelliGenetics Molecular Biology software package (IntelliGenetics, Inc., Mountain View, CA).

(e) Enzymes and reagents

Restriction endonucleases were purchased from International Biotechnologies, Inc. (New Haven, CT) and Bethesda Research Laboratories (Gaithersburg, MD) and were used as recommended by the suppliers. T4 DNA ligase and deoxynucleoside and dideoxynucleoside triphosphates were from Pharmacia P-L. PolIk was purchased from Boehringer Mannheim Biochemicals (Indianapolis, IN), and [α-^{32}P]deoxynucleoside triphosphates (300–800 Ci/mmol, 11.1–29.6 TBq/mmol) were from Amersham (Arlington Heights, IL).

(f) Computer analysis of nucleotide sequence and protein secondary structure

The sequence in pPA26 of *B. anthracis* DNA was analyzed by several computer software packages. The MOLGENJR programs (Lowe, 1986) were run

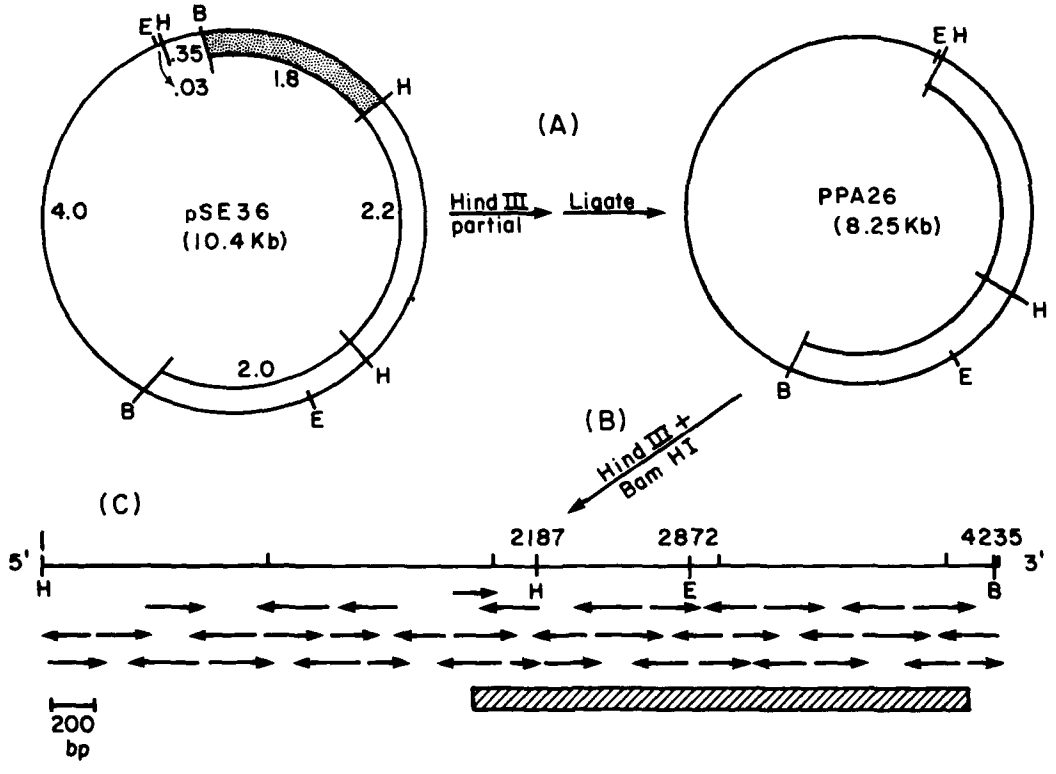


Fig. 1. Construction of plasmid containing the *PA* gene and sequencing strategy. (A) Plasmid pPA26 was constructed from pSE36 by partial *Hin*dIII digestion. The 4.2-kb *Hin*dIII-*Bam*HI portion of the plasmid (open box) contains the *PA* gene. The distances (in kb) between the *Bam*HI (B) and *Hin*dIII (H) sites are indicated; *Eco*RI sites (E) are included. (B) To sequence this insert, pPA26 was digested with *Bam*HI + *Hin*dIII. The 2.2-kb *Hin*dIII and 2.0-kb *Hin*dIII-*Bam*HI fragments were isolated and cloned into M13mp10 and M13mp11. (C) The arrows indicate the direction and extent of sequencing of the DNA fragments, totalling 4235 bp. The hatched bar indicates the structural gene for the mature PA protein.

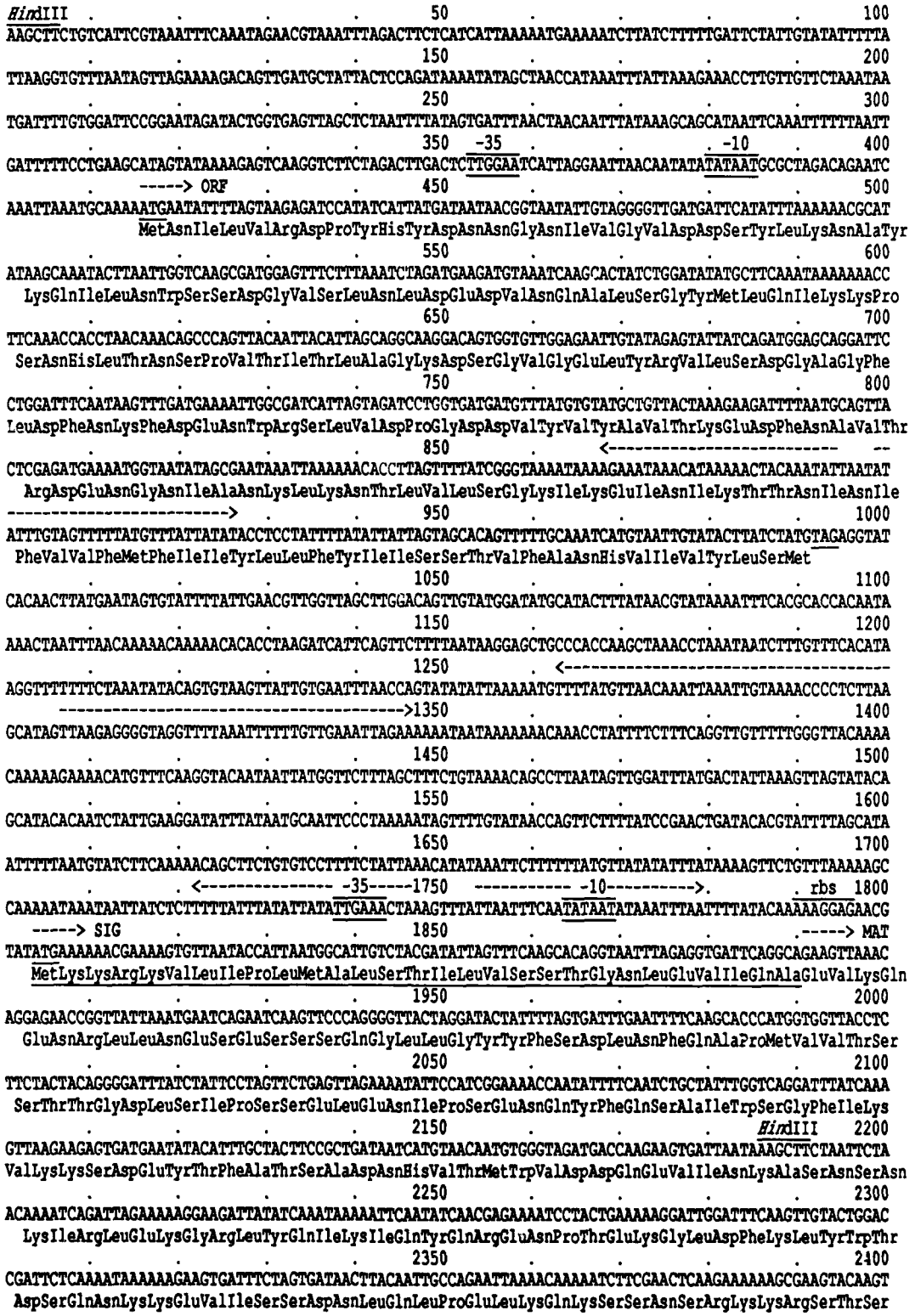
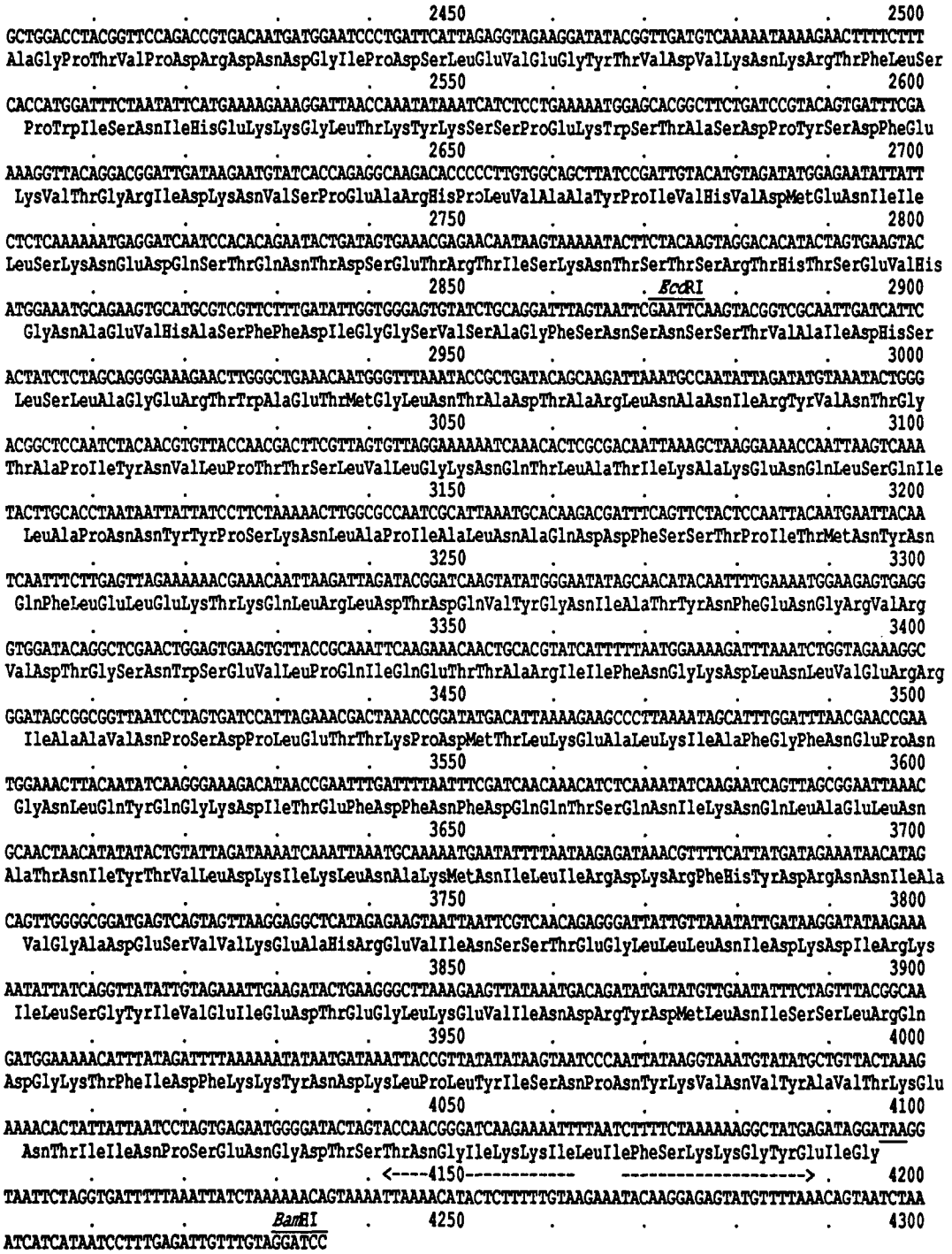


Fig. 2. Nucleotide and deduced amino acid sequence of the *PA* gene and 5′ and 3′ flanking sequences. The sequence shown corresponds to nt 1–4235 on the map in Fig. 1. Restriction endonuclease sites described in Fig. 1 and in RESULTS, section **b**, are indicated. The presumptive −35 and −10 sequences, and Shine–Dalgarno ribosome-binding site (rbs) of the *PA* gene and of the potential 192-nt ORF are underlined, as are the start (ATG) and stop (TAG, TAA) codons. Arrows above the nucleotide sequence indicate initiation of



translation of the potential ORF upstream from the *PA* gene (ORF1) and of the signal sequence (SIG) and mature protein (MAT) of the *PA* gene. The 29-aa signal peptide is underlined. The translated amino acid sequences of the 192-nt ORF and the *PA* gene, only, are shown. The potential stem-loop termination structure flanking the 3′ end of the *PA* gene, and the three palindromic sequences on the 5′ side of the *PA* gene are indicated by dashed lines between outward pointing arrowheads above the sequences.

on an IBM-PC microcomputer to confirm experimental restriction enzyme cleavage patterns, deduce ORFs, and translate the primary nucleotide sequence into amino acid sequences. Other programs from the MOLGENJR package were used to examine the translated peptide sequences, calculate codon usage and polypeptide M_rs, and plot hydropathy and secondary structure histograms. We used a VAX750 minicomputer, executing program SEQ in the IntelliGenetics Molecular Biology software package, to search for regions of dyad symmetry and to calculate free energies of base pairing in potential DNA hairpin secondary structure. Other unpublished programs and algorithms were used to search for potential activation sequences (ENHANCE2.MSB) and significant open reading frames (ORFREAL.MSB) and to create condensed, dot-matrix hydropathy and secondary structure histograms (AGNAKDCF.MSB). These are available from J.R. Lowe.

RESULTS AND DISCUSSION

(a) Subcloning of the *PA* gene and sequencing strategy

The *PA* gene of *B. anthracis* was originally cloned into pBR322 as a 6-kb insert (Vodkin and Leppla, 1983). Digestion and religation of the recombinant plasmid pSE36 yielded a smaller plasmid with a 4.2-kb insert (pPA26), which retained the *PA* gene (Fig. 1A). *E. coli* transformants of pSE36 and pPA26 both produced proteins of about 83 kDa on SDS–PAGE which reacted specifically with anti-PA antibody on Western-blot analysis and were biologically active in the CHO cell-elongation assay (Leppla, 1984; data not shown). To determine the location and direction of transcription of the *PA* gene, the 4.2-kb insert was excised, digested with *Hin*dIII + *Bam*HI into two fragments of 2.0 kb and 2.2 kb, and sequenced as indicated in Fig. 1,B and C.

(b) Analysis of the coding and regulatory regions of the *PA* gene

The nucleotide sequence of the *PA* gene is shown in Fig. 2. Analysis of the sequence revealed an ORF 2319 bp long. The structural gene for the mature protein began at nt 1891, coding for a glutamic acid residue, and the translated sequence was in agreement with both the N-terminal amino acid sequence and the amino acid composition determined previously. The coding region for this portion of the *PA* gene was 2205 bp long, encoding a 735-aa protein with a calculated M_r of 82 684. The size of the mature PA protein as determined by sequence analysis was similar to that estimated by SDS–PAGE analysis of PA from *Bacillus* culture supernatants, 83–85 kDa (Ivins and Welkos, 1986; Leppla et al., 1985; Vodkin and Leppla, 1983). The final residue of the coding region (glycine) was followed by a TAA stop codon (nt 4096). Thus, as indicated in Fig. 1C, all except the N-terminal 53 aa were encoded within the 2.0-kb *Hin*dIII-*Bam*HI fragment at the 3′ end of the 4.2-kb *B. anthracis* insert. This location of the gene at the end of the insert confirms the position of the *PA* gene mapped in recently isolated *B. subtilis* recombinants (Ivins and Welkos, 1986). In that study, cloning of the *B. anthracis* insert into *B. subtilis* (pUB110) yielded two plasmid recombinants with deletions at the 5′ end of the insert. The smaller recombinant plasmid retained just 2.6-kb of DNA at the 3′ end of the *PA* insert but produced full-sized, functional PA protein (Ivins and Welkos, 1986).

Preceding the sequence encoding the 83-kDa PA protein (starting at nt 1891) were two ATG codons in phase with the ORF, at nt 1834 and 1804. Similar to other *Bacillus* proteins, PA is a secreted protein and is probably synthesized as a precursor having a signal peptide. The methionine codon at nt 1804 appears to be the likely starting point for translation. It would initiate a sequence having several characteristics in common with other *Bacillus* signal sequences that have been identified. The 29-aa peptide that would be encoded is typical of the size of other *Bacillus* signal sequences (Blobel et al., 1979; Kreil, 1981; Lampen and Nielsen, 1982; Mikesell et al., 1983; Wells et al., 1983; Yang et al., 1983). Also, the positively charged, N-terminal 5 aa (Met-Lys-Lys-Arg-Lys), the hydrophobic central region (aa 6–21), and the terminal alanine residue are characteristic of bacterial signal peptides (Blobel et al., 1979; Kreil, 1981; Lampen and Nielsen, 1982; Mikesell et al., 1983; Wells et al., 1983; Yang et al., 1983).

A putative Shine–Dalgarno ribosome-binding site, indicated in Fig. 2, is located 7 bp upstream

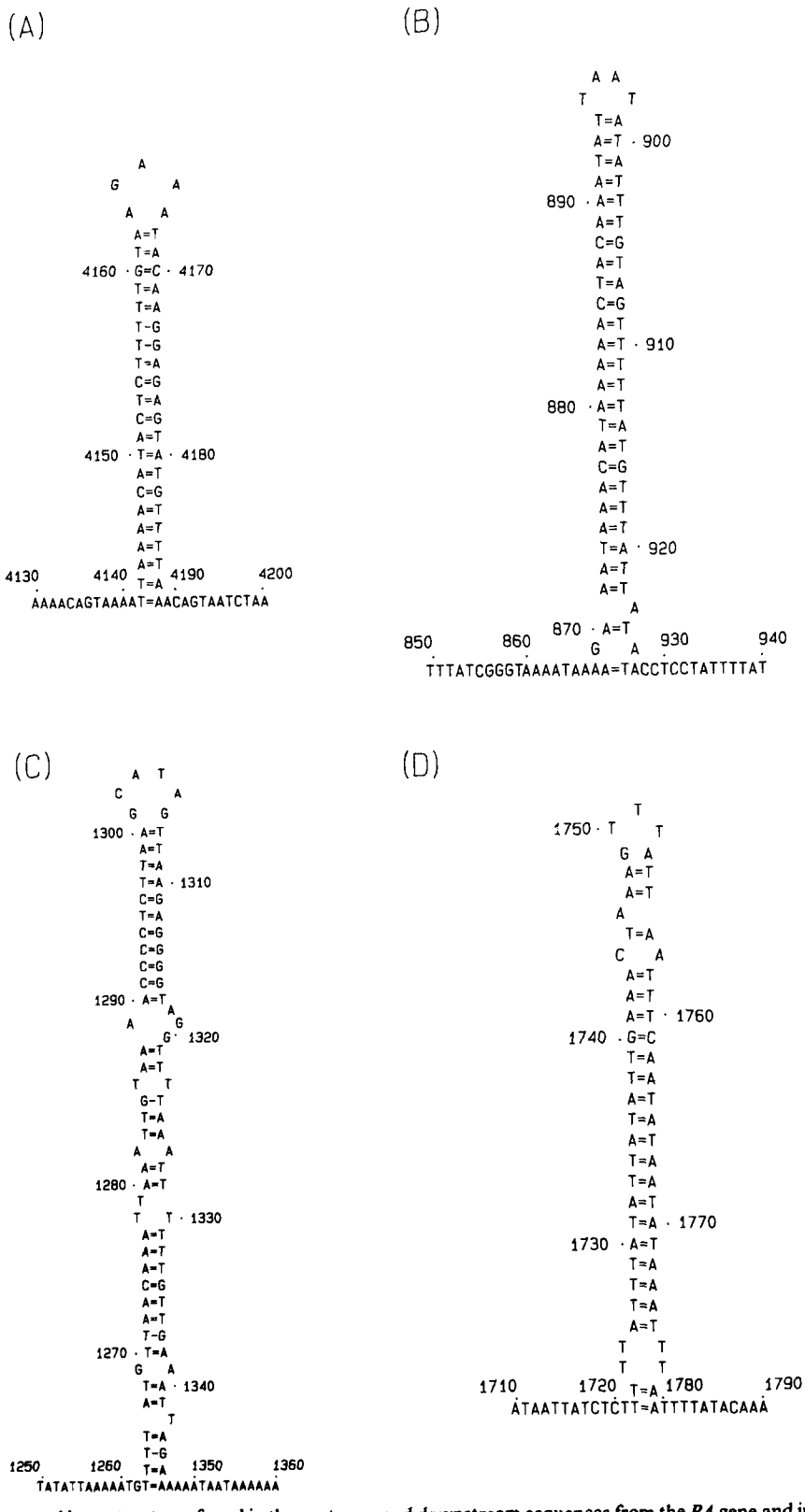


Fig. 3. Possible stem and loop structures found in the upstream and downstream sequences from the *PA* gene and in the putative peptide coding region. Numbering corresponds with that of Fig. 2. The calculated free energies of these conformations were **(A)** $\Delta G_f = -22.2$ kcal/mol, **(B)** $\Delta G_f = -25.4$ kcal/mol, **(C)** $\Delta G_f = -19.6$ kcal/mol, and **(D)** $\Delta G_f = -15.8$ kcal/mol.

from the ATG codon at nt 1804. The sequence of this site, AAAGGAG, and the distance separating it from the start codon, closely resemble the characteristics of the Shine–Dalgarno sites reported for several other *Bacillus* sp. genes (Duvall et al., 1984; McLaughlin et al., 1982; Ohmura et al., 1983; Waalwijk et al., 1985; Yang et al., 1983). The Shine–Dalgarno sequence has a calculated binding energy with *B. subtilis* 16S rRNA of −14.0 kcal/mol (Band and Henner, 1984; McLaughlin et al., 1981; Tinoco et al., 1973). Possible promoter sequences are underlined in Fig. 2. The putative RNA polymerase recognition site (TATAAT) at nt 1764 is identical to the *E. coli* and *B. subtilis* σ^{43} −10 consensus sequence. The 6-bp sequence starting at nt 1738, and separated by 20 bp from the −10 site, resembles the conserved −35 site of *E. coli* and the −35 site reported for genes of Gram-positive organisms (Rosenberg and Court, 1979). The optimal distance between the −10 and −35 RNA polymerase recognition regions in *B. anthracis* genes is unknown. In *E. coli*, these sequences are separated by 16 to 19 bp, with 17 bp being the most frequent and resulting in maximum promoter strength (Rosenberg and Court, 1979). *Bacillus* promoters, especially those recognized by σ^{43}-containing RNA polymerase, are often similar in their sequence and spacing to *E. coli* promoters; however, several different promoter sequences have been identified (Fliss and Setlow, 1984; Johnson et al., 1983; Waalwijk et al., 1985; Wells et al., 1983). Also, distances between the two promoter regions as long as 21 bp have been reported for other sequences, e.g. the pertussis toxin gene (Locht and Keith, 1986). In vitro and in vivo transcription analyses will be necessary to locate the precise promoter region for the *PA* gene.

An inverted repeat forming a potential termination structure was located 3′ of the translation stop codon as shown in Fig. 3A. The putative hairpin structure contained 19 complementary bp and two T-G mismatches between nt 4142 and 4188. The structure had a strong predicted free energy of bp formation (ΔG_f = −22.2 kcal/mol).

We observed three additional regions forming potential stem-and-loop structures having large negative free energies of formation. The sequence from nt 868–926 (Fig. 3B), was inside the 192-codon ORF and had a strong calculated ΔG_f of −25.4 kcal/mol. The second region of dyad symmetry (Fig. 3C), from nt 1263–1346, had a predicted ΔG_f = −19.6 kcal/mol. The third region (Fig. 3D) spanned the *PA* promoter from nt 1722–1779 and had a predicted ΔG_f = −15.8 kcal/mol. If any or all of these regions is recognized as a transcriptional terminator in *E. coli*, their presence could possibly explain the low PA synthesis from the original clones (5–10 ng PA/ml) (Vodkin and Leppla, 1983).

(c) Base composition and codon usage of the *PA* gene

The base composition of the coding strand of the *PA* gene was: A = 39%, T = 30% (A + T = 69% of total), G = 17%, C = 14% (G + T = 31%). The codon usage is shown in Table I. There was a preference for A and T at the third position in the codons, which might reflect the high A + T content. The codon usage was similar to that of another *Bacillus* gene of plasmid origin, the crystal protein toxin gene of the related species *B. thuringiensis* (Schnepf et al., 1985). However, in contrast to the crystal protein gene, the *PA* gene has no codons for cysteine. The codon usage in the *PA* gene differed from that in genes for toxins and other proteins produced by other Gram-positive and Gram-negative bacteria (Table I and data not shown).

(d) Analysis of protein structure from the nucleotide sequence

The prediction of the amino acid sequence of PA and the deduction of protein structural information were performed by algorithms of the computer programs described above (Lowe, 1986, and other unpublished programs). The algorithms used to predict the hydropathic profile and the protein secondary structure are based on the methods of Kyte and Doolittle (1982) and Chou and Fasman (1978), respectively. These predictions for the mature protein as well as putative signal sequence are shown in Fig. 4. The signal sequence (region A of Fig. 4), has a hydrophilic N-terminal end and a hydrophobic central core, as expected from comparisons with similar analyses of other proteins with confirmed signal sequences (not shown).

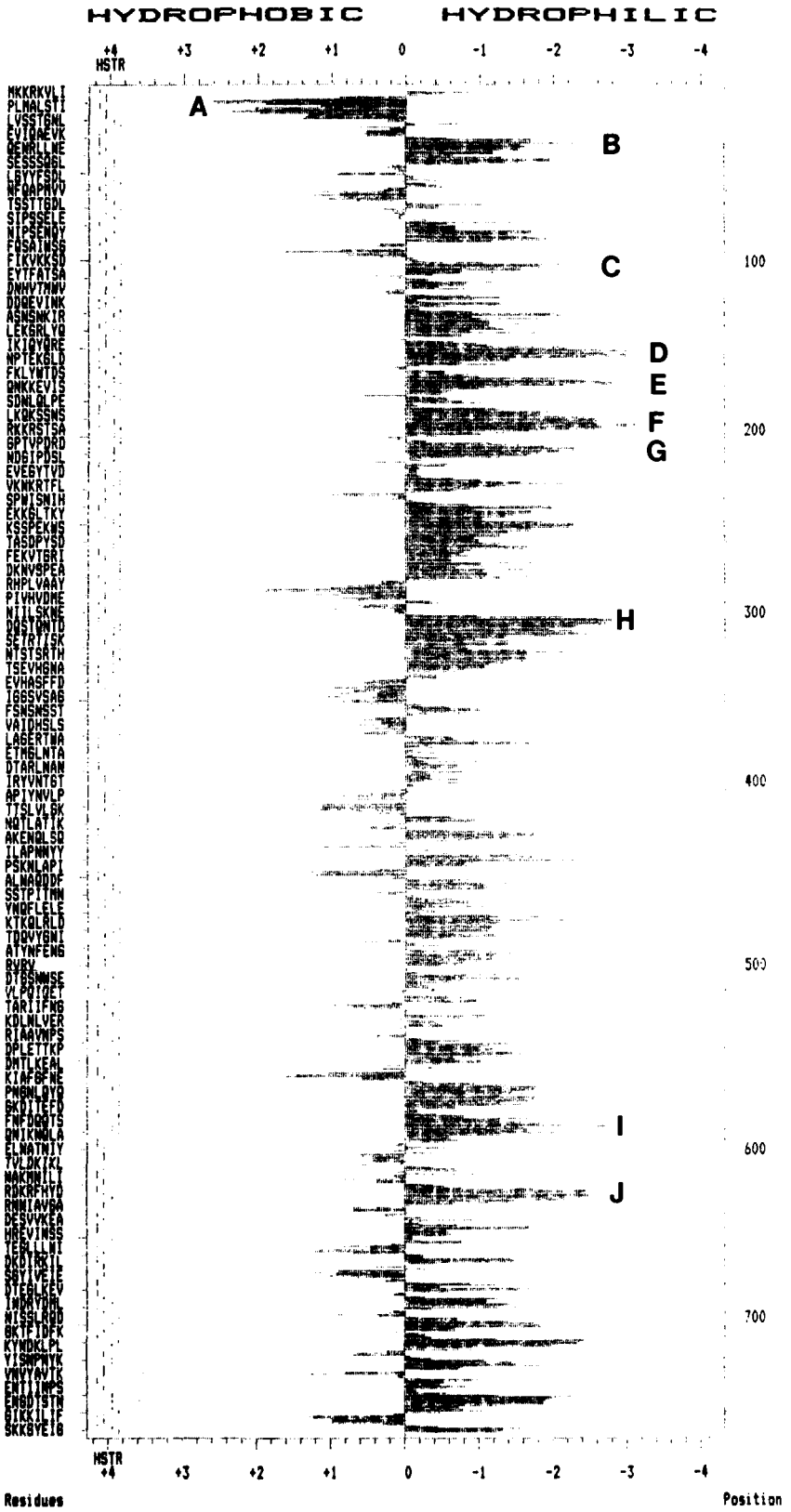


Fig. 4. Kyte–Doolittle hydropathic analysis of 764 residue polypeptide from PA. Combination hydropathy/secondary structure plot of the PA protein. The left margin contains the amino acid sequence of PA. Residue numbers are scaled on the right ordinate. The abscissa units are hydropathy values. Dot positions on the left portion of the plot indicate the most probable secondary structure feature predicted. Headings are: H for helix, S for β-sheet, T for β-turn, and R for random coil. Region A is the hydrophobic signal sequence with its highly charged N terminus. Regions B–J are potential antigenic sites in the sequence. Dot-matrix output was computer-generated with program AGNAKDCF.EXE. Algorithms developed according to the schemes of Kyte and Doolittle (1982) and Chou and Fasman (1978). Polypeptide molecular size = 85 786 Da.

TABLE I

Amino acid composition of PA and codon usage comparisons[a]

		Species and gene[b]								
aa[c]	Codon	B.a. PA	B.a. ORF1	B.t. CryPro	C.t. TetTox	C.d. DTox	S.a. EntB	B.p. PToxS3	E.c. ToxA	V.c. CTxA
Ala	GCU	31.7	12.5	31.3	37.5	42.2	42.9	4.0	28.6	23.1
41[c]	GCC	4.9	0.0	12.5	12.5	11.1	0.0	56.0	0.0	7.7
	GCA	46.3	75.0	32.8	41.7	20.0	57.1	8.0	50.0	53.8
	GCG	17.1	12.5	23.4	8.3	26.7	0.0	32.0	21.4	15.4
Arg	CGU	10.3	0.0	21.3	5.3	35.3	16.7	0.0	18.2	0.0
29	CGC	0.0	0.0	5.3	0.0	5.9	0.0	64.3	0.0	0.0
	CGA	10.3	50.0	13.3	5.3	11.8	33.3	0.0	4.5	0.0
	CGG	13.8	0.0	2.7	0.0	5.9	16.7	7.1	9.1	0.0
	AGA	55.2	50.0	44.0	57.9	17.6	33.3	7.1	54.5	33.3
	AGG	10.3	0.0	13.3	31.6	23.5	0.0	21.4	13.6	66.7
Asn	AAU	76.8	71.4	74.4	81.7	80.0	71.4	0.0	87.5	85.7
69	AAC	23.2	28.6	25.6	18.3	20.0	28.6	100.0	12.5	14.3
Asp	GAU	87.2	93.8	79.4	92.3	71.4	70.8	12.5	62.5	100.0
47	GAC	12.8	6.2	20.6	7.7	28.6	29.2	87.5	37.5	0.0
Cys	UGU	—	—	64.7	75.0	50.0	100.0	0.0	100.0	100.0
0	UGC	—	—	35.3	25.0	50.0	0.0	100.0	0.0	0.0
Gln	CAA	83.9	100.0	81.6	84.6	87.5	100.0	30.0	45.5	100.0
31	CAG	16.1	0.0	18.4	15.4	12.5	0.0	70.0	54.5	0.0
Glu	GAA	74.5	100.0	70.7	82.1	59.5	66.7	71.4	53.8	85.7
51	GAG	25.5	0.0	29.3	17.9	40.5	33.3	28.6	46.2	14.3
Gly	GGU	11.1	41.7	25.0	35.7	37.0	44.4	9.5	35.0	50.0
36	GGC	5.6	8.3	12.5	10.7	13.0	0.0	61.9	15.0	0.0
	GGA	52.8	41.6	45.0	50.0	21.7	44.4	14.3	45.0	50.0
	GGG	30.6	8.3	17.5	3.6	28.3	11.1	14.3	5.0	0.0
His	CAU	90.0	66.7	90.9	85.7	58.8	66.7	25.0	50.0	75.0
10	CAC	10.0	33.3	9.1	14.3	41.2	33.3	75.0	50.0	25.0
Ile	AUU	50.9	50.0	56.3	36.8	36.1	57.1	18.8	38.9	66.7
57	AUC	17.5	0.0	23.9	1.8	25.0	0.0	56.3	0.0	8.3
	AUA	31.6	50.0	19.7	61.4	38.9	42.9	25.0	61.1	25.0
Leu	UUA	67.7	52.6	45.0	60.4	20.0	52.4	0.0	61.1	33.3
62	UUG	12.9	5.3	11.0	13.2	15.0	23.8	12.5	5.6	11.1
	CUU	9.7	10.5	22.0	11.3	25.0	4.8	8.3	16.7	11.1
	CUC	3.2	5.3	2.0	0.0	10.0	4.8	20.8	5.6	0.0
	CUA	4.8	21.0	15.0	11.3	20.0	9.5	0.0	0.0	33.3
	CUG	1.6	5.3	7.0	3.8	10.0	4.8	58.3	11.1	11.1
Lys	AAA	78.3	75.0	72.7	92.0	72.5	76.5	0.0	83.3	63.6
60	AAG	21.7	25.0	27.3	8.0	27.5	23.5	100.0	16.7	36.4
Met	AUG	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
10										
Phe	UUU	79.2	77.8	75.9	95.7	78.9	85.7	0.0	66.7	100.0
24	UUC	20.8	22.2	24.1	4.3	21.1	14.3	100.0	33.3	0.0

TABLE I (continued)

aa[c]	Codon	Species and gene[b] B.a. PA	B.a. ORF1	B.t. CryPro	C.t. TetTox	C.d. DTox	S.a. EntB	B.p. PToxS3	E.c. ToxA	V.c. CTxA
Pro	CCU	31.0	50.0	33.9	41.2	39.1	42.9	0.0	7.7	66.7
29	CCC	10.3	0.0	5.4	5.9	13.0	14.3	18.2	23.1	0.0
	CCA	37.9	50.0	42.9	47.1	26.1	42.9	18.2	53.8	33.3
	CCG	20.7	0.0	17.9	5.9	21.7	0.0	63.6	15.4	0.0
Ser	UCU	30.6	21.4	19.8	45.7	27.8	43.8	0.0	26.3	22.2
72	UCC	4.2	0.0	15.1	2.2	7.4	0.0	30.0	15.8	0.0
	UCA	20.8	35.7	27.9	23.9	13.0	12.5	0.0	31.6	33.3
	UCG	9.7	7.1	7.0	2.2	18.5	18.8	20.0	5.3	11.1
	AGU	31.9	14.3	22.1	17.4	13.0	18.8	0.0	10.5	33.3
	AGC	2.8	21.4	8.1	8.7	20.4	6.3	50.0	10.5	0.0
Thr	ACU	32.8	33.3	28.4	33.3	40.0	64.3	5.9	41.7	30.0
58	ACC	10.3	11.1	16.2	16.7	20.0	0.0	47.1	16.7	10.0
	ACA	37.9	55.6	31.1	50.0	23.3	14.3	5.9	41.7	40.0
	ACG	19.0	0.0	24.3	0.0	16.7	21.4	41.2	0.0	20.0
Trp	UGG	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
7										
Tyr	UAU	78.6	81.8	76.9	88.2	72.2	77.3	42.1	73.9	100.0
28	UAC	21.4	18.2	23.1	11.8	27.8	22.7	57.9	26.1	0.0
Val	GUU	27.9	52.6	25.9	42.9	36.2	42.1	9.1	63.6	33.3
43	GUC	4.7	0.0	14.8	0.0	12.8	5.3	54.5	9.1	16.7
	GUA	39.5	42.1	40.7	50.0	31.9	31.6	18.2	18.2	50.0
	GUG	27.9	5.3	18.5	7.1	19.1	21.1	18.2	9.1	0.0
M_r		85787	21610	133047	65900	60753	31433	24989	29862	13909

[a] Within-group percentage codon usage calculated with MOLGENJR software package (Lowe, 1986).

[b] The following genes from the species listed were examined.

B.a. PA, *Bacillus anthracis* protective antigen gene (*PA*).

B.a. ORF1, *Bacillus anthracis* hypothetical protein gene 1 on pXO1 plasmid.

B.t. CryPro, *Bacillus thuringiensis* crystal protein gene (Sanger and Coulson, 1977)

C.t. TetTox, *Clostridium tetani* tetanus toxin gene (Fairweather et al., 1986).

C.d. DTox, *Corynebacterium diphtheriae* diphtheria toxin gene (Greenfield et al., 1983).

S.a. EntB, *Staphylococcus aureus* enterotoxin B gene (Jones and Khan, 1986).

B.p. PToxS3, *Bordetella pertussis* pertussis toxin S3 binding subunit gene (Locht and Keith, 1986).

E.c. ToxA, *Escherichia coli* heat-labile enterotoxin A gene (Yamamoto et al., 1984).

V.c. CTxA, *Vibrio cholerae* cholera toxin alpha subunit gene (Mekalanos et al., 1983).

[c] Total number of specific aa residues deduced from the nucleotide sequence of the *PA* gene is shown below each aa.

(e) Regions of the sequence upstream from the *PA* gene

Other ORFs, in addition to the longest one of 2319 bp encoding PA, were found in the 4.2-kb sequence. The only ORF at least 100 codons long was a 576-nt sequence (ORF1) beginning with an ATG at nt position 416 upstream of the *PA* gene. The 192-codon ORF encodes a polypeptide with a calculated M_r of 21610 Da. The codon usage of the translated region is similar to that observed for the *PA* gene (Table I). A computer analysis (ORFREAL.MSB)

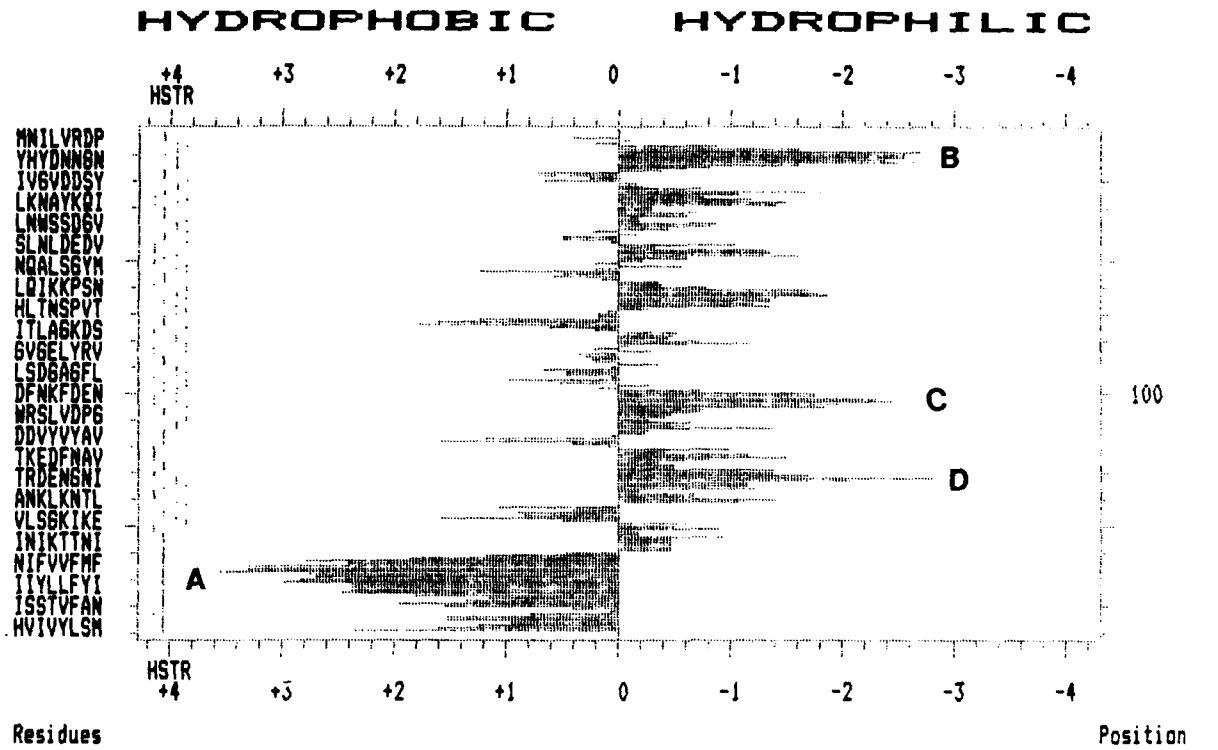


Fig. 5. Kyte–Doolittle hydropathic analysis of 192 residue polypeptide from PA260RF1. Combination hydropathy/secondary structure plot of the putative peptide. Plot organization and generation are the same as described in Fig. 4. Region A is a long, highly hydrophobic 28-aa sequence with mostly β-sheet structure predicted. Regions B–D are potential antigenic sites in the sequence. Polypeptide molecular size = 21 609 Da.

of this ORF according to the method of Fickett (1982) calculated a 92% coding probability. A similar analysis of the PA-coding region also gave a 92% coding probability. Potential −10 and −35 RNA polymerase recognition sites, but no consensus Shine–Dalgarno site, on the 5′ side of the cryptic ORF were identified. The ORF terminated with a TAG stop codon. Fig. 5 is a hydropathy plot and secondary structure analysis of the putative protein. The sequence does not appear to encode a signal peptide but does have an interesting C terminus rich in hydrophobic residues embedded in a region with a high probability for β-sheet structure. This suggests that the protein could be membrane-bound at its C terminus. The significance of this putative gene is unknown and awaits analysis of expression experiments using the cloned plasmid DNA.

(f) Conclusions

The nucleotide sequence of the ORF encoding PA, one of the three protein components of anthrax toxin, was determined. A region encoding a putative 29-bp signal sequence, regulatory sequences upstream from the *PA* gene, and a newly identified ORF were also deduced. The codon usage in the *PA* gene differed from that in genes for other bacterial proteins compared, except for a crystal protein toxin gene in the related species *B. thuringiensis*. The availability of the complete nucleotide sequence of the *PA* gene of *B. anthracis* will serve several useful purposes. For example, the *PA* promoter sequence is being probed by promoter-probing vectors and the sequence being altered by site-specific mutagenesis. Thus, enhanced production of cloned PA in the *B. subtilis* or *E. coli* hosts will become feasible. Also, specific mutagenesis of the PA-coding region will be done to: (1) produce immunogenic, biologically inactive cross-reactive proteins for vaccine studies (Hambleton et al., 1984; Little and Knudson, 1986; Turnbull et al., 1986); and (2) examine the role of different domains of PA on binding to target cell membranes and to the EF and LF components of anthrax toxin (Friedlander, 1986; Leppla, 1984;

O'Brien et al., 1985). Finally, segments of the *PA* nucleotide sequence will be used as probes to examine the genetic organization of PA in variant strains of *B. anthracis* (Little and Knudson, 1986; Turnbull et al., 1986).

ACKNOWLEDGEMENTS

We thank D. Becker for assistance in processing the manuscript. The views of the authors do not purport to reflect the positions of the Department of the Army or the Department of Defense. Approved for public release; distribution unlimited. Clearance date: 29 January, 1988.

REFERENCES

Band, L. and Henner, D.J.: *Bacillus subtilis* requires a 'stringent' Shine–Dalgarno region for gene expression. DNA 3 (1984) 17–21.

Beall, F.A., Taylor, M.J. and Thorne, C.B.: Rapid lethal effect in rats of a third component found upon fractionating the toxin of *Bacillus anthracis*. J. Bacteriol. 83 (1962) 274–1280.

Blobel, G., Walter, P., Chang, C.N., Goldman, B.M., Erikson, A.H. and Lingappa, V.R.: Translocation of proteins across membranes: the signal hypothesis and beyond. Symp. Soc. Exp. Biol. 33 (1979) 9–36.

Chou, P.Y. and Fasman, G.D.: Prediction of the secondary structure of proteins from their amino acid sequence. In Meister, A. (Ed.), Advances in Enzymology and Related Areas of Molecular Biology, Vol. 47. Wiley, New York, 1978, pp. 45–148.

Duvall, E.J., Williams, D.M., Mongkolsuk, S. and Lovett, P.S. Regulatory regions that control expression of two chloramphenicol-inducible *cat* genes cloned in *Bacillus subtilis*. J. Bacteriol. 158 (1984) 784–790.

Fairweather, N.F., Lyness, V.A., Pickard, D.J., Allen, G. and Thomson, R.O.: Cloning, nucleotide sequencing, and expression of tetanus toxin fragment C in *Escherichia coli*. J. Bacteriol. 165 (1986) 21–27.

Fickett, J.W.: Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. 10 (1982) 5303–5318.

Fliss, E.R. and Setlow, P.: Complete nucleotide sequence and start sites for transcription and translation of the *Bacillus megaterium* protein C gene. J. Bacteriol. 158 (1984) 809–813.

Friedlander, A.M.: Macrophages are sensitive to anthrax lethal toxin through an acid-dependent process. J. Biol. Chem. 261 (1986) 7123–7126.

Green, B.D., Battisti, L., Koehler, T.M., Thorne, C.B. and Ivins, B.E.: Demonstration of a capsule plasmid in *Bacillus anthracis*. Infect. Immun. 49 (1985) 291–297.

Greenfield, L., Bjorn, M.J., Horn, G., Fong, D., Buck, G.A., Collier, R.J. and Kaplan, D.A.: Nucleotide sequence of the structural gene for diphtheria toxin carried by corynebacteriophage β. Proc. Natl. Acad. Sci. USA 80 (1983) 6853–6857.

Grosjean, H. and Fiers, W.: Preferential codon usage in prokaryotic genes: the optimal codon–anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18 (1982) 199–209.

Hambleton, P., Carman, J.A. and Melling, J.: Anthrax: the disease in relation to vaccines. Vaccine 2 (1984) 125–132.

Ivins, B.E. and Welkos, S.L.: Cloning and expression of the *Bacillus anthracis* protective antigen gene in *Bacillus subtilis*. Infect. Immun. 54 (1986) 537–542.

Johnson, W., Moran, E. and Losick, R.: Two RNA polymerase sigma factors from *Bacillus subtilis* discriminate between overlapping promoters for a developmentally regulated gene. Nature 302 (1983) 800–804.

Jones, C.L. and Khan, S.A.: Nucleotide sequence of the enterotoxin B gene from *Staphylococcus aureus*. J. Bacteriol. 166 (1986) 19–33.

Kreil, G.: Transfer of proteins across membranes. Annu. Rev. Biochem. 50 (1981) 317–348.

Kyte, J. and Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157 (1982) 105–132.

Lampen, J.O. and Nielsen, J. B.K.: Penicillinase and the secretion of proteins by *Bacilli*. In Ganesan, A.T., Chang, S. and Hoch, J.A. (Eds.), Molecular Cloning and Gene Regulation in *Bacilli*. Academic Press, New York, 1982, pp. 99–109.

Leppla, S.H.: *Bacillus anthracis* calmodulin-dependent adenylate cyclase: chemical and enzymatic properties and interactions with eucaryotic cells. In Greengard, P. and Robinson, G.A. (Eds.), Advances in Cyclic Nucleotide and Protein Phosphorylation Research. Raven Press, New York, 1984, pp. 189–198.

Leppla, S.H., Ivins, B.E. and Ezzell, J.W.: Anthrax toxin. In L. Leive (Ed.), Microbiology — 1985. American Society for Microbiology, Washington, DC, 1985, pp. 63–66.

Lincoln, R.E. and Fish, D.C.: Anthrax toxin. In Montie, T.C., Kadis, S. and Ajl, S.I. (Eds.), Microbial Toxins, Vol. 3. Academic Press, New York, 1970, pp. 361–414.

Little, S.F. and Knudson, G.B.: Comparative efficacy of *Bacillus anthracis* live spore vaccine and protective antigen vaccine against anthrax in the guinea pig. Infect. Immun. 52 (1986) 509–512.

Locht, C. and Keith, J.M.: Pertussis toxin gene: nucleotide sequence and genetic organization. Science 232 (1986) 1258–1264.

Lowe, J.R.: Transportable microcomputer programs for DNA and protein analysis. Fed. Proc. 45 (1986) 1582.

Maniatis, T., Fritsch, E.F. and Sambrook, J.: Molecular Cloning. A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1982.

McLaughlin, J.R., Murray, C.L. and Rabinowitz, J.G.: Unique features in the ribosomal binding site sequence of the Gram-positive *Staphylococcus aureus* B-lactamase gene. J. Biol. Chem. 256 (1981) 11283–11291.

McLaughlin, J.R., Chang, S.-Y. and Chang, S.: Transcriptional analyses of the *Bacillus licheniformis penP* gene. Nucleic Acids Res. 10 (1982) 3905–3919.

Mekalanos, J.J., Swartz, D.J., Pearson, G.D.N., Harford, N., Groyne, F. and De Wilde, M.: Cholera toxin genes: nucleotide sequence, deletion analysis and vaccine development. Nature 306 (1983) 551–557.

Messing, J.: New M13 vectors for cloning. Methods Enzymol. 101 (1983) 20–78.

Messing, J., Crea, R. and Seeburg, P.: A system for shotgun DNA sequencing. Nucleic Acids Res. 9 (1981) 309–321.

Mikesell, O.P., Ivins, B.E., Ristroph, J.D. and Dreier, T.M.: Evidence for plasmid-mediated toxin production in *Bacillus anthracis*. Infect. Immun. 39 (1983) 371–376.

O'Brien, J.O., Friedlander, A., Drier, T., Ezzell, J. and Leppla, S.: Effects of anthrax toxin components on human neutrophils. Infect. Immun. 47 (1985) 306–310.

Ohmura, K., Yamazaki, H., Takeichi, Y., Nakayama, A., Otozai, K., Yamane, K., Yamasaki, M. and Tamura, G.: Nucleotide sequence of the promoter and NH_2-terminal signal peptide region of *Bacillus subtilis* α-amylase gene cloned in pUB110. Biochem. Biophys. Res. Commun. 112 (1983) 678–683.

Palva, I., Pettersson, R.F., Kalkkinen, N., Lehtovaara, P., Sarvas, M., Soderlund, H., Takkinen, K. and Kaariainen, L.: Nucleotide sequence of the promoter and NH_2-terminal signal peptide region of the α-amylase gene from *Bacillus amyloliquefaciens*. Gene 15 (1981) 43–51.

Robertson, D.L. and Leppla, S.H.: Molecular cloning and expression in *Escherichia coli* of the lethal factor gene of *Bacillus anthracis*. Gene 44 (1986) 71–78.

Rosenberg, M. and Court, D.: Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. 13 (1979) 319–353.

Sanchez-Pescador, R. and Urdea, M.: Use of unpurified synthetic oligodeoxyribonucleotide primers for rapid deoxynucleotide chain termination sequencing. DNA 3 (1984) 339–343.

Sanger, F., Nicklen, S. and Coulson, A.R.: DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74 (1977) 5463–5467.

Schnepf, H.E., Wong, H.C. and Whiteley, H.R.: The amino acid sequence of a crystal protein from *Bacillus thuringiensis* deduced from the DNA base sequence. J. Biol. Chem. 260 (1985) 6264–6272.

Tinoco Jr., I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J.: Improved estimation of secondary structure in ribonucleic acids. Nature 246 (1973) 40–41.

Turnbull, P.C.B., Broster, M.G., Carman, J.A., Manchee, R.J. and Melling, J.: Development of antibodies to protective antigen and lethal factor components of anthrax toxin in humans and guinea pigs and their relevance to protective immunity. Infect. Immun. 52 (1986) 356–363.

Vodkin, M.H. and Leppla, S.H.: Cloning of the protective antigen gene of *Bacillus anthracis*. Cell 34 (1983) 693–697.

Waalwijk, C., Dullemans, A.M., Van Workum, M.E.S. and Visser, B.: Molecular cloning and the nucleotide sequence of the M_r 28000 crystal protein gene of *Bacillus thuringiensis* subsp. *israelensis*. Nucleic Acids Res. 13 (1985) 8207–8217.

Wells, J.A., Ferrari, E. Henner, D.J., Estell, D.A. and Chen, E.Y.: Cloning, sequencing, and secretion of *Bacillus amyloliquefaciens* subtilisin in *Bacillus subtilis*. Nucleic Acids Res. 11 (1983) 7911–7925.

Yamamoto, T., Tamura, T. and Yokota, T.: Primary structure of heat-labile enterotoxin produced by *Escherichia coli* pathogenic for humans. J. Biol. Chem. 259 (1984) 5037–5044.

Yang, M., Galizzi, A. and Henner, D.: Nucleotide sequence of the amylase gene from *Bacillus subtilis*. Nucleic Acids Res. 11 (1983) 237–249.

Communicated by R.E. Yasbin.

N.B. The views of the author do not purport to reflect the positions of the Department of the Army or the Department of Defense (Para. 4-3, AR 360-5).