Nucleotide sequence of the essential region of bacteriophage P4

Ching-Shwun Lin*

Department of Microbiology, University of Iowa, Iowa City, IA 52242, USA

Received 16 August 1984; Revised and Accepted 19 October 1984

ABSTRACT

Nucleotide sequence of one-third of the genome of coliphage P4 has been obtained and mutations vir1, εaml04, ci405, sid1, and §35 identified. The ε gene likely encodes a 10 kd protein with εaml04 being located at the beginning of the gene. ci405, a proposed repressor gene mutation, is located in a sequence capable of coding for a 15 kd protein. A new class of P4 mutations, ash, is located in the neighborhood of ci405. Two TATA-like sequences are mapped 5' to this ci (ash) sequence. Vir1 is possibly a promoter-up mutation and is located near or within the replication origin, which is about 400 bp long and AT rich. A sid1 mutation is amber that shortens the sid protein by 9 amino acids. The § gene may encode a 17 kd protein and appears to be coupled with the sid gene translationally. In the 5' side of the sid gene a sequence of CACAAT is the best TATA-like sequence. Sequences of two possible genes that are previously unrecognized and part of the α and psu genes are also identified.

INTRODUCTION

P4 is a coliphage that requires a helper phage such as P2 to complete the life cycle (1). P4 can replicate its DNA and lysogenize in the absence of the helper (1) but depends on all 18 late gene products of the helper to grow lytically (2).

For replication, the α gene of P4 is essential (3). For lysogenization, P4 has a repressor gene (4), an int gene and the attachment site (5). For acquiring late gene products from the helper, P4 uses two distinct mechanisms, namely, derepression and transactivation, that are governed by the ε and δ genes, respectively (6,7).

The derepression mechanism causes the helper prophage to replicate and subsequently express the late genes (4). The transactivation mechanism activates the helper's late genes in the absence of replication of the helper (2).

Fig. 1. Genetic map of P4 and sequenced regions in nine P4 strains. The nonessential region is located approximately between 8 kb and 11 kb. Each line under the genetic map represents a sequenced region and each x represents a mutation.

Although P4 acquires its capsid protein from the helper, it possesses the ability to assemble capsids that are only one-third the size of the helper's capsid. Such an ability derives from its *sid* gene (8).

P4 is also capable of suppressing the polarity of amber mutations of the late genes of the helper (9). The responsible gene is *psu*; its role in the growth of P4 is not known (10).

The genome of P4 is an 11.2 kb linear DNA (11) with one-third of it being 'nonessential' because it can be replaced by foreign DNA or partially deleted (12). This nonessential region contains the *int* gene, the attachment site, and an early gene β whose function is unknown (3,13).

The other two-thirds of the genome contains the two early genes, α and ϵ , that are transcribed in one direction, the three late genes, *sid*, δ , and *psu*, that are transcribed in the other direction (14), and the replication origin in between (15). This 'essential' region also contains a *vir1* locus and the putative repressor gene *cI* (5). The *vir1* mutation confers partial insensitivity to P4 immunity (1,16); its molecular nature will be discussed in this report.

In studying the derepression mechanism of P4, we (Lin and Six (17)) have isolated a group of mutants called *ash* for their

adaptation to a secondary helper, phage P3. We found that ash and cI mutations are identical and are associated with the vir1 locus and the ϵ gene. To understand the molecular relationship among vir1, ash, and ϵ , I sequenced the DNA of the vir1-ash- ϵ region for several P4 strains. During this sequencing work, many DNA fragments in the sid- δ -psu region were also generated. Since there has been interest in the relationship among these three genes and comparing them with P2 genes, I decided to sequence these fragments. In the end, I have obtained a continuous wildtype P4 DNA sequence of 3819 nucleotide pairs that covers most of the essential region except for the large body of gene α . Sequences covering the respective mutational sites in 8 P4 mutants were also obtained.

Because our lab was primarily interested in the vir1-ash- ϵ region, I also attempted to map promoters in this region and to identify the product of the ϵ gene. Their results are also presented here.

MATERIALS AND METHODS

Bacterial Strains

C-1a (prototropic, 18), C-1792 (supF, 19), C-3102 (rho102, 20), and HF4704 (uvrA, 21).

Phage Strains

P4 wt (1), P4 vir1 (16), P4 vir1 sid1 (8), P4 cI405 (5), P4 ϵ aml04 (6; has an ash9 mutation, 17), P4 ash8, P4 ash9, P4 ash10, P4 vir1 α aml ash3, and P4 vir1 635 ash26 (17).

Extraction of P4 DNA from Lysogens Carrying P4 as Plasmids

P4 ash8 and P4 ϵ aml04 ash9 were found to exist as high copy number plasmids in lysogens derived from C-1a and C-3102, respectively (17). Their DNA could thus be obtained from these lysogens, following the standard procedures for extraction of plasmids as described by Davis *et al.* (22). In brief, cells were grown in LB to about 5×10^8 /ml, concentrated, and then lysed with lysozyme and Triton X-100. Chromosomal and Plasmid DNA in the lysate were then separated in CsCl gradient. From 1 liter of culture about 150 μ g of P4 ash8 DNA and 500 μ g of P4 ϵ aml04 ash9 DNA were obtained.

Extraction of P4 DNA from Infected Cells

Kahn *et al.* (23) have performed extraction of P4 DNA from infected cells previously. In this work, C-la had C-1792 were used as hosts for P4 am⁺ and P4 am⁻, respectively. The procedures are briefly described as follows. Cells were grown in LB to about 1.5×10^8 /ml and then concentrated 2-fold in LB supplemented with 5 mM CaCl₂. P4 was then added at an MOI of 3. Following 7 min of adsorption period, 9 volume of LB were added. The culture was allowed to grow for another 2 hours. Thereafter, standard plasmid extraction procedures were followed to extract small circular P4 DNA. From 1 liter of the final culture, usually 30 μ g of DNA could be obtained for P4 *vir*⁺ and 150 μ g for P4 *vir*⁻.

Cloning of P4 DNA

In order to obtain sufficient amount of DNA for sequencing, the 6.2 kb *Bam*HI fragment which contains most of the P4 essential genes (13) was cloned into pBR322, using C-la as the recipient for transformation. The procedures were essentially that of Davis *et al.* (22). This cloning was performed for P4 *wt*, P4 *vir1*, P4 *ash8*, P4 ϵ aml04 *ash9*, P4 *ash10*, P4 *vir1* α aml *ash3*, and P4 *vir1* 635 *ash26*. Only the attempt for P4 *vir1* α aml *ash3* was unsuccessful.

DNA Sequencing

Most DNA sequencing work followed the method of Maxam and Gilbert (24). The sequencing strategy is summarized in Figure 2. In this strategy, each restriction fragment preparation was divided into two portions and labeled separately with γ -P³²-ATP for 5' ends and α -P³²-dNTP for 3' ends. Each labeled fragment was then treated with a restriction enzyme to remove one end or occasionally strand-separated. The products were therefore singly labeled DNA molecules that are ready for sequencing reactions. With this procedure, all DNA fragments were sequenced on both strands with most regions sequenced more than once to achieve unequivocal readings.

The region between the two far right *Hinf*I sites in Fig. 2 does not contain restriction sites that would generate convenient subfragments for sequencing. *Nci*I (CC⁺CGG or CC⁺GGG) cuts in this region; however, it generates fragments as small as 12 bp in

Fig. 2. Sequencing strategy. The entire fragment is about 4 kb long. Relevant restriction sites are labeled as A for AvaI, D for DdeI, F for HinfI, H for HindIII, M for BamHI, N for NciI, S for BstNI, and T for TaqI. Each arrow represents about 250 to 300 bp of both strands with the direction corresponding to the direction of sequencing (see text for details). Mutations identified in this fragment are labeled as a for ash, δ for 635, ϵ for ϵ aml04, s for sid1, and v for vir1.

length. To overcome this problem, I devised a special sequencing strategy as follows. The HinfI fragment was partially digested with NciI until a wide spectrum of fragments of various lengths had appeared. The digest was then divided into 3 portions. One large portion (45 μ l) was labeled with α - P^{32} -dCTP, another with α - P^{32} -dGTP, and a small portion (11 μ l) was labeled with α - P^{32} -dATP (C or G labels one of the two possible NciI ends, A labels the HinfI end). After electrophoresis in a polyacrylamide gel, the labeled fragments were visualized by autoradiography (Figure 3). Any fragments that appeared in (i) both the C and G panels, (ii) both the C and A panels, or (iii) both the G and A panels must be labeled only at one end in each of the C, G and A panels. They are marked as C1 to C6 and G1 to G6 in Figure 3. (Those in the A panel did not need to be sequenced because sequences of both ends of the original HinfI fragment were determined with the standard sequencing strategy.) DNA sequencing was then conducted for fragments G1 to G6, and their sequences allowed the determination of which of the C1 to C6 fragments needed to be sequenced. The same HinfI fragment and a TaqI-BamHI fragment in the sid- δ region (Figure 2) were also sequenced with the method of Sanger *et al.* (25) with HaeIII subfragments cloned into the HincII site of M13 mp8 (23).

Fig. 3. A 5% polyacrylamide gel showing partially NciI digested and differentially labeled DNA fragments. The far right panel displays HinfI fragments of pBR322 as size markers. C, G, and A panels are described in the text.

Mung Bean Nuclease Mapping

Sollner-Webb and Reeder (27) used the S1 mapping method developed by Berk and Sharp (28) to map the transcriptional initiation site. Green and Roeder (29) have shown that mung bean nuclease tends to "end-nibble" duplex nucleic acids much less than S1 nuclease. Mung bean nuclease was therefore used to map

the transcriptional initiation sites in P4 DNA.

RNA was extracted from C-1a, C-1a(pCS322) (pCS322 is a derivative of pBR322 with a 6.2 kb BamHI fragment insert from P4 wt), and C-1a infected with P4 for 10 min. The extraction method was that of Baker and Yanofsky (30).

DNA fragments were labeled at 5' ends and strand-separated according to Maxam and Gilbert (24). A total of 200 cps (counts per second) of single-stranded DNA was mixed and precipitated with 10 μ g of RNA and then resuspended into 10 μ l of 0.4 M NaCl, 1 mM EDTA, 40 mM PIPES, pH 6.4. The suspension was incubated in a 65°C waterbath for 12 hours, followed by the addition of 400 μ l of mung bean nuclease buffer (0.05 M sodium acetate, 0.05 M NaCl, 1 mM ZnCl₂, pH 5) containing 10 units (or as specified in Results) of mung bean nuclease. The digestion proceeded at 37°C for 2 hours. The nucleic acids were recovered by precipitation with ethanol and run alongside with the sequencing product of the single-stranded DNA in a DNA sequencing gel.

Detection of P4 Proteins

P4 protein synthesis was detected according to Barrett et al. (31) except that C¹⁴-amino acid mixture (specific activity: 55 mCi/mmole, New England Nuclear) was used instead of C¹⁴-reconstituted algal protein hydrolysate.

RESULTS and DISCUSSION

Shown in Figure 1 are regions that I have sequenced in nine P4 strains. A continuous wild type sequence of 3819 bp was obtained. Based on the completion of sequencing the psu end of P4 DNA by Emily Dale (personal communication), this sequence was determined to extend from nucleotide 573 to nucleotide 4391 of P4 DNA. Due to the fact that the two early genes, ϵ and α (only part of it), and the putative ash gene (17) are located in the right end of this segment and appear to be transcribed in one direction, whereas the three late genes, sid, δ , and psu (only part of it), are located in the left end of this segment and appear to be transcribed in the other direction, this continuous sequence will be presented in two separate figures (Figures 4 and 5) with the numbers being from 2201 to 4391 for the early gene sequence and from 2200 to 573 for the late gene sequence.

The ϵ Gene

The ϵ gene is defined by an amber mutation ϵ aml04 (6). In the DNA of P4 ϵ aml04 ash9, the amber codon (3487, Figure 4) was found. In the open reading frame in phase with the amber codon, there are two ATG codons upstream to the amber codon. However, neither one is preceded by a Shine and Dalgarno sequence. Instead, a GTG codon (3466, Figure 4), which is preceded by a Shine and Dalgarno sequence and leads a run of 95 amino acid residues, appears to be the initiation codon for the ϵ gene.

To identify the ϵ gene product, I used C^{14} -amino acid mixture as label instead of S^{35} -methionine because of the lack of methionine codon in the gene. In the gel shown in Figure 6, a protein of the size predicted for the ϵ protein (10 kd) is seen to be synthesized by P4 ash9. The inability of P4 wt to synthesize this protein in detectable amount is perhaps due to repression of the ϵ gene by the repressor which is believed to be the product of the ash gene (17). Another protein of the size of about 20 kd was also strongly expressed by P4 ash9. It is also visible in P4 wt lane but not in P4 ϵ aml04 ash9 lane. Based on the similarity in size and level of expression, this protein is perhaps the same as one of the two "extra" P4 proteins identified by Barrett et al. (32). Its absence in the ϵ mutant suggests that its expression may depend on the ϵ gene. Also noticeable in this gel is the overproduction of the α protein by P4 ash9. This also may be due to derepression of the α gene caused by the ash mutation.

The ash and cI Mutations

The cI and ash mutations are apparently identical and were both mapped between vir1 and ϵ aml04 (17). The cI405 mutation and five ash mutations are located within a stretch of 33 nucleotides (between 3167 and 3199, Figure 4). An open reading frame capable of coding for a 15 kd protein encompasses this region. A promoter-like sequence was identified in front of this open reading frame by mung bean nuclease mapping (see below). These features suggest the existence of an ash (or cI) gene. The genetic data of Calendar et al. (5) and Lin and Six (17) further suggest that this gene is the repressor gene. However, it has to be pointed out that mutations ash3 and ash8 do

not change the amino acid residues and therefore other possibilities (such as a controlling sequence rather than a gene) must be considered as to how ash mutations affect the gene expression of P4. Also not explainable at present is the finding of a single base deletion in the DNA of P4 vir1 635 ash26 at nucleotide 3386 shown in Figure 4.

The vir1 Mutation.

A base pair change (2543, Figure 4) was found in the DNA of both P4 vir1 and P4 vir1 635 ash26 in the region where the vir1 mutation is mapped (13). The vir1 mutation is known to cause partial insensitivity to P4 immunity (1,16). E. W. Six has shown that the vir1 mutation is cis-acting and causes constitutive expression of the α and ϵ genes but not the ζ gene (unpublished). The vir1 mutation is therefore conceivably an operator or a promoter-up mutation of genes α and ϵ . As shown in Figure 4, the sequence surrounding the vir1 site bears no resemblance to a typical procaryotic operator. The alternative, i.e., a promoter-up mutation, is favorable because (i) vir1 changes a sequence of CAAAAT to TAAAAT, which (the latter sequence), because of its close approach to the consensus sequence, should be a good conventional promoter, and (ii) the virulence suppression phenomenon exhibited by an insertional ash mutation can be rather satisfactorily explained, as will be discussed in more detail elsewhere (17).

The Replication Origin

The replication origin of P4 has been mapped to the vicinity of the vir1 site (15). I found a region of about 400 nucleotides in the vicinity of the vir1 site to be rich in AT (65%) and contain several dyad symmetries and tandem repeats and a sequence (TTTTATATTTT, 2220, Figure 4) homologous to the yeast ori consensus sequence (33). These unusual features suggest that this region is the replication origin of P4. Moreover, a stretch of 6 C's which occurs 25 nucleotides left of the vir1 site may be used by the α protein as template to initiate DNA replication (32) since P4 α protein has a transcriptional activity specific for poly-(dC) as template (34).

Part of the α Gene

About 330 nucleotides downstream from the termination codon

GCTTACTGTTTGTGTTTATATTTACTGTTACCTCTGTTACCTTAATAAAAAAGATAAGTAATACAGTAAGTTAAAGGGTGAACAAT 2291
 CGCAGTAATGACTGTTACCGTCTGTTACCACTGTTACCCGTTGATGGGCTTTTTGTGCTGTTTACTACTGTTTGCCTTTTATTAATTCG 2382
 CCGGGAATGAATAAGAAGAACGATTGTTATTTACTATAAAATTTACAAATGCTTTAAGTGGCTTTAAATGACTTTAAGTGAGATGAA 2473
 T vlr1 2564
 CAAAAACACACAGCCATTGTAAGACAGCCTGAACAAATCCCCCTGTTGCGTCTGCTGAAAAATATTCACAAAATAAGCGCTACCCGAAG 2644
 CCGGACGGACTTATCCGGTCTGTATGGACATTAACGAGGTAGCCG orf
 MET Gln Ala Val Phe Ser Ser Pro Ser Pro Ala 2713
 CCA GTG ACG CCA CTG ATG CCG CTG CCG GAC ATC ACG CAG GAG CGT TTT TTA CGT GTG CCG GAA GTG ATG 2782
 Pro Val Thr Pro Leu Met Pro Leu Pro Asp Ile Thr Gln Glu Arg Phe Leu Arg Val Pro Glu Val Met
 CAC CTG TGC GGC CTG TCA CGC TCG ACC ATC TAC GAA CTC ATC CGT AAG GGG GAA TTT CCG CCG CAG GTG 2851
 His Leu Cys Gly Leu Ser Arg Ser Thr Ile Tyr Glu Leu Ile Arg Lys Gly Glu Phe Pro Pro Gln Val
 AGT CTT GGC GGT AAA AAT GTG GCC TGG CTG CAC TCT GAA GTC ACC GCA TGG ATG GCC GGA CGC ATT GCC 2920
 Ser Leu Gly Gly Lys Asn Val Ala Trp Leu His Ser Glu Val Thr Ala Trp Met Ala Gly Arg Ile Ala
 GGC GCA AAC GGG GGT ACG ACG CAT GAT GAT GCC CGC TCT GCA AAA ACT CCC TTT TTC TGG CTT GCC TTT 3004
 Gly Ala Asn Gly Gly Thr Thr His Asp Asp Ala Arg Ser Ala Lys Thr Pro Phe Phe Trp Leu Ala Phe
 TTC CGG CAT TTG CGG ATA TAG TTTTTGCGCTGCCGCAAAATCGGACGGCGGTGAGAACCCGAGCAATTCACAGCGACATA 3086
 Phe Arg His Leu Arg Ile * ash
 TGACGCGCCACGCGTCTTTTTTGTGTCGCAATCAATGCCACAGAGCGCCAGATT ATG GTG TGG TGT GTG GTC AGT CGT GCA 3155
 MET Val Trp Cys Val Val Ser Arg Ala
 GAT GGC ATT CCG TGC ATC CTG CCC GCG TCC GCT CAT TAT GCG GCA GAG TCA ATG GTA GCT CAG GCG GGG 3224
 Asp Gly Ile Pro Cys Ile Leu Pro Ala Ser Ala His Tyr Ala Ala Glu Ser Met Val Ala Gln Ala Gly
 A ash3 ash8
 CAG CCT CCG GGC TGG CCG GTA TCC TGT GAA GCC GGT ATT CTC ACC CCC GTC TGG GCT ATC GCC ATC GAG 3293
 Gln Pro Pro Gly Trp Pro Val Ser Cys Glu Ala Gly Ile Leu Thr Pro Val Trp Ala Ile Ala Ile Glu
 CGT GAG AAC TCC GGC GAT AGC GTT ATT TGC TAT TCA CAG GAG GCT GCC ATC ATG GCT ACA ACC CTT ACC 3362
 Arg Glu Asn Ser Gly Asp Ser Val Ile Cys Tyr Ser Gln Glu Ala Ala Ile Met Ala Thr Thr Leu Thr
 CCC TCA CAC CCT GAA TTT GTC TTT GTG TTT GCG GCT GTC CGT CGC GCA GAC CGT CAT CCC CGT ATC TGT 3431
 Pro Ser His Pro Glu Phe Val Phe Ala Ala Val Arg Arg Ala Asp Arg His Pro Arg Ile Cys
 ATG CTT CGC ACC GTC GCC GGT GAT GAA CGC AGT GCC CGC CGT TCC CTT GTC CGT GAC TAT GTG CTC TCC 3498
 Met Leu Arg Thr Val Ala Gly Asp Glu Arg Ser Ala Arg Arg Ser Leu Val Arg Asp Tyr Val Leu Ser
 CTT GCT GCC CGT CTG CCG GTG GTG GAG GTG TCC CGT GCG TAA T AAA AAA GCC CCT CAG ACC GTC TCA 3567
 Leu Ala Ala Arg Leu Pro Val Val Glu Val Ser Arg Ala * Lys Lys Ala Pro Gln Thr Val Ser
 Val Arg Asn
 GCG CGT CAT GAC GCC CGT GAA CAC CTC AGC ATT GAG GCT TAC CAT AAG CTC AAC CGC GCC AGC GCC GTA 3636
 Ala Arg His Asp Ala Arg Glu His Leu Ser Ile Glu Ala Tyr His Lys Leu Asn Arg Ala Ser Ala Val
 TCC CCG TTT GTT GGG GGT GAT TTG ATT CAC CGT GAA CTC TCC GGC CTG CAT CAG CTC TAC ATT CCA CAT 3705
 Ser Arg Phe Val Gly Gly Asp Leu Ile His Arg Glu Leu Ser Gly Leu His Gln Leu Tyr Ile Pro His
 ATT TTC AGC TAC CTG AAT GAA GAT ATT GAT TTT GTG CTG AAT GAG CTG AAA GCC AAA GGC CTG TGC CGC 3772
 Ile Phe Ser Tyr Leu Asn Glu Asp Ile Asp Phe Val Leu Asn Glu Leu Lys Ala Lys Gly Leu Cys Arg
 GAT TTT CTC GCC CAG CAG AAA GAC CCG GGA GAC AGC CAT GTT TGA T TTT CCC CAG CCC GGT GAG 3841
 Asp Phe Leu Ala Gln Gln Lys Asp Arg Gly Asp Arg Thr His Val * Phe Pro Gln Pro Gly Glu
 MET Phe Asp
 ATT TAC CGT TCT GCC GGT TTT CCC GAT GTG GCC GTG GTC GGC ATT CTG GAA GAC GGT ATT CCG TGG GAA
 Ile Tyr Arg Ser Ala Gly Phe Pro Asp Val Ala Val Val Gly Ile Leu Glu Asp Gly Ile Pro Trp Glu

```

3910
ATG CCG TAC CGC TGC CCG GAC ATT GTC TGG AAC CCG TAC CGC CGT AAA TTC AGT ATC CTT GTG CGT ATC
Met Pro Tyr Arg Cys Pro Asp Ile Val Trp Asn Pro Tyr Arg Arg Lys Phe Ser Ile Leu Val Arg Ile
3979
CTC GCT GAC GGA CGC ACC ACA GAC ATC CCG CTG GGG CGT TTT CTG CGG GAA TTT ACC TGT GAC CGT CCT
Leu Ala Asp Gly Arg Thr Thr Asp Ile Pro Leu Gly Arg Phe Leu Arg Glu Phe Thr Cys Asp Arg Pro
4048
GAC CTG TTT AAA CGC AGC CCC GTA AAC CCG CAT GCG GTA CTG AAA GAA ATG GCC GGA GAC CCG GAA TTA
Asp Leu Phe Lys Arg Ser Pro Val Asn Arg His Ala Val Leu Lys Glu Met Ala Gly Asp Pro Glu Leu
4117
CAG AAA TGG CCG GAA AAA TAT CTG GAT ATT TAC CCG CAG GAC ACT GTT CCG GTC AGC CGG GCA GCA CCG
Gln Lys Trp Arg Glu Lys Tyr Leu Asp Ile Tyr Pro Gln Asp Thr Val Pro Val Ser Arg Ala Ala Pro
4186
GTG GCG CCG GAA TGG CCG GAA ATT CCC CGC ACG GAG CCT GAC CCG GAA ACC ACC CCG GAT AAC AGT TAC
Val Ala Arg Glu Trp Arg Glu Ile Pro Arg Thr Glu Pro Asp Pro Glu Thr Thr Pro Asp Asn Ser Tyr
4272
CGC AAT TAT CTG TAA TAAAAACGACCCCGAAAAATTAATGTGCGTATTGCGCAGGGATACGCACGTCTTCAGGAGACGCAG
Arg Asn Tyr Leu *
-10
ATATGCCTTATCAGTTAATGCAACCGGACGGAATGCAGTCATCTGTACAGGGAGAAAAGCAA ATG AAA ACA CCC TTA CCG
4345
MET Lys Thr Pro Leu Pro
4391
CCC GTC TTA CGC GCT GCC CTT TAC CGT CGC GCT GTC G
Pro Val Leu Arg Ala Ala Leu Tyr Arg Arg Ala Val

```

Fig. 4. DNA sequence of P4 early genes. The sequence is shown 5'→3' with nucleotide number 1 referring to the far left nucleotide of the 5' end of P4 DNA oriented as that in Fig.1. Points of interest in this sequence are marked for the consensus years ori sequence, → for direct repeats, ---> for dyad symmetries, -10 for the consensus -10 sequence, and — for the beginning of genes. Some genes are overlapping and their deduced amino acid sequences cannot be printed in the same line. Wherever this occurs, the first few amino acids of the succeeding gene are printed one line lower. Mutations ash9, ash8, and cI405 are next to each other and in this order. Ash10 is an insertional mutation. Ash26 has a mutation similar to ash10 plus a deletion at 3386.

identified for the ϵ gene, an ATG codon (4337, Figure 4), preceded by a stretch of purine nucleotides, leads a short open reading frame to the end of the sequence. About 140 nucleotides upstream from this ATG codon there is a sequence TGTAAT, which is quite similar to the consensus -10 sequence, and 30 nucleotides behind this sequence there is one dyad symmetry. Although the open reading frame led by this ATG codon is too short to assert that it is the initiation codon of the α gene, these special sequence features suggest such a possibility.

The sid Gene

The sid gene product has been shown to be P4 gp 3 of molecular weight 27 kd (35). I have found an open reading frame in a region where the sid gene is mapped. It is led by an ATG codon (2059, Figure 5), which is preceded by a Shine and Dalgarno

2110
 GGTTTTCTCTGCGACCTTTTCTGGCTTCCGGTCTGAGGATGAGTCTCTGTGTGACGGGTGGCACATCTGCAATGCGTGTGTTGTG
 -10
 CCGGTGTACGT CACAA TTTCTTAACCTGAAGTGACGAGGAGCCGAAA sid 2030
 MET Ser Asp His Thr Ile Pro Glu Tyr Leu
 1961
 CAA CCC GCA CTG GCA CAA CTG GAA AAG GCC AGA GCC GCC CAT CTT GAG AAC GCC CGC CTG ATG GAT GAG
 Gln Pro Ala Leu Ala Gln Leu Glu Lys Ala Arg Ala Ala His Leu Glu Asn Ala Arg Leu Met Asp Glu
 1886
 ACC GTC ACG GCC ATT GAA CGG GCA GAG CAG GAA AAA AAT GCG CTG GCG CAG GCC GAC GGA AAC GAC GCT
 Thr Val Thr Ala Ile Glu Arg Ala Glu Gln Glu Lys Asn Ala Leu Ala Gln Ala Asp Gly Asn Asp Ala
 1823
 GAC GAC TGG CGC ACG GCC TTT CGT GCA GCC GGT GGT GTC CTG AGC GAC GAG CTG AAA CAG CGC CAC ATT
 Asp Asp Trp Arg Thr Ala Phe Arg Ala Ala Gly GGT Gly Val Leu Ser Asp Glu Leu Lys Gln Arg His Ile
 1754
 GAG CGC GTG GCA CGC CGG GAG CTG GTA CAG GAA TAT GAC AAT CTG GCC GTG GTG CTG AAT TTC GAA CGT
 Glu Arg Val Ala Arg Arg Glu Leu Val Gln Glu Tyr Asp Asn Leu Ala Val Val Leu Asn Phe Glu Arg
 1685
 GAA CGC CTG AAA GGG GCG TGT GAC AGC ACG GCC ACC GCC TAC CGG AAG GCA CAT CAT CAC CTT CTG AGT
 Glu Arg Leu Lys Gly Ala Cys Asp Ser Thr Ala Thr Ala Tyr Arg Lys Ala His His His Leu Leu Ser
 1616
 CTG TAT GCA GAG CAT GAG CTG GAA CAC GCC CTG AAT GAA ACC TGT GAG GCG CTT GTC CGG GCA ATG CAT
 Leu Tyr Ala Glu His Glu Leu Glu His Ala Leu Asn Glu Thr Cys Glu Ala Leu Val Arg Ala Met His
 1547
 CTG AGC ATT CTG GTA CAG GAA AAT CCG CTC GCC AAC ACC ACC GGC CAT CAG GGC TAC GTC GCA CCG GAA
 Leu Ser Ile Leu Val Gln Glu Asn Pro Leu Ala Asn Thr Thr Gly His Gln Gly Tyr Val Ala Pro Glu
 1478
 AAG GCT GTC ATG CAG CAG GTG AAA TCA TCG CTG GAA CAG AAA ATT AAA CAG ATG CAA ATC AGC CTC ACC
 Lys Ala Val Met Gln Gln Val Lys Ser Ser Leu Glu Gln Lys Ile Lys Gln Met Gln Ile Ser Leu Thr
 sid1 T1409
 GGC GAG CCT TGG CTC CGG CTG ACC GGA CTG TCA GCG GCA ACA CTC CCG CAC ATG GAT TAT GAG GTG GCA
 Gly Glu Pro Trp Leu Arg Leu Thr Gly Leu Ser Ala Ala Thr Leu Pro His Met Asp Tyr Glu Val Ala
 T sid1 1340
 GGC ACA CCG GCA CAG CGC AAG GTG TGG CAG GAC AAA ATA GAC CAG CAG GGA GCA GAG CTT AAG GCC AGA
 Gly Thr Pro Ala Gln Arg Lys Val Trp Gln Asp Lys Ile Asp Gln Gln Gly Ala Glu Leu Lys Ala Arg
 1272
 GGG CTG CTG TCA TGA TT TAC TGT CCG TCG TGT GGA CAT GTT GCT CAC ACC CGT CGC GCA CAT TTC ATG
 Gly Leu Leu Ser * Tyr Cys Pro Ser Cys Gly His Val Ala His Thr Arg Arg Ala His Phe Met
 MET Ile
 1203
 GAC GAT GGC ACC AAG ATA ATG ATT GCA CAG TGC CGG AAT ATT TAT TGC TCT GCG ACA TTT GAA GCG AGT
 Asp Asp Gly Thr Lys Ile Met Ile Ala Gln Cys Arg Asn Ile Tyr Cys Ser Ala Thr Phe Glu Ala Ser
 1134
 GAA AGC TTT TTC TCT GAC AGT AAA GAT TCA GGA ATG GAA TAC ATT TCA GGC AAA CAG AGA TAC CGC GAT
 Glu Ser Phe Phe Ser Asp Ser Lys Asp Ser Gly Met Glu Tyr Ile Ser Gly Lys Gln Arg Tyr Arg Asp
 1065
 TCA CTG ACG TCA GCC TCC TGC GGT ATG AAA CGC CCG AAA AGA ATG CTT GTT ACC GGA TAT TGT TGT CGG
 Ser Leu Thr Ser Ala Ser Cys Gly Met Lys Arg Pro Lys Arg Met Leu Val Thr Gly Tyr Cys Cys Arg
 A 835
 AGA TGT AAA GGC CTT GCA CTG TCA AGA ACA TCG CGG CGT CTG TCT CAG GAA GTC ACC GAG CGT TTT TAT
 Arg Cys Lys Gly Leu Ala Leu Ser Arg Thr Ser Arg Arg Leu Ser Gln Glu Val Thr Glu Arg Phe Tyr
 927
 GTG TGC ACG GAT CCG GGC TGT GGT CTG GTG TTT AAA ACG CTT CAG ACC ATC AAC CGC TTC ATT GTC CGC
 Val Cys Thr Asp Pro Gly Cys Gly Leu Val Phe Lys Thr Leu Gln Thr Ile Asn Arg Phe Ile Val Arg
 CCG GTC ACG CCG GCG AAC TGG CAG AAC GCC TGC ATG AAA AAC AGG AAC TGC CGC CAG TAC GGT TAA AAA
 Pro Val Thr Pro Ala Asn Trp Gln Asn Ala Cys Met Lys Asn Arg Asn Cys Arg Gln Tyr Gly *
 767
 CACAATCATATTCTGCTGCTGGAATGAGGCTGCCGGTTAACACCGCGCTGCCGCACACCGTATTTTATTCTTCAGCATGATGAGA
DSU
 AAGAGATAACG ATG GAA AGC ACA GCC TTA CAG CAG GCC TTT GAC ACC TGT CAG AAT AAC AAG ACA GCA TGG
 MET Glu Ser Thr Ala Leu Gln Gln Ala Phe Asp Thr Cys Gln Asn Asn Lys Thr Ala Trp
 693
 CTG CAA CGC AAA AAT GAG CTG GCA GCG GCC GAA CAG GAA TAT CTG CGG CTT CTG TCA GGA GAA GGC AGA
 Leu Gln Arg Lys Asn Glu Leu Ala Ala Ala Glu Gln Glu Tyr Leu Arg Leu Leu Ser Gly Glu Gly Arg
 624
 AAC GTC AGT CGC CTG GAC GAA TTA CGC AAT ATT ATC GAA GTC AGA AAA TGG CAG
 Asn Val Ser Arg Leu Asp Glu Leu Arg Asn Ile Ile Glu Val Arg Lys Trp Gln
 573

sequence, and is capable of coding for a 27 kd protein. I have also found two base pair changes in the DNA of P4 vir1 sid1 within the sequence of this open reading frame (within the HindIII-HaeII fragment in which Kahn et al. (13) mapped the sid1 mutation). One (1408, Figure 5) is at the third base of a GTG codon with no change of the amino acid residue. The other (1354, Figure 5) changes a glutamine (GAG) codon to an amber codon. This latter finding is consistent with the observation by Lee (35) that P4 vir1 sid1 is an amber mutant.

The δ Gene

The δ gene product has not been identified. I have found an open reading frame in a region where the δ gene is mapped. It begins with an ATG codon (1328, Figure 5), which is preceded by a Shine and Dalgarno sequence, and is capable of coding for a 17 kd protein. I have also found a base pair change (1049, Figure 5) in the DNA of P4 vir1 δ 35 ash26 in the sequence of this open reading frame (within the BamHI-HindIII fragment in which Kahn et al. (13) mapped the δ 35 mutation). The termination codon identified for the sid gene overlaps with the initiation codon for the δ gene. The overlapping sequence, ATGA, is the most common overlapping sequence found in λ phage DNA (36) and overlapping genes are often translationally coupled (37,38). Both translational coupling and transcriptional polarity (due to the sid1 amber mutation) may explain why P4 vir1 sid1 is phenotypically δ^- (13,35).

The psu Gene

About 105 nucleotides downstream from the termination codon identified for the δ gene, an ATG codon, which is preceded by a stretch of purine nucleotides, leads an open reading frame to the end of the sequence. Emily Dale (personal communication) has recently completed the psu gene sequence and agreed with the choice of this codon to be the initiation codon. There are a few dyad symmetries at the end of the δ gene and in front of the psu

Fig. 5. DNA sequence of P4 late genes. The sequence is shown 5'-->3' with nucleotide number 1 referring to the far left nucleotide of the 5' end of P4 DNA oriented as that in Fig. 1. It is opposite to the sequence of Fig. 4 in strand polarity. Points of interest and amino acid sequences are marked and arranged in the same way as Fig. 4.

Fig. 6. A 10-20% gradient SDS-polyacrylamide gel showing P4 proteins at 20 minutes after infection.

gene. They may relate to the observation that δ mutants are partially psu^- (10,13).

Two Possible Coding Sequences

There are two open reading frames (orf) in the early gene sequence with which no genes have been identified. One begins at

Fig. 7. P4 DNA fragments used for mung bean nuclease mapping. Restriction sites are labeled as A for *Ava*I, H for *Hin*I, and T for *Tag*I with numbers corresponding to P4 DNA sequence. The *vir*1 and *ε*aml04 mutations are labeled as v and ε, respectively.

2612 of Figure 4 and encodes 109 amino acid residues; the other begins at 3746 and encodes 151 amino acid residues. Considering that many aspects regarding the regulation of early genes are not yet understood, these two possible coding sequences require attention.

Possible Promoters

I used mung bean nuclease mapping to look for promoters of the early genes. In total, 14 probes were prepared from 7 P4 *wt* DNA fragments that are shown in Figure 7. The 2 probes from fragment 2 were tested only with RNA of C-la(pCS322) and only the slow strand ("slow" and "fast" refer to the migration pattern of the complementary strands of a DNA fragment in a strand-separation gel) yielded positive mapping results. Two possible transcriptional initiation sites were detected, as indicated by the arrows in Figure 8a. Ten nucleotides upstream from the lower site there is a sequence whose complementary sequence is homologous to the consensus -10 sequence. In Figure 4 this -10 sequence is marked (3125). This possible transcriptional initiation site was also detected by the slow strand of fragment 5 with RNA of C-la(pCS322), P4 *wt*-infected C-la, and P4 *vir*1-infected C-la (Figure 8b). RNA of C-la(pCS322) gave a much stronger protection signal than RNA of the other two cases, indicating that the possible promoter is perhaps used more efficiently in pCS322 than in the phage. RNA of P4 *vir*1 *ash*26 did not protect the probe; this is not unexpected since *ash*26 is a single base pair insertion and consequently the mRNA of the *ash*

Fig. 8. Mung bean nuclease mapping results. (a) Lanes G, AG, C and CT display the sequencing products of the slow strand of fragment 2. Lanes 1 to 4 are the products of the probe hybridized with the RNA of C-1a(pCS322) and treated with the following amounts of mung bean nuclease, respectively: (1) 16 units, (2) 8 units, (3) 4 units, (4) 2 units. The two arrows indicate the two possible transcription initiation sites. The region marked as -10 is complementary to the consensus -10 sequence. (b) The probe was the slow strand of fragment 5. Lane 1 is the untreated probe. Lanes 2 to 6 are the products of the probe hybridized with RNA of C-1a, C-1a(pCS322), P4 *wt*-infected C-1a, P4 *vir1*-infected C-1a, and P4 *vir1 ash26*-infected C-1a, respectively, and treated with 10 units of mung bean nuclease. (c) Lanes AG and CT display the sequencing products of the fast strand of fragment 4. Lanes 1 to 6 are the same as those described in (b) except that the probe is different.

region has an extra nucleotide which is subjected to attack by mung bean nuclease.

Since the upper site in Figure 8a is located in a crowded region, shorter probes were made from fragments 4 and 6. Only the fast strand of fragment 4 and the slow strand of fragment 6 yielded positive results. These two strands are of the same strand of P4 DNA and detected the same possible transcriptional initiation site, which is shown in Figure 8c. Ten nucleotides upstream from this site there is a sequence whose complementary sequence is homologous to the consensus -10 sequence. In Figure 4 this -10 sequence is marked (2934). No difference was found between C-la(pCS322) and C-la infected with P4 wt, P4 vir1, or P4 vir1 ash26 in this case.

For the late genes, there is a possible promoter in front of the sid gene with the -10 sequence being CACAAT (2023, Figure 5) (suggested by B. Lindqvist). S1 mapping results of Dale, Christie and Calendar (personal communication) agree with this idea.

ACKNOWLEDGMENT

This work is part of the author's Ph.D. thesis under the supervision of E. W. Six with support from NIH grant AI04043. The protein labeling experiment was done in R. Calendar's laboratory at U.C. Berkeley and the M13 DNA sequencing was done at the Linus Pauling Institute. The author is very grateful to R. Calendar for providing the protein labeling facilities and commenting on the manuscript.

*Present address: Linus Pauling Institute of Science and Medicine, 440 Page Mill Road, Palo Alto, CA 94306, USA

REFERENCES

1. Six, E.W. and Klug, C.A.C. (1973) *Virology* 51, 327-344.
2. Six, E.W. (1975) *Virology* 67, 249-263.
3. Gibbs, W., Goldstein, R., Weiner, R., Lindqvist, B.H. and Calendar, R. (1973) *Virology* 53, 24-39.
4. Six, E.W. and Lindqvist, B. (1978) *Virology* 87, 217-230.
5. Calendar, R., Ljungquist, E., Deho, G., Usher, D.C., Goldstein, R., Youderian, P., Siorni, G. and Six, E. W. (1981) *Virology* 113, 20-38.
6. Geisselsoder, J., Youderian, P., Deho, G., Chidambaram, M. and Goldstein, R. (1981) *J. Mol. Biol.* 148, 1-19.

7. Souza, L., Calendar, R., Six, E.W. and Lindqvist, B.H. (1977) *Virology* 81, 81-89.
8. Shore, D., Deho, G., Tsipis, J. and Goldstein, R. (1978) *Virology* 29, 106-125.
9. Sunshine, M.G., Six, E.W., Barrett, K. and Calendar, R. (1976) *J. Mol. Biol.* 106, 673-682.
10. Sauer, B., Ow, D., Ling, L. and Calendar, R. (1981) *J. Mol. Biol.* 145, 29-46.
11. Inman, R., Schnos, M., Simon, L., Six, E.W. and Walker, D. (1971) *Virology* 44, 67-72.
12. Souza, L., Geisselsoder, J., Hopkins, A. and Calendar, R. (1978) *Virology* 85, 335-341.
13. Kahn, M., Ow, D., Sauer, B., Rabinowitz, A. and Calendar, R. (1980) *Molec. Gen. Genet.* 177, 399-412.
14. Harris, J.D. and Calendar, R. (1978) *Virology* 85, 343-358.
15. Krevolin, M.D., Inman, R., Roof, D., Kahn, M. and Calendar, R. (Submitted to *J. Mol. Biol.*)
16. Lindqvist, B.H. and Six, E.W. (1971) *Virology* 43, 1-7.
17. Lin, C.-S. and Six, E.W. (Manuscript in preparation)
18. Sasaki, I. and Bertani, G. (1965) *J. Gen. Microbiol.* 40, 365-376.
19. Sunshine, M.G., Thorn, M., Gibbs, W., Calendar, R. and Kelly, B. (1971) *Virology* 46, 691-702.
20. Geisselsoder, J., Chidamparam, M. and Goldstein, R. (1978) *J. Mol. Biol.* 126, 447-456.
21. Lindqvist, B.H. and Sinsheimer, R.L. (1967) *J. Mol. Biol.* 28, 87-94.
22. Davis, R.W., Botstein, D. and Roth, J.R. (1980) In *Advanced Bacterial Genetics* (Cold Spring Harbor Laboratory) 116-125.
23. Kahn, M. and Hopkins, A. (1978) *Virology* 85, 359-363.
24. Maxam, A.M. and Gilbert, W. (1980) *Methods Enzymol.* 65, 499-560.
25. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA.* 74, 5463-5467.
26. Messing, J. and Vieira, J. (1982) *Gene* 19, 269-276.
27. Sollner-Webb, B. and Reeder, R.H. (1979) *Cell* 18, 485-499.
28. Berk, A.J. and Sharp, P.A. (1977) *Cell* 12, 721-732.
29. Green, M.R. and Roeder, R.G. (1980) *Cell* 22, 231-242.
30. Baker, R.F. and Yanofsky, C. (1968) *Proc. Natl. Acad. Sci. USA* 60, 313-320.
31. Barrett, K.J., Marsh, M.L. and Calendar, R. (1976) *J. Mol. Biol.* 106, 683-707.
32. Barrett, K.J., Blinkova, A. and Arnold, G. (1983) *J. Virol.* 48, 157-169.
33. Broach, J.R., Li, Y.-Y., Feldman, J., Jayaram, M., Abraham, J., Nasmyth, K. A. and Hicks, J. B. (1982) *Symposia on Quantitative Biology XLVII*, 1165-1173.
34. Barrett, K.J., Gibbs, W. and Calendar, R. (1972) *Proc. Natl. Acad. Sci. USA* 69, 2986-2990.
35. Lee, S.-J. (1981) Bachelor thesis, Harvard College.
36. Sanger, F., Coulson, A.R., Hong, G.f., Hill, D.F. and Petersen, G.B. (1982) *J. Mol. Biol.* 162, 729-773.
37. Oppenheim, D.S. and Yanofsky, C. (1980) *Genetics* 95, 758-795.
38. Schumperli, D., McKenney, K., Sobieski, D.A. and Rosenberg, M. (1982) *Cell* 30, 865-871.