

Nucleotide sequence of a novel δ -endotoxin gene *cryIg* of *Bacillus thuringiensis* ssp. *galleriae*

S.V. Smulevitch, A.L. Osterman, A.B. Shevelev, S.V. Kaluger, A.I. Karasin, R.M. Kadyrov, O.P. Zagnitko, G.G. Chestukhina and V.M. Stepanov

Institute of Microbial Genetics, Lab. of Protein Chemistry, Moscow, 113545, USSR, 1st Dorozhny, 1

Received 18 August 1991

A gene *cryIg* coding for entomocidal protein δ -endotoxin of *Bacillus thuringiensis* ssp. *galleriae* str. 11-67 named CryIG has been cloned and sequenced (EMBL accession number X58120). The deduced amino acid sequence that contains 1156 amino acid residues shows only 28% of identical residues, when compared with other δ -endotoxins of the CryI family. The extent of identity is substantially higher for some regions of the sequence ('conserved blocks'), that presumably bear important structural or functional properties. This implies that CryIG δ -endotoxin follows the same type of polypeptide chain folding as other CryI proteins, whereas peculiarities of primary structure help to explain its unique specificity.

CryIG; δ -Endotoxin; Primary structure; Genomic library; *Bacillus thuringiensis*

1. INTRODUCTION

Entomocidal protein δ -endotoxins are intensively studied in many laboratories. Their ability to selectively kill only certain species of insects creates a basis for their utilization in plant protection. Nevertheless the mode of the toxic action remains unknown.

Proteinaceous crystals of *B. thuringiensis* ssp. *galleriae* str. 11-67 are composed of at least two δ -endotoxins of $M_r \sim 130$ kDa differing strongly in their immunological properties and entomocidal specificity, named 'positive' and 'negative' components according to the electrophoretic behavior of the respective true toxins [1]. Earlier we reported cloning of the gene coding for the 'positive' component, the endotoxin with unique specificity towards larvae of *Galleria mellonella* [2]. The partial amino acid sequence of this protein was determined by

Edman's method. Here we present the complete sequence of this gene named *cryIg*, according to the principles of δ -endotoxin classification proposed in [3].

2. METHODS

2.1. *cryIg* gene cloning

A genomic library of *Bacillus thuringiensis* ssp. *galleriae* str. 11-67 was obtained on phasmid vector λ -pSL5, especially designed for prokaryotic genomic library construction, using partial *EcoRI* digestion of total DNA [5]. Immunoscreening was performed with affinity-purified monospecific antiserum against pure CryIG protein. Restriction mapping of three immunopositive clones and preliminary localization of the toxin gene were reported earlier [1]. pOC10 phasmid, containing the full-size gene was used in further cloning procedures. The phage part of the pOC10 was removed by *XbaI* deletion, yielding pOK10 plasmid. Exact mapping and localization of the *cryIg* gene were performed using immunological tests (Fig. 1). *KpnI-XbaI* insert, carrying the full-size gene, the 521 base-pair 5'-flank and the 1.3 kb 3'-flank was cloned into the pUC18 cloning vector, resulting in the

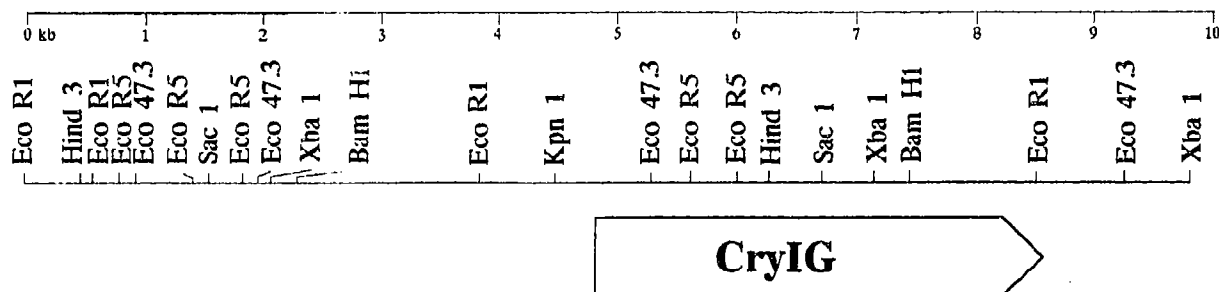


Fig. 1. Structural organization of the insert of pOK10 plasmid, containing full-size CryIG gene.

Correspondence address: A.L. Osterman, Institute of Microbial Genetics, Lab. of Protein Chemistry, Moscow, 113545, USSR, 1st Dorozhny, 1. Fax: (7) (095) 315 05 01.

Kpn1 10 20 30 40 50 60 70 80 90 100 110 120
GGTACCTTCACTGTATTGTTTCACCTCCATCAGGTTTCAAATTGAAAACTAAATGTTCTACGATGTAAGCTATTATTTATGACGAAAGATACGTGTA AAAATCGTATTGAGATTGA

130 140 150 160 170 180 190 200 210 220 230 240
TGAATGTAGATAAGCAGAAATCAACTTACAAGTATTAGGGATTTCCTCGACCCGTTTTATACAAAACAGTATACTGCGATGGGTAGATATTCGTCATTTCCTTAAGTTTTAAG

250 260 270 280 290 300 310 320 330 340 350 360
ATGCAGTATATGTAGATCATATTTAAATATAGCGTTCATCACTACTATATTATGTAATTGACGGTAATCATATTCAAGTGCCTGATTTACAAATCAAAATAATGAAAGAACATGCTC

370 380 390 400 410 420 430 440 450 460 470 480
AATCTGCTCAAATATCTGTTTTTATTTATGAGTAAAAACCGAAGGTTGTGGAACGTAACGTAAATAAACGTAATGGCGAAGATATATAGATGTC AATATAAAAAAGTTAACCCAAA

490 500 SD-seq. 520 530 540 550 560 570 580 590 600
TAATGTTTTAAAAATTTAAAAATAATGTAGGAGGAAAAATTAATAATAAAACACGGAATATTGCGCTTCCAATTGTGGTGTGCATCTGATGATGTTCCGAAATATCCTTTAG
M N Q N K H G I I G A S N C G C A S D D V A K Y P L

610 620 630 640 650 660 670 680 690 700 710 720
CCAACAATCCATATTCATCTGCTTTAAATTTAAATCTTGCAAAAATAGTAGTATTCTCAACTGGATTAAACATAATAGCGGATGCAGCAAAAGAACAGTATCTATTGGBACAACCATAG
A N N P Y S S A L N L N S C Q N S S I L N W I N I I G D A A K E A V S I G T T I

730 740 750 760 770 780 790 800 810 820 830 840
TCTCTTATACAGCACCTTCTTACTGGATTAATTTCAATAGTATATGACCTTATAGTAAAGTACTAGGAGTAGTAGTGACAAATC CATATCAGATTTGCTATATGTACTTAT
V S L I T A P S L T G L I S I V Y D L I G K V L G G S S G Q S I S D L S I C D L

850 860 870 880 890 900 910 920 930 940 950 960
TATCTATTATGATTACGGGTAAGTCAGAGTGTAAAAATGATGGGATTCAGATTTTAAATGTTCTGTACTCTTATACAGGAACATTTAGAGGCTCTGGATAGCTGGAATAAGAATC
L S I I D L R V S Q S V L N D G I A D F N G S V L L Y R N Y L E A L D S W N K N

970 980 990 1000 1010 1020 1030 1040 1050 1060 1070 1080
CTAATTCGCTTCTGCTGAGAACCTCGTACTCGTTTTAGAAATCGCCGACTCAGAAATTTGATAGAATTTAACCCGAGGGTCTTTAACGAATGGTGGCTCGTTAGCTAGACAAAATGCC
P N S A S A E E L R T R F R I A D S E F D R I L T R G S L T N G G S L A R Q N A

1090 1100 1110 1120 1130 1140 1150 1160 1170 1180 1190 1200
AAATATTATTATACCTTCTTTTGGAGGCGTGCATTTTCCATTTATTACTACTAAGGGATGCTACTAGATATGGCACTAATTTGGGGGCTATACAATGCTACACCTTTTATAAAATATC
Q I L L L P S F A S A A F F H L L L L R D A T R Y G T N W G L Y N A T P F I N Y

1210 1220 1230 1240 1250 1260 1270 1280 1290 1300 1310 1320
AATCAAACTAGTAGAGCTTATTGAATATACTGATATTGCGTACATGGTATAATCGAGGTTTCAACGAACCTAAGACAACGAGGCAC TAGTGC TACAGC TTGGTTAGAATTCATA
Q S K L V E L I E L Y T D Y C V H W Y N R G F N E L R Q R G T S A T A W L E F H

1330 1340 1350 1360 1370 1380 1390 1400 1410 1420 1430 1440
GATATCGTAGAGAGATGACATTGATGGTATTAGATATAGTAGCATCATTTCAGTCTTGATATTACTAAITACCAATAGAAACAGATTTTCAGTTGAGTAGGGTCAITTTACAGATC
R Y R R E M T L M V L D I V A S F S S L D I T N Y P I E T D F Q L S R V I Y T D

1450 1460 1470 1480 1490 1500 1510 1520 1530 1540 1550 1560
CAATGGTTTTGTACATCGTAGTCTTAGGGGAGAAAGTTGGTTTAGCTTTGTTAATAGAGCTAATTTCTCAGATTTAGAAAATGCAATACCTAATCCTAGACCGTCTTGGTTTTTAA
P I G F V H R S S L R G E S W F S F V N R A N F S D L E N A I P N P R P S W F L

1570 1580 1590 1600 1610 1620 1630 1640 1650 1660 1670 1680
ATAATGATTATATCTACTGGTTCACCTTACATTCGCCGGTAGCCCAAGTACTGATAGAGCGAGGGTATGGTATGGAAGTCGAGATCGAATTTCCCTGCTAATTCACAATTTATTACTG
N N M I I S T G S L T L P V S P S T D R A R V W Y G S R D R I S P A N S Q F I T

1690 1700 1710 1720 1730 1740 1750 1760 1770 1780 1790 1800
AACTAATCTCGGACAACATACGACTGCTACACAAAATATTTAGGGCGAAATATATTTAGAGTAGATTCACAGCTGTAATTTAAATGATACCACATATGGAGTGAATAGGGCGGTAT
E L I S G Q H T T A T Q T I L G R N I F R V D S Q A C N L N D T T Y G V N R A V

1810 1820 1830 1840 1850 1860 1870 1880 1890 1900 1910 1920
TTTATCATGATCGGAGTGAAGGTTCTCAAAGATCCGTGACGAGGGTATATTCGAAACACTGGGATAGATAACCCTAGAGTTCAAATATTAACACTTATTTACCTGGAGAAAATTCAG
F Y H D A S E G S Q R S V Y E G Y I R T T G I D N P R V Q N I N T Y L P G E N S

1930 1940 1950 1960 1970 1980 1990 2000 2010 2020 2030 2040
ATATCCAACTCCAGAACTATACATATATTAAGCACAACAATAAATTTAACAGGAGGACTTAGACAAGTAGCATCTAATCCCGTTCATCTTTAGTAATGTATGGTTGGACACATA
D I P T P E D Y T H I L S T T I N L T G G L R Q V A S N R R S S L V M Y G W T H

2050 2060 2070 2080 2090 2100 2110 2120 2130 2140 2150 2160
AAAGTCTGGCTCGTAACAATACCAITTAATCCAGATAGAATTACACAGATACCATTGACGAAGGTTGATACCCGAGGCACAGGTGTTCTTATGTAATGATCCAGGTTTATAGGAGGAG
K S L A R N N T I N P D R I T Q I P L T K V D T R G T G V S Y V N D P G F I G G

2170 2180 2190 2200 2210 2220 2230 2240 2250 2260 2270 2280
CTCTACTTCAAAGGACTGACCATGGTTCGCTGGAGTATTGAGGGTCCAATTTCCACTTCACTTAAGACAACAATATCGTATTAGAGTCCGTTATGCTTCAACAACAATATTCGATTGA
A L L Q R T D H G S L G V L R V Q F P L H L R Q Q Y R I R V R Y A S T T N I R L

586

```

2290 2300 2310 2320 2330 2340 2350 2360 2370 2380 2390 2400
GTGTGAATGGCAGTTTCGGTACTATTCTCAAAATCTCCTAGTACAATGAGATTAGGAGAGGATTAAAGATACGGATCTTTTGTATAAGAGAGTTTAATACTTCTATTAGACCCACTG
S V N G S F G T I S Q N L P S T M R L G E D L R Y G S F A I R E F N T S I R P T
826
2410 2420 2430 2440 2450 2460 2470 2480 2490 2500 2510 2520
CAAGTCCGGACCAAATTCGATTGACAATAGAACCATCTTTATTAGACAAGAGGCTTATGTAGATAGAATTGAGTTCATTCCAGTTAATCCGACCGAGAGGCCGAAAGAGGATCTAGAAG
A S P D Q I R L T I E P S F I R Q E V Y V D R I E F I P V N P T R E A K E D L E
866
2530 2540 2550 2560 2570 2580 2590 2600 2610 2620 2630 2640
CAGCAAAAAAGCGGTGGCGAGCTTGTTCACGCACAAGGGACGATTACAAGTAAATGTGAAAGATTATCAAGTCGATCAAGCGGCAAATTTAGTGTATGCTTATCAGATGAACAA
A A K K A V A S L F T R T R D G L Q V N V K D Y Q V D Q A A N L V S C L S D E Q
876
2650 2660 2670 2680 2690 2700 2710 2720 2730 2740 2750 2760
ATGGGTATGACAAANAGATGTTATTGGAGCGGTACGTCCGCAAAACGACTTAGCCGAGAACGCAACTTACTTCAGGATCCAGATTTAATACAATCAATAGTACAGAAGAAATGGAT
Y G Y D K K M L L E A V R A A K R L S R E R N L L Q D P D F N T I N S T E E N G
886
2770 2780 2790 2800 2810 2820 2830 2840 2850 2860 2870 2880
GGAAAGCAAGTAACGGCGTTACTATTAGTGGGGCGGGCCATTCATAAAGGCGTGCATTCAGCTAGCAAGTGCACGAGAAAATACCCAAACATACATCTATCAAAAAGTAGATGCAT
W K A S N G V T I S E G G P F Y K G R A I Q L A S A R E N Y P T Y I Y Q K V D A
896
2890 2900 2910 2920 2930 2940 2950 2960 2970 2980 2990 3000
CGGAGTTAAAGCGTATACACGTTATAGACTGGATGGGTTCTGTGAAAGTAGTCAAGATTAGAAATGATCTCATTACCATCATAAAGTCCATCTTGTGAAAAATGTACCAGATAATT
S E L K P Y T R Y R L D G F V K S S Q D L E I D L I H H H K V H L V K N V P D N
906
3010 3020 3030 3040 3050 3060 3070 3080 3090 3100 3110 3120
TAGTATCTGACTTACCAGATGATCTTGTAGTGGAAATCAATCGATGTGAGGAAACAAGATGGTAAATGCGCAACTGGAAACAGAGCATCATCATCCGATGGATGCTGTGAAAGCAG
L V S D T Y P D D S C S G I N R C Q E Q Q M V N A Q L E T E H H H P M D C C E A
916
3130 3140 3150 3160 3170 3180 3190 3200 3210 3220 3230 3240
CTCAAAACATGAGTTTCTTCTATATTGATACAGGGATTTAAATTCGAGTGTAGACCAGGGAATCTGGGCGATCTTAAAGTTCGAAACAACCGATGGTTATGCGACGTTAGGAAATC
A Q T H E F S S Y I D T G D L N S S V D Q G I W A I F K V R T T D G Y A T L G N
926
3250 3260 3270 3280 3290 3300 3310 3320 3330 3340 3350 3360
TTGAATGGTAGAGGTCCGACCGTTCGGGTGAATCTTTAGAACSTGAACAAGGGATAATACAAAATGGAGTGCAGAGCTAGGAAGAAAGCGTGCAGAAACAGATCGCGGTGATCAAG
L E L V E V G P L S G E S L E R E Q R D N T K W S A E L G R K R A E T D R V Y Q
936
3370 3380 3390 3400 3410 3420 3430 3440 3450 3460 3470 3480
ATGCCAAACAATCCATCAATCATTATTTGTGGATTATCAAGATCAACAATTAATCCAGAATAGGGATGGCAGATATATGGACGCTCAAAAATCTTGTCCGATCAATTTAGATGTAT
D A K Q S I N H L F V D Y Q D Q Q L N P E I G M A D I M D A Q N L V A S I S D V
946
3490 3500 3510 3520 3530 3540 3550 3560 3570 3580 3590 3600
ATAGCGATGCCGTACTGCAAAATCCCTGGAATTAACATGAGATTACACAGAGCTGTCCAATCGCTTACAACAAGCATCGTATCTGTATACGCTCGAAATCGGGTGCAAAATGGGACT
Y S D A V L Q I P G I N Y E I Y T E L S N R L Q Q A S Y L Y T S R N A V Q N G D
956
3610 3620 3630 3640 3650 3660 3670 3680 3690 3700 3710 3720
TTAACAACGGCTAGATAGCTGGAATGCAACAGCGGGTGCATCGGTACAACAGGATGGCAATACGCAATTTCTTAGTCTTCTTCTCATTGGGATGCACAAGTTTCTCAACAATTTAGAGTGC
F N N G L D S W N A T A G A S V Q Q D G N T H F L V L S H W D A Q V S Q Q F R V
966
3730 3740 3750 3760 3770 3780 3790 3800 3810 3820 3830 3840
AGCCGAATGTAATATGATTACGTGTAAACAGCAGAGAAAGTAGCGCGGAGACGGATACGCTACTATCCGGATGATGCTCATACACAGAAACGCTTACATTTAATGCATGTGATT
Q P N C K Y V L R V T A E K V G G G D G Y V T I R D D A H H T E T L T F N A C D
976
3850 3860 3870 3880 3890 3900 3910 3920 3930 3940 3950 3960
ATGATATAAATGGCAGTACGTGACTGATAAATACGTATCAACAAAAGAGTGGTATTCATCCGGAGACACAACACATGTGGGTAGAGGTAATGAAACAGAAAGGTGCATTTTCATATAG
Y D I N G T Y V T D N T Y L T K E V V F H P E T Q H M W V E V N E T E G A F H I
986
3970 EcoRI 3980 3990 4000 4010 4020 4030 4040 4050 4060 4070 4080
ATAGTATTGAATTCGTTGAAACAGAAAAGTAACGGGATGATGTTCCGAACATATAAGGTATAAGGACGATACCGCGTATAAAGATCTCAAAACAGAAATGTGAAATAATGAGGACCCTC
D S I E F V E T E K *
996
4090 4100 4110 4120 4130 4140 4150
CGGTAGTCGTACATGAAAGTACACGACTACCGGAGGGTATTTTATATAAAAATGTGGTTTTCACACGGTCT
IR-> <-IR

```

Fig. 2. The nucleotide sequence of a 4156 bp fragment of *Bacillus thuringiensis* ssp. *galleriae* str. 11-67 DNA and the deduced amino acid sequence of CryIG δ -endotoxin. The N-terminal amino acid sequence determined with the Edman method is doubly underlined. Inverted repeat in 3'-flanking region is underlined. Also shown are the SD-sequence, *Kpn*I and *Eco*RI sites.

pKP7 plasmid, which allowed expression of *cryIlg* gene in *E. coli* using the Lac-promoter. Immunological and toxicity assays of the expression product confirmed its identity with CryIG toxin.

2.2. Sequence analysis

Sequencing of pKP7 was carried out by the SEQUENASE version [4] of Sanger dideoxynucleotide method using a number of subclones in single- or double-strand form obtained in pUC or M13 vectors. The

sequence protocol included use of standard and custom synthetic primers. Sequence data were submitted to EMBL Data Library - accession number X58120.

3. RESULTS AND DISCUSSION

The phasmid library on the λ -pSL5 vector has shown good expression properties, producing a quantity of protein, sufficient for immunoscreening in the phage form, although *EcoRI* site of λ -pSL5 is not situated in the region transcribed from the *Lac*-promoter. The strategy based on primary cloning of long inserts with consequent localization and subcloning of the full-size gene appeared to be productive.

Analysis of the DNA-derived amino-acid sequence shows its identity with the N-terminal fragment determined with Edman's method (Fig. 2). The sequence contains an open reading frame of 1156 codons with the AUG initiator codon and TAA terminator ochre-codon. The calculated M_r of the protein product is 129 740 kDa; experimental measurement of the M_r of CryIG has given essentially the same value [1].

Computer assay of 5'-non-coding region of the CryIG gene failed to reveal sequences similar to known bacillar promoters, which might imply that expression of the *cryI* gene proceeds via polycistronic mRNA. The Shine-Delgarno sequence GGAGGA was found 7 base pairs upstream from the initiator codon (Fig. 2). In the 3'-non-coding region 85 bp downstream from

terminator codon, a 17-bp inverted repeat was found. The hairpin structure that might be presumed for this repeat cannot be considered a good terminator because the loop part of it is too long, i.e. 13 base pairs.

CryIG differs markedly from all other members [3] of the CryI family. Its primary structure reveals only a marginal extent of identity when compared with other known members of the CryI family; it does not exceed 21% for the N-terminal half (1-670 bp) and 35% for the C-terminal half (671-1156 bp). The identical residues are not evenly distributed along the sequence forming pronounced stretches. In particular, all five conserved blocks of amino-acid residues, common for all CryI δ -endotoxins, have also been found in the CryIG sequence. Characteristically, the identity extent for these blocks is substantially higher, i.e. 60-80%.

A more detailed discussion of these data, especially of the alignment of sequences of CryI proteins will appear elsewhere.

REFERENCES

- [1] Chestukhina, G.G., Kostina, L.I., Zalunin, I.A., Khodova, O.M. and Stepanov, V.M. (1988) FEBS Lett. 232, 249-251.
- [2] Osterman, A.L. et al. (1989) Mol. Biol. (USSR) 23, 463-472.
- [3] Hofte, H. and Whiteley, H.R. (1989) Microbiol. Revs. 53, 242-255.
- [4] Tabor, S. and Richardson, C.C. (1987) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
- [5] Yankovsky, N.K., Fonstein, M.Yu. and Lashina, S.Y. (1989) Gene 81, 203-207.