Gene, 16 (1981) 35–58
Elsevier/North-Holland Biomedical Press

Nucleotide sequence and genome organisation of filamentous bacteriophages f1 and fd

(Restriction maps; DNA sequences; genes; reading frames; regulatory signals; near identity of f1 and M13 phages)

Ewald Beck and Barbara Zink

Mikrobiologie, Universität Heidelberg, Im Neuenheimer Feld 230, 6900 Heidelberg (F.R.G.)

(Received April 6th, 1981)
(Accepted August 4th, 1981)

SUMMARY

The DNA sequence of the filamentous phage f1, consisting of 6407 nucleotides, has been determined. When compared with the DNA sequence of the related filamentous phage fd (Beck et al., 1978), the f1 sequence is one nucleotide shorter and differs in 180 positions from the fd DNA. Only ten of these base exchanges cause amino acid exchanges in the known gene products. Most of the exchanges in f1 are the same as in M13 (Van Wezenbeek et al., 1980), showing a near identity of these two phage (there are only 59 nucleotide differences). Regulatory units for replication, transcription, and translation are in their essential parts identical in all three phage.

INTRODUCTION

The genomes of the filamentous *Escherichia coli* phage, e.g. fd, f1, and M13, consist of single-stranded circular DNAs of about 6400 nucleotides. These code for at least nine genes, whose products are involved in phage DNA replication, phage assembly and phage capsid synthesis (Marvin and Hohn, 1969; Ray, 1977). Since the propagation of the filamentous phage is catalysed mainly by host functions, their genomes have been studied for many years as model systems of regulation in *E. coli* for replication, transcription, and translation (recent reviews Ray, 1977; Schaller, 1979). In addition, the DNA of these phages was also used early as model system in the development of methods for the structural analysis of genomes. Ling (1972) sequenced a number of large pyrimidine oligonucleotides; Oertel and Schaller (1972) determined the sequence and the order of pyrimidine tracts in a pyrimidine rich segment of the fd DNA, and Sanger et al. (1973; 1974) deduced the sequence of 89 nucleotides in the f1 DNA (pos. 6321–6408) using the ribo-substitution technique. Three ribosome-binding sites were sequenced by Pieczenik et al. (1974) and the promoter site of gene X by Schaller et al. (1975) and Sugimoto et al. (1975). The DNA from the origin of replication, first isolated and characterized from a pre-initiation complex (Schaller et al., 1976) was the first continuous stretch of fd DNA to be analysed (Gray et al.,

Abbreviations: bp, base pairs; IG, intergenic region; pos., position; pDNA, DNA protected by RNA polymerase against pancreatic DNase digestion; RF, replicative form.

0378-1119/81/0000–0000/$02.75 © 1981 Elsevier/North-Holland Biomedical Press

1978), using the rapid methods for DNA sequencing newly developed by Sanger and Coulson (1975) and Maxam and Gilbert (1977). The DNA sequence of the origin of replication of defective interfering particles of f1 was also analysed (Ravetch et al., 1979). Takanami et al. (1976) and Sugimoto et al. (1977) analysed the region of the genes VII and VIII and the central terminator of transcription, sequencing RNA produced by in vitro transcription of restriction fragments. In 1978 the total fd DNA sequence was determined and published first in a preliminary version (Schaller et al., 1978), followed by the final sequence in a short publication (Beck et al., 1978). At that time about 90% of the DNA sequence of the related phage f1 also had been analysed mainly to confirm the gene reading frames by identifying the numerous silent base exchanges between the f1 and fd DNA. In this paper we discuss the experimental details of the fd and f1 DNA sequence analysis and the derived structures of genes and regulatory signals. In addition, we present the completed f1 DNA sequence. It differs by 180 base exchanges from the fd DNA sequence, only few of which cause amino acid changes in the gene products. As the DNA of the other closely related filamentous phage M13 has also been sequenced completely (Van Wezenbeek et al., 1980), a comparison of the three sequences is presented.

MATERIALS AND METHODS

(a) Bacteriophage and enzymes

The wild-type bacteriophage f1 and the f1 nonsense mutants amR5, amR7, amR124, and amR143 were from N.D. Zinder, New York. The bacteriophage fd was from H. Hoffmann-Berling, Heidelberg. The fd strain 478 which was isolated as a single plaque from the fd stock and used in the sequence analysis differs in at least one position (1859) from the fd phage from ATCC which was sequenced in part in the laboratory of M. Takanami, Kyoto. The viral DNA was converted into the double-stranded form (RF) in vitro by oligonucleotide primed synthesis as described (Gray et al., 1978). The restriction endonucleases *Hpa*II, *Hae*III, *Hin*fI, *Hha*I, *Hga*I, *Alu*I, and *Taq*I were prepared essentially as described by Roberts et al. (1976), *Acc*II, *Hph*I, and *Mbo*II were purchased from New England Biolabs, and *Sau*3A was a gift from H. Streeck, Munich. Polynucleotide kinase and calf intestinal phosphatase were from Boehringer GmbH, Mannheim. [γ-^{32}P]-ATP (spec. act. approx. 6000 Ci/mmol) was prepared as described by Johnson and Walseth (1979).

(b) 5′-End-labeling of DNA

Restriction fragments were dephosphorylated either by adding phosphatase into the cleavage mixture together with the restriction endonuclease, or in cases of flush-ended or 3′-extended ends in 50 mM Tris pH 8 at 60°C (0.02 units phosphatase per 20 μl assay; incubation time 30–60 min). The samples were phenol-extracted, desalted on a small Sephadex G75 column (2 ml disposable pipette) in 10 mM ammonium-bicarbonate pH 8.6 and lyophilised. This was found to be the best method for complete removal of the phosphatase. Phosphorylation with [γ-^{32}P]ATP and polynucleotide kinase was carried out essentially as described by Maxam and Gilbert (1980). In general 1–2 pmol cleaved RF DNA were used per assay.

(c) DNA sequencing methods

Gel electrophoresis, elution of DNA from polyacrylamide gels, separation of labeled fragment ends either by a secondary restriction enzyme cleavage or by separation of denatured strands, and the base-specific chemical modification were performed essentially as described by Maxam and Gilbert (1980). The depurination was carried out in 66% formic acid for 2–8 min at 20°C, followed by 3-fold dilution with water, three ether extractions, lyophilisation and hydrolysis in 1 M piperidine at 90°C for 1 h in an oven. Some fragments analysed on long (1 m) sequencing gels (0.4 mm thick) could be read up to position 450.

RESULTS AND DISCUSSION

(a) Sequencing strategy

For the complete analysis of the DNAs of fd and f1 ten different restriction endonucleases were used (see Fig. 1). Usually the DNA was cleaved with a particular restriction enzyme and the resulting fragments were end-labeled as a mixture and separated on polyacrylamide gels. Most of the radioactive fragments were used for the sequence analysis. Many restriction maps (e.g. from *Hha*I, *Hin*fI, *Hph*I, *Mbo*II, *Sau*3A, and *Taq*I) were not established prior to sequencing but resulted from matching overlapping sequences.

In the case of restriction endonucleases *Hga*I and *Taq*I, which each have only ten cleavage sites in the fd DNA, the restricted and end-labeled DNA was further cleaved by a second restriction enzyme before separation on a gel. By comparing fragments present before and after the second cleavage it could be deduced which new fragments had been generated by the second digestion and which were thus labeled at only one end and which could be used directly in the subsequent analysis. Using this method separation of the re-cleaved fragments on a second gel was unnecessary. However, there was often a higher degree of contamination by neighbouring bands or background which interfered with extended reading of nucleotide sequences.

The restriction enzyme *Hae*III also cleaves single-stranded DNA efficiently (Blakesley and Wells, 1975). With this enzyme single-stranded fragments could be prepared directly and used for sequencing without secondary cleavage.

Although it was usually possible to read the DNA sequences clearly, 85% of the fd DNA was sequenced in both strands to avoid mistakes that could occur at methylated bases (Ohmori et al., 1978), in regions with a distinct secondary structure, or by incorrect reading or processing of the sequence information. Care was taken that all restriction sites used to generate fragments were read through from alternative starts. This is particularly important in repetitive sequences that may contain closely spaced repeating restriction sites. Such an example occurs in the fd DNA sequence around position 2390, where a sequence of 18 nucleotides consisting of two small *Hpa*II fragments was not included in the preliminary version of the fd DNA sequence (Schaller et al., 1978). The nucleotide sequences were stored and processed, using computer programmes written in the computer language APL and established by Osterburg and Sommer (1981).

(b) The DNA sequence

The fd sequence was derived by reading serially overlapping fragments, making use of most of the different restriction cuts and both DNA strands as shown in Fig. 1. Sequencing of f1 DNA was started somewhat later. Therefore, only about 50% of the sequence analysis was carried out in both strands, since we could refer to the completed fd sequence. Nevertheless, some regions of f1 were analysed in more detail than in fd, and the f1 sequence was also used to confirm fd DNA sequences which had been determined in one strand only.

Fig. 2 shows combined sequences of the fd and the f1 DNA. The continuous sequence corresponds to the fd DNA sequence as published in 1978 (Beck et al.). About 97% of the f1 DNA is identical to the fd DNA. There exist 180 base changes, which are indicated above the fd DNA sequence. Whereas about 150 of them lie within genes, only 10 actually cause amino acid changes. The others are "silent" alterations, i.e., they involve variable bases in the codons. This fact was used already earlier as indirect evidence for the correct reading frames of the genes in the filamentous phage genome (Schaller et al., 1978; and see below). Base changes present in the M13 DNA sequence (Van Wezenbeek et al., 1980) are also included in Fig. 2. Many of them coincide with the changes in f1, demonstrating that f1 and M13 are more closely related to each other than to fd.

A series of partial sequences from fd DNA and f1 DNA published earlier (see above) could be fitted into the complete sequences. All agree essentially with our data. Two changes had to be made in regulatory regions: one at the promoter of gene VIII, where the sequence at the start of transcription is a G_5 run (not G_4 as in Takanami et al., 1976), the other at the central terminator, where the sequence at the end point of transcription is C_2T_9 (not C_2T_8C as in Ling, 1972, or C_2T_8 as in Sugimoto et al., 1977). An f1 DNA sequence of the intergenic region (IG) between genes IV and II (position 5500–6000; Fig. 3) analysed by Ravetch et al. (1977; 1979) contains several deletions of one or two nucleotides when compared with the corresponding fd DNA sequence. None of these deletions could be confirmed in our f1 DNA sequence. The corresponding region in M13, analysed first by Suggs and Ray (1978) and confirmed by the M13 DNA sequence of Van Wezenbeek

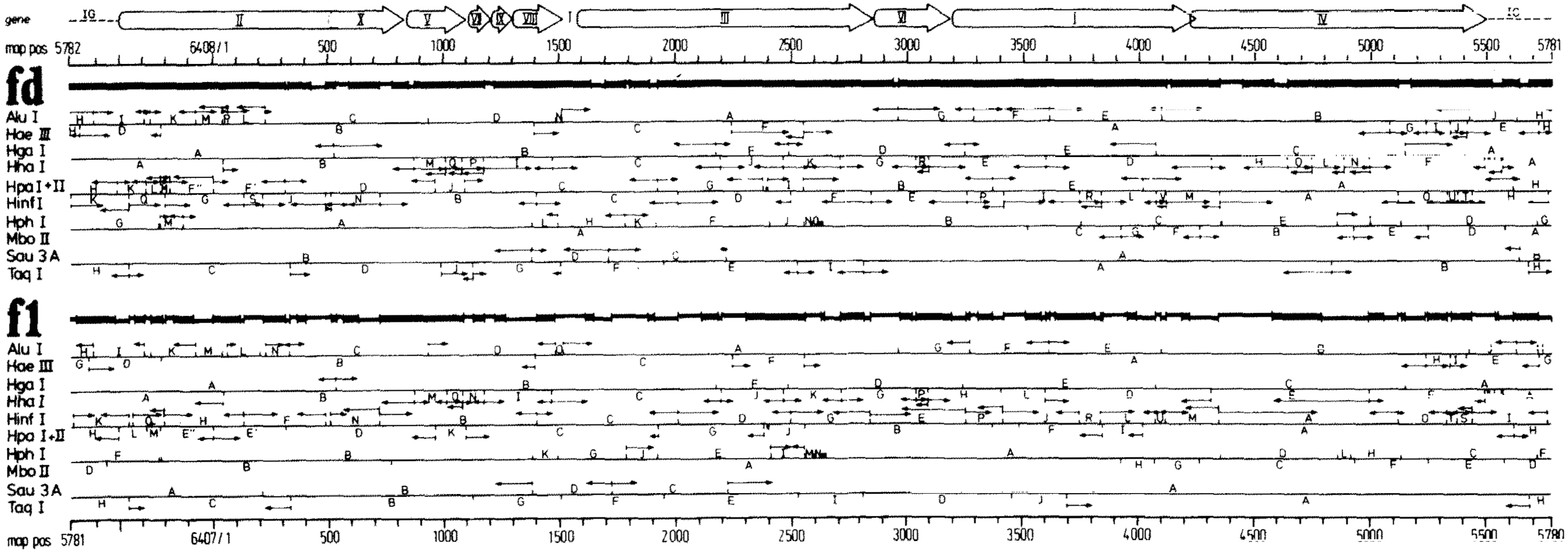

Fig. 1. Genetic and physical linearized maps of bacteriophages fd and f1, including the sequencing strategy applied. The circular phage genome is opened at the start point of viral strand replication in the intergenic region between genes IV and II (IG). The genes (roman numerals) and the central terminator of transcription (T) are indicated in the top line. The zero point of the map (see 6408/1 and 6407/1 map positions) is the single *Hin*dII (*Hpa*I) cleavage site in fd. The bars for fd and f1 each indicate how far the DNA sequence is determined in both strands (solid bar) or in the plus (= viral) strand (gap in the lower part of the bar) or in the minus strand only (gap in the upper part of the bar). The length and the orientation of the individual sequencing runs are indicated by arrows in the restriction enzyme cleavage maps.

et al. (1980) agrees with the f1 sequence except for two positions. Whereas M13 and f1 are almost identical in this region, fd differs in 23 positions of the intergenic region from these two phage.

In addition to the published results, other fd DNA sequence data for the region between position 300 and 1600 were made available (M. Takanami, personal communication). In this analysis a difference between our fd DNA and that used by Takanami was noticed: a G → A exchange at position 1859 creating an additional *Hin*fI site in Takanami's DNA. The altered restriction fragment pattern of this enzyme was demonstrated experimentally (M. Takanami, personal communication).

(c) Restriction maps

When the work on fd sequence analysis started in 1977, several restriction maps (*Hpa*II, *Hae*III, *Alu*I, *Hga*I) had already been completed for fd, f1 and/or M13. During the sequence analysis these maps were refined and maps for many other restriction enzymes established. The maps of *Taq*I, *Hha*I, *Hin*fI, *Hae*III, *Sau*3A, *Bam*HI, *Hph*I, *Mbo*II, and *Acc*II (*Tha*I) were checked experimentally by comparison of the fragment length derived from the DNA sequence with the corresponding fragment patterns on polyacrylamide gels. The recognition sites in the three filamentous phage DNAs of the best known restriction enzymes are listed in Table I. In nondenaturing polyacrylamide gels some fragments (e.g. in fd *Hpa*II-B (pos. 2552–3371), *Hpa*II-H (pos. 5615–5996) and *Taq*I-H (pos. 5648–6041) migrate more slowly than other fragments of comparable length. Such fragments usually contain extended inverted repeats, which may cause secondary structures divergent from the normal double helical form of DNA.

TABLE I

Restriction endonuclease recognition sites

The sites for fd, f1 and M13, as found by computer analysis, are compiled (c.f. also Fuchs et al., 1980). Italic numbers represent cleavage sites that are experimentally proven. No cleavage sites exist in all three phages for the enzymes *Ava*II, *Ava*III, *Bcl*I, *Bgl*I, *Bgl*II, *Eco*RI, *Hin*dIII, *Kpn*I, *Mst*I, *Pst*I, *Pvu*I, *Pvu*II, *Sac*I, *Sac*II, *Sal*I, *Sma*I, *Xho*I, and *Xba*I.

Name	Sequence	Position							
AccI	GTAGAC	6091							
AluI	AGCT	*38*	*63*	*203*^{bc}	*229*	*333*^{bc}	*934*	*1488*	*1517*
		2963	*3277*	*3613*	*4097*	*5427*	*5631*	*5888*	*6108*
		6135	*6336*						
AsuI	GGGCC	5725							
AvaI	CTCGGG	5826^{c}							
BalI	TGGCCA	5081^{ac}							
BamHI	GGATCC	*2220*	*5645*^{a}						
BbvI	$GC_{T}^{A}GC$	932	1367	2521	3132	4872	5537		
ClaI	ATCGAT	2527^{ac}	6040						
DdeI	CT.AG	233	1099	1371	1417	1784	1847	1862	1877
		1901	1973	2015	2318	2333	2348	2363	2678^{a}
		3362	4014	4041	4080	4094	4122	4282^{c}	4882^{a}
		5263	5371^{b}	6066^{bc}	6219^{bc}	6347			
EcoB	$TGA(N_8)TGCT$	*959*^{ab}	*6348*						
EcoRII	$CC_{T}^{A}GG$	1014	1966						
Fnu4HI	GC.GC	932	1367	1394	1871^{b}	2285	2288	2312	2327^{a}
		2357	2384^{ab}	2521	3132	4872	4888^{a}	5501	5515
		5537							
HaeII	RGCGCY	*2710*^{ac}	*3039*^{bc}	*4743*^{a}	*5560*	*5568*			
HaeIII	GGCC	*1396*	*2245*	*2554*	*5082*^{ac}	*5240*	*5346*	*5415*	*5726*
		5829^{a}	*5868*^{bc}	*6181*					
HgaI	GACGC,GCGTC	*526*	*2164*	*2479*	*3238*	*4084*	*5159*		
HgiAI	$G_{T}^{A}GC_{T}^{A}C$	4744^{bc}	5466						
HhaI	GCGC	*44*	*873*^{ab}	*1011*	*1085*	*1177*	*2195*	*2467*	*2711*
		3040	*3096*	*3409*^{bc}	*3599*	*4313*	*4642*^{a}	*4744*^{a}	*4886*^{a}
		4996	*5491*	*5504*	*5513*	*5535*	*5561*	*5569*	
HindII	GTYRAC	4715^{b}	*6406*						
HinfI	GA.TC	*136*	*216*^{ac}	*490*	*511*	*723*	*1403*^{ab}	*2011*	*2497*
		2845	*3259*	*3419*	*3743*	*3839*	*4073*	*4118*	*4350*
		5121	*5330*	*5376*	*5439*	*5767*	*5789*	*6043*	*6062*^{bc}
		6199	*6406*						
HpaI	GTTAAC	*6406*							
HpaII	CCGG	*314*	*966*	*1095*	*1924*	*2378*	*2390*^{ab}	*2396*	*2552*
		3371	*3843*^{bc}	*4019*	*5615*	*5996*	*6119*	*6179*	*6221*^{a}
HphI	GGTGA,TCACC	*1376*	*1503*	*1774*	*1909*	*2398*	*2542*	*2581*	*2620*
		2626	*2635*	*3740*^{a}	*4347*^{ab}	*4365*^{a}	*4849*	*4924*^{bc}	*5118*
		5707	*6163*	*6189*	*6286*^{a}				
MboII	GAAGA,TCTTC	*781*^{bc}	*3529*^{a}	*3913*	*4076*	*4272*	*4938*	*5256*	*5588*
		5963^{bc}							
MnlI	CCTC,GAGG	254	331^{a}	373	484	560	587	625	655
		782^{a}	1039	1088	1231	1297	1318	1326	1345
		1373	1416	1506	1663^{a}	1732	1834	1849	1864
		1879	1897	1945	2008	2020	2218	2263	2269
		2320	2335	2350	2365	2369^{c}	2673^{c}	2677	2894
		3052	3322	3337	3353	3704	4022	4308^{ac}	4399^{a}
		4699	4773^{bc}	4821^{bc}	4836	4922^{bc}	4927	5037^{a}	5348
		5417	5448	5682	5688^{a}	6096	6114	6244	6349
RsaI	GTAC	173	280	1022	1165	1769	1796	1889	1905
		1970	2133	3468	3669^{bc}	4191	4381	5385^{bc}	5462^{a}
		5487	6001	6323	6390^{a}				
Sau3A	GATC	*216*^{b}	*1382*	*1714*	*2221*	*5646*^{a}			
SfaNI	GATGC,GCATC	25	388	1031^{a}	1354	3980	4851		
ThaI	CGCG	43	347	1119	1176	2466	2710^{b}	3356	3410^{bc}
		3600	3953	4314	4425^{b}	4641^{a}	4887^{a}	4995	5490
		5514	5534	5910					
TaqI	TCGA	*336*	*988*^{a}	*1127*	*1508*	*1949*	*2528*	*2635*^{ab}	*3456*^{bc}
		3695^{bc}	4666^{c}	*4834*^{a}	*5684*	*6041*			
XhoII	RGATCY	215^{b}	2220	5645^{a}					

[a] Site exists in fd only; [b] site exists in f1 only; [c] site exists in M13 only.

(d) Genes and gene products

A genetic map of the eight known genes of the filamentous phage was established by Lyons and Zinder (1972) and correlated later with the size of the gene products determined on SDS gels (Model and Zinder, 1974) and the physical maps (Vovis et al., 1975). In addition to these approximate positions and gene lengths the amino acid sequences of gene V and gene VIII proteins (Nakashima and Koningsberg, 1974; Nakashima et al., 1974) were determined,

```
              . T       .         .         .         .         .         .         .         .         .
   1 AACGCTACTACCATTAGTAGAATTGATGCCACCTTTTCAGCTCGCGCCCCAAATGAAAATATAGCTAAACAGGTTATTGACCATTTGCGAAATGTATCTA
     AsnAlaThrThrIleSerArgIleAspAlaThrPheSerAlaArgAlaProAsnGluAsnIleAlaLysGlnValIleAspHisLeuArgAsnValSerAsn   R1
                                                                  ***                                         R2
                   ******   ***                           ***         ***         ***                  ***    R3

              .         .         .         .         .         .         .         .         .         .
 101 ATGGTCAAACTAAATCTACTCGTTCGCAGAATTGGGAATCAACTGTTACATGGAATGAAACTTCCAGACACCGTACTTTAGTTGCATATTTAAAACATGT
       GlyGlnThrLysSerThrArgSerGlnAsnTrpGluSerThrValThrTrpAsnGluThrSerArgHisArgThrLeuValAlaTyrLeuLysHisVal    R1
                                                                                   ***         ***            R2
               ***                                    MetGlu***                                               R3
        G     .       (C).        .         .         .         .         .         .     C   .         .
 201 TGAACTACAGCACCAGATTCAGCAATTAAGCTCTAAGCCATCCGCAAAAATGACCTCTTATCAAAAGGAGCAATTAAAGGTACTGTCTAATCCTGACCTG
      GluLeuGlnHisGlnIleGlnGlnLeuSerSerLysProSerAlaLysMetThrSerTyrGlnLysGluGlnLeuLysValLeuSerAsnProAspLeu    R1
                                ***                     ***                     ***                           R2
     ***                              ***                                                   ***   ***         R3
          G   .         .         .  A      . [A]     .         .         .         .         .     C   .
 301 TTGGAATTTGCTTCCGGTCTGGTTCGCTTTGAGGCTCGAATTGAAACGCGATATTTGAAGTCTTTCGGGCTTCCTCTTAATCTTTTTGATGCAATTCGCT
     LeuGluPheAlaSerGlyLeuValArgPheGluAlaArgIleGluThrArgTyrLeuLysSerPheGlyLeuProLeuAsnLeuPheAspAlaIleArgPhe  R1
                                                            ***                                               R2
                                  ***         ***                                 ***      *** gene X start   R3
              .         T         .         .         .         .         .         .         .         .
 401 TTGCTTCTGACTATAATAGACAGGGTAAAGACCTGATTTTTGATTTATGGTCATTCTCGTTTTCTGAACTGTTTAAAGCATTTGAGGGGGATTCAATGAA
       AlaSerAspTyrAsnArgGlnGlyLysAspLeuIlePheAspLeuTrpSerPheSerPheSerGluLeuPheLysAlaPheGluGlyAspSerMetAsn    R1
                                      ***                                                           ***       R2
            ***   ******      ***            ***                     ***      ***      ***                    R3
              .         .         .         .         . T       .         .     T   .         .         .
 501 TATTTATGACGATTCCGCAGTATTGGACGCTATCCAGTCTAAACATTTTACAATTACCCCCTCTGGCAAAACTTCCTTTGCAAAAGCCTCTCGCTATTTT
      IleTyrAspAspSerAlaValLeuAspAlaIleGlnSerLysHisPheThrIleThrProSerGlyLysThrSerPheAlaLysAlaSerArgTyrPhe    R1
          -----------------------------------------------------------------------------------------------     R2
           ***                              ***                                                               R3
          T   .         .A  C     .         .         .T        .         .         .         .     A   .
 601 GGTTTCTATCGTCGTCTGGTTAATGAGGGTTATGATAGTGTTGCTCTTACCATGCCTCGTAATTCCTTTTGGCGTTATGTATCTGCATTAGTTGAGTGTG
     GlyPheTyrArgArgLeuValAsnGluGlyTyrAspSerValAlaLeuThrMetProArgAsnSerPheTrpArgTyrValSerAlaLeuValGluCysGly  R1
     ----------------------------------------------------------------------------------------***              R2
                         ******      ******                     ***                              ***          R3
              .       C .         . T       .         .         .         .         .  T      .         .
 701 GTATTCCTAAATCTCAATTGATGAATCTTTCCACCTGTAATAATGTTGTTCCGTTAGTTCGTTTTATTAACGTAGATTTTTCCTCCCAACGTCCTGACTG
       IleProLysSerGlnLeuMetAsnLeuSerThrCysAsnAsnValValProLeuValArgPheIleAsnValAspPheSerSerGlnArgProAspTrp    R1
                       ******                              ***               ***                              R2
            ***                           ******                        ***                        ***        R3
                       gene II end            gene V start
              .         .         .         .C        .         .     A   .   [C]   .(C)  [T] .         .
 801 GTATAATGAGCCAGTTCTTAAAATCGCATAAGGTAATTCAAAATGATTAAAGTTGAAATTAAACCGTCTCAAGCGCAATTTACTACCCGTTCTGGTGTTT
      TyrAsnGluProValLeuLysIleAla***            ***                                                           R1
                                                    ***   ***   ***                                           R2
        ******         ***            ***      MetIleLysValGluIleLysProSerGlnAlaGlnPheThrThrArgSerGlyValSer   R3
              .         .         .         .         .         .         .[T]      .        T. T      .
 901 CTCGTCAGGGCAAGCCTTATTCACTGAATGAGCAGCTTTGTTACGTTGATTTGGGTAATGAATATCCGGTGCTTGTCAAGATTACTCTCGACGAAGGTCA
```

```
                             ***-------------------------------------------------------------------------     R1
                                 ***               ***      ******                                            R2
       ArgGlnGlyLysProTyrSerLeuAsnGluGlnLeuCysTyrValAspLeuGlyAsnGluTyrProValLeuValLysIleThrLeuAspGluGlyGln    R3
           C  .         .         .T        . [T]     .         .      C  .         .         .         .
1001 GCCAGCGTATGCGCCTGGTCTGTACACCGTGCATCTGTCCTCGTTCAAAGTTGGTCAGTTCGGTTCTCTTATGATTGACCGTCTGCGCCTCGTTCCGGCT
     -----------------------------------------------------------------------***                              R1
                                                                                ***                     ***   R2
      ProAlaTyrAlaProGlyLeuTyrThrValHisLeuSerSerPheLysValGlyGlnPheGlySerLeuMetIleAspArgLeuArgLeuValProAla     R3
  gene V end       gene VII start .         .         .         .         .         .         .         .
1101 AAGTAACATGGAGCAGGTCGCGGATTTCGACACAATTTATCAGGCGATGATACAAATCTCCGTTGTACTTTGTTTCGCGCTTGGTATAATCGCTGGGGGT
            MetGluGlnValAlaAspPheAspThrIleTyrGlnAlaMetIleGlnIleSerValValLeuCysPheAlaLeuGlyIleIleAlaGlyGly     R1
                                                    ***                                    ***                R2
     Lys***gene IX start                                                                                      R3
gene VII end            .         .         .         .         .         .         .         .  gene IX end
1201 CAAAGATGAGTGTTTTAGTGTATTCTTTCGCCTCTTTCGTTTTAGGTTGGTGCCTTCGTAGTGGCATTACGTATTTTACCCGTTTAATGGAAACTTCCTC
     GlnArg***MetPhe***                        ***                                    ***                     R1
                                                               ***                                            R2
          MetSerValLeuValTyrSerPheAlaSerPheValLeuGlyTrpCysLeuArgSerGlyIleThrTyrPheThrArgLeuMetGluThrSerSer    R3
            gene VIII start       I         .         .         .         .         .         .         .
1301 ATGAAAAAGTCTTTAGTCCTCAAAGCCTCCGTAGCCGTTGCTACCCTCGTTCCGATGCTGTCTTTCGCTGCTGAGGGTGACGATCCCGCAAAAGCGGCCT
                                                                            ***   ***                         R1
     MetLysLysSerLeuValLeuLysAlaSerValAlaValAlaThrLeuValProMetLeuSerPheAlaAlaGluGlyAspAspProAlaLysAlaAlaPhe   R2
      ***         ***               ***                                          -------------------------    R3
       [A]    .         .         .         .         .         .         .         .         .         .
1401 TTGACTCCCTGCAAGCCTCAGCGACCGAATATATCGGTTATGCGTGGGCGATGGTTGTTGTCATTGTCGGCGCAACTATCGGTATCAAGCTGTTTAAGAA
      ***                                                                                          ***        R1
       AspSerLeuGlnAlaSerAlaThrGluTyrIleGlyTyrAlaTrpAlaMetValValValIleValGlyAlaThrIleGlyIleLysLeuPheLysLys    R2
     ----------------------------------------------------------------------------------------------------     R3
           gene VIII end.         .         .         .         .               gene III start .
1501 ATTCACCTCGAAAGCAAGCTGATAAACCGATACAATTAAAGGCTCCTTTTGGAGCCTTTTTTTTTGGAGATTTTCAACGTGAAAAAATTATTATTCGCAA
                                         ***                                       MetLysLysLeuLeuPheAlaIle   R1
      PheThrSerLysAlaSer******                                                        ***                     R2
     ----------------------------------------------------------------------------------------------------     R3
              .         .         .         .         .         .    C    .         .        (T)        .
1601 TTCCTTTAGTTGTTCCTTTCTATTCTCACTCCGCTGAAACTGTTGAAAGTTGTTTAGCAAAACCTCATACAGAAAATTCATTTACTAACGTCTGGAAAGA
       ProLeuValValProPheTyrSerHisSerAlaGluThrValGluSerCysLeuAlaLysProHisThrGluAsnSerPheThrAsnValTrpLysAsp    R1
           ***                                             ***                                                R2
     ----------------------------------***      ***                                       ***                 R3
              .         .         .     [T] .         .         .   A     .         .         .         .
1701 CGACAAAACTTTAGATCGTTACGCTAACTATGAGGGCTGTCTGTGGAATGCTACAGGCGTTGTGGTTTGTACTGGTGACGAAACTCAGTGTTACGGTACA
      AspLysThrLeuAspArgTyrAlaAsnTyrGluGlyCysLeuTrpAsnAlaThrGlyValValValCysThrGlyAspGluThrGlnCysTyrGlyThr     R1
                ***              ------------------------------------------------------------------------     R2
                             ***   ***                                          ***                           R3
              .         .         .         .         .         .         .    (C)  .         .         .
1801 TGGGTTCCTATTGGGCTTGCTATCCCTGAAAATGAGGGTGGTGGCTCTGAGGGTGGCGGTTCTGAGGGTGGCGGTTCTGAGGGTGGCGGTACTAAACCTC
     TrpValProIleGlyLeuAlaIleProGluAsnGluGlyGlyGlySerGluGlyGlyGlySerGluGlyGlyGlySerGluGlyGlyGlyThrLysProPro   R1
     ----------------------------------------------------------------------------------------------------     R2
                               ***   ***            ***            ***            ***            ***          R3
```

```
                                (T)
1901 CTGAGTACGGTGATACACCTATTCCGGGCTATACTTATATCAACCCTCTCGACGGCACTTATCCGCCTGGTACTGAGCAAAACCCCGCTAATCCTAATCC
       GluTyrGlyAspThrProIleProGlyTyrThrTyrIleAsnProLeuAspGlyThrTyrProProGlyThrGluGlnAsnProAlaAsnProAsnPro    R1
     ----------------------------------------------------------------------------------------------------      R2
      ***      ***                                                           ***            ***   ***          R3
                                                                          G
2001 TTCTCTTGAGGAGTCTCAGCCTCTTAATACTTTCATGTTTCAGAATAATAGGTTCCGAAATAGGCAGGGTGCATTAACTGTTTATACGGGCACTGTTACT
      SerLeuGluGluSerGlnProLeuAsnThrPheMetPheGlnAsnAsnArgPheArgAsnArgGlnGlyAlaLeuThrValTyrThrGlyThrValThr     R1
     -------------------------------------------------------------------------***                          R2
            ***               ***                    ******           ***                                  R3

2101 CAAGGCACTGACCCCGTTAAAACTTATTACCAGTACACTCCTGTATCATCAAAAGCCATGTATGACGCTTACTGGAACGGTAAATTCAGAGACTGCGCTT
     GlnGlyThrAspProValLysThrTyrTyrGlnTyrThrProValSerSerLysAlaMetTyrAspAlaTyrTrpAsnGlyLysPheArgAspCysAlaPhe   R1
                                                                  ---------------------------------------      R2
              ***      ***                                            ***                ***               R3
                                                                                (T)
2201 TCCATTCTGGCTTTAATGAGGATCCATTCGTTTGTGAATATCAAGGCCAATCGTCTGACCTGCCTCAACCTCCTGTCAATGCTGGCGGCGGCTCTGGTGG
       HisSerGlyPheAsnGluAspProPheValCysGluTyrGlnGlyGlnSerSerAspLeuProGlnProProValAsnAlaGlyGlyGlySerGlyGly    R1
     ----------------------------------------------------------------------------------------------------      R2
                  ******                ***                    ***                                         R3
                                T                                          [A]           [T]  .[T]
2301 TGGTTCTGGTGGCGGCTCTGAGGGTGGCGGCTCTGAGGGTGGCGGTTCTGAGGGTGGCGGCTCTGAGGGTGGCGGTTCCGGTGGCGGCTCCGGTTCCGGT
      GlySerGlyGlyGlySerGluGlyGlyGlySerGluGlyGlyGlySerGluGlyGlyGlySerGluGlyGlyGlySerGlyGlyGlySerGlySerGly     R1
     ----------------------------------------------------------------------------------------------------      R2
                        ***            ***            ***            ***                                  *** R3
                      G
2401 GATTTTGATTATGAAAAAATGGCAAACGCTAATAAGGGGGCTATGACCGAAAATGCCGATGAAAACGCGCTACAGTCTGACGCTAAAGGCAAACTTGATT
     AspPheAspTyrGluLysMetAlaAsnAlaAsnLysGlyAlaMetThrGluAsnAlaAspGluAsnAlaLeuGlnSerAspAlaLysGlyLysLeuAspSer   R1
     ------------------------------------------***     ------------------------------------------------      R2
           ***   ***              ******                        ***                ***   ***         ***   R3
                                   (C)
2501 CTGTCGCTACTGATTACGGTGCTGCTATCGATGGTTTCATTGGTGACGTTTCCGGCCTTGCTAATGGTAATGGTGCTACTGGTGATTTTGCTGGCTCTAA
       ValAlaThrAspTyrGlyAlaAlaIleAspGlyPheIleGlyAspValSerGlyLeuAlaAsnGlyAsnGlyAlaThrGlyAspPheAlaGlySerAsn    R1
     ----------------------------------------------------------------------------------------------------      R2
                ***                              ***                ***   ***              ***          ***R3
                                                                                 [C]
                                                                              CC (T)    A
2601 TTCCCAAATGGCTCAAGTCGGTGACGGTGATAATTCACCTTTAATGAATAATTTCCGTCAATATTTACCTTCTTTGCCTCAGTCGGTTGAATGTCGCCCT
      SerGlnMetAlaGlnValGlyAspGlyAspAsnSerProLeuMetAsnAsnPheArgGlnTyrLeuProSerLeuProGlnSerValGluCysArgPro     R1
     ----------------------------------------******                                                        R2
                           ***    ******                ***                                      ***       R3
             [A]
      T      (C)
2701 TATGTCTTTGGCGCTGGTAAACCATATGAATTTTCTATTGATTGTGACAAAATAAACTTATTCCGTGGTGTCTTTGCGTTTCTTTTATATGTTGCCACCT
     TyrValPheGlyAlaGlyLysProTyrGluPheSerIleAspCysAspLysIleAsnLeuPheArgGlyValPheAlaPheLeuLeuTyrValAlaThrPhe   R1
                                                          ***          ------------------------------------    R2
                       ***      ***        ***   ***         gene VI start                                 R3
                     [T]                     gene III end
2801 TTATGTATGTATTTTCGACGTTTGCTAACATACTGCGTAATAAGGAGTCTTAATCATGCCAGTTCTTTTGGGTATTCCGTTATTATTGCGTTTCCTCGGT
       MetTyrValPheSerThrPheAlaAsnIleLeuArgAsnLysGluSer***                                                 R1
     ----------------------------------------------------------------------------------------------------      R2
                               ***          ******         MetProValLeuLeuGlyIleProLeuLeuLeuArgPheLeuGly   R3
              .         .         . (A)    T.         .         .         .         .         .         .
2901 TTCCTTCTGGTAACTTTGTTCGGCTATCTGCTTACTTTCCTTAAAAAGGGCTTCGGTAAGATAGCTATTGCTATTTCATTGTTTCTTGCTCTTATTATTG
               ***                                                ***                                        R1
     -----------------------------------------***            ***                                             R2
     PheLeuLeuValThrLeuPheGlyTyrLeuLeuThrPheLeuLysLysGlyPheGlyLysIleAlaIleAlaIleSerLeuPheLeuAlaLeuIleIleGly  R3
              .         .         .         .   T     .        C.         .[T]      .         .         .
3001 GGCTTAACTCAATTCTTGTGGGTTATCTCTCTGATATTAGCGCACAATTACCCTCTGATTTTGTTCAGGGCGTTCAGTTAATTCTCCCGTCTAATGCGCT
                                                                                   ***                       R1
         ***                        ***   ***               ***                                 ***          R2
       LeuAsnSerIleLeuValGlyTyrLeuSerAspIleSerAlaGlnLeuProSerAspPheValGlnGlyValGlnLeuIleLeuProSerAsnAlaLeu   R3
                                                                                        gene VI end Δ gene I start
              .         .         .         .         .         .         .         .         .         .
3101 TCCCTGTTTTTATGTTATTCTCTCTGTAAAGGCTGCTATTTTCATTTTTGACGTTAAACAAAAAATCGTTTCTTATTTGGATTGGGATAAATAAATATGG
                               ***                                                                           R1
                                                     ***   ***                              ***      MetAla  R2
      ProCysPheTyrValIleLeuSerValLysAlaAlaIlePheIlePheAspValLysGlnLysIleValSerTyrLeuAspTrpAspLys***          R3
              .         .         .         .         .         .         .         .         .         .
3201 CTGTTTATTTTGTAACTGGCAAATTAGGCTCTGGAAAGACGCTCGTTAGCGTTGGTAAGATTCAGGATAAAATTGTAGCTGGGTGCAAAATAGCAACTAA
                                                   ***      ***         ***                           ***    R1
       ValTyrPheValThrGlyLysLeuGlySerGlyLysThrLeuValSerValGlyLysIleGlnAspLysIleValAlaGlyCysLysIleAlaThrAsn   R2
                 ***         ***                                                ***            ***           R3
              .         .         .         .         .         .         .         .       A .         .
3301 TCTTGATTTAAGGCTTCAAAACCTCCCGCAAGTCGGGAGGTTCGCTAAAACGCCTCGCGTTCTTAGAATACCGGATAAGCCTTCTATTTCTGATTTGCTT
        ***                                       ***               ***         ***            ***           R1
      LeuAspLeuArgLeuGlnAsnLeuProGlnValGlyArgPheAlaLysThrProArgValLeuArgIleProAspLysProSerIleSerAspLeuLeu    R2
             ***                                                                                             R3
             G. C       .         T         .    C    .      C  .  G      .         .         T        T.
3401 GCTATTGGTCGTGGTAATGATTCCTACGACGAAAATAAAAACGGTTTGCTTGTTCTTGATGAATGCGGTACTTGGTTTAATACCCGTTCATGGAATGACA
                   ******               ***                  ******               ***               ***      R1
     AlaIleGlyArgGlyAsnAspSerTyrAspGluAsnLysAsnGlyLeuLeuValLeuAspGluCysGlyThrTrpPheAsnThrArgSerTrpAsnAspLys  R2
               ------------------------------------------------------------------------------------------    R3
              .         .         . A       .      A  .         .         .      C  .         .         .
3501 AGGAAAGACAGCCGATTATTGATTGGTTTCTTCATGCTCGTAAATTGGGATGGGATATTATTTTTCTTGTTCAGGATTTATCTATTGTTGATAAACAGGC
                        ***                  ***                                             ******          R1
       GluArgGlnProIleIleAspTrpPheLeuHisAlaArgLysLeuGlyTrpAspIleIlePheLeuValGlnAspLeuSerIleValAspLysGlnAla   R2
     ----------------------------------------------------------------------------------------------------    R3
              .       (A) T       .      T  .         .      (G).T        T         .     A   .      G  .
3601 GCGTTCTGCATTAGCTGAACACGTTGTTTATTGTCGCCGTCTGGACAGAATTACTTTACCCTTTGTCGGCACTTTATATTCTCTTGTTACTGGCTCAAAA
                    ***                                                                                      R1
      ArgSerAlaLeuAlaGluHisValValTyrCysArgArgLeuAspArgIleThrLeuProPheValGlyThrLeuTyrSerLeuValThrGlySerLys    R2
     -----------***                                                                                          R3
              .         .      C  .         . C       .         .         .         .         .     G   .
3701 ATGCCTCTGCCTAAATTACATGTTGGTGTTGTTAAATATGGTGATTCTCAATTAAGCCCTACTGTTGAGCGTTGGCTTTATACTGGTAAGAATTTATATA
                ***                  ***      ***                     ***                  ***         ***   R1
     MetProLeuProLysLeuHisValGlyValValLysTyrGlyAspSerGlnLeuSerProThrValGluArgTrpLeuTyrThrGlyLysAsnLeuTyrAsn  R2
                                                         ***                                                 R3
              .T        .      [T].         .   C     .     T   .    G    .         .         .         .
3801 ACGCATATGACACTAAACAGGCTTTTTCCAGTAATTATGATTCAGGTGTTTATTCATATTTAACCCCTTATTTATCACACGGTCGGTATTTCAAACCATT
```

```
            ***   ***               ***   ***                                                                R1
       AlaTyrAspThrLysGlnAlaPheSerSerAsnTyrAspSerGlyValTyrSerTyrLeuThrProTyrLeuSerHisGlyArgTyrPheLysProLeu   R2
                                                                 ***                                    ***  R3
              .         .         .         .         .         .         .  T      .         .         .
3901 AAATTTAGGTCAGAAGATGAAATTAACTAAAATATATTTGAAAAAGTTTTCTCGCGTTCTTTGTCTTGCGATAGGATTTGCATCAGCATTTACATATAGT
                                ***                                                                ***       R1
      AsnLeuGlyGlnLysMetLysLeuThrLysIleTyrLeuLysLysPheSerArgValLeuCysLeuAlaIleGlyPheAlaSerAlaPheThrTyrSer    R2
          ***         ***   ***            ***                              ***                              R3
              .         .         .         .         .         .         .         .         .         .
4001 TATATAACCCAACCTAAGCCGGAGGTTAAAAAGGTAGTCTCTCAGACCTATGATTTTGATAAATTCACTATTGACTCTTCTCAGCGTCTTAATCTAAGCT
                   ***         ***                     ***   ******         ***               ***            R1
     TyrIleThrGlnProLysProGluValLysLysValValSerGlnThrTyrAspPheAspLysPheThrIleAspSerSerGlnArgLeuAsnLeuSerTyr  R2
         ***                           ***                                                         ***       R3
              .         .         .         .         .         .         .AC       .         .         .
4101 ATCGCTATGTTTTCAAGGATTCTAAGGGAAAATTAATTAATAGCGACGATTTACAGAAGCAAGGTTATTCCATCACATATATTGATTTATGTACTGTTTC
                           ***            ******                                       ***                   R1
       ArgTyrValPheLysAspSerLysGlyLysLeuIleAsnSerAspAspLeuGlnLysGlnGlyTyrSerIleThrTyrIleAspLeuCysThrValSer   R2
                        gene IV start ***                                                                    R3
                                    gene I end
     C        .         .         .         .         .         .         .         .    [G]  .         .
4201 AATTAAAAAAGGTAATTCAAATGAAATTGTTAAATGTAATTAATTTTGTTTTCTTGATGTTTGTTTCATCATCTTCTTTTGCTCAAGTAATTGAAATGAA
        ***      ***      ***      ***   ***               ***                              ***      ***     R1
      IleLysLysGlyAsnSerAsnGluIleValLysCysAsn***                                                ***          R2
                         MetLysLeuLeuAsnValIleAsnPheValPheLeuMetPheValSerSerSerSerPheAlaGlnValIleGluMetAsn   R3
                            [A]                                       [T]
           (T).         .[T](T)   .         .  [T] [C].   [C]   .  [T](G)[G]  [A](G).    T [T].         A
4301 TAATTCGCCTCTGCGCGATTTCGTGACTTGGTATTCAAAGCAAACAGGTGAATCTGTTATTGTCTCACCTGATGTTAAAGGTACAGTGACTGTATATTCC
                            ***                                                           ***                R1
     ***                                             ***                  ***   ***                          R2
      AsnSerProLeuArgAspPheValThrTrpTyrSerLysGlnThrGlyGluSerValIleValSerProAspValLysGlyThrValThrValTyrSer    R3
              . A       . C    (G).        T.         .  (A)    .         .    T    .(C)      (T)      G.
4401 TCTGACGTTAAGCCTGAAAATTTACGCAATTTCTTTATCTCTGTTTTACGTGCTAATAATTTTGATATGGTTGGCTCAATTCCTTCCATAATTCAGAAAT
                                                                                             ***             R1
       ***   ***   ***                                    ******   ***                                       R2
     SerAspValLysProGluAsnLeuArgAsnPhePheIleSerValLeuArgAlaAsnAsnPheAspMetValGlySerIleProSerIleIleGlnLysTyr  R3
         T    .C A     (C)        .         .        A.         .         .         .         .         .
4501 ATAACCCAAATAGTCAGGATTATATTGATGAATTGCCATCATCTGATATTCAGGAATATGATGATAATTCCGCTCCTTCTGGTGGTTTCTTTGTTCCGCA
                                -------------------------------------------------------------------------    R1
      ***      ***            ******            ***            *********                                     R2
       AsnProAsnSerGlnAspTyrIleAspGluLeuProSerSerAspIleGlnGluTyrAspAspAsnSerAlaProSerGlyGlyPhePheValProGln   R3
              .         . T       .         .  G      .        C.A   [C]  .      [A]. [G]     .T        .
4601 AAATGATAATGTTACTCAAACATTTAAAATTAATAACGTTCGCGCAAAGGATTTAATAAGGGTTGTAGAATTGTTTGTTAAATCTAATACATCTAAATCC
     -----------------------------------------------------******      ***                                    R1
        ******               ***   ******                                          ***   ***      ***        R2
      AsnAspAsnValThrGlnThrPheLysIleAsnAsnValArgAlaLysAspLeuIleArgValValGluLeuPheValLysSerAsnThrSerLysSer    R3
                                                   [A]
              .    [A]  .C  C     TC       T.    T (T).         .         .    T    .C    (A) .         .
4701 TCAAATGTATTATCTGTTGATGGTTCTAACTTATTAGTAGTTAGCGCCCCTAAAGATATTTTAGATAACCTTCCGCAATTTCTTTCTACTGTTGATTTGC
```

```
                                       ******                     ***                                        R1
                      ***      ***            ***      ***            ***                        ***         R2
     SerAsnValLeuSerValAspGlySerAsnLeuLeuValValSerAlaProLysAspIleLeuAspAsnLeuProGlnPheLeuSerThrValAspLeuPro  R3
              .         .  G  T  G. A  T    .         .         .       A .         .        T.         .
4801 CAACTGACCAGATATTGATTGAAGGATTAATTTTCGAGGTTCAGCAAGGTGATGCTTTAGATTTTTCCTTTGCTGCTGGCTCTCAGCGCGGCACTGTTGC
                    ***         ***                           ***                                            R1
         ***            ***                           ***                                                    R2
       ThrAspGlnIleLeuIleGluGlyLeuIlePheGluValGlnGlnGlyAspAlaLeuAspPheSerPheAlaAlaGlySerGlnArgGlyThrValAla   R3
     A  C     .         . C  C    .         .     T   .         .         T         .         .         .
4901 TGGTGGTGTTAATACTGACCGTCTAACCTCTGTTTTATCTTCTGCGGGTGGTTCGTTCGGTATTTTTAACGGCGATGTTTTAGGGCTATCAGTTCGCGCA
                            ***                                                      ***                     R1
              ***   ***                                                ***                                   R2
      GlyGlyValAsnThrAspArgLeuThrSerValLeuSerSerAlaGlyGlySerPheGlyIlePheAsnGlyAspValLeuGlyLeuSerValArgAla    R3
              .         .         .        A.         .         .         .   [C]   .  (T)    .         .
5001 TTAAAGACTAATAGCCATTCAAAAATATTGTCTGTGCCTCGTATTCTTACGCTTTCAGGTCAGAAGGGTTCTATTTCTGTTGGCCAGAATGTCCCTTTTA
      ***                                                                                                    R1
             ******                                                                                          R2
     LeuLysThrAsnSerHisSerLysIleLeuSerValProArgIleLeuThrLeuSerGlyGlnLysGlySerIleSerValGlyGlnAsnValProPheIle  R3
              .   G     .         .         .         .  [A]    .         .A        C     C   .   [T]   .
5101 TTACTGGTCGTGTAACTGGTGAATCTGCCAATGTAAATAATCCATTTCAGACGGTTGAGCGTCAAAATGTTGGTATTTCTATGAGTGTTTTTCCCGTTGC
                 ***                  ***                                             ***i                   R1
                        ***               ***               ***                                              R2
       ThrGlyArgValThrGlyGluSerAlaAsnValAsnAsnProPheGlnThrValGluArgGlnAsnValGlyIleSerMetSerValPheProValAla   R3
              .         .  C G    .T     C  .         .         .         .         .         .         .
5201 AATGGCTGGCGGTAATATTGTTTTAGATATAACCAGTAAGGCCGATAGTTTGAGTTCTTCTACTCAGGCAAGTGATGTTATTACTAATCAAAGAAGTATT
                            ***   ***                  ***                  -----------------------------    R1
                 ***                     ***      ***                        ***         ***                 R2
      MetAlaGlyGlyAsnIleValLeuAspIleThrSerLysAlaAspSerLeuSerSerSerThrGlnAlaSerAspValIleThrAsnGlnArgSerIle    R3
      [T]     .         .      A  .      [A].         .         T         .   (G)   .   C  A  .         .
5301 GCGACAACGGTTAATTTGCGTGATGGTCAGACTCTTTTGCTCGGTGGCCTCACTGATTACAAAAACACTTCTCAAGATTCTGGTGTGCCGTTCCTGTCTA
     ----------------------------------------------------------------------------------------------------    R1
                ***      ***                              ***                                          ***   R2
     AlaThrThrValAsnLeuArgAspGlyGlnThrLeuLeuLeuGlyGlyLeuThrAspTyrLysAsnThrSerGlnAspSerGlyValProPheLeuSerLys  R3
                                                                                              gene IV end
              .         .         .    C    .  [C]    .         . A       .         .            ====== .
5401 AAATCCCTTTAATCGGCCTCCTGTTTAGCTCCCGTTCTGATTCTAACGAGGAAAGCACGTTGTACGTGCTCGTCAAAGCAACCATAGTACGCGCCCTGTA
     ---------***                                                                        ***                 R1
                              ***         ***   ***                                                          R2
       IleProLeuIleGlyLeuLeuPheSerSerArgSerAspSerAsnGluGluSerThrLeuTyrValLeuValLysAlaThrIleValArgAlaLeu***   R3
              .         .         .         .         .         .         .         .         .         .
5501 GCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTT
                            -----------------------------------------------------------------------------    R1
              ***                            ***                     ***                                     R2
                                                                                                             R3
              .   G     .         .         .      C  .         .         .         .         C A       .
5601 TCTCGCCACGTTCTCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGATCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCTCCAAAAACTT
     ----------------------------------------------------------------------------------------------------    R1
                                        ***               ***           ---------------------------------    R2
                                                                                                        ***  R3
```

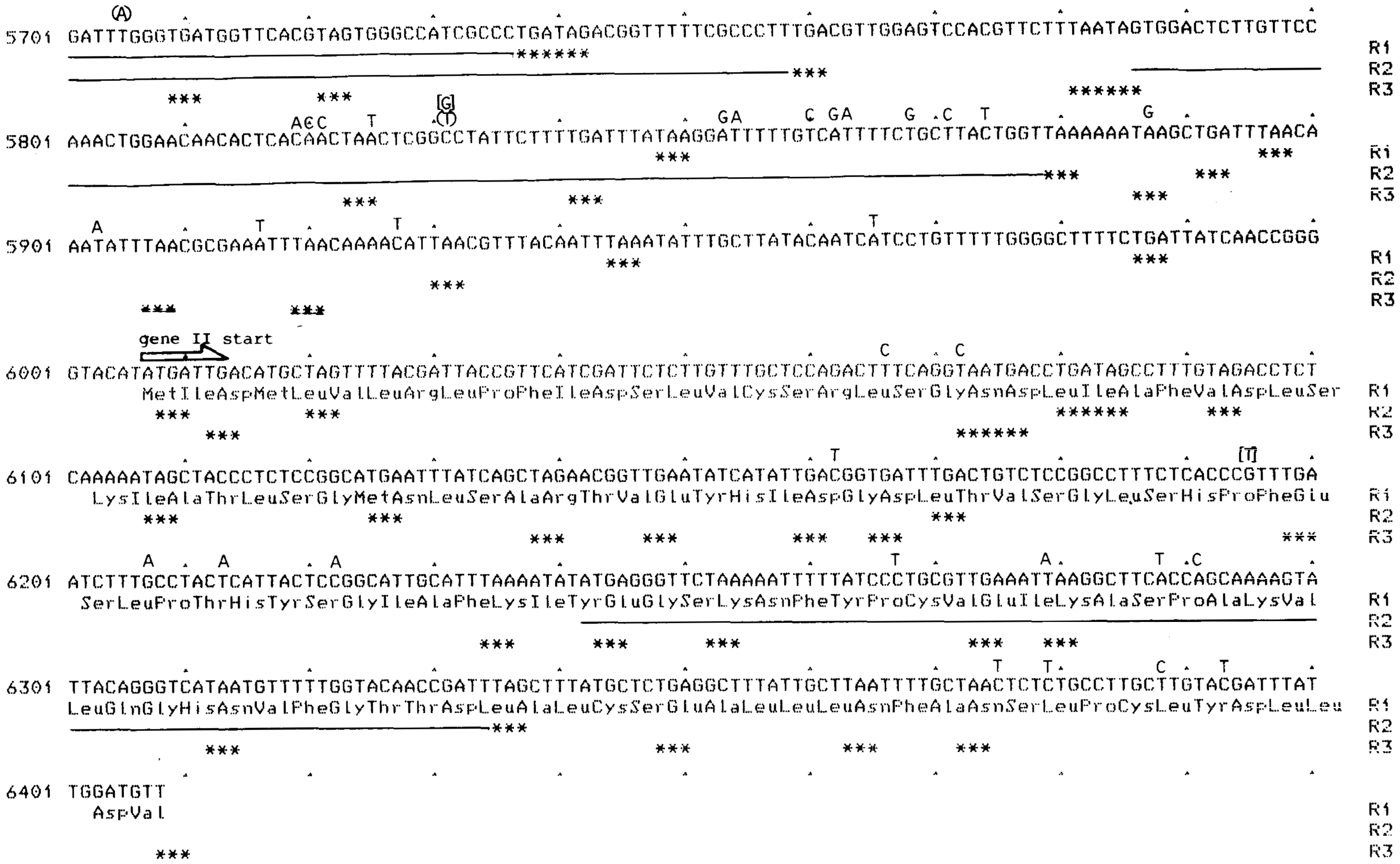
(A)
5701 GATTTGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCC
R1
R2
R3
[G]
AЄC T (T) GA C GA G C T G
5801 AAACTGGAACAACACTCACAACTAACTCGGCCTATTCTTTTGATTTATAAGGATTTTTGTCATTTTCTGCTTACTGGTTAAAAAATAAGCTGATTTAACA
R1
R2
R3
A T T T
5901 AATATTTAACGCGAAATTTAACAAAACATTAACGTTTACAATTTAAATATTTGCTTATACAATCATCCTGTTTTTGGGGCTTTTCTGATTATCAACCGGG
R1
R2
R3
gene II start
C C
6001 GTACATATGATTGACATGCTAGTTTTACGATTACCGTTCATCGATTCTCTTGTTTGCTCCAGACTTTCAGGTAATGACCTGATAGCCTTTGTAGACCTCT
MetIleAspMetLeuValLeuArgLeuProPheIleAspSerLeuValCysSerArgLeuSerGlyAsnAspLeuIleAlaPheValAspLeuSer
R1
R2
R3
T [T]
6101 CAAAAATAGCTACCCTCTCCGGCATGAATTTATCAGCTAGAACGGTTGAATATCATATTGACGGTGATTTGACTGTCTCCGGCCTTTCTCACCCGTTTGA
LysIleAlaThrLeuSerGlyMetAsnLeuSerAlaArgThrValGluTyrHisIleAspGlyAspLeuThrValSerGlyLeuSerHisProPheGlu
R1
R2
R3
A A A T A T C
6201 ATCTTTGCCTACTCATTACTCCGGCATTGCATTTAAAATATATGAGGGTTCTAAAAATTTTTATCCCTGCGTTGAAATTAAGGCTTCACCAGCAAAAGTA
SerLeuProThrHisTyrSerGlyIleAlaPheLysIleTyrGluGlySerLysAsnPheTyrProCysValGluIleLysAlaSerProAlaLysVal
R1
R2
R3
T T C T
6301 TTACAGGGTCATAATGTTTTTGGTACAACCGATTTAGCTTTATGCTCTGAGGCTTTATTGCTTAATTTTGCTAACTCTCTGCCTTGCTTGTACGATTTAT
LeuGlnGlyHisAsnValPheGlyThrThrAspLeuAlaLeuCysSerGluAlaLeuLeuLeuAsnPheAlaAsnSerLeuProCysLeuTyrAspLeuLeu
R1
R2
R3
6401 TGGATGTT
AspVal
R1
R2
R3

Fig. 2. Nucleotide sequence of the DNA of bacteriophage fd. Base exchanges for f1 and M13 are indicated: no brackets = exchange is common for f1 and M13, parentheses ( ) = exchange exists for f1 only; brackets [ ] = exchange exists for M13 only. The starts and stops and the amino acid sequences for the gene products are shown. Stop codons in the three reading frames are indicated as asterisks. The lines represent possible reading frames of more than 30 amino acid residues.

allowing an exact correlation of these genes with the nucleotide sequence. The start sites of two genes were determined, or later on confirmed, respectively, by sequence analysis of the N-terminal amino acids of the proteins (gene III: Goldsmith and Koningsberg, 1977; gene II: Meyer et al., 1980). Start sites of three other (unknown) genes were determined by three ribosome-binding sites sequenced by Pieczenik et al. (1974). Two of these sites were subsequently correlated to genes V and VIII, the best characterized products of the filamentous phage. The third site defines the start of gene IV.

The exact positions of the genes are shown in the final nucleotide sequence (Fig. 2). There is only one possible reading frame for each gene within the limits derived from the genetic map. The other reading frames contain many stop codons. Continuous reading frames longer than 30 amino acids that are not allied with known genes are also indicated in Fig. 2. In most cases translation is made unlikely by the absence of a Shine-Dalgarno sequence except for two theoretical peptides of 65 and 42 amino acids starting at positions 3417 and 4528, respectively. There is no genetic evidence for additional genes in the filamentous phage.

The arrangement of the genes reflects an economical usage of DNA. The only two noncoding regions between genes VIII and III and between genes IV and II contain the central terminator of transcription and the origins of DNA replication, respectively. Gene IX overlaps with one nucleotide at each end with the neighbouring genes VII and VIII. Most other genes are separated by one or two nucleotides from each other. These one or two nucleotides obviously function to change the reading frame for the following gene and to avoid the synthesis of fusion proteins in case of readthrough by suppression. This principle is also demonstrated by the deletion of one of the two nucleotides in the intergenic space between genes VI and I in the f1 and M13 DNA compared with the fd DNA (pos. 3195). In both cases a change of the reading frames is maintained. In contrast to the eicosahedral phages (ϕX174, G4), there are no gene overlaps of more than one nucleotide in the filamentous phage genome except of a short run of 20 bases, which is common for genes I and IV (pos. 4221–4240). There is obviously no selection pressure to limit the genome length of these phages. By insertion of heterogeneous DNA into the genome of fd, hybrid phages could be constructed which were several times longer than wild-type fd (Herrmann et al., 1978).

Since mistakes can occur in establishing a DNA sequence for different reasons – as mentioned above – it is necessary to have useful criteria to control the derived reading frame of the genes. The best controls are data from protein sequence analysis. Amino acid sequences, partial or complete, were available for the genes V, VIII, III, and II. Another control that contributes greatly to the credibility of the complete sequence derived from the DNA, amino acid composition, was available for the gene III product.

There are also several possible ways to check the correctness of a reading frame at the nucleotide level. A *first* simple possibility is based on the fact that about 50% of all triplets within the genes end with a T residue similarly to the eicosahedral phage ϕX174 and G4 (Sanger et al., 1977; Godson et al., 1978). The filamentous phage mostly use codons with the highest number of T's for all amino acids (see Table II). Although it cannot be used as an exact proof for the correctness of a sequence in a specific short region, this phenomenon can be used to confirm the reading frame over a longer distance. *Secondly*, the filamentous phage obviously do not have overlapping genes. There are in the unused reading frames stop codons every 30–40 nucleotides on the average. A *third* test used exists in a comparison of the DNA sequences of the closely related phages fd and f1: we determined 280 base exchanges, 120 of them within the genes. Only ten of these result in amino acid exchanges, the others are "silent" in the correct reading, i.e. they concern the variable bases in the codons. The *fourth* and most conclusive but also most elaborate method used is the determination of base exchanges to amber mutants. For almost all genes the DNA sequence of one or several amber mutants was analysed (see Table IV).

The genes are arranged in three functional groups in the genome: replication (genes II and V), capsid (genes IX, VIII, III, and VI), and morphogenesis (genes I and IV). According to this, the gene VII protein (unknown function) could either be involved in replication or be part of the virion from its position on the DNA.

The most significant functional and biochemical features of the gene products and the criteria for

TABLE II

Codon usage in fd

Phe	TTT	67	Ser	TCT	92	Tyr	TAT	65	Cys	TGT	16
	TTC	39		TCC	33		TAC	14		TGC	8
Leu	TTA	65		TCA	35	ochre	TAA	5	opal	TGA	3
	TTG	32		TCG	9	amber	TAG	1	Trp	TGG	18
	CTT	49	Pro	CCT	46	His	CAT	12	Arg	CGT	32
	CTC	17		CCC	9		CAC	6		CGC	16
	CTA	6		CCA	13	Gin	CAA	35		CGA	5
	CTG	26		CCG	18		CAG	43		CGG	1
Ile	ATT	72	Tor	ACT	60	Asn	AAT	82	Ser	AGT	14
	ATC	16		ACC	23		AAC	23		AGC	11
	ATA	20		ACA	15	Lys	AAA	73	Arg	AGA	11
Met	ATG	33		ACG	11		AAG	34		AGG	5
Val	GTT	98	Ala	GCT	59	Asp	GAT	72	Gly	GGT	95
	GTC	18		GCC	16		GAC	38		GGC	51
	GTA	25		GCA	28	Glu	GAA	40		GGA	5
	GTG	11		GCG	17		GAG	31		GGG	9

localisation of the reading frames are summarized in Table III. Gene II shows two possible **ATG** start sites in positions 6007 and 6016. Based on a better Shine-Dalgarno sequence we have predicted that the former one must correspond to the protein start (Schaller et al., 1978). This was confirmed by determination of the N-terminus for 90% of the gene II product, using radiolabel Edman degradation (Meyer et al., 1980). However, about 10% of the protein showed amino acids in positions that correspond to a start at the second **ATG** codon. Whether the two proteins which were co-isolated from a membrane fraction and which are not to be distinguished on **SDS** gels have different biological functions is not known.

The existence of a ninth gene in the filamentous phage genome between the genes **VII** and **VIII** was predicted already from the preliminary fd sequence by Schaller et al. (1978). A gap of 94 nucleotides with no known coding or regulatory function shows a continuous reading frame, whereby the first and the last triplet each have an overlap of one nucleotide with the adjacent genes. The protein predicted from the sequence consists of 32 amino acids with a composition (6 Ser, 2 Arg, no His) that is similar to that of the C-protein, a minor capsid component which has been detected in highly purified **f1** and M13 phage (Simons et al., 1979). There are no amber mutants known for the gene **IX.** This is explained by the **DNA** sequence, which shows that possible amber codons can only be created by transversions in positions 1223, 1249, and 1274. Hydroxylaminic treatment used to construct amber mutants of the filamentous phage (Lyons and Zinder, 1972) could induce only transitions.

Gene III protein contains a remarkably high degree of glycine residues (16%). Most of these are clustered in repetitive sequences: the sequence Glu-Gly-Gly-Gly-Ser appears three times around amino acid position 95 and four times around position 255, accompanied by repetitions of Gly-Gly-Gly-Ser at both sites. In the **DNA** of an fd Tn5 derivative a stretch of 30 nucleotides, corresponding to two of the Glu-Gly-Gly-Gly-Ser repeats (amino acids 253-262) is deleted (Auerswald, 1979). The deleted amino acids are obviously not essential for gene III function. Around amino acid 375 the protein seems to be variable too, since base exchanges in positions 2699, 2702, and 2710 result in amino acid changes in **f1** and M13 relative to fd (Table V).

M_r estimations of the gene III protein differed between 55 000 and 68 000 depending on the SDS gel system used (Goldsmith and Koningsberg, 1977). Even the lowest value differs markedly from that derived from the DNA sequence (Mr 42 660). The unusual clustering of glycine residues may alter the binding of SDS and therefore the migration of the protein on gels.

The reading frame of gene IV extends 20 nucleotides back into the 3′ end of gene I. Most of the nucleotide sequence of the ribosome binding site is homologous to the ribosome binding site of gene V. Sequence homologies are also recognizable in the first 20 nucleotides of the two genes. Moreover, parts of this homologous sequence are repeated within the coding region of gene IV (positions 3901–3931 and 4285–4305), in both cases centered around ATG codons, probably reflecting an evolutionary pathway. The reading frame of this gene ends with a TAG codon in position 5499. This stop codon lies directly at the beginning of the largest hairpin in the viral DNA. This structure may help to terminate transcription and/or translation, resulting in the correct length of the gene product even in UAG suppressor strains.

(e) Regulatory signals

Structures of regulatory signals concerning all three levels of phage development, replication, transcription and translation are recognizable on the DNA. The regulatory unit of replication lies in an intergenic region of 508 bp (= IG) between the end of gene IV and the start of gene II. This DNA segment can be folded into several large hairpin structures (Fig. 3). The existence of these structures in the viral DNA was demonstrated by their resistance to S1 nuclease (Gray et al., 1978). Their significance is indicated by the fact that their sequence is conserved in all three filamentous phage. Although the IG is the most variable region in the filamentous phage genome – more than 5% of the bases of fd differ from f1 and M13 – none of these base changes lies in the stem of a hairpin but only in the regions between. Most of the region between the end of gene IV (pos. 5501) and the hairpin C does not seem to be necessary for propagation of the DNA. By cloning parts of the IG in plasmid pBR322 under conditions not permissive for ColE1-directed replication it has been shown that positions 5727–5868, containing the start site of the *ori*-RNA (Geider et al., 1978) and the nicking site of the gene II protein (Meyer et al., 1979), is sufficient in the presence of helper virus for replication of the hybrid replicon (Cleary and Ray, 1980; Sommer, 1981). In a pseudo wild-type fd, a revertant from a transposon-containing phage, a deletion of 64 nucleotides (pos. 5553–5618) was observed which removed part of hairpin A and the pyrimidine-rich region between hairpins A and B (Auerswald, 1979). This region is obviously dispensable for phage multiplication. A ColE1 vector containing the left half of the IG from the end of gene IV to the *Hae*III fragment mentioned above (pos. 5489–5868) not only replicates under phage control but also packs single (plus)-strands of the vector into phage-coat protein efficiently (Sommer, 1981). The "packing origin" must therefore be localized in the left half of the IG; this indicates that hairpin A is involved in this function, as speculated earlier (Schaller, 1979).

In vitro transcription starts at several promoters which are located in front of almost each gene (except of genes VII and IX) and proceeds unidirectionally to a single Rho-independent stop signal immediately after gene VIII. In this way more RNA copies are produced of the genes proximal to the central terminator than of the more distal genes. This polar effect is amplified by the fact that the strongest promoters are located in the region preceding the termination signal. The products of the genes encoded by this region are the most abundantly needed proteins of the phage. In vivo this "cascade" model of transcription could be proved only in the region between the IG and the central terminator. Some of the RNA species of this region are obviously processed in the cell (Smits et al., 1980). In the other part of the genome there exist at least two additional Rho-dependent termination signals which cease transcription behind genes VI and IV (Smits et al., 1980). In addition to the promoters known already (reviewed e.g. by Edens et al., 1976) H. Schaller (unpublished data) mapped some weaker start points of RNA synthesis. The mixture of all RNA-polymerase-binding sites protected against pDNA was isolated, 5′ end-labeled and used to prime repair synthesis on fd DNA single strands. The mixture of the extended promoter regions was then cleaved with several different restriction enzymes and separated on polyacrylamide gels. By secondary cleavage with other restriction enzymes most of the resulting fragments could be positioned on the physical map. In addition, partial DNA sequences of nine from eleven binding sites isolated by this approach were determined. All these sites are listed in Fig. 4. The last nucleotide of each sequence in this figure corresponds to the first base of the (complementary) coding

TABLE III

The gene products of filamentous bacteriophages

Gene	Start and stop codon (position)	Function and appearance	No. of amino acid residues and most frequent amino acids (%)	M_r (dalton)	Silent base exchanges fd - f1 fd - M13	T in 3rd pos. (%)	Protein data
II	ATG 6007 (90% start) ATG 6016 (10% start)	initiation of replication of viral strand DNA; nicks the plus strand of RF DNA between nucleotides pos. 5763 and 5764	410 Leu 11.5 Ser 11 Ala 6.8	46260	30 31	45	M_r on SDS gel 46000 daltons (Konings et al., 1975); N-terminal amino acid sequene determined (Meyer et al., 1980)
X	ATG 496 TAA 880	unknown; not detected *in vivo*	111 Ser 10.8 Val 9.0 Asn 7.2 Leu 7.2	12680	10 10	56	12000 dalton protein in coupled transcription-translation systems encoded by f1 DNA <u>HpaII</u> fragment C (pos. 314 - 966)(Model and Zinder, 1974)
V	ATG 843 TAA 1004	single strand specific DNA binding protein; concentration in cell ($\sim 10^5$ copies) determines the fraction of viral DNA that replicates or is packaged into capsid protein	87 Leu 11.5 Val 9.2 Gly 8 Ser 8	9690	7 10	45	amino acid sequence determined (Nakashima et al., 1974)
VII	ATG 1108 TGA 1207	unknown	33 Ile 15.2 Ala 12.1 Glu 12.1 no His	3600	0 0	30	

IX	ATG 1206 TGA 1302	capsid protein few copies per phage	32 Ser 6 Arg 2 no His	3650	0 0	47	amino acid compos. and M_r of C protein agress with data derived from DNA (Simons et al., 1979)
VIII	ATG 1301 double stop TGATAA pos. 1520	major capsid protein; about 1900 copies per phage (Asbeck et al., 1969)	precursor 73 mature 50 Ala 20	prec. 7630 mat. 5240	1 1	26	amino acid sequence of mature protein determined (Asbeck et al., 1969; Nakashima and Koningsberg, 1974)
III	GTG 1579 TAA 2851	minor capsid protein; essential for adsorption to the F-pilus; 5 copies per phage; (Goldsmith and Koningsberg, 1977, corrected by the real M_r)	precursor 424 mature 406 Gly 16	prec. 44640 mat. 42609	14 14	51	amino acid composition and N-terminal amino acid sequence of mature protein determined; M_r from SDS gels 55000-68000 daltons (Goldsmith and Koningsberg, 1977)
VI	ATG 2856 TAA 3193	capsid protein few copies per phage	112 Leu 21.4 Ile 11.6 no Glu no His	12350	4 4	52	amino acid compos. and M_r of D protein agrees with data derived from DNA (Simons et al., 1979)
I	ATG 3197 TAA 4242	morphogenesis	348 Leu 10.9 Lys 9.2 Ser 8.3 Val 7.8	39530	27 28	50	M_r from SDS gels 35000 - 36000 daltons (Model and Zinder, 1974; Konings et al., 1975)
IV	ATG 4221 TAG 5499	morphogenesis	426 Ser 13.6 Val 10.8 Leu 9.6 no Cys	45780	61 70	51	M_r from SDS gels 48000 daltons (Konings et al., 1975)

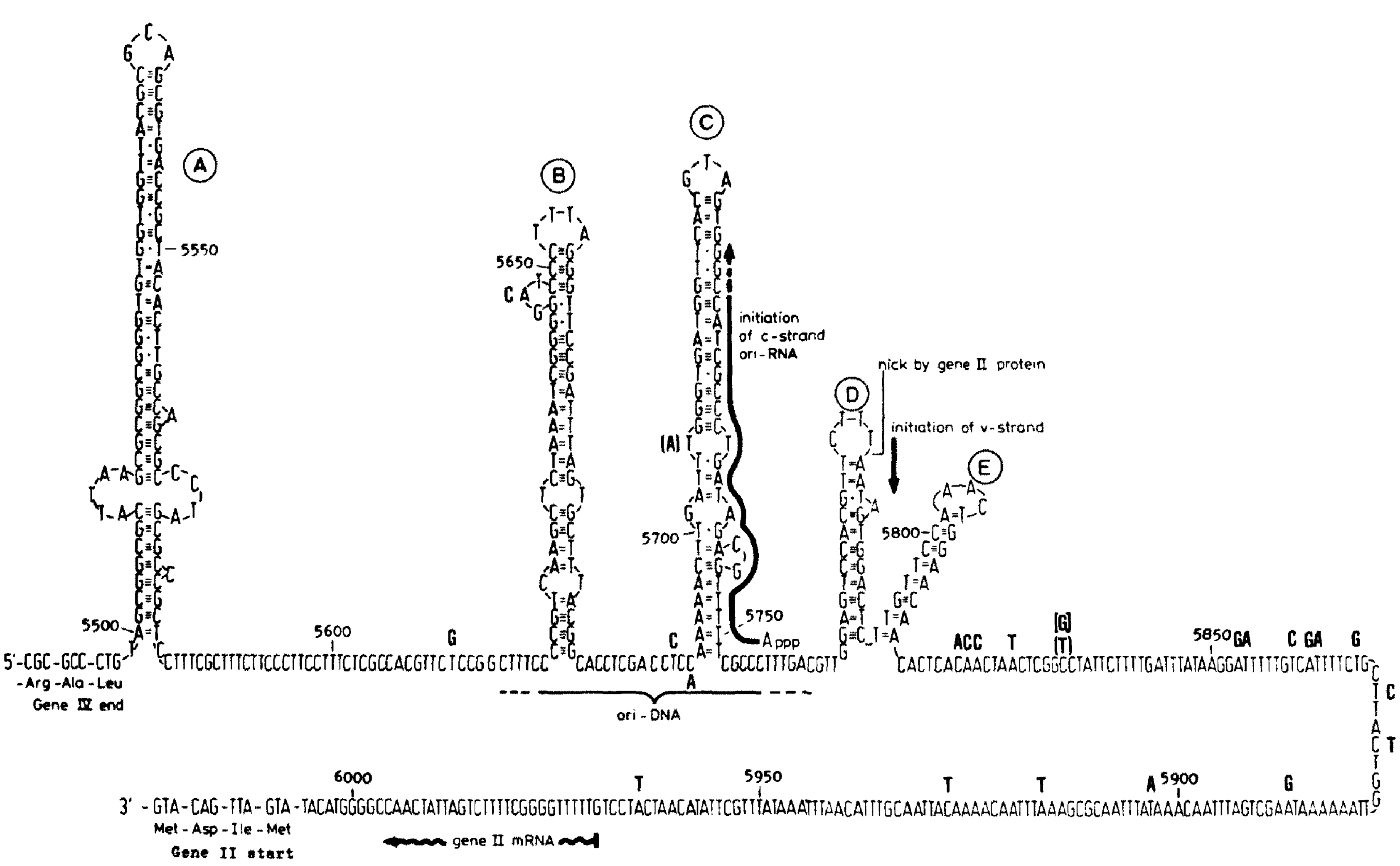
A
B
C
D
E
5500
5550
5600
5650
5700
5750
5800
5850
5900
5950
6000
5'-CGC-GCC-CTG
CTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCTCCGGCTTTCC
CACCTCGACCCTCC
CGCCCTTTGACGTT
CACTCACAACTAACTCGGCCTATTCTTTTGATTTATAAGGATTTTTGTCATTTTCTG
-Arg -Ala -Leu
Gene IV end
ori - DNA
initiation of c-strand ori - RNA
Appp
nick by gene II protein
initiation of v-strand
3' - GTA-CAG-TTA-GTA-TACATGGGGCCAACTATTAGTCTTTTCGGGGTTTTTGTCCTACTAACATATTCGTTTATAAATTTAACATTTGCAATTACAAAACAATTTAAAGCGCAATTTATAAACAATTTAGTCGAATAAAAAATT
Met - Asp - Ile - Met
Gene II start
gene II mRNA
(A)
[G]
(T)
ACC
GA
C GA

Fig. 3. Secondary structure of the DNA in the intergenic region between genes IV and II. The positions of the RNA-polymerase-protected DNA fragment (*ori*-DNA; Schaller et al., 1976), the primer RNA for complementary-strand replication (*ori*-RNA; Geider et al., 1978), and the initiation point of viral (v) strand replication (Meyer et al., 1979) are indicated. Base exchanges in f1 and M13 DNA are marked as described in the legend to Fig. 2.

```
                  -35                             -10        +1
                   .                               .          .
          A        C                           T
        AA  T      TTGACA                         TATAAT     CAT
          T        G                           G              →

        375
                             *                           *
X       TCTTAATCTTTTTGATGCAATTCGCT TTGCTTCTGACTATAATAG  ACAGGGTAAAGACCTGATTTTTGA
                                                           →→→
        5129
           *                                         *
II      ACAAAACATTAACGTTTACAATTTAA ATATTTGCTTATACAATCA  TCCTGTTTTTGGGGCTTTTCTGATTA
                                                            →
        1149
VIII    GATACAAATCTCCGTTGTACTTTGTT  TCGCGCTTGGTATAATCG CTGGGGGTCAAAGATGAGTG
                                                          →→→→
        6194
                  *      *        *
II'     TTTGAATCTTTGCCTACTCATTACTCCGGCATTGCATTTAAAATAT ATGAGGGTTCTAAAA
        4057
IV      TGATAAATTCACTATTGACTCTTCTCAGCGTCTTAATCTAAGCTAT  CGCTATGTTTTCAAGGATT
        764
                        *
V       TTATTAACGTAGATTTTTCCTCCCAACGTCCTGACTGGTATAATGA  GCCAGTTCTTAAAATCGCA
        1494
III     TTAAGAAATTCACCTCGAAAGCAAGC TGATAAACCGATACAATTA AAGGCTCCTTTTGGA
                                                            →
        2710
        *
VI      GGCGCTGGTAAACCATATGAATTTTCTATTGATTGTGACAAAATAA  ACTTATTCCGTGGTGTCTTTGCGTTTC
        3079
I       TAATTCTCCCGTCTAATGCGCTTCCC TGTTTTTATGTTATTCTCT CTGTAAAGGCTGCTATTT
        2318
I'      GGCAAATTAGGCTCTGGAAAGACGCTCGTTAGCGTTGGTAAGATTCAGGATAATTGTAGCTG
        4628
                   *                    * *                    *
IV'     AATTAATAACGTTCGCGCAAAGGATTTAATAAGGGTTGTAGAATTGTTTGTTAAATCTAATACA

                           ----------RNA polymerase protected------------------
```

Fig. 4. Nucleotide sequences of promoter sites in fd DNA. The sequences are aligned with respect to the known initiation nucleotides (→) and the RNA polymerase recognition sites. The rightmost nucleotide in each line corresponds to the first protected base in the pDNA (minus)-strand determined by H. Schaller, as described in the text. Homologies to the consensus sequences around positions −35 and −10 (top line), as compiled by Siebenlist et al. (1980), are underlined. The upper four (strong G-start) promoters have been ordered according to their relative strength (Seeburg et al., 1977). Base exchanges to f1 and M13 DNA are marked by asterisks (see Fig. 2).

strand protected in the RNA polymerase-promoter complex.

The sequences show homologies around positions −10 and −35 to other promoter sites of *E. coli* RNA polymerase (reviews: Rosenberg and Court, 1979; Siebenlist et al., 1980). In the weaker promoters (lower part of Fig. 4) this homology is less pronounced.

In most cases promoters are integrated into the end of the preceding gene. The positions of the polymerase binding sites in front of genes II, X, V, and VII (positions 5940–5980; 400–440; 790–830, and 1170–1210, respectively) coincide with four strong G-start promoters determined by in vitro transcription of restriction fragments with RNA polymerase (Seeburg and Schaller, 1975; Edens et al., 1976). The gene V promoter was positioned incorrectly in previous publications (Schaller et al., 1978; Van Wezenbeek et al., 1980), since a provisional fd sequence was used in the initial interpretation of the mapping of the RNA polymerase binding sites mentioned above. The end of the polymerase protected DNA was determined 105 nucleotides away from the next *Hin*fI cleavage site, which is at position 723 and not at position 741. The thus derived polymerase-binding site shows a perfect Pribnow

TABLE IV

Amber mutants of fd, f1 and M13

Gene	Phage	Name of mutant	Position	Base exchange
II	f1	R124	6349	G→T
V	fd	fd122[a]	906	C→T
	M13	5H1[a] 5H3[a] 5H27[a]	999	C→T
	f1	R13	999	C→T
VII	M13	7H2[b]	1114	C→T
	M13	7H3[b]	1141	C→T
VIII	M13	8H1[c]	1373	G→T
III	M13	3H1[a] 3H4[a]	2017	C→T
	M13	3H5[a]	2473	C→T
VI	f1	R5 R7	3066	C→T
	M13	6H1[a] 6H2[a] 6H3[a] 6H6[a]	3066	C→T
I	M13	1H7[a]	3263	C→T
IV	f1	R143	5265	C→T

[a] Van Wezenbeek et al. (1980); [b] Hulsebos and Schoenmakers (1978); [c] Boeke and Model (1979).

TABLE V

Amino acid exchanges in proteins betweed fd, f1 and M13

Gene	Amino acid pos.	Amino acid			Pos. of exchange in DNA
		fd	f1	M13	
II	249	Glu	Glu	Lys	343
	274	Arg	Ser	Ser	420
VIII	35	Asp	Asp	Asn	1403
III	374	Pro	Leu[a]	Pro	2699[a]
	375	Tyr	Phe	Phe	2702
	378	Gly	Arg	Ser	2710
I	142	His	Asn	His	3620
	164	Val	Ile	Ile	3686
	326	Ile	Leu	Leu	4172
IV	30	Pro	Ser	Pro	4308
	42	Thr	Thr	Ser	4344
	70	Asn	Asp	Asn	4428
	98	Ser	Asn	Asn	4513
	110	Ile	Asn	Asn	4549
	166	Val	Val	Ile	4716

[a] This exchange was observed only in f1 amber mutant R5.

hexamer. A corrected assignment of this promoter in the fd DNA sequence is presented (Siebenlist et al., 1980), but also has to be shifted in the –35 region by one nucleotide, since in M13, as well as in f1, there is a T at position 783 instead of a C. An exchange of the corresponding C in position –32 for a T causes a down mutation in the λp_L promoter (*sex*1 mutant; Kleid et al., 1976). The alignment to the –35 region proposed here results in an equivalent homology pattern as that shown by Siebenlist et al. (1980), with the base exchange at a point of nonhomology. In the II′ promoter, which is a strong RNA start site in vitro, position –33 is converted by a T → A base exchange in f1 and M13 (pos. 6213) into the "ideal" form. There is no mutant known in this position in other promoters, but the potentially different frequency of RNA initiation at this site in fd on the one hand and in f1 and M13 on the other hand is not yet measured.

There are several other changes within the promoters of the filamentous phages, indicated in Fig. 3. All of them concern positions outside the polymerase recognition sites except for a C → G exchange in the –35 region of promoter IV′. It is questionable whether this promoter and the I′ promoter have any function in vivo since these sites were only detected as polymerase binding sites in vitro and not by a transcription product.

The position of the gene VIII promoter was also established by sequence analysis of the gene transcript, which starts with pppG$_4$ at position 1196 (Takanami et al., 1976). Gene IX is encoded by the same mRNA. The ATG start codon lies 10 nucleotides downstream from the beginning of the mRNA. Transcripts probably do not start exclusively at one single position: like the "wobbling" start of gene X mRNA (Nüsslein and Schaller, 1975), a percentage may initiate either one nucleotide before or one after, giving rise to varying numbers of G residues at the 5′ end. Only the longest RNA chains starting with G5 may offer an efficient ribosome-binding site, which perhaps accounts for the low expression of gene IX.

Three RNA polymerase binding sites in front of genes VI, I, and IV (pos. 2740–2780, 3100–3140, and 4080–4120, respectively) confirm the positions of three A-start promoters, also determined by in vitro transcription of restriction fragments (Edens et al., 1976). A further binding site (pos.

1510–1550) overlaps partially with the central termination signal for transcription and defines the position of the gene III promoter. The mRNA probably starts at position 1544 with pppU (M. Takanami, personal communication; Edens et al., 1978).

Sequences preceding the start codons of the genes listed in Fig. 5 show varying degrees of complementarity to the 3′ end of the 16s rRNA (Shine and Dalgarno, 1974). Three ribosome-binding sites of phage f1 were isolated as early as 1974 from ribosome-RNA complexes, and the nucleotide sequence was analysed (Pieczenik et al., 1974). In establishing the DNA sequence it appeared that these sites belonged to the genes V, VI, and VIII.

The start codons for all genes are ATG except for gene III, which starts with GTG, possibly contributing to the low expression rate of this gene. In genes that are efficiently expressed (e.g. genes II, V, and VIII) an A follows the ATG codon, which is in

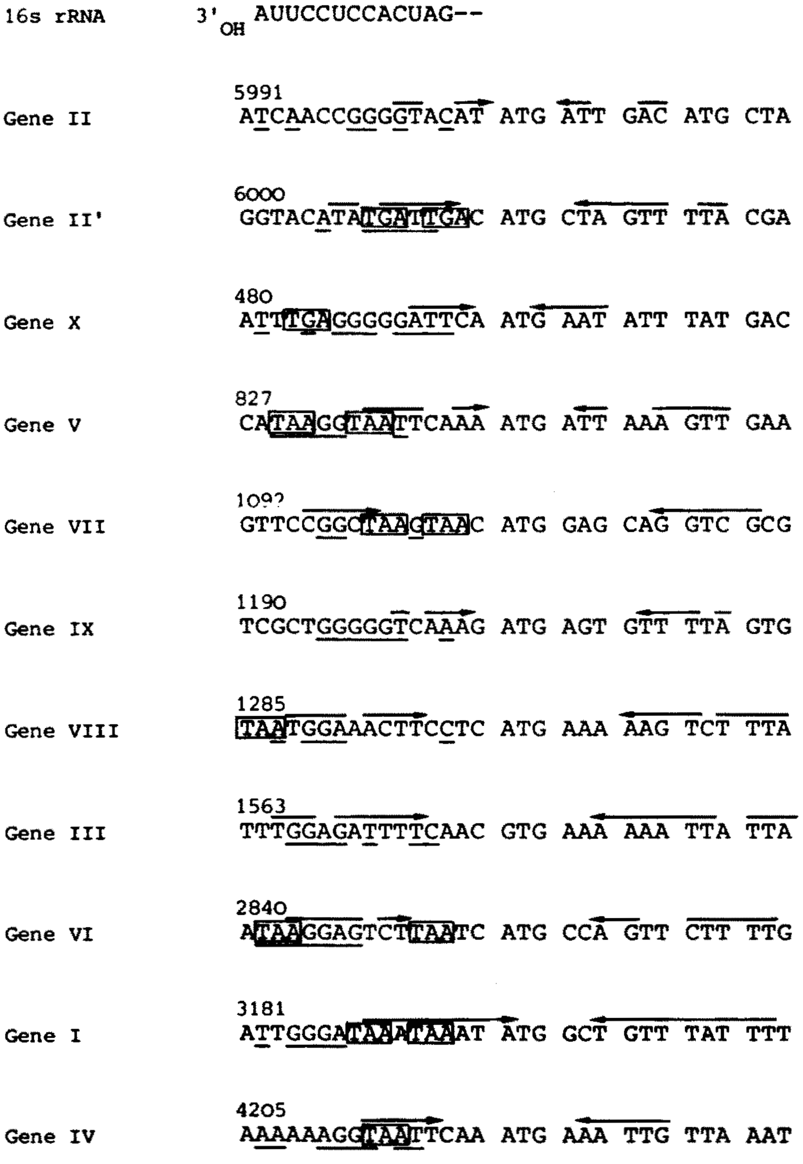
16s rRNA
3'OH AUUCCUCCACUAG--
Gene II
5991
ATCAACCGGGGTACAT ATG ATT GAC ATG CTA
Gene II'
6000
GGTACATATGATTGAC ATG CTA GTT TTA CGA
Gene X
480
ATTTGAGGGGGATTCA ATG AAT ATT TAT GAC
Gene V
827
CATAAGGTAATTCAAA ATG ATT AAA GTT GAA
Gene VII
1092
GTTCCGGCTAAGTAAC ATG GAG CAG GTC GCG
Gene IX
1190
TCGCTGGGGGTCAAAG ATG AGT GTT TTA GTG
Gene VIII
1285
TAATGGAAACTTCCTC ATG AAA AAG TCT TTA
Gene III
1563
TTTGGAGATTTTCAAC GTG AAA AAA TTA TTA
Gene VI
2840
ATAAGGAGTCTTAATC ATG CCA GTT CTT TTG
Gene I
3181
ATTGGGATAAATAAAT ATG GCT GTT TAT TTT
Gene IV
4205
AAAAAAGGTAATTCAA ATG AAA TTG TTA AAT

Fig. 5. Nucleotide sequences of ribosome binding sites in fd DNA. Nucleotides complementary to the 3′-terminus of 16s rRNA (Shine and Dalgarno, 1974) are underlined. Palindrome structures near the start codon are indicated by arrows, and stop signals preceding the start codons are boxed.

agreement with the hypothesis that the fourth base in the f-Met-tRNA anticodon is involved in the formation of the translation-initiation complex (Taniguchi and Weissmann, 1978).

In all these ribosome binding sites a palindrome can be observed more or less evidently (indicated by arrows in Fig. 5), which allows part of the sequence upstream from the start codon to base-pair with the sequence downstream, thus exposing the ATG triplet on top of a small hairpin structure. Such structures were first considered as possible translation recognition signals in other systems (Steitz and Jakes, 1975), but this idea was later rejected (Steitz, 1979). Similar structures are also recognized at other ribosome-binding sites (coat and A proteins of f2, MS2, Qβ, genes C and F of ϕX174, genes *lacI, galE, galT* of *E. coli*) as listed in Steitz (1979).

Stop codons immediately precede or overlap with translational start signals due, primarily, to the close packing of genes in the filamentous phage genome (see above). However, this arrangement may also provide a helper function for translation-promoted re-initiation of translation: The ribosomes stop in a position that allows Shine-Dalgarno base pairing to occur anew.

The completed f1 DNA sequence shows this phage to be closely related to the two other filamentous bacteriophage fd and M13. The small gene products are almost all identical to their various counterparts, whereas the amino acid sequences of the larger proteins diverge from one another by as much as 2%. Regulatory elements also vary only slightly in their essential parts. More variable regions lie in the IG between highly conserved segments, the latter probably representing structurally functional domains. Such variable regions can, in part, be deleted or replaced by heterologous DNA, which allowed the filamentous phage to be used as efficient cloning vehicles (Messing et al., 1977; Herrmann et al., 1980).

ACKNOWLEDGEMENTS

We thank Dr. H. Schaller for discussion, Dr. M. Takanami for unpublished data, R. Sommer for help with the computer processing of sequence data and P. Shaw for help with the manuscript. This work was supported by a grant from the Deutsche Forschungsgemeinschaft (Scha 134/8).

REFERENCES

Asbeck, F., Beyreuther, K., Köhler, H., von Wettstein, G. and Braunitzer, G.: Virusproteine, IV. Die Konstitution des Hüllproteins des Phagen fd. Hoppe Seylers Z. Physiol. Chem. 350 (1969) 1047–1066.

Auerswald, E.A.: Struktur des Transposons Tn*5* und Analyse seiner Integration und Exzision im Bakteriophagen fd. Ph. D. Thesis, University of Heidelberg, 1979.

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) 4495–4504.

Blakesley, R.W. and Wells, R.D.: Single-stranded DNA from ϕX174 and M13 is cleaved by certain restriction endonucleases. Nature 257 (1975) 421–422.

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

Cleary, J.M. and Ray, D.S.: Replication of the plasmid pBR322 under the control of a cloned replication origin from the single-stranded DNA phage M13. Proc. Natl. Acad. Sci. USA 77 (1980) 4638–4642.

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 (1978) 825–842.

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

Fuchs, C., Rosenvold, E.C., Honigman, A. and Szybalski, W.: Identification of palindromic sequences 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-strand to replicative form conversion. Proc. Natl. Acad. Sci. USA 75 (1978) 645–649.

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

Gray, C.P., Sommer, R., Polke, C., Beck, E. and Schaller, H.: Structure of the origin of DNA replication of bacteriophage fd. Proc. Natl. Acad. Sci. USA 75 (1978) 50–53.

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. 177 (1980) 231–242.

Herrmann, R., Neugebauer, K., Zentgraf, H. and Schaller, H.: Transposition of a DNA sequence determining kanamycin resistance into the single-stranded genome of bacteriophage fd. Mol. Gen. Genet. 159 (1978) 171–178.

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.

Johnson, R.A. and Walseth, T.F.: The enzymatic preparation of [α-^{32}P]ATP, [α-^{32}P]GTP, [^{32}P]cAMP, [^{32}P]cGMP and their use in the assay of adenylate and guanylate cyclases and cyclic nucleotide phosphodiesterases. Adv. Cycl. Nucleotide Res. 10 (1979) 135–167.

Kleid, D., Humayun, Z., Jeffrey, A. and Ptashne, M.: Novel properties of a restriction endonuclease isolated from *Haemophilus parahaemolyticus*. Proc. Natl. Acad. Sci. USA 73 (1976) 293–297.

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

Ling, V.: Fractionation and sequences of the large pyrimidine oligonucleotides from bacteriophage fd DNA. J. Mol. Biol. 64 (1972) 87–102.

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

Marvin, D.A. and Hohn, B.: Filamentous bacterial viruses. Bacteriol. Rev. 33 (1969) 172–209

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

Maxam, A.M. and Gilbert, W.: Sequencing end-labeled DNA with base-specific chemical cleavages, in Grossman, L. and Moldave, K. (Eds.), Methods in Enzymology, Vol. 65. Academic Press, New York, 1980, pp. 499–560.

Messing, J., Gronenborn, B., Müller-Hill, B. and Hofschneider, P.H.: Filamentous coliphage M13 as a cloning vehicle. Insertion of a *Hin*dII fragment of the *lac* regulatory region in M13 replicative form in vitro. Proc. Natl. Acad. Sci. USA 74 (1977) 3642–3646.

Meyer, T.F., Beyreuther, K. and Geider, K.: Recognition of two initiation codons for the synthesis of phage fd gene II protein. Mol. Gen. Genet. 180 (1980) 489–494.

Meyer, T.F., Geider, K., Kurz, Ch. and Schaller, H.: Cleavage site of bacteriophage fd gene II protein in the origin of viral strand replication. Nature 278 (1979) 365–366.

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

Nakashima, Y., Dunker, A.K., Marvin, D.A. and Koningsberg, W.: The amino acid sequence of a DNA binding protein, the gene 5 product of fd filamentous bacteriophage. FEBS Lett. 43 (1974) 125.

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

Nüsslein, C. and Schaller, H.: Stabilisation of promoter complexes with a single ribonucleoside triphosphate. Eur. J. Biochem. 56 (1975) 563–569.

Oertel, W. and Schaller, H.: A new approach to the sequence analysis of DNA. FEBS Lett. 27 (1972) 316–320.

Ohmori, H., Tomizawa, J. and Maxam, A.M.: Detection of 5-methylcytosine in DNA sequences. Nucl. Acids Res. 5 (1978) 1479–1485.

Osterburg, G. and Sommer, R.: Computer support of DNA sequence analysis. Comp. Prog. Biomed. 13 (1981) 101–109.

Pieczenik, 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.

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 (1977) 4219–4223.

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

Ray, D.S.: Replication of filamentous bacteriophages, in Fraenkel-Conrat, H. and Wagner, R. (Eds.), Comprehensive Virology, Vol. 7. Plenum, New York, 1977, pp. 105–178.

Roberts, R.J., Myers, P.A., Morrison, A. and Murray, K.: A specific endonuclease from *Haemophilus haemolyticus*. J. Mol. Biol. 103 (1976) 199–208.

Rosenberg, M. and Court, D.: Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. 13 (1979) 319–353.

Sanger, F., Air, G., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchinson, C.A., Slocombe, P.M. and Smith, M.: Nucleotide sequence of bacteriophage ϕX174 DNA. Nature 265 (1977) 687–695.

Sanger, F., Coulson, A.R.: A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94 (1975) 441–448.

Sanger, F., Donelson, J.E., Coulson, A., Kössel, H., Fischer, D.: Use of DNA Polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage f1. Proc. Natl. Acad. Sci. USA 70 (1973) 1209–1213.

Sanger, F., Donelson, J.E., Coulson, A., Kössel, H., Fischer, D.: Determination of a nucleotide sequence in bacteriophage f1 DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 90 (1974) 315–333.

Schaller, H.: The intergenic region and the origins for filamentous phage DNA replication. Cold Spring Harbor Symp. Quant. Biol. 43 (1979) 401–408.

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

Schaller, H., Gray, C.P. and Herrmann, R.: Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd. Proc. Natl. Acad. Sci. USA 72 (1975) 737–741.

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

Scherer, G.E.F., Walkinshaw, M.D. and Arnot, S.: A computer aided oligonucleotide analysis provides a model for RNA polymerase-promoter recognition. Nucl. Acids Res. 5 (1978) 3759–3773.

Seeburg, P.H., Nüsslein, C. and Schaller, H.: Interaction of RNA polymerase with promoters from bacteriophage fd. Eur. J. Biochem. 74 (1977) 107–113.

Seeburg, P.H. and Schaller, H.: Mapping and characterization of promoters in phage 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., Simpson, R.B. and Gilbert, W.: *E. coli* RNA polymerase interacts homologously with two different promoters. Cell 20 (1980) 269–281.

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.

Smiths, 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. 112 (1980) 309–321.

Sommer, R.: Klonierung und Analyse der für die Replikation und Phagenmorphogenese essentiellen DNA Strukturen des Bakteriophagen fd. Ph. D. Thesis, University of Heidelberg, 1981.

Steitz, J.A.: Genetic signals and nucleotide sequences in messenger RNA, in Goldberg, R.F. (Ed.), Biological regulation and development, I. Gene Expression. Plenum, New York, 1979, pp. 349–399.

Steitz, J.A. and Jakes, K.: How ribosomes select initiator regions in mRNA: Base pair formation between the 3'-terminus of 16s rRNA and the mRNA during initiation of protein synthesis in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 72 (1975) 4734–4738.

Suggs, S.V. and Ray, D.S.: Nucleotide sequence of the origin for bacteriophage M13 DNA replication. Cold Spring Harbor Symp. Quant. Biology 43 (1978) 379–388.

Sugimoto, K., Sugisaki, H., Okamoto, T. and Takanami, M.: Studies on bacteriophage fd DNA, III. Nucleotide sequence preceding the RNA start site on a promoter-containing fragment. Nucl. Acids Res. 2 (1975) 2091–2100.

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

Takanami, M., Sugimoto, H., Sugisaki, H. and Okamoto, T.: Sequence of the promoter for the coat protein gene of the bacteriophage fd. Nature 260 (1976) 297–302.

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 Wezenbeek, P.M.G.F., Hulsebos, T.J.M. and Schoenmakers, J.G.G.: Nucleotide sequence of the filamentous bacteriophage M13 genome: Comparison with phage fd. Gene 11 (1980) 129–148.

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.