Volume 14 Number 21 1986 Nucleic Acids Research

Nucleotide sequence and analysis of the 58.3 to 65.5-kb early region of bacteriophage T4

Kristoffer Valerie[1,3,4], John Stevens[1], Mark Lynch[1,5], Earl E.Henderson[1,2] and Jon K.de Riel[1]

[1]Fels Research Institute, and [2]Department of Microbiology and Immunology, Temple University School of Medicine, Philadelphia, PA 19140, USA and [3]Department of Biochemistry and Biotechnology, Royal Institute of Technology, S-100 44 Stockholm, Sweden

Received 21 July 1986; Revised and Accepted 30 September 1986

ABSTRACT

The complete 7.2-kb nucleotide sequence from the 58.3 to 65.5-kb early region of bacteriophage T4 has been determined by Maxam and Gilbert sequencing. Computer analysis revealed at least 20 open reading frames (ORFs) within this sequence. All major ORFs are transcribed from the left strand, suggesting that they are expressed early during infection. Among the ORFs, we have identified the ipIII, ipII, denV and tk genes. The ORFs are very tightly spaced, even overlapping in some instances, and when ORF interspacing occurs, promoter-like sequences can be implicated. Several of the sequences preceding the ORFs, in particular those at ipIII, ipII, denV, and orf61.9, can potentially form stable stem-loop structures.

INTRODUCTION

Recently, considerable progress has been made studying the bacteriophage T4 genome at the molecular level due to the development of T4 strains with unmodified DNA suitable for digestion with restriction endonucleases (1). Many T4 genes have since been cloned, sequenced, and their gene products overproduced in Escherichia coli (for review see ref. 2). A majority of the T4 genes studied so far have been essential and non-essential genes with well-characterized phenotypes. Many non-essential genes are involved with metabolic functions, and their gene products are believed to augment T4 infection by providing more abundant substrates for the phage (3). Perhaps two thirds of the non-essential metabolic genes are without known functions, which has hampered the progress of studying these genes and their gene products (3). One way to begin to investigate the non-essential regions of phage T4 is to identify the unknown genes by nucleotide sequencing, clone and express them in E.coli, and study their gene products to try and elucidate their functions.

Several non-essential genes have been mapped to the early region between the rI and e loci on the T4 genome (3). These include the ipII and ipIII genes (4,5), the denV gene (6), the regB gene (7), the vs gene (8), and the tk gene (9). The stI and stIII genes have a less precise association with this region

(3,10). An origin of replication has also been assigned to this part of the T4 genome (11,12).

In our attempts to clone and sequence the denV gene (13,14), we determined the nucleotide sequence between approximately 61 and 65-kb on the T4 map, and in collaboration with Dwight Hall and coworkers in search of the T4 tk gene (15) we sequenced the remaining 58 to 61-kb region. The analysis of the nucleotide DNA sequence from this region reveals a total of 20 tightly spaced open reading frames (ORFs), all transcribed from the early strand (16). Of these ORFs four have been identified, namely the ipII and ipIII genes (17), the denV gene (13,14), and the tk gene (15). A number of putative promoters and stem-loop signals can also be implicated from this analysis.

MATERIALS AND METHODS

Strains, growth conditions and DNA purifications

The bacteriophage T4 strain 56(amE51,dCTPase$^-$)denA(nd28,endoII$^-$)denB (delrIIH23B,endoIV$^-$)alc8 (12) was the source of dC-T4 DNA after growth in E.coli strains K803 (r_K^-,m_K^-,su$^+$,rgl$^-$) and B834 (r_B^-,m_B^-,su$^-$,rgl$^+$,galU$^-$) as described (1) except that L-broth (1% Bacto-tryptone, 0.5% Bacto yeast extract, 0.5% NaCl, 0.1% glucose) was used instead of M9S medium. Lysis was completed by the addition of chloroform, and the phage was concentrated by polyethyleneglycol (PEG)-6000 precipitation as described (18). The phage-PEG precipitate was layered on to a CsCl step-gradient and centrifuged to separate the phage from chromosomal DNA and debris (19). The E.coli strains used for cloning experiments were MM294 (r_K^-,m_K^+) (21) and GM119 (dam$^-$, dcm$^-$) (22), from B. Bachmann at the E.coli Genetic Stock Center, Yale University, New Haven, CT and HB101 (r_B^-,m_B^-, recA$^-$) (20). The plasmids used for cloning experiments were pBR322 (23), and pUN121 (24). The plasmid pKG1.3, containing the T4 tk gene on a 1.45-kb EcoRI-HindIII fragment, will be described elsewhere (15). Transformation of E.coli cells was performed as described (13). E.coli cells were grown in L-broth supplemented with ampicillin (50 µg/ml) and/or tetracycline (10 µg/ml) when propagating plasmids. Cleared lysates of plasmid-containing E.coli cells were produced essentially as described by Godson and Vapnek (25), after the plasmid had been amplified overnight by the addition of 150 µg/ml of chloramphenicol according to Clewell and Helinski (26). The lysates were phenol-extracted before the DNA was ethanol-precipitated and banded on buoyant-density CsCl gradients overnight at 50,000 rpm using a Beckmann VTi65 rotor.

Cloning and mapping of T4 fragments

Restriction endonucleases, T4 DNA ligase, T4 DNA polymerase, E.coli DNA

polymerases, and bacterial alkaline phosphatase were purchased from New England Biolabs, Bethesda Research Laboratories or Boehringer-Mannheim Biochemicals, and were used as recommended by the manufacturers. [α-^{32}P]dATP and [α-^{32}P]dCTP (>3000 Ci/mmole) for nick-translation and sequencing were obtained from Amersham or ICN. DNA digests were fractionated by electrophoresis on agarose or acrylamide gels as described (27,28), and fragments were recovered from gel slices by electrophoresis into dialysis bags in 0.1 X electrophoresis buffer. Colony hybridizations and Southern blots were carried out essentially as described (28-30).

DNA sequencing and sequence analysis

The cloned T4 fragments were sequenced by the method of Maxam and Gilbert (31) with the modifications of Maniatis et al. (30). Fragments were labeled by extending the recessed 3'-end with the use of DNA polymerase I (Klenow fragment) and either [α-^{32}P]dATP or [α-^{32}P]dCTP and appropriate dNTPs. The HindII site at 3414 was end-labeled with T4 DNA polymerase. Primer-extension sequencing was used to confirm the sequence at the junction of the EcoRI site at 6551. This was performed by end-labeling this site in the plasmid pRI1-1 and re-cutting with Aha III to isolate the 1466-bp EcoRI-Aha III fragment. The ^{32}P-labeled primer was then denatured and annealed to a denatured dC-T4 DNA Aha III digest and extended with deoxy- and dideoxynucleoside triphosphates (32) and thereafter digested with EcoRI to transfer the labeled phosphate to the extended fragments.

The 8008-bp sequence was processed on a VAX computer (Digital Equipment Corp.) to search for ORFs, restriction enzyme sites and inverted repeats, using programs in the SEQ package at the Fox Chase Cancer Institute, Philadelphia, PA compiled by P. Young and H. Cael.

RESULTS AND DISCUSSION

To clone fragments from the denV region of T4 we initially isolated and tried to clone the 15-kb Sal I (58.0 to 73.2-kb on the T4 map) or Xho I (49.0 to 64.2-kb) fragments (35) directly into Sal I-digested and phosphatase-treated pBR322. This strategy proved to be unsuccessful in our hands probably because several of the genes from this region of T4 might be deleterious to E.coli metabolism, thereby making these large DNA fragments virtually unclonable in E.coli without causing deletions or rearrangements. As an alternative, smaller fragments of partially EcoRI, HindIII or Taq I digested dC-T4 DNA were cloned into pBR322 or pUN121 (Taq I fragments were cloned in the single Cla I site of pBR322). Clones containing T4 DNA from the denV region were identified by

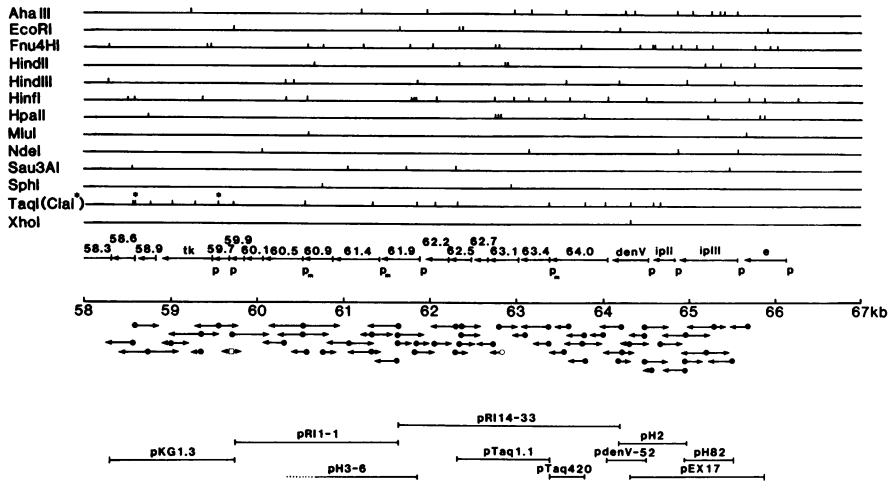


Fig 1. Restriction enzyme map, deduced genetic map, sequencing strategy, and T4 plasmid map of the 58 to 66-kb early region. ORFs and known genes have been indicated. Arrows show the direction of transcription and size of each ORF. Putative early (p) and middle mode (p_m) promoters are also included. Sequencing was performed from ends labeled with Klenow enzyme (●); T4 polymerase (○), or by primer-extension (□).

colony hybridization, using the gel-purified 15-kb Sal I fragment or 6.2-kb Sal I-Xho I (58.0 to 64.2-kb) fragment as nick-translation probes. A number of clones were isolated and further characterized by restriction enzyme mapping and Southern blotting, and the clones were mapped in respect to each other. The clone map in Fig 1 shows the clones we more closely examined and sequenced in this study. The T4 inserts in some of the plasmid clones, like pRI14-33 and pRI1-1, were used as hybridization probes to search for other clones containing inserts from this region. We have later determined that the insert in pH3-6 contained rearrangments (indicated by the dotted line in Fig 1) and only the right-hand portion of the ~3.5-kb insert seem to be intact. During our cloning experiments we noticed that some regions seem to be difficult to clone and others are possible to clone but the T4 DNA adversely affects the growth of the host cells. One region from which we have not been able to obtain extensive overlapping clones spans the EcoRI site at 4626. We believe that the reason for this is a promoter located in the upstream 688-bp EcoRI fragment which presumably acts on downstream detrimental T4 genes and thus prevent the cloning of fragments spaning this region. This hypothesis is also supported

by a 40-bp (4451-4491) tandem duplication found in plasmid pRI14-33 (data not shown). Another reason for these difficulties might be that this promoter together with downstream sequences interfere with plasmid replication. This has been reported to occur when attempts have been made to clone strong promoters on high-copy plasmids (36). We have also noticed that the clones carrying the ipIII or ipII genes such as pH2, pH82 and pEX17 (see Fig 1) have a reduced growth rate, suggesting some inconvenience for the host cells, but not severe enough to select for rearrangements or deletions of the T4 insert.

Sequencing

The sequencing of the denV region was performed as described in **Materials and Methods**, and the sequencing strategy is outlined in Fig 1. The sequence of the e gene is that reported by Owen et al. (37) and has been included in this report to obtain a better overall view of the DNA organization from this region. We have sequenced and confirmed the sequence between the HinfI site

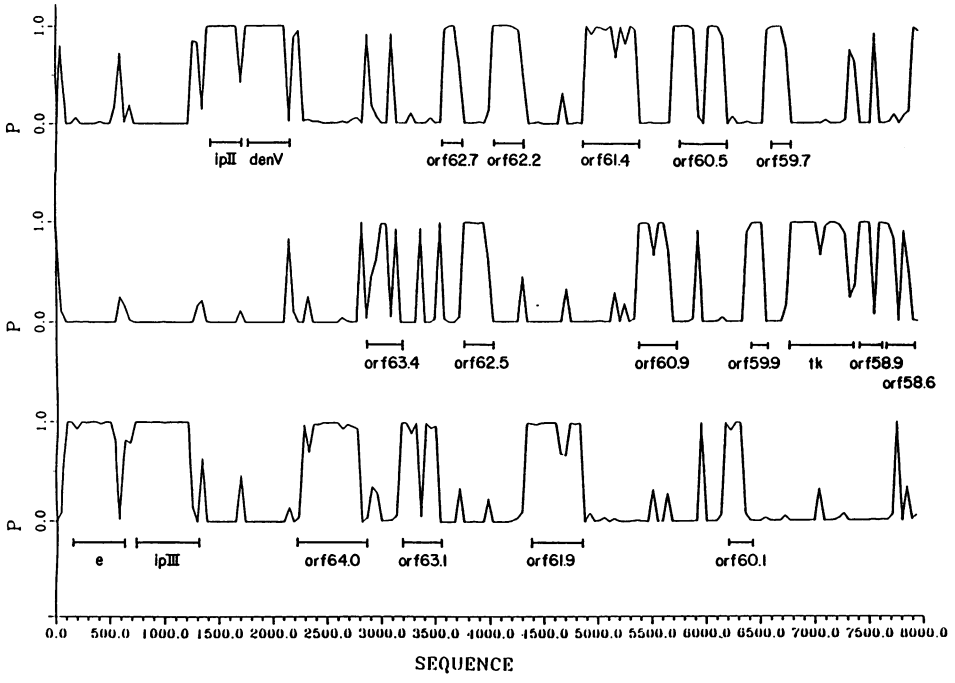
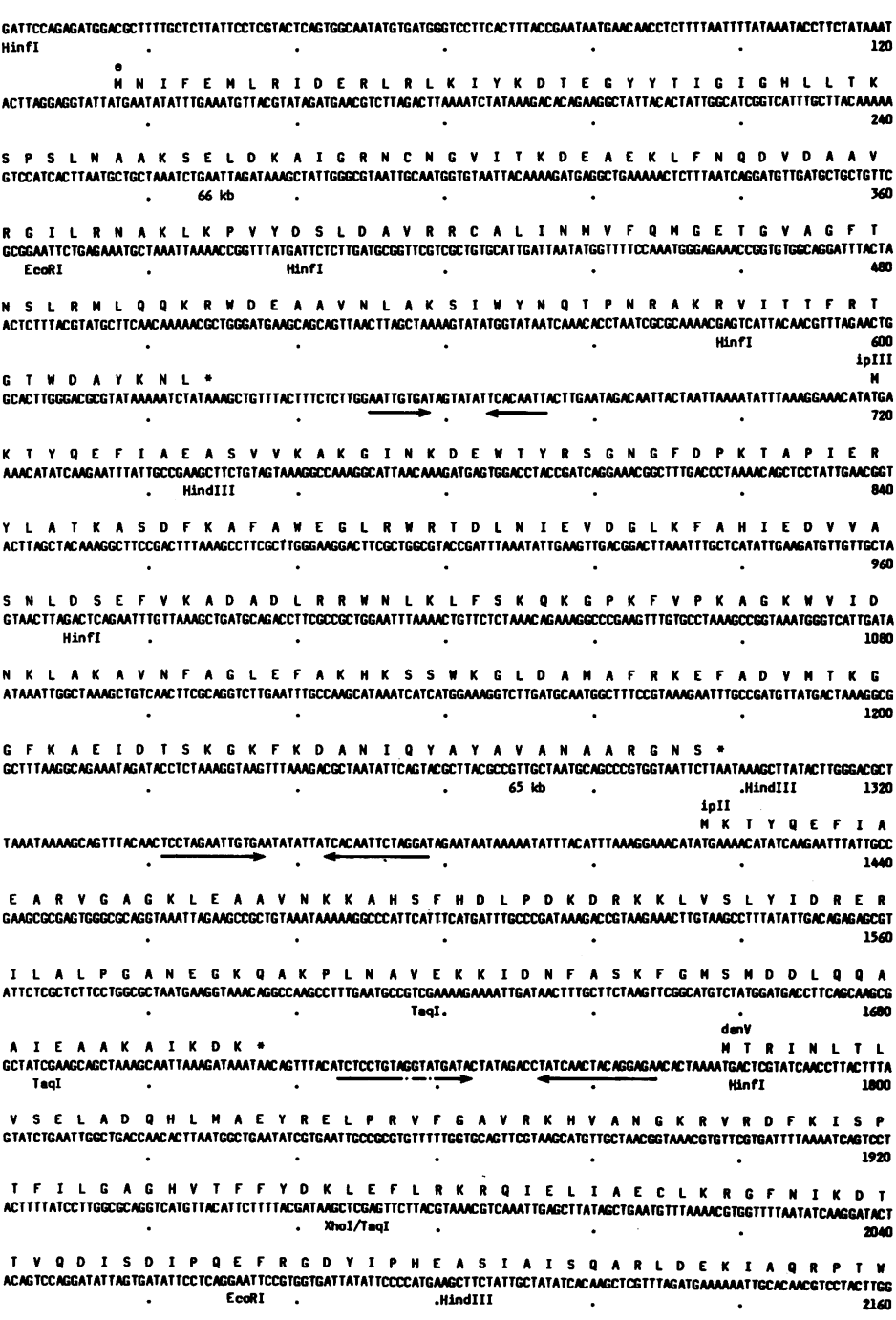
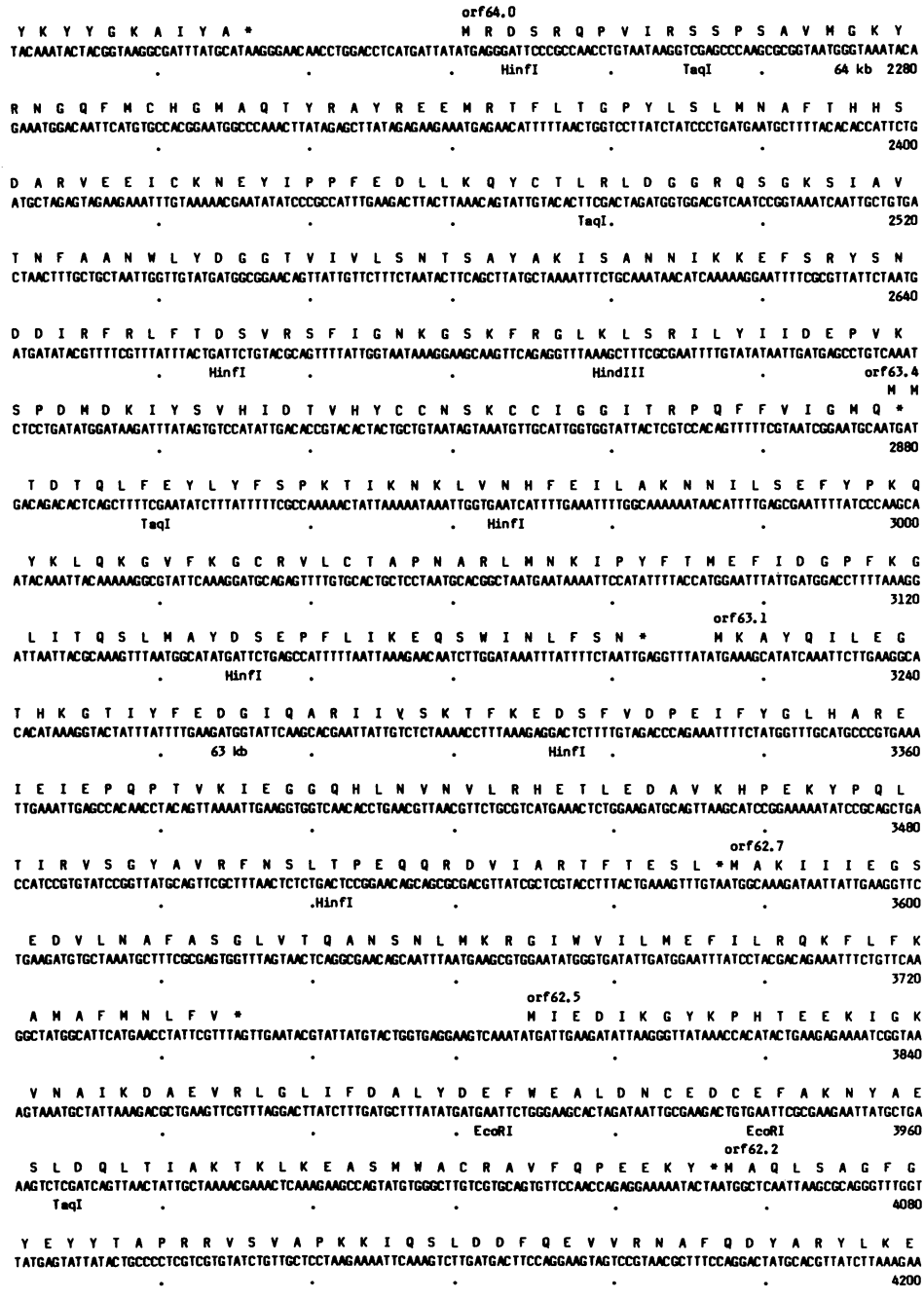
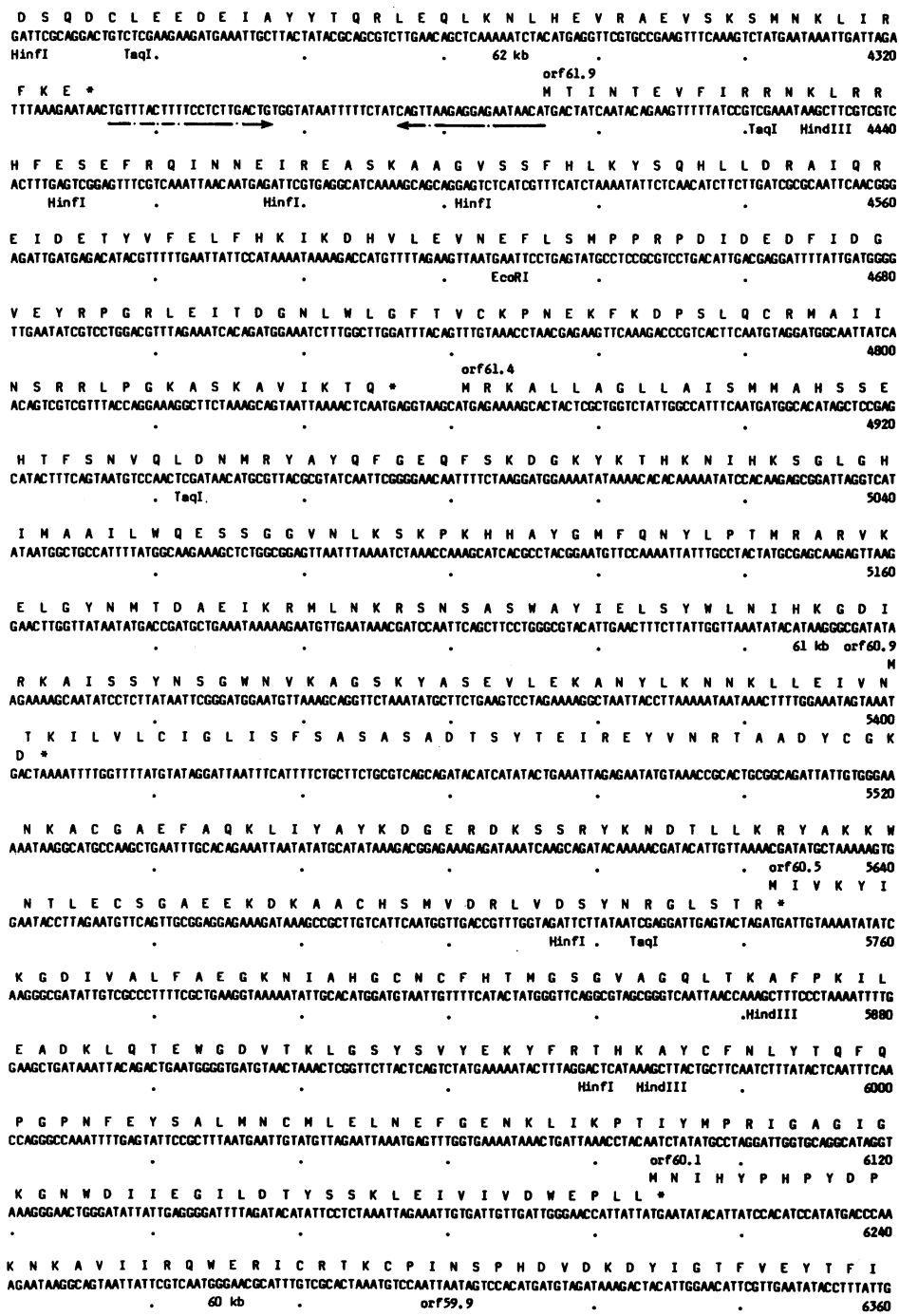


Fig 2. Putative genes in the 58 to 66-kb region. Shown are the plots of probabilities (0.0 to 1.0) for each of the three reading frames to code for a gene, based on the codon usage in the e, ipIII, ipII, and denV genes. Indicated under the peaks are the different ORFs deduced from the nucleotide sequence.


HinfI
66 kb
HinfI
HinfI
ipIII
HindIII
65 kb
.HindIII
ipII
TaqI.
dənV
TaqI
HinfI
XhoI/TaqI
EcoRI
.HindIII


orf64.0
Y K Y Y G K A I Y A *
M R D S R Q P V I R S S P S A V M G K Y
TACAATACTACGGTAAGGCGATTTATGCATAAGGGAACAACCTGGACCTCATGATTATATGAGGGATTCCCGCCAACCTGTAATAAGGTCGAGCCCAAGCGCGGTAATGGGTAAATACA
HinfI
TaqI
64 kb 2280
R N G Q F M C H G M A Q T Y R A Y R E E M R T F L T G P Y L S L M N A F T H H S
GAAATGGACAATTCATGTGCCACGGAATGGCCCAAACTTATAGAGCTTATAGAGAAGAAATGAGAACATTTTTAACTGGTCCTTATCTATCCCTGATGAATGCTTTTACACACCATTCTG
2400
D A R V E E I C K N E Y I P P F E D L L K Q Y C T L R L D G G R Q S G K S I A V
ATGCTAGAGTAGAAGAAATTTGTAAAAACGAATATATCCCGCCATTTGAAGACTTACTTAAACAGTATTGTACACTTCGACTAGATGGTGGACGTCAATCCGGTAAATCAATTGCTGTGA
TaqI.
2520
T N F A A N W L Y D G G T V I V L S N T S A Y A K I S A N N I K K E F S R Y S N
CTAACTTTGCTGCTAATTGGTTGTATGATGGCGGAACAGTTATTGTTCTTTCTAATACTTCAGCTTATGCTAAAATTTCTGCAAATAACATCAAAAAGGAATTTTCGCGTTATTCTAATG
2640
D D I R F R L F T D S V R S F I G N K G S K F R G L K L S R I L Y I I D E P V K
ATGATATACGTTTTCGTTTATTTACTGATTCTGTACGCAGTTTTATTGGTAATAAAGGAAGCAAGTTCAGAGGTTTAAAGCTTTCGCGAATTTTGTATATAATTGATGAGCCTGTCAAAT
HinfI
HindIII
orf63.4
M M
S P D M D K I Y S V H I D T V H Y C C N S K C C I G G I T R P Q F F V I G M Q *
CTCCTGATATGGATAAGATTTATAGTGTCCATATTGACACCGTACACTACTGCTGTAATAGTAAATGTTGCATTGGTGGTATTACTCGTCCACAGTTTTTCGTAATCGGAATGCAATGAT
2880
T D T Q L F E Y L Y F S P K T I K N K L V N H F E I L A K N N I L S E F Y P K Q
GACAGACACTCAGCTTTTCGAATATCTTTATTTTTCGCCAAAAACTATTAAAAATAAATTGGTGAATCATTTTGAAATTTTGGCAAAAAATAACATTTGAGCGAATTTTATCCCAAGCA
TaqI
HinfI
3000
Y K L Q K G V F K G C R V L C T A P N A R L M N K I P Y F T M E F I D G P F K G
ATACAAATTACAAAAGGCGTATTCAAGGATGCAGAGTTTTGTGCACTGCTCCTAATGCACGGCTAATGAATAAAATTCCATATTTTACCATGGAATTTATTGATGGACCTTTTAAAGG
3120
orf63.1
L I T Q S L M A Y D S E P F L I K E Q S W I N L F S N *
M K A Y Q I L E G
ATTAATTACGCAAAGTTTAATGGCATATGATTCTGAGCCATTTTTAATTAAAGAACAATCTTGGATAAATTTATTTTCTAATTGAGGTTTATATGAAAGCATATCAAATTCTTGAAGGCA
HinfI
3240
T H K G T I Y F E D G I Q A R I I V S K T F K E D S F V D P E I F Y G L H A R E
CACATAAAGGTACTATTTATTTTGAAGATGGTATTCAAGCACGAATTATTGTCTCTAAAACCTTTAAAGAGGACTCTTTTGTAGACCCAGAAATTTTCTATGGTTTGCATGCCCGTGAAA
63 kb
HinfI
3360
I E I E P Q P T V K I E G G Q H L N V N V L R H E T L E D A V K H P E K Y P Q L
TTGAAATTGAGCCACAACCTACAGTTAAAATTGAAGGTGGTCAACACCTGAACGTTAACGTTCTGCGTCATGAAACTCTGGAAGATGCAGTTAAGCATCCGGAAAAATATCCGCAGCTGA
3480
orf62.7
T I R V S G Y A V R F N S L T P E Q Q R D V I A R T F T E S L *M A K I I I E G S
CCATCCGTGTATCCGGTTATGCAGTTCGCTTTAACTCTCTGACTCCGGAACAGCAGCGCGACGTTATCGCTCGTACCTTTACTGAAAGTTTGTAATGGCAAAGATAATTATTGAAGGTTC
.HinfI
3600
E D V L N A F A S G L V T Q A N S N L M K R G I W V I L M E F I L R Q K F L F K
TGAAGATGTGCTAAATGCTTTCGCGAGTGGTTTAGTAACTCAGGCGAACAGCAATTTAATGAAGCGTGGAATATGGGTGATATTGATGGAATTATCCTACGACAGAAATTTCTGTTCAA
3720
orf62.5
A M A F M N L F V *
M I E D I K G Y K P H T E E K I G K
GGCTATGGCATTCATGAACCTATTCGTTTAGTTGAATACGTATTATGTACTGGTGAGGAAGTCAAATATGATTGAAGATATTAAGGGTTATAAACCACATACTGAAGAGAAAATCGGTAA
3840
V N A I K D A E V R L G L I F D A L Y D E F W E A L D N C E D C E F A K N Y A E
AGTAAATGCTATTAAAGACGCTGAAGTTCGTTTAGGACTTATCTTTGATGCTTTATATGATGAATTCTGGGAAGCACTAGATAATTGCGAAGACTGTGAATTCGCGAAGAATTATGCTGA
. EcoRI
EcoRI
3960
orf62.2
S L D Q L T I A K T K L K E A S M W A C R A V F Q P E E K Y *M A Q L S A G F G
AAGTCTCGATCAGTTAACTATTGCTAAAACGAAACTCAAAGAAGCCAGTATGTGGGCTTGTCGTGCAGTGTTCCAACCAGAGGAAAAATACTAATGGCTCAATTAAGCGCAGGGTTTGGT
TaqI
4080
Y E Y Y T A P R R V S V A P K K I Q S L D D F Q E V V R N A F Q D Y A R Y L K E
TATGAGTATTATACTGCCCCTCGTCGTGTATCTGTTGCTCCTAAGAAAATTCAAAGTCTTGATGACTTCCAGGAAGTAGTCCGTAACGCTTTCCAGGACTATGCACGTTATCTTAAAGAA
4200


HinfI
TaqI.
62 kb
4320
orf61.9
4440
.TaqI HindIII
HinfI
HinfI.
HinfI
4560
EcoRI
4680
orf61.4
4800
4920
5040
61 kb orf60.9
5160
5280
orf60.5
5640
HinfI . TaqI
5760
.HindIII
5880
HinfI HindIII
6000
orf60.1
6120
6240
60 kb
orf59.9
6360

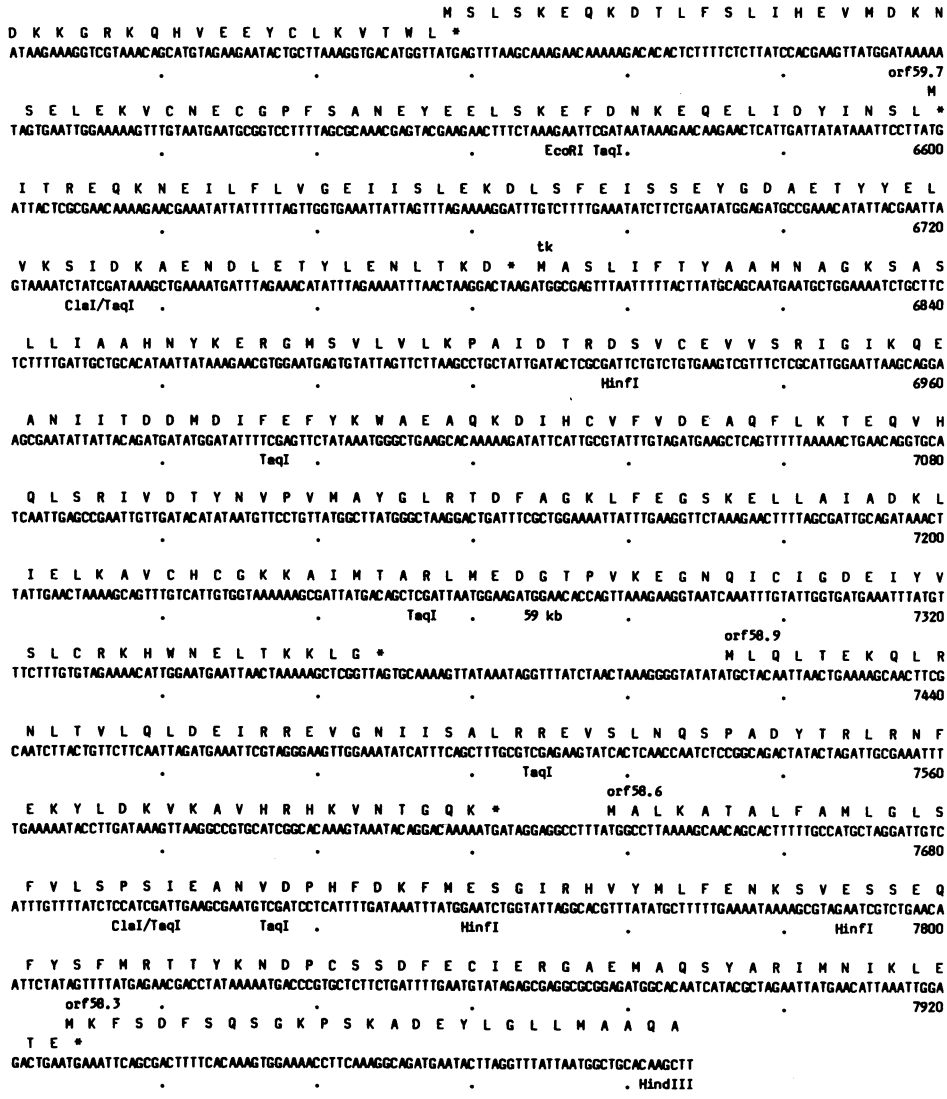


Fig 3. Complete nucleotide sequence of the 58.3 to 66.3-kb region. Shown are the complete nucleotide sequence of the 58.3 to 66.3-kb region with deduced amino acid sequences and gene designations at the start of each putative gene. Common restriction enzyme sites and T4 map coordinates are included. Significant inverted repeats are indicated with arrows.

at 577 and the HindIII site at 745. The EcoRI-HindIII insert in the tk plasmid pKG1.3 was cloned from a different T4 strain from the one used for the rest of the work (15). Approximately 200 nucleotides of this sequence has

been confirmed in our dC-T4 strain by primer extension of a phage DNA template from the EcoRI site at 6551.

Analysis of the nucleotide sequence

A computerized search for ORFs revealed at least 20 ORFs, all transcribed from the early strand of T4 (16). We have earlier described the mapping and nucleotide sequence of the denV gene (13), and we could identify both the ipIII and ipII genes from published primary amino acid sequence of the ipII and amino acid composition of the ipIII gene products (38,39). Together with the nucleotide sequence available for the e gene (37) we had four identified genes from this part of the T4 genome, which could serve as references genes to search for other genes by using the computer program FRAMESCAN (34). This program assesses the likelihood that an ORF may actually code for a protein, based on the known codon usage within the same genome. The plot of the FRAMESCAN program is shown in Fig 2. A very good correlation between this plot and ORFs preceded by putative ribosome binding sites can be observed, indicating with high probability that these ORFs are true T4 genes. Using the same technique, no significant ORFs were found on the late strand. The complete DNA sequence obtained from cloned T4 fragments is shown in Fig 3, with the deduced primary amino acid sequences for the ORFs indicated above the nucleotide sequence. Selected restriction enzyme sites obtained from the computer search are also indicated in the sequence. Many of these sites have been confirmed by restriction enzyme analysis of dC-T4 DNA and of cloned T4 fragments. To standardize our sequence with the T4 physical map, T4 map coordinates have been included, using the start of the e gene (66.13-kb on the T4 map) (3,37) as reference coordinates. Unknown genetic characters have been assigned as orfs followed by the T4 coordinates for respective putative start sites. The computer analysis has revealed that the potential ORFs are very tightly packed, with very little spacing between genes, indicating that phage T4 seems to use its genetic capacity economically. Where spacing between genes occurs one can recognize putative promoter and terminator sequences (33). A summary of the identified ORFs and their characteristics is given in Table I. Overall, 93% of the sequence is assigned to structural genes. Most of the ORFs have an A-T content of ~64%, which is close to the overall A-T content reported for T4 (40). The genes on which we have gathered some information are discussed individually below starting with the ipIII gene.

ipIII gene. The 5' portion of the ipIII gene has previously been identified by Owen et al. (37). We have been able to confirm and extend their results by sequencing the entire gene and compare the deduced amino acid sequence (17), with the partial primary amino acid sequence and amino acid

Table I. Summary of genes and open reading frames and their characteristics.

Position Sequence[a]	T4 map[b]	Ribosome binding sequence[c]	Designation[d]	Predicted mol. mass[e] (dalton)	Net charge[f]	% AT	Polarity index[g]	Term./init. codon
135- 626	66.13-65.64	TACTTAGGAGGTATTATGAAT	e	18,691	+9	63.2	48	TAA
717-1295	65.55-64.97	ATTTAAAGGAAACATATGAAA	ipIII	21,686	+11	59.0	47	TAATAA
1414-1713	64.85-64.55	ATTTAAAGGAAACATATGAAA	ipII	11,085	+6	58.6	49	TAA
1777-2190	64.49-64.07	CAGGAGAACACTAAAATGACT	denV	16,078	+9	61.8	46	TAA
2220-2876	64.05-63.39	GGACCTCATGATTATATGAGG	orf64.0	25,025	+15	63.2	47	ATGATGA
2268-	64.00-	AGCCCAAGCGCGGTAATGGGT		23,257	+13	64.6	47	
2295-	63.97-	AGAAATGGACAATTCATGTGC		22,175	+11	64.6	47	
2307-	63.96-	TTCATGTGCCACGGAATGGCC		21,747	+10	65.0	47	
2340-	63.93-	GCTTATAGAGAAGAAATGAGA		20,347	+10	65.2	47	
2376-	63.89-	CCTTATCTATCCCTGATGAAT		18,967	+9	65.2	48	
2825-3202	63.44-63.06	CTGCTGTAATAGTAAATGTTG	orf63.4	15,048	+8	69.4	44	TGA
2876-	63.39-	CGTAATCGGAATGCAATGATG		12,834	+5	71.0	44	
3213-3572	63.05-62.57	CTAATTGAGGTTTATATGAAA	orf63.1	13,813	+1	59.2	51	TAATG
3575-3748	62.69-62.52	TACTGAAAGTTTGTAATGGCA	orf62.7	6,611	+2	62.0	34	TAG
3788-4051	62.48-62.21	GTGAGGAAGTCAAATATGATT	orf62.5	10,210	-6	63.2	48	TAATG
4054-4329	62.21-61.94	AGAGGAAAAATACTAATGGCT	orf62.2	10,904	-1	60.2	53	TAA
4392-4850	61.87-61.42	TAAGAGGAGAATAACATGACT	orf61.9	17,976	+7	61.8	53	TGA
4861-5403	61.40-60.86	CTCAATGAGGTAAGCATGAGA	orf61.4	20,680	+20	64.0	50	ATGACTAA
5399-5743	60.87-60.52	TTTTGGAAATAGTAAATGACT	orf60.9	13,055	+6	65.6	53	ATGA
5743-6207	60.52-60.06	AGGATTGAGTACTAGATGATT	orf60.5	17,488	+1	65.6	39	ATGA
6207-6416	60.06-59.85	TTGGGAACCATTATTATGAAT	orf60.1	8,506	+7	65.8	50	ATGA
6416-6598	59.85-59.67	AAAGGTGACATGGTTATGAGT	orf59.9	7,134	-8	69.4	61	ATGA
6598-6783	59.67-59.48	TTATATAAATTCCTTATGATT	orf59.7	7,238	-10	73.6	56	TAAGATG
6788-7366	59.48-58.90	TAACTAAGGACTAAGATGGCG	tk	21,621	+3	65.8	44	TAG
7412-7621	58.85-58.64	ACTAAAGGGGTATATATGCTA	orf58.9	8,273	+8	63.4	59	TGATAG
7637-7927	58.63-58.34	TGATAGGAGGCCTTTATGGCC	orf58.6	11,124	-3	63.6	47	ATGA
7927-	58.34-	TAAATTGGAGACTGAATGAAA	orf58.3	>2,921				
		E.coli 16S RNA 3' AUUCCUCCACAUG				Average % AT: 64.0		

a, The sequence between 1 and 745 has been published previously (37); b, The T4 map coordinates have been adjusted to the start site of the e gene at 66.13-kb; c, Putative ribosome binding sequences (43) complementary to the E.coli 16S rRNA have been underlined, and the initiation codons are indicated with an arrow; d, The unidentified ORFs have been designated as orfs followed by the T4 coordinate for respective start site (see b), and identified genes have been designated with their proper name (3); e, Deduced from the nucleotide sequence; f, Net charge after adding up the acidic (asp and glu), and basic (arg, lys, and his) amino acid residues; g, Polarity index calculated as in (54).

composition of the ipIII gene product (38). The gene encodes a protein of 21,686 dalton, which is subsequently processed to a mature protein of 20,444 dalton by cleavage between the glutamic acid-10 and alanine-11 residues by the protease encoded by T4 gene 21 (41).

ipII gene. The ipII gene was identified in the same manner as the ipIII gene (17), by comparing the deduced amino acid sequence from the nucleotide sequence with the primary amino acid sequence of internal protein II (39). The gene codes for an unprocessed protein of 11,085 dalton, which after cleavage by gp21 between amino acid residues 10 and 11 will be reduced to 9,843 dalton.

Both ipIII and ipII encode proteins which are fairly basic in nature, a property not unexpected for proteins believed to have DNA-binding capabilities and functions similar to histones in eukaryotic cells.

denV gene. The denV sequence (13), cloning and expression of the denV gene in E.coli (14) have been reported previously. The gene codes for the 16,078 dalton endonuclease V, a pyrimidine dimer-specific DNA N-glycosylase (42).

orf64.0. This open reading frame has 6 potential start sites beginning with ATGs, encoding proteins with deduced molecular masses between 25,025 and

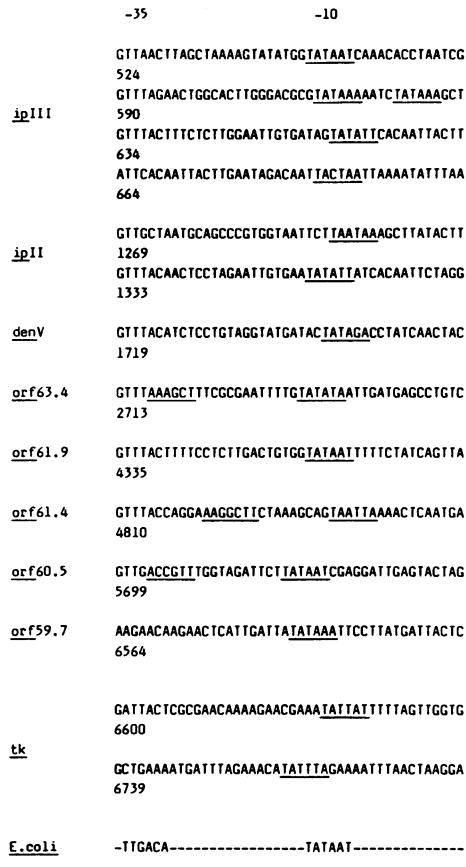


Fig 4. Putative early and middle mode promoters. Sequences which resemble the E.coli consensus promoter sequence (33) have been lined up at the -35 region. The genes (ORFs) following the promoter-like sequences have been indicated to the left. The 'Pribnow-box' at position -10 has been underlined as well as the A(AT)TGCTT sequences associated with middle mode promoters (47).

18,967 dalton. By examining the sequences preceding the initiation codons, probably one of the first two ATGs is utilized *in vivo*, since they seem to have the best complementary fit to the *E.coli* 16S rRNA-binding site (43) (see Table I). The genetic pressure to utilize most of the DNA and the observation that the most favorable and likely nucleotide at the +4 position of an *E.coli* gene is a purine (44) also suggest that one of the earlier initiation codons are being utilized. Of course, one cannot eliminate the possibility that T4 might use different initiation codons at different times during the infectious cycle. It has been reported that a 23 kDa protein is absent from extracts of *denV*$^-$ $T4_{v1}$-infected *E.coli*, as judged by SDS-polyacrylamide electrophoresis (42). We have reported the mutation in the *denV* gene from the T4 strain $T4_{v1}$ to be a -T frame shift, creating a premature termination codon in the *denV* gene and potentiating the production of a 13.5 kDa nonsense endonuclease V protein (13). If strong polarity effects are envisioned as part of T4 translational regulation, which is suggested by the tight coupling of adjacent genes reported here and elsewhere (45,46), it is possible that the 23 kDa protein missing in $T4_{v1}$-infected *E.coli* extracts is the *orf*64.0 gene product downstream from the *denV* gene.

orf63.4. This ORF has two potential initiation codons. The most likely one to be utilized is the second one starting at position 2876 for the same reasons discussed earlier for *orf*64.0 (see Table I). Here we also find a tight coupling between the termination codons ATGATGA from the preceding *orf*64.0 gene with the initiation codon of *orf*63.4 at 2876. As mentioned for *orf*64.0, the first initiation codon at 2825 could still be used at certain times during infection. It is tempting to speculate that the initiation codon at 2825 is being utilized later in infection since the putative promoter at 2713 which is located within the structural *orf*64.0 gene, seems to have a middle mode character suggested by the AAGCTT sequence in the -35 region (47) (Fig 4).

orf63.1. This putative gene encodes a slightly basic (+1) protein with a deduced molecular mass of 13,813 dalton. The gene has a well defined ribosome binding site preceding the ATG codon (see Table I). We have observed that the DNA from this gene hybridizes with chromosomal DNA from uninfected *E.coli* on Southern blots, indicating a fair amount of DNA homology between T4 and *E.coli* within this gene (data not shown). In fact, we reported the isolation of a plasmid that hybridized to the 62-kb region of T4 (Fed. Proc., 1983, 42:1985), which we now believe does not contain T4 DNA but instead *E.coli* DNA. We found that only the ~500-bp flanking sequences of the 3.4-kb *Sal* I insert in this plasmid hybridized to a 559-bp *Hind*II probe (3414-3973, see Fig 3) partially spanning *orf*63.1. We have determined the nucleotide sequence of part of the homologous

flanking sequence and found that the recovered plasmid shared sequence at least between positions 3440 and 3565 with orf63.1 (data not shown). The homology ended at position 3565, while the putative gene in the 3.4-kb insert continued for another 13 nucleotides before terminating. As can be seen from the sequence in Fig 3, orf63.1 terminates 7 nucleotides downstream from where the shared homology with the 3.4-kb insert ends. The implications of this preliminary finding need to be investigated further, since most of the genes from this region of T4 still have not been associated with a phenotypic property.

orf59.7. While all of the other identified and unidentified genes from this region have recognizable ribosome binding sites preceding their respective initiation codon, orf59.7 has not (see Table I). Instead there seems to be a promoter-like sequence 12 bp upstream from the ORF, potentiating the synthesis of a transcript without any leader sequence (see Fig 4). This has been reported for another putative T4 gene and is believed to destabilize the interaction between the mRNA and the 16S rRNA and limit the utilization of the mRNA from such genes (45).

tk gene. The cloning, sequence, and expression of the tk gene in E.coli will be described elsewhere (15). The gene encodes a moderately basic (+3) protein with a deduced molecular mass of 21,621 dalton (Table I). It may be interesting that the tk gene and the following gene, orf58.9 show some similarity of nucleotide sequence with the splicing sequences of the td gene intron (48). If the exon-intron sequences of the td gene is lined up with the spacing sequence between the tk gene and orf58.9, one can see that there are similarities at the junctions:

```
                                     Splice
                            |---------------------------|
td            -GTTTTCTTGGGT TAATTGAGG-(999 bp)-AACATAATG CTACCGTTT-
                      * *** **  *  *                  *** ****  **
tk - orf58.9  -AAAAAGCTCGGT TAGTGCAAA-( 30 bp)-GTATATATG CTACAATTA-
               7355
```

None of the other ORFs identified in this region (see Table I and Fig 3) show such homology with the td splicing sequence as the tk-orf58.9 junction. However, spliced mRNA from this region has yet to be detected.

Putative promoters and transcriptional terminators

A number of putative promoter sequences can be found in the 8-kb sequence (Fig 4). The most outstanding ones are found in the e-denV gene region. This group of genes are all known as early genes and should therefore be recognized by unmodified E.coli RNA polymerase (16,47). As can be seen in Fig 4 these promoters show a very good resemblance to the E.coli consensus promoter sequence (33). The sequence at the -35 position seems to be a little different from the consensus sequence by having a GTTTACA instead of TTGACA, while the -10 sequence

TATAAT seems to be more consistent. Other putative promoters, like the ones preceding orf63.4, orf61.4, and orf60.5, seem to have middle mode characteristics with the sequence A(AT)TGCTT close to the -35 region (47). In these cases the promoters are buried in the upstream structural genes, a feature which seems logical if the upstream gene is an early one and is being switched off before transcription initiates at middle promoters. Some genes which are known to be expressed throughout the infectious cycle, like the ipIII and ipII genes, seem to have a number of different and sometimes overlapping promoters. The distance between the -35 and -10 regions among these putative promoters also vary to some extent. Testing fragments of pRI14-33 (Fig 1) for RNA polymerase binding in vitro, we have confirmed the existence of the promoter preceding orf61.9. Binding was assayed by retention of ^{32}P-labeled DNA fragments on nitrocellulose filters after incubation with E.coli RNA polymerase and heparin (49). Our findings correlate well with the observation we have made that the region spanning the EcoRI site at 4626 is difficult if not impossible to clone. This promoter might be the early promoter mapped by Gram et al. (50) who determined by in vitro transcription of T4 DNA that a sequence close to the 688-bp EcoRI fragment initiates transcription very efficiently.

The late promoter consensus sequence TATAAATA (51) occurs at positions 101 and 114, 34 and 21 bp upstream from the start codon of the e gene (52), and again at 7379, 33 bp upstream from the start codon of orf58.9. Transcription from the late promoters upstream from the e gene has been confirmed by S_1 mapping and primer extension (52).

Searching for inverted repeats to pinpoint potential transcriptional terminators (33), three or four regions stand out. These include stem-loops preceding the ipIII, ipII, denV and orf61.9 genes, all indicated with arrows in the nucleotide sequence (Fig 3). The Gibbs free energy of these inverted repeats are -5.8, -20.8, -19.9, and -16.4 kcal/mole respectively. Interestingly, all four of the inverted repeats are formed in such a way that the -10 'Pribnow-box' will be at the top of the loop and very much exposed to a RNA polymerase. A potential, very strong, stem-loop between coordinates 649-669 and 1352-1372 is also possible, but since there would be a loop of approximately 700 nucleotides between the complementary sequences, it is questionable if this structure occurs in vivo.

CONCLUSIONS

The analysis of this nucleotide sequence has revealed, like other investigations on large segments of the T4 genome (45,46), that T4 seems to encode a

myriad of small and genetically unidentified genes. Many of these genes might be small accessory proteins involved in metabolic functions during T4 infection of E.coli. Our analysis also shows that the T4 genetic material is being utilized maximally for structural genes (93%) and where intergene sequences can be found, transcriptional control elements can be implicated. Interestingly, many of the putative genes seem to be coupled to each other by sharing overlapping termination and initiation codons, like ATGA. This overlap has been observed by other investigators (45), and seems to assure an effective translation of a whole operon.

So far we have been able to identify genetically only 4 out of 20 putative genes from this region. Another two genes, namely the vs and regB genes, map most certainly to this region. The vs gene has been genetically mapped between the tk and regB genes (3). It has been observed that an E.coli strain which has a ts valyl-tRNA synthetase and is carrying a plasmid containing the 1925-bp (4626-6551) EcoRI fragment grows at the non-permissive temperature, indicating that the T4 fragment carries the T4 vs gene (G. Marchin, personal communication). The regB gene is most likely either orf62.7, orf62.5 or orf62.2, since the regB gene has been mapped at 61-kb between the denV and vs genes (3) and the molecular mass of the RegB protein is believed to be fairly small.

We do not have any physical evidence, except for the four genes we have identified, that the ORFs from this region produce proteins with the molecular masses we present in Table I. A previous study has shown the synthesis of five polypeptides with molecular masses of 22, 21, 14, 11, and 9 kDa from a λ-T4 hybrid phage containing a partial HindIII fragment (1.5 and 2.0-kb) (53), which most likely are the 1433, 105 and 2037-bp HindIII fragments (4428-5861, 5861-5966 and 5966-8003) in our sequence. The molecular masses of the ORFs in these three adjacent HindIII fragments deduced from our sequence are 22, 21, 17.5, 13.1, 11.1, 8.5, 8.3, 7.2 and 7.1 kDa (Table I), which are in good agreement to the ones produced in the λ-T4 hybrid.

ACKNOWLEDGMENTS

The authors wish to thank the staff at the Computer Center at the Fox Chase Cancer Institute, Phila., Pa, particularly Lee-Ann Wiedemann, Holly Cael and Bob Stodola, for providing help with the computer analysis. We also thank George Marchin for giving us information regarding the vs gene prior to publication. This work was supported by a Basic Research Grant (I-899) to J.K. de Riel from the March of Dimes Foundation.

Present addresses: [4]Department of Molecular Genetics and [5]Department of Tumor Biology, Smith Kline and French Laboratories, Philadelphia, PA 19101, USA

REFERENCES

1. Snyder, L., Gold, L., and Kutter, E. (1976) Proc. Natl. Acad. Sci. USA 73, 3098-3102.
2. Mathews, C., Kutter, E., Mosig, G., and Berget, P. (eds.) (1983) Bacteriophage T4, ASM, Washington, D.C.
3. Wood, W.B. and Revel, H.R. (1976) Bacteriol. Rev. 40, 847-868.
4. Black, L.W. and Ahmad-Zadeh C. (1971) J. Mol. Biol. 57, 71-92.
5. Black, L.W. (1974) Virology 60, 166-179.
6. Harm, W. (1963) Virology 19, 66-71.
7. Chace, K.V. and Hall, D.H. (1975) J. Virol. 15, 929-945.
8. Marchin, G.L. (1980) Science 209, 294-295.
9. Chace, K.V. and Hall, D.H. (1973) J. Virol. 12, 343-348.
10. Krylov, V.N. (1971) Genetika 7, 112-119.
11. King, G.J. and Huang, W.M. (1982) Proc. Natl. Acad. Sci. USA 79, 7248-7242.
12. Yee, J.K. and Marsh, R.C. (1985) J. Virol. 54, 271-277.
13. Valerie, K., Henderson, E.E., and de Riel, J.K. (1984) Nucleic Acids Res. 12, 8085-8096.
14. Valerie, K., Henderson, E.E., and de Riel, J.K. (1985) Proc. Natl. Acad. Sci. USA 82, 4763-4767.
15. Stevens, J., Gutekunst, K.A., Hall, D.H., Valerie, K., and de Riel, J.K. To be submitted.
16. Rabussay, D. (1983) in Bacteriophage T4. Mathews, C., Kutter, E., Mosig, G., and Berget, P. (eds.) ASM, Washington, DC. pp.167-173.
17. de Riel, J.K., Valerie, K., Lynch, M., and Henderson, E.E. Submitted.
18. Yamamoto, K.R., Alberts, B., Benzinger, R., Lawhorne, L., and Treiber, G. (1976) Virology 40, 734-744.
19. Davis, R.W., Botstein, D., and Roth, J.R. (1980) Advanced Bacterial Genetics, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
20. Boyer, H.W. and Roulland-Dussoix, D. (1969) J. Mol. Biol. 41, 459-472.
21. Meselson, M. and Yuan, R. (1968) Nature 217, 1110-1114.
22. Marinus,M.G. and Morris, N.R. (1973) J. Bacteriol. 114, 1143-1150.
23. Bolivar, F., Rodriguez, R.L., Greene, P.J., Betlach, M.C., Heynecker, H.L., Boyer, H.W., Crosa, J.H., and Falkows, S. (1977) Gene 2, 95-114.
24. Nilsson, B., Uhlen, M., Josephson, S., Gatenbeck, S., and Philipson, L. (1983) Nucleic Acids Res. 11, 8019-8030.
25. Godson, G.N., and Vapnek, D. (1973) Biochem. Biophys. Acta 299, 516-520.
26. Clewell, D.B. and Helinski, D.R. (1972) J. Bacteriol. 110, 1135-1146.
27. Peacock, A.C. and Dingman, C.W. (1968) Biochemistry 7, 668-674.
28. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
29. Hanahan, D. and Meselson, M. (1980) Gene 10, 63-67.
30. Southern, E.M. (1975) J. Mol. Biol. 98, 503-517.
31. Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
32. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
33. Rosenberg, M. and Court, D. (1979) Ann. Rev. Genet. 13, 319-353.
34. Staden, R. and McLachlan, A.D. (1982) Nucleic Acids Res. 10, 141-156.
35. O'Farrell, P.H., Kutter, E., and Nakanishi, M. (1980) Mol. Gen. Genet. 179, 421-435.
36. Stueber, D. and Bujard, H. (1982) EMBO J. 1, 1399-1404.

37. Owen, J.E., Schultz, D.W., Taylor, A., and Smith, G.R. (1983) J. Mol. Biol. 165, 229-248.
38. Tsugita, A., Black, L.W., and Showe, M.K. (1976) J. Mol. Biol. 98, 271-275.
39. Isobe, T., Black, L.W., and Tsugita, A. (1976) J. Mol. Biol. 102, 349-365.
40. Sober, H.A. (1970) in Handbook of Biochemistry, 2nd edit. p. H100, CRC Press, Cleveland, Ohio.
41. Black, L.W., and Showe, M.K. (1983) in Bacteriophage T4. Mathews, C., Kutter, E., Mosig, G., and Berget, P. (eds.) ASM, Washington, D.C. pp. 219-245.
42. Nakabeppu, Y. and Sekiguchi, M. (1981) Proc. Natl. Acad. Sci. USA 78, 2742-2746.
43. Shine, J. and Dalgarno, L. (1975) Nature 254, 34-38.
44. Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Swebelius-Singer, B., and Stormo, G. (1981) Ann. Rev. Microbiol. 35, 365-403.
45. Gram, H. and Ruger, W. (1985) EMBO J. 4, 257-264.
46. Broida, J. and Abelson, J. (1985) J. Mol. Biol. 185, 545-563.
47. Brody, E., Rabussay, D., and Hall, D.H. (1983) Bacteriophage T4. Mathews, C., Kutter, E., Mosig, G., and Berget, P. (eds.), ASM, Washington, D.C., pp. 174-183.
48. Chu, F.K., Maley, G., Maley, F., and Belfort, M. (1984) Proc. Natl. Acad. Sci. USA 81, 3049-3053.
49. Fukada, K., Gossens, L., and Abelson, J. (1980) J. Mol. Biol. 137, 213-234.
50. Gram, H., Liebig, H.D., Hack, A., Niggemann, E., and Ruger, W. (1984) Mol. Gen. Genet. 194, 232-240.
51. Christensen, A.C. and Young, E.T. (1983) in Bacteriophage T4. Mathews, C., Kutter, E., Mosig, G., and Berget, P. (eds.), ASM, Washington, D.C., pp. 184-188.
52. McPheeters, D.S., Christensen, A., Young, E.T., Stormo, G., and Gold, L. (1986) Nucleic Acids Res. 14, 5813-5826.
53. Mileham, A.J., Murray, N., and Revel, H.R. (1984) J. Virol. 50, 619-622.
54. Capaldi, R.A. and Vanderkooi, G. (1972) Proc. Natl. Acad. Sci. USA 69, 930-932.