Gene, 11 (1980) 129–148

Nucleotide sequence of the filamentous bacteriophage M13 DNA genome: comparison with phage fd

(Restriction endonucleases; DNA sequence; gene structure; regulatory signals)

Peter M.G.F. van Wezenbeek, Theo J.M. Hulsebos and John G.G. Schoenmakers

Laboratory of Molecular Biology, University of Nijmegen, Nijmegen (The Netherlands)

(Received March 3rd, 1980)
(Accepted May 27th, 1980)

SUMMARY

The 6407 nucleotide-long sequence of bacteriophage M13 DNA has been determined using both the chemical degradation and chain-termination methods of DNA sequencing. This sequence has been compared with that of the closely related bacteriophage fd (Beck et al., 1978). M13 DNA appears to be only a single nucleotide shorter than fd DNA. There is an average of 3.0% of nucleotide-sequence differences between the two genomes, but the distribution of these changes is not random; the sequence of some genes is more conserved than of others. In contrast, the nucleotide sequences and positions of the regulatory elements involved in transcription, translation and replication appear to be identical in both filamentous phage DNA genomes.

INTRODUCTION

The small filamentous, male-specific coliphages like M13, fd and f1 are characterized by a small single-stranded circular DNA genome of about 6400 bases (for reviews, see Denhardt, 1975; Ray, 1977). Since their genomes code for only a limited number of proteins, the reproduction of these phages and their concomitant processes of DNA replication and gene expression are largely dependent on host functions. For this reason the small DNA phages are attractive models for unravelling the complexity of larger reproductive systems and the analysis of replication and expression of the more complex host genome.

By combined genetic data (Lyons and Zinder, 1972), ordering of conditionally lethal mutants on restriction-enzyme cleavage maps (van den Hondel et al., 1975; 1976; Seeburg and Schaller, 1975; Horiuchi et al., 1975; Vovis et al., 1975) and protein size data (Model and Zinder, 1974; Konings et al., 1975) eight genes have been ordered on the filamentous M13 genome. Very recently we have provided solid evidence for the existence of another small M13 gene, designated gene IX (Simons et al., 1979). The genes III, VIII and IX code for virion structural proteins whereas gene VI specifies a small hydrophobic protein that is most probably also an integral part of the viral capsid (Simons et al., 1979). Gene II is required for double-stranded DNA replication whereas gene V-protein functions as a helix-destabilizing protein in single-stranded viral DNA synthesis (Pratt and Erdahl, 1968). The remaining genes I, IV and VII specify proteins that are most probably involved in phage assembly. Their exact function,

however, has still to be ascertained. To understand the genetic organisation of this small genome at the nucleotide level and to get a better insight into the processes that regulate its replication and viral gene expression we undertook to sequence the M13 DNA genome.

We report here the complete nucleotide sequence of bacteriophage M13 DNA, which comprises 6407 bases. Earlier results were obtained by analysing highly labelled RNA transcribed in vitro from restriction fragments (Hulsebos and Schoenmakers, 1978). All of our more recent results described here and in other reports (Hulsebos and Schoenmakers, 1978; van Wezenbeek and Schoenmakers, 1979) were obtained using the base-specific, partial degradation technique of terminally labelled DNA fragments (Maxam and Gilbert, 1977). In some cases the established sequence was confirmed using the enzymatic priming and chain-termination method of DNA sequencing (Sanger et al., 1977).

Schaller and co-workers in collaboration with Takanami's group have pursued similar lines of investigation and their results on the entire nucleotide sequence of phage fd have recently been presented elsewhere (Beck et al., 1978). The determination of the complete nucleotide sequence of two closely related filamentous phages now makes it possible to assign precise genomic locations to biological functions that have been discovered and analysed by many research groups. Since, in addition, the complete sequences of coliphage ϕX174 (Sanger et al., 1978) and G4 (Godson et al., 1978) and of Simian Virus SV40 (Fiers et al., 1978, Reddy et al., 1978) have been reported, further comparative studies can be made on the interrelationships of the DNA structure of these classes of small DNA-containing viruses (Fuchs et al., 1978; see also Fuchs et al., 1980).

MATERIALS AND METHODS

(a) Phages

The DNA subjected to sequencing was obtained from bacteriophage M13. The phage originated from P.H. Hofschneider, Munich. The M13 nonsense mutants am2-H2, am5-H1, am5-H3, am5-H27, am7-H2, am7-H3, am8-H1, am3-H1, am3-H4, am3-H5, am6-H1, am6-H2, am6-H3, am6-H6, am6-H7, am1-H7 and am4-H38, the characteristics of which have been described (Pratt and Erdahl, 1968), were kindly provided by D. Pratt, Davies, CA. The f1 nonsense mutants R124, R13, R99, R148 and R143 were kindly supplied by N. Zinder and his colleagues, Rockefeller University, New York, the fd mutant fd122 was a kind gift from H. Schaller, Heidelberg.

(b) Enzymes and substrates

Sources of restriction endonucleases and the other enzymes applied in this sequence study have been described previously (van Wezenbeek and Schoenmakers, 1979). The dideoxynucleoside triphosphate inhibitors were purchased from P.L. Biochemicals. [γ-^{32}P]ATP (spec. act. >1000 Ci/mmol) was routinely prepared by the procedure of Glynn and Chappel (1964).

(c) M13 DNA and restriction fragments

The procedures for the propagation and purification of wild-type and amber mutant phages and the preparative methods for the isolation of single-stranded viral DNA and of circularly closed double-stranded RF have been described in detail elsewhere (van den Hondel et al., 1975, 1976, van den Hondel and Schoenmakers, 1975).

The isolation of restriction fragments by preparative polyacrylamide gel electrophoresis was performed as described by van den Hondel et al. (1975).

(d) Labelling of fragments with ^{32}P at a single 5′-OH terminus

Suitable restriction fragments (3–4 pmol) were dephosphorylated with bacterial alkaline phosphatase and labelled at their 5′-ends with [γ-^{32}P]ATP and T4-polynucleotide kinase as described previously (van Wezenbeek and Schoenmakers, 1979). In later experiments labelling with kinase was performed by an exchange reaction. Non-dephosphorylated restriction fragments were dissolved in 50 μl of 10 mM Tris · HCl, 7 mM $MgCl_2$, 7 mM β-mercaptoethanol, pH 7.4 and transferred to a polythene tube containing 100 pmol of dried [γ-^{32}P]ATP. The exchange reaction was started by adding 2–3 units of T4 polynucleotide

kinase. After 45 min at 37°C the reaction was terminated with phenol. Carrier tRNA (10 μg) was added and after two extractions with phenol, the labelled fragments were precipitated with ethanol. To generate fragments labelled at only one end, the ^{32}P-labelled fragments were either cleaved with a second restriction endonuclease followed by electrophoretic separation of the products or the DNA strands were directly separated on polyacrylamide gels according to the procedure of Maxam and Gilbert (1977).

(e) DNA sequencing methods

Nucleotide sequence analysis was performed according to the chemical modification method of Maxam and Gilbert (1977). The dideoxynucleoside triphosphate chain-termination method of sequencing was carried out essentially as described by Sanger et al. (1977).

The nucleotide sequences were stored and studied using the computer programmes devised by Staden (1977).

RESULTS AND DISCUSSION

(a) Cleavage maps and nucleotide sequence

Restriction-enzyme cleavage maps are essential not only for analysing the details of organisation and expression of viral genomes but also for sequencing purposes. Several cleavage maps of M13 DNA have been reported previously (van den Hondel et al., 1975; 1976), and several new maps have been constructed during the course of this sequencing study. The sites where several restriction endonucleases cleave the M13 replicative form DNA are presented in Fig. 1.

The procedure followed for nucleotide sequence analysis was invariably the same for each restriction fragment: fragments with only one ^{32}P-labelled 5′-terminus were prepared and worked up by the procedure of Maxam and Gilbert (1977). The partially cleaved DNA samples were routinely loaded on three different types of gels. Electrophoresis on a 25% gel allowed an unambiguous reading of about 30 nucleotides starting at the second base after the restriction enzyme cleavage site. On a 20% gel it was possible to read from positions 25 to about 70 whereas on a 10% or 15% gel, after a large number of nucleotides were allowed to run off, we could normally read until position 150–170.

The restriction fragments that were subjected to sequence analysis are presented in the lower part of Fig. 1. All DNA regions have been analysed at least in duplicate, and every fragment has been analysed on several gels such that the critical areas were appropriately spread out. In some cases an additional confirmation of the deduced sequence was obtained by applying the enzymic priming and chain termination method with the various restriction fragments as primers for limited DNA synthesis on the viral DNA strand as the template (Sanger et al., 1977).

As the map provides more than 140 cleavage sites throughout the M13 DNA molecule at distances generally less than 200 base pairs, a large part of the sequence could be approached in the same direction from two different restriction sites. It also enabled us to determine a very large part of the final sequence independently on both the viral and complementary strand. Only a combination of these sequence data allowed the elimination of certain sequence ambiguities around a few cleavage sites or at certain DNA regions with a high secondary structure.

Few cleavage sites, for instance, are found around the nucleotide positions 850, 1700 and 4600. Consequently, very small parts near these sites could only be sequenced in a single direction. As repeated sequencing of these regions by the chemical degradation method and the enzymic dideoxytriphosphate chain-termination method resulted in completely identical read-outs, we consider their sequence reliable. Some cleavage sites (positions 1396, 1714, 2845, 4665) have not been confirmed by sequencing across these sites. On the other hand, electrophoresis of digested fragments covering these sites indicated that very small fragments were absent and therefore the presence of a cluster of identical cleavage sites had to be ruled out in such cases.

Sometimes regions of autoradiograms revealed a peculiar band-to-band spacing, a phenomenon that can be accounted for by intramolecular secondary structure of the DNA during electrophoresis. An example of this is shown in Fig. 2A within a set of four T-residues. It is of interest to note that these residues are directly preceded by a self-complementary region that can form a stable hairpin-loop (posi-

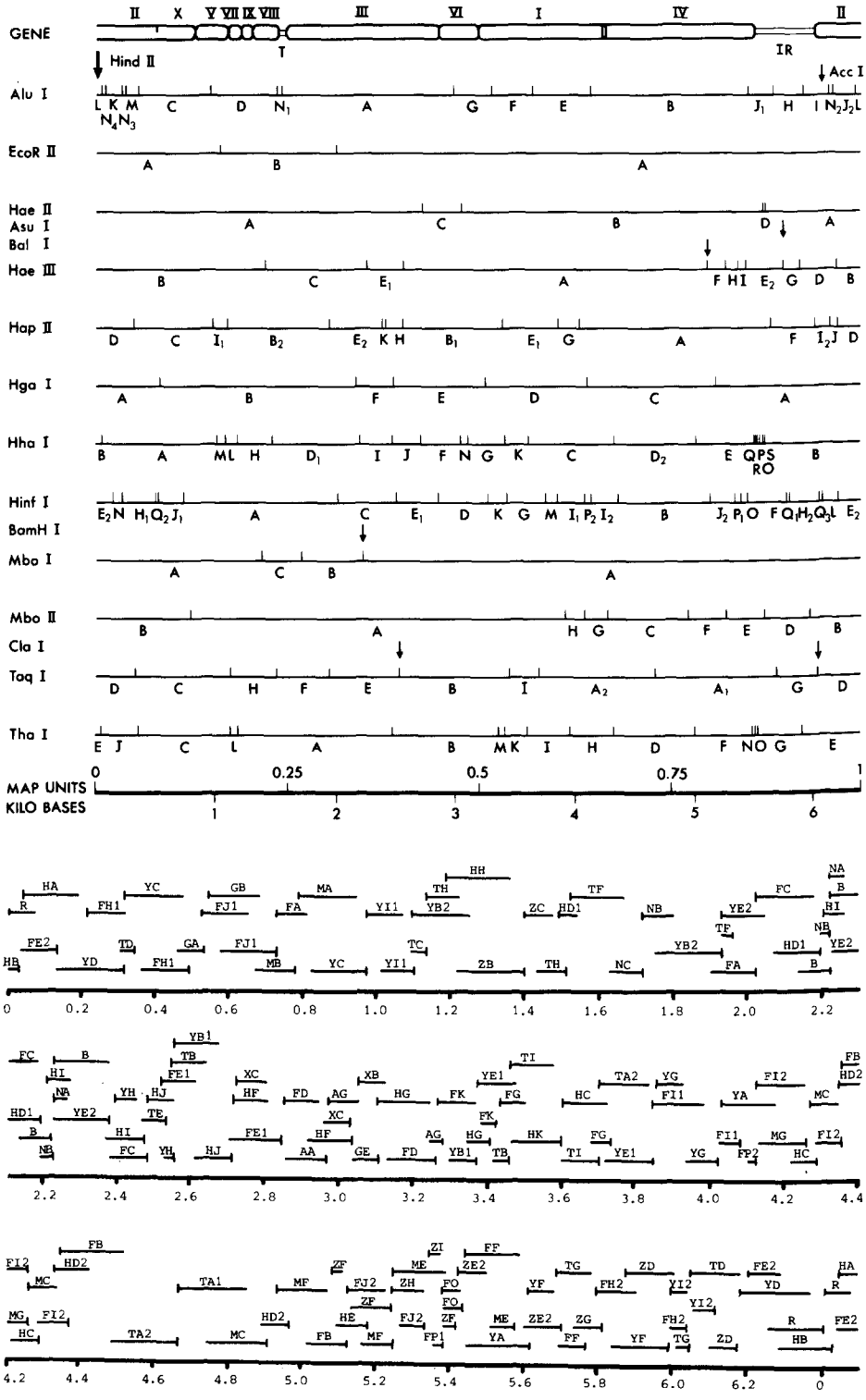

Fig. 1. (Upper part) Genetic map and restriction enzyme cleavage maps of bacteriophage M13 DNA. The circular phage genome is presented in a linear form with the unique *Hin*dII cleavage site as zero point. The Roman numerals refer to the M13 genes. T stands for the rho-independent termination site of transcription. IR refers to the intergenic region in which the origin of replication of viral and non-viral DNA strands is located. (Lower part) The sequencing strategy applied: The solid line represents the length of M13 DNA in kilobases. The extent of the individual sequencing runs are represented by the small horizontal lines. The vertical bars represent the location of the single ^{32}P-labelled 5′-terminal ends. The capital letters above each line denote the restriction enzyme fragment which is used for sequencing at the 5′-end. The second capital letter refers to the fragment which is obtained after digestion with the restriction endonuclease coded by the first letter. The lettercode used is: A, *Alu*I; B, *Bam*HI; F, *Hin*fI; G, *Hga*I; H, *Hha*I; M, *Mbo*II; N, *Mbo*I; R, *Hin*dII; T, *Taq*I; X, *Hae*II; Y, *Hap*II; Z, *Hae*III.

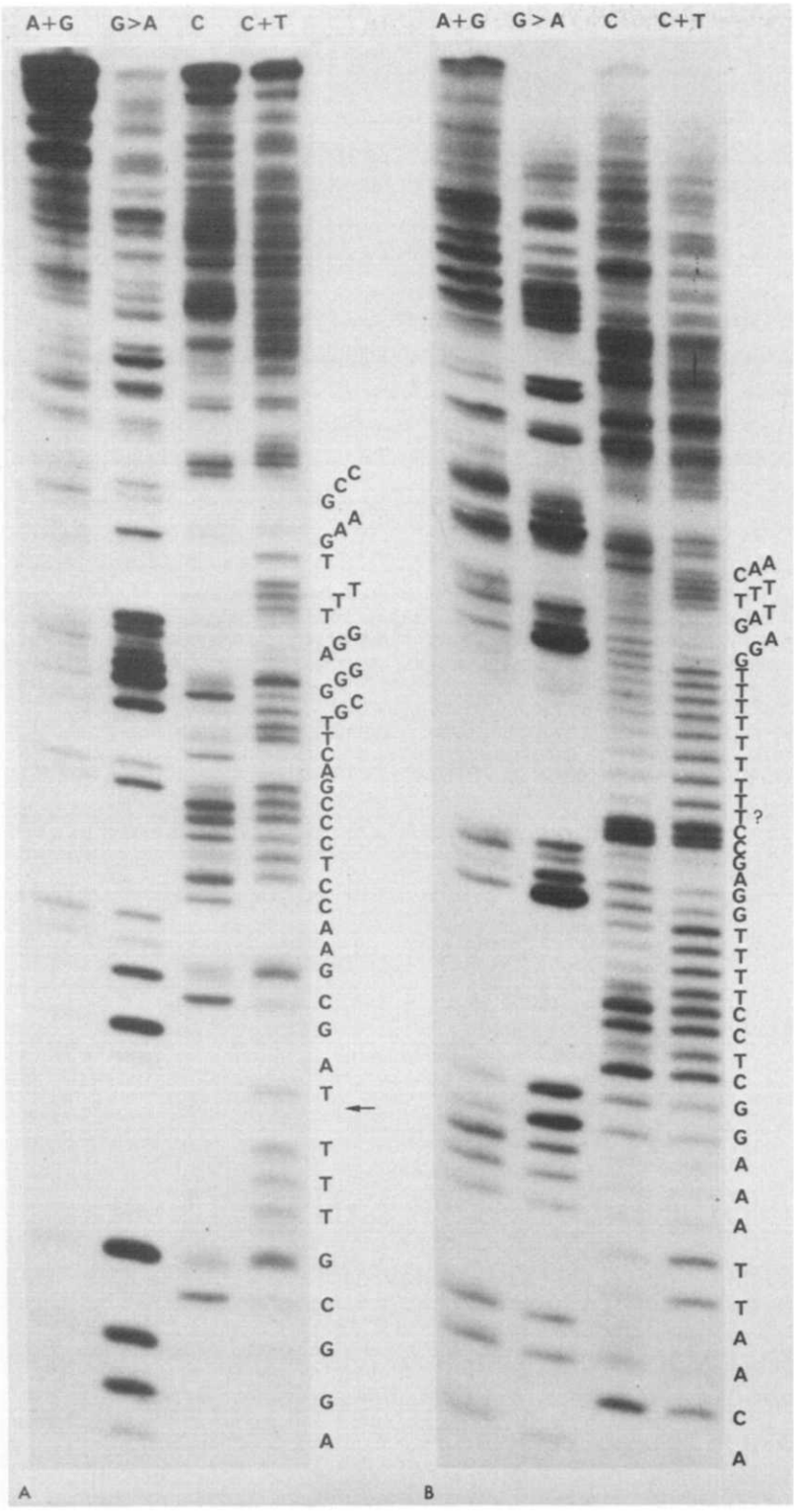

Fig. 2. Autoradiographs of DNA sequencing gels prepared according to the procedure of Maxam and Gilbert (1977). (A) Sequence of the right hand 5′-terminal end of fragment *Hap*-B1. The nucleotide positions 3311–3355 on the viral strand of this fragment are designated. (B) Sequence of the left hand 5′-terminal end of fragment *Taq*I-F. The nucleotide positions 1512–1575 are from the viral strand of this fragment.

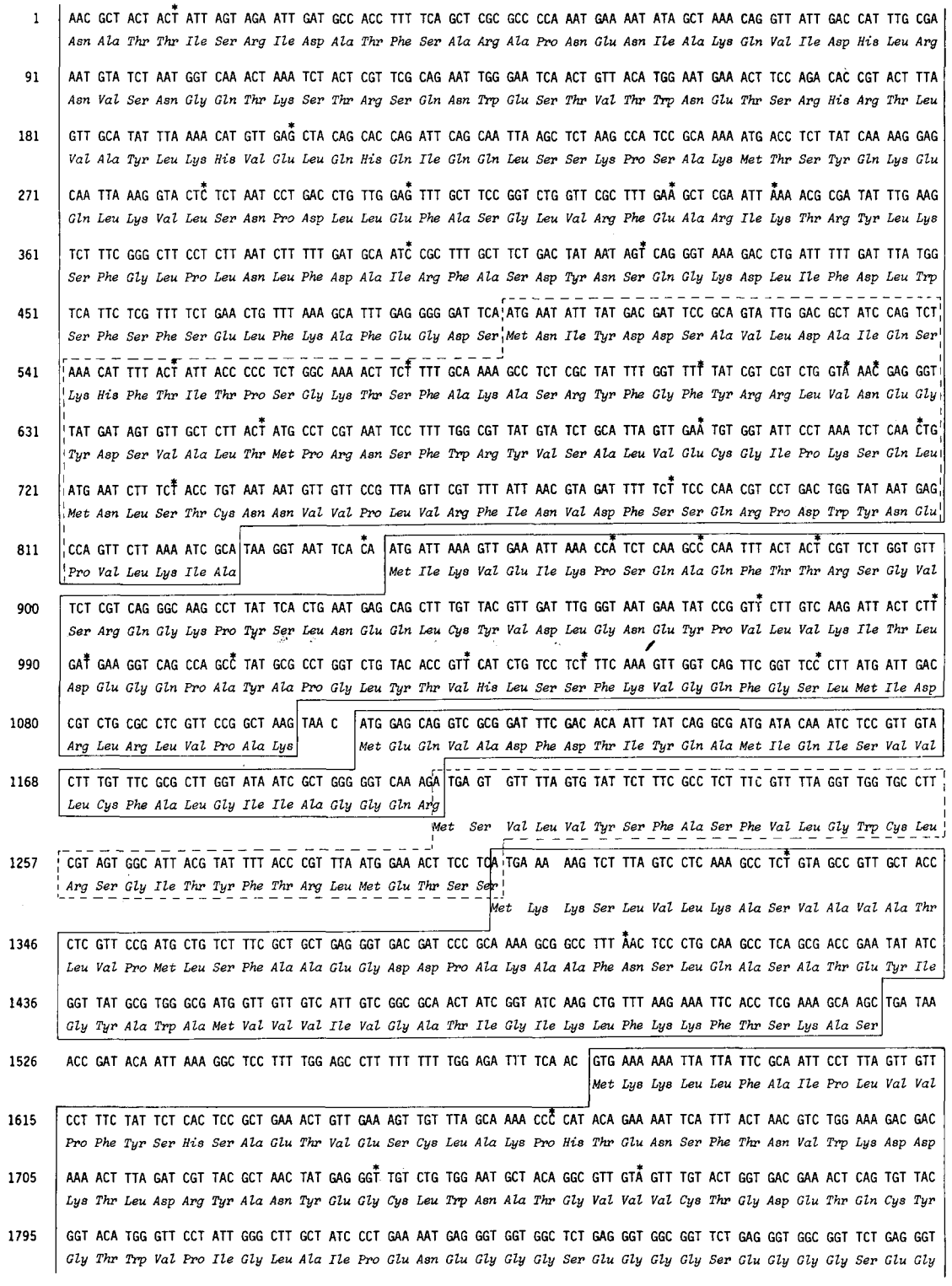

Fig. 3a.

Fig. 3. Nucleotide sequence and amino acid sequence of bacteriophage M13. The numbering of nucleotides is in the 5′-3′-direction of the viral DNA strand and starts at the unique *Hin*dII-cleavage site. The M13 genes are boxed in. The asterisks denote the differences observed between the nucleotide sequence of phage M13 DNA and that of the closely related phage fd (Beck et al., 1978). The restriction-enzyme recognition sites in M13 DNA, as found by computer-analysis data, are compiled.

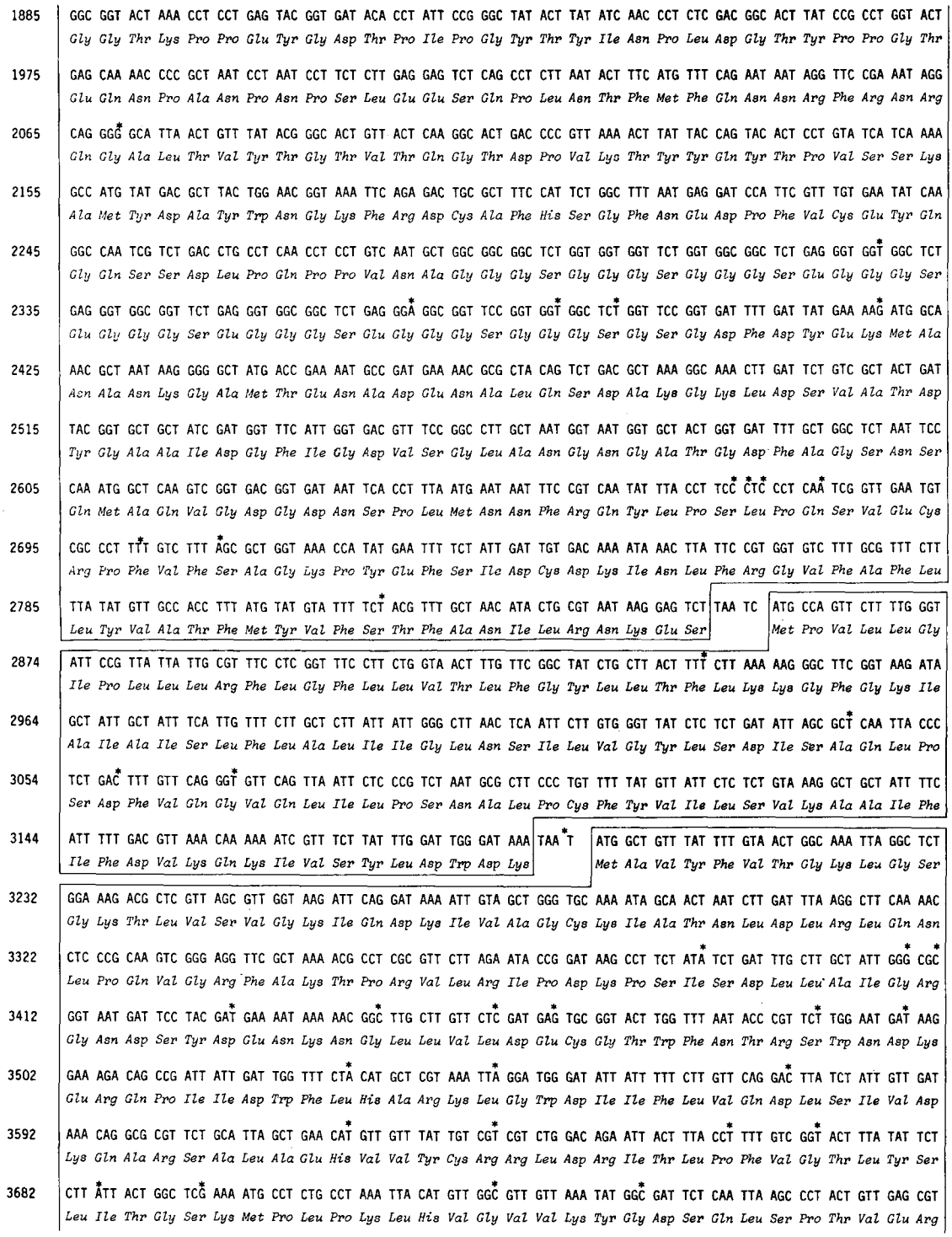
1885 GGC GGT ACT AAA CCT CCT GAG TAC GGT GAT ACA CCT ATT CCG GGC TAT ACT TAT ATC AAC CCT CTC GAC GGC ACT TAT CCG CCT GGT ACT
Gly Gly Thr Lys Pro Pro Glu Tyr Gly Asp Thr Pro Ile Pro Gly Tyr Thr Tyr Ile Asn Pro Leu Asp Gly Thr Tyr Pro Pro Gly Thr
1975 GAG CAA AAC CCC GCT AAT CCT AAT CCT TCT CTT GAG GAG TCT CAG CCT CTT AAT ACT TTC ATG TTT CAG AAT AAT AGG TTC CGA AAT AGG
Glu Gln Asn Pro Ala Asn Pro Asn Pro Ser Leu Glu Glu Ser Gln Pro Leu Asn Thr Phe Met Phe Gln Asn Asn Arg Phe Arg Asn Arg
2065 CAG GGG GCA TTA ACT GTT TAT ACG GGC ACT GTT ACT CAA GGC ACT GAC CCC GTT AAA ACT TAT TAC CAG TAC ACT CCT GTA TCA TCA AAA
Gln Gly Ala Leu Thr Val Tyr Thr Gly Thr Val Thr Gln Gly Thr Asp Pro Val Lys Thr Tyr Tyr Gln Tyr Thr Pro Val Ser Ser Lys
2155 GCC ATG TAT GAC GCT TAC TGG AAC GGT AAA TTC AGA GAC TGC GCT TTC CAT TCT GGC TTT AAT GAG GAT CCA TTC GTT TGT GAA TAT CAA
Ala Met Tyr Asp Ala Tyr Trp Asn Gly Lys Phe Arg Asp Cys Ala Phe His Ser Gly Phe Asn Glu Asp Pro Phe Val Cys Glu Tyr Gln
2245 GGC CAA TCG TCT GAC CTG CCT CAA CCT CCT GTC AAT GCT GGC GGC GGC TCT GGT GGT GGT TCT GGT GGC GGC TCT GAG GGT GGT GGC TCT
Gly Gln Ser Ser Asp Leu Pro Gln Pro Pro Val Asn Ala Gly Gly Gly Ser Gly Gly Gly Ser Gly Gly Gly Ser Glu Gly Gly Gly Ser
2335 GAG GGT GGC GGT TCT GAG GGT GGC GGC TCT GAG GGA GGC GGT TCC GGT GGT GGC TCT GGT TCC GGT GAT TTT GAT TAT GAA AAG ATG GCA
Glu Gly Gly Gly Ser Glu Gly Gly Gly Ser Glu Gly Gly Gly Ser Gly Gly Gly Ser Gly Ser Gly Asp Phe Asp Tyr Glu Lys Met Ala
2425 AAC GCT AAT AAG GGG GCT ATG ACC GAA AAT GCC GAT GAA AAC GCG CTA CAG TCT GAC GCT AAA GGC AAA CTT GAT TCT GTC GCT ACT GAT
Asn Ala Asn Lys Gly Ala Met Thr Glu Asn Ala Asp Glu Asn Ala Leu Gln Ser Asp Ala Lys Gly Lys Leu Asp Ser Val Ala Thr Asp
2515 TAC GGT GCT GCT ATC GAT GGT TTC ATT GGT GAC GTT TCC GGC CTT GCT AAT GGT AAT GGT GCT ACT GGT GAT TTT GCT GGC TCT AAT TCC
Tyr Gly Ala Ala Ile Asp Gly Phe Ile Gly Asp Val Ser Gly Leu Ala Asn Gly Asn Gly Ala Thr Gly Asp Phe Ala Gly Ser Asn Ser
2605 CAA ATG GCT CAA GTC GGT GAC GGT GAT AAT TCA CCT TTA ATG AAT AAT TTC CGT CAA TAT TTA CCT TCC CTC CCT CAA TCG GTT GAA TGT
Gln Met Ala Gln Val Gly Asp Gly Asp Asn Ser Pro Leu Met Asn Asn Phe Arg Gln Tyr Leu Pro Ser Leu Pro Gln Ser Val Glu Cys
2695 CGC CCT TTT GTC TTT AGC GCT GGT AAA CCA TAT GAA TTT TCT ATT GAT TGT GAC AAA ATA AAC TTA TTC CGT GGT GTC TTT GCG TTT CTT
Arg Pro Phe Val Phe Ser Ala Gly Lys Pro Tyr Glu Phe Ser Ile Asp Cys Asp Lys Ile Asn Leu Phe Arg Gly Val Phe Ala Phe Leu
2785 TTA TAT GTT GCC ACC TTT ATG TAT GTA TTT TCT ACG TTT GCT AAC ATA CTG CGT AAT AAG GAG TCT TAA TC ATG CCA GTT CTT TTG GGT
Leu Tyr Val Ala Thr Phe Met Tyr Val Phe Ser Thr Phe Ala Asn Ile Leu Arg Asn Lys Glu Ser Met Pro Val Leu Leu Gly
2874 ATT CCG TTA TTA TTG CGT TTC CTC GGT TTC CTT CTG GTA ACT TTG TTC GGC TAT CTG CTT ACT TTT CTT AAA AAG GGC TTC GGT AAG ATA
Ile Pro Leu Leu Leu Arg Phe Leu Gly Phe Leu Leu Val Thr Leu Phe Gly Tyr Leu Leu Thr Phe Leu Lys Lys Gly Phe Gly Lys Ile
2964 GCT ATT GCT ATT TCA TTG TTT CTT GCT CTT ATT ATT GGG CTT AAC TCA ATT CTT GTG GGT TAT CTC TCT GAT ATT AGC GCT CAA TTA CCC
Ala Ile Ala Ile Ser Leu Phe Leu Ala Leu Ile Ile Gly Leu Asn Ser Ile Leu Val Gly Tyr Leu Ser Asp Ile Ser Ala Gln Leu Pro
3054 TCT GAC TTT GTT CAG GGT GTT CAG TTA ATT CTC CCG TCT AAT GCG CTT CCC TGT TTT TAT GTT ATT CTC TCT GTA AAG GCT GCT ATT TTC
Ser Asp Phe Val Gln Gly Val Gln Leu Ile Leu Pro Ser Asn Ala Leu Pro Cys Phe Tyr Val Ile Leu Ser Val Lys Ala Ala Ile Phe
3144 ATT TTT GAC GTT AAA CAA AAA ATC GTT TCT TAT TTG GAT TGG GAT AAA TAA T ATG GCT GTT TAT TTT GTA ACT GGC AAA TTA GGC TCT
Ile Phe Asp Val Lys Gln Lys Ile Val Ser Tyr Leu Asp Trp Asp Lys Met Ala Val Tyr Phe Val Thr Gly Lys Leu Gly Ser
3232 GGA AAG ACG CTC GTT AGC GTT GGT AAG ATT CAG GAT AAA ATT GTA GCT GGG TGC AAA ATA GCA ACT AAT CTT GAT TTA AGG CTT CAA AAC
Gly Lys Thr Leu Val Ser Val Gly Lys Ile Gln Asp Lys Ile Val Ala Gly Cys Lys Ile Ala Thr Asn Leu Asp Leu Arg Leu Gln Asn
3322 CTC CCG CAA GTC GGG AGG TTC GCT AAA ACG CCT CGC GTT CTT AGA ATA CCG GAT AAG CCT TCT ATA TCT GAT TTG CTT GCT ATT GGG CGC
Leu Pro Gln Val Gly Arg Phe Ala Lys Thr Pro Arg Val Leu Arg Ile Pro Asp Lys Pro Ser Ile Ser Asp Leu Leu Ala Ile Gly Arg
3412 GGT AAT GAT TCC TAC GAT GAA AAT AAA AAC GGC TTG CTT GTT CTC GAT GAG TGC GGT ACT TGG TTT AAT ACC CGT TCT TGG AAT GAT AAG
Gly Asn Asp Ser Tyr Asp Glu Asn Lys Asn Gly Leu Leu Val Leu Asp Glu Cys Gly Thr Trp Phe Asn Thr Arg Ser Trp Asn Asp Lys
3502 GAA AGA CAG CCG ATT ATT GAT TGG TTT CTA CAT GCT CGT AAA TTA GGA TGG GAT ATT ATT TTT CTT GTT CAG GAC TTA TCT ATT GTT GAT
Glu Arg Gln Pro Ile Ile Asp Trp Phe Leu His Ala Arg Lys Leu Gly Trp Asp Ile Ile Phe Leu Val Gln Asp Leu Ser Ile Val Asp
3592 AAA CAG GCG CGT TCT GCA TTA GCT GAA CAT GTT GTT TAT TGT CGT CGT CTG GAC AGA ATT ACT TTA CCT TTT GTC GGT ACT TTA TAT TCT
Lys Gln Ala Arg Ser Ala Leu Ala Glu His Val Val Tyr Cys Arg Arg Leu Asp Arg Ile Thr Leu Pro Phe Val Gly Thr Leu Tyr Ser
3682 CTT ATT ACT GGC TCG AAA ATG CCT CTG CCT AAA TTA CAT GTT GGC GTT GTT AAA TAT GGC GAT TCT CAA TTA AGC CCT ACT GTT GAG CGT
Leu Ile Thr Gly Ser Lys Met Pro Leu Pro Lys Leu His Val Gly Val Val Lys Tyr Gly Asp Ser Gln Leu Ser Pro Thr Val Glu Arg
GENE III
GENE VI
GENE I

Fig. 3b.

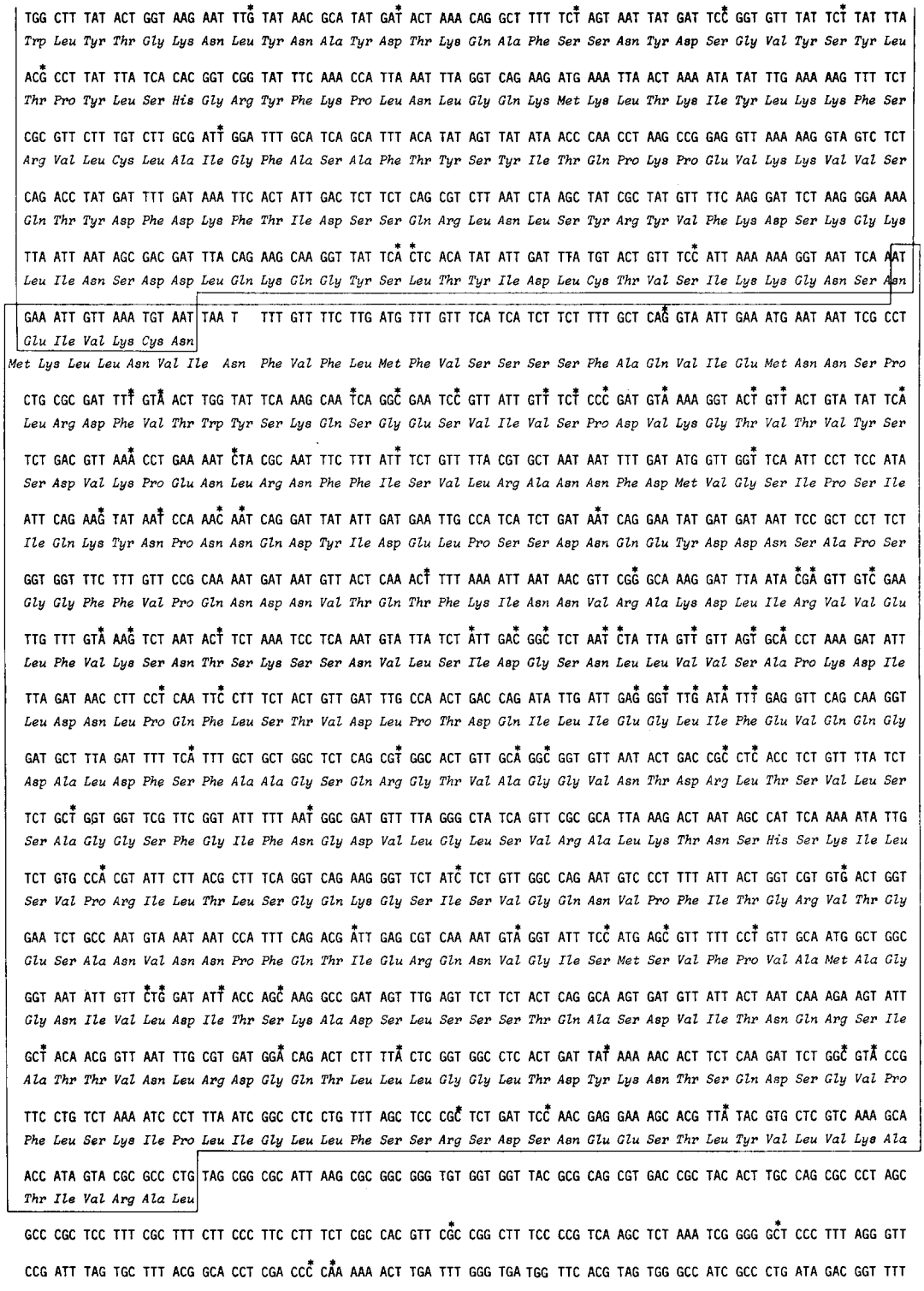

Fig. 3c.

```
5750  TCG CCC TTT GAC GTT GGA GTC CAC GTT CTT TAA TAG TGG ACT CTT GTT CCA AAC TGG AAC AAC ACT CAA CCC TAT CTC GGG CTA TTC TTT
5840  TGA TTT ATA AGG GAT TTT GCC GAT TTC GGC CTA TTG GTT AAA AAA TGA GCT GAT TTA ACA AAA ATT TAA CGC GAA TTT TAA CAA AAT ATT
5930  AAC GTT TAC AAT TTA AAT ATT TGC TTA TAC AAT CTT CCT GTT TTT GGG GCT TTT CTG ATT ATC AAC CGG GGT ACA T  ATG ATT GAC ATG
                                                                                                          Met Ile Asp Met
6018  CTA GTT TTA CGA TTA CCG TTC ATC GAT TCT CTT GTT TGC TCC AGA CTC TCA GGC AAT GAC CTG ATA GCC TTT GTA GAC CTC TCA AAA ATA
      Leu Val Leu Arg Leu Pro Phe Ile Asp Ser Leu Val Cys Ser Arg Leu Ser Gly Asn Asp Leu Ile Ala Phe Val Asp Leu Ser Lys Ile
6108  GCT ACC CTC TCC GGC ATG AAT TTA TCA GCT AGA ACG GTT GAA TAT CAT ATT GAT GGT GAT TTG ACT GTC TCC GGC CTT TCT CAC CCT TTT
      Ala Thr Leu Ser Gly Met Asn Leu Ser Ala Arg Thr Val Glu Tyr His Ile Asp Gly Asp Leu Thr Val Ser Gly Leu Ser His Pro Phe
6198  GAA TCT TTA CCT ACA CAT TAC TCA GGC ATT GCA TTT AAA ATA TAT GAG GGT TCT AAA AAT TTT TAT CCT TGC GTT GAA ATA AAG GCT TCT
      Glu Ser Leu Pro Thr His Tyr Ser Gly Ile Ala Phe Lys Ile Tyr Glu Gly Ser Lys Asn Phe Tyr Pro Cys Val Glu Ile Lys Ala Ser
6288  CCC GCA AAA GTA TTA CAG GGT CAT AAT GTT TTT GGT ACA ACC GAT TTA GCT TTA TGC TCT GAG GCT TTA TTG CTT AAT TTT GCT AAT TCT
      Pro Ala Lys Val Leu Gln Gly His Asn Val Phe Gly Thr Thr Asp Leu Ala Leu Cys Ser Glu Ala Leu Leu Leu Asn Phe Ala Asn Ser
6378  TTG CCT TGC CTG TAT GAT TTA TTG GAT GTT
      Leu Pro Cys Leu Tyr Asp Leu Leu Asp Val
```

GENE II

RESTRICTION ENZYME RECOGNITION SITES IN PHAGE M13 DNA

Name	Sequence	Position
Acc I	GTAGAC	6090
Alu I	AGCT	39 63 203 229 333 934 1488 1517 2963 3276 3612 4096 5426 5630 5887 6107 6134 6335
Asu I	GGGCC	5724
Bal I	TGGCCA	5080
BamH I	GGATCC	2220
Cla I	ATCGAT	2527 6039
EcoR II	CCTGG	1014 1966
Hae II	AGCGCT	2710 3039
	AGCGCC	5559 5567
Hae III	GGCC	1396 2245 2554 5081 5239 5345 5414 5725 5867 6180
Hap II	CCGG	314 966 1095 1924 2378 2396 2552 3370 3842 4018 5614 5995 6118 6178
Hga I	GACGC	526 2164 2479 3237
	GCGTC	4083 5158
Hha I	GCGC	44 1011 1085 1177 1470 2195 2467 2711 3040 3096 3408 3598 4312 4995 5490 5503 5512 5534 5560 5568
Hinf I	GAATC	136 723 4349 5120 6198
	GAGTC	2011 2845 5766
	GACTC	4072 5329 5788 6061
	GATTC	216 490 511 2497 3258 3418 3742 3838 4117 5375 5438 6042
Hph I	GGTGA	1376 1774 1909 2398 2542 2581 2620 2626 4847 5117 5706 6162
	TCACC	1503 2635 4923 6188
Hpa I	GTTAAC	6405
Mbo I	GATC	1382 1714 2221
Mbo II	GAAGA	3912
	TCTTC	781 4075 4271 4937 5255 5587 5962
Taq I	TCGA	336 1127 1508 1949 2528 3455 3694 4665 5683 6040
Tha I	CGCG	43 347 1119 1176 2466 3355 3409 3599 3952 4313 4994 5489 5513 5533 5909

Cleavage sites which are absent in phage M13 DNA : Ava II, Ava III, Bcl I, Bgl II, EcoR I, Hind III, Kpn I, Mst I, Pvu I, Pvu II, Pst I, Sma I, Sac I, Sac II, Sal I, Xho I, Xba I, Eca I.

Fig. 3d.

tion 3319–3341, cf. Fig. 4C) and that, by applying the chain-termination method of sequencing, it was not possible to proceed alongside such a region with a strong secondary structure. Such sequence ambiguities could easily be solved by the observation of a regular band-to-band spacing after sequencing the opposite DNA strand.

Peculiarities were also found in the characteristic set of nine consecutive T-residues of the central termination site for transcription (position 1557–1565). In most experiments a C_2T_9 sequence was clearly indicated. In some experiments, as shown in Fig. 2B, a C_3T_8 sequence can be read, but if one takes into account that the last nucleotide added decreases the mobility more in case this nucleotide is a T instead of a C, a C_2T_9 sequence remains possible. The presence of only two C residues is in accordance with the nucleotide sequence of RNA transcripts terminated at this site (Sugimoto et al., 1977; Edens, 1978). Another uncertainty was an A-residue at position 5538 which is very difficult to detect. The reason for this is still unclear but in all these cases it is probably the secondary structure of these regions that is responsible for such an irregular sequencing behaviour.

The results of these combined sequence studies are presented in Fig. 3. It shows that the entire M13 DNA sequence encompasses 6407 nucleotides.

(b) Gene structure

Complementation studies with conditionally lethal phage mutants have indicated that the M13 genome consists of eight genes. Very recently, a ninth small gene (gene IX) has been detected (Simons et al., 1979), whereas from protein synthesis data there is suggestive evidence for the existence of still another gene (gene X) (Konings et al., 1975, Model and Zinder, 1974). Its product, however, has been detected so far only in vitro. The products of gene V, a helix-destabilizing protein, and of gene VIII, the major capsid protein of the virion, have been characterized by their amino acid sequence (Cuypers et al., 1974, Nakashima and Koningsberg, 1974). Hence, their position can easily be traced within the nucleotide sequence. Such amino acid sequence data are lacking for the remaining M13 genes. To locate the positions of these genes within the nucleotide sequence we have applied several hydroxylamine-induced amber mutants and have analysed the nucleotide changes introduced in the DNA sequence of these mutants. This approach allowed us to determine the reading frame of each gene and, consequently, to predict its initiation and termination signals and the amino acid sequence of its product. The precise location and length of each M13 gene, as deduced from these analyses, are included in Fig. 3.

The nucleotide sequence of gene II ranges from position 6006 till 831. This is predicted from our sequence analysis of two amber mutants of gene II, M13am2-H2 and f1R124. In the former a C → T transition was found at position 214 which changes the glutamine codon CAG into an amber codon. The f1-mutant is characterized by a G → T transversion at position 6348, which changes the glutamic codon GAG into an amber codon. Our results do not discriminate between a starting position of gene II at one of the two closely ATG triplets at position 6006 and 6015. However, the former is more likely since it is preceded by a sequence that is characteristic for a ribosome-binding site. Gene II then codes for a protein of 410 amino acids (mol.wt. 46 117), which is in good agreement with the estimated size of the in vitro synthesized gene II-protein (Konings et al., 1975).

DNA fragments containing the C-terminal part of gene II code for a protein termed "X-protein". As its synthesis has only been demonstrated in vitro and conditionally lethal X-mutants distinguishable from late gene II-mutants have not yet been found, the existence of a separate though overlapping gene X is still not proven. The start of "gene X" is most likely the ATG triplet at position 496, as this is preceded by a potential ribosome-binding site. From this ATG codon until the end of gene II is the only sequence which upon reading gives rise to a protein (mol.wt. 12 670), the size of which corresponds to the in vitro synthesized product. That X-protein is the result of an initiation event within gene II in phase with the rest of gene II-protein is in accordance with observations of Model and McGill (cf. Horiuchi et al., 1978), who demonstrated that synthesis of X-protein in vitro is only affected by a late amber mutant (R21) in gene II.

Gene V is located from position 843 up to 1106. The nucleotide sequence of this gene fully supports the amino acid sequence estimated for gene V-protein (Cuypers et al., 1974) and the sequence of its preceding ribosome binding site (cf. Pieczenick et al.,

1974). Also, the gene V amber mutants analysed so far fit with the nucleotide sequence data. In fd122 we found a C→T transition at position 906, whereas the M13 mutants am5-H1, am5-H3, am5-H27 and the f1 mutants R13 and R99 all showed a C→T change at the same position, namely at 999.

Gene VII is located from position 1108–1209. Its reading frame has been established by sequencing the C→T transitions in the gene VII amber mutants am7-H2 and am7-H3 (Hulsebos and Schoenmakers, 1978). These changes were found at positions 1114 and 1141, respectively. The protein encoded by gene VII has not yet been observed either in M13 infected *E. coli* cells (Henry and Pratt, 1969) or in minicells carrying M13 RF as a plasmid (Smits et al., 1978). Also the in vitro synthesis under the direction of M13 DNA or gene VII containing DNA fragments failed to demonstrate the products of this small gene (Model and Zinder, 1974; Konings et al., 1975; Edens et al., 1978). From the nucleotide sequence data the product is assumed to be a short peptide of only 33 amino acids long. It is of interest to note that the f1 mutant R148, which has been considered to be a gene V amber mutant, showed a C→T change within the CAG codon of gene VII at position 1114. That this mutant is indeed a gene VII mutant is supported by our observations that, during infection of *E. coli* *su*III cells with this mutant, the gene V-proteins formed are of wild-type character (T. Hulsebos, unpublished data).

Gene IX is located from position 1206–1304. The region covering this gene has previously been considered as a noncoding "leader" sequence of gene VIII-mRNA (cf. Sugimoto et al., 1977). The fact that this region can be read from an ATG triplet in position 1206 in a continuous reading frame to yield a protein of 32 amino acids, which overlaps its contiguous gene VIII by only a single nucleotide, led Schaller et al. (1978) to postulate that this region might represent an additional M13 gene. Definite proof, however, was lacking as no conditionally lethal mutants are available that originate from this region of the M13 genome. Very recently, we have demonstrated that gene IX really exists and that this gene codes for an additional small virion capsid protein (C-protein) of the mature phage (Simons et al., 1979).

Genes III and VIII code for the virion capsid proteins A and B, respectively. The in vitro products of these genes are synthesized in a precursor form (Konings et al., 1975). The existence of a gene VIII-protein precursor was also inferred from the observation that the sequence of the ribosome-binding site on gene VIII-mRNA did not coincide with the N-terminal amino acid sequence of the mature capsid protein (Pieczenick et al., 1974). The length and sequence of the precursor has been predicted from the nucleotide sequence of gene VIII-mRNA (Sugimoto et al., 1977) and is confirmed by DNA and amino acid sequence data (Hulsebos and Schoenmakers, 1978; Horiuchi et al., 1978). The structural gene ranges from position 1301 up to 1525 and codes for a protein precursor that contains 23 extra amino acids at its N-terminal end. The only gene VIII amber mutant known so far, i.e. am8-H1, shows a G→T transversion at position 1371, which is very near the Ala-Ala bond cleaved in the processing reaction (Boeke and Model, 1979).

Gene III ranges from position 1579 up to 2853. It codes for a protein of 424 amino acids. The reading frame of this gene was confirmed by analysing the gene III mutants am3-H1, am3-H4 and am3-H5. The former two showed a C→T transition at position 2017, whereas the am3-H5 showed a similar change at position 2473.

Since the N-terminal amino acid sequence NH2-Ala-Glu-Thr-Val-Glu-Ser-, as determined for the mature minor capsid protein of phage fd (Goldsmith and Konigsberg, 1977), corresponds to the nucleotide sequence at position 1633–1650 and the first in phase initiation codon preceding this sequence is the GTG triplet at position 1579, it is predicted that the gene III protein precursor contains 18 additional amino acids at its N-terminal end. It is of interest to note that the N-terminal "signal" peptides of both gene III- and VIII-protein are of rather hydrophobic character (Table I). The calculated molecular weight of gene III-protein is 42675. This is substantially below the values of about 59000–70000 daltons observed in SDS-polyacrylamide gels. This discrepancy is probably due to the unusual clustering of the amino acids glycine and serine as the nucleotide sequence of this gene is characterized by two clusters of a four-fold repeat of a quindecanucleotide (positions 1834–1893 and 2320–2379), which code for the polypeptide Glu-Gly-Gly-Gly-Ser. In addition, the second cluster is preceded by another unusual cluster of nucleotides (position 2284–2319), which code for

TABLE I

Coding capacity of M13 DNA and hydrophobicity of its products

The values in parentheses refer to the processed virion protein products of genes VIII and III. The N-terminal peptides cleaved off during this processing reaction at the membrane are presented by VIIIp and IIIp respectively. Hydrophobicity of the protein products has been calculated as described by Dayhoff et al. (1976).

Gene	Nucleotides	Stop codon	Amino acids	Mol. weight of protein	Hydrophobicity (%)
II	1230	TAA	410	46,117	31.0
X	333	TAA	111	12,670	30.6
V	261	TAA	87	9,666	31.0
VII	99	TGA	33	3,587	42.4
IX	96	TGA	32	3,654	40.6
VI	336	TAA	112	12,264	50.9
I	1044	TAA	348	39,500	29.6
IV	1278	TAG	426	45,791	33.5
VIII	219	TGA	73	7,622	34.2
	(150)	–	(50)	(5,234)	28.0
VIIIp.	69	–	23		47.8
III	1272	TAA	424	44,748	20.0
	(1218)	–	(406)	(42,675)	18.7
IIIp.	54	–	18		50.0

a threefold repeat of the tetrapeptide Gly-Gly-Gly-Ser.

Gene VI is located from position 2856 up to 3194. Its reading frame was established by analysing the M13 gene VI amber mutants am6-H1, am6-H2, am6-H3, am6-H6 and am6-H7. Interestingly, they all showed a C → T change at position 3066 of the DNA sequence (van Wezenbeek and Schoenmakers, 1979). The gene VI-protein is predicted to be 112 amino acids long. It is characterized by a very high Leu (21.4 mol%) and Ile (11.6 mol%) content, and the protein appears to be extremely hydrophobic in nature (Table I). The latter character might be a reason for the failure to demonstrate its synthesis so far by in vitro translation experiments and its detection in M13-infected cells as well. However, our recent Edman degradation analysis of the capsid proteins present in M13 virions have indicated that one of the two hitherto unidentified additional phage protein components, i.e. D-protein, might be the mature product of gene VI (G. Simons, unpublished data).

The in vitro product of gene I is about 35 000 daltons. Analysis of gene I amber mutant amI-H7 showed a C → T transition at position 3262 (van Wezenbeek and Schoenmakers, 1979). The only continuous translational reading frame in the gene I region extends from 3196 up to 4242. This corresponds to a product size of 348 amino acids (mol.wt. 39 500) in agreement with the size of the in vitro product. Gene IV extends from position 4220–5500, implying a 23-nucleotide overlap between gene IV and gene I. The deduced sequence of the ribosome binding site of gene IV (cf. Ravetch et al., 1977a) corresponds exactly to positions 4204–4227, in agreement with the reading frame for gene IV as deduced by analysis of the gene IV amber mutant R143. The latter mutation was found at the CAG codon at position 5264. The molecular weight 45 791 of gene IV, as based on DNA sequence, corresponds to the size of the gene IV product synthesized in vitro (Konings et al., 1975). For a synopsis of the M13 coding regions and products see Table I.

(c) Non-coding regions

The nucleotide sequence presented in Fig. 3 shows that most M13 genes are located close to each other. Apart from the intragenic location of "X" within gene II, there is only a substantial overlap between the C-terminal part of gene I and the N-terminal part

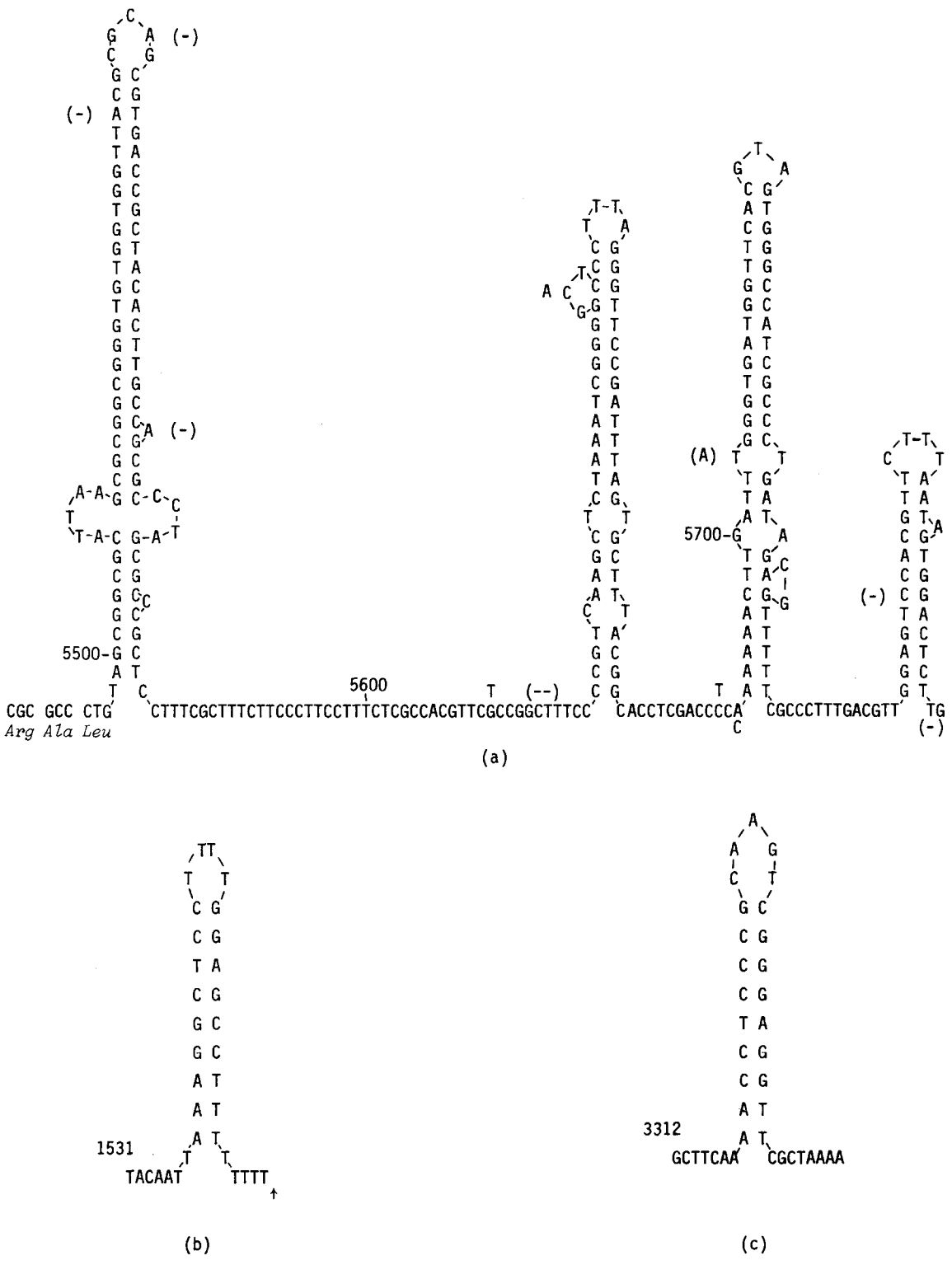

Fig. 4. (a) Secondary structure in and around the origin of replication of viral and complementary M13 DNA strand synthesis. Included are the nucleotides which differ in f1 (in parentheses) and in fd (without parentheses). Nucleotides which are deleted in f1 are indicated by (–). (b) Secondary structure of the rho-independent terminator of transcription between genes VIII and III. The position at which mRNA synthesis terminates is indicated by an arrow. (c) Secondary structure of a postulated rho-dependent termination site immediately distal to the gene VI-boundary.

of gene IV, whereas the overlaps between genes VII, IX and VIII are limited to a single nucleotide within the translational start and stop signals of these genes. The noncoding part of M13 DNA amounts to 573 nucleotides, which is about 8.9% of the genome. The largest intergenic regions are between genes VIII and III and between IV and II. Both regions encompass important regulatory elements for either replication or expression of the M13 viral genome.

The region between gene IV and II has been the subject of study in several research groups, as it contains the sites of initiation and termination of both the complementary and viral DNA strand synthesis (Tabak et al., 1974; Horiuchi and Zinder, 1976; Van den Hondel, 1976; Suggs and Ray, 1977). Initiation of complementary strand synthesis involves synthesis of an RNA primer by rifampicin-sensitive *E. coli* RNA polymerase. Recognition of the replication origin by this enzyme is most probably determined by the high secondary structure of the origin region, which prevents it from being covered by *E. coli* DNA-binding protein (Schaller et al., 1976). The energetically most favourable hairpin-like structures present in this region are presented in Fig. 4. The first large hairpin is located immediately distal to gene IV at position 5499 till 5576. Its secondary structure seems not to be correlated with replication since insertion of DNA fragments of various sizes at the positions 5563–5564 and 5571–5572 do not destroy the capability of phage replication (Herrmann et al., 1978; 1980; Ray and Kook, 1978; J. Schoenmakers, unpublished data). The latter hairpin might contain, however, an important regulatory signal for transcription since there is convincing evidence for the existence of a rho-induced termination signal immediately after gene IV (Edens, 1978; Smits et al., 1980). In the presence of *E. coli* binding protein, RNA polymerase binds to phage fd DNA and protects a unique DNA region from nuclease digestion (Schaller et al., 1976). This unique region contains the two large hairpins at position 5624–5678 and 5691–5750, and the latter are therefore considered as the target structures for recognition by RNA polymerase and synthesis of the initiating primer-RNA. The nucleotide sequence of the RNA primer has been determined (Geider et al., 1978). The RNA initiated at nucleotide 5756, is approx. 30 nucleotides long and hence complementary to one strand only of the third hairpin loop.

Viral DNA strand synthesis is thought to occur by a rolling-circle mechanism of replication. Accordingly, initiation and termination of viral DNA synthesis would be at the nick-site in the viral strand of supercoiled RFI. The site at which gene II-protein, which acts as a nickase, introduces a discontinuity in the viral strand of phage fd, has been determined (Meyer et al., 1979). This nick corresponds to position 5780–5781 of the M13 sequence and lies in a region with a two-fold symmetry (Fig. 4).

The second large intergenic region is located between genes VIII and III and contains the central terminator of transcription. Analysis of in vitro transcripts and coupled transcription-translation data have clearly shown that initiation of transcription occurs at nine different sites along the circular M13 genome but, in the absence of rho, all in vitro transcripts terminate at this unique site (Konings and Schoenmakers, 1979). Consequently, all M13 transcripts share an identical 3′-OH terminal nucleotide sequence. RNA sequence analyses have shown that the transcripts terminate in a stretch of eight U residues and contain a tight hairpinlike structure very near their 3′-OH end (Sugimoto et al., 1977; Edens, 1978) (Fig. 4). The nucleotide sequence that corresponds to this terminator is located immediately distal to gene VIII at position 1538–1564.

Several M13 intergenic regions are extremely short. The appearance of such short silent regions of either one or two nucleotides between the translational start and stop signals of M13 cistrons is rather intriguing since they appear to occur between genes that by genetic criteria form an operon, i.e. between genes V and VII and within the gene cluster III, VI and I. We do not know the possible function of these small silent regions, but it remains possible that upon reading of a message the 30S ribosomal subunit that 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, in turn, suggests that translational starts at very short intergenic regions are determined by the translation frequency of its proximal gene. Although direct proof for this assumption is still lacking, it seems more than a coincidence that the start signal of gene IX overlaps with the termination signal of its proximal gene VII, as there exists solid evidence now that the recently discovered gene IX also forms

part of the gene V–VII operon (G. Simons, unpublished data).

(d) M13 promoters

M13 has nine promoters. Initially, a number of promoter sites were identified by binding RNA polymerase to purified restriction fragments (Seeburg and Schaller, 1975; Okamoto et al., 1975). By length measurements of RNA synthesized on various DNA fragments and subsequent coupled and uncoupled transcription-translation, Edens et al. (1976; 1978a,b) showed that three mRNAs start with pppA and come from promoters preceding genes VI, I and IV and that one mRNA, probably starting with pppU, is derived from the gene III-promoter. In analogous experiments they demonstrated that five RNA transcripts start with pppG and come from promoters preceding genes VIII, V, "X" and II and from an internal startpoint within gene II. The sequence of the 5′-end of the gene VIII message has been determined and located on the M13 DNA sequence (Sugimoto et al., 1977; Hulsebos and Schoenmakers, 1978). Its position is shown in Table II along with the other promoter regions in the M13 DNA sequence. The identification of the M13 promoters is based on homology with fd (Schaller et al., 1978) and with the common features identified in several *E. coli* promoters. These common features are: (*i*) a sequence similar to TATAATPu centered about 8 nucleotides from the mRNA initiation point and (*ii*) a sequence similar to TGTTGACAATT centered about 35 nucleotides from the mRNA starting point (Siebenlist, 1979). The presence of both characteristic sequences at positions where the 5′-ends of the individual M13 transcripts already have been mapped supports the identification of the M13 promoters. It should be noted that no promoter sequence homology is found in front of gene VII and that the gene VIII messages are formed from an initiation point preceding gene IX.

All the fd promoter sequences and their M13 analogues given in Table II have been deduced from in vitro studies and, until recently, any evidence was lacking as to whether these in vitro promoters function as such in the infected cell. However, studies by Rivera et al. (1978) have now demonstrated that the 5′-end of the gene VIII message in vivo starts at exactly the same promoter position as its in vitro counterpart. Moreover, our recent investigations on the distribution and lengths of M13 phage messages in the infected cell (Smits et al., 1980) and our cloning experiments with various M13 promoters inserted into the promoter-deficient pBR322-derived plasmid pBRH2 (P. van Wezenbeek, unpublished data) have demonstrated that the given promoters preceding genes II, V and VIII are operative in vivo.

The same argument holds for the transcriptional termination site(s). Rivera et al. (1978) and Smits et al. (1980) have clearly shown that termination of in vivo transcription occurs immediately distal to gene VIII, at exactly the same sequence position as found for the in vitro RNA transcripts. It is also clear that in the infected cell termination of transcription is not limited to this site. Evidence is accumulating that termination also occurs immediately distal to gene IV and distal to the gene III (VI) region (Smits et al., 1980). A very good candidate for the latter termination signal would be the region containing the

TABLE II

Promoter sequences of M13 DNA

M13 sequences containing promoter sites for *E. coli* RNA polymerase (Konings and Schoenmakers, 1978) are lined up for maximal homology in the recognition site region and the Pribnow Box region; consequently the distance between these two regions varies by ± one basepair. Frequent and less frequent occurring nucleotides within the "ideal" sequence of promoter sites, as shown at the top of the table, are denoted by capital and small letters respectively (Siebenlist, 1979). Nucleotides which are idential to this "ideal" sequence are underlined. Bases which differ with fd are denoted by asterisks.

		t TGTTGACAAtTT T t t TG$^{G}_{t}$TATAAT g c T	
		. -30 -20 -10 . 0	
$G_{0.18}$	(VIII)	*1155* AATCTCCGTTGTACTTTGTT TCGCGCTTGGTATAATCGCTGGGGGTC	
$A_{0.49}$	(I)	*3086* CCCGTCTAATGCGCTTCCCT GTTTTTATGTTATTCTCTCTGTAAAGG	
$G_{0.06}$	(X)	*381* TCTTTTTGATGCAATCCGCT TTGCTTCTGACTATAATAGTCAGGGTAA	
$G_{0.92}$	(II)	*5926* TATTAACGTTTACAATTTAA ATATTTGCTTATACAATCTTCCTGTTTT	
$X_{0.25}$	(III)	*1500* AATTCACCTCGAAAGCAAGC TGATAAACCGATACAATTAAAGGCTCCT	
$A_{0.64}$	(IV)	*4055* TTGATAAATTCACTATTGAC TCTTCTCAGCGTCTTAATCTAAGCTATCG	
$G_{0.12}$	(V)	*786* CCAACGTCCTGACTGGTATAATGAGCCAGTTCTTAAAATCGCATAAGGTA	
$A_{0.44}$	(VI)	*2716* GGTAAACCATATGAATTTTC TATTGATTGTGACAAAATAAACTTATTCC	
$G_{0.99}$	(II′)	*6201* TCTTTACCTACACATTACTC AGGCATTGCATTTAAAATATATGAGGGTT	

self-complementary sequence from position 3319–3341 (cf. Fig. 4).

(e) M13 ribosome binding sites

Table III shows the sequences preceding the initiation codons of the M13 genes. They all show complementarity to the 3′-OH terminal sequence of 16S ribosomal RNA which is characteristic for ribosome binding sites (Shine and Dalgarno, 1974). The ribosome binding sites, as isolated from phage f1 transcripts, of genes VIII, V and IV have previously been confirmed by RNA sequencing (Pieczenick et al., 1974; Ravetch et al., 1977a).

All M13 genes use ATG as initiation codon with the exception of gene III, which initiates at a GTG codon. Furthermore, it is obvious that gene VII has the lowest potential of Shine-Dalgarno base pairing among all filamentous phage genes. Taniguchi and Weissmann (1978) demonstrated that interaction of the ribosome binding site with fMet-tRNA plays an essential role in the formation of 70S initiation complexes. Ribosome binding was substantially enhanced in case the first base following the ATG initiation codon was mutated from G to A. From Table III it can be seen that all M13 genes have an A following the initiation codon with the exception of genes VII, VI and I. Protein synthesis under the direction of the latter three genes is extremely low both in vivo and in vitro (Model and Zinder, 1974; Konings et al., 1975; Smits et al., 1978).

Table III

DNA sequences of ribosome binding sites in M13 DNA

Nucleotides complementary to the 3′-OH terminus of 16S ribosomal RNA are underlined.

16S RNA	$3'_{OH}$AUUCCUCCACUAG-- --- --- --- ---	
GENE V	843 CATAAGGTAATTCACA·ATG ATT AAA GTT *Met Ile Lys Val*	
GENE VIII	1301 TAATGGAAACTTCCTC ATG AAA AAG TCT *Met Lys Lys Ser*	
GENE IV	4220 AAAAAAGGTAATTCAA ATG AAA TTG TTA *Met Lys Leu Leu*	
GENE II	6006 ATCAACCGGGGTACAT ATG ATT GAC ATG *Met Ile Asp Met*	
GENE X	496 ATTTGAGGGGGATTCA ATG AAT ATT TAT *Met Asn Ile Tyr*	
GENE VII	1108 GTTCCGGCTAAGTAAC ATG GAG CAG GTC *Met Glu Gln Val*	
GENE IX	1206 TCGCTGGGGGTCAAAG ATG AGT GTT TTA *Met Ser Val Leu*	
GENE III	1579 TTTGGAGATTTTCAAC GTG AAA AAA TTA *Met Lys Lys Leu*	
GENE VI	2856 ATAAGGAGTCTTAATC ATG CCA GTT CTT *Met Pro Val Leu*	
GENE I	3196 GATTGGGATAAATAAT ATG GCT GTT TAT *Met Ala Val Tyr*	

(f) Codon use

The base composition of M13 viral DNA as deduced from its nucleotide sequence is A, 24.58%; G, 20.52%; C, 20.23% and T, 34.67%. The frequent occurrence of T is not randomly distributed along the phage genome. As is the case for φX174 (Sanger et al., 1978) and to a lesser extent for G4 (Godson et al., 1978), there is a striking preference for codons that have a T in the third position. Overall 50.7% of the codons used in M13 DNA show this behaviour as compared to 43.0% in φX174 and 33.8% in G4. This preferential occurrence of T in the third position of codons is observed in all M13 genes with the exception of genes VII and VIII. The highest value has gene VI (52.2%), the lowest has gene VIII (26.7%). The overall spectrum of codon use in M13 is shown in Table IV. Some codons are used only rarely: out of a total of 196 leucine codons only eight are CUA, out of 161 glycine codons six are GGA and out of 69 arginine codons two are CGG. It is probable that these rare codons have a modulating role in translation if their corresponding tRNA is also rare, as suggested by Fiers et al. (1976). Some distributions, however, suggest otherwise. The AUA codon use in M13 is rather high despite the fact that its corresponding $tRNA^{Ile}$ is only a minor component of the bulk of isoleucine tRNAs (Harada and Nishimura, 1974).

TABLE IV

Use of codons in M13

Phe	TTT	71	Ser	TCT	99	Tyr	TAT	66	Cys	TGT	16
	TTC	36		TCC	30		TAC	12		TGC	8
Leu	TTA	64		TCA	33	ochre	TAA	6	opal	TGA	3
	TTG	28		TCG	8	amber	TAG	1	Trp	TGG	18
Leu	CTT	47	Pro	CCT	48	His	CAT	13	Arg	CGT	31
	CTC	22		CCC	9		CAC	5		CGC	16
	CTA	8		CCA	14	Gln	CAA	35		CGA	6
	CTG	27		CCG	15		CAG	43		CGG	2
Ile	ATT	72	Thr	ACT	66	Asn	AAT	86	Ser	AGT	13
	ATC	16		ACC	19		AAC	22		AGC	13
	ATA	21		ACA	11	Lys	AAA	72	Arg	AGA	10
Met	ATG	33		ACG	12		AAG	36		AGG	4
Val	GTT	96	Ala	GCT	61	Asp	GAT	74	Gly	GGT	92
	GTC	18		GCC	17		GAC	35		GGC	52
	GTA	29		GCA	29	Glu	GAA	37		GGA	6
	GTG	6		GCG	13		GAG	33		GGG	11

(g) Comparison of the M13 and fd DNA sequence

The complete nucleotide sequence of fd-DNA, as well as some parts of f1-DNA, have been reported (Beck et al., 1978; Ravetch et al., 1977b; 1979; Boeke and Model, 1979). The nucleotide sequence of the origin region of phage M13 DNA has independently been determined by Suggs and Ray (1979). The overall M13 DNA sequence appears to be only one nucleotide shorter than the fd sequence. This deletion is located within a noncoding region of M13 DNA, i.e. in the region between genes VI and I (position 3194–3195).

Between M13 and fd a total of 3.0% of the nucleotides have been interchanged. Their positions have been marked with an asterisk in the final M13 sequence (Fig. 3). Only 12 of these substitutions result in a change of the corresponding amino acid sequence (6.25%). Most of the interchanges, however, appear to be third-base changes of codons, in such a way that the amino acid sequence remains unaltered. In three cases (position 2676, 4660 and 5225) a third base change is accompanied by a first base change in such a way that the codon capacity remains the same. Of the 192 base changes observed 118 appear to be transitions (A↔G, 36; C↔T, 82) and 74 are transversions (G↔T, 13; A↔T, 36; C↔G, 7; A↔C, 18). There are 131 interchanges that are of the nature X↔T. Interchanges appear most frequently within serine, glycine and leucine codons, are low within arginine, histidine, glutamine and tyrosine, whereas the codons for tryptophane, methionine and cysteine remain unchanged. As shown in Fig. 3, the frequency of base substitutions is different among the filamentous phage genes. Genes VII, IX and VIII are rather conservative. The former two show no base changes at all, whereas in gene VIII only two base changes occur, of which one leads to an Asp→Asn interchange in the major coat protein of M13 as compared to fd and f1. A similar conservative character is apparent for genes III and VI. This is in contrast to the other M13 genes that show high frequencies of substitutions.

Base changes are also either very limited or completely absent within the control sequences of both phages. The nucleotide sequence of the central terminator of transcription located immediately distal to gene VIII and the sequences of all ribosome binding sites in M13 and fd are completely identical (except for the one base deletion at position 3194–3195). Furthermore, the majority of promoter sequences of both phages are identical. The few base changes noted in the promoter regions (cf. Table II) are all located outside the sequences that are considered as the targets for RNA polymerase recognition and binding. As far as the replication origins is concerned it is striking to note that the sequences of the three major loops in M13 and fd are exactly identical. The only

difference is a C→A substitution at position 5646, which is in that part of the loop that is not involved in base-pairing. The other nucleotide changes all occur in sequences located between the loop structures. In phage f1 several single-base deletions and one two-base deletion are noted.

Now that the detailed base sequences of M13 and fd are known, the overall picture emerges that these two phages and very probably also phage f1 have their regulatory elements as well as the sizes of their encoded products conserved, which is in agreement with the well-known homology between these class of phages. Diversification of the filamentous phage genomes is expected to occur only at the level of their synthesized products. This is in contrast to the closely related isometric phages ϕX174, G4 and S13. Despite their identical genome structure and organisation of gene function, these phages show marked differences in base sequence as well as in length of gene products with similar function (Godson et al., 1978).

Knowledge of the M13 DNA sequence now provides easy access to well-defined parts of the phage DNA molecule. It will encourage further studies on site-directed mutagenesis and cloning of regulatory elements and the construction of suitable cloning vehicles for sequencing purposes.

ACKNOWLEDGEMENTS

We thank Drs. Ken Horiuchi and Gerald Vovis for supplying the f1 amber mutants, and Dr. Heinz Schaller for his gift of the fd 122 mutant. We also acknowledge the help of Dr. Hans Meyer in setting up the computer programme. This work was supported by a grant from the Netherlands Foundation for Chemical Research (SON) with financial aid from The Netherlands Organization of Pure Research (ZWO).

REFERENCES

Beck, E., Sommer, R., Auerswald, E.A., Kurz, Ch., Zink, B., Osterburg, G., Schaller, H., Sugimoto, K., Sugisaki, H., Okamoto, T. and Takanami, M.: Nucleotide sequence of bacteriophage fd DNA. Nucl. Acids Res. 5 (1978) 4495–4503.

Boeke, J.D. and Model, P.: Molecular basis of the am8HI lesion in bacteriophage M13. Virology 96 (1979) 299–301.

Cuypers, T., Van der Ouderaa, F.J. and de Jong, W.W.: The amino acid sequence of gene 5 protein of bacteriophage M13. Biochem. Biophys. Res. Commun. 59 (1974) 557–563.

Dayhoff, M.O., Dayhoff, R.E. and Hunt, L.E.: Composition of proteins, in Dayhoff, M.O. (Ed.), Atlas of Protein Sequence and Structure. The National Biomedical Research Foundation, Silver Spring, MD, Vol. 5, Suppl., 1976, pp. 301–310.

Denhardt, D.T.: The single-stranded DNA phages. CRC Crit. Rev. Microbiol. 4 (1975) 161–223.

Edens, L.: Regulation of expression of the bacteriophage M13 genome. Initiation and termination of transcription. Ph.D. Thesis, University of Nijmegen, the Netherlands, 1978.

Edens, L., van Wezenbeek, P., Konings, R.N.H. and Schoenmakers, J.G.G.: Mapping of promoter sites on the genome of bacteriophage M13. Eur. J. Biochem. 70 (1976) 577–587.

Edens, L., Konings, R.N.H. and Schoenmakers, J.G.G.: A cascade mechanism of transcription in bacteriophage M13 DNA. Virology 86 (1978a) 354–367.

Edens, L., Konings, R.N.H. and Schoenmakers, J.G.G.: Transcription of bacteriophage M13 DNA: Existence of promoters directly preceding genes III, VI, and I. J. Virol. 28 (1978b) 835–842.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., van den Berghe, A., Volckaert, G. and Ysebaert, M.: Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260 (1976) 500–507.

Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., van de Voorde, A., van Heuverswyn, H., van Herreweghe, J., Volckaert, G. and Ysebaert, M.: Complete nucleotide sequence of SV40 DNA. Nature 273 (1978) 113–120.

Fuchs, C., Rosenvold, E.C., Honigman, A. and Szybalski, W.: A simple method for identifying the palindromic sequences recognized by restriction endonucleases: The nucleotide sequence of the *Ava*II site. Gene 4 (1978) 1–23.

Fuchs, C., Rosenvold, E.C., Honigman, A. and Szybalski, W.: Identification of palindromic sequences recognized by restriction endonucleases, as based on the tabularized sequencing data for seven viral and plasmid DNAs. Gene 10 (1980) 357–370.

Geider, K., Beck, E. and Schaller, H.: An RNA transcribed from DNA of the origin of phage fd single-stranded to replicative form conversion. Proc. Natl. Acad. Sci. USA 75 (1978) 645–649.

Glynn, I.M. and Chapell, J.B.: A simple method for the preparation of ^{32}P-labelled adenosine triphosphate of high specific activity. Biochem. J. 90 (1964) 147–149.

Godson, G.N., Barrell, B.G., Staden, R. and Fiddes, J.C.: Nucleotide sequence of bacteriophage G4 DNA. Nature

276 (1978) 236–247.

Goldsmith, M.E. and Konigsberg, W.H.: Adsorption protein of bacteriophage fd. Isolation, molecular properties, and location in virus. Biochemistry 16 (1977) 2686–2694.

Harada, F. and Nishimura, S.: Purification and characterization of AUA-specific isoleucine transfer ribonucleic acid for *Escherichia coli* B. Biochemistry 13 (1974) 300–307.

Henry, T.J. and Pratt, D.: The proteins of bacteriophage M13. Proc. Natl. Acad. Sci. USA 62 (1969) 800–807.

Herrmann, R., Neugebauer, K., Schaller, H. and Zentgraf, H.: Integration of DNA fragments coding for antibiotic resistance into the genome of phage fd in vivo and in vitro, in Denhardt, D.T., Ray, D.S. and Dressler, D. (Eds.), Single-stranded DNA Phages, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 1978, pp. 473–476.

Herrmann, R., Neugebauer, K., Pirkl, E., Zentgraf, H. and Schaller, H.: Conversion of bacteriophage fd into an efficient single-stranded DNA vector system. Mol. Gen. Genet. (1980) in press.

Horiuchi, K. and Zinder, N.D., Origin and direction of synthesis of bacteriophage f1 DNA. Proc. Natl. Acad. Sci. USA 73 (1976) 2341–2345.

Horiuchi, K., Vovis, G.F., Enea, V. and Zinder, N.D.: Cleavage map of bacteriophage f1: Location of the *Escherichia coli* B-specific modification sites. J. Mol. Biol. 95 (1975) 147–165.

Horiuchi, K., Vovis, G.F. and Model, P.: The filamentous phage genome: Genes, physical structure, and protein products, in Denhardt, D.T., Ray, D.S. and Dressler, D. (Eds.), Single-stranded DNA Phages, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1978), pp. 113–137.

Hulsebos, T. and Schoenmakers, J.G.G.: Nucleotide sequence of gene VII and of a hypothetical gene (IX) in bacteriophage M13. Nucl. Acids Res. 5 (1978) 4677–4698.

Konings, R.N.H. and Schoenmakers, J.G.G.: Transcription of the filamentous phage genome, in Denhardt, D.T., Ray, D.S. and Dressler, E. (Eds.), Single-stranded DNA Phages, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1978, pp. 507–530.

Konings, R.N.H., Hulsebos, T. and van den Hondel, C.A.: Identification and characterization of the in vitro synthesized gene products of bacteriophage M13. J. Virol. 15 (1975) 570–584.

Lyons, L.B. and Zinder, N.D.: The genetic map of the filamentous bacteriophage f1. Virology 49 (1972) 45–60.

Maxam, A.M. and Gilbert, W.: A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA 74 (1977) 560–564.

Meyer, T.F., Geider, K., Kurz, C. and Schaller, H.: Cleavage site of bacteriophage fd gene II-protein in the origin of viral strand replication. Nature 278 (1979) 365–367.

Model, P. and Zinder, N.D.: In vitro synthesis of bacteriophage f1 proteins. J. Mol. Biol. 83 (1974) 231–251.

Nakashima, Y. and Konigsberg, W.: Reinvestigation of a region of the fd bacteriophage coat protein sequence. J. Mol. Biol. 88 (1974) 598–600.

Okamoto, T., Sugimoto, K., Sugisaki, H. and Takanami, M.: Studies on bacteriophage fd DNA, II. Localization of RNA initiation sites on the cleavage map of the fd genome. J. Mol. Biol. 95 (1975) 33–44.

Pieczenick, G., Model, P. and Robertson, H.D.: Sequence and symmetry in ribosome-binding sites of bacteriophage f1 RNA. J. Mol. Biol. 90 (1974) 191–214.

Pratt, D. and Erdahl, W.S.: Genetic control of bacteriophage M13 DNA synthesis. J. Mol. Biol. 37 (1968) 181–200.

Ravetch, J.V., Horiuchi, K. and Model, P.: Mapping of bacteriophage f1 ribosome binding sites to their cognate genes. Virology 81 (1977a) 341–351.

Ravetch, J.V., Horiuchi, K. and Zinder, N.D., Nucleotide sequences near the origin of replication of bacteriophage f1, Proc. Natl. Acad. Sci. USA 74 (1977b) 4219–4222.

Ravetch, J.V., Horiuchi, K. and Zinder, N.D.: DNA sequence analysis of the defective interfering particles of bacteriophage f1. J. Mol. Biol. 128 (1979) 305–318.

Ray, D.S., Replication of filamentous bacteriophage, in Fraenkel-Conrat, H. and Wagner, R.R. (Eds.), Comprehensive Virology, Plenum, New York, 1977, pp. 105–178.

Ray, D. and Kook, K.: Insertion of the Tn*3*-transposon into the genome of the single-stranded DNA phage M13. Gene 4 (1978) 109–119.

Reddy, V.B., Thimmappaya, B., Dhar, R., Subramanian, K.N., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L. and Weissman, S.M.: The genome of simian virus 40. Science 200 (1978) 494–502.

Rivera, M.J., Smits, M.A., Quint, W., Schoenmakers, J.G.G. and Konings, R.N.H.: Expression of bacteriophage M13 DNA in vivo. Localization of the transcription initiation and termination signal of the mRNA coding for the major capsid protein. Nucl. Acids Res. 5 (1978) 2895–2912.

Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchinson (III), C.A., Slocombe, P.M. and Smith, M.: The nucleotide sequence of bacteriophage ϕX174. J. Mol. Biol. 125 (1978) 225–246.

Sanger, F., Nicklen, S. and Coulson, A.R.: DNA sequencing with chain-terminating inhibitors, Proc. Natl. Acad. Sci. USA 74 (1977) 5463–5467.

Schaller, H., Uhlmann, A. and Geider, K.: A DNA fragment from the origin of single to double strand DNA replication of bacteriophage fd. Proc. Natl. Acad. Sci. USA 73 (1976) 49–53.

Schaller, H., Beck, E. and Takanami, M., Sequence and regulatory signals of the filamentous phage genome, in Denhardt, D.T., Ray, D.S. and Dressler, D. (Eds.), Single-stranded DNA Phages, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1978, pp. 139–163.

Seeburg, P.H. and Schaller, H.: Mapping and characterization of promoters in bacteriophage fd, f1 and M13. J. Mol. Biol. 92 (1975) 261–277.

Shine, J. and Dalgarno, L.: The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome-binding sites. Proc. Natl. Acad. Sci. USA 71 (1974) 1342–1346.

Siebenlist, U.: Nucleotide sequence of the three major

early promoters of bacteriophage T7. Nucl. Acids Res. 6 (1979) 1895–1907.

Simons, G.F.M., Konings, R.N.H. and Schoenmakers, J.G.G.: Identification of two new capsid proteins in bacteriophage M13. FEBS Lett. 106 (1979) 8–12.

Smits, M.A., Simons, G., Konings, R.N.H. and Schoenmakers, J.G.G.: Expression of bacteriophage M13 DNA in vivo, I. Synthesis of phage-specific RNA and protein in minicells. Biochim. Biophys. Acta 521 (1978) 27–44.

Smits, M.A., Schoenmakers, J.G.G. and Konings, R.N.H.: Expression of bacteriophage M13 DNA in vivo, IV. Isolation, identification and characterization of phage-specific mRNA species. Eur. J. Biochem. (1980) accepted for publication.

Staden, R.: Sequence data handling by computer. Nucl. Acids Res. 4 (1977) 3037–3051.

Suggs, S. and Ray, D.S.: Replication of bacteriophage M13, XI. Localization of the origin of M13 single-strand synthesis. J. Mol. Biol. 110 (1977) 147–163.

Suggs, S. and Ray, D.S.: Nucleotide sequence of the origin for bacteriophage M13 DNA replication. Cold Spring Harbor Symp. Quant. Biol. 43 (1979) 379–388.

Sugimoto, K., Sugisaki, H., Okamoto, T. and Takanami, M.: Studies on bacteriophage fd DNA, IV. The sequence of messenger RNA to the major coat protein gene. J. Mol. Biol. 111 (1977) 487–507.

Tabak, H.F., Griffith, J., Geider, K., Schaller, H. and Kornberg, A.: Initiation of deoxyribonucleic acid synthesis, VII. A unique location of the gap in the M13 replicative duplex synthesized in vitro. J. Biol. Chem. 249 (1974) 3049–3054.

Taniguchi, T. and Weissmann, C.: Site-directed mutations in the initiator region of the bacteriophage Qβ coat cistron and their effect on ribosome binding. J. Mol. Biol. 118 (1978) 533–565.

Van den Hondel, C.A.M.J.J.: Localization of structural genes and regulatory elements on the genome of bacteriophage M13, Ph.D. Thesis, University of Nijmegen, The Netherlands, 1976.

Van den Hondel, C.A. and Schoenmakers, J.G.G.: Studies on bacteriophage M13 DNA, 1. A cleavage map of the M13 genome. Eur. J. Biochem. 53 (1975) 547–558.

Van den Hondel, C.A., Weyers, A., Konings, R.N.H. and Schoenmakers, J.G.G., Studies on bacteriophage M13 DNA, 2. The gene order of the M13 genome. Eur. J. Biochem. 53 (1975) 559–567.

Van den Hondel, C.A., Pennings, L. and Schoenmakers, J.G.G.: Restriction enzyme-cleavage maps of bacteriophage M13. Existence of an intergenic region on the M13 genome. Eur. J. Biochem. 68 (1976) 55–70.

Van Wezenbeek, P. and Schoenmakers, J.G.G.: Nucleotide sequence of the genes III, VI and I of bacteriophage M13. Nucl. Acids Res. 6 (1979) 2799–2818.

Vovis, G.F., Horiuchi, K. and Zinder, N.D.: Endonuclease R · *Eco*RII restriction of bacteriophage f1 DNA in vitro: Ordering of genes V and VII, location of an RNA promoter for gene VIII. J. Virol. 16 (1975) 674–684.

Communicated by W. Fiers.