Nucleotide sequence of the genes III, VI and I of bacteriophage M13

P. van Wezenbeek and J. G. G. Schoenmakers

Laboratory of Molecular Biology, University of Nijmegen, Nijmegen, Netherlands

Received 12 January 1979

ABSTRACT

A DNA region of 2750 base pairs encompassing the genes III, VI and I of bacteriophage M13 has been sequenced by the Maxam-Gilbert procedure. By establishing the nucleotide changes introduced by several amber mutations, the coding region and the regulatory signals of each gene have been deduced. The genes appear to span 1275 base pairs (gene III; mol. wt. 44,748) 339 base pairs (gene VI; mol. wt. 12,264) and 1047 base pairs (gene I; mol. wt. 39,500). Their separating non-codogenic regions are extremely short, namely two and one base pair, respectively. The C-terminal end of gene I, however, intrudes 23 nucleotides into gene IV. From the nucleotide sequence it appears that the minor capsid protein of the phage, which is encoded by gene III, is synthesized in a precursor form containing 18 extra amino acids at its N-terminal end. Furthermore, in this capsid protein two clusters of a fourfold repeat of the sequence Glu-Gly-Gly-Gly-Ser are apparent. Gene VI appears to code for a small, extremely hydrophobic polypeptide. Its total hydrophobic amino acids content of 51% suggests that this protein can only function in the host cell membrane.

INTRODUCTION

The genome of the male-specific filamentous coliphage M13 is a single-stranded circular DNA of about 6400 bases which codes for at least eight genes (1). In the infected cell this DNA is rapidly converted into a double-stranded circularly closed replicative form (RF) which then acts as a template for viral gene expression and production of progeny phage DNA.

In the last few years the mechanisms of these processes and their regulatory control have been studied extensively. The conversion of single-stranded viral DNA into RF and the specific role of the host proteins and phage-encoded products in this replication process are now reasonably well understood (reviewed in ref. 2). In addition, the nucleotide sequence of the replication origin, its secondary structure and the replication starts and stops have been elucidated (3-6). Progress has also been made in unraveling the transcription process in vitro (7) and in vivo (8,9). Transcription of
replicative form DNA occurs in a cascade-like fashion as a result of initiation of RNA synthesis at many different promoter sites and termination of transcription at only a single site (10-15). Since transcription proceeds in only one direction along the circular genome, a gradient of gene transcripts is formed which increases towards the central termination site.

The region distal to this site, encompassing the genes III, VI and I, is of special interest since there is suggestive evidence that these genes form an operon which only can be transcribed from a promoter immediately preceding gene III (16). In addition, Konings et al. (17) have evidenced that the minor capsid protein of the virion, which is encoded by gene III, is smaller than the direct translational product of this gene. This suggests that gene III-protein, as has been shown for the gene VIII-encoded major capsid protein (17,34), is subject to posttranslational processing. The biological function of the two other structural genes VI and I are unknown, although the synthesis of a protein (mol.wt. 36,000) encoded by gene I has clearly been demonstrated in DNA-dependent protein synthesizing systems (17, 18). The synthesis of gene VI-protein, however, has not yet been detected either in vitro (17-19) or in vivo (20,21).

To elucidate the structural features of gene VI and to get a better understanding of the control mechanism(s) which are pertinent to this gene and its adjacent genes III and I we have established the nucleotide sequence of this segment of the M13 DNA genome. DNA sequencing analysis was performed by the Maxam-Gilbert method (25). By establishing the nucleotide changes introduced by several amber mutations in each gene, their coding regions could be determined and their regulatory regions deduced.

MATERIALS AND METHODS

Materials

The restriction endonucleases R.Hap II, R.Hae III and R.Alu I were prepared as described previously (22). R.Hinf I was isolated by a slight modification of the method of Roberts et al. (23), R.Hha I and R.Mbo I were purchased from New England Biolabs, Beverly, Mass., R.BamH I from Boehringer, Mannheim and R.Taq I was from Microbiol. Res. Establishment, Porton Down. T4 DNA polymerase was prepared according to the method of Goulian et al. (24). E.coli DNA polymerase (Klenow) was obtained from Boehringer. Polynucleotide kinase from T4-infected E.coli cells was purchased from P.L.Biochemicals, and bacterial alkaline phosphatase (BAPF) from Worthington B.C. Hydrazine
was from Eastman Kodak Co. and dimethylsulphate from Aldrich Co. (Y-32P)-ATP (spec.act. >1000 Ci/mmol) was prepared by the procedure of Glynn and Chappell (37).

**M13 DNA and Restriction Fragments**

The procedures for the propagation and purification of wild-type and amber mutant M13 phages and the preparative methods for the isolation of single-stranded viral DNA and of circularly closed double-stranded M13 RF have been described elsewhere (26,27). The preparation and purification of the various restriction fragments of M13 RF was performed according to the procedure described by Van den Hondel *et al.* (26).

**Labeling of fragments with 32P at a single 5'-OH terminus**

The 5'-ends of restriction fragments were dephosphorylated with bacterial alkaline phosphatase essentially as described by Maxam and Gilbert (25). Labeling of the 5'-OH ends of fragments was performed with (Y-32P)-ATP and polynucleotide kinase (25). The dephosphorylated fragments (3-4 pmol) were dissolved in 45 μl of 10 mM glycine-NaOH, pH 9.5, 1 mM spermidine, 0.1 mM EDTA. The fragments were denatured by heating at 100° for 3 min, then quickly chilled and transferred to an Eppendorf tube containing 100 pmol of dried (Y-32P)-ATP. After addition of 5 μl of 0.5 M glycine-NaOH, pH 9.5, 0.1 M MgCl₂, 50 mM dithiothreitol, the phosphorylation was started by adding 2-3 units of polynucleotide kinase. After 30 min at 37° the reaction was terminated with phenol. Carrier tRNA (10 μg) was added and after two extractions with phenol the labeled fragments were precipitated with ethanol. The precipitate was dissolved in 70 μl of 10 mM Tris-HCl, pH 7.6, 1 mM EDTA (buffer A), the solution was heated at 100° for 3 min and the DNA fragments were renatured by incubation at 67° for 2 h. Thereafter the appropriate restriction enzyme and restriction enzyme buffer was added and the volume adjusted to 100 μl with buffer A. After a digestion at 37° for 2 h the fragments labeled on only one end were separated on 5% polyacrylamide gels which were prepared in 40 mM Tris, 20 mM sodium acetate, 2 mM EDTA, pH 7.8. After electrophoresis and autoradiography the bands containing the 32P-labeled fragments were excised and the DNA was eluted as described by Van den Hondel *et al.* (26).

**DNA sequencing methods**

The nucleotide sequence of 5'-32P-labeled restriction fragments was
determined by the method of Maxam and Gilbert (25). Purine residues were partially methylated by dimethylsulphate. Cleavage at guanine residues was obtained by heating at neutral pH and subsequent treatment with 0.1 N alkali at 90°. Preferential cleavage at adenine residues was achieved by treatment with 0.1 N HCl followed by treatment with 0.1 N alkali at 90°. Cleavage at cytosine and thymine was obtained by partial hydrazinolysis followed by treatment with 0.5 M piperidine. Hydrazinolysis at thymine residues was suppressed by the presence of 2 M NaCl.

The products of the individual cleavage reactions were separated by electrophoresis on 15%, 20% and 25% polyacrylamide slab gels (40 cm x 25 cm x 0.1 cm) which were prepared in 50 mM Tris-borate, pH 8.3, 1 mM EDTA and 7 M urea (acrylamide/bisacrylamide ration 30:1).

The "plus and minus" method for DNA sequencing was essentially as described by Sanger and Coulson (28). Routinely, DNA hybrids were prepared by annealing denatured M13 DNA restriction fragments (2-3 pmol) with intact single-stranded viral M13 DNA (0.5 pmol). The annealed reaction mixtures (10-20 μl) were then used for limited DNA synthesis with DNA polymerase I and (α-32P)-dATP as the radioactive precursor. Digestions with restriction enzymes were carried out before the plus and minus reactions. For the plus and minus reactions and the electrophoretic separation of the reaction products the detailed procedure of Air et al. (31) was followed.

RESULTS

Nucleotide sequence of the region covering the genes III, VI and I

The dimethylsulphate-hydrazine DNA sequencing procedure of Maxam and Gilbert (25) has been applied to restriction fragments covering the genes III, VI and I. These genes have been mapped next to each other and are located distal to the central, rho-independent termination site of transcription (Fig. 1).

A detailed restriction enzyme cleavage map of this region is shown in Fig. 1. It is clear that this map provides sufficient cleavage sites at distances less than 200-250 base pairs for establishing an unambiguous sequence of this part of the M13 DNA genome. Several sets of successive and overlapping restriction fragments were labeled at their 5'-hydroxy termini with (γ-32P)-ATP and polynucleotide kinase. Each fragment was then cleaved with the appropriate restriction endonuclease to produce fragments with a single 5'-labeled end. After electrophoretic separation, each labeled...
Figure 1: Genetic map of bacteriophage M13 and the restriction enzyme cleavage map of the DNA region encompassing the genes III, VI and I. The symbol T stands for the rho-independent termination site of transcription. IG refers to the intergenic region in which the origin of replication of viral and non-viral DNA strands is located. Regions which have been sequenced on either the viral (+) or non-viral (-) strands of fragments are indicated. Each horizontal line represents the extent of individual sequence runs carried out at least in duplicate. Lines grouped under (+) have their 5'-ends at the left-hand side and their start positions coincide with an restriction enzyme cut; lines grouped under (-) start with an enzyme cut at the left.

fragment was subjected to the chemical degradation procedure and the partial products were analysed on the DNA sequencing gels. For this purpose, the partially cleaved DNA samples were divided into three portions and loaded on three different types of polyacrylamide gels. Electrophoresis on a 25% slab gel allowed an unambiguous reading of about 30 nucleotides starting at the second base after the restriction enzyme cleavage site. On a 20%
slab gel it was possible to read from positions 25 to about 70 whereas on
a 15% slab gel (in some cases a 10% gel) we could normally read until
position 130-150. Whenever feasible, sequencing was performed on both
strands of fragments and was carried out at least twice to insure accuracy
of the sequence data. Occasionally a peculiar band-to-band spacing was
observed, a phenomenon that can be accounted for by intramolecular secondary
structure of the DNA during electrophoresis. In such cases either multiple
sequence runs were carried out under different electrophoresis conditions or
the sequence was confirmed by establishing the sequence at the opposite
region of the complementary strand. The sequence strategy applied in this
study is depicted in Fig. 1 which shows which DNA strand was actually
sequenced and which restriction sites were used for determining the over-
lapping sequences. From these cumulative results it is clear that about 90%
of the DNA sequence was determined on both strands. Clear exceptions are
the flanking regions of the MboI-site between fragments B and C and the left-
hand regions of the HapII-B2/E2 junction and of AluX-A for which insufficient
cleavage sites are present in this part of the DNA genome.

The nucleotide sequence established by these procedures and which com-
prised about 2750 base pairs, is shown in Fig. 2.

Mapping of genes within the established nucleotide sequence

To locate the coding regions of the genes III, VI and I within the
nucleotide sequence we have made use of several amber mutants in each gene.
Knowledge of the exact position of the amber mutation provides information
about the different reading frames and consequently about the exact position
of each gene within the deduced nucleotide sequence. All amber mutants
applied in this study were obtained after treatment of M13 phage with
hydroxylamine, a mutagenic agent which preferentially introduces C → T
transitions.

Map position of gene III

Previously we have demonstrated that the amber mutation am3-H5 within
gene III can be rescued by the wild-type restriction fragment HapII-H (26).
Interestingly, in the nucleotide sequence of the viral strand of this
wild-type fragment there is only one CAG sequence present. This is located
25 base pairs to the left of the Hinf-I cleavage site between the fragments
HinfI-C and E1 (cf. Fig. 1). Therefore, it is very probable that this CAG
sequence is mutated into a TAG (amber) codon contained in am3-H5 phage DNA. To obtain evidence for this, we sequenced part of the HinfI-C fragment which was isolated from M13 RF bearing the am3-H5 mutation. The amber mutant fragment was terminally labeled with (γ-32P)-ATP and subsequently digested with R.Hap II. After fractionation, isolation and sequencing of the fragment which bears the 32P-label at the 5'-end of the complementary (non-viral) strand, a single change of a G + A was observed (Fig. 3). This change occurred in the non-viral strand sequence 5'-CTA-3', a sequence which corresponds to the amber codon 5'-TAG-3' in the viral strand. The mutation, am3-H5, which therefore originates from a C + T transition in the viral strand, is located at position 2473 (Fig. 2).

In the same way we sequences part of fragment TaqI-E derived from M13 RF bearing a second amber mutation in gene III, i.e. am3-H1. Also in this case a single G + A change was observed in the non-viral strand sequence of this fragment (data not shown). This change which corresponded to a C + T transition in a 5'-CAG-3' sequence of the wild-type viral strand, was found at position 2017 (Fig. 2). Both the am3-H5 and am3-H1 mutations were found in phase with each other thereby confirming that the deduced sequence for this part of the genome is correct.

With the aid of these genetic data the amino acid sequence of the gene III product can now be deduced. Previously we have demonstrated that gene III, which codes for the minor capsid protein of the phage, is contained on fragment AluI-A (19). Since the molecular weight of this protein is about 45,000 and AluI-A has only a length of 1446 base pairs, the start codon for gene III protein must be located very near the terminal end of this fragment. Recently, the N-terminal amino acid sequence of the minor capsid protein of the closely related phage fd has been established (29). This sequence, Met-Ala-Glu-Thr-Val-Glu-Ser-, correspond exactly with the nucleotide sequence at the positions 1633-1650 (Fig. 2). The reading frame for this N-terminal polypeptide is in phase with the mutated amber codons in phage M13 am3-H5 and am3-H1. Since no ATG initiation codon is found immediately preceding this N-terminal sequence the start codon of gene III must be located more upstream but before the Alu I cleavage site between the fragments N and A at position 1517 (cf. Fig. 1). No ATG codon, however, is present in this region. Thid only candidate for the start of gene III protein is the GTG codon at position 1579. That this GTG triplet is the initiation codon is strongly supported by the observation that an in phase stop codon (TAA) is found 14 triplets more upstream this initiation

2805
potential ribosome binding site. Within the proposed ribosome binding site there are eight bases which are complementary to the 3'-OH terminal codon (position 1537). Furthermore, the GTG codon is in the center of a sequence of ribosomal 16S RNA (cf. Table 3). This complementarity is a
common feature of ribosome binding sites (35).

Our finding that the polypeptide initiated at this GTG initiation codon is 18 amino acid residues longer than the minor capsid protein present in the mature phage particle, therefore strongly suggests that this capsid protein is synthesized in vivo as a precursor molecule. This conclusion is fully supported by previous observations of Konings et al. (17), who observed a difference in electrophoretic mobility on SDS gels between the gene III product synthesized in vitro and the minor capsid protein isolated from the mature M13 virion.

Reading in phase from the amber codon am3-H5 in the 3'-direction we find a nonsense TAA codon at position 2851 (Fig. 2). From these nucleotide sequence data it therefore can be concluded that the precursor of the minor capsid protein is 424 amino acids long whereas the mature capsid protein has only a length of 406 amino acids.
From our previous marker-rescue experiments we know that several amber mutants of gene VI are salvaged by the wild-type DNA fragment Alul-G (22). For an exact localization of these mutation sites within the established nucleotide sequence we made use of the "plus and minus" sequencing method of Sanger and Coulson (28). Restriction fragments which partially overlap fragment Alul-G were applied as primers whereas the viral DNA strands isolated from five different amber mutants in gene VI were used as DNA templates. By using wild-type fragment HhaI-G as a primer for a limited DNA synthesis reaction on am6-H3 single-stranded viral DNA, we were able to show unambiguously that the sequence 5'-CTG-3' of the wild-type complementary strand has changed into a 5'-CTA-3' sequence in the mutant complementary strand (Fig. 4). This change, which corresponds to a mutation of 5'-CAG-3' into 5'-TAG-3' in
Figure 4. DNA sequence around the \textit{am6-H3} mutation site in phage M13 DNA. Sequencing procedure was the "plus and minus" method (28) with fragment \textit{HhaI-G} as the primer.

the viral strand, was found 31 base pairs to the left of the \textit{HhaI} cleavage site between the fragments \textit{HhaI-N} and \textit{HhaI-G} (Fig. 2; position 3066). Although several other CAG codons are present within the sequence of gene VI, we have observed that in all of the independently isolated amber-6 mutants we have tested so far, \textit{i.e.} \textit{am6-H1}, \textit{am6-H2}, \textit{am6-H3}, \textit{am6-H6}, and \textit{am6-H7}, surprisingly it is only the codon at position 3066 which is mutated into an amber codon in these mutant M13 phages (data not shown).

The initiation codon of gene VI protein was found by reading in the 5'-direction and in phase from the mutated \textit{am-6} codon till the nonsense codon at position 2841 (Fig. 2). The first ATG codon beyond this TAA codon has been considered the initiation codon of gene VI (position 2856). This codon, which is located only 5 base pairs beyond the C-terminal end of gene III, is in fact the only in phase ATG codon present in gene VI which is located upstream of the mutated amber codon. The polypeptide encoded by gene VI is terminated at a TAA codon which is found at position 3192.
The amino acid sequence of gene VI protein, as deduced from the nucleotide sequence data, is shown in Fig. 2. Gene VI protein has a length of 112 amino acids and a molecular weight of 12,264.

Map position of gene I

The mutant aml-H7, which is an amber mutation in gene I, is rescued by the wild-type fragment AluI-G (22). In the derived nucleotide sequence of fragment AluI-G there are, irrespective of the reading frame chosen, only three CAG codons. One of these codons has already been characterized in the previous section as the amber codon present in the M13 mutant am6-H3. The second CAG codon is in the same reading frame at three triplets downstream of the am6-H3 codon (position 3075) whereas the third is found in a different reading frame 16 nucleotides from the 3'-terminal end of the viral strand in fragment AluI-G (Fig. 2, position 3262). Genetic studies have indicated that amber mutations within genes VI and I complement each other (16). For this reason we conclude that the first CAG codon which was identified as the codon which is mutated in the gene VI amber mutants, at the same time cannot be the codon which is mutated into an amber codon in aml-H7 mutant phages. For identical reasons the second CAG codon, which is in the same phase at an extremely short distance from the first CAG codon, cannot be the mutated codon in aml-H7 RF. Hence, the third CAG codon on fragment AluI-G is most probably the codon which is mutated into an amber codon in aml-H7 phages.

From the established nucleotide sequence it appears that the latter codon is part of the Hinf I cleavage site between fragments D and K (Fig. 1). The sequence of this particular region is 5'-GATTCAG-3'. If indeed the third CAG codon is mutated into the amber codon in aml-H7 RF, then this sequence should be 5'-GATTAG-3' resulting in a disappearance of this particular Hinf I cleavage site. As shown in Fig. 5, this turned out to be the case. The fragments D an K present in the digestion pattern of wild-type M13 RF are absent in the pattern of aml-H7 RF. In the latter digest, however, a fragment now can be observed which has no counterpart among the fragments of wild-type RF and which has a length equivalent to the sum of the lengths of fragments HinfI-D and K. From these observations we conclude that the amber mutation in aml-H7 RF is positioned within the Hinf I cleavage site between the fragments D and K at position 3262.

Reading in phase from this amber codon in the 5'-direction the first nonsense (TAA) codon is found at position 3127. Only a single ATG codon
Figure 5.
Restriction cleavage pattern of wild-type M13 RF and aml-H7 RF after digestion with R. Hinf-I.
The fragments were displayed on a discontinuous gel consisting of 3% polyacrylamide on top of a 10% polacrylamide gel layer (26) and visualized by staining with ethidium bromide followed by ultraviolet irradiation.

(position 3196) is located between this nonsense codon and the mutated codon in aml-H7 RF whereas an in phase GTG codon is absent. It is therefore almost certain that this ATG codon which is located only 4 nucleotides beyond the C-terminal end of gene VI, functions as the initiation codon for the synthesis of gene I protein. Its position at only 23 triplets upstream of the mutated amber codon is also in accordance with our previous marker-rescue results which demonstrated that the aml-H7 mutation must be located very near the N-terminal end of gene I (22).

The coding region of gene I terminates at a TAA codon at position 4240. This position is quite striking since it implies that gene I extends into the sequence coding for the N-terminal part of its adjacent gene, gene IV. Ravetch et al. (36) have sequenced ribosome-protected RNA fragments from phage f1 RF-directed RNA transcripts. The sequence of the ribosome binding site of gene IV as determined by this group corresponds exactly with the DNA sequence at position 4204 - 4227 (Fig. 2). Therefore, the N-terminal part of gene IV is well defined. From these combined data we conclude that the C-terminal end of gene I overlaps with the first 23 base pairs of the N-terminal end of gene IV.

The deduced length of gene I, the sequence of which is presented in
Fig. 2, is 1047 base pairs. Its encoded protein has a molecular weight of 39,500.

DISCUSSION

The single-stranded DNA of the male-specific, filamentous coliphages, like M13, fd and fl, is characterized by a rather high thymine content (35.8%, cf. 2). Surprisingly, from the nucleotide sequence of the M13 genes it now can be concluded that this frequent occurrence of T-residues is not at random but that there exists a significant preference for codons with a T in the third position. The codons of gene III have in 213 out of 424 cases (50%) a T in the third position. For gene VI this percentage is even 53% whereas for gene I a percentage of 51% is calculated. The reason for this high T content of the viral DNA strand is unknown but an analogous preference for T in the third position of codons has already been noted for phage ΦX174 (31).

During the course of our work the complete nucleotide sequence of the DNA of the closely related phage fd has been established (40). Comparison of our nucleotide sequence data of phage M13 with the corresponding regions in phage fd indicates that the sizes of the genes in M13 and fd are exactly identical. The only difference in size noted so far between M13 and fd is located within a non-coding part of the genome, i.e. in the intergenic region of genes VI and I. In phage M13 a single base pair is deleted in this region. Furthermore, it is observed that within the established sequence 2.2% of the bases are interchanged. Only in four cases a base change resulted in a change of the corresponding amino acid sequence (bases at positions 2702, 3685, 4171 and 4333). Most of the changes, however, are in the third position of the codon in such a way that as a result of "wobble" base pairing the amino acid sequence remains unaltered. In one case only, two bases within the same codon are different (position 2674 and 2776). This difference, i.e. a CTC codon in M13 and a TTG codon in fd does not result, however, in a change of the codon capacity. Table 1 summarizes the positions and the nature of the observed interchanges. Of the 63 substitutions observed, 40 are transitions in nature and the remainders are transversions. Furthermore 45 of the base interchanges are of the nature X→T (71%).

Gene III appears to span 1275 base pairs. Its initiation codon CTG is located immediately distal to the central terminator of in vitro transcription which is located at position 1538 to 1564. This ρho-independent termination signal, which recently has been shown to operate also as such
TABLE I

Nucleotide interchanges in the genes III, VI and I of bacteriophage M13 and fd

<table>
<thead>
<tr>
<th>M13 → fd</th>
<th>number of substitutions</th>
<th>sequence position</th>
</tr>
</thead>
<tbody>
<tr>
<td>G → A</td>
<td>5</td>
<td>2418 3462 3696 3795 4285</td>
</tr>
<tr>
<td>A → G</td>
<td>6</td>
<td>1764 2682 2710 3546 3685 4332</td>
</tr>
<tr>
<td>C → T</td>
<td>13</td>
<td>1665 2673 2674 3059 3411 3444 3456</td>
</tr>
<tr>
<td>T → C</td>
<td>16</td>
<td>3576 3726 3741 4348 4354 4369</td>
</tr>
<tr>
<td>A → G</td>
<td>2</td>
<td>1737 2328 2385 2391 2939 3071 3429</td>
</tr>
<tr>
<td>T → A</td>
<td>8</td>
<td>3498 3621 3636 3660 3669 3810 3828</td>
</tr>
<tr>
<td>G → T</td>
<td>2</td>
<td>4321 4363</td>
</tr>
<tr>
<td>T → G</td>
<td>2</td>
<td>2070 3408</td>
</tr>
<tr>
<td>A → T</td>
<td>4</td>
<td>2817 4387</td>
</tr>
<tr>
<td>T → A</td>
<td>8</td>
<td>2370 3387 3531 4375</td>
</tr>
<tr>
<td>C → G</td>
<td>1</td>
<td>2702 3044 3489 3855 3972 4333 4366 4384</td>
</tr>
<tr>
<td>C → A</td>
<td>3</td>
<td>3843 4171 4200</td>
</tr>
<tr>
<td>G → C</td>
<td>1</td>
<td>4363 3408</td>
</tr>
<tr>
<td>A → C</td>
<td>2</td>
<td>4170 4399</td>
</tr>
</tbody>
</table>

deletion: position 3194-3195 A

in vivo (9), is characterized by a GC-rich region with a two-fold rotational symmetry immediately preceding a long stretch of T residues (cf. 34). From the deduced nucleotide sequence of gene III it further can be concluded that its primary protein product is subject to post-translational cleavage. Apparently, during phage assembly at the host cell membrane 18 mainly hydrophobic amino acid residues have to be split off from the N-terminal end of the minor capsid protein precursor before packaging into the virion can occur. From these data a molecular weight of 42,675 for the minor capsid protein is calculated. This value is substantially lower than the average molecular weight of about 60,000-70,000 estimated for this protein from the relative electrophoretic mobility on SDS-polyacrylamide gels (17,18,29). As already pointed out by Schaller and Takanami (32) this discrepancy is probably due to the unusual structural features of gene III protein. In the nucleotide sequence of gene III of phage M13 and fd two clusters of a four-
fold repeat of a quindecanucleotide are present which code for the polypeptide Glu-Gly-Gly-Gly-Ser (positions 1834-1893 and 2320-2379). In addition, the second cluster is preceded by another unusual cluster of nucleotides (position 2284-2319) which codes for a three-fold repeat of the tetrapeptide Gly-Gly-Gly-Ser and is followed (position 2380-2391) by a single tetrapeptide coding sequence (cf. Fig. 2). It is probably due to this unusual clustering of the amino acids glycine and serine that aberrant molecular weights have been found on SDS-gels.

It should be emphasized, however, that both sequence repeats are also very rich in GC-pairs. It therefore cannot be ruled out that such unusual GC-rich regions exert additional effects such as pausing of RNA polymerase during transcription of this gene. Pratt et al. (16) have shown that genes III, VI and I form an operon with gene III amber mutations polar on genes VI and I. Since gene III mutants with the strongest polar effect have been mapped proximal to the first repeat whereas - as shown in this study - the less-polar mutant am3-H1 and the non-polar mutant am3-H5 are located, respectively, in between both repeats and distal to these repeats, it is very attractive to correlate these unusual structures with the polarity phenomenon.

We now can conclude that gene VI codes for a rather small protein of 112 amino acids long. Its amino acid composition together with that of gene III and gene I protein is presented in Table 2. Hitherto several groups

<table>
<thead>
<tr>
<th></th>
<th>pre-III</th>
<th>III</th>
<th>VI</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phe</td>
<td>23</td>
<td>21</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>Ser</td>
<td>35</td>
<td>33</td>
<td>8</td>
<td>29</td>
</tr>
<tr>
<td>Tyr</td>
<td>21</td>
<td>20</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>Cys</td>
<td>8</td>
<td>8</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Leu</td>
<td>22</td>
<td>19</td>
<td>24</td>
<td>40</td>
</tr>
<tr>
<td>Trp</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Pro</td>
<td>27</td>
<td>25</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>His</td>
<td>3</td>
<td>2</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Arg</td>
<td>9</td>
<td>9</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>Gln</td>
<td>15</td>
<td>15</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>Glu</td>
<td>26</td>
<td>26</td>
<td>-</td>
<td>7</td>
</tr>
<tr>
<td>Ile</td>
<td>10</td>
<td>9</td>
<td>13</td>
<td>21</td>
</tr>
<tr>
<td>Thr</td>
<td>26</td>
<td>26</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Asp</td>
<td>24</td>
<td>24</td>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>Asn</td>
<td>29</td>
<td>29</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>Lys</td>
<td>16</td>
<td>14</td>
<td>7</td>
<td>33</td>
</tr>
<tr>
<td>Met</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Val</td>
<td>22</td>
<td>20</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>Ala</td>
<td>29</td>
<td>28</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>Gly</td>
<td>67</td>
<td>67</td>
<td>8</td>
<td>23</td>
</tr>
</tbody>
</table>

**Total** 424 406 112 348
have not succeeded in demonstrating the *in vivo* or *in vitro* synthesis of gene VI protein (17-21). One of the possibilities for not detecting this protein could be a very low rate of synthesis due to a very weak ribosome binding site of gene VI. It has been postulated that the efficiency of ribosome recognition is determined by the number of bases at the initiation site which can form Watson-Crick base pairs with the 3'-OH terminal end of 16S ribosomal RNA (35). As is shown in Table 3 the complementarity between the Shine-Dalgarno sequence and the ribosome binding site of gene VI is by no means weak and of the same order of magnitude as found for the initiation sites of genes III, I and IV. Hence, the low level of gene VI protein synthesis is not explained at the level of ribosome recognition capacity. On the other hand, it now appears that gene VI protein is one of the most extreme hydrophobic proteins known so far (33). It has a very high content of leucine and isoleucine (21% and 12%, respectively). It is also striking that most leucine residues are located in the N-terminal part of this protein. Due to this strongly hydrophobic character it is very probable that gene VI protein can only be synthesized at and only can function in the host cell membrane. Since proteins with such a high hydrophobic character are also hardly soluble in 8 M urea, it might be that for these two reasons gene VI protein synthesis has escaped detection on the various polyacrylamide gel systems used for the identification of the *in*

| TABLE 3 |
| DNA sequences of ribosome binding sites in M13 DNA |

<table>
<thead>
<tr>
<th>RNA</th>
<th>3′OH-</th>
<th>AUUCCUCACUAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene III</td>
<td>TTTGGAGAA</td>
<td>ATG AAA AAA TTA</td>
</tr>
<tr>
<td></td>
<td>ATTCTT</td>
<td>Lys Lys Leu</td>
</tr>
<tr>
<td>gene VI</td>
<td>ATAGGAGTCTTAATC</td>
<td>ATG CCA GTT CTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Met Pro Val Leu</td>
</tr>
<tr>
<td>gene I</td>
<td>GATTGGATATAAAA</td>
<td>ATG GCT GTT TAT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Met Ala Val Tyr</td>
</tr>
<tr>
<td>gene IV</td>
<td>AAAAGGTAATTCAA</td>
<td>ATG AAA TTG TTA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Met Lys Leu Leu</td>
</tr>
</tbody>
</table>

Nucleotides complementary to the 3′-OH terminus of 16S ribosomal RNA are underlined.
Another interesting feature of the nucleotide sequence presented is that the intergenic regions between genes III and VI and between VI and I are very short. These silent regions comprise only two base pairs and a single base pair only. If it is assumed that genes III, VI and I are transcribed via a single polycistronic mRNA as suggested by the polarity between these genes, then it is very probable that these short intergenic regions will enhance the polarity effects. Studies by Martin and Webster (38) and Steege (39) have evidenced that upon reading of a polycistronic mRNA the 30S ribosomal subunit which has just completed termination at the termination site of a short intergenic region is not released per se but instead in its attached form is capable of initiating translation at the adjacent start codon with a new 50S particle. This suggests that translational starts at very short intergenic regions are determined by the translation frequency of the proximal genes. Consistent with this is our finding of very short intergenic regions between the consecutive genes V, VII and IX (30) for which also a polarity has been demonstrated (G. Simons et al., unpublished data).

We are aware of some conflicting evidence against the general assumption that genes III, VI and I form an operon which only can be transcribed via a polycistronic mRNA. By in vitro transcription studies with M13 RF and M13 DNA restriction fragments we have demonstrated that gene III but also genes VI and I are preceded by well-defined sites where RNA synthesis preferentially starts in vitro and which we have considered as promoter sites (11,19). Nucleotide sequence analyses have indicated that promoter sites are rich in AT-base pairs and that they all contain regions partially homologous to the sequences TGTTGACAATT and TATAATPu. The latter sequences represent the recognition and binding sequence, respectively, which are the primary targets with which E.coli RNA polymerase interacts. Regions which contain these characteristic sequences are found, as expected, in front of gene III (position 1490-1544) but also in front of gene VI (position 2711-2765) and gene I (position 3082-3136). The same regions were recently identified as in vitro RNA polymerase binding sites by Schaller and Takanami (32). It should be emphasized, however, that within the regions preceding genes VI and I two additional stretches can be traced (2160-2664 and 2129-3183) with similar promoter sequence characteristics and which show even a more pronounced sequence homology with the presumed recognition and binding sequences.
Whether these sequences represent in vivo promoters awaits further sequence studies on M13-specific RNA species present in the infected cell.

ACKNOWLEDGMENTS

We thank Dr. Theo Hulsebos for his mutual interest and participation in these sequence studies, and Ben Cornelissen, Rob Moormann and Norbert Groot for their sequence contributions. The generous help of Dr. Hans Meijer in setting up the computer program which was kindly provided to us by Dr. Staden, Cambridge, is greatly acknowledged. We thank Dr. Heinz Schaller for communicating his sequence results of fd and Dr. Ruud Konings for his advises with regard to this manuscript.

REFERENCES