Nucleotide sequence of *Bacillus* phage Nf terminal protein gene

---

Mark C. Leavitt and Junetsu Ito

Department of Microbiology and Immunology, College of Medicine, University of Arizona,
Tucson, AZ 85724, USA

Received May 18, 1987; Accepted June 2, 1987

Accession no. Y00363

ABSTRACT

The nucleotide sequence of *Bacillus* phage Nf gene E has been determined. Gene E codes for phage terminal protein which is the primer necessary for the initiation of DNA replication. The deduced amino acid sequence of Nf terminal protein is approximately 66% homologous with the terminal proteins of *Bacillus* phages PZA and Ø29, and shows similar hydropathy and secondary structure predictions. A serine which has been identified as the residue which covalently links the protein to the 5' end of the genome in Ø29, is conserved in all three phages. The hydropathic and secondary structural environment of this serine is similar in these phage terminal proteins and also similar to the linking serine of adenovirus terminal protein.

INTRODUCTION

Bacteriophage Nf belongs to a collection of small lytic phages which infect *Bacillus*. Closely related phages include Ø29, PZA, Ø15, GA-1 and M2 (which differs from Nf by a small internal deletion). All these phages contain small, linear duplex DNA genomes of approximately 18-20 kilobase pairs with viral encoded terminal proteins covalently linked to the 5' termini of both DNA strands (1,2). One of the distinguishing features of these phages is that the terminal protein is used as a primer for the initiation of DNA replication.

All known DNA polymerases require a primer which supplies a 3'-hydroxyl group in order to initiate DNA synthesis (3). Usually the primer is a small RNA molecule which is later removed and the resulting gap repaired by DNA polymerase. This type of initiation creates a problem in the replication of linear DNA molecules. Removal of the RNA primer from the 5' end of the genome results in an irreparable gap since all the known DNA polymerases synthesize only in a 5' to 3' direction (3). To circumvent this problem, a model was first proposed in adenovirus (which has a similar terminal protein) where the terminal protein primes the polymerization by providing an hydroxyl group, namely the hydroxyl side group of an internal amino acid (4,5). In the case of adenovirus and Ø29, the beta-hydroxyl group of a serine residue serves this

function (6,7). The initiation reaction results in the formation of a phosphodiester bond linkage between this internal residue and the first nucleotide of the newly synthesized DNA strand. In vivo and in vitro studies of Ø29 and M2 have shown that the initiation reaction requires; DNA-terminal protein template, viral DNA polymerase, Mg^{++} , dATP, and free terminal protein (8,9,10). Elongation of the DNA strand from this terminal protein-nucleotide initiation complex proceeds by parental strand displacement (11,12).

Although protein-primed DNA replication was initially described in adenovirus and Ø29, other examples of linear DNA molecules with terminal proteins have been found in phage PRD1, which infects gram-negative bacteria (13), certain Streptococcus pneumoniae phages (14,15), Streptomyces plasmid pSLA2 (16), and Maize mitochondrial DNA S1 and S2 (17). These and other systems have identified at least 2 other amino acids (threonine and tyrosine) which can also form covalent linkage during the initiation of DNA replication.

In order to more fully understand the mechanism of protein-primed DNA synthesis, it is necessary to learn more about the terminal proteins themselves. Thus far the terminal protein genes of Ø29 (18,19), PZA (20) and adenovirus (21) have been sequenced. However, Ø29 and PZA both belong to group A of the small Bacillus phages (based on similar serological properties, physical maps, peptide maps, DNA sequence of the terminal portion of the genome and genome size) and as a consequence the amino acid sequence of their terminal proteins are nearly identical. The sequence of a more distantly related group B phage such as Nf would provide better insight to the common structural features of terminal proteins and to the evolutionary relationship of the Bacillus phages.

In this paper we present the nucleotide sequence of Nf gene E and the deduced amino acid sequence of its product, terminal protein. We also compare these sequences with those of related phages PZA and Ø29 to determine the degree of conservation of the terminal proteins. The structural characteristics of the Bacillus phage terminal proteins are also compared to reveal consensus structures and possible DNA linking sites for both Nf and PZA terminal proteins.

MATERIALS AND METHODS

DNA Preparation

Bacteriophage Nf was grown in Bacillus amyloliquefaciens H strain and purified phage particles were isolated by cesium chloride centrifugation (22). Deproteinized DNA was prepared as described (23).

Sequencing

DNA sequencing was done by the method of Maxam and Gilbert (24) as modified by Smith and Calvo (25). Restriction enzymes were obtained from New England Biolabs, International Biotechnologies Inc., and Bethesda Research Laboratories. Calf intestinal phosphatase was obtained from Boehringer Mannheim and IBI. [Gamma-³²P]ATP was obtained from ICN and New England Nuclear.

Computer Analysis

Secondary structure was determined by a program based on the method of Chou and Fasman (26). Hydrophobicity data was determined by a program based on the method of Kyte and Doolittle (27) in which the hydrophobic indices are averaged over a span of seven residues. Both analyses were done using micro-computer programs of Mount and Conrad (28).

RESULTS

Sequence Comparisons

Nucleotide sequence determination of Nf terminal protein coding region reveals an open reading frame of 801 nucleotides which initiates with GTG (Fig. 1). This would code for a protein of about 31.2 kDa which is in very good agreement with previous estimates of Nf terminal protein size (30 kDa) from gel migration analysis (29). The coding region of the transcriptional activator, gene F (30), overlaps the first 112 nucleotides of the terminal protein, gene E. The terminal protein gene and the transcriptional activator gene of PZA and Ø29 have an identical coding frame overlap (18,19,20). Thus, the terminal protein genes of Nf, PZA and Ø29 are all identical in size, 801 nucleotides coding for a protein of 266 amino acid residues. These results suggest that neither an insertion nor frameshift mutation have occurred within these genes since divergence from their common progenitor.

Both PZA and Ø29 belong to group A of the small *Bacillus* phages and consequently their terminal proteins are highly homologous, 91.7% at the nucleotide level and 97.7% at the amino acid level. Nf, on the other hand, is a member of group B and contains a smaller genome than group A phages (1). The nucleotide sequence of Nf gene E is 65.8% and 66.0% homologous to PZA and Ø29 respectively. The amino acid sequence of Nf terminal protein is 62.8% homologous to PZA and 62.4% homologous to Ø29.

In Nf and PZA terminal proteins, neither the location nor the amino acid species is known for the residue which is covalently linked to the 5' ends of the genome. However, in Ø29 the linking residue has been identified as a serine at position 232 (31). Both Nf and PZA translations show a serine at

```

1
Nf  HARNSRIRITWNDKALYAKLVKNTKAKISRTKKKYGIDLSEIELPPLESFQTRFEFNKVKQKQESFTNRANQNYQFVKKQYGVASKAKINEIAKNTKE 100
PZA  SP  KD  E  R      A      V  A  NI  D  E  AQ  EQAS  HR  E  A  V      A  ER
Ø29  SP  KD  E  R      A      V  TA  DI  D  D  E  AQ  EQAS  HR  E  A  V      A  ER

101
AQRIVDEQRREEISDKPFISGGKQOGTVGORMOILSPSQVTGISRPDSFNFDDVRSYARLRTLEEGMAEKASPDYDRRNTQNHQMFIEIVEKSFNSDWLS 200
V  L  KINANK  EYVA  P  IE  IANT  AH  N  H  D  SK  T  S      S  ENRTD  Q  EKK  I  LQL  KS  G  FDAA
V  L  KINANK  EYVA  P  IE  IANT  AH  N  H  D  SK  T  S      S  ENRTD  Q  EKK  I  LQL  KS  G  FDAA

201
DELVERLKKIPDDOFFELVNFDEISFEYFDSEGEVKAESAMLNKINSYLDYERGDVMDLKGFF 266
I  E      Y  F  RIS  E      NT  NV  GVVY  L  EQ  R  FD  S
I  E      Y  F  RIS  E      NT  NV  GVVY  L  EQ  R  FD  S

```

Figure 2. Amino acid sequence of terminal proteins. Amino acid sequence of Nf is shown on the top line. Empty spaces indicate residues identical with Nf. The position of the linking serine in Ø29 is indicated by an arrow.

position 232 which lies in the center of a conserved sequence 5 amino acid residues in length (Fig. 2).

Hydropathy

Hydropathic profiles of Nf, PZA, and Ø29 terminal proteins show very similar regions of hydrophobicity and hydrophilicity (Fig. 3). As noted earlier by Hermoso and Salas, serine 232 lies in a hydrophilic region of Ø29 terminal protein (31). This is also the case for serine 232 of Nf and PZA terminal proteins. Interestingly, the linking serine of adenovirus terminal protein is also found in a hydrophilic region (31,21).

Figure 3. Hydropathic profiles of Nf, PZA and Ø29 terminal proteins. The position of the linking serine of Ø29 is indicated by an arrow.

Figure 4. Secondary structure predictions of Nf, PZA and Ø29 terminal proteins based on the method of Chou and Fasman (26). Helical regions are represented by , beta sheet regions by --, and beta turns by / \. The position of the linking serine of Ø29 is indicated by an arrow.

Secondary Structure

The secondary structure of Nf terminal protein as predicted by Chou and Fasman analysis is very much like that of PZA and Ø29 (Fig. 4). All are primarily alpha helical: Nf 61%; PZA and Ø29 60%. The linking serine residues of Ø29 and adenovirus terminal protein are located in a beta-turn (31). Similarly, serine 232 of Nf and PZA terminal proteins is located in a beta-turn.

Transcription Termination Signals

Davidson et al. (32) described termination sites of *in vitro* Ø29 transcription. One of the terminators of early transcription, TE2, has been tentatively identified by its characteristic DNA sequence for a transcription terminator (18). These sequences are typically composed of a T cluster preceded by a region of dyad symmetry with high G+C content (33). TE2 of Ø29 is located near the translational start signal of gene 4 (the transcriptional activator). Analysis of Nf genome in this region also reveals a putative terminator with similar sequence characteristics (Fig. 5). There are 11 differences in the 2 stem-loop structures, 8 of which occur in loop regions.

Figure 5. Sequence of transcriptional termination sites of Nf and Ø29 (TE2). Numbers indicate position of nucleotide from figure 1. Initiation codons of transcriptional activator genes are lined.

DISCUSSION

We have determined the nucleotide sequence of gene E of *Bacillus* phage Nf which codes for the DNA terminal protein. Nf gene E is approximately 65% homologous with the terminal protein genes of PZA and Ø29. On the other hand, Nf gene F (transcriptional activator) is approximately 78% homologous with the transcriptional activator genes of PZA and Ø29. The 112 base pair sequence where genes E and F overlap shows an 82% and 85% homology to the same region in PZA and Ø29. Since the homology of the overlap region is considerably greater than the overall homologies of either gene E or gene F, the overlap region is probably more evolutionarily conserved due to the constraints of 2 reading frames. A mutation in this region of the terminal protein gene, albeit silent, may have a deleterious effect on the transcriptional activator.

Comparison of the nucleotide changes between Nf terminal protein gene and those of Ø29 and PZA genes showed that most changes result in amino acid substitution (table 1). This is sharply contrasted with nucleotide changes observed between Ø29 and PZA terminal protein genes in which most changes are silent. Transversions account for 54% of the base changes between Nf-Ø29 and for 53% in Nf-PZA, whereas only 35% of the base changes are transversions in Ø29-PZA.

Translation of gene E initiates with the codon GUG instead of the typical initiation codon AUG. GUG and UUG, which usually code for valine and leucine, are frequently used as initiation codons in *Bacillus subtilis* and other gram-positive organisms. *E.coli* on the other hand initiates almost exclusively with AUG (34). It has been suggested that the reduced stability of base pairing between anticodon of the initiator tRNA-fMet and GUG may be compensated for by stronger ribosomal-mRNA interactions. Since *Bacillus* in general requires a stronger Shine-Dalgarno complementarity to the 3' end of 16S rRNA than *E.coli* (35), this increased stability at the initiation site may permit *Bacillus* to have higher variability in initiation codons. The initiation site of Nf terminal protein may support this contention since the Shine-Dalgarno sequence

Table 1. Comparison of the sequence differences among the *Bacillus* phage terminal proteins and terminal protein genes.

CHANGES	Nf-Ø29	Nf-PZA	Ø29-PZA
Nucleotide differences	272(34.0%)	274(34.2%)	66(8.3%)
Amino acid differences	100(37.6%)	99(37.2%)	6(2.3%)
Silent nucleotide changes	78(28.7%)	83(30.2%)	58(87.9%)

preceding the GUG initiation codon is highly complementary to the 3' end of 16S rRNA (Fig. 1). The strength of the complementarity of the Shine-Dalgarno sequence, measured as the free energy of binding (36) is -19.3 kcal/mol in Nf as compared to -11.8 and -16.2 in PZA and Ø29 respectively. The free energy of binding of Nf terminal protein Shine-Dalgarno sequence is stronger than any previously identified in the small *Bacillus* phages. To our knowledge this is the first gene identified from a *Bacillus subtilis* phage which initiates with GUG.

Perhaps the most intriguing feature of the terminal proteins is the linking site because of its direct involvement in DNA replication. The linking sites of Ø29 and adenovirus terminal proteins share certain structural characteristics although no apparent sequence homology exists. In both cases the linking residue is a serine located near the carboxy terminus in a beta-turn region preceded by an alpha-helix. Hydropathy profiles of this area show it to be hydrophilic and therefore probably near the protein's surface (27). The linking residues of Nf and PZA terminal proteins have yet to be identified. Our results suggest that these terminal proteins have a serine in position 232 which is identical with the position of the linking serine of Ø29. Furthermore we show that serine 232 of Nf and PZA terminal proteins is located in a beta-turn preceded by an alpha-helix, and that this residue lies in a hydrophilic region near the carboxy terminus. Since the phage terminal proteins are identical in size, and have very similar secondary structure and hydropathy predictions, it is most likely that serine 232 is also the linking residue in Nf and PZA terminal proteins.

ACKNOWLEDGEMENTS

We are grateful to Dr. D. Mount for the use of his DNA sequence analysis program and to Dr. M. S. Saedi for critical reading of this manuscript.

This research was supported in part by a National Institutes of Health Grant, GM 28013.

REFERENCES

1. Yoshikawa, H., Elder, J.H., and Ito, J. (1986) *J. Gen. Appl. Microbiol.* 32, 39-49.
2. Mizukami, Y., Sekiya, T. and Hirokawa, H. (1986) *FEBS Lett.* 197, 311-314.
3. Kornberg, A. (1974) *DNA Synthesis*, W.H. Freeman and Co., San Francisco.
4. Rekosh, D.M.K., Russell, W.C., Bellett, A.J.D. and Robinson, A.J. (1977) *Cell* 11, 283-295.
5. Friefeld, B.R., Lichy, J.H., Field, J., Gronostajski, R.M., Guggenheimer, R.A., Krevolin, M.D., Nagata, K., Hurwitz, J. and Horwitz, M.S. (1986) *Curr. Top. Microbiol. Immunology.* 110, 221-225.

6. Hermoso, J.M. and Salas, M. (1980) *Proc. Natl. Acad. Sci. USA* 77, 6425-6428.
7. Desiderio, S.V. and Kelly, T.J. (1981) *J. Mol. Biol.* 145, 319-337.
8. Shih, M-F., Watabe, K. and Ito, J. (1982) *Biochem. Biophys. Res. Commun.* 105, 1031-1036.
9. Matsumoto, K., Saito, T. and Hirokawa, H. (1983) *Mol. Gen. Genet.* 191, 26-30.
10. Watabe, K., Leusch, M. and Ito, J. (1984) *Proc. Natl. Acad. Sci. USA* 81, 5374-5378.
11. Harding, N.E. and Ito, J. (1980) *Virology* 104, 323-338.
12. Inciarte, M.R., Salas, M., and Sogo, S.M. (1980) *J. Virol.* 34, 187-199.
13. Bamford, D.H., McGraw, T., MacKenzie, G. and Mindich, L. (1983) *J. Virol.* 47, 311-316.
14. Garcia, E., Gomez, A., Ronda, C., Escarmis, C. and Lopez, R. (1983) *Virology* 128, 92-104.
15. Lopez, R., Ronda, C., Garcia, P., Escarmis, C. and Garcia, E. (1984) *Mol. Gen. Genet.* 197, 67-74.
16. Hirochika, H., Natamura, K., Sakaguchi, K. (1984) *EMBO J.* 3, 761-766.
17. Kremble, R.J. and Thompson, D. (1982) *Nucl. Acids Res.* 10, 8181-8190.
18. Yoshikawa, H. and Ito, J. (1982) *Gene* 17, 323-335.
19. Escarmis, C. and Salas, M. (1982) *Nucl. Acids Res.* 10, 5785-5798.
20. Paces, V., Vleck, C., Urbanek, P. and Hostomsky, Z. (1985) *Gene* 38, 45-56.
21. Smart, J.E. and Stillman, B.W. (1982) *J. Biol. Chem.* 257, 13499-13506.
22. Ito, J., Kawamura, F. and Yanofsky, S. (1976) *Virology* 70, 37-51.
23. Ito, J. (1978) *J. Virol.* 28, 895-904.
24. Maxam, A.M. and Gilbert, W. (1980) *Methods Enzymol.* 65, 499-650.
25. Smith, D.R. and Calvo, J.M. (1980) *Nucl. Acids Res.* 8, 2255-2274.
26. Chou, P.Y., and Fasman, G.D. (1978) In Meister, E. (ed) *Advances in Enzymol.* Vol. 47, pp.45-148. Wiley, New York.
27. Kyte, J. and Doolittle, R.F. (1982) *J. Mol. Biol.* 157, 105-132.
28. Mount, D.W. and Conrad, B. (1984) *Nucl. Acids Res.* 12, 811-817.
29. Yoshikawa, H. and Ito, J. (1981) *Proc. Natl. Acad. Sci. USA* 78, 2596-2600.
30. Mizukami, Y., Sekiya, T. and Hirokawa, H. (1986) *Gene* 42, 231-235.
31. Hermoso, J.M., Mendez, E., Soriana, F. and Salas, M. (1985) *Nucl. Acids Res.* 13, 7715-7728.
32. Davidson, B.L., Murray, C.L. and Rabinowitz, J.C. (1980) *J. Biol. Chem.* 255, 8819-8830.
33. Rosenberg, M. and Court, D. (1979) *Ann. Rev. Genet.* 13, 319-353.
34. Hagar, P. and Rabinowitz, J.C. (1985) *J. Biol. Chem.* 260, 15163-15167.
35. McLaughlin, J.R., Murray, C.L. and Rabinowitz, J.C. (1981) *J. Biol. Chem.* 256, 11283-11291.
36. Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M., Gralla, J. (1973) *Nature (New Biology)* 246, 40-41.