Volume 11 Number 19 1983 Nucleic Acids Research

The complete nucleotide sequence of the gene coding for diphtheria toxin in the corynephage omega (tox+) genome

G.Ratti, R.Rappuoli and G.Giannini

Sclavo Research Centre, Via Fiorentina 1, 53100 Siena, Italy

Received 11 August 1983; Accepted 7 September 1983

SUMMARY

A segment of corynephage omega (tox^+) DNA, containing the gene for diphtheria toxin (tox) was fragmented with restriction enzymes and the fragments cloned into M13 vectors for nucleotide sequence determination. A long open reading frame was shown to encode the tox gene by comparing the predicted amino acid sequence with that of peptides derived from the mature toxin molecule. Analysis of the nucleotide sequence shows RNA polymerase and ribosome binding signals preceding a GTG codon in the open reading frame: if this is the correct starting signal for translation, then a 25 amino acid signal peptide can be predicted for the toxin molecule.

INTRODUCTION

Diphtheria toxin is a protein which can bind to most eukaryotic cell surfaces and upon entry to the cell blocks protein synthesis by adding an ADP-ribosyl group onto a modified histidine residue of the eukaryotic ribosomal elongation factor EF2. The molecule can be separated in vitro, by mild treatment with trypsin and a reducing agent, into two fragments, A and B. Fragment A retains the enzymatic activity, whereas fragment B comprises the region required for binding to and entry into eukaryotic cells. Fragment B also contains most of the antigenic determinants of the complete toxin molecule. Isolated A or B fragments are not toxic (for review see refs 1 and 2).

Diphtheria toxin is the product of a gene (tox) present in the genome of lysogenic corynephages which infect the gram-positive bacterium Corynebacterium diphtheriae: after lysogenisation the tox gene is expressed under bacterial control (3). Corynephages possess a single linear ds-DNA molecule, ranging in size from 35 to 38 Kb according to the strain (4-6). As with lambda phage, corynephage DNA can circularize through cos sites and integrate into two specific sites of the C. diphtheriae genome, attB1 and attB, (7) by recombination with a corresponding phage site, attP. Restriction maps for several phage strains have been determined and the cos site

and *att* site as well as the *tox* gene have also been physically mapped (4-6).

In this paper we report the complete primary structure of the *tox* gene from the corynephage omega (*tox*$^+$) and the first complete amino acid sequence of a wild-type diphtheria toxin as deduced from the nucleotide sequence of the gene.

METHODS

Cloning and sequencing strategy

Phage omega (tox$^+$) was prepared from *C. diphtheriae* C7 (omega tox$^+$) strain (6), grown in 2 l flasks or 10 l fermenters. Growth conditions and phage purification procedures were as previously described (7). Phage DNA was extracted with phenol and purified by CsCl isopycnic centrifugation as described (7).

For safety reasons, in order to avoid cloning the entire *tox* gene into M13 vectors, the following strategy was adopted. Omega phage DNA was digested with the restriction nuclease *Bam*HI and the fragment Bam-3 (7.8 Kb), known to contain the *tox* gene (6), was isolated by agarose gel electrophoresis followed by elution and purification of the corresponding band by standard procedures (8). The 7.8 Kb Bam-3 fragment is unique to corynephage omega, since the restriction maps of other corynephage strains are, in this portion of the genome, different. The *tox* gene was further localized (6) by a *Hin*dIII/*Eco*RI digestion of the 7.8 Kb Bam-3 fragment, and the resulting 2 Kb fragment dissected into A and B portions by Sau3A digestion (9,10). With this approach it was possible to obtain a set of restriction fragments safe for cloning under P1/EK1 conditions (9) and the set of clones used for sequencing is shown schematically in Fig. 1.

Sequencing was performed by the dideoxynucleotide chain termination technique in M13mp8 and mp9 vectors (11) in the host *E.coli* JM101 (12) as described by Sanger et al. (13). Long fragments were sequenced from both ends by recloning into an M13 vector by the "clone-turnaround" technique (14).

Searches for gene signals by weight matrix methods and analysis of gene properties were performed by using some of the options of the computer program ANALYSEQ (15).

RESULTS

The nucleotide sequences of the various fragments shown in Fig. 1 were

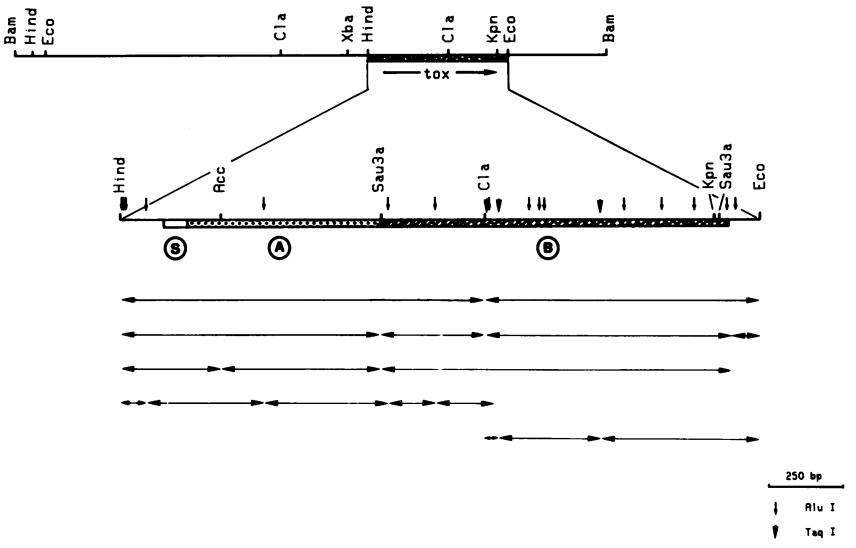

Fig.1 Restriction map of the 7.8 Kb Bam-3 fragment from phage omega DNA and of the Hind-Eco subfragment, described in the text, containing the tox gene. Underneath the map are indicated the coding regions for three diphtheria toxin domains: S, signal peptide; A, fragment A; B, fragment B. In the lower half of the figure are shown fragments which have been cloned in M13mp8/9 vectors for sequencing.

aligned by comparing overlapping fragments; the published data on the sequence of several diphtheria toxin peptides were also used for checking the correct alignment of the nucleotide sequences. The complete primary structure (1880 nucleotides) of the omega phage Eco-Hind fragment is presented in Fig. 2: only the "sense" strand is shown. Only one extensive open reading frame (ORF)is present in this fragment (from base 101 to base 1801).

A search for prokaryotic promoters and ribosomal binding sites by a weight matrix method, according to Hawley and McLure (16) and Stormo et al. (17,18), indicated the presence of transcription and translation signals close to a GTG start codon at position 122. The amino acid sequence coded for by the ORF starting from this position is also shown in Fig. 2 (560 amino acids, corresponding to 60,782 daltons). The broken line above the amino acid symbols indicate those portions that agree with published amino acid sequence data on diphtheria toxin fragments A and B (19,20). Single amino acid discrepancies within identified peptides are marked with asterisks. Fragment A is encoded from base 197 to base 775 and corresponds

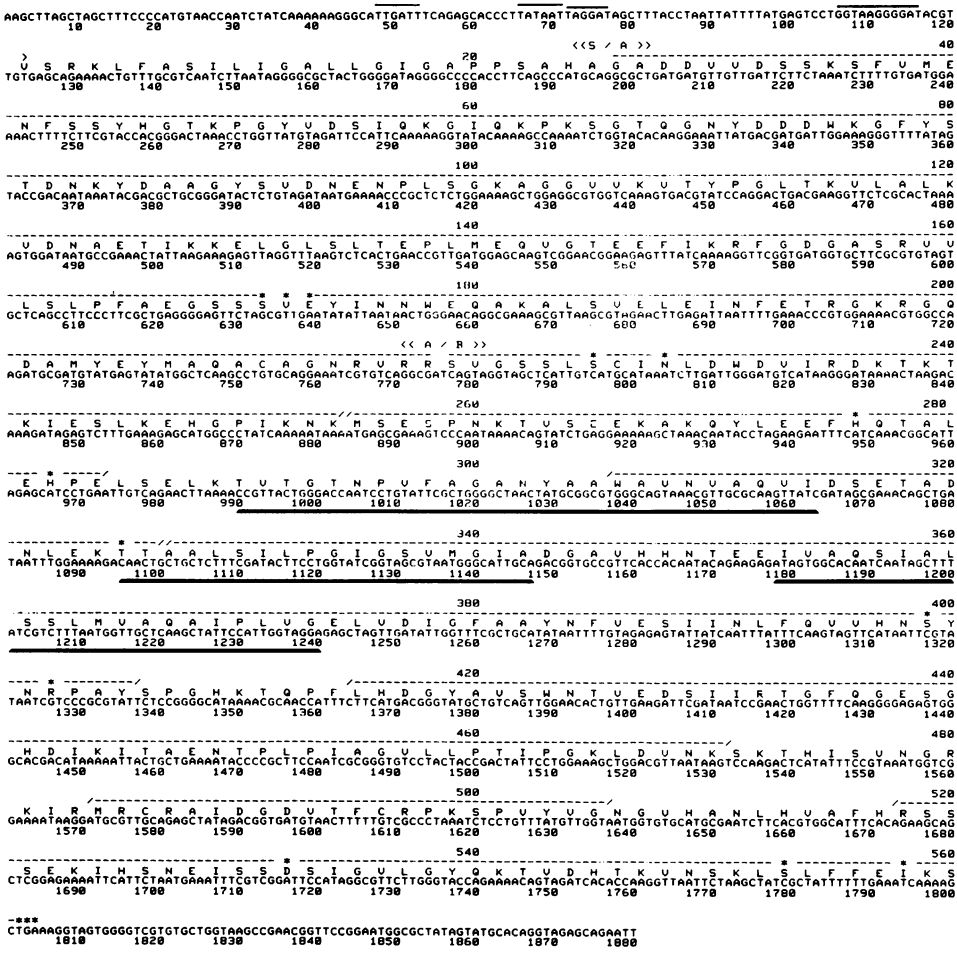

Fig.2 The nucleotide sequence of the Hind-Eco fragment from omega phage DNA, comprising the gene for diphtheria toxin. An open reading frame goes from position 101 to position 1799. Possible promoter and ribosomal binding sites preceding the GTG codon at position 122 are marked. The translation product starting from this position is shown in one-letter symbols above the nucleotide sequence. The sign > over the first residue indicates that this codon would be expected to be translated as f-Met rather that Val. The broken line above the amino acid sequence shows sequences in general agreement with published peptide sequencing data. Single amino acid differences are indicated by an *. Gaps in the line indicate lack of data or more extensive discrepancies. Three continuous lines below the nucleotide sequence indicate three regions coding for uninterrupted stretches of hydrophobic amino acid residues. The boundaries between the three major domains of the toxin (S, A and B) are indicated by letters above the amino acid sequence. The arrows above the top line indicate an inverted repeat in the -10 region.

to a polypeptide of 21,164 daltons. Fragment B is encoded from base 776 to base 1801 and corresponds to a polypeptide of 37,194 daltons. Without the 25 amino acid sequence (2,460 daltons) preceding the N-terminus of fragment A, the mature toxin molecule has 535 amino acid residues and a molecular weight of 58,340. Three portions of the gene code for long hydrophobic stretches (25, 18 and 21 residues) in fragment B and are marked in Fig. 2 by heavy continuous lines below the DNA sequence.

DISCUSSION

The amino acid sequence of diphtheria toxin as predicted from the nucleotide sequence of the *tox* gene in phage omega (Fig. 2) is in general agreement with the sequences of peptide fragments of the toxin (19,20). The single amino acid discrepancies (asterisks in Fig. 2) might be due to strain variation, although the precise source of toxin used is not clearly stated in the papers quoted above. The best starting signal for translation is probably the GTG codon at base 122. There is a typical ribosomal binding site, as found in gram- positive bacteria (21), five nucletides before the GTG. Further upstream we find sequences in good agreement with the -10 and -35 regions described for prokaryotic promoters (16). In Fig. 2 we have marked the sequence TTGATT as the -35 signal. Two possible -10 sequences are suggested because, although the first of the two (TATAAT) is actually identical to the accepted consensus signal, the distance between this sequence and the -35 sequence is not optimal. The second sequence (TAGGAT) would have the advantage of a correct spacing, while retaining four bases of the consensus and in particular the conserved T in the last position.

If the GTG codon at base 122 is the correct starting signal for the translation of the *tox* gene, then the product shows a stretch of 25 amino acids preceding the known N-terminus of the mature protein. This is most likely to represent the signal peptide for the secretion of diphtheria toxin (22) and indeed it displays all the features expected for such a function: two positively charged residues, Arg(3) and Lys(4), preceding a hydrophobic core, and the sequence Ala-His-Ala at its end as the expected cutting site for a signal peptidase (23). Diphtheria toxin appears therefore to be translated as a 60,782 dalton protein, and, after losing a 2,460 dalton N-terminal peptide, to be secreted as a 58,340 dalton mature protein.

An outstanding feature of the toxin sequence is the well known hydro-

phobic domain in the B fragment (1,2): three portions of the gene coding for stretches of 25, 18 and 21 hydrophobic amino acids, regularly spaced by ten amino acid residues are shown in Fig. 2. All the hydrophilic residues in the spacing sequences are negatively charged, with the exception only of Lys(324). This domain of diphtheria toxin appears therefore well suited for interacting with cell membranes and, in particular, the length of the hydrophobic amino acid stretches is sufficient to span the whole length of a cell membrane (24).

We find a possible eukaryotic ribosome binding site (25) preceding the codon for Met(255) shortly after the beginning of the B portion of the gene (sequence AAATAAAATG at position 877). This is probably purely coincidental, although it has been proposed (26) that the tox gene might actually have an eukaryotic origin. This stemmed from the consideration that, firstly, the tox gene is not essential either for corynephages or for the host bacterium and it maps in the phage genome very closely to the phage attP site, i.e. in the position where a transduced gene would be expected to be. Secondly, fragment B interacts with most eukaryotic cell membranes. The presence of a eukaryotic gene start signal before the hydrophobic domain would imply that the tox gene might be the result of the fusion of two or more genes and that the portion of diphtheria toxin that interacts with the eukaryotic cell membrane might have been derived from an ancestral eukaryotic gene.

ACKNOWLEDGEMENTS

The authors wish to thank Drs J.R. Murphy and G.F.Hatfull for useful discussions.

REFERENCES

1. Collier, R.J. (1982) In "ADP-ribosylation Reactions", Hayaishi, D. and Ueda, K., Eds, pp 575-592, Academic Press, New York
2. Pappenheimer, A.M.Jr (1977) Ann.Rev.Biochem. 46, 69-91
3. Murphy, J.R., Michel, J. and Teng, M. (1978) J.Bacteriol. 135 ,511-516
4. Costa, J., Michel J. and Murphy, J.R. (1981) J.Bacteriol. 148, 125-130
5. Michel, J.L., Rappuoli R., Murphy, J.R. and Pappenheimer, A.M.Jr (1982) J.Virol. 42, 510-518
6. Rappuoli, R., Michel, J.L. and Murphy, J.R. (1983) J.Virol 45, 524-530
7. Rappuoli R., Michel, J.L. and Murphy, J.R. (1983) J.Bacteriol. 153, 1202-1210
8. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning - A Laboratory Manual. Cold Spring Harbor Laboratory. Cold Spring Harbor, New York
9. Recombinant DNA Tech. Bull. (1982) Vol.5, No 1, p.33 and Vol.5, No 3, p.153

10. Leong, D., Coleman K.D. and Murphy, J.R. (1983) Science 220, 515-517
11. Messing, J. and Vieira, J. (1982) Gene 19, 269-276
12. Messing J. (1979) Recombinant DNA Tech. Bull. 2, 43-48
13. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Petersen, G.B. (1982) J.Mol.Biol. 162, 729-773
14. Winter, G. and Fields, S. and Ratti, G.(1981) Nucleic Acids Res. 9, 6907-6915
15. Staden, R. et al. (1983) Nucleic Acids Res., in the press
16. Hawley, D.K. and McLure, R. (1983) Nucleic Acids Res. 11, 2237-2255
17. Stormo, G.D., Schneider, T.D. and Gold, L.M.(1982) Nucleic Acids Res. 10,2971-2996
18. Stormo, G.D., Schneider, T.D., Gold, L. and Ehrenfeucht, A. (1982) Nucleic Acids Res. 10, 2997-3011
19. Delange, R.J., Williams, L.C., Drazin, R.E. and Collier R.J. (1979) J.Biol.Chem. 254, 5838-5842
20. Falmagne, P., Capiau, C., Zanen, J., Kayser, G. and Ruysschaert, J-M. (1982) Toxicon 1, 243-246
21. Loefdahl, S., Guss, B., Uhlen, M., Philipson, L. and Lindberg, M. (1983) 80, 697-701
22. Smith, W.P., Tai, P.C., Murphy, J.R. and Davis, B.D. (1980) J.Bacteriol. 141,184-189
23. Perlman, D. and Halvorson, H.O. (1983) J.Mol.Biol. 167, 391-409
24. Segrest, J.P. and Jackson, R.L. (1977) In "Membrane Proteins and their Interactions with Lipids". Capaldi R.A. Ed., pp.21-45. Marcel Dekker, New York
25. Sargan, D.R., Gregory, S.P., Butterworth, P.H.W. (1982) FEBS Lett. 147,133-136
26. Pappenheimer, A.M.Jr (1981) In "The Harvey Lectures, 1980-1981", pp.45-73. Academic Press, New York