J. Biochem. 2009;146(6)775–785
doi:10.1093/jb/mvp125

The Genomic Structure of *Thermus* Bacteriophage φIN93

Isao Matsushita[1],* and Hideshi Yanase[2]

[1]Energy Technology Laboratories, Osaka gas Co., Ltd, 6-19-9, Torishima Konohana-ku, Osaka 554-0051; and [2]Department of Biotechnology, Tottori University, 4-101 Koyamacho-minami, Tottori 680-8552, Japan

Received May 19, 2009; accepted July 24, 2009; published online August 12, 2009

We have determined the complete nucleotide sequence of the phage φIN93 is 19,604-bp long and contains 39 putative open reading frames. The functions for 20% of φIN93 gene products are similar to those expressed by other known phages and bacteria, and include peptidase, lytic enzymes, integrase, repressor protein and replication protein. The structural proteins of the φIN93 virion were identified through sodium dodecyl sulphate–polyacrylamide gel electrophoresis and found to have no similarity to those of other phages. We also determined the transcription initiation sites and classified four transcription units using the primer extension method. Three transcription units were transcribed in the same direction as part of the lytic cycle, while the remaining unit was transcribed in the opposite direction as part of the lysogenic cycle.

Key words: bacteriophage, genome, replication protein, *Thermus*, transcription unit.

Abbreviations: PCR, Polymerase chain reaction; Ile-tRNA, Isoleucine-tRNA.

Thermus sp. are Gram-negative bacteria that grow at temperatures ranging from 47°C to 85°C (*1*). We recently isolated a new extremely thermophilic phage, φIN93, which was induced from a lysogenic strain, *Thermus aquaticus* TZ2 (*2*). Electron micrographs of φIN93 showed polyhedral particles without tails, confirming that φIN93 differs from φYS40, the extremely thermophilic phage infectious to *T. thermophilus* HB8, which has a tail (*3*). In addition to their morphologies, φYS40 and φIN93 also differ with respect to the length and GC content of the genomes.

The only known thermophilic bacteria susceptible to φIN93 infection are *T. aquaticus* TZ2 and *T. thermophilus* HB8 (*T. thermophilus* HB27, *T. aquaticus* YT1, *T. rubber* and *T. filiformis* are not susceptible). Comparison of the 16S rRNA sequences of *T. aquaticus* TZ2 and *T. thermophilus* HB8 (*4*) revealed the two sequences to be completely identical, indicating the two organisms are of the same species. Thus, we renamed *T. aquaticus* TZ2 and *T. thermophilus* TZ2. It was therefore concluded that the same species serves as host for both φIN93 and φYS40.

To better understand the genomic structure of φIN93 and the functions of its genes, we determined the sequence of the φIN93 genome, searched for gene and protein homologs using the respective databases, and analysed the transcription units. Here, we describe the results of our study of the φIN93 genome and the proteome of the φIN93 virion.

MATERIALS AND METHODS

Cell Growth and Phage Infection—Thermus thermophilus TZ2 was grown overnight at 70°C in A-2 medium consisting of 0.1% tryptone, 0.1% yeast extract and Castenholtz basal salts (pH 7.0) (*2*, *5*). To amplify φIN93, the overnight culture was added to fresh A-2 medium to a final concentration of 1% v/v (10^8 cells/ml), and then incubated for 3 h at 70°C while shaking in a Taitec BioShaker BR-21FH at 220 r.p.m. To this fresh culture, a stock solution of φIN93 was added to a multiplicity of infection of about 0.4, after which the culture was incubated under the same conditions until complete lysis had occurred (~2.5 h). The cell debris was then removed by centrifugation at 2,800*g* for 10 min at room temperature using a T4SS rotor in a Hitachi Himac CT6E centrifuge. The φIN93-containing supernatant was then collected and stored at 4°C.

Purification of φIN93 Virions—φIN93 virions were prepared as described earlier (*6*). DNase I and RNase A were added to the φIN93-containing lysate to final concentrations of 1 μg/ml each, and the mixture was incubated for 30 min at 30°C. Thereafter, 2.5 M NaCl solution containing 25% polyethylene glycol 8,000 was added to a final concentration of 20% (v/v), and the mixture was incubated on ice for 1 h before being centrifuged at 15,000*g* for 30 min at 4°C using an RPR12-2 rotor in a Hitachi Himac CR20B3 centrifuge. The precipitated pellets were suspended in 2 ml of 10-mM ammonium acetate buffer containing 5-mM $MgSO_4$ (pH 6.0), after which the resultant suspension was subjected to 1.15–1.7 g/ml CsCl equilibrium density gradient centrifugation at 45,000*g* for 1 h at 4°C using an 80Ti rotor in a Beckman L8M centrifuge. The purified φIN93 suspension was collected

*To whom correspondence should be addressed. Tel: +81-(0)6-6205-3089, Fax: +81-(0)6-6462-3433,
E-mail: imatsu@osakagas.co.jp
Note: The nucleotide sequence of φIN93 genome are available in the DDBJ/EMBL/GenBank databases under the accession number AB063393.

and dialyzed overnight against a 1,000-fold volume of the ammonium acetate buffer.

Extraction of φIN93 DNA—An equal volume of phenol–chloroform (1 : 1 v/v) was added to the purified phage suspension, mixed and centrifuged at 2,800*g* for 10 min at room temperature using a T4SS rotor in a Hitachi Himac CT6E centrifuge, after which the aqueous phase was extracted. The DNA was then precipitated in ethanol, collected and dissolved in TE buffer (10 mM Tris–hydrochloride and 1 mM EDTA, pH 8.0).

DNA Sequencing—*Escherichia coli* JM109 and plasmid vector pUC19 were purchased from TakaraBio Co. Samples of φIN93 DNA were then separately digested with BamHI, Hind III, PstI, Hind III/XbaI, SphI or SphI/KpnI, after which the fragments were cloned into pUC19 using standard methods (*7*). *Escherichia coli* JM109 harbouring pUC19 or its derivatives were selected on Luria-Bertani plates containing 50 μg/ml Ap, 40 μg/ml X-gal and 0.4 mM IPTG. In addition, φIN93 DNA fragments were amplified by PCR (Gene Amp PCR System 9700 Applied Biosystems Co.) using synthesized primers using a PCR kit (TOYOBO Co.). The nucleotide sequences of the plasmids and amplified DNA fragments were then determined in both directions using universal primer, reverse primer and synthesized primers (around 90 primers) with an ALF DNA sequencer (Amersham Pharmacia Co).

Blotting of Phage Virion Proteins and Determination of N-terminal Amino acid Sequences—Purified φIN93 virions heated for 5 min at 90°C in sodium dodecyl sulphate (SDS) and 2-mercaptoethanol were separated by 15% SDS–polyacrylamide gel electrophoresis (SDS–PAGE) and blotted (2 mA/cm^2 for 30 min) onto a PVDF membrane using a Horize Blott AE-6677 according to the manual (ATTO Co.). The N-terminal amino acid sequences of the proteins were determined using PROCISE-cLC (Applied Biosystems Co.).

PCR Analysis of attP, attL and attR—To isolate φIN93 lysogenic host cells, in the above mentioned phage infection experiment, the culture was further incubated for 16 h under the same conditions after complete lysis had occurred. To isolate the living cells in the culture, the culture (100 μl) was spread on A-2 medium containing 3% agar and incubated at 70°C for 16 h. Then, the isolates were examined to be lysogenic host cells by φIN93-induction with mitomycin C (*2*). The genomic DNA of the lysogenic host cells was prepared using Isoplant (NIPPON GENE Co.). PCR was performed on the genomic DNA using synthetic primers (primer A: 5′-CCTCGGCGGATCGGCGACCC-3′, primer B: 5′-CTGG CAGGACCACGGGCTC-3′, primer C: 5′-CCCGGGTGCA GTACGCCAC-3′ and primer D: 5′-CCCTGGGCATCCCC GCGAGGCC-3′). The PCR products were analysed by electrophoresis through a 1.5% agarose gel and sequenced.

Primer Extension Experiments—Total RNA was prepared from φIN93-infected *T. thermophilus* TZ2 cells. Cells infected with φIN93 were harvested before complete lysis occurred (incubation time for infection: ~1 h), and RNA was extracted using an Aurum Total RNA Mini Kit (BioRad Laboratories, Inc.). Primer extension analysis was performed using synthetic biotinylated primers (No. 28: 5′-GCGCACCGCCCCTAAGCG-3′, No. 5: 5′-GGGCCGCCACGATCTCG-3′, No. 12: 5′-CCACGCTCA CGCTGACCAC-3′, No. 36: 5′-GCCTCACCCCATGCTGGT C-3′) with a Primer Extension System-AMV Reverse Transcriptase Kit (Promega Co.). Following reverse transcription, the synthesized DNA was analysed by electrophoresis on an 8% SDS–PAGE containing 8 M urea. As a standard, a sequence ladder was prepared with φIN93 DNA using a Sequencing High Non-Radio Isotopic DNA Sequencing Kit (TOYOBO Co.). An Imaging High Non-Radio Isotopic Detection Kit (TOYOBO Co.) was then used to detect the synthesized DNA.

Computer Analysis—Open reading frames (ORFs) and hairpin structures in the φIN93 DNA were predicted using GENETYX WIN software (Software Development Co.). BLAST and PSI-BLAST were used to detect homologous genes and proteins in the respective DNA and protein databases.

RESULTS

Complete Nucleotide Sequence of Phage φIN93 DNA—φIN93 is a double-stranded and a circular DNA phage whose genome contains 19,604 bp, the sequence of which has now been determined. The φIN93 genome is predicted to contain 39 ORFs (Table 1) with lengths ranging from 114 to 1,239 bp, occupying 91% of the tightly packed genome. There are 15 instances of overlap (from 1 to 35 bases long) between neighbouring ORFs. Thirty-five of the predicted ORFs start at the AUG codon, while four ORFs start with the GUG codon. At the ends of four of the φIN93 genes are TAA stop codons; 24 genes end with TGA codons and 11 end with TAG codons.

Thirty-five ORFs (ORFs 1–35), or ~80% of the φIN93 genome, are consecutively transcribed in the same direction, rightward on the gene map (Fig. 1). These ORFs are predicted to be involved in the lytic cycle and to encode a replication protein, structural proteins making up the phage virion, and lytic enzymes. The remaining four ORFs (ORFs 36–39) are consecutively transcribed in the opposite direction (Fig. 1), and encode integrase, endonuclease and repressor proteins are related to the lysogenic cycle.

Replication Proteins—Encoded by ORF1, the protein gp1 is comprised of 412 amino acid residues and is similar to the replication protein repA (402 amino acid residues) encoded in the *Thermus sp*. plasmid (*8*). Examination of pair-wise alignment with repA indicated that gp1 shows 28.1% similarity to the repA (Fig. 2A). In addition, a 116-bp high AT region (AT content: 54.3%) was present (Fig. 2B), which based on the analysis of repA, is predicted to be the region of the replicative origin.

Proteins Making up the Phage Virion—SDS–PAGE analysis of the φIN93 virion purified as described in the MATERIALS AND METHODS section revealed three thick protein bands and seven thin bands (Fig. 3). The N-terminal amino acid sequences were determined for all but one of these proteins; the N-terminal amino acid of the protein in band 'g' was blocked. We found that eight of the proteins correspond to ORF2, 5, 12, 13, 17, 20, 21 and 23, respectively (Table 2). These ORFs

Table 1. **Gene products of phage φIN93 and their predicted molecular functions.**

ORF No.	ORF strand/ position[a]	Length (a.a.)	Function	Organism source	Score (bits)/ *E*-value	Identity	Accession number
1	23–1261	412	Replication protein repA 402 a.a.	*Thermus sp*. Plasmid	64 2e–08	28.1	O05481
2	1251–1595	114	Unknown				
3	1615–2004	129	Unknown				
4	2021–2314	97	Unknown				
5	2416–2976	186	Unknown				
6	3027–3341	104	Unknown				
7	3342–3863	173	Unknown				
8	3856–4446	196	m23/m37 peptidase domain protein 214 a.a.	*Carboxydothermus hydrogenoformans*	65 2e–09	20.9	Q3A8W6
9	4443–4646	67	Unknown				
10	4646–4840	64	Unknown				
11	4830–5510	226	Unknown				
12	5647–6063	138	Unknown				
13	6074–6589	171	Unknown				
14	6598–7473	291	Unknown				
15	7486–7899	137	Unknown				
16	7902–8156	84	Unknown				
17	8153–8806	217	Unknown				
18	8830–9030	66	Unknown				
19	9020–9322	100	Unknown				
20	9331–9564	77	Unknown				
21	10274–10726	150	Unknown				
22	10804–11766	320	wd40-like beta-propeller protein 560 a.a.	*Thermus* phage P74–26	72 6e–11	18.1	A7XXT3
			Putative uncharacterized protein 560 a.a.	*Thermus* phage P23–45	69 3e–10	18.4	A7XXD9
23	11925–12842	305	Putative uncharacterized protein 297 a.a.	*Thermus* phage P23–45	50 2e–04	31.5	A7XXE1
			Putative uncharacterized protein 297 a.a.	*Thermus* phage P74–26	50 2e–04	29.8	A7XXT5
24	12844–13119	91	Putative uncharacterized protein 82 a.a.	*Thermus* phage P23–45	47 5e–04	25.3	A7XXE2
			Putative uncharacterized protein 82 a.a.	*Thermus* phage P74–26	44 0.004	23.1	A7XXT7
25	13131–13397	88	Unknown				
26	13363–14082	239	Lytic transglycosilase (= lysozyme like) 576 a.a.	*Shigella flexneri serotype* 5b	54 1e–05	13.4	Q0T7X0
27	14093–14350	85	Unknown				
28	17960–18163	67	Excisionase/xis, DNA binding protein 74 a.a.	*Saccharopolyspora erythraea*	46 9e–04	34.3	A4FPZ5
29	18160–18336	58	Unknown				
30	18333–18446	37	Unknown				
31	18450–18758	102	Unknown				
32	18767–18991	74	Unknown				
33	18988–19167	59	Unknown				
34	19146–19430	94	Unknown				
35	19403–27	75	Unknown				
36	–/17843–17187	218	lexA-like repressor 236 a.a.	*Deinococcus geothermalis*	132 3e–29	39.7	Q1IYM3
			Putative phage repressor 261 a.a.	*Desulfovibrio vulgaris*	47 2e–06	45.5	B8DJI7
37	–/17185–16721	154	Unknown				
38	–/16589–15696	297	Endonuclease 177 a.a.	*Gluconobacter oxydans*	53 2e–05	10.4–9.8	Q5FN72
			Putative Mrr restriction system protein 307 a.a.	*Streptomyces avermitilis*	48 9e–04		Q82P74
39	–/15531–14482	395	Phage integrase 390 a.a.	*Moorella thermoacetica*	197 1e–48	34.9	Q2RHJ2

[a]The positions of the ORFs in the phage φIN93 genome are given by nucleotide number from the *Bam*HI site of 0 position. "–" indicates a leftwards transcription orientation.

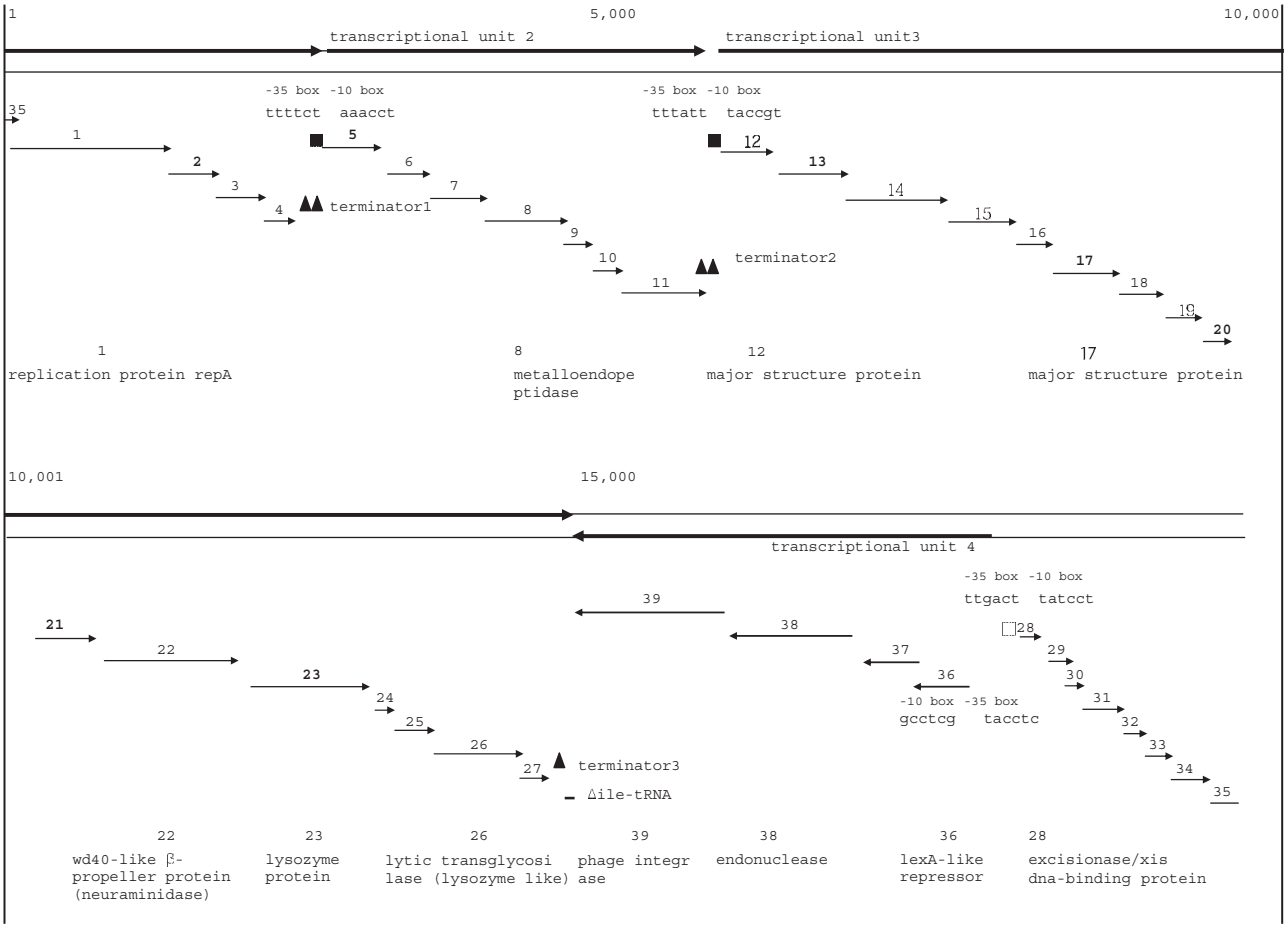

Fig. 1. **Gene map of the φIN93 genome.** The ORFs are indicated by thin arrows and numbered from 1 to 39. The transcription units are indicated by thick arrows. The numbers of ORFs corresponding to phage structural proteins are shown in bold. Filled triangles (filled triangle) indicate hairpin structures regarded as terminators of transcription units. Open squares show the *Thermus* promoters. Filled squares show the phage promoters and the −35 box and −10 box of the phage promoters are indicated

are oriented in the same direction, and homology analysis revealed that their products are not similar to proteins from any other known phages; that is, they are inherent to φIN93. In the result of homology analysis with protein databases on the N-terminal amino acid of the thick protein band 'a', it corresponded to the DNA binding protein HU, encoded in the genome of *T. thermophilus* HB8 but not in that of φIN93. So, the existence of the identical HU gene in the host *T. thermophilus* TZ2 genome was also confirmed by PCR and sequence analysis (data not shown). However, to clarify the function of the HU in φIN93, we have to examine whether the HU simply binds to the surface of the phage virion or contributes to the stabilization of the φIN93 DNA within phage virion.

ORF12, encoding the protein in the thick band 'b' (gp12), immediately follows a φIN93 inherent promoter, which has strong transcriptional activity (described later). ORF17, encoding the protein in the thick band 'h' (gp17), contains a TTT sequence adjacent to a Shine-Dalgarno (SD) sequence. In other phages, including T7 (gene 10), T4 (gene 23) and Qβ (gene C) (*9*), this sequence with consecutive Ts functions as an enhancer during transcription of some coat protein genes (Table 3). This likely explains why gp12 and gp17 are expressed at high levels, and we predict that they function as coat proteins. For all but two bands, the molecular weights determined by SDS–PAGE were in good agreement with those deduced from the ORFs. The two exceptions were ORFs 2 and 20. The accuracy of the nucleotide sequences of those two ORFs and their surrounding region was considered to be good because the confidence levels for the results of the sequencing, which was carried out several times in both directions, were high. Notably, if a −1 frame shift in ORF2 and a +1 frame shift in ORF20 had occurred, the deduced molecular weights would have been in good agreement with those of gp2 and gp20, respectively. However, further study will be needed to determine whether such frame shifting actually occurred.

Lytic Proteins—Encoded by ORF22, gp22 is similar to the WD40-like β-propeller protein from *Thermus* phage P74-26 (*10*). β-Propeller proteins have a propeller-like structure comprised of 4–8 blade-shaped β-sheets around a central point, which functions as the active site when the protein is an enzyme. WD40 domains are also believed to play a crucial role in the interaction with

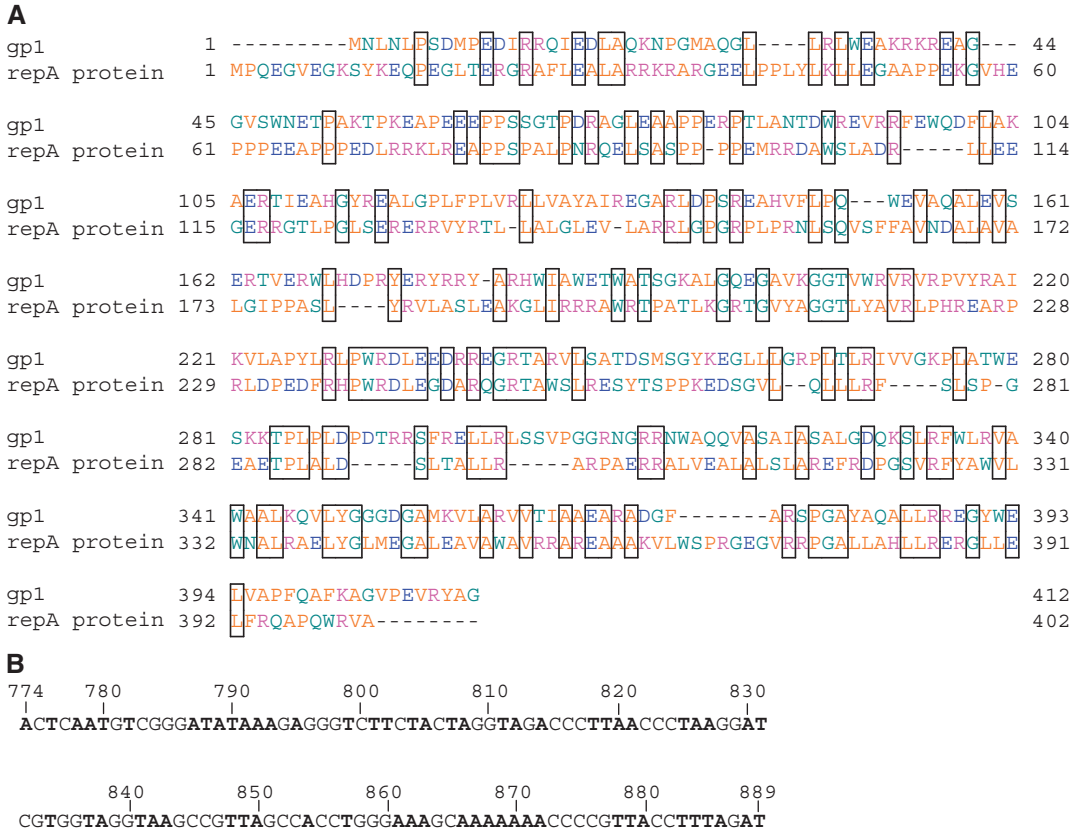

Fig. 2. **Replication origin region of φIN93.** (A) Pair-wise alignment of repA from *Thermus* plasmid with gp1. Identical amino acid residues are boxed. The yellow type indicates the hydrophobic residues, blue type indicates the acidic residues, pink type indicates the basic residues and green type indicates the neutral residues. (B) The AT-rich region in ORF1. The A and T nucleotides are shown in bold.

other proteins. The neuraminidase of the influenza virus is a β-propeller protein with six blade-shaped β-sheets. It is a lysis-related enzyme found on the surface of the virion. gp22 also may be a lysis-related neuraminidase.

Encoded by ORF23, gp23 is comprised of 305 amino acid residues and is similar to an uncharacterized protein (297 amino acid residues) from *Thermus* phages p23-45 and p74-26 (*10*). We previously isolated a lysozyme from φIN93 lysate (*11*); the N-terminal sequence of gp23 was determined to be identical to the corresponding sequence of the isolated lysozyme.

Encoded by ORF24, gp24 is comprised of 91 amino acid residues and is similar to another uncharacterized protein (82 amino acid residues) from *Thermus* phages P23-45 and P74-26 (*10*). In contrast, gp25, encoded by ORF25, is not similar to any other known proteins.

Encoded by ORF26, gp26 is similar to lytic transglycosylase (*i.e.* it is lysozyme-like) from *Shigella flexneri* serotype 5b, and is therefore predicted to be a lytic transglycosylase. Transglycosylases are lysis-related enzymes that cleave the β-1, 4-glycosidic bonds of *N*-acetylmuramic acid and *N*-acetylglucosamine.

The results summarized above suggest that ORF22, ORF23 and ORF26 all encode lysis-related proteins. In *E. coli* phage λ, genes encoding three lysis-related proteins (S′, R, Rz) have been identified and shown to be situated in tandem within the genome. Thus ORF24 and ORF25, situated between ORF22 and ORF26, may also be lysis-related proteins.

Regulatory Proteins—Encoded by ORF36, gp36 has 218 amino acid residues and is similar to the lexA-like repressor (236 amino acid residues) from *Deinococcus geothermalis*, and is also similar to putative phage repressor (261 amino acid residues) from *Desulfovibrio vulgaris*. The locus of ORF36 in φIN93 genome is the first ORF involved in progression of the lysogenic cycle and gp36 is thought to be a repressor having a function involved in transcriptional regulation between the lysogenic cycle and lytic cyle.

Excisionase, Endonuclease, Integrase, and Target Site of Integration—Encoded by ORF28, gp28 has 67 amino acid residues and is similar to excisionase (74 amino acid residues) *from Saccharopolyspora erythraea*, which catalyses DNA excision and, along with integrase, contributes to lysogenization. Although in the *E. coli* phage λ genome, the excisionase gene is situated immediately downstream of integrase gene, in the φIN93 genome it is at a distance from integrase gene (ORF39, described later).

Encoded by ORF38, gp38 has 298 amino acid residues and is similar to the endonuclease (177 amino acid residues) from *Gluconobacter oxydans* and the putative

Mrr restriction system protein (307 amino acid residues) from *Streptomyces avermitilis*. We therefore predict that gp38 is likely an enzyme involved in the restriction system mediating phage infection of host cells.

Encoded by ORF39, gp39 has 396 amino acid residues and is similar to the phage integrase (390 amino acid residues) from *Moorella thermoacetica*. We therefore predict that gp39 is an integrase needed for integration of φIN93 DNA into the host chromosome. In the case of *E. coli* phage λ, the integrase gene is located immediately downstream of the endonuclease gene, and a similar configuration was noted in the φIN93 genome, as well.

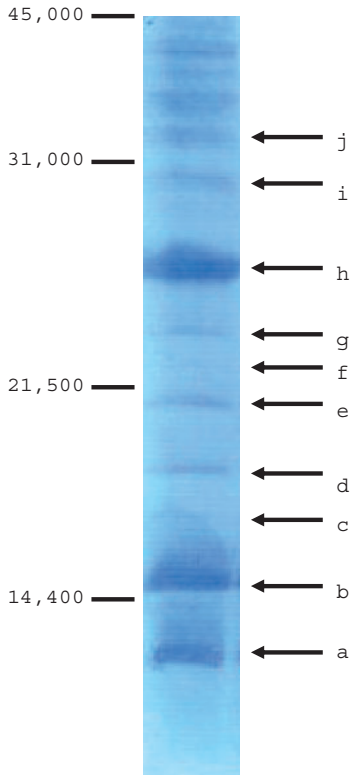

Fig. 3. **SDS–PAGE analysis of φIN93 phage virion proteins.** SDS gel shows the protein composition of purified φIN93 virions. Phage virions dissolved in 2% SDS and 2-mercaptoethanol were subjected to SDS–PAGE. Protein bands were stained with Coomassie Brilliant Blue R-250.

Within the genome of *E. coli* phage λ, there is an attachment site, attP, downstream of the integrase gene. However, studies of the bacteriophages P4 and P22 and the virus-like particle SSV1 have shown that, in some cases, tRNA genes are used as integration sites (*12–14*). In the case of φIN93, a partial *Thermus* Ile-tRNA1a sequence (*15*) with a 32-bp deletion at the 3′-terminus was located downstream of the integrase gene (ORF39) (Fig. 4). In the results of PCR experiments on the genomic DNA of lysogenic host cells, the DNA fragment (709 bp) containing attL was amplified by primers A and B, while the DNA fragment (881 bp) containing attR was amplified by primers C and D. Sequence analysis of the boundaries (*att*L and *att*R) of the φIN93 DNA integrated into the genome of the lysogenic host cell revealed that the partial Ile-tRNA1a sequence was used for attachment of the core sequence. φIN93 integration occurred through recombination of the partial Ile-tRNA1a sequence (*att*P) in the φIN93 genome and the identical sequence of the Ile-tRNA gene in the host genome (Fig. 5A–C).

Endopeptidases—Encoded by ORF8, gp8 has 196 amino acid residues and is similar to M23/M37 peptidase domain protein (216 amino acid residues) and the Gly–Gly endopeptidase from *Carboxydothermus hydrogenoformans*. We therefore predict that gp8 is an endopeptidase.

Polymerases—We found no ORFs within the φIN93 genome encoding a protein similar to a DNA polymerase or RNA polymerase from any known phage or bacterium.

Table 3. **The nucleotide sequences adjacent to SD sequences of φIN93 ORF17 and the other phage coat protein genes.**

Gene	
T7 gene 10 (a coat protein)	TTTGTTTAAC**TTT**AAGAAGGAGA
T4 gene 23 (a coat protein)	CGTTTCTAAT**TTT**AAAGGTTAAC
Qβ gene C (a coat protein)	GCGTTGAAAC**TTT**GGGTCAATTT
φIN93 ORF17	GCCGTGCTGG**TTT**GGAGGTGGA

SD sequences are underlined. Ts nucleotides, considered enhancer sequence, are shown in bold letters.

Table 2. **The NH_2-terminal amino acid sequences of phage virion proteins and the corresponding ORF.**

Protein band	Molecular weight estimated by SDS–PAGE	NH2-terminal a.a. sequence	ORF No.	Molecular weight deduced from ORF
a	13,000	AAKKTVTKAD	fonud in the host genome	10,321
b	14,400	ADTAAIAAQD	12	14,569
c	16,500	MQGAWIVAVA	20	7,700
d	18,000	VLNLLGIKQE	21	16,119
e	20,500	AYQRVPVDPN	5	19,252
f	22,500	MQEAFERIKR	13	19,293
g	24,000	not determined	–	–
h	26,000	MSENTLLGIAA	17	23,278
i	30,000	MRVDRV	2	12,813
j	32,500	MSVRITNFGLD	23	33,141

```
Thermus thermophilus HB8          10         20         30         40         50         60         70
  Ile-tRNA1a              TGGTGGGCGA TGGTGGACTT GAACCACCGA CCTCACGCTT ATCAGGCGTG CGCTCTAACC AGCTGAGCTA ATCGCCC
                          ********** ********** ********** ********** ***** **    *     *    * *  *       *
φIN93 partial             TGGTGGGCGA TGGTGGACTT GAACCACCGA CTCACGCTTA TCAGCCCGCA AGGGAGGAGG GGTTAGGAAC CGGGTTA
  Ile-tRNA1a
```

Fig. 4. **Sequence comparison of Ile-tRNA from *T. thermophilus* HB8 and the partial Ile-tRNA from φIN93.** The asterisks indicate identical sequences.

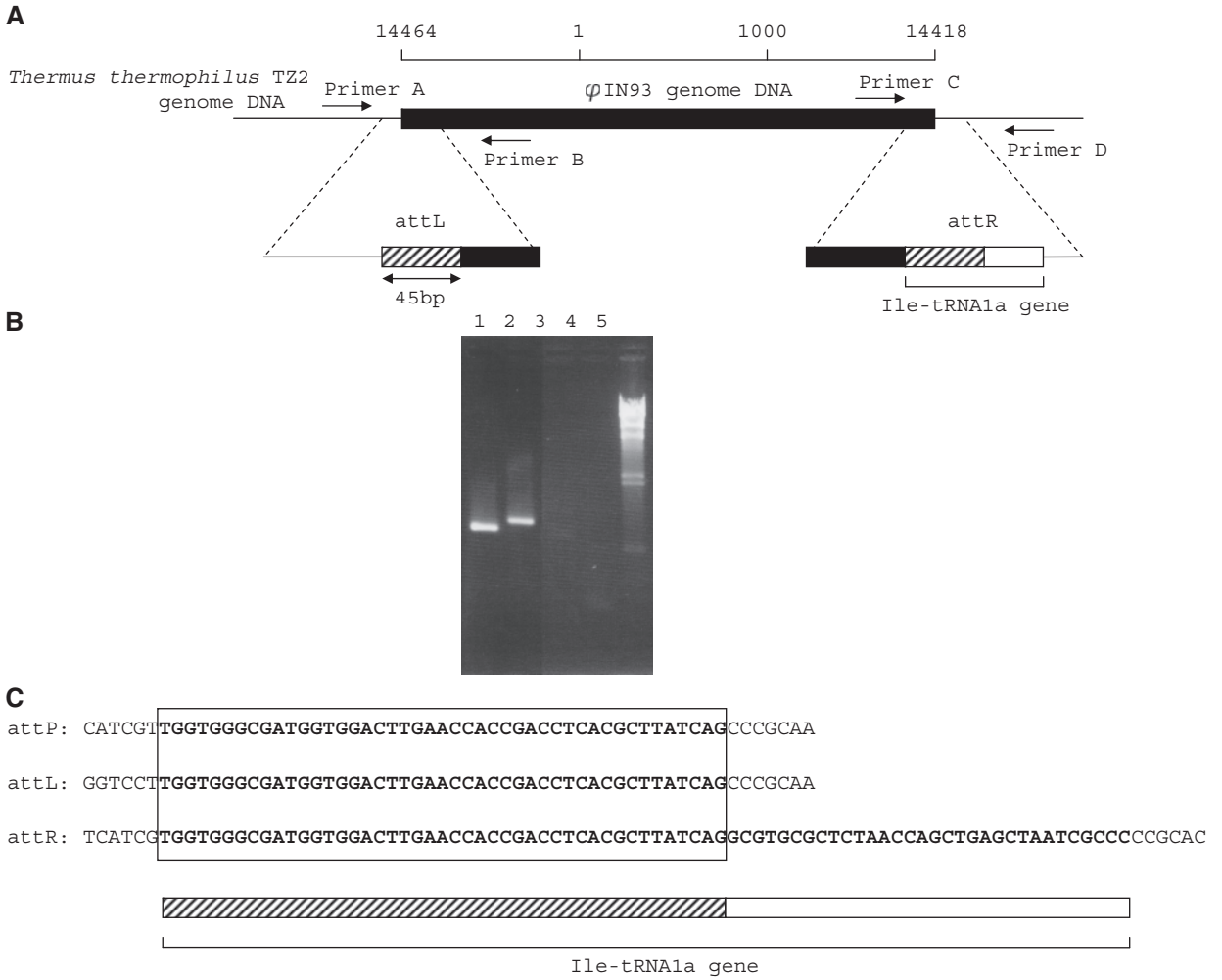

Fig. 5. **Integration of φIN93 into the host cell.** (A) Physical map of the integrated φIN93 genome. Filled bars represent the φIN93 genome; hatched bars represent the 45-bp attachment core sequence (*att*P). (B) PCR analysis on attR and attL. Lane 1: with primers A and B for attL (709bp), lane 2: with primers C and D for attR (881bp), lane 3: with primers A and C, lane 4: with primers B and D, lane 5: λ DNA/HinDβ. PCR of lanes 3 and 4 were used as a control. (C) DNA sequence comparison between the boundaries of the integrated φIN93 DNA (*att*L and *att*R) and the corresponding region in the φIN93 genome (*att*P). The 45-bp attachment core sequence is boxed.

We therefore suggest that φIN93 utilizes polymerases expressed by the host *T. thermophilus* TZ2.

Terminators of Transcription—To determine the transcription units in the φIN93 genome, we first used the Genetyx program to search for terminators of transcription that have hairpin structures (estimated on the basis of >90% matching stem parts) and identified three such structures downstream of ORF4, ORF11 and ORF27, respectively. The predicted terminator sequence behind ORF4 consisted of two consecutive hairpin loops [found at nucleotide positions 2,324 and 2,349, 10 nucleotides downstream from the translational stop codon (TAA)], with predicted energies of −19.5 kcal/mol (about −82 kJ/mol) for the 9-base-long inverted repeat and −22.0 kcal/mol (about −92 kJ/mol) for the 10-base-long inverted repeat, followed by a T-rich stretch (4 Ts) (Fig. 6A).

The predicted terminator sequence behind ORF11 also consisted of two consecutive hairpin loops [found at nucleotide positions 5,517 and 5,561, seven nucleotides

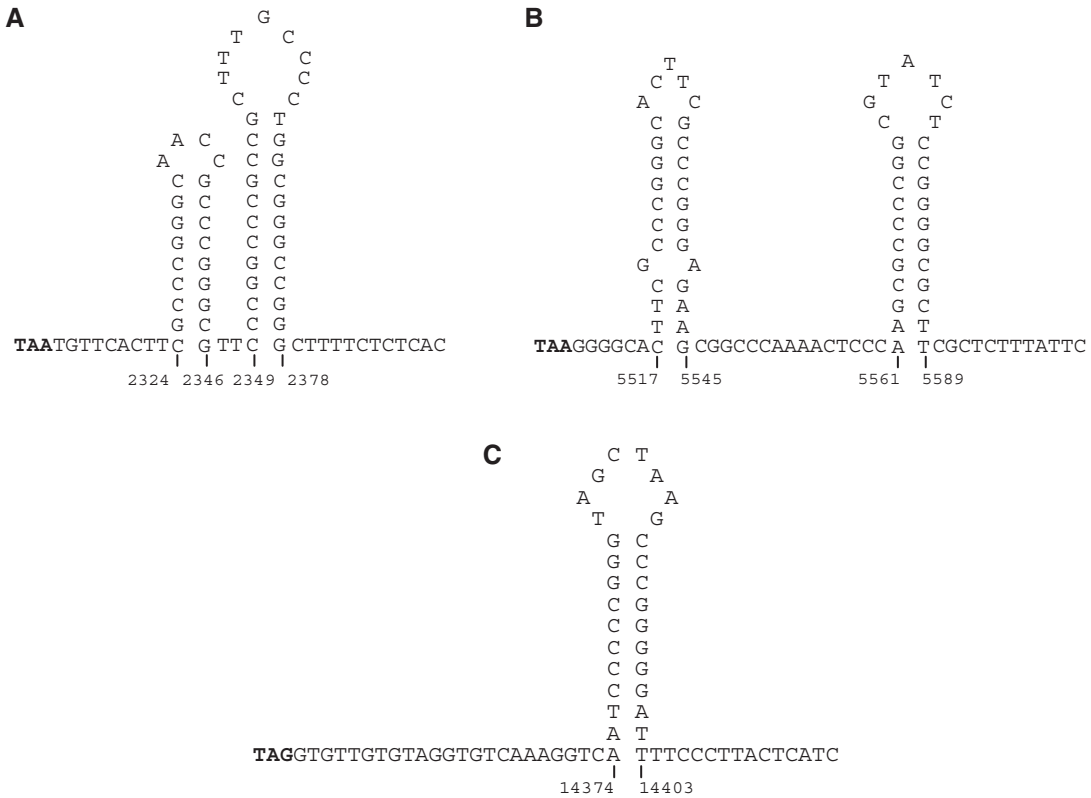

Fig. 6. **The hairpin structures regarded as terminators.** Terminator (A) is located behind ORF4. Terminator (B) is located behind ORF11. Terminator (C) is located behind ORF27 and behind ORF39. Stop codons are shown in bold.

downstream from the translational stop codon (TAA)], with predicted energies of −18.1 kcal/mol (about −75.6 kJ/mol) for the 10-base-long inverted repeat and −20.3 kcal/mol (about −84.8 kJ/mol) for the 11-base-long inverted repeat, followed by a T-rich stretch (5 Ts). Notably, the hairpin structures downstream of ORFs 4 and 11 were very similar, with 45 of 64 bp matching (Fig. 6B).

The predicted terminator sequence behind ORF27 consisted of one hairpin loop [found at nucleotide position 14,374, 24 nucleotides downstream from the translational stop codon (TAG)], with a predicted energy of −21.8 kcal/mol (about −91 kJ/mol) for the 11-base-long inverted repeat, followed by a T-rich stretch (4 Ts) (Fig. 6C). Because all the sequences in the aforementioned hairpin structures had GC-rich stem parts and were accompanied by Ts, we predict they are Rho-independent terminators.

Transcription Initiation Sites and Transcription Units—Given the results of our analysis of gene function and transcription terminators, we next carried out a set of primer extension experiments to search for the 5′-termini of the mRNAs [the transcription initiation sites (+1)] (Fig. 7). In the lytic cycle, transcription initiation sites were deduced to be upstream of ORF28, ORF5 and ORF12. In the lysogenic cycle, a transcription initiation site was deduced to be upstream of ORF36.

The transcription initiation site (+1) upstream of ORF28 was found at guanine 17,957, three nucleotides upstream of the translational start codon (ATG), and the fragment synthesized by reverse transcription was detected as a thin band (Fig. 7A). Seven nucleotides upstream of the transcription initiation site (+1) is the sequence TATCCT, which is a putative −10 box, and another 18 nucleotides distant is the sequence TTGAC T, which is a putative −35 box. These nucleotide sequences are identical to the promoter sequences recognized by host *Thermus* sp. RNA polymerase (*16*, *17*).

Upstream of ORF5, behind the hairpin loop (transcription terminator downstream of ORF4), the transcription initiation site (+1) was found at adenine 2,416, the same position as the translational start codon (ATG), and the fragment formed a thick band (Fig. 7B). Seven nucleotides farther upstream is a putative −10 box (AAACCT), and another 17 nucleotides distant is a putative −35 box (TTTTCT).

Similarly, upstream of ORF12, behind the hairpin loop (transcription terminator downstream of ORF11), the transcription initiation site (+1) was found at adenine 5,632, 15 nucleotides upstream from the translational start codon (ATG), and the fragment formed a thick band (Fig. 7C). Eight nucleotides farther upstream is a putative −10 box (TACCGT), and another 18 nucleotides distant is a putative −35 box (TTTATT). A characteristic common to the promoter nucleotide sequences upstream of ORF5 and ORF12 is the continuous Ts in the −35 box. This makes the promoters upstream of ORF5 and ORF12 inherent to φIN93 and different from host *Thermus* sp.

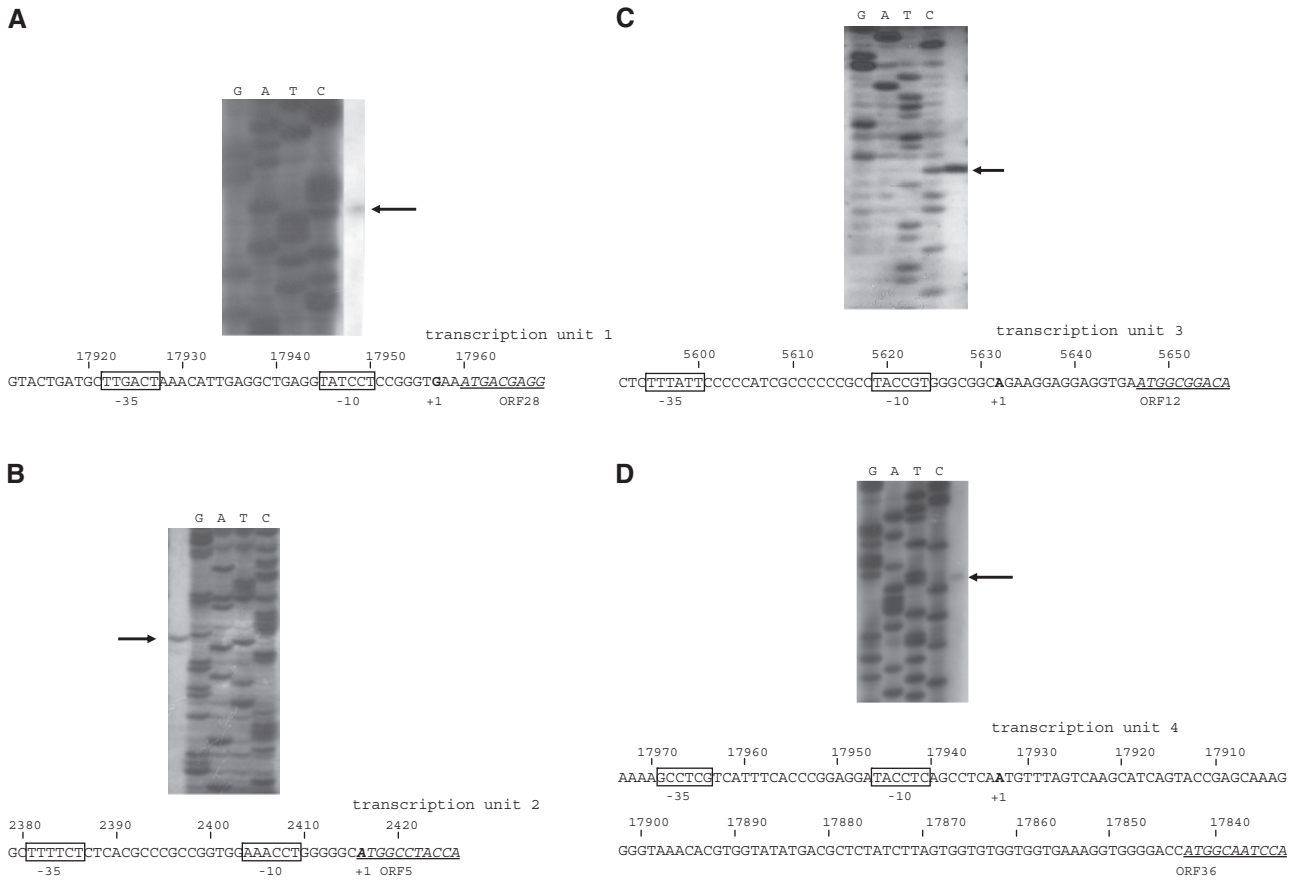

Fig. 7. **Primer extension analysis of each transcription unit.** The arrow indicates the estimated 5′-end of the mRNA in each transcription unit. The DNA sequence ladders shown were synthesized using the same primer. The putative −10 and −35 promoter regions are boxed, and the transcription initiation sites are indicated in bold (+1). The nucleotide sequences of ORFs are shown in italics and underlined. (A) transcription Unit 1, (B) transcription Unit 2, (C) transcription Unit 3 and (D) transcription Unit 4.

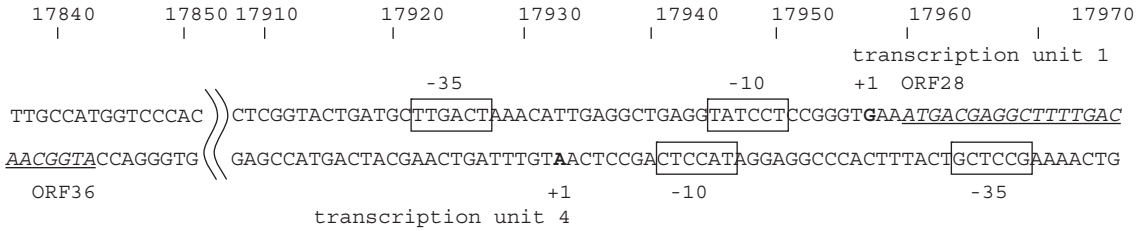

Fig. 8. **Promoter regions of transcription Units 1 and 4.** The putative −10 and −35 promoter regions of transcription Units 1 and 4 are boxed, and the transcription initiation sites are indicated in bold. The nucleotide sequences of the ORFs are shown in italics and underlined.

promoter sequences. In addition, the thickness of the detected bands is indicative of the strong transcriptional activities of these promoters.

Upstream of ORF36, at the first ORF contributing to the lysogenic cycle, the transcription initiation site (+1) was found at adenine 17,843, and the synthesized fragment formed a thin band (Fig. 7D). Eight nucleotides farther upstream is a putative −10 box (TACCTC); however, only a GCCTCG sequence with a low AT content was found at the location corresponding to the −35 box, another 17 nucleotides distant. If indeed the sequence of the −10 box is TACCTC, the TA dinucleotide overlaps the putative −10 box TATCCT upstream of ORF28 (the first ORF contributing to the lytic cycle). This may indicate that this −10 box is an important region for transcriptional regulation of the lytic and lysogenic cycles (Fig. 8).

From the above analysis, we arrived at the conclusion that the ORFs of the φIN93 genome can be divided into four transcription units. Transcription Unit 1, the first

unit in the progression of the lytic cycle, is ORF28–ORF4, which encode excisionase, replication protein and a transcription factor. Transcription Unit 2 is ORF5–ORF11, which encode structural proteins. Transcription Unit 3 is ORF12–ORF27, which encode proteins related to lytic enzyme, thus completing the requirements of a lytic cycle. Finally, transcription Unit 4 is ORF36–ORF39, which are transcribed in the direction opposite to that of transcription Units 1–3, and encode the repressor proteins, endonuclease and integrase required for the lysogenic cycle.

DISCUSSION

Only three ORFs within the genome of φIN93 showed similarity to ORFs of *Thermus* P23-45 and *Thermus* P74-26, and no similarity was found to the genome of φYS40 (*18*). The GC content (65.9%) of the φIN93 DNA is almost as high as in the host *Thermus* sp. DNA, while *Thermus* P23-45 DNA (57.8%) and P74-26 DNA (57.8%) have lower GC contents than φIN93, and φYS40 DNA has the lowest GC content (32.6%) among these phages, approximately the same as mesophilic bacteria. A comparison of genome sizes revealed that φIN93 has the smallest genome (19,604 bp), followed by *Thermus* P23-45 (84,201 bp), *Thermus* P74-26 (83,319 bp) and φYS40 (152,372 bp). Among these four phages, φIN93 appears to be the most thermophilic and the simplest, possessing only the basic functions needed for survival in a thermophilic environment. Thus, φIN93 may have evolved in a manner different from the other three phages, despite being able to infect the same *Thermus* sp.

When we digested the φIN93 DNA with various restriction enzymes, cohesive end site (*cos*) was not detected. From homology search, terminal protein was not found, either. In addition, a similarity between the gp1 of φIN93 and the replication protein of a *Thermus sp*. plasmid was found. To analyse the function of gp1, we constructed a recombinant plasmid that the ORF1 and kat genes were inserted into pBluescriptII SK(+), and we clarified that the gp1 has the function as a replication protein because the plasmid was amplified in *T. thermophilus* HB27. Thus, φIN93 may have the same replication system as the *Thermus sp*. plasmid, which differs from that of rolling circle form used by *E. coli* phage λ. Now we are preparing for submission on this result.

Our analysis of the transcription initiation sites and the promoters suggest that transcription Units 1 and 4 are both transcribed by host *Thermus* RNA polymerase as an early event after infection of a *Thermus* cell by φIN93. Thereafter, transcription Units 2 and 3 may be transcribed by host RNA polymerase recognizing inherent φIN93 promoter sequences instead of *Thermus* promoters. This is similar to the case of *Bacillus subtilis* bacteriophage SPO1, which develops via transcription driven by a different promoter sequence at every stage (*19*).

In φYS40, Sevostyanova *et al*. (*20*) described that most of middle/late are leaderless, and in contrast the vast majority of host as well as early phage transcripts contain SD sequences. In φIN93, however, most of the ORFs are translated with SD sequence. The only three ORFs (ORF5, ORF11 and ORF28) seem to be translated without SD sequence. In addition, these ORFs are included in transcription Unit 1 or 2 that is early/middle transcripts different from the case of φYS40. Now, we have no definite information on the mechanism of translation without SD sequence in φIN93, so further work is required.

ACKNOWLEDGEMENT

We thank Y. Fujiie, M. Kajimoto, Y. Morikawa, S. Oyama, A. Nakata and M. Moriya for technical assistance and discussion. We are grateful to Professor emeritus H. Kawasaki of Oosaka Prefecture University for valuable discussion.

REFERENCES

1. Oshima, T. and Imahori, K. (1974) Description of *Thermus thermophilus* (Yoshida and Oshima) comb. nov., a nonsporulating thermophilic bacterium from a Japanese thermal Spa. *Int. J. Syst. Bacteriol.* **24**, 102–112
2. Matsushita, I., Yamashita, N., and Yokota, A. (1995) Isolation and characterization of bacteriophage induced from a new isolate of *Thermus aquaticus*. *Microbiol. Cult. Collect.* **11**, 133–138
3. Sakaki, Y. and Oshima, T. (1975) Isolation and characterization of a bacteriophage infectious to an extreme thermophile, *Thermus thermophilus* HB8. *J. Virol.* **15**, 1449–1453
4. Hartmann, R.K. and Erdmann, V.A. (1989) *Thermus thermophilus* 16S rRNA is transcribed from an isolated transcription unit. *J. Bacteriol.* **171**, 2933–2941
5. Brock, T.D. and Freeze, H. (1969) *Thermus aquaticus* gen. nor. and sp. nor., a nonspolulating extreme thermophile. *J. Bacteriol.* **98**, 289–297
6. Yamato, K.R., Alberts, B.M., Benzinger, R., Lawwhorne, L., and Treiber, G. (1970) Rapid bacteriophage sedimentation in the presence of polyethylene glycol and its application to large scale virus purification. *Virology* **40**, 734–744
7. Maniatis, T., Fitsch, E.F., and Sambrook, J. (1982) *Molecular Cloning, A laboratory manual*, pp. 249–251, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
8. de Grado, M., Lasa, I., and Berenguer, J. (1998) Characterization of a plasmid replicative origin from an extreme thermophile. *FEMS Microbiol. Lett.* **165**, 51–57
9. Olins, P. O. and Rangwala, S.H. (1989) A novel sequence element derived from bacteriophage T7 mRNA acts as an enhancer of translation of the *lacZ* gene in *Escherichia coli*. *J. Biol. Chem.* **264**, 16973–16976
10. Minakhin, L., Berdygulova, Z., Severinov, K., Goel, M., Florens, L., Washburn, M., Mushegian, A., Ramanculov, E., Glazko, G., Karamychev, Valeri N., Slesarev, Alexei I., Kozyavkin, Sergei A., Khromov, I., Severinov, K., Ackermann, H., Mushegian, A., and Severionv, K. (2008) Genomic comparison and proteomic characterization of *Thermus thermophilus* bacteriophages P23–45 and P74–26: siphoviruses with triplex-forming sequences and the longest known tails. *J. Mol. Biol.* **378**, 468–480
11. Matsushita, I. and Yanase, H. (2008) A novel thermophilic lysozyme from bacteriophage. *Biochem. Biophys. Res. Commun.* **377**, 89–92
12. McShan, W. Michael, Tang, Y.-F., and Ferretti, Joseph J. (1997) Bacteriophage T12 of *Streptococcus pyogenes* integrase into the encoding a serine tRNA. *Mol. Microbiol.* **23**, 719–728

13. Pierson, L. S. III. and Kahn, M. L. (1987) Integration of satellite bacteriophage P4 in *Escherichia coli*. DNA sequences of the phage and host regions involved in site-specific recombination. *J. Mol. Biol.* **196**, 487–496
14. Wolf-Dieter, R., Peter, P., and Siobhan, Y. (1989) Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic Acids Res.* **17**, 1907–1914
15. Horie, N., Hara-Yokoyama, M., Yokoyama, S., Watanabe, K., Kuchino, Y., Nishimura, S., and Miyazawa, T. (1985) Two tRNA Ile1 species from an extreme thermophile, *Thermus thermophilus* HB8: effect of 2-thiolation of ribothymidine on the thermostability of tRNA. *Biochem.* **24**, 5711–5715
16. Faraldo, M.M., de Pedro, M.A., and Berenguer, J. (1992) Sequence of the S-layer gene of *Thermus thermophilus* HB8 and functionality of its promoter in *Escherichia coli*. *J. Bacteriol.* **174**, 7458–7462
17. Maseda, H. and Hoshino, T. (1995) Screening and analysis of DNA fragments that show promoter activities in *Thermus thermophilus*. *FEMS Microbiol. Lett.* **128**, 127–134
18. Naryshkina, T., Liu, J., Florens, L., Swanson, S.K., Pavlov, A.R., Pavlova, N.V., Inman, Kozyavkin. S.A., Washburn, M., Mushegian, A., and Severinov, K. (2006) *Thermus thermophilus* bacteriophage φYS40 genome and proteomic characterization of virions. *J. Mol. Biol.* **364**, 667–677
19. Talkington, C. and Pero, J. (1979) Distinctive nucleotide sequences of promoters recognized by RNA polymerase containing a phage-coded "σ-like" protein. *Proc. Natl Acad. Sci. USA* **76**, 5465–5469
20. Sevostyanova, A., Djordjevic, M., Kuznedelov, K., Naryshkina, T., Gelfand, M.S., Severinov, K., and Minakhin, L. (2007) Temporal regulation of viral transcription during development of *Thermus thermophilus* bacteriophage φYS40. *J. Mol. Biol.* **366**, 420–435