Primary structure of the DNA terminal protein of bacteriophage PRD1

Jui-Cheng Hsieh, Guhung Jung, Mark C. Leavitt and Junetsu Ito

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

Received August 4, 1987; Accepted September 22, 1987

ABSTRACT

The genome of a lipid-containing phage, PRD1, is replicated by a protein-priming mechanism. We have determined the nucleotide sequence of the PRD1 gene 8 which specifies the terminal protein, the protein primer for DNA synthesis. The coding region is 780 base pairs long and encodes for 259 amino acids (29,326 daltons). The predicted amino acid sequence of the PRD1 terminal protein reveals no substantial homology with that of any known terminal protein. However, hydropathy profiles of the PRD1, Ø29, and Nf terminal proteins are remarkably similar, suggesting a common evolutionary origin. A particular tyrosine residue is predicted to be covalently linked to the 5' end of the PRD1 DNA. The initiation codon ATG of gene 8 is preceded by the identifiable ribosome binding site, and putative promoter sequences. There are unique palindromic sequences between the ribosome binding site and "-10" region.

INTRODUCTION

Bacteriophage PRD1 is the prototype of a group of small icosahedral, lipid-containing bacterial viruses (1,2). This phage infects a wide variety of gram-negative bacteria harboring plasmids of the P, N or W incompatibility type (1,3). The PRD1 genome is a linear, double-stranded DNA molecule of about 14.7 kilobases with a 28,000 dalton terminal protein covalently linked at both 5' ends (4,5). The PRD1 DNA contains a long inverted terminal repeat (111 bp) which does not bear substantial homology to that of any known viral DNAs with terminal protein (6).

The DNA terminal protein is an intriguing protein. It provides a primer for the initiation of linear DNA replication. Because all known DNA polymerases act only in the 5' to 3' direction and require primer, linear non-redundant DNA poses the special problem of preserving the 5' end of the DNA if RNA is used as a primer and is removed (7). Protein-priming provides a mechanism by which the 5' end of linear DNA is preserved. Furthermore, since terminal protein is required stoichiometrically for the initiation of DNA replication, the regulation of the terminal protein synthesis directly affects the rate of linear DNA synthesis. In addition, the terminal protein

may enhance the fidelity of DNA replication at the 5' termini of the linear DNA. Since chemical mismatches between the bases can occur with a frequency of 10^{-5} , which is substantially higher than the frequency of spontaneous mutation, it can be inferred that the accuracy of DNA synthesis is maintained (8). Finally, it has been well documented that the terminal protein is essential for DNA encapsidation in the $\phi 29$ system (9).

Recently it has been shown that the DNA polymerases which use protein primers have partial, but significant amino acid sequence homologies with DNA polymerases which use RNA primers (10,11). These findings suggest strongly that both types of DNA polymerases have evolved from a common ancestor. In view of these findings, it becomes of great interest to investigate evolutionary origins of the terminal proteins. Where do terminal proteins come from? What is the role of the terminal protein in the fidelity of DNA replication?

Because PRD1 is the smallest bacteriophage infecting gram-negative bacteria that contains linear duplex DNA, this system is especially amenable to biochemical genetic studies of linear DNA replication and of the DNA terminal protein. Recent studies from this laboratory indicate that PRD1 DNA replication starts at either end of the DNA molecule and proceeds by strand-displacement synthesis (Yoo and Ito, in preparation). This type of DNA replication resembles the systems described for adenovirus (12,13,14) and Bacillus phage $\phi 29$ (15,16). In this paper, we present the nucleotide sequence of the PRD1 terminal protein gene and the responding amino acid sequence. We also compared the amino acid sequence of the PRD1 terminal protein with that of other DNA terminal proteins. Further, we predict the location of the tyrosine residue that is linked to the DNA.

MATERIALS AND METHODS

Bacterial Strain and Plasmid

Phage PRD1 was kindly provided by R. Olsen and L. Mindich, independently. PRD1 DNA was prepared from purified phage grown in Salmonella typhimurium LT2 carrying a plasmid, pLM2 (17). E. coli HB94 (pLM2,pLM3) which contains PRD1 genes 1 and 8 was kindly provided by L. Mindich (17). Plasmid DNA was prepared by the method of Clewell (18).

DNA Sequencing

The pLM3 plasmid was first cleaved with Pst I and the resulting DNA fragment containing genes 1 and 8 of PRD1 was subcloned into phage M13mp9. A nested set of deletions was generated by the method of Dale et al., (19). Each set of deletions was sequenced by the dideoxy chain termination method of

Figure 1. Genetic and physical maps of region of PRD1 which contains the genes for terminal protein (gene 8) and for DNA polymerase (gene 1). The genetic map was adapted from that of McGraw et al. (17). Arrows indicate the direction of translation. Open circles at both 5' ends of the genome indicate terminal protein.

Sanger et al., (20). For confirmation, most regions were also sequenced by the method of Maxam and Gilbert (21).

Computer Analysis

The hydropathic profile was determined by a program based on the method of Kyte and Doolittle (22). Secondary structure was determined by a program based on the method of Chou and Fasman (23). Both analyses were done using microcomputer programs of Mount and Conrad (24).

RESULTS AND DISCUSSION

Nucleotide Sequence

The PRD1 genes coding for the DNA terminal protein (gene 8) and DNA polymerase (gene 1) are mapped at the left end of the PRD1 genome, as shown in Figure 1 (17). The direction of early transcription has recently been determined to be left to right (Gerendasy and Ito, unpublished). The nucleotide sequence of the left end of the PRD1 DNA (1,046 bp) and the deduced amino acid sequence is diagrammed in Figure 2. The first 150 bp of the DNA sequence which contains the 111 bp terminal inverted repeat previously described (6) and is included here for clarity.

Figure 2. The nucleotide sequence and predicted amino acid sequence of the PRD1 terminal protein gene. Initiation and termination codons for the PRD1 terminal protein and initiation codon for the PRD1 DNA polymerase are boxed. Putative promoter sequences "-35 region" and "-10 region" are also boxed. Shine-Dalgarno sequences are dotted. The vertical arrow points is the end nucleotide of the inverted terminal repeat sequences. The horizontal arrows indicate unique repeat sequences. Numbers indicate nucleotide positions from the left end of PRD1 genome.

In agreement with genetic studies by Mindich and coworkers (5,17), nucleotide sequence data reveal that there is only one open reading frame that is of sufficient length to encode the entire PRD1 terminal protein (Figure 2). This open reading frame is 780 nucleotides long and ends in an ochre termina-

Figure 3. Possible secondary structures at the 5' flanking nucleotide sequence of the PRD1 terminal protein gene.

tion codon TAA. A protein encoded by this open reading frame is 259 amino acids in length and has a calculated molecular weight of 29,326 daltons, a value which is in good agreement with molecular weight estimates from sodium dodecyl sulphate-polyacrylamide gel electrophoresis (3). The initiation codon of the PRD1 terminal protein is preceded by an identifiable Shine-Dalgarno (25) ribosome binding site. A presumptive RNA polymerase recognition sequence "-35 region" can be found at nucleotides 145-150 (TGTGTA) and a "-10 region" at nucleotides 169-174 (TAAAAT). Because PRD1 infects a wide variety of gram-negative bacteria, it was of interest to find out what kind of promoter sequences the phage has. Although these putative promoter sequences must be confirmed by S1 mapping experiments, these nucleotide sequences rather resemble consensus *E. coli* promoter sequences (26).

Between the "-10 region" and initiation codon, there are two palindromic sequences that can form various secondary structures (Figure 3). The biological significance of these sequences is not known, but they may represent transcriptional or translational regulatory signals. Since there is

Table 1. Predicted amino acid composition of terminal protein

Amino acid	Comparison	
	Numbers in PRD1	Numbers in Ø29
Basic side chains		
Lys	36	25
Arg	18 (21.2%)	18 (16.9%)
His	1	2
Acidic side chains		
Asp	7	19
Glu	17 (9.2%)	27 (17.3%)
Uncharged polar side chains		
Gly	20	11
Asn	11	17
Gln	15	14
Cys	0 (31.6%)	0 (31.2%)
Ser	9	20
Thr	14	10
Tyr	13	11
Nonpolar side chains		
Ala	20	14
Val	20	12
Leu	15	17
Ile	15	18
Pro	12 (37.8%)	8 (34.5%)
Phe	11	15
Met	4	6
Trp	1	2
Total	259	266

neither an obvious transcriptional termination site at the end of the terminal protein gene, nor promoter sequences for the DNA polymerase gene, it seems likely that a polycistronic mRNA specifies both PRD1 terminal protein and DNA polymerase. It will be interesting to elucidate regulatory mechanism by which the synthesis of these two proteins is controlled.

The unusual features of the nucleotide sequence of the PRD1 terminal protein gene are the consecutive runs of the single base adenine. There are 7A, 6A, 5A, 4A, and 3A sequences. These poly A sequences run 1, 3, 6 and 20 times, respectively. These sequences are attributed to 27 lysine residues in the PRD1 terminal protein (see below).

Amino Acid Sequence of the PRD1 Terminal Protein

The distribution of the amino acid residues over the polypeptide chain of the PRD1 terminal protein exhibits some interesting features (Figure 2). The terminal protein has two strongly basic regions: at the amino-terminus, there are 17 positively charged residues (Arg, Lys and His) in the first 53 residues (about 32%) and at the carboxyl-terminus, there are 9 positively charged amino acids in the last 17 residues (about 53%). The rest of the polypeptide chain contains about 13% basic amino acid residues.

No substantial amino acid sequence homology was found when the deduced amino acid sequence of the PRD1 terminal protein was compared with those of terminal proteins of Ø29, Nf and adenovirus (27,28,39,30,31,32). Comparison of the amino acid composition of the terminal proteins of PRD1 and Ø29 are shown

Table 2. Codon usage in terminal protein genes

PRD1		Ø29	PRD1		Ø29	PRD1		Ø29	PRD1		Ø29
TTT/Phe	9	9	TCT/Ser	0	1	TAT/Tyr	11	9	TGT/Cys	0	0
TTC/Phe	2	6	TCC/Ser	2	2	TAC/Tyr	2	4	TGC/Cys	0	0
TTA/Leu	3	5	TCA/Ser	2	4	TAA/TER	1	0	TGA/TER	0	0
TTG/Leu	5	4	TCG/Ser	1	0	TAG/TER	0	1	TGG/Trp	1	0
CTT/Leu	5	2	CCT/Pro	4	5	CAT/His	0	1	CGT/Arg	5	5
CTC/Leu	0	1	CCC/Pro	4	1	CAC/His	1	1	CGC/Arg	10	1
CTA/Leu	1	2	CCA/Pro	2	1	CAG/Gln	12	3	CGA/Arg	1	4
CTG/Leu	1	2	CCG/Pro	2	1	CAG/Gln	3	7	CGG/Arg	2	3
ATT/Ile	8	7	ACT/Thr	3	1	AAT/Asn	4	8	AGT/Ser	0	5
ATC/Ile	6	2	ACC/Thr	4	2	AAC/Asn	7	5	AGC/Ser	4	6
ATA/Ile	1	9	ACA/Thr	2	8	AAA/Lys	22	11	AGA/Arg	0	5
ATG/MET	4	7	ACG/Thr	5	1	AAG/Lys	14	16	AGG/Arg	0	2
GTT/Val	8	3	GCT/Ala	3	8	GAT/Asp	4	10	GGT/Gly	1	6
GTC/Val	1	1	GCC/Ala	7	2	GAC/Asp	3	8	GGC/Gly	12	0
GTA/Val	7	4	GCA/Ala	3	4	GAA/Glu	16	19	GGA/Gly	3	3
GTG/Val	4	4	GCG/Ala	7	5	GAG/Glu	1	8	GGG/Gly	4	2

Numbers indicate the frequency of codons used in the terminal protein genes.

in Table 1. The Ø29 terminal protein contains 7 additional amino acid residues. The most significant difference is the presence of 22 additional acidic amino acid residues (Asp, Glu) in the Ø29 terminal protein. On the other hand, the PRD1 terminal protein contains 10 additional basic amino acid residues. Clearly, the PRD1 terminal protein has a greater net positive charge. It is remarkable that more than 60% of the aspartic acid and 46% of the glutamic acid residues are present in the amide form in the PRD1 terminal protein. This is in contrast to the Ø29 terminal protein in which only 34% of the glutamic acid residues are present in the amide form. Neither protein contains cysteine residues.

Codon Usage

Codon frequencies in the terminal protein gene of the PRD1 and Ø29 are summarized in Table 2. Comparison of codon usage reveals some striking differences between the terminal proteins. Most notable is the use of the glycine codon GGC by PRD1. Of the 20 glycine codons present in the PRD1 terminal protein, GGC is used 12 times (60%). On the other hand, Ø29 terminal protein does not use GGC at all, even though it contains 11 glycine residues (Table 2). This can be explained by the fact that the Ø29 DNA does not contain the GGCC sequence perhaps by the evolutionary avoidance of the restriction site (GGCC) of Bsu I endonucleases which are produced by its host, *Bacillus subtilis* (33).

Figure 4. Hydropathic profiles of C-terminal regions of the PRD1, Ø29, and Nf terminal proteins. Data on Ø29 and Nf terminal proteins were obtained from reference (31). PRD1 terminal protein (———), Ø29 (— · — · —), and Nf (-----).

In addition, the PRD1 terminal protein gene shows a strong bias toward the arginine codon, CGC and the glutamine codon, CAA.

Hydropathy

It is known that the linkage of the PRD1 terminal protein to the DNA is a phosphodiester bond between tyrosine and a dGMP of the DNA (4). The nucleotide sequence data predict the existence of 13 tyrosine residues in the PRD1 terminal protein (Figure 2). To obtain information on which tyrosine residue would likely be the linking amino acid to the DNA, we have determined the hydropathic profile of the PRD1 terminal protein following the algorithm of Kyte and Doolittle (22). When the hydropathic profiles of PRD1, Ø29, and Nf terminal proteins were compared, there seemed to be no common similarity at first glance. However, when 7 residue gaps were introduced into the PRD1 terminal protein, the three profiles were superimposable at the C-terminal

Figure 5. The predicted secondary structure of PRD1 terminal protein. (Ω), (Λ), and (—) denote alpha-helical, beta-sheet, and random-coil residues, respectively. Beta-turns are indicated by chain reversal. Arrows indicate the three methionine positions. The most likely tyrosine, residue 226, involved in linkage with DNA is boxed. The + and - signs refer to charged residues, and the numbers refer to residues in the terminal protein.

portion (residues 210-260) of the terminal proteins (Figure 4). This correspondence suggest that the carboxyl termini of the three terminal proteins are functionally homologous. The linking amino acids have been characterized in two cases, adenovirus terminal protein (31) and Ø29 terminal protein (34). In both cases, the linking sites are situated in the C-terminal portion of the proteins which are also hydrophilic areas. By analogy with these two systems, it is likely that the tyrosine at 226 is the linking site in the PRD1 terminal protein.

Secondary Structure

The secondary structure of the PRD1 terminal protein was predicted by the empirical method of Chou and Fasman (23). The PRD1 terminal protein has a 37% alpha-helical, 24% beta pleated sheet, 14% beta-turn and 25% random coil structure (Figure 5). The distribution of the alpha-helical and beta-sheet structure is not symmetrical: the N-terminal, one-third of the protein, is predicted to contain about 37% alpha-helical and 10% beta-sheet, while the C-terminal one-third contains 44% alpha-helix and 20% beta-sheet.

The predicted secondary structure of the PRD1 terminal protein contains two antiparallel beta-beta dimer structures at residues 116-136 and 163-180. The beta-beta dimers have been suggested to form a polypeptide double helix that could play a role in nucleic acid binding (35,23). The predicted linking site, tyrosine 226 is located in a beta-turn probably located at the surface of the terminal protein. Since there are only 3 internal methionine residues in the PRD1 terminal protein, the dGMP linking peptide can be identified using cyanogen bromide degradation.

Recent evidence indicates that the PRD1 DNA polymerase has segmental, but significant amino acid sequence homology with other DNA polymerases which include DNA polymerases of phage T4, Herpes virus, Vaccinia virus and adeno-

virus (11). These findings strongly suggest that those DNA polymerases which use either an RNA or protein primer have evolved from a common ancestral gene.

In contrast to DNA polymerase, there is no substantial amino acid sequence homology between the PRD1 terminal protein and the other terminal proteins so far described. However, there is significant similarity between the hydropathy profiles of the PRD1, Ø29, and Nf terminal proteins. Thus, it is likely that these three terminal proteins evolved from a common ancestor. The lack of substantial amino acid sequence homology may be explained by weak selective pressure on the terminal protein gene, allowing it to diverge considerably during evolution compared to the DNA polymerase gene.

ACKNOWLEDGEMENTS

We are grateful to Dr. Leonard Mindich for supplying phage PRD1 and strains. We are also grateful to Dr. H. Bernstein and Connie Innis for critical reading of this manuscript. This research was supported by National Institute of Health Grant GM28013.

REFERENCES

1. Olsen, R.H., Siak, J. and Gray, R.H. (1974) *J. Virol.* 14, 689-699.
2. Bamford, D.H., Rouhiaman, L., Takkinen, K. and Soderlund, H. (1981) *J. Gen. Virol.* 57, 365-373.
3. Mindich, L., Bamford, D., Goldthwaite, C., Laverty, M. and Mackenzie, G. (1982) *J. Virol.* 44, 1013-1020.
4. Bamford, D.H. and Mindich, L. (1984) *J. Virol.* 50, 309-315.
5. McGraw, T., Yang, H. and Mindich, L. (1983) *Mol. Gen. Genet.* 190, 237-244.
6. Gerendasy, D.D. and Ito, J. (1987) *J. Virol.* 61, 594-596.
7. Watson, J. (1972) *Nature* 239, 197-201.
8. Topal, M.D. and Fresco, J.R. (1976) *Nature* 236, 285-289.
9. Bjornsti, M.A., Reilly, B.E. and Anderson, D.L. (1984) *J. Virol.* 50, 766-772.
10. Larder, B.A., Kemp, S.D. and Darby, G. (1987) *EMBO J.* 6, 169-175.
11. Jung, G., Leavitt, M.C., Hsieh, J.C. and Ito, J. (1987) *Proc. Natl. Acad. Sci. USA* (in press).
12. Challberg, M.D. and Kelly, T.J. Jr (1979) *Proc. Natl. Acad. Sci. USA* 76, 655-659.
13. Nagata, R., Guggenheimer, A. and Huruntz, J. (1983) *Proc. Natl. Acad. Sci. USA* 80, 6177-6181.
14. Stillman, B.W. (1983) *Cell* 35, 7-9.
15. Watabe, K., Shih, M., Sugino, A. and Ito, J. (1982) *Proc. Natl. Acad. Sci. USA* 79, 5245-5248.
16. Salas, M. (1983) *Curr. Top. Microbiol. Immunol.* 109, 89-106.
17. Mindich, L. and McGraw, T. (1983) *Mol. Gen. Genet.* 190, 233-236.
18. Clewell, D.B. (1972) *J. Bacteriol.* 110, 667-672.
19. Dale, R.M.K., McClure, B.A. and Honchius, J.P. (1985) *Plasmid* 13, 31-40.
20. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5469.
21. Maxam, A.M. and Gilbert, W. (1980) *Methods in Enzymol.* 65, 499-650.
22. Kyte, J. and Doolittle, R.F. (1982) *J. Mol. Biol.* 157, 105-132.
23. Chou, P. and Fasman, G. (1978) *Adv. Enzymol.* 47, 45-148.

24. Mount,D.W. and Conrad,B. (1984) Nucl. Acids Res. 12, 811-817.
25. Shine,J. and Dalgarno,L. (1974) Proc. Natl. Acad. Sci. USA 71, 1342-1346.
26. Pribnow,D. (1975) J. Mol. Biol. 99, 419-443.
27. Yoshikawa,H. and Ito,J. (1982) Gene 17, 323-335.
28. Escarmis,C. and Salas,M. (1982) Nucl. Acids Res. 10, 5785-5798.
29. Shih,M., Watabe,K., Yoshikawa,H. and Ito,J. (1984) Virology 133, 56-64.
30. Paces,V., Vlcek,C., Urbanek,P. and Hostowsky,Z. (1985) Gene 38, 45-56.
31. Leavitt,M.C. and Ito,J. (1987) Nucl. Acids Res. 15, 5251-5259.
32. Smart,J.E. and Stillman,B.W. (1982) J. Biol. Chem. 257, 13499-13506.
33. Ito,J. and Roberts,R.J. (1979) Gene 5, 1-7.
34. Hermoso,J.M., Mendez,E., Soriano,F. and Salas,M. (1985) Nucl.Acids Res. 13, 7715-7728.
35. Carter,C.W. and Kraut,J. (1974) Proc. Natl. Acad. Sci. USA 71, 283-287.