Volume 14 Number 23 1986 Nucleic Acids Research

The region of phage T4 genes 34, 33 and 59: primary structures and organization on the genome

Sabine Hahn, Ulrich Kruse and Wolfgang Rüger+

Arbeitsgruppe Molekulare Genetik, Lehrstuhl Biologie der Mikroorganismen, Ruhr-Universität Bochum, D-4630 Bochum 1, FRG

Received 3 October 1986; Revised and Accepted 11 November 1986

ABSTRACT

The product of gene 33 is essential for the regulation of late transcription and gene product 59 is required in recombination, DNA repair and replication. The exact functions of both proteins are not known. Restriction fragments spanning the genomic area of genes 33 and 59 have been cloned into phage M13 and a 4.9 kb nucleotide sequence has been determined. Translation of the DNA sequence predicted that gp33 contains 112 amino acids with a mol.wt. of 12.816 kd while gp59 is composed of 217 amino acids adding up to a mol.wt. of 25.967 kd. The genomic area studied here also contains 3 open reading frames of genes not identified to date and it is thought to include the NH_2-terminal part of g34. One of the open reading frames seems to code for the 10 kd protein, probably involved in the regulation of transcription of bacteriophage T4. This protein is predicted to consist of 89 amino acid residues with a mol.wt. of 10.376 kd. Gene 33 and the gene for the 10 kd protein were cloned separately on high expression vectors resulting in over-production of the two proteins.

INTRODUCTION

In bacteria, the DNA-dependent RNA polymerase (RNAP, EC 2.7.7.6) is the only enzyme known to synthesize messenger, ribosomal, and transfer RNA. As recently reviewed (1,2) the enzyme consists of a catalytic "core" (α_2, ß, ß') which forms a tight complex. Associated with the "core" is a regulatory subunit, the σ-factor. This polypeptide enables the holoenzyme to interact specifically with the promoter and to initiate transcription. As soon as the nascent RNA chain is elongated, σ is eliminated from the transcription complex and can be recycled. Though the primary structures of all RNAP subunits of the E. coli enzyme are known, their catalytical function during initiation, elongation and termination of RNA chains and the process of transcriptional regulation remain to be further elucidated.

One approach is to study transcription in cells of E. coli infected with bacteriophage T4. The rifampicin sensitivity of all steps in the replication of bacteriophage T4 and labeling experiments at different times in the infection cycle indicate that the host enzyme remains the basic unit performing

viral transcription. However, various phage coded proteins modify the transcription apparatus (for a review, see 3, 4). Thus the α-subunits are ADP-ribosylated, first one by the T4-coded protein gpalt and later both by gpmod. Three proteins, gp33, gp45 and gp55, are required for late transcription *in vivo*. Two other proteins, with mol.wts. of 10 and 15 kd co-purify with the RNAP at different stages of the infection cycle (5, 6). The products of at least four other genes (motA, motB, motC and alc) have also been implicated in transcription control, although they are not found bound to polymerase (7, 8). It is inferred that the regulatory events including the modification of the core enzyme and the apparent association of auxiliary proteins with the core, guide the enzyme to stop transcribing from bacterial promoters and to recognize instead sets of early, middle and late phage promoters in a temporal sequence.

Immediately after infection unmodified *E. coli* RNAP binds to T4 early promoters which by comparison of their "-10 and -35 regions"are closely related to regular *E. coli* promoters. Under the influence of gpmot the enzyme seems to recognize middle mode promoters which differ from T4 early promoters in their "-35 region" as well as in its slightly shorter distance to the "-10 region" (9,10). In T4 late promoters the "-10 region" usually is TATAAATA while common "-35 regions" are not observed (11).

At this point of understanding T4-regulated transcription it would be highly desirable to identify the genetic loci of all gene products involved in T4-directed RNA synthesis: the determination of the nucleotide sequence of the corresponding genes should facilitate their cloning into expression vectors. This would make available gene products which are otherwise difficult to isolate in amounts necessary possibly to reconstitute the T4 transcriptional apparatus. Site-directed mutagenesis (12) as well as protein-protein (13) and protein-DNA crosslinking experiments (14, 15) then could contribute to the elucidation of the arrangement and the function of the T4-coded subunits in the process of transcription.

Recently we sequenced T4 gene 55 and overexpressed its gene product (16). We now focussed our attention on gene 33 and its immediate environment: often functionally related proteins are clustered on the same transcript.

In the near vicinity of gene 33 there are two more regions of general interest. Firstly, a region which separates areas of early transcription from areas of late transcription. Secondly, this genomic region also harbours genes 59 and "das". Gp59 was reported to act in DNA repair (UV and alkylation damage;17), in recombination, and in the organization of the DNA-membrane complex (18),

while the "das" mutation partially suppresses the effects of g46⁻ and g47⁻ mutations on host DNA breakdown (19,20).
This paper deals with the nucleotide sequence and the organization of genes 33, 59 and of four open reading frames on the genome of bacteriophage T4 with respect to the position of possible promoter and termination sites. One of the orfs (orfB) is strongly suspected to represent the gene coding for the 10 kd protein which is involved in the regulation of T4 transcription, possibly acting as an antagonist to the host σ-factor (21). Gene 33 and orfB were cloned on high expression vectors, both gene products can be over-expressed.

MATERIALS AND METHODS

The materials and methods used in this study were essentially the same as described previously (16,22) with the following exceptions:

Phages and Plasmids: Amber mutants used in the marker rescue tests came from Dr.A.H.Doermann, Seattle. The vector for cloning a T4 early promoter was pLBU3 which was supplied by Dr.H.Bujard, Heidelberg (23). The expression vector pPLc2833 was kindly donated by Dr.E.Remaut, Ghent (24). Plasmid pT7-3 and E. coli K38 carrying plasmid gGP1-2 were a gift of Dr.S.Tabor, Harvard (25).

DNA Sequencing: The nucleotide sequence of gene 33 inserted into the expression vector pPLc 2833 was analyzed following the plasmid sequencing technique (26,27). The DNA primer corresponding to the sequence between positions 3312 and 3332 of the nucleotide sequence was composed on an Applied Biosystems DNA Synthesizer, Model 381A, according to Sinha et al.(28).

RESULTS

Nucleotide Sequencing and the Primary Structure of the Region

Genes das, 33 and 59 are located on a 6.95 kb SalI restriction fragment (Fig.1A) lying between 143.570 and 150.522 kb on the genomic map of bacteriophage T4 (29,30). After cleavage of T4 dC-DNA with restriction endonuclease SalI and electrophoresis of the digest on preparative agarose gels, this DNA fragment (SalI-8) can be recovered in pure form. Further digestion with restriction endonuclease EcoRI generates a 3.0 kb EcoRI fragment (EcoRI-14, 147.500-150.500 kb) and two EcoRI-SalI termini, one of 3.93 kb length (143.570-147.500 kb), the other one so small as to run off the gel upon electrophoresis on agarose. Fragment EcoRI-14 and the large EcoRI-SalI terminus were again recovered from agarose gels.
First, fragment EcoRI-14 which should carry the "switch" region between early and late transcription, was cloned and amplified in pLBU3. This plasmid was

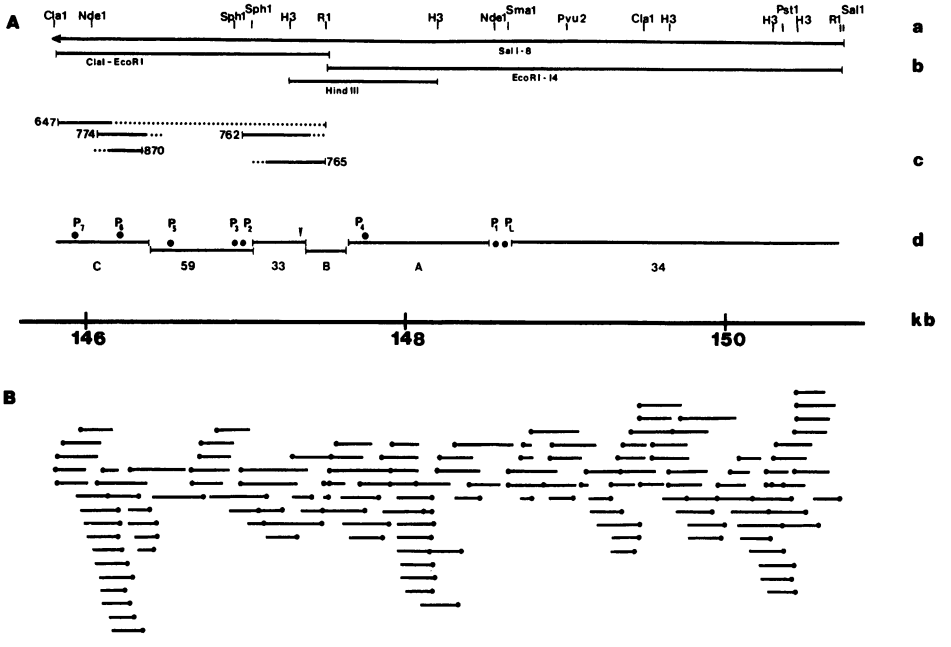

Fig.1: The Position of Genes and Open Reading Frames on the Physical Map of Bacteriophage T4 and the Sequencing Strategy A. The schematic outline represents the genomic region between 143.570 and 150.522 kb on the physical map of bacteriophage T4 (29,30). Line a: positions of the restriction enzyme cleavage sites. H3 stands for HindIII and RI for EcoRI. Line b: position of restriction fragments serving as starting material for the sequencing experiments. Line c: positions of subclones used in the marker rescue experiments. Solid lines represent the portions sequenced, dotted lines indicate the direction of extension of each individual clone. Line d: Genes, orfs and the putative promoter● and termination ▼ sites found in the sequence. B. The positions of 135 gel readings from which the total sequence was composed. Points represent the 5'end of each individual sequence, bars and points together represent the approximate extent of each reading.

designed to identify and to tolerate strong viral promoter sites (23): a promoter region incorporated into the polylinker site of this plasmid triggers the transcription downstream through to the gene of the ß-galactosidase α-fragment and is stopped at the fd terminator inserted next to it. Upon transformation of the ligation mix into E. coli DZ 291, which carries the deletion lacZΔM15, deep blue colonies indicate that the DNA incorporated into the plasmid possibly harbours a promoter site. Blue, tet^r colonies were picked and recultured separately. Their plasmids were isolated and the DNA inserts were tes-

ted with respect to the following items: i. molecular weight, ii. hybridization with restriction fragments known to span the genomic region around gene 33 and iii. size and number of sub-fragments generated after digestion with restriction endonucleases KpnI, XhoI and SmaI. The positions of their cleavage sites on the genomic map of T4 and hence the size of the sub-fragments generated can be predicted. One of the clones meeting all requirements, pSB1/17, was selected for sequencing.

After amplification of plasmid pSB1/17, the inserted fragment EcoRI-14 was isolated and digested with restriction endonuclease Sau3AI, RsaI, HindIII or HaeIII. The subfragments obtained were cloned into the appropriate restriction sites of phage M13mp8 (31) and sequenced by chain termination (32).

Second, the large EcoRI-SalI terminus was further cleaved with restriction endonuclease ClaI. This treatment generates two sub-fragments, a SalI-ClaI terminus (2.28 kb, 143.570-145.850 kb) and an EcoRI-ClaI terminus (1.65 kb, 145.850-147.500 kb). Both fragments were again separated by electrophoresis on agarose gels. While the SalI-ClaI terminus was discarded, the EcoRI-ClaI terminus which should harbour gene 33 and possibly gene 59, was isolated from the agarose. The DNA was sonicated (16,33), treated with nuclease SI to generate blunt ends, sub-cloned in M13mp8 and sequenced at random.

The nucleotide sequence at the internal EcoRI site was verified by analysis of a HindIII fragment which was known from hybridization experiments to overlap the region. The HindIII fragment was obtained from a digest of restriction fragment SalI-8. It was amplified in M13mp8, and sequenced as described above for the EcoRI-ClaI terminus.

Fig.1 indicates the positions of 135 clones analyzed to obtain the entire sequence from both strands (B) as well as the positions of the restriction fragments, the genes, the unidentified open reading frames, the promoters and the possible termination site within this 4.9 kb genomic region. The complete nucleotide sequence is available from the authors upon request.

Open Reading Frames, Transcription and Translation Signals

The sequence was screened for open reading frames and probable promoter and termination sites. Six open reading frames (orf) were identified (Table 1). In all six the codon usage is T4 specific, i.e. certain codons that are rarely used in strongly expressed genes of E. coli (34) are used more frequently in the genes sequenced here. Codons terminating with A or T are most abundant. Among these codons are also those specified by tRNAs coded for on the T4 genome (35). The codons CCC and GGG do not appear (data available upon request). Marker rescue tests performed with T4 33^- (amN134) or (amN134,amC18) and T4

Table 1: Genes and Open Reading Frames.

Position	Leader sequence	Designation	% Amino acids charged +	% Amino acids charged -	Polarity index	% Alpha helicity	% Beta sheet	Mol.wt. of the product (daltons)
2084-1	ACAATAGGAGCCCGGGAGAATGGCC	Gene 34 (partial)						
2189-3103	ACTAATTGAACGAGGTTCATATGGA	OrfA	17.4	14.1	50	54.8	11.8	35 517
3115-3381	ATGAATTTTGAGGTGAATAATGGCT	OrfB	20.2	23.6	53	83.1	1.1	10 367
3362-3697	ACTGAAGAGGTAGTTGAACTTTATG	Gene 33	11.6	20.5	54	78.6	2.7	12 816
3697-4346	TAATACATTAGATTTTCTACTATGA	Gene 59	18.0	12.0	48	58.1	12.9	25 967
4352--	GAAATCTTGCAAGTATTGATATGAA	OrfC (complete)	13.8	12.9	48	41.0	14.8	24 509
16S RNA	3'AUUCCUCCACAUG							

We show the positions of genes 33, 59 and of three orfs, A-C, which were identified within the sequence outlined here. Twenty five bases of the mRNA identical strand comprising the possible ribosome binding sites (49) and the translation starts, both underlined, are presented for each gene. The mol.wts. for the gene products, the percentages of positively and negatively charged amino acids, the polarity indices (61), and the percentage of α-helicity (62) are indicated. The data for orfC were completed, taking advantage of the sequence worked out by Krisch and Allet (47). The data for the tentative gp34 are omitted since the nucleotide sequence of the gene is still incomplete.

Table 2: Putative Promoter and Termination Sites.

	-35		-10
E.coli promoters (consensus)	TTGACA	(15-21 bp)	TATAAT
P1 (position 2150)	TTTACT	17 bp	TATAAT
P2 (position 3750)	GATACA	17 bp	TATAAA
P3 (position 3805)	TGGAAA	16 bp	TATAAT
T4 middle mode promoters (consensus)	AATGCTT	(12-17 bp)	TATAAT
P3 (position 3805)	AATGGAA	17 bp	TATAAT
P4 (position 3000)	ATTGCTT	17 bp	TATAAT
P5 (position 4200)	CTTGCTT	17 bp	TATAAT
P6 (position 4540)	AATGCTT	14 bp	TAAAAT
P7 (position 4820)	ATTGCTT	15 bp	TATTAT
T4 late promoters (consensus)			TATAAATATCTATT
PL (position 2120)			TATAAATACTTATT
P2 (position 3760)			TATAAATTATACTT
Termination site			
Position 3400	CCGGTCGATGAGACCGG		-12.7 Kcal/mol

Shown are the putative promoter and termination sites. The consensus sequence for the E. coli promoter is as given by Rosenberg and Court and by Hawley and McClure (37,38), and that for the T4 middle mode promoter was taken from Christensen and Young (10). The free energy content of the pin and loop structure was calculated using the computer program designed by Zuker and Stiegler (44). Inverted repeats are underscored. For further details see text.

59^- (amHL628) indicated the positions of genes 33 and 59 within the sequence. The ratios of wild-type progeny to total progeny differed by 4 (g33) and 2 (g59) orders of magnitude as compared to the corresponding values obtained with the controls (16). Clone 870 did not rescue g59 (Fig.1A).

In addition to the open reading frames two promoter sites were identified: one early (P1) and one late promoter (PL), lie in close vicinity at about 148.3 kb on the map. This is one of the regions on the T4 genome separating genes transcribed early in the infection cycle from those being transcribed late (36). The -10 region of the early promoter is TATAAT, the -35 region is TTTACT with a distance of 17 bp in between. The sequence of the "-10" region of the late promoter is TATAAATACTTAT. Both -10 regions are thus in agreement with the corresponding consensus sequences (37,38,11,39) while the -35 region of the early promoter shows 2 base pairs changed (Table 2).

On the grounds of their base sequence at least 6 more regions might represent additional promoter sites (P2-P7; Table 2). They are situated within open reading frames. The possible importance of these sites for the regulation of the gene expression will be discussed.

The 2100 bp open reading frame identified on the r-strand of the sequence, downstream from the late promoter, probably corresponds to gene 34 coding for the large tail fiber protein. This polypeptide was reported to have a molecular weight of 140-150 kd (40,41) corresponding to 1200-1300 amino acid residues. The sequence analyzed here codes for 694 residues with a mol.wt. of 74.521 kd representing about 50% of the gene product. Marker rescue experiments performed with T4 34^-(am N58) and with T4 34^-(am A455) were negative, but the mol.wts. of the two amber fragments resulting from these mutations were reported to be 103 and 129 kd, respectively (42,43). Thus it can be assumed that both amber mutations tested are situated beyond the region sequenced here.

OrfB has a 21 bp overlap with gene 33. At the beginning of g33, downstream from the TAG stop codon of orfB there exists a palindromic sequence within the reading frame of gene 33. The stem and loop structure which can be formed has a ΔG value of -12,7 kcal/mol (44). The stem consists of 1 A-T and 5 G-C pairs, the loop is formed by 5 bases. This structure might represent a factor-dependent termination signal. The importance of this signal for the expression of gene 33 will be discussed.

Characterization of the Putative Gene Products

The amino acid sequence of a gene product allows to deduce physicochemical properties as the molecular weight, the overall charge and the pola-

rity index. Secondary structure prediction (45) and hydropathy calculation (46) further characterize a protein.

Gp 33 with 112 amino acid residues has a molecular weight of 12.816 kd. The protein is negatively charged due to a high content of glutamic acid, it contains no His or Trp residues, and exhibits a high percentage of α-helicity. Hydropathy calculations suggest that it is a soluble protein.

In the case of the protein coded for by orfB, 89 amino acid residues contribute to a mol.wt. of 10.367 kd with a slightly negative net charge. Secondary structure predictions indicate that the α-helicity exceeds that of gp33. The protein contains no Pro, Cys or Trp residues.

Gp 59, consisting of 217 amino acid residues, shows a mol.wt. of 25.967 kd and a positive charge. This protein has the only interesting hydropathy profile: at amino acid positions 71-90 and 146-170 the protein exhibits two hydrophobic domains. The hydropathies of the segments average +1.1 and +1.3, respectively, leading to the idea that the protein might possibly bind to the membrane.

GporfA has a mol. wt. of 35.517 kd and is composed of 305 amino acid residues. Its net charge is positive.

GporfC was analyzed by complementing our sequence with the one of Krisch and Allet (47). Both sequences have an overlap of 194 bases. The protein is composed of 210 amino acid residues adding up to a mol.wt. of 24.509 kd. The polarity index is 48, the hydropathy plot shows rather regularly alternating hydrophobic and hydrophilic domains.

Cloning Gene 33 into a High Expression Vector

The expression vector pPLc2833, a derivative of pBR322, was designed to overexpress proteins under the control of the P_L promoter (24). Gene 33 was isolated from M13 SB762 as a 504 bp EcoRI-BamHI fragment and cloned into the corresponding restriction sites of vector pPLc2833. (Note that the BamHI site originates from M13mp8, it does not appear in our sequence). This procedure forces the T4 fragment into the expression vector in the proper orientation. Since in pPLc2833 there is no system selective for DNA inserted at the multilinker site, a number of amp^r colonies were isolated and selected for DNA inserts of the appropriate size. The DNA inserts were again cleaved out and characterized with respect to internal restriction sites. In spite of the positive response of all parameters tested, cells carrying plasmid pPLcSB33 did not visibly overproduce gp33 upon induction and electrophoresis of cell extracts on polyacrylamide gels stained with Coomassie blue. This result was surprising since, in earlier experiments with vector pPLc2833, cloning of T4

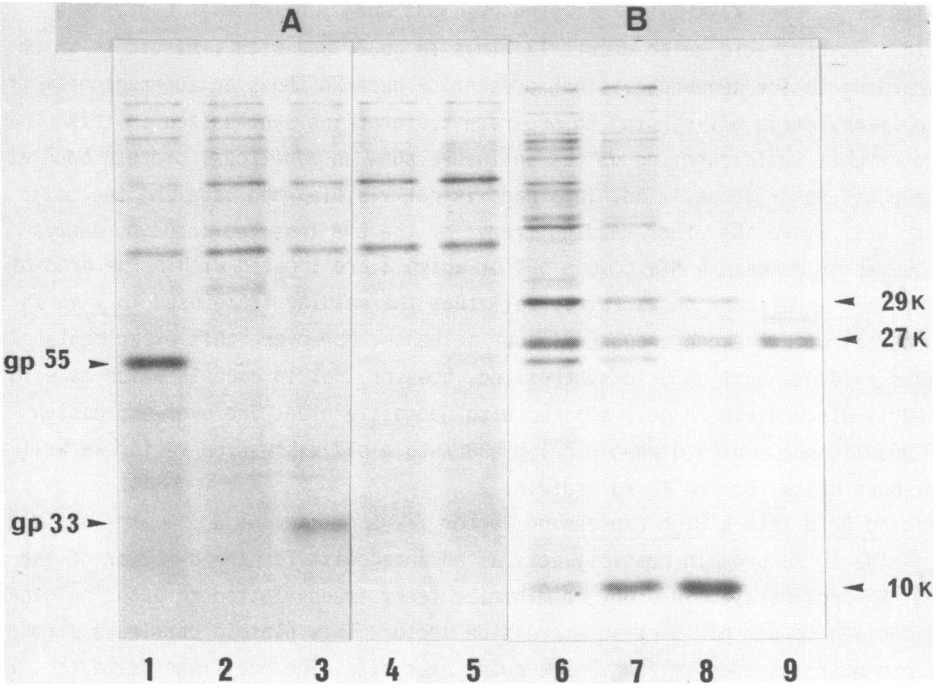

Fig.2: Over-Expression of Gene Products. A. Expression of gp33 in vector pPLc2833 (24). Channel 1: Control, pPLcHG55, induced at 42°C (16), channel 2: control pPLc 2833, induced at 42°C, channel 3: pPLcSB33, induced at 42°C, channel 4: pPLcSB33, induced at 42°C, channel 5: pPLc2833, induced at 42°C. Channels 1-3 labeled with ^{14}C-glu and channels 4 and 5 labeled with ^{14}C-his. B. Expression of the 10 kd protein in expression vector pT7-3 (25). Channel 6: control, 30°C, channel 7: induced at 42°C, channels 8 and 9: induced at 42°C in the presence of rifampicin, final concentration 200ug/ml. Channels 6-8 labeled with ^{14}C-glu, channel 9 labeled with ^{14}C-pro. The gel concentrations were 12.5% and 15%, respectively (74). For further explanation see text.

genes 55, ßgt and αgt (16,22) resulted in a vast over-production of the proteins. We therefore constructed a 20 bp DNA primer, corresponding to positions 3312-3332 of the sequence and analyzed the insert of pPLc2833 with the aid of the double strand sequencing technique. This experiment proved that the DNA insert contained gene 33 and its putative ribosome binding site in proper orientation. The nucleotide sequence of g33 remained unchanged in the cloning procedure. Complementation tests showed that gp33 was biologically active. At 42°C growth rate and colony survival of cells carrying gene 33 are not reduced as compared to the controls without a g33 insert, indicating that the presence of gp33 does not markedly affect the survival of cells carrying the plasmid (data not shown).

Cells of E. coli K12ΔHIΔtrp carrying pPLcSB33 were labeled with L-glutamic acid, an amino acid which occurs 17 times in gp33, and with L-histidine which, according to the sequence, is not present. Figure 2A shows an autoradiogram of a polyacrylamide gel serving to separate the proteins synthesized in this experiment. Cells carrying g33 as an insert show an additional protein band at the position of about 12 kd. This band is not visible when labeling the cells with His. Since the other reading frames of the DNA inserted into the expression vector contain 6 His codons out of which 3 are located within the area of g33, this experiment shows fairly well that the reading frame used in vivo is identical with the one proposed in our sequence. Moreover, this experiment gives evidence that gp33 is synthesized, however, not in amounts which make it clearly discernible on gels stained with Coomassie blue. The over-expression of gp33 is not improved when cloning g33 into a pT7 expression vector as will be shown below for the 10 kd protein.

Cloning OrfB into a High Expression Vector

The 10 kd protein possibly acts as an antagonist for the σ-factor of the host RNA polymerase. In order to minimize leaky transcription through the gene we decided to use pT7-3 as an expression vector. This plasmid carries a strong T7 promoter inserted in front of a polylinker site. The polylinker site is followed by the gene for the ß-lactamase. Transcription of this gene serves as an internal control for the functioning of the system and at the same time over-expressed ß-lactamase is a convenient mol.wt. marker (29 and 27 kd) for the identfication of the gene product synthesized under the control of the T7 promoter. T7 RNA polymerase is supplied by a second plasmid pGP1-2 harbouring T7 g1 under the control of the heat inducible P_L promoter of bacteriophage λ. This plasmid carries a kanamycin resistance marker.

OrfB was excised from the T4-M13 hybrid phage SB893 as a coherent HindIII fragment and cloned into the HindIII site of pT7-3. The vector was then transformed into E. coli K38 already carrying pGP1-2 (25). The cells were grown under the selective pressure of ampicillin and kanamycin. Out of 100 colonies tested (48), 50% carried pT7-3 with a HindIII insert. Out of these about 50% had the insert in the orientation as to over-express the 10 kd protein. Figure 4B shows that the over-expressed protein is labeled with ^{14}C-glutamic acid but not with ^{14}C-proline (5).

There appear 13 proline codons in the other reading frames of the DNA insert out of which 4 are situated in the area of orfB. In contrast to the results obtained when over-expressing gp33 the 10 kd protein becomes clearly visible when staining the gel with Coomassie blue.

DISCUSSION

We present a 4910 bp nucleotide sequence which includes T4 genes 33, 59, possibly part of gene 34 and the region dividing genes being transcribed early in the infection cycle from those being transcribed late. The DNA section in vestigated harbours six orfs as determined according to the following criteria: i. an uninterrupted reading frame of more than 50 amino acid residues deduced from the nucleotide sequence among the 6 possible reading frames, ii. the beginning of the hypothetical orf with the start codons ATG or GTG, iii. the appearance of a Shine-Dalgarno sequence 5-9 bp upstream of an open reading frame, complementary to the sequence of the E. coli 16S RNA (49,50) and, iv. the T4-directed codon usage as specified by the high A-T content of its DNA and by the utilization of eight tRNAs coded for on the phage genome (34,35).

Two out of the six orfs were identified by marker rescue experiments as genes 33 and 59. A third (partial) orf in this sequence most probably represents the NH_2-terminal part of gene 34, coding for the large tail fiber protein. A dimer of gp34 forms the proximal half fiber (for an overview see 51). The identity of the orf and g34 is inferred from the size of the gene, its position on the physical map of phage T4 and the position of the gene with respect to the late promoter PL. Moreover, a segment of 25 residues appears within the amino acid

Table 3: Repetitive Amino Acid Sequence in Gene Product 34.

Position in nucleotide sequence	Amino acid sequence
686-612	AITPETLANRTATETRRGIARIATT
566-492	IITPKKLNERTATETRRGVAEIATQ
458-384	IITPKKLQARQGSESLSGIVTFVST
311-237	VVSPKALDQYKATPTQQGAVILAVE
191-117	VVTPETLHKKTSTDGRIGLIEIATQ
083-009	AVTPKTLNDRRATESLSGIAEIATQ
	*** * * * * *

Identical and homologous amino acid sequences are underscored. Homologies were defined as exchanges of amino acid residues within the groups (I,L,V,M), (F,Y,W), (K,R), (S,T), and (D,E). Some positions (*) are highly conserved or allow an exchange between a positive (K) and a negative charge (E).

sequence of this part of the gene which is, with modifications, repeated 6 times at regular distances. Within these repeats some amino acid positions are highly conserved (Table 3). Hydropathy calculations (46) show a proximal hydrophylic and a distal hydrophobic component within the repeated sequences. On the other hand, neither the distribution of the electrical charges nor secondary structure predictions (45) result in a common pattern. When sequencing this area of the genome from DNA of another T4 mutant strain (T4,SaΔ9; 52) we found differences in the nucleotide sequence at two positions. They lead to an exchange of one amino acid residue each. At position 934 Pro is replaced by Leu (CCA-►CTA) and at position 889 Asp is exchanged against Gly (GAT-►GGT). With respect to the dimerization of gp34 it also should be mentioned that there appear no Cys residues in this part of gp34.

The only gene mapped in genetic analyses between g34 and g33 is gdas (36,53). The nucleotide sequence reveals two open reading frames in this area: orfA and orfB. Since orfB is likely to represent the hitherto unidentified gene of the 10 kd protein which is involved in T4 late transcription, orfA might probably code for gpdas.

The identity of orfB with the gene coding for the 10 kd protein is inferred from the position of the gene in one transcription unit with gene 33 and the absence of proline in the primary structure of its gp (5). A negative net charge, a high proportion of α-helicity as well as similarities in the amino acid sequences with proteins like some σ-factors of bacterial RNA polymerases (Table 4A) further support this idea. The codon CGA (Arg) which might mutate by transition to TGA (and hence be detected only when using an "opal" suppressor strain) is the only codon in orfB which could be changed in one step into a termination codon. This may be a reason why this gene was not mapped with the aid of conditional lethal T4 mutants; in vitro mutagenesis will help to determine if it is essential.

The genetic information coded for on the DNA stretch sequenced is closely packed leaving only short non-coding sequences between the genes. The sequence TATAAATA specifying a late promoter and the -35 region of the neighbouring early promoter are separated by only 4 base pairs. The reading frames of orfA and of orfB are separated by 11 bps, orfB and gene 33 overlap by 21 bps, gene 33 and gene 59 are linked by an ATGA element. As pointed out earlier, the initiation codon ATG and the termination codon TGA of the preceding orf overlap in this sequence which might lead to an effective translation of this messenger (54). G59 and orfC are separated by only 4 bps.

Two more features within this sequence are noteworthy. First, there is a

Table 4: Similarities among Different Proteins.

		1.	2.
A	E. coli HtpR (66)	(48) EAAKTLILSHLRFVVH-IARNYA	EAAKTLILSHLRFVVH-IARNYA
	B. subt. spoIIG (67)	(58) AARAILIERNLRLVVY-IARKFE	AARAILIERNLRLVVY-IARKFE
	B. subt. rpoD (68)	(132) ESKRRLAEANLRLVVS-IAKRYV	ESKRRLAEANLRLVIS-IAKRYV
	E. coli rpoD (69)	(374) RAKKEMVEANLRLVIS-IAKKYT	RAKKEMVEANLRLVIS-IAKKYT
	T. acidoph. HTa (70)	(2) VGIS--ELS--KE----VAKKAN	VGIS--ELS--KE----VAKKAN
	SP01 gp28 (71)	(131) EGL-EK-LKETLESD--IAKSLL	EGL-EK-LKETLESD--IAKSLL
	T4 gp55 (16)	(101) ACFNAFVQRIKKERKE-VAKKYS	ACFNAFVQRIKKERKE-VAKKYS
	T4 gp33	(15) TGLSEKELSIKKEKDE-IAK-LL	TGLSEKELSIKKEKDE-IAK-LL
	T4 10 kd	(1) MAKKEMVEFDEAIHGEDLAKFIK	MAKKEMVEFDEAIHGEDLAKFIK
B	Gp36k	(123) VMDIDKYEADDHIAVLVKKFSLEGHKILIISSDGDFTQL +++++ **** *++*+++ + +*+ +** ++* *++**	
	DNA polI (72)	(106) LLAVSGVEADDVIGTLAREAEKAGRPVLISTGDKDMAQL	
	Gp59	(131) KHDEQTDNLVWNNYSIKLKAYRKILNI ** +** ++ ++*+ +** ++* *+*	
	Gp28 (71)	(191) KHIDQTLGISNKQYDSELKKFVKRLTI	
	GporfC	(182) DTTAKSMVCYFNSGWIPLEDPEYCELCQL +*+**+ + *++*+++*** + **+	
	GpssbPf3 (73)	(14) GTSAKGNPYTFQEGFLHLEDKPFPLQCQF	

A. Similarities found between several σ-like factors known from the literature and from this study. Column 1: comparison of E. coli rpoD with several σ-factors and the DNA binding protein of T. acidophilum, column 2: comparison of T4 gp33 with the same proteins. Identical and homologous amino acids are underscored. No similarities were detected with SP01 gp33 and gp34 (63) and with 29 gp4 (64). B. Similarities found between gporfA, gp59, and gporfC, and several proteins known from the literature. Amino acid identities are marked *, homologous exchanges, as defined by the mutation data matrix, are marked + (59). Numbers written in parentheses refer to literature and to amino acid positions, respectively. For further details see text. When this manuscript had been completed Gribskov and Burgess (65) showed that the - factors of E. coli, B. subtilis, phage SP01, and phage T4 are homologous proteins.

palindromic sequence found at the beginning of gene 33, giving rise to a stem and loop structure. This structure might possibly represent a factor-dependent termination signal halting transcription at the end of orfB; gene 33 mRNA then would only be available in a large scale if an antitermination mechanism operated. It might be this signal which prevents an effective over-expression of gp33 in the two vector systems used in this study. Moreover, the spacing

between the ATG start codon and the putative Shine-Dalgarno sequence of g33 is extended to 11 bp which might adversely affect translation (50).
Second, at the beginning of gene 59 there are two regions in tandem which might represent additional early or one middle and one late promoter (P2 and P3 in Fig.1 and Table 2). These sites are situated within the reading frame of g59 and either one is followed at a distance of several bases by ATG or GTG start codons in register with orf59, but no good ribosome binding sites are apparent. Since the putative promoters are positioned at the beginning of the gene, it might be that transcripts possibly initiated at these signals would lead to proteins 'modified at the level of transcription'. A similar arrangement was already found within the open reading frame preceding gene αgt (16) and mRNA molecules missing ribosome binding sites are known. It was argued that messengers lacking these sites would be translated less frequently (55). Alternatively, transcription from promoters situated within an orf might aim at the expression of more distal genes. This seems probable in case of P4 and P5 which were detected at the ends of orfA and orfg59, respectively, and in case of P6 and P7 at the end of orfC (Fig.1 and Table 2). The -35 regions of these promoter structures share an identity of at least 5 bases with the consensus sequence for middle mode promoters, AATGCTT (9,10). During middle mode transcription promoter P4 should either serve orfB through to g32 or, if termination was effective, only orfB. OrfC and g32 should be transcribed from the putative promoter P5. P6 as well as P7 could serve g32 exclusively.
Krisch and Allet (47) located a structure typical for rho-independent termination signals near the end of g32. Pulitzer et al. showed that the transcription of g32 is entirely dependent on gpmot (7). Therefore, we suggest that transcription of orfg32 is not started at the early promoter P1 but at the putative middle mode promoters found in g59 and/or orfC. It should be recalled that transcription of gprIIB is started from a middle mode promoter located within orfrIIA (56).
In-vitro transcription performed with E. coli RNAP and T4w DNA (57) showed that the transcript synthesized most abundantly in these experiments could be initiated by dinucleotide elongation with UpC followed in sequence by pU and pC. The length of this transcript was roughly estimated to be 4.2 kb and hybridization experiments revealed that it was complementary to the genomic region between map positions 148.6 and 144.6 (58). Comparison with the sequence presented here makes it probable that this transcript is started at promoter P1. The sequence TCTC appears at position -1. The distance from here to the termination signal at the end of g32 is about 3820 bps.

The amino acid sequences coded for by the open reading frames were compared with the database of the Protein Identification Resource (59,60). The similarities found with a number of eucaryotic and procaryotic proteins were not clearly above statistical margins, they are difficult to interpret and in no case it was possible to assign functions to the proteins. This would have been particularly interesting for orfA, B and C. Therefore, we restrict ourselves to point out arbitrarily three similarities which might lead to further experimentation (Table 4B).

The organization of genes and promoter sites as deduced from the nucleotide sequence presented here raises a number of interesting questions. Thus it would be highly desirable to know which are the exact functions of the gene products coded for by the open reading frames identified in this study. Is the 10 kd protein or gp33 involved in antitermination and, why is there an apparently strong early promoter controlling a gene cluster harbouring genes which seem to be required predominantly in the replicative phase of the infection cycle? We also might learn about the regulation of T4 transcription if we could find an answer to the questions as to whether the presence of the A-T-rich region of a late promoter in close vicinity of the -35 region of an early promoter enhances the activity of the latter. The modification of RNAP in the course of the infection cycle is known to be an important element in T4 gene expression. The arrangement of early, middle and late promoters within a transcription unit and mutual influences might additionally contribute to the regulation of T4 transcription.

ACKNOWLEDGMENTS

We wish to thank all persons who provided us with bacteria and phage strains. We are grateful to Dr.Staden for allowing us to use his sequence analysis program, to Dr.E.M.Kutter for discussion of parts of this manuscript and to Drs. E.P.Geiduschek, E.M.Kutter, and K.P.Williams for sending us their manuscripts before publication. Our thanks are also due to H.D.Liebig for sequencing the gene 33 insert in pPLcSB33 and to Dr.U.Kück for synthesizing the 20 bp oligonucleotide used as a primer for the sequence analysis. W.R. thanks Mrs.U.Aschke for expert technical assistance. The financial support of this study by the Deutsche Forschungsgemeinschaft is gratefully acknowledged.

Additional data i.e. the hydropathy profiles of the proteins and the computed secondary structure predictions are available from the authors upon request.

Abbreviations: bp = base pair, g = gene, gp = gene product, kb = kilobase pairs, kd = kilodaltons, m.o.i. = multiplicity of infection, orf = open reading frame, RNAP = RNA polymerase.

+ To whom correspondence should be addressed

REFERENCES

1. Chamberlin,M.J. (1982) The Enzymes, 15, 61-86.
2. von Hippel,P.H., Bear,D.G., Morgan,W.D., and McSwiggen,J.A (1984) Ann. Rev. Biochem. 53, 389-446.
3. Rabussay,D. (1982) In "Molecular Aspects of Cellular Regulation" Cohen,P. and van Heynigen,S. (eds) pp 219-331, Elsevier/North Holland Biomedical, N.Y.
4. Rabussay,D. (1983) In "Bacteriophage T4", Mathews,C.K., Kutter,E.M., Mosig,G., and Berget,P.B. (eds) pp 167-173, ASM Publications.
5. Stevens,A. (1972) Proc. Natl. Acad. Sci. USA 69, 603-607.
6. Malik,S., Dimitrov,M., and Goldfarb,A. (1985) J. Mol. Biol. 185, 83-91.
7. Pulitzer,J.F., Colombo,M., and Ciaramella,M. (1985) J. Mol. Biol. 182, 249-263.
8. Drivdahl,R.H. and Kutter,E.M. (1986) manuscript in preparation.
9. Brody,E., Rabussay,D., and Hall,D.H. (1983) In "Bacteriophage T4", Mathews,C.K., Kutter,E.M., Mosig,G., and Berget,P.B. (eds) pp 174-183, ASM xublications.
10. Christensen,A.C. and Young,E.T. (1983) In "Bacteriophage T4", Mathews,C.K., Kutter,E.M., Mosig,G., and Berget,P.B. (eds) pp 184-188, ASM Publications.
11. Elliott,T. and Geiduschek,E.P. (1984) Cell, 36, 211-219.
12. Shortle,D., DiMaio,D., and Nathans,D. (1981) Ann. Rev. Genet. 15, 265-94.
13. Han,K.K., Richard,C., and Delacourte,A. (1984) Int. J. Biochem. 16, 129-145.
14. Harrison,C.A., Turner,D.H., and Hinkle,D.C. (1982) Nucleic Acid Res. 10, 2399-2414.
15. Hockensmith,J.W., Kubasek,W.L., Vorachek,W.R., and von Hippel,P.H. (1986) J. Biol. Chem. 261, 3512-3518.
16. Gram,H. and Rüger,W. (1985) EMBO J. 4, 257-264.
17. Wu,R., Wu,J.-L., and Yeh,Y.-C. (1975) J. Virol. 16, 5-16.
18. Shah,D.B. (1976) J. Virol. 17, 175-182.
19. Hercules,K. and Wiberg,J.S. (1971) J. Virol. 8, 603-6012.
20. Mickelson,C. and Wiberg,J.S. (1981) J. Virol. 40, 65-77.
21. Stevens,A. and Rhoton,J.C. (1975) Biochemistry 14, 5074-5079.
22. Tomaschewski,J., Gram,H., Crabb,J.W., and Rüger,W. (1985) Nucleic Acids Res. 13, 7551-7568.
23. Gentz,R., Langner,A., Chang,A.C., Cohen,S.N., and Bujard,H. (1981) Proc. Natl. Acad. Sci. USA 78, 4936-4940.
24. Remaut,E., Stanssens,P., and Fiers,W. (1981) Gene 15, 81-93.
25. Tabor,S. and Richardson,C.C. (1985) Proc. Natl. Acad. Sci. USA 82, 1074-1078.
26. Chen,E. and Seeburg,P.H. (1985) DNA 4, 165-170.
27. Hattori,M. and Sakaki,Y. (1986) Anal. Biochem. 152, 232-238.
28. Sinha,N.D., Biernat,J., McManus,J., and Köster,H. (1984) Nucleic Acids Res. 12, 4539-4557.
29. Kutter,E.M. and Rüger,W. (1983) In "Bacteriophage T4", Mathews,C.K., Kutter,E.M., Mosig,G., and Berget,P.B. (eds) pp 277-290, ASM Publications.
30. Rüger,W. and Kutter,E. (1984) In "Genetic Maps/1984" Vol 3, O'Brian,S.J. (ed) Laboratory of Viral Carcinogenesis, Natl. Cancer Institute, pp 28-34, Cold Spring Harbor.
31. Messing,J.,Crea,R., and Seeburg,P.H.(1981) Nucleic Acids Res. 9, 309-321.
32. Sanger,F., Nicklen,S., and Coulson,A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
33. Rand,K.N. and Gait,M.J. (1984) EMBO J. 3, 397-402.
34. Grosjean,H. and Fiers,W. (1982) Gene 18, 199-209.
35. Fukada,K. and Abelson,J. (1980) J. Mol. Biol. 139, 377-391.

36. Wood,W.B. and Revel,H.R.(1976) Bacteriol. Rev., 40, 847-868.
37. Rosenberg,M. and Court,R. (1979) Annu. Rev. Genet. 13, 319-353.
38. Hawley,D.K. and McClure,W.R. (1983) Nucleic Acids Res. 11, 2237-2255.
39. Christensen,A.C. and Young,E.T. (1982) Nature 299, 369-371.
40. Ward,S. and Dickson,R.C. (1971) J. Mol. Biol. 62, 479-492.
41. Vanderslice,R.W. and Yegian,C.D. (1974) Virology 60, 265-275.
42. Beckendorf,S.K. and Wilson,J.H. (1972) Virology 50, 315-321.
43. Revel,H.R. (1981) Mol. Gen. Genet., 182, 445-455.
44. Zuker,M. and Stiegler,P, (1981) Nucleic Acids Res. 9, 133-148.
45. Rawlings,N., Ashman,K., and Wittmann-Liebold,B. (1983) Int. J. Pept. Prot. Res. 22, 515-524.
46. Kyte,J. and Doolittle,R.F. (1982) J. Mol. Biol. 157, 105-132.
47. Krisch,H.M. and Allet,B. (1982) Proc. Natl. Acad. Sci. USA 49, 4937-4941.
48. Birnboim,H.C. and Doly,J. (1979) Nucleic Acids Res. 7, 1513-1523.
49. Shine,J. and Dalgarno,L. (1974) Proc. Natl. Acad. Sci. USA 71, 1342-1346.
50. Gold,L., Pribnow,D., Schneider,T., Shinedling,S., Singer,.B.S., and Stormo,G. (1981) Ann. Rev. Microbiol. 35, 365-403.
51. Wood,W.B. and Crowther,R.A. (1983) In "Bacteriophage T4", Mathews,C.K., Kutter,E., Mosig,G., and Berget,P.B. (eds) pp 259-269, ASM Publications.
52. Mattson,T., van Houwe,G., Bolle,A., Selzer,G., and Epstein,R. (1977) Mol. Gen. Genet. 154, 319-326.
53. Hercules,C. and Wiberg,J. (1971) J. Virol. 8, 603-612.
54. Kröger,M. and Hobom,G. (1982) Gene 20, 25-38.
55. Walz,A., Pirrotta,V., and Ineichen,K. (1976) Nature 262, 665-669.
56. Pribnow,D., Sigurdson,D.C., Gold,L., Singer,B.S., Napoli,C., Brosius,J., Dull,T.J., and Noller,H.F. (1981) J. Mol. Biol. 149, 337-376.
57. Rüger,W. (1978) Eur. J. Biochem. 88, 109-117.
58. Gram,H., Liebig,H.D., Hack,A., Niggemann,E., and Rüger,W. (1984) Mol. Gen. Genet. 194, 232-240.
59. Dayhoff,M.O., Barker,W.C., and Hunt,L.T. (1983) Methods Enzymol. 91, 524-545.
60. Lipman,D.J. and Pearson,W.R. (1985) Science 227, 1435-1441.
61. Capaldi,R.A. and Vanderkooi,G. (1972) Proc. Natl Acad. Sci. USA 69, 930-932.
62. Robson,B. and Suzuki,E. (1976) J. Mol. Biol. 107, 327-356.
63. Costanzo,M., Brzustowicz,L., Hannett,N., and Pero,J. (1984) J. Mol. Biol. 180, 533-547.
64. Escarmis,C. and Salas,M. (1982) Nucleic Acids Res. 10, 5785-5798.
65. Gribskov,M. and Burgess,R.R. (1986) Nucleic Acids Res. 14, 6745-6763.
66. Landick,R., Vaughn,V., Lau,E.T., Van Bogelen,R.A., Erickson,J.W., and Neidhardt,F.C. (1984) Cell 38, 175-182.
67. Stragier,P., Bouvier,J., Bonamy,C., and Szulmajster,J. (1984) Nature 312, 376-378.
68. Gitt,M.A., Wang,L.-F., and Doi,R.H. (1985) J. Biol. Chem. 260, 7178-7185.
69. Burton,Z., Burgess,R.R., Lin,J., Moore,D., Holder,S., and Gross,C.A. (1981) Nucleic Acids Res. 9, 2889-2903.
70. Delange,R.J., Williams,L.C., and Searcy,D.G. (1981) J. Biol. Chem. 256, 905-911.
71. Costanzo,M. and Pero,J. (1983) Proc. Natl. Acad. Sci. USA 80, 1236-1240.
72. Joyce,C.M., Kelley,W.S., and Grindley,N.D.F. (1982) J. Biol. Chem. 257, 1958-1964.
73. Putterman,D.G., Casadevall,A. Boyle,P.D., Yang,H.-L., Frangione,B., and Day,L.A. (1984) Proc. Natl. Acad. Sci. USA 81, 699-703.
74. Laemmli,U.K. (1970) Nature 227, 680-685.